XI. Magyar Számítógépes Nyelvészeti Konferencia

MSZNY 2015

Szerkesztette:

Tanács Attila
Varga Viktor
Vincze Veronika

Szeged, 2015. január 15-16.
http://rgai.inf.u-szeged.hu/mszny2015

ISBN: 978-963-306-359-0

Szerkesztette: Tanács Attila, Varga Viktor és Vincze Veronika
{tanacs, vinczev}@inf.u-szeged.hu
viktor.varga.1991@gmail.com

Felelős kiadó: Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.

Nyomtatta: JATEPress
6722 Szeged, Petőfi Sándor sugárút 30–34.

Szeged, 2015. január

Előszó

Idén immár tizenegyedik alkalommal rendezzük meg Szegeden a Magyar Számítógépes Nyelvészeti Konferenciát 2015. január 15-16-án. A konferencia fő célkitűzése a kezdetek óta állandó maradt: a rendezvény fő profilja a nyelv- és beszédtechnológia területén végzett legújabb, illetve folyamatban levő kutatások eredményeinek ismertetése és megvitatása, mindemellett lehetőség nyílik különféle hallgatói projektek, illetve ipari alkalmazások bemutatására is.

Nagy örömömre szolgál, hogy a hagyományoknak megfelelően a konferencia nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A konferenciafelhívásra idén is nagy számban beérkezett tudományos előadások közül a programbizottság 36-ot fogadott el az idei évben, így 24 előadás, 8 poszter-, illetve 4 laptopos bemutató gazdagítja a konferencia programját. A programban a magyar számítógépes nyelvészet rendkívül széles skálájáról találhatunk előadásokat a számítógépes szintaxis és szemantika területétől kezdve a véleménykinyerésen át a klinikai szövegek számítógépes feldolgozásáig.

Nagy örömet jelent számomra az is, hogy Tihanyi László, az Európai Bizottság gépi fordítással foglalkozó szakértője, elfogadta meghívásunkat, és plenáris előadása is a konferenciaprogram szerves részét képezi.

Ahogy az már hagyománnyá vált, idén is tervezzük a „Legjobb Ifjú Kutatói Díj” odaítélését, mellyel a fiatal korosztály tagjait kívánjuk ösztönözni arra, hogy kiemelkedő eredményekkel járuljanak hozzá a magyarországi nyelv- és beszédtechnológiai kutatásokhoz.

Ezúton szeretném megköszönni a Neumann János Számítógép-tudományi Társaságnak szíves anyagi támogatásukat.

Szeretnék köszönetet mondani a programbizottságnak: Vámos Tibor programbizottsági elnöknek, valamint Alberti Gábor, Kornai András, László János, Németh Géza, Prószéky Gábor és Váradi Tamás programbizottsági tagoknak. Szeretném továbbá megköszönni a rendezőbizottság és a kötetszerkesztők munkáját is.

Csirik János, a rendezőbizottság elnöke

Szeged, 2015. január

Tartalomjegyzék

I. Fordítás

II. Szintaxis, szemantika

III. Morfológia, korpusz

VII. Poszterbemutatók

VIII. Laptopos bemutatók

IX. Angol nyelvű absztraktok

I. Fordítás

Gépi fordítás minőségének becslése referencia nélküli módszerrel

Yang Zijian Győző[1], Laki László[1,2], Prószéky Gábor[1,2,3]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport
[3] MorphoLogic
{yang.zijian.gyozo, laki.laszlo,
proszeky.gabor}@itk.ppke.hu

Kivonat: A gépi fordítás elterjedésével a gépi fordítók kimenetének automatikus kiértékelése is középpontba került. A hagyományos kiértékelési módszerek egyre kevésbé bizonyultak hatékonyaknak. A legfőbb probléma a hagyományos módszerekkel, hogy referenciafordítást igényelnek. A referenciafordítás előállítása idő- és költségigényes, ezért nem tudunk velük valós időben kiértékelni, és a kiértékelés minősége erősen függ a referenciafordítás minőségétől. A jelen kutatás célja, hogy olyan minőségbecslő módszert mutasson be, ami nem igényel referenciafordítást, tud valós időben kiértékelni és magasan korrelál az emberi kiértékeléssel. Az új módszer a QuEst, ami két modulból áll: tulajdonságkinyerés és modelltanítás. A tulajdonságok kinyerése során a QuEst különböző szempontok alapján minőségi mutatószámokat nyer ki a forrás- és a célnyelvi mondatokból. Majd a kinyert mutatók, illetve regressziós modell segítségével a QuEst emberi kiértékeléssel tanítja be a minőségbecslő modellt. A rendszer a betanított minőségbecslő modellel képes valós időben kiértékelni, nem használ referenciafordítást és nem utolsó sorban, mivel emberi kiértékeléssel tanított, magasan korrelál az emberi kiértékeléssel.

1 Bevezetés

Hogyan mérjük a gépi fordítás minőségét? A gépi fordítás széles körben elterjedt a hétköznapokban. Azonban a legtöbb gépi fordító minősége megbízhatatlan. Ezért egyre több helyen merül fel igényként a gépi fordítás minőségének becslése. Elsősorban vállalati és kutatási környezetben van rá nagy szükség. Cégek esetében igen nagy segítséget tud nyújtani egy minőségi mutató a gépi fordítás utómunkáját végző szakemberek számára. Másik alkalmazása a gépi fordító rendszerek kimenetének vegyítése. Egy helyes minőségbecsléssel több gépi fordítást tudunk összehasonlítani és a jobb fordítást kiválasztva javíthatjuk a végső fordítás minőségét. Végül, de nem utolsó sorban, ismerve a fordítás minőségét ki tudjuk szűrni a használhatatlan fordításokat, illetve figyelmeztetni tudjuk a végfelhasználót a megbízhatatlan szövegrészletekre.

A gépi fordítás minőségének helyes becslése nem könnyű feladat. A hagyományos módszerek legnagyobb problémája, hogy referenciafordítást igényelnek, amelynek létrehozása igen drága és időigényes. Ezek a módszerek nem tudnak valós időben

kiértékelni és mivel ember által fordított referenciafordítás alapján értékelnek, a minőség jelentős mértékben függ a fordítás minőségétől.

A jelen kutatás ezekre a problémákra keres megoldást. A cikk egy olyan módszert mutat be, ami nem használ referenciafordítást, képes valós időben kiértékelni és magasan korrelál az emberi kiértékeléssel.

2 Gépi fordítás kiértékelő módszerek

2.1 Referenciafordítással történő kiértékelés

Kétféle módszert különböztetünk meg a gépi fordítás minőségének kiértékeléséhez: referenciafordítással történő és referenciafordítás nélküli kiértékelés.

Referenciafordítással történő kiértékelésre több módszer is rendelkezésünkre áll. A kiértékeléshez szükség van referenciamondatokra, melyeket emberek fordítottak le a forrásnyelvi korpusz alapján, majd a rendszer összehasonlítja a referenciamondatokat a gépi fordító által lefordított mondatokkal. Fontosabb referenciafordítással történő kiértékelő módszerek:

A BLEU (BiLingual Evaluation Understudy) [3] az egyik legnépszerűbb kiértékelő módszer. A BLEU azt vizsgálja, hogy a gépi fordító által lefordított mondatokban szereplő szavak és kifejezések mennyire illeszkednek pontosan a referenciafordításhoz. Az algoritmus az n-gramokból számolt értékek súlyozott átlagát adja eredményül. A módszer előnye, hogy olcsó és gyors. Hátránya, hogy nem érzékeny a szórendi átalakításokra.

Az OrthoBleu algoritmus [2] a BLEU algoritmus elméletén alapszik. A különbség, hogy amíg a BLEU szavakat kezel, addig az OrthoBleu karakterek szintjén keresi az egyezést. Ez a módszer a ragozásos nyelveknél különösen előnyös, hiszen ha két szónak csak a toldaléka különbözik, a BLEU két külön szónak kezeli, és nem talál egyezést a két szó között, az OrthoBleu ezzel szemben a karakterek szintjén sokkal több egyezést talál.

A NIST (NIST Metrics for Machine Translation - MetricsMATR) [10] szintén a BLEU módszeren alapul, de pontosabb közelítést eredményez nála. Minden fordítási szegmenshez megadott módszerek alapján két független bírálatot rendelnek, majd ebből a két értékből állítják fel a végső pontszámot, amit hozzárendelnek minden fordítási szegmenshez. A NIST nem a referenciafordítást használja, hanem ezeket a bírálatok által kiszámolt pontszámokat. A NIST a szegmensekre számolt pontokból átlagot és súlyozott átlagot számol, majd ezek kombinálásával kiad egy dokumentum szintű pontszámot, ezután pedig a dokumentum szintű pontszámokkal végez rendszerszintű kiértékelést. A NIST mérték a korreláció értéke lesz az így kapott pontszámok és a bírálatok által számolt értékek között.

A TER (Translation Edit Rate / Translation Error Rate) [8] fordítási hibaarányt számol a gépi fordítás és az emberi referenciafordítás között, az alapján, hogy mennyi javítást (szó beszúrása, törlése, eltolása, helyettesítése) kell végezni, majd a javítások számát elosztja a referenciafordítás átlagos hosszával. A TER nem kezeli a szemantikai problémákat, mert a gépi fordítás csak azt számolja ki, hogy mennyi az eltérés a

referenciafordítás és a gépi fordítás között. De közben lehet, hogy kevesebb javítással létrehozható olyan mondat, ami jelentésben megegyezik a referenciafordítással. Erre a problémára dolgozták ki a HTER (Human-targeted Translation Edit Rate / Human-targeted Translation Error Rate) módszert. A HTER módszer során célnyelvi anyanyelvű embereket kértek fel, hogy minimális lépéssel javítsák ki a gépi fordító által generált mondatokat úgy, hogy megegyezzen a jelentése a referenciamondattal. Majd az így keletkezett új referenciamondatra számolják ki a TER értéket.

2.2 Referenciafordítás nélkül történő kiértékelés

Az eddigi módszerek mind referenciafordítást igényelnek. Hátrányuk, hogy óriási emberi erőforrást igényelnek, továbbá nincsen lehetőség futási időben kiértékelni a fordítást. A referenciafordítás nélküli kiértékelő módszereket más néven minőségbecslésnek hívják. A minőségbecslés egy felügyelet nélküli automatikus kiértékelő módszer. Alapvetően statisztikai módszerekkel közelítik a problémát. A NAACL 2012 Seventh Workshop On Statistical Machine Translation keretében kiadott osztott feladatra [7] mutattak be egy teljesen újszerű, referenciafordítás nélkül történő kiértékelő módszert. Az új módszer amellett, hogy nem igényel referenciafordítást, képes futási időben kiértékelni a fordítás minőségét, továbbá a módszer segítségével minőségi mutatót adhatunk az olvasó és az utójavítást végző ember számára. A módszert a Lucia Specia által vezetett QUEST [5] és a QTLaunchPad [12] projekt keretében dolgozták ki. A két projekt közös terméke a QuEst keretrendszer [4]. A QuEst keretrendszer megvalósítja a referenciafordítás nélküli kiértékelést.

3 QuEst

A referenciafordítás nélküli kiértékeléséhez a QuEst (Quality Estimation) [6] keretrendszert kutattuk illetve használtuk fel, aminek segítségével készítettünk egy működő, referenciafordítás nélküli angol–magyar minőségbecslő és kiértékelő rendszert.

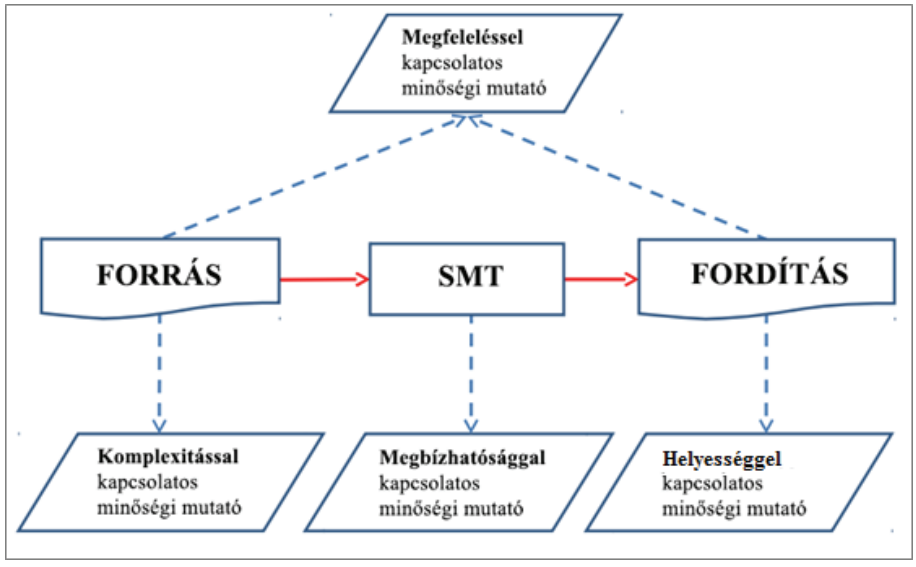
A QuEst mind a forrásnyelvi, mind a célnyelvi szövegből számtalan tulajdonságot tud kiértékelni, a nyelvfüggetlen tulajdonságoktól a nyelvspecifikus tulajdonságokig széles körben, így nem csak a fordítás pontosságára, hanem a mondat helyességére és egyéb problémákra is tud megoldást nyújtani, amelyekre más kiértékelő, mint a BLEU vagy NIST nem képesek. A QuEst keretrendszerben lévő, nyelvtől független tulajdonságok kiértékeléséhez készült funkciók felhasználhatóak a magyar fordítás kiértékelésére. A nyelvspecifikus tulajdonságok kiértékeléséhez viszont magyar nyelvre jellemző kiegészítő eszközökkel kell bővíteni a rendszert. A QuEst keretrendszer JAVA illetve Python nyelven írták, szükséges hozzá a JAVA környezet és Python osztálykönyvtárak. Két főmodulból áll: *tulajdonságkinyerő* modul és *modelltanító* modul.

3.1 A tulajdonságok kinyerése

A tulajdonságok kinyeréséhez (Feature Extraction / Feature Sets) a QuEst-nek a forrásmondatokra illetve a gépi fordító által lefordított mondatokra van szüksége. Mivel a QuEst mondatokkal dolgozik, itt a szegmens egy mondatot jelöl. A QuEst nyelvtől függő és nyelvtől független tulajdonságokat is ki tud értékelni. A nyelvtől független tulajdonságok bármilyen nyelvre használhatóak, viszont a nyelvtől függő tulajdonságok előállításához nyelvspecifikus eszközökre is szükség van, mint például szófaji egyértelműsítő.
Az 1. ábrán azt láthatjuk, hogy a QuEst többféle típusú tulajdonságokat is ki tud értékelni: *megfelelés* (adequacy), *komplexitás* (complexity), *megbízhatóság* (confidence) és *helyesség* (fluency).

A QuEst a kinyert tulajdonságokból minőségi mutatószámokat számol minden szegmensre, így kapunk egy táblázatot, amiben a sorok az egyes szegmensek, az oszlopok a tulajdonságok.

A tulajdonságok kiértékelésére számtalan lehetőség nyílik, de nem biztos, hogy mindegyik tulajdonság releváns a minőségbecslés szempontjából. Lucia Specia kutatása [1] alapján, az angol–spanyol nyelvpárra egy 17 alaptulajdonságból (baseline) álló készletet állítottak össze. Ezek a tulajdonságok a leginkább relevánsak a minőség szempontjából. További tulajdonságok hozzáadásával nem javult jelentősen a minőség. Ebből az következik, hogy nem az a cél, hogy minél több tulajdonságot kiértékeljünk, hanem „a kevesebb néha több” elv alapján, a feladat: minőség szempontjából releváns tulajdonságokat kell keresni.



1. ábra. QuEst által kezelhető tulajdonságok típusai.

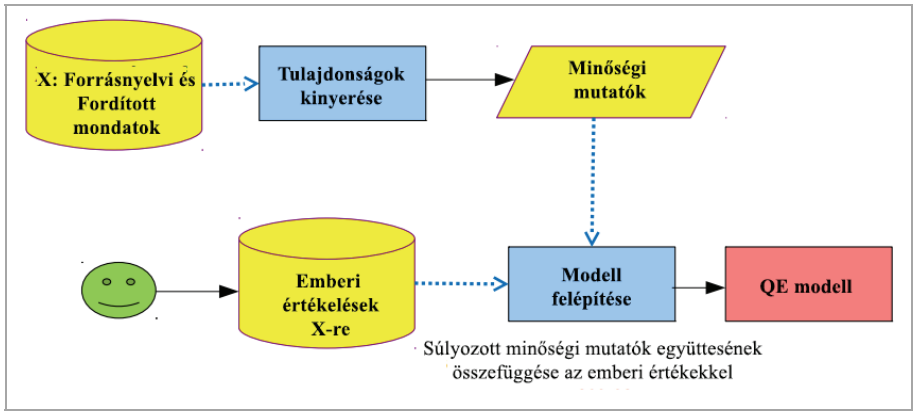
3.2 A modell felépítése

A QuEst másik főmodulja a modell felépítése, ami két részből áll: *tanulás* és *becslés*.

A tanuláshoz szükségünk van egy tanítóhalmazra. A tanítóhalmaz tartalmazza a forrásszöveget, a gépi fordító által lefordított szöveget és emberi értékeléseket. Az emberi értékelés úgy készül, hogy a gépi fordító által lefordított mondatokat emberi szakértők pontozzák két szempont alapján:

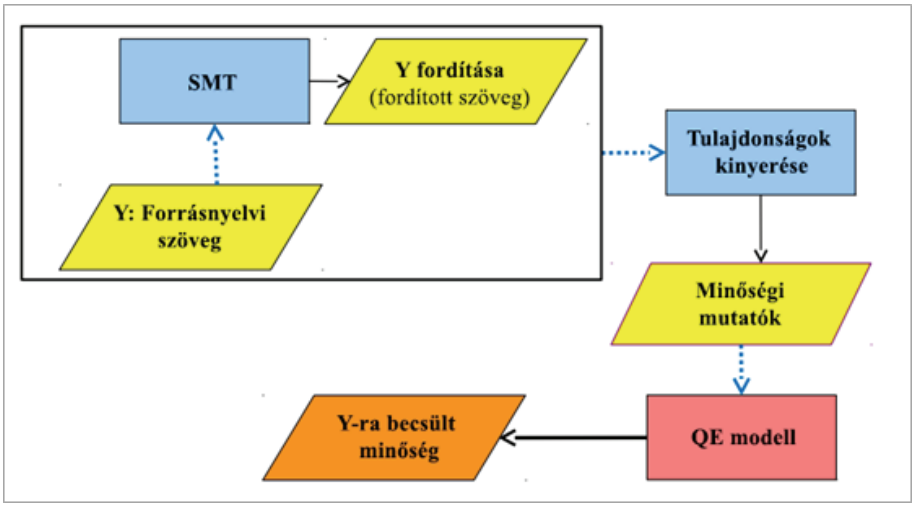
- *megfelelés* (adequacy): a lefordított célnyelvi szöveget értékeli 1–5 pontos skálán, az alapján, hogy mennyire pontos a fordítás a forrásnyelvi mondathoz képest.
- *helyesség* (fluency): a lefordított célnyelvi szöveget értékeli 1–5 pontos skálán, az alapján, hogy mennyire helyes a célnyelvi mondat.

A QuEst a tulajdonságkinyerő modell mutatószámai és az emberi értékek alapján regresszió modellel betanítja a minőségbecslő modellt. A 2. ábrán látható a tanulás folyamata.



2. ábra. Minőségbecslő modell tanításának folyamata.

A QuEst a gépi fordító által generált kimenetére – a tulajdonság kiértékelővel – mutatószámokat számol. Ezután a mutatószámokkal és az emberi értékelésekkel felépíti a kiértékeléshez szükséges modellt. Majd a tanulás során betanított modell segítségével tudja az új bemeneti mondatok minőségét megbecsülni. A minőség becslésének folyamatában már nincsen szükség emberi értékelésre. A 3. ábrán látható a becslés folyamata.



3. ábra. Minőségbecslő modell becslésének folyamata.

4 Módszerek bemutatása

A kiértékeléshez lefordított mondatokat vettünk. A forrásnyelv angol, a célnyelv magyar. A mondatokat négy különböző gépi fordítóval (Google, Bing, MetaMorpho, MOSES) lettek lefordítva, illetve a tanítóanyagban szerepel még ember által lefordított mondat is. Majd betanítottuk és kiértékeltük a QuEst keretrendszerrel. A QuEst kiértékelés minőségének mérésére a MAE (1) (Mean Absolute Error – Átlagos abszolút eltérés), RMSE (2) (Root Mean Squared Error – Átlagos négyzetes eltérés gyöke) [13] és Pearson-féle korreláció értékeket használtunk.

$$MAE = (1/N) * \sum | H(si) - V(si) | \quad (1)$$

$$RMSE = \sqrt{((1/N) * \sum (H(si) - V(si))2)} \quad (2)$$

4.1 Az emberi értékelés létrehozása

A QuEst emberek által értékelt pontszámokat használ a tanításhoz, ezért a QuEst működéséhez szükség van emberek által értékelt tanítóhalmazra. Az emberi értékelés pontszámainak létrehozásához készítettünk egy weboldalon elérhető kérdőívet[1]. A kiértékeléshez önkénteseket kértünk fel, akik közép- illetve felsőfokú angoltudással

[1] http://nlpg.itk.ppke.hu/node/65

rendelkeznek. A mostani eredmények 500 kiértékelt mondattal jöttek létre, de a tanítóhalmaz folyamatosan bővül.

Kettő értékelési szempontot vettünk figyelembe: *megfelelés* és *helyesség*. A megfeleléssel azt mértük, hogy a lefordított mondat tartalmilag mennyire adja vissza a forrásnyelvi mondat mondanivalóját. A helyességgel azt mértük, hogy a lefordított mondat szerkezetileg és nyelvtanilag mennyire helyes, mennyire közelít egy anyanyelvi mondathoz. A minőséget 1–5-ig terjedő skálán osztályoztuk [11] (lásd 1. táblázat).

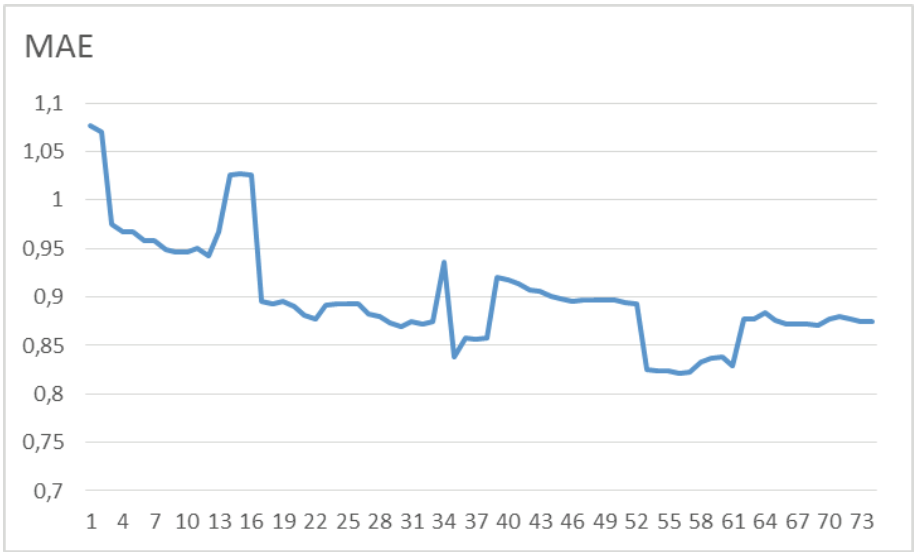
1. táblázat. Értékelési szempontok.

Megfelelés	Helyesség
0 – Nem tudom értelmezni az eredeti (angol) mondatot	
1 – egyáltalán nem jó	1 – érthetetlen a mondat
2 – jelentésben egy kicsit pontos	2 – nem helyes a mondat
3 – közepesen jó a pontosság	3 – több hibát tartalmaz a mondat
4 – jelentésben nagyrészt pontos	4 – majdnem jó a mondat
5 – jelentésben tökéletesen pontos	5 – hibátlan a mondat

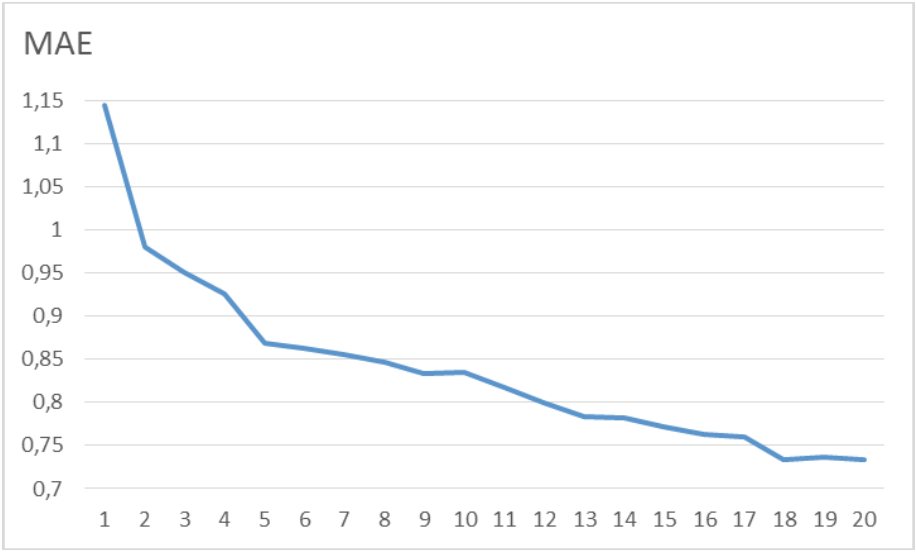
4.2 A tulajdonságok kinyerése

A tulajdonságok kinyeréséhez a QuEst keretrendszert használtunk. A Lucia Specia 2013-as cikkében [1], a QuEst kutatás során kiértékeltek több mint 160 tulajdonságot, de ami igazán releváns, az csak 17 tulajdonság volt az angol–spanyol nyelvpárra. A feladat megtalálni az angol–magyar gépi fordító minőségének kiértékelése szempontjából releváns tulajdonságokat. Elsőként az angol–spanyol alaptulajdonságokkal értékeltem ki az angol–magyar mondatpárokra.

Második lépésként, kipróbáltunk további 57 tulajdonságot, majd ezekből a tulajdonságokból kivettük a nem releváns tulajdonságokat. A kiválasztás folyamata: véletlenszerűen megkevertük a 74 tulajdonságot, majd vettük az elsőt és kiértékeltük. Ezután betanítottuk a minőségbecslő modellt a kiértékelés alapján és kiszámoltuk a MAE értéket a teszthalmazra. Ezek után hozzávettük a második tulajdonságot és újra elvégeztük a kiértékelés folyamatát. Majd így tovább egy ciklussal mindig eggyel több tulajdonságot hozzávettünk és kiértékeltük (lásd 4. ábra). Ha a kiértékelés során az újonnan hozzáadott tulajdonság növelte az MAE értéket, eltároltuk, ha nem, akkor elvetettük. Amikor a ciklus a végére ért, elölről kezdtük a folyamatot. A ciklust elvégeztük 15-ször és a végén megvizsgáltuk, hogy melyek azok a tulajdonságok, amelyek legalább 3 alkalommal javították az eredményt. Ezeket a tulajdonságokat összegyűjtöttük, és az egész kiválasztás folyamatot elölről kezdtük a kiválasztott tulajdonsághalmazon. Így a végére maradt 20 tulajdonság, amelyekből nem tudott az algoritmus többet kizárni és ezzel a 20 tulajdonsággal sikerült elérni a legjobb MAE értéket (lásd 5. ábra).



4. ábra. Kiválasztás folyamata: 74 tulajdonsággal számoló ciklus MAE értékei.



5. ábra. Kiválasztás folyamata: 20 tulajdonsággal számoló ciklus MAE értékei.

4.3 A tanulás és a tesztelés

A gépi fordító minőségének becsléséhez a kinyert tulajdonságokkal és az emberi értékekkel a QuEst betanítja a minőségbecslő modellt. A modell teszteléséhez az 500 mondatos tanítóhalmazt bontottuk 80%-20% arányban tanító-, illetve teszthalmazra. Az így létrehozott tanítóhalmazzal betanítottuk SVR (Szupport Vektor Regresszió) [9] módszerével a kiértékelő modellt, majd a betanított modellel megbecsültük a minőségi mutatókat a teszthalmaz minden sorára. Végül a teszthalmazra kiszámolt minőségi mutatók és a teszthalmazra számolt emberi értékek alapján számoltunk MAE, RMSE és Pearson-féle korreláció értékeket.

5 Eredmények

Az optimalizáló algoritmussal egy 20 tulajdonságból álló alapkészletet állítottunk össze angol–magyar nyelvpárra. A 20 tulajdonságra angol–magyar nyelvpárra optimalizált QuEst rendszert 400 mondattal tanítottuk be és 100 mondattal teszteltük. Az alábbi táblázatban láthatóak az általunk optimalizált és javasolt 20 alaptulajdonság eredményei, összehasonlítva az angol–spanyol alaptulajdonság-készlettel kiértékelt eredményeivel, valamint a 74 tulajdonság által kapott eredményekkel.

A 2. táblázat alapján láthatjuk, hogy az angol–magyar nyelvpárra optimalizált 20 alaptulajdonság valóban jobb eredményt adott mind a 17 angol–spanyol nyelvpárra optimalizált alaptulajdonság-készlethez képest, mind a 74 alaptulajdonsághoz képest. Az eredmény alapján a QuEst a 20 alaptulajdonság készlettel körülbelül 18%-os átlag hibamértékkel tudja megközelíteni az emberi értékeket és a korreláció is elég magas (~71%). Az angol–magyar nyelvpárra optimalizált 20 alaptulajdonság-készlet a 3. táblázatban látható.

2. táblázat. Eredmények összehasonlítása.

	20 alaptulajdonság (angol–magyar)	17 alaptulajdonság (angol–spanyol)	74 tulajdonság (angol–magyar)
MAE	**0,7340**	0,9079	0,8746
RMSE	**0,9341**	1,1148	1,0573
Pearson-féle korreláció	**0,7131**	0,5369	0,6154

3. táblázat. A 20 alaptulajdonság angol–magyar nyelvpárra.

Tokenek száma a forrásmondatban.
Tokenek száma a célmondatban.
Átlagos tokenhossz a forrásmondatban.
Forrásmondat perplexitása.

Célmondat perplexitása.
Átlagos száma minden forrásszó fordításának a mondatban (giza küszöb: valószínűség > 0,5).
Átlagos száma minden forrásszó fordításának a mondatban (giza küszöb: valószínűség > 0,2)
Fordítások átlaga minden forrásszóra a mondatban, súlyozva a forrásnyelvi korpuszban lévő minden szó inverz gyakoriságával.
Átlagos unigram gyakoriság a második kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
Átlagos trigram gyakoriság a második kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
Forrásnyelvi korpuszban lévő negyedik kvartilisben lévő forrásszó trigramjának gyakorisága százalékban.
A korpuszban előforduló különböző trigramok százaléka.
A forrásmondatban és a célmondatban lévő kettőspontok számának különbsége abszolút értékben.
A forrásmondatban és a célmondatban lévő pontosvesszők számának különbsége abszolút értékben.
A forrásmondatban és a célmondatban lévő pontosvesszők számának különbsége abszolút értékben, célmondat hosszával normalizálva.
Írásjegyek száma a célmondatban.
Tokenek száma a forrásmondatban, amelyek nem csak a-z betűt tartalmaznak.
Forrásmondatban lévő a–z tokenek százalékának és a célmondatban lévő a–z tokenek százalékának aránya.
Igék százaléka a célmondatban.
Igék százalékának aránya a forrás és a célmondatban.

6 Összefoglalás

A kutatás során felépítettünk egy QuEst keretrendszert, és optimalizáltuk angol–magyar nyelvpárra. A kiértékeléshez szükség volt emberi értékelésekre, amihez készítettünk egy fordításkiértékelő weboldalt.

Az optimalizálás során kipróbáltunk 74 tulajdonságot, amiből felállítottuk az optimalizált 20 tulajdonságból álló alapkészletet angol–magyar nyelvpárra.

A rendszer további tulajdonságok kipróbálásával tovább optimalizálható. Az általunk felépített QuEst keretrendszer megfelelő alapul szolgál a referenciafordítás nélkül történő angol–magyar gépi fordítás kiértékeléséhez és ezen a területen való további kutatásokhoz.

Hivatkozások

1. Beck, D., Shah, K., Cohn, T., Specia, L.: SHEF-Lite: When Less is More for Translation Quality Estimation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation (2013) 337–342
2. FTSK, OrthoBLEU – MT Evalution Based on Orthographic Similarities [Online] Elérhető: http://www.fask.uni-mainz.de/user/rapp/comtrans/d05orthobleu.html. [Hozzáférés dátuma: 2014. december 1.]
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL) (2002) 311–318
4. Specia, L., Shah, K., de Souza, J. G. C., Coh, T.: QuEst – A translation quality estimation framework. In: Proceedings of the 51st ACL: System Demonstrations (2013) 79–84
5. Specia, L.: QuEst – an open source tool for translation quality estimation [Online] Elérhető: http://staffwww.dcs.shef.ac.uk/people/L.Specia/projects/quest.html [Hozzáférés dátuma: 2014. december 1.]
6. Specia, L.: QuEst [Online] Elérhető: http://www.quest.dcs.shef.ac.uk. [Hozzáférés dátuma: 2014. december 1.]
7. Specia, L.: Shared Task: Quality Estimation [Online] Elérhető: http://www.statmt.org/wmt12/quality-estimation-task.html [Hozzáférés dátuma: 2014. december 1.]
8. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (2006) 223–231
9. Welling, M.: Support Vector Regression. University of Toronto (2004)
10. NIST, The NIST 2008: Metrics for MAchine TRanslation" Challenge (MetricsMATR) (2008)
11. Koehn, P.: Statistical Machine Translation. 1st ed. Cambridge University Press, New York, NY, USA (2010)
12. QTLaunchPad, „QTLaunchPad," [Online] Elérhető: http://www.qt21.eu/launchpad [Hozzáférés dátuma: 2014. december 1.]
13. Hyndman, R. J., Koehler, A. B.: Another look at measures of forecast accuracy (2005)

Synonym Acquisition from Translation Graph

Judit Ács

Budapest University of Technology and Economics,
HAS Research Institute for Linguistics
e-mail: judit.acs@aut.bme.hu

Abstract. We present a language-independent method for leveraging synonyms from a large translation graph. A new WordNet-based precision-like measure is introduced.

Keywords: synonyms, translation graph, WordNet, Wiktionary

1 Introduction

Semantically related words are crucial for a variety of NLP tasks such as information retrieval, semantic textual similarity, machine translation etc. Since their construction is very labor-intensive, very few manually constructed resources are freely available. The most notable example is WordNet [4]. WordNet organizes words into synonym sets (*synsets*) and defines several types of semantic relationship between the synsets. Although WordNet has editions in low-density languages, its construction cost keeps these WordNets quite small. One way to overcome the high construction cost is using crowdsourced resources such as Wiktionary [7] for the automatic construction of synonymy networks.

Wiktionary is a rich source of multilingual information, with rapidly growing content thanks to the hundreds or thousands of volunteer editors. A Wiktionary entry corresponds to one word form or expression. Cross-lingual homonymy is dealt with one section per language (e.g. the article *doctor* in the English Wiktionary has sections about the word's usage in different languages: English, Asturian, Dutch, Latin, Romanian and Spanish). Wiktionary also has a rich synonymy network that was leveraged by Navarro et al. [7] but unfortunately they have not made their results publicly available. They also leveraged Wiktionary's translation graph (see Section 2) for extending this network. Their method, the Jaccard similarity of two words' translation links is used as a baseline in this paper. Instead of the synonymy network, we only utilize the translation graph because it is richer and easier to parse.

2 Translation graph

We define the *translation graph* as an undirected graph, where vertices correspond to words or expressions (we shall refer to one vertex as a word even if it is a multiword expression) and edges correspond to the translation relations

between them. We consider the translation relation symmetric for simplicity, thus rendering the translation graph undirected, unlike graphs acquired from lexical definitions such as [3]. Same-language edges are possible, but self-loops are filtered.

Wiktionary is a constantly growing source of information, therefore leveraging it again and again may yield significantly better and richer results. In [1] we developed a tool called *wikt2dict*[1] for extracting translations from more than 40 Wiktionary editions, which we ran on Wiktionary dumps from November 2014 in the present paper. Although wikt2dict supports dozens of languages and the list can easily be extended, we filtered the translation graph to a smaller set of languages. The languages chosen were[2]: English (en), German (de), French (fr), Hungarian (hu), Greek (el), Romanian (ro) and Slovak (sk). The latter three are supported by Altervista Thesaurus, helping us in evaluation. We present the results on two graphs: the 7 language graph of all languages and a subset of it containing only the first four languages (en, de, hu, fr). The full graph has 385,022 vertices and 514,047 edge with 2,67 average degree, the smaller graph has 299,895 vertices and 359,949 edges with 2,4 average degree.

According to our previous measure in [9], translations acquired from Wiktionary are around 90% correct. Most errors are due to parsing errors or the lack of lexicographic expertise of Wiktionary editors. It is a popular method to use a pivot language for dictionary expansion, see [8] for a comparison of such methods. The results are known to be quite noisy due to polysemy and this has been addressed in [9] by accepting only those pairs that are found via several pivots. However, this aggressive filtering method prunes about half of the newly acquired translations especially in the case of low-density languages. By allowing longer paths between two words, the number of candidates greatly increases, and filtering for candidates having at least two paths prunes fewer good results. The longer the path, the worse quality the translation candidates are (see Section 4), therefore we only accept very short paths. Two disjoint paths between vertices constitute a short cycle in the graph.

The main assumption of this paper is that edges on short cycles are very similar in meaning and using longer cycles than 4, prunes fewer results than the simple triangulation. We require the vertices of a cycle to be unique. We assume that same-language edges are synonyms or closely related expressions. We will discuss this relation in Section 4. An example of this phenomenon is illustrated in Figure 1.

There is no polynomial algorithm for finding all cycles in a graph, but given the low average degrees, the extraction of short cycles using DFS is feasible.

The main downside of this method that it is unable to link vertices found in different biconnected components, since they do not have two unique routes between them.

[1] https://github.com/juditacs/wikt2dict

[2] with their respective Wiktionary code

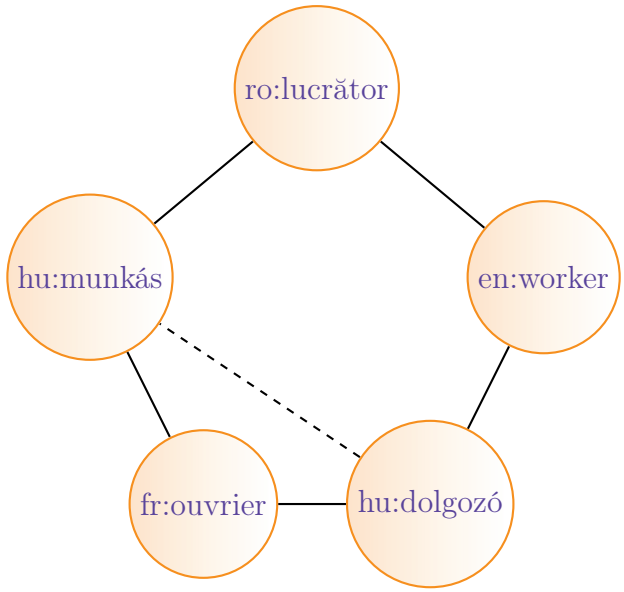


Fig. 1. Example of a pentagon found in the translation graph. The two Hungarian words are synonyms.

3 Results

Finding all k long cycles turned out to be feasible for $k <= 7$ with the given graph size. The baseline method was the Jaccard similarity of two vertices' neighbors:

$$J(w_a, w_b) = \frac{|N_a \cap N_b|}{|N_a \cup N_b|}, \tag{1}$$

where N_a is the set of word w_a's neighbors and N_b is the set of word w_b's neighbors. All pairs with non-zero Jaccard similarity were flagged as candidate pairs. Since every vertex on a square or pentagon is surely at most 2 edges away from each other, the baseline covers all candidates acquired via squares and pentagons. One can expect new results in the main diagonals of hexagons and more from heptagons. It turns out that only heptagons could outperform the baseline in sheer numbers.

We present the results in Table 3.

4 WordNet relation of translations

WordNet covers a wide range of semantic relations between synsets, such as hypernymy, hyponymy, meronymy, holonymy and synonymy itself between lemmas in the same synset. We compared our synonym candidates to WordNet relations

Table 1. Results

Method	Synonym candidates	
	4 languages (en,de,fr,hu)	7 languages (el,sk,ro)
Baseline	398,525	469,071
Squares	25,945	31,819
Pentagons	64,703	84,516
Hexagons	175,313	223,180
Heptagons	411,879	525,106

and found that many candidates correspond to at least one kind of WordNet relation if both words are present in WordNet. Since many words are absent from WordNet (denoted as *OOV*, out-of-vocabulary), these numbers do not reflect the actual precision of the method, but they are suitable for comparing different methods' precision.

The relations considered were:

Synonymy : both words are lemmas of the same synset.
Other : we group other WordNet relations such as hypernymy, hyponymy, holonymy, meronymy, etc. Most candidates in this group are hypernyms.
OOV : we flag a pair of words out-of-vocabulary if at least one of them is absent from WordNet.

We computed the measures on Princeton WordNet as well as on the Hungarian WordNet [5]. The results are illustrated in Figure 2 and Figure 3. In each run, more than half of the candidates have some kind of relation in WordNet. Shorter cycles have a lower no relation ratio than the baseline or longer cycles but they are clearly inferior in the number of pairs generated. We have fewer candidates flagged 'other WN relation' in the Hungarian WordNet, which suggests that – unsurprisingly – the English WordNet has more inter-WN relations. It also suggests that our methods perform worse on a medium-density language such as Hungarian than it does on English.

5 Manual precision evaluation

We performed manual evaluation on a small subset of Hungarian results. Since the baseline covers all pairs generated by $k < 6$ long cycles, we compared the results with and without the baseline. The results are summarized in Table 5.

We also did a manual spot check on the Hungarian pairs flagged OOV or 'other WN relation' when comparing with the Hungarian WordNet. Candidates found in heptagons were excluded. Out of the 100 samples, 53 were synonym, 22 were similar and 25 candidates were incorrect. The results suggest that WordNet coverage by itself is indeed insufficient for precision measurement.

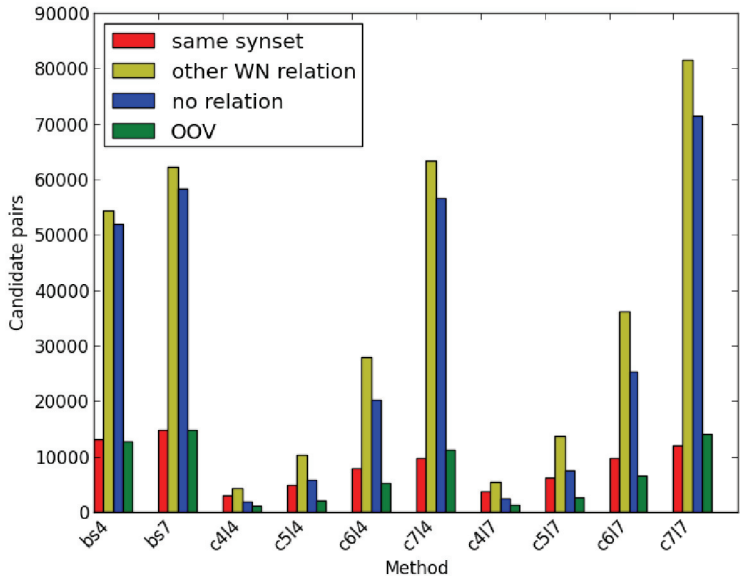


Fig. 2. Types of WN relations between English synonym candidate pairs. Method abbreviations: bs (baseline), c*K*l*N* (K long cycles, N languages).

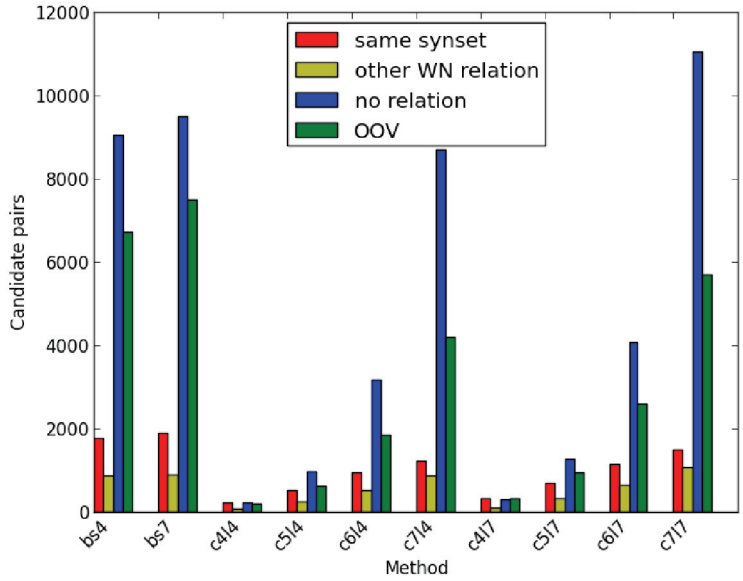


Fig. 3. Types of WN relations between Hungarian synonym candidate pairs. Method abbreviations: bs (baseline), c*K*l*N* (K long cycles, N languages).

Table 2. Results of manual precision evaluation

Data set	Correct	Similar	Incorrect
Baseline disjoint	32	12	56
Cycles disjoint	37	17	46
Intersection	54	25	21

6 Recall

Automatic synonymy acquisition is known to produce very low recall compared to traditional resources, due to the input's sparse structure and the method's shortcomings. We collected synonyms from several resources: WordNet (English and Hungarian), Big Huge Thesaurus (English)[3] and Altervista Thesaurus (English, French, German, Greek, Romanian and Slovak)[4]. We collected 84,069 English, 30,036 Hungarian, 14,444 French, 8,742 German, 8,199 Romanian, 7,868 Greek and 4,624 Slovak synonym pairs. We consider these resources silver standard.

Table 3 illustrates the recall of the baseline, the cycle detection and their combined recall on all resources. It is clear that our methods – while yielding fewer results – outperform the baseline. Although the combined results have the best recall, we have our doubts about their precision. As mentioned earlier, the greatest downside of our method that it is unable to explore synonyms found in different connected components of the graph. This fact reduces the number of possible candidates thus limiting recall. Still, when taking into consideration the fact that some pairs are theoretically impossible to find, the achieved recall remains quite low, although higher the numbers presented by Navarro et al. [7]. In Table 3 we present the non-OOV maximum (when both words of the pair from the silver standard are present in the translation graph) and the recall on pairs where both words are in the same connected component. There is some variance between the languages, most notably, German stands out. This may be due to the German Wiktionary's high quality and the small size of the German silver standard.

The baseline is limited to words at most two edges apart, and its coverage is 0.115 on known words. Cycles over length 5 are able to produce additional pairs, and their combined recall is 0.159 on known words. The two methods combined achieve almost 0.2 but the results become quite noisy.

7 Conclusions

We presented a language-independent method for exploring synonyms in a multilingual translation graph acquired from Wiktionary. We compared the syn-

[3] https://words.bighugelabs.com/
[4] http://thesaurus.altervista.org/

Table 3. Recall of silver standard synonym lists

Method	Language	4 languages			7 languages		
		all	in vocab	same comp	all	in vocab	same comp
Baseline	English	0.07	0.108	0.123	0.076	0.115	0.13
	Hungarian	0.037	0.135	0.147	0.04	0.143	0.154
	French	0.054	0.065	0.077	0.058	0.067	0.078
	German	0.159	0.218	0.247	0.163	0.222	0.247
	Greek	-	-	-	0.045	0.076	0.084
	Romanian	-	-	-	0.034	0.081	0.087
	Slovak	-	-	-	0.019	0.074	0.076
	All	**0.066**	**0.113**	**0.129**	**0.067**	**0.115**	**0.129**
Cycles	English	0.099	0.153	0.174	0.116	0.174	0.197
	Hungarian	0.042	0.155	0.168	0.051	0.182	0.195
	French	0.084	0.101	0.12	0.097	0.113	0.13
	German	0.146	0.2	0.227	0.16	0.218	0.242
	Greek	-	-	-	0.038	0.064	0.07
	Romanian	-	-	-	0.037	0.088	0.093
	Slovak	-	-	-	0.012	0.044	0.045
	All	**0.088**	**0.149**	**0.17**	**0.093**	**0.159**	**0.178**
Combined	English	0.121	0.187	0.213	0.137	0.206	0.233
	Hungarian	0.062	0.225	0.244	0.069	0.249	0.267
	French	0.103	0.123	0.146	0.116	0.135	0.156
	German	0.183	0.252	0.286	0.192	0.261	0.29
	Greek	-	-	-	0.063	0.106	0.117
	Romanian	-	-	-	0.055	0.133	0.141
	Slovak	-	-	-	0.026	0.098	0.101
	All	**0.11**	**0.187**	**0.213**	**0.114**	**0.195**	**0.219**

onym candidates to WordNet and found that most candidates either appear in the same synset or have a very close relationship such as hypernymy in WordNet. Precision was examined both manually and by comparing the candidates to WordNet. Recall was measured against manually built synonym lists. Our method outperforms the baseline in both precision and recall.

Acknowledgment

I would like to thank Prof. András Kornai for his help in theory and Gergely Mezei for his contribution on cycle detection. I would also like to thank my annotators, Gábor Szabó and Dávid Szalóki.

References

1. Ács, J., Pajkossy, K., Kornai, A. Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora,

Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
2. Bird, S. Nltk: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions, pages 69–72. Association for Computational Linguistics, 2006.
3. Blondel, V.D., Senellart, P.P. Automatic extraction of synonyms in a dictionary. vertex, 1:x1 (2011)
4. Fellbaum, C. WordNet. Wiley Online Library (1998)
5. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T. Methods and results of the Hungarian WordNet project. In: Proceedings of the Fourth Global WordNet Conference (GWC-2008) (2008)
6. Miller, G.A. Wordnet: a lexical database for English. Communications of the ACM, Vol. 38., No. 11 (1995) 39–41
7. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, Y., Magistry, P., Huang, C.-R. Wiktionary and NLP: Improving synonymy networks. In: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Association for Computational Linguistics (2009) 19–27
8. Saralegi, X., Manterola, I., San Vicente, I. Analyzing methods for improving precision of pivot based bilingual dictionaries. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 846–856
9. Ács, J. Pivot-based multilingual dictionary building using wiktionary. In: The 9th edition of the Language Resources and Evaluation Conference (2014)

Comparison of Distributed Language Models on Medium-resourced Languages

Márton Makrai

Research Institute for Linguistics of the Hungarian Academy of Sciences
e-mail: makrai.marton@nytud.mta.hu

Abstract. word2vec and GloVe are the two most successful open-source tools that compute distributed language models from gigaword corpora. word2vec implements the neural network style architectures skip-gram and cbow, learning parameters using each word as a training sample, while GloVe factorizes the cooccurrence-matrix (or more precisely a matrix of conditional probabilities) as a whole. In the present work, we compare the two systems on two tasks: a Hungarian equivalent of a popular word analogy task and word translation between European languages including medium-resourced ones e.g. Hungarian, Lithuanian and Slovenian.

Keywords: distributed language modeling, relational similarity, machine translation, medium-resourced languages

1 Introduction

The empirical support for both the syntactic properties and the meaning of a word form consists in the probabilities with that the word appears in different contexts. Contexts can be documents as in latent semantic analysis (LSA) or other words appearing within a limited distance (window) from the word in focus. In these approaches, the corpus is represented by a matrix with rows corresponding to words and columns to contexts, with each cell containing the conditional probability of the given word in the given context. The matrix has to undergo some regularization to avoid overfitting. In LSA this is achieved by approximating the matrix as the product of special matrices.

Neural nets are taking over in many filed of artificial intelligence. In natural language processing applications, training items are the word tokens in a text. Vectors representing word forms on the so called embedding layer have their own meaning: Collobert and Weston [1] trained a system providing state of the art results in several tasks (part of speech tagging, chunking, named entity recognition, and semantic role labeling) with the same embedding vectors. Mikolov et al. [2] trained an embedding with the skip-gram (sgram) architecture, that not only encode similar word with similar vectors but reflects *relational similarities* (similarities of relations between words) as well. The system answers analogical questions. For more details see Section 2.

The two approaches, one based on cooccurrence matrices and the other on neural learning are represented by the two leading open-source tools for computing distributed language models (or simply vector space language models, VSM) from gigaword corpora, GloVe and word2vec respectively. Here we compare them on a task related to statistical machine translation. The goal of the EFNILEX project has been to generate protodictionaries for European languages with fewer speakers. We have collected translational word pairs between English, Hungarian, Slovenian, and Lithuanian.

We took the method of Mikolov et al. [3] who train VSMs for the source and the target language from monolingual corpora, and collect word translation by learning a mapping between these supervised by a seed dictionary of a few thousand items.

Before collecting word translations, we test the models in an independent and simpler task, the popular analogy task. For this, we created the Hungarian equivalent of the test question set by Mikolov et al. [2, 4].[1]

The only related work evaluating vector models of a language other than English on word analogy tasks we know is Sen and Erdogan [5] that compares different strategies to deal with the morphologically rich Turkish language. Application of GloVe to word translations seems to be a novelty of the present work.

2 Monolingual analogical questions

Measuring the quality of VSMs in a task-independent way is motivated by the idea of representation sharing. VSMs that capture something of language itself are better that ones tailored for the task. We compare results in the monolingual and the main task in Section 5.4.

Analogical questions (also called relational similarities [6] or linguistic regularities [2]) are such a measure of merit for vector models. This test has gained popularity in the VSM community in the recent year. Mikolov et al. observe that analogical questions like *good* is to *better* as *rough* is to ... or *man* is to *woman* as *king* is to ... can be answered by basic linear algebra in neural VSMs:

$$\text{good} - \text{better} \approx \text{rough} - \mathbf{x} \tag{1}$$

$$\mathbf{x} \approx \text{rough} - \text{good} + \text{better} \tag{2}$$

So the vector nearest to the right side of (2) is supposed to be *queen*, which is really the case.

We created a Hungarian equivalent of the analogical questions made publicly available by Mikolov et al. [2, 4][2].

[1] For data and else visit the project page `http://corpus.nytud.hu/efnilex-vect`.

[2] More precisely, we follow the main ideas reported in Mikolov et al. [2] and target the sizes of the data-set accompanying Mikolov et al. [4].

Analogical pairs are divided to morphological ("grammatical") and semantic ones. The morphological pairs in Mikolov et al. [2] were created in the following way:

> [We test] base/comparative/superlative forms of adjectives; singular/plural forms of common nouns; possessive/non-possessive forms of common nouns; and base, past and 3rd person present tense forms of verbs. More precisely, we tagged 267M words of newspaper text with Penn Treebank POS tags [7]. We then selected 100 of the most frequent comparative adjectives (words labeled JJR); 100 of the most frequent plural nouns (NNS); 100 of the most frequent possessive nouns (NN POS); and 100 of the most frequent base form verbs (VB).

Table 1. Morphological word pairs

English		Hungarian	
plural	singular	plural	singular
decrease	decreases	lesznek	lesz
describe	describes	állnak	áll
eat	eats	tudnak	tud
enhance	enhances	kapnak	kap
estimate	estimates	lehetnek	lehet
find	finds	nincsenek	nincs
generate	generates	kerülnek	kerül

The Hungarian morphological pairs were created in the following way: For each grammatical relationship, we took the most frequent inflected forms from the Hungarian Webcorpus [8]. The suffix in question was restricted to be the last one. See sizes in Table 2. In the case of opposite, we restricted ourselves to forms with the derivational suffix *-tlan* (and its other allomorphs) to make the task morphological rather then semantic. plural-noun includes pronouns as well.

For the semantic task, data were taken from Wikipedia. For the capital-common-countries task, we choose the one-word capitals appearing in the Hungarian Webcorpus most frequently. The English task city-in-state contains USA cities with the states they are located in. The equivalent tasks `county-center` contains counties (*megye*) with their centers (*Bács-Kiskun – Kecskemét*) currency contains the currencies of the most frequent countries in the Webcorpus. The `family` task targets gender distinction. We filtered the pairs where the gender distinction is sustained in Hungarian (but dropping e.g. *he – she*). We put some relational nouns in the possessive case (*bátyja – nővére*). We note that this category contains the royal "family" as well, e.g. the famous *king – queen*, and even *policeman – policewoman*.

Both morphological and semantic questions were created by matching every pair with every other pair resulting in e.g. $\binom{20}{2}$ questions for family.

Table 2. Sizes of the question sets

	English		Hungarian
	# questions	# pairs	# questions
gram1-adjective-to-adverb	32	992	40
gram2-opposite	812	29	30
gram3-comparative	37	1332	40
gram4-superlative	34	1122	40
gram5-present-participle	33	1056	40
gram6-nationality-adjective	41	1599	41
gram7-past-tense	40	1560	40
gram8-plural-noun	37	1332	40
gram9-plural-verb	30	870	40
capital-common-countries	23	506	20
capital-world	116	4524	166
city-in-state	2467	68	
county-center			19
county-district-center			175
currency	30	866	30
family	23	506	20

Table 3. Semantic word pairs

English		Hungarian	
Athens	Greece	Budapest	Magyarország
Baghdad	Iraq	Moszkva	Oroszország
Bangkok	Thailand	London	Nagy-Britannia
Beijing	China	Berlin	Németország
Berlin	Germany	Pozsony	Szlovákia
Bern	Switzerland	Helsinki	Finnország
Cairo	Egypt	Bukarest	Románia

Table 4. Analogical questions

English				Hungarian			
Athens	Greece	Baghdad	Iraq	Budapest	Magyarország	Moszkva	Oroszország
Athens	Greece	Bangkok	Thailand	Budapest	Magyarország	London	Nagy-Britannia
Athens	Greece	Beijing	China	Budapest	Magyarország	Berlin	Németország
Athens	Greece	Berlin	Germany	Budapest	Magyarország	Pozsony	Szlovákia
Athens	Greece	Bern	Switzerland	Budapest	Magyarország	Helsinki	Finnország
Athens	Greece	Cairo	Egypt	Budapest	Magyarország	Bukarest	Románia

3 Word translations with vector models

For collection of word translations, we take the method of Mikolov et al. [3] that starts with creating a VSM for the source and the target language from monolingual corpora in the magnitude of billion(s) of words. VSMs represent words in vector spaces of some hundred dimensions. The key point of the method is learning a linear mapping from the source vector space to the target space supervised by a seed dictionary of 5 000 words. Training word pairs are taken from among the most frequent ones skipping pairs with a source of target word unknown to the language model. The learned mapping is used to find a translation for each word in the source model. The computed translation is the target word with a vector closest to the image of the source word vector by the mapping. The closeness (cosine similarity) between the image of the source vector and the closest target vector measures the goodness of the translation, the similarity of the source and the computed target word. Best results are reported when the dimension of the source model is 2–4 times the dimension of the target model, e.g. 800 → 300.

We generate word translations between the following language pairs: Hungarian-Lithuanian, Hungarian-Slovenian, and Hungarian-English.

The method provides a measure of confidence for each translational pair, namely the distance of the vector computed by mapping the source word vector, and the nearest target word vector. This measure makes a tuning between precision and recall possible (Table 10). With a higher cosine similarity cut-off (column cos >), we get word translations for a smaller vocabulary (vocab) with a higher precision, while lower cosine similarities produce a greater vocabulary with translations of a lower precision. prec@1 is the ratio of words, for which the first candidate translation coincides with that provided in the seed dictionary, prec@5 is the ratio of words with the seed translation in the first 5 candidates. These are strict metrics, as synonyms of the gold translation count as incorrect. gold is the number of words with a gold translation in the corresponding part of the test data.

We follow Mikolov et al. [2] in using least squares of the Euclidean distance for training, and, surprisingly, cosine similarity for translation generation, which is the only combination of the two distances that works.

4 Data

4.1 Corpora and vectors

For English, we use vector models downloaded from the home pages of the tools, while for the medium-resourced languages, we train new models on the corpora in Table 5, using the tokenization provided by the authors of the corpora.

4.2 Seed dictionaries

Mikolov et al. [3] use Google translate as a seed dictionary. We have been experimenting with three seed dictionaries: (1) efnilex12, the protodictionaries collected

Table 5. Corpora for medium-resourced languages. Word counts are given in billions.

language	corpus	# words
Lithuanian	webcorpus [9]	1.4 B
Slovenian	slWaC [10]	1.6 B
Hungarian	webcorpus [8]	0.7 B
Hungarian	HNC [11]	0.8 B

within the EFNILEX project [12], (2) word pairs collected using wikt2dict with and without triangulation (See Ács et al. [13], and, for sizes, Table 6), and (3) dictionaries from the opus collection (Europarl, OpenSubtitles2012 and OpenSubtitles2013)[3]. efnilex12 contains directed dictionaries (ranked by the conditional probability of the (cooccurrence of the) target word conditioned on the source word).

Table 6. Number of translational word pairs in the seed dictionaries

	efnilex12	wikt	wikt triang	OSub12	OSub13	Europarl
en-hu	83 K	47 K	+134 K	97 K	19 K	21 K
hu-lt	152 K	6 K	+21 K	11 K	9 K	27 K
hu-sl	235 K	2 K	+26 K	63 K	45 K	29 K

5 Results

Throughout the following two sections, these abbreviations will be used: d for dimension, w for window radius ($w = 15$ means that (a maximum of) 15 words are considered on both sides of the word in focus), i for number of training iterations over the corpus (epochs), m for minimum word count in the vocabulary cutoff, and n for number of negative samples (in the case of word2vec).

5.1 Analogical questions

For comparing the Hungarian analogical questions to the English ones, we trained sgram models on the concatenation of HNC and the Hungarian Webcorpus with $d = 300, m = 5$ comparing negative sampling to hierarchical softmax (two techniques to avoid computing the denominator of softmax that is a sum with as many terms as there are words in the embedding) and the effect of subsampling of frequent words, see [14] for details. In Table 7, it can be seen that we (bellow the line) get similar results in the Hungarian equivalent of the original tasks

[3] http://opus.lingfil.uu.se/

(Mikolov et al. [14] are above the line) in the morphological questions, while Hungarian results in the semantic questions are worse. This suggests that the semantic questions are too hard. This problem has to be investigated further.

Table 7. Comparison of results in word translations to those of Mikolov et al [3]

		morph		semant		total	
en [14]	$n = 5$	61		58		60	
	$n = 15$	61		61		61	
	HS	52		59		55	
hu	$n = 5$	63.0	3419/5430	38.5	269/699	60.2	3688/6129
	$n = 15$	61.9	3359/5430	39.2	274/699	59.3	3633/6129
	HS	48.9	2653/5430	22.5	157/699	45.8	2810/6129

5.2 Protodictionary generation

In this section we report our results in Slovenian/Hungarian/Lithuanian to English protodictionary generation. We take four source embeddings: two Slovenian ones trained on slWaC, one trained on the Hungarian Webcorpus, and one on the Lithuanian webcorpus by Zséder et al. [9], all in $d = 600$. One of the Slovenian models is a GloVe one, the other models are cbow models with $n = 15$ and $w = 10$. The target model is always glove.840B.300d from the GloVe site, the seed dictionary is OpenSubtitles2012. The source (rs), the target (rt) embedding, or both (rst) was restricted to words accepted by Hunspell. In Table 8 we compare our results (bellow the line) to those of Mikolov et al. [3] (above the line) with slightly different metaparameters. The vocabulary cutoff m of the source embedding is specified for each word2vec model we trained.

Table 8. Results in protodictionary collection

	prec@1	prec@5
en → sp	33	51
sp → en	35	52
en → cz	27	47
cz → en	23	42
en → vn	10	30
vn → en	24	40
glove-sl → en rs	44.80	63.40
word2vec-sl → en $m = 100$ rs	41.70	60.40
word2vec-hu → en $m = 50$ rst	32.80	54.70
word2vec-lt → en $m100$ rt	21.20	36.50

Table 9. Example word translations. cos is the cosine similarity of the image of the source word vector by the learned mapping and the nearest target vector. Words in the target language are listed in the (descending) order of their similarity to the image vector.

source word	cos	translations			
öt	0.9101	five	six	eight	three
jó	0.8961	good	really	too	very
de	0.8957	but	though	even	just
bár	0.8955	though	but	even	because
hit	0.8904	faith	belief	salvation	truth
ugyan	0.888	though	but	even	because
vöröshagymát	0.8878	onion	garlic	onions	tomato

Table 10. Trade-off between precision and recall in Hungarian to English word translation.

cos >	vocab	gold	prec@1	prec@5
0.7	3803	301	68.4%	84.4%
0.6	9967	711	54.7%	74.1%
0.5	12949	958	46.6%	65.6%
0.4	13451	988	45.3%	64.0%

5.3 word2vec, LBL4word2vec and GloVe

We compared word2vec, its modification LBL4word2vec[4], and GloVe with two parameter settings in the two tasks. The two parameter settings were needed because the default (recommended) values of d, w, i and m are different in the two architectures, see Table 11 with the more computation-intensive setting in bold. We trained two models with each architecture on HNC: a small one with

Table 11. Default values of parameters shared by word2vec and GloVe

	word2vec	GloVe
d	**100**	50
w	5	**15**
i	5	**25**
m	**5**	10

the less computation-intensive one of the two default values and a big one with the lesser one (except for using $d = 52$ in small for historical reasons). For the number of negative samples, which is specific for word2vec, we use the default

[4] `https://github.com/qunluo/LBL4word2vec`

$n = 5$. See results in Table 12. Note that GloVe results could be further improved by taking the average other the two vectors learned by the model for each word.

Table 12. Comparison of models trained in different architectures Rows within each model "size" are sorted by precision in semantic task that we consider more relevant to lexicography than morphology. The total number of questions that do not contain out-of-vocabulary words is 5514 in morphological questions and 6283 in semantic ones.

		morph		sem		total	
small	word2vec sgram	49.0%	2703	20.3%	156	45.5%	2859
	LBL4word2vec sgram	46.6%	2567	19.4%	149	43.2%	2716
	word2vec cbow	49.9%	2751	15.7%	121	45.7%	2872
	glove	41.3%	2277	11.1%	85	37.6%	2362
big	word2vec sgram	57.8%	3186	42.0%	323	55.8%	3509
	LBL4word2vec sgram	55.5%	3058	36.3%	279	53.1%	3337
	glove	58.1%	3206	31.3%	241	54.9%	3447
	word2vec cbow	57.8%	3187	30.7%	236	54.5%	3423

5.4 Comparison of results in the two tasks

In Figure 1 we show the results of some Hungarian VSMs in the analogical and the word translation task plotted against each other. The horizontal axis shows precision in the semantic analogical questions, while the vertical axis shows precision (@5) in protodictionary generation to the Google News model[5] restricted to words accepted by Hunspell and using seed pairs collected with wikt2dict. It can be seen that result in the two tasks are unfortunately uncorrelated.

6 Parameter analysis

6.1 Corpus

Quality In Table 13, we compare on analogical questions models trained on the Hungarian National Corpus (September 12 snapshot) [11] that is a curated corpus of Hungarian, and on the Hungarian Webcorpus [8] that is a similarly sized webcorpus. The numbers suggest that a curated corpus is more suitable for the analogical task.

Size Table 14 shows how the performance depends on the size of the corpus. It is clear that a much larger corpus is needed to answer semantic questions.

[5] `https://code.google.com/p/word2vec/#Pre-trained_word_and_phrase_vectors`

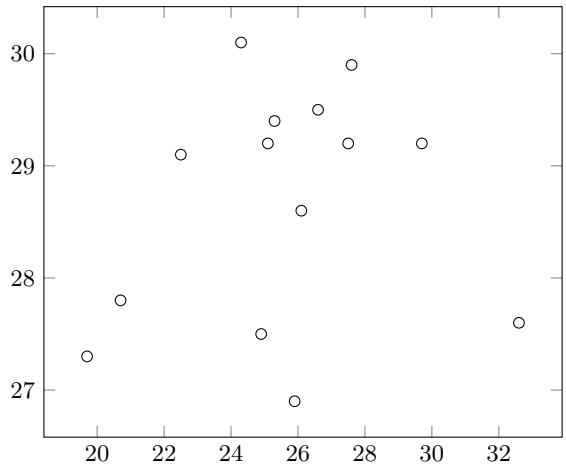


Fig. 1. Precision in monolingual (horizontal axis) vs. bilingual (vertical axis) task

Table 13. Comparison of results on two different corpora. The denominator of each fraction is the number of questions with all three words known to the vector model, while the numerator is the number of correct answers for these questions. Parameters: $d = 152$, $m = 10$, $i = 5$ in both models. For word2vec, $w = 5$ and $n = 5$ while for glove, $w = 3$. The different window sizes mean that these results are not suitable for comparing the models just the corpora.

model	question type	Webcorpus		HNC	
word2vec	morphological	**54.9**	**2924**/5326	51.8	2856/5514
	semantic	8.3	40/482	**16.0**	**123**/769
	total	**51.0**	2964/5808	47.4	**2979**/6283
glove	morphological	47.4	2525/5326	**48.2**	**2658**/5514
	semantic	9.3	45/482	**14.4**	**111**/769
	total	**44.2**	2570/5808	44.1	**2769**/6283

Table 14. The effect of corpus size.

	morph		sem		total	
1M	1.8	58/3256	0.0	0/84	1.7	58/3340
2M	6.1	191/3130	0.0	0/60	6.0	191/3190
10M	24.9	986/3954	7.4	8/108	24.5	994/4062
100M	55.1	2530/4594	31.4	37/118	54.5	2567/4712
754M	63.2	3486/5514	49.8	383/769	61.6	3869/6283

6.2 word2vec

Hierarchical softmax and negative samples We also tried whether hierarchical softmax (HS) and negative sampling can be combined to get better result with either of the techniques. A negative answer can be seen in Table 15 (HNC, $d = 100, w = 5, i = 5, m = 5$).

Table 15. Hierarchical softmax (HS) and negative sampling.

	morph		semant		total	
cbow $hs = 0, n = 5$	59.4%	**3276**/5514	24.1%	185/769	55.1%	3461/6283
cbow $hs = 1, n = 0$	49.0%	2702/5514	13.9%	107/769	44.7%	2809/6283
cbow $hs = 1, n = 5$	49.5%	2730/5514	14.3%	110/769	45.2%	2840/6283
sgram $hs = 0, n = 5$	59.1%	3261/5514	33.6%	**258**/769	56.0%	3519/6283
sgram $hs = 1, n = 0$	49.8%	2744/5514	23.1%	178/769	46.5%	2922/6283
sgram $hs = 1, n = 5$	50.4%	2781/5514	23.1%	178/769	47.1%	2959/6283

6.3 Protodictionaries: Seed dictionary

We compare result obtained in the protodictionary generation task with different English-Hungarian seed dictionaries in Table 16. The source language model is always glove.840B.300d[6], the target model is also a GloVe model trained on HNC ($d = 300, m = 1, w = 15, i = 25$). For details of the seed dictionaries see Section 4.2.

Table 16. Accuracy of protodictionary generation with different seed dictionaries

seed dictionary	prec@1	prec@5
Europarl	17.70%	34.10%
wikt triang	13.10%	25.30%
wikt	12.50%	25.40%
OpenSubtitles2012	10.30%	23.40%
efnilex12 en→hu	10.10%	23.80%

Acknowledgements

I would like to thank Tamás Váradi for supervision, András Kornai, Csaba Oravecz and Attila Zséder for ideas and advices, and Mehmet Umut Sen for translating the essence of [5] to English. Work was supported by the EFNILEX project.

[6] `http://nlp.stanford.edu/projects/glove/`

References

1. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08, New York, NY, USA, ACM (2008) 160–167
2. Mikolov, T., Yih, W.t., Geoffrey, Z.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT-2013. (2013)
3. Mikolov, T., Le, Q.V., Sutskeve, I.: Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168 (2013)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In Bengio, Y., LeCun, Y., eds.: Proc. ICLR 2013. (2013)
5. Sen, M., Erdogan, H.: Learning word representations for Turkish. In: Signal Processing and Communications Applications Conference (SIU), 2014 22nd. (2014) 1742–1745
6. Turney, P.D.: Similarity of semantic relations. Computational Linguistics **32** (2006) 379–416
7. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn treebank. Computational Linguistics **19** (1993) 313–330
8. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004). (2004) 203–210
9. Zséder, A., Recski, G., Varga, D., Kornai, A.: Rapid creation of large-scale corpora and frequency dictionaries. In: Proceedings to LREC 2012. (2012) 1462–1465
10. Ljubešić, N., Erjavec, T.: hrwac and slwac: Compiling web corpora for Croatian and Slovene. In Habernal, I., Matousek, V., eds.: Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings. Lecture Notes in Computer Science, Springer (2011) 395–402
11. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian gigaword corpus. In: Proceedings of LREC 2014. (2014)
12. Héja, E., Takács, D.: An online dictionary browser for automatically generated bilingual dictionaries. In: Proceedings of EURALEX2012. (2012) 468–477
13. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., eds.: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 3111–3119

Statisztika megbízhatósága a nyelvészetben Széljegyzetek egy szótárbővítés ürügyén

Naszódi Mátyás

MorphoLogic, 1122 Ráth György utca 36.
naszodim@morphologic.hu

Kivonat: Manapság szinte korlátlan mennyiségben lehet természetes nyelvű szövegeket elérni a www jóvoltából. Emiatt a nyelvi kutatásoknál, eszközök fejlesztésénél erősen támaszkodnak nyelvi statisztikákra. A megbízhatóság kérdésével viszont kevesen foglalkoznak, pedig ez kulcskérdése a tömeges adatok felhasználhatóságának. Ez a cikk azzal foglalkozik, milyen jellegű objektív korlátai vannak a statisztikáknak, és hogyan lehet becsülni a megbízhatóságot.

1 Bizonytalanságok a nyelvben

A 80-as években, mikor természetes nyelvű szövegek feldolgozásával kezdtem foglalkozni, matematikusként azt a tényt kellett tudomásul vennem, hogy semmi sem százszázalékos. Korábban olyan témakörben dolgoztam, ahol egy állítás vagy igaz, vagy nem, esetleg az adott axiómarendszerben nem eldönthető. Nem így a nyelvészetben. A természetes nyelv tele van gyengén definiált fogalommal, szabályokkal, melyekre mindig találunk kivételeket, többértelműséget, redundanciát.

1.1 Bizonytalanságok szószinten

Már az a kérdés, hogy egy szó magyarnak tekinthető, vagy esetleg a karaktersorozat hibás, nem egyszerű kérdés. Ha nem lettek volna a tizenkilencedik századi nyelvújítók, nem lenne egységes írásunk, *mozdony* szavunk. Ha Karinthy nem írta volna le a *patyolat* szót, azt sem ismernénk. Ha Kellér Dezső nem alkotta volna meg a *maszek* szót, nem használnánk. Egy újságcikkben megjelenhetnek tulajdonnevek időnként nem magyar abc betűit használva. Idegen szavak kerülnek a köznyelvbe, melyeknek lenne ugyan magyar megfelelője, de mégis a jövevény terjed el. Ezeregy oka lehet, hogy egy-egy új szót bevesz a nyelv, másokat elfelejt. A toldalékolás is változik, illetve bizonytalan. A régies múlt *kihala vala* a köznyelvből. Megállapodott szabályokat rúg fel a gyakorlat, illetve fel nem ismerhető régi szabályok felülírnak szokásosakat. (*gyorsan*, *boldogan*, de *nagyon*, *fiatalon*, *gazdagon*; *mondta*, vagy *mondotta*; *falsság*, *nyersség*, de *rosszaság*, *gyorsaság*, sőt *frissesség*, *bölcsesség*; *gondtalan, de gondatlan*) De nem csak időbeni változások léteznek. Ami megengedett egy szaknyelvi szövegben, helytelen lehet egy köznapi mondatban.

A szavak jelentése sem egyértelmű. Egy szóalaknak nem csak azért lehet több értelme, mert azonos alakú, lényegesen más szóról van szó (*lépnek* FN, *lépnek* IGE; *értem* = *érik* IGE+ME1 /*én*+*ért*/ *ért* IGE+E1 tárgyas…), hanem egyes képzett, összetett szó új jelentést kaphat, de megtarthatja eredeti nyelvtani struktúrából eredő jelentését is. (*lovagol – lovon közlekedik* / *lovagol valamin – ragaszkodik egy érvhez*). Számos más jelenség is okozhat többértelműséget. Az esetek többségében az egyik értelmezésnek sokkal nagyobb a valószínűsége, mint a többinek, ha másként nem, a szövegkörnyezet függvényében. Angol nyelvben a szavak többértelműségének többsége abból ered, hogy a szavaknak névszói és igei jelentése is van, de ezt a szórend egyértelműsíti. Az erősen ragozó nyelveknél a szóalakok nagyszámú változata miatt keletkeznek különböző módon generálható, de azonos alakú szavak.

1.2 Bizonytalanság a nyelvtanban

A mai nyelvgyakorlat elüt a korábbitól. Mostanában számos angol nyelvből átvett forma jelentkezik a hétköznap és a sajtó gyakorlatában. Ezen túl nyelvészeti cikkeket olvasva sok olyan mintapéldát találtam, melyet a szerző helytelennek, esetleg kérdésesnek talál, nekem meg nyelvileg tökéletesnek mutatkozik, és viszont, helyesnek jelzett mondatban találok javítandó hibákat. Néha egymásnak ellentmondó szabályokat kell egy mondatra alkalmazni. A mondatok nyelvtani elemzése köztudottan nem egyértelmű. A gépi elemzők nagyságrenddel több alternatívát tárnak fel, mint az ember feltételezi olvasás közben. Ezek nagy hányada fel sem merül egy olvasónak, mert a szemantikai megkötések, az ismert és gyakori minták elnyomják a ritka lehetséges elemzést. Az itt is igaz, hogy az esetek többségében egy-két elemzés dominál, a többinek kis súlya van. Arról már nem is beszélek, hogy egy nyelvre több lehetséges nyelvtant lehet készíteni. Míg a szavakról, minősítésükről egyöntetűbb rendszerek vannak, a nyelvtan, főként a formalizált nyelvtan, szerzőnként eltér. A nyelv egy objektív jelenség, de a nyelvtan ember által alkotott modell, ami lehet jó, pontos, de sohasem a valóság maga [5]. Azért alkotja a nyelvész, hogy áttekinthetőbbek, kezelhetőbbek legyen a nyelv jelenségei.

1.2 Bizonytalanság jelentésben, fordításoknál

Egy mondatnak számos interpretációja lehet egy másik nyelven, illetve szemantikai reprezentációban. Léteznek jó és gyenge fordítások. Ritkán beszélhetünk tökéletesről, de megfelelő fordításról gyakran. Ebben az esetben olyan valószínűségi modellt lehet alkalmazni, ami szerint azt mérjük, hogy két különböző nyelven írt mondatnak mi a korrelációja, vagyis ugyanabban a helyzetben, ahol az egyik elhangzik, a másik nyelven milyen valószínűséggel hangzik el a másik. Mivel a mondatok száma gyakorlatilag végtelen, ezt nem lehet számba venni, viszont az egyes mondatoknál kevés a minimális valószínűségi küszöböt elérő mondatpárok száma. Ezt használják ki a statisztikai és memória alapú fordítók.

Sorolhatnék még számos bizonytalansági kérdést a természetes nyelveknél. Tulajdonképpen a természetes nyelv velejárója a nem teljesen meghatározottság, illetve a többértelműség [5]. A nyelvészeti kérdésekre általában nem tudunk igennel, nemmel válaszolni. Megfelelőbb az olyan valószínűségi modell, melyben a legtöbb esethez nagy (egyhez közeli) vagy kicsi (nulla körüli) valószínűséget rendelünk. Számos esetben nem tudjuk minősíteni határozottan a nyelvi jelenséget. Ezért érdemes valószínűségi modelleket alkalmazni.

2 A nyelvi statisztikák karaktere

Ezek dacára nem ismerek olyan helyesírás-ellenőrzőt, amely nem kategorikus választ ad arra, hogy egy szó helyes vagy nem. A stílusellenőrzők zöme is határozott állítással ítél, esetleg egy-két esetben ad figyelmeztető jellegű az üzenete. A fordítóprogramok sem zavarják a felhasználót azzal a közléssel, mennyire biztos a szöveg interpretációja, esetleg hány száz, ezer egyéb alternatívát ismer az adott mondat áttételére.

A tapasztalat azt mutatja, hogy nyelvi esetek túlnyomó többségénél a bizonytalanság karakterisztikája olyan, hogy a gyakori esetek elnyomják a ritkábbakat. Értsd ezen azt, hogyha két-három választék van a megoldásra, az egyik általában nagyon gyakori, a többi ritka. Úgy is lehet magyarázni, hogy ha egy konfidenciatartományt jelölünk ki, akkor ebbe kevesen jutnak be. Ha sok lehetséges eset van, akkor azok kis százaléka lefedi a futó szövegek nagy százalékát. Ez egy közismert jelenség, amely igaz természeti törvényszerűségekből eredhet.

2.1 A Zipf-törvény

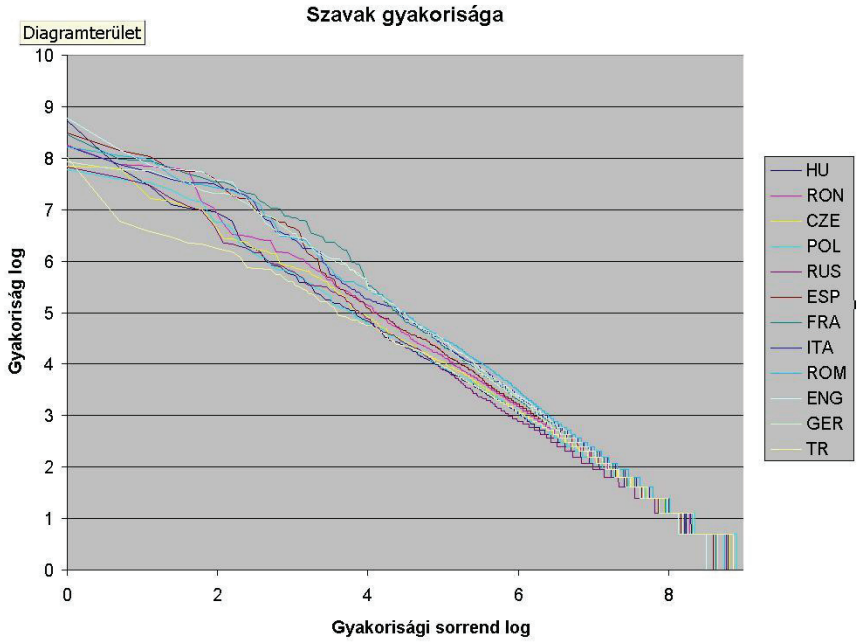
Számos természetben előforduló sokaságnál igaz a következő összefüggés: Ha megkülönböztethető csoportokat alkotunk az előforduló egyedekből, és a csoportokat előfordulásuk gyakoriságának sorrendjében rendezzük, akkor az `n`-edik csoport előfordulásának relatív gyakorisága a következőképp becsülhető:

$$f(n) \approx C/n^s \qquad (1)$$

ahol `C` egy normáló konstans, `s` pedig egy egynél kicsit nagyobb kitevő [1].

A törvény nagyobb sokaságoknál igaz. Ha azt nézzük, hogy a vagyon hogy oszlik el az emberek közt, vagyis gazdagság szerint sorrendezzük az embereket; a népesség hogyan oszlik el a lakóhelyeken, vagyis a településeket sorrendezzük a lakosok száma szerint, hasonló jelenséggel találkozunk. A lényeg, hogy a csoportok száma meghaladja az ezret.

Kis `n`-nekre, tehát a leggyakoribb csoportokra nem jó ez a becslés. A leggazdagabbak eloszlása sem igazán követi a képletet, ott más törvények is belejátszanak az adatok alakulásába. Ha például a szövegekben előforduló szavakat vagy szóalakokat vizsgáljuk, akkor a leggyakoribb szavak relatív gyakoriságára nem állja meg a helyét a becslés, a ritkább értékeknél sem használható a képlet – a mérési, mintavételezési hiba miatt – de a legalább tízszer előforduló egyéb szavaknál elég jó a közelítés.



1. ábra. A szavak gyakorisági grafikonja jól mutatja, hogy a középmezőnyben követi a Zipf törvényt.

Az s kitevő értéke függ a sokaság típusától. Ha a szógyakoriságot nézzük, akkor magyar nyelvnél ez közelebb van az egyhez, mint az angolnál. Ha megnézzük a függvény grafikonját, monoton csökkenő, alulról konvex függvényt kapunk, ami az ordinátához simul. A nyelvi esetekben az s annál közelebb van az 1-hez, minél nagyobb az elvi csoportok száma. Ez magyarázza a magyar és angol közti eltérést. 1 nem lehet az s, mert akkor nem lehetne normálni, mert a görbe alatti terület korlátlan (1. táblázat).

Az s-et tapasztalati mérések alapján becsülik, de ennek a becslésnek is vannak megbízhatatlansági tényezői. Ha viszont az s-et ismerjük, becsülhetjük a teljes sokaság számát is. Például a szavak, szóalakok számát egy nyelven. A becslés nem pontos, de nagyságrendi adatokat meg lehet állapítani. A fenti táblázat adatai nem a nyelvről, hanem a regényről, illetve a fordításról/fordítóról adnak tájékoztató adatot.

A becslés nem alkalmazható egyszerű betűstatisztikára betűírás esetén, mert az egy nyelven használt elemek száma erősen ezer alatt van. Ideografikus (pl. kínai) írás esetén viszont már megfelelő az alaphalmaz számossága. Általában digráf statisztikára sem igazán jó, de a lehetséges betűhármasok száma egy nyelven már kellően nagy, hogy megkísérelhessük a Zipf-képlet konstansainak meghatározását.

1. táblázat. s becsült értéke Orwell 1984 fordításai alapján.

Nyelv	Szavak	Szóalakok	>1	>2	>3	s
magyar	88054	19236	6575	3378	2638	0,9327
török	73224	19186	7046	4164	2944	1,0059
orosz	74109	18560	6443	3873	2485	0,9284
cseh	79681	17630	6673	3991	2817	0,9832
lengyel	80186	19598	7399	4237	2898	0,9411
roszin	89560	16874	6651	3963	2863	1,0059
francia	111395	11367	5629	3748	2842	1,1586
spanyol	95380	11442	5390	3513	2602	1,1384
olasz	102578	13771	6321	4025	2983	1,1060
román	107477	14154	6479	4217	3100	1,0754
angol	104759	9211	4906	3397	2639	1,2258
német	100052	14048	5754	3598	2628	1,2502

2.2 A nyelvi statisztikák abszolút pontatlansága

Minden nyelvre jellemző a betűstatisztikája. Már itt is gondban lehetünk, mert ha például megnézzük Arany János és Petőfi Sándor írásait, akkor a két azonos korban élő költőnél eltérések fedezhetők fel. Arany János nagy hangsúlyt helyezett arra, hogy mély hangrendű szavakat használjon, emiatt az *a* és az *e* betűk gyakorisága alig különbözik, míg Petőfinél lényegesen több az *e* betű. Tehát egy nyelvnek nincs jól meghatározható betűvalószínűsége, illetve ha azt feltételezzük, hogy van, akkor nehéz az általános statisztikának megfelelő mintavételezést alkalmazni.

Ha meg akarják határozni, hogy hány és milyen hal van a Balatonban, akkor azt úgy teszik, hogy egy mérhető területen lehalásszák a halakat, és az itt nyert eredményből extrapolálnak. Nem mindegy viszont, hogy ezt hol teszik. A part közelében más az eredmény, mint a parttól távol, és Siófoknál eltér az összetétel a keszthelyi öblétől. Nevezzük ezt a bizonytalanságot ***haleffektus****nak.*

Habár Arany Jánosnál és Petőfi Sándornál eltérnek a statisztikák, mégis nagy vonalakban hasonló az eredmény. A hasonlóság azt jelenti, hogy a gyakoribb betűk relatív gyakorisága legfeljebb 20-30%-kal tér el egymástól a két írónál és általában a magyar szövegekben. A ritka betűknél ennél nagyobb eltérés is lehetséges.

Az egyszerű karakterstatisztikáknál sokkal jobban jellemzi a nyelveket, milyen karakterek követik egymást, ezért betűkettősök, betűhármasok felmérésével elég jól meg lehet határozni a nyelvet, melyen a szöveg íródott [2]. Itt sem abszolút kiértékelésről van szó. A módszer inkább az, hogy a kérdéses szöveg gyakoriságvektorát veti össze a szóba jövő nyelvek előzetesen elkészített gyakoriságvektorával, és amelyikhez a legközelebb áll, azt a nyelvet kiálltja ki győztesnek. A módszer tehát a Bayes-módszer egyik alkalmazása.

Ha nem karakter-, hanem például szóstatisztikát készítenek, akkor az eltérések jelentősebbek. A leggyakoribb szavak nyelvtani funkciós szavak. Például a névelők, kötőszavak. Ezek aránya minden szövegben hasonló. A továbbiak viszont eltérnek.

A sorrendben első húsz-harminc szó minden nagyobb szövegben előfordul, ráadásul egyenletesen eloszolva a szövegtörzsben. A gyakoriak, de nem az élbolyban szereplők már nem. Egy vaskosabb receptkönyv elemzésénél vettem észre, hogy a könyv első felében a *só* majdnem annyiszor szerepel, mint a másodikban, de a *cukor* a második felében sokkal többször. Hát persze, mert az édességek a könyv végén szerepeltek.

Ez azt jelenti, hogy a szavak gyakoriságát a **haleffektus** miatt nehéz becsülni. Inkább csak kvalitatív, mintsem kvantitatív eredményeket várhatunk. Ilyen eszközökkel viszont már szerzőket is fel lehet ismerni, illetve szaktémákat lehet megkülönböztetni. Ennek finomított változata elég a plágiumkereséshez.

Érdekes, hogy az azonos jelentésű szavaknál hogyan lehet megkülönböztetni az alaktól a szót. Füredi Mihály Magyar nyelv szépprózai gyakorisági szótárában [1] az *az* névelő és az *az* mutató névmást külön számlálta. Miután ilyen munkát csak egyszerű gépi módszerrel lehet elvégezni – a nyolcvanas években gépi egyértelműsítőről legfeljebb csak álmodhattunk – egyszerű emberi beavatkozással oldották meg a problémát: mivel mindkettő értelmezés gyakori a magyar nyelvben, ezért elég egy részmintában emberi erővel elvégezni a statisztikát, és a két értelmezés arányát fixnek véve az alakok alapján lehet jól becsülni a kettő gyakoriságát a teljes korpuszon. De mi van, ha az egyik értelmezés ritka, mint a *meg* igekötő (gyakori) és kötőszó (ritka) használatánál. Minden esetet megnézni nem lehet emberi erővel, tehát nem tudhatjuk, mi a ritka változat gyakorisága. Ha mindkettő ritka, akkor persze átnézhetőek az adott szóalak előfordulásai, emberi döntéssel megadható, hol, melyik a helyes értelmezés. Ilyen szóból viszont – a Zipf-törvénynek megfelelően – rengeteg van. Tehát az eseti döntés megoldható, de a tömegfeldolgozásuk nem.

A tapasztalatom az, hogy a nyelvi adatok abszolút pontossága egy anyagon belül nem függ az eset gyakoriságától, tehát állandó. Akár gyakori esetről van szó, akár ritkáról, a tévedés gyakorisága ugyanaz, tehát a relatív tévedés fordított arányba van az illető gyakoriságával. Emiatt gyakori esetekről sokkal megbízhatóbbak az adataink, mint a ritkákról.

2.3 Anyaggyűjtés matematikája

A nyelvi feladatok többsége olyan, mint egy szótár készítése. Szavakat gyűjtünk, és hozzárendelünk információkat. Amennyire lehet, ezt géppel vagy gép segítségével végezzük. A szótár – ha helyesírás-ellenőrzőről van szó – csak szógyűjtés. Erősen ragozó nyelveknél további osztályozás is nélkülözhetetlen. A szógyűjtés manapság abból áll, hogy nagy korpuszokban fel nem dolgozott szavakat, kifejezéseket keresünk, majd feldolgozzuk azokat. A gyakori szavakra gyorsan rátalálunk, és feldolgozása is egyszerűbb, hisz az ember a felmerülő kérdésekre biztos választ tud adni. A ritkábbakra nehezebb rálelni, és nem mindig egyszerű a szóhoz rendelt tulajdonságokat pontosan megadni.

Tehát a szótár bővítése annál nehezebb, minél ritkább szavakkal kell törődni. Ha ezt nem vennénk figyelembe, akkor a szótárbővítés sebessége a tételszámokkal lineárisan nőne:

$$T(N) = \sum_i e_i = N*e \quad (2)$$

ahol N a feldolgozott szavak száma, e pedig az egy szóra fordított idő. Ha így lenne, akkor is gond a szövegek lefedettsége, mivel a Zipf-törvény alapján a szöveg nagy hányadát ugyan lefedhetjük a gyakori szavakkal, de ha még nagyobb lefedettséget akarunk elérni, akkor hatványozottan nagyobb korpuszokat kell vizsgálni.

Emiatt a szótár mennyiségével rohamosan nő az elvégzendő munka. Új elem megtalálásához a gyakoriság reciprok arányában nő a szükséges korpusz mérete, így a feldolgozás ideje a tételszámokkal legalább négyzetesen nő:

$$T(N) = \sum_i e_i = C*\sum_i i^s = C* N^{1+s}/(1+s) > O(N^2) \quad (3)$$

Ráadásul ahhoz, hogy javuljon a lefedettség – a ritkább elemek kevéssel javítják ezt a paramétert – a pontosság növeléséhez ezen érték hatványát kell munkába fektetnünk. Persze a talált objektumról véleményt is kell mondani. Ha csak helyesírás-ellenőrzőről van szó, akkor meg kell ítélni, az újonnan talált szó eleme-e a nyelv szókészletének, vagy sem. Ez azért fontos, mert nem minden meglelt sztring helyes szó. Még az sem segít, ha a gyakoriságát nézzük. A *hüje* szó ötször több esetben szerepel a Google szerint a szövegekben, mint például az *oktondi*. Ennek ellenére az előző antiszó, míg az utóbbi teljesen köznyelvi, helyes alak. Ha feltételezzük, hogy egy szó felvétele, osztályozása is fordított arányban van a gyakoriságával, akkor ez a korábbi képletet erősíti. Így egy minőség előállításához szükséges idő nagyságrendjének becslése:

$$T(q) > O(1/q^2) \quad (4)$$

ahol $1-q$ a lefedettség, tehát valamilyen minősítéshez tartozó mérőszám.

2.4 A tévedések matematikája – a mennyiségi korlát

A Zipf-törvény magyarázataiban szerepel az is, hogy a kis valószínűségű osztályok értékei nem igazán követik a képletet. Ez nagyszámú osztályok esetén természetes, hisz a ritka elemek gyakorisága 1 körüli, vagy annál kisebb, akkor – egész értékű előfordulás miatt – eleve létezik egy pontatlanság. A pontatlanságnak más okai is vannak:

1. a korábban említett mintavételezési hiba – a haleffektus
2. az emberi döntések hibája
3. a forrásanyagban (korpuszban) fellelhető hibák

Ha ezt is beszámítjuk, akkor a szótár minősége – az egyedek helyes és helytelen felvételének aránya – fokozatosan romlik a szótár növekedtével. Ha a szótárkészítő egyenletes minőségben dolgozna, akkor a szótár minőségének felső határa a szótárkészítő helyes döntéseinek arányával azonos. Ha viszont feltételezzük, hogy a helyes döntés aránya csökken a felvett szó gyakoriságának csökkenésével, akkor azt a képletet kapjuk, hogy van egy olyan határ, amikor a bővítés már ront. Ha valaki nem hiszi, akkor nézze meg egy keresőben, hányszor szerepel a *kéttannyelvű* szó a weben, és hányszor a *két tannyelvű* helyes alak. A hibás sokkal többször. Ez a konkrét példa nem szerepel ugyan a helyesírási szótárakban, de az általános nyelvtani szabályok miatt külön kell írni. Nem úgy, mint a *négykerék-meghajtású*t. (Házi feladat, miért?)

És e két példa nem az igazán bonyolult eset a magyar nyelvben. Vagyis bárhogy állapítjuk meg a szótár minőségi követelményét, egy mennyiség elérése után már rosszabb eredményt kapunk a követelménynél.

A minőséget persze lehet javítani. Az egyik lehetséges módszer, hogy a döntéseket két, esetleg több független (kis korrelációval rendelkező) döntnök (algoritmus) hozza. Ha kicsi a hibaszázalék, ez két független döntés esetén közel felezi a hibás kódolásokat. Mivel a hibák aránya a 3. képlet alapján négyzetesen nő a gyűjtés mennyiségével, előbb utóbb akkor is objektív korlátba ütközik a minőség megtartása, és ezen a sok független döntnök sem segít.

Ha például szótárkészítésről van szó, akkor tapasztalatom szerint egy ember belátható idő alatt maximum 10 000 tételből álló szótár készítésére képes. A nagyobb szótárakhoz közösségekre, lektori rendszerekre van szükség. A nagy – 100 000-200 000 vagy több tételt tartalmazó szótárak többgenerációs munkák. Tehát a minőség eléréséhez sok ember paralel döntésére van szükség. Ha egy ilyen szótárat módosítani akarnak, akkor a hagyományos szócikkiosztás módszere nem célravezető, hisz ilyenkor már ritkább szavak kerülnek sorra, és lehetséges, hogy többet rontunk a meglévő készleten, mint javítunk. Ez látszik is az utóbbi időkben megjelent összevethető szótárak minőségén.

Érdekes probléma például a helyesírás-ellenőrzők adatbázisának építése. Nem feltétlen a nagyobb szótár a jobb. Lefedettségben ugyan igaz, de a helyesírás-ellenőrzőnek az a feladata, hogy hiba esetén jelezzen. A hiba pedig azt jelenti, hogy a szó a szövegben helytelen. Ha minden helyes formát megengednénk, akkor olyan elütések, melyek helyes szóalakhoz vezetnek, elrejtik a hibát. Ilyenből rengeteg van. Például a *tanit* helyes nyelvileg, *tan*+*i*+*t*, de nem gyakori az előfordulása. Ezzel szemben a *tanít* elég gyakori, és a hibás, rövid *i*-vel történő írása is gyakrabban fordul elő, mint a korábbi eset.

Egy ellenőrzőprogram jóságát abban szokták mérni, mennyi az elsődleges és másodlagos hibák száma. Elsődleges hiba, ha nem ismer fel egy helyes szót, másodlagos, ha jónak tekint egy hibásat. A két hibának más a súlya. Tehát a minősítés (az eszköz rosszaságának mérőszáma) lehet a következő:

$$M = C_1*E_1+C_2*E_2 \quad (5)$$

ahol a C-k a súlyok, és az E-k a hibák relatív gyakoriságai. Általában a másodlagos hibát nagyobb súllyal szokták számítani, mint az elsődlegest. $C_1<<C_2$. A felmérést a legritkább esetben végzik futó szövegeken, inkább csak kigyűjtött szókészleteken. A szóellenőrzők nem képesek figyelembe venni a szövegkörnyezetet.

Ennél finomabb súlyozás logikusabb lehet. Tudniillik a konkrét hibáknak más a hatásuk a szövegekben. Ha azt írjuk, hogy *hüje*, vagy *talpalattnyi*, az nem okoz akkora galibát, mintha a *kóros* helyett *koros* kerül a szövegbe. Míg a korábbiaknál az olvasó helyesen értelmezi a hibás karakterláncot, az utóbbinál értelemzavaró lehet az elírás. Ezért ha egy helyesírás-ellenőrzőbe felveszünk egy új szót, akkor azt is meg kell vizsgálni, hogy milyen gyakori szavakhoz van közel a szó vagy a szóból előállított szóalak, és ez mekkora gondot jelenthet. Ezért nem szerepel egy helyesírás-ellenőrzőben sem a *suly* (*betegség*) szavunk. Egy új szó felvételének haszna a következő:

$$M = C_1*E - \sum_j C_{2j}*m_j - \sum_i C_{3i}*m_i \quad (6)$$

ahol C_{2j} a másodlagos hiba kára, ha a szó helyett a j-edik szó kerül a szövegbe, C_{3i} a másodlagos hiba kára, ha a j-edik szó helyett az új szó kerül a szövegbe, a két szó közti távolság pedig m_{ij}, vagyis arányos annak a valószínűségével, hogy az i-edik szó helyett a j-edik írjuk le.

A szótár minőségre a következő becslést adhatjuk.

$$M = \sum_i C_{1i}*E_i - \sum_i \sum_j C_{2ji}*m_{ij} \quad (7)$$

ahol C_{2ji} a másodlagos hiba kára, ha az i-edik szó helyett a j-edik szó kerül a szövegbe, a két szó közti távolság pedig m_{ij}.

Összefoglalva, a több gyakran vezet minőségromláshoz, túl a kódolási hibákon. Az a mennyiség, amelyen felül már nem javul a nyelvi gyűjtemény, függ a gyűjtemény típusától, a kódolás módjától, de munka közben becsülhető. Ezen túl nem szabad menni, mert többet árthatunk, mint használunk!

2.5 A fordítások minősítése

Fordítások esetén – ha fordítómemóriáról vagy statisztikai fordítóról van szó – két nagy eltérés van az emberi, fordításhoz képest:

1. Az adatokat nem ember gyűjti, hanem meglevő korpuszok szolgáltatják, melyet általában kontroll nélkül fogad be a rendszer.
2. Az eredményekről nem igen-nem jelleggel minősítünk, hanem valamilyen jósági mértéket lehet, kell mondani.

Mindkét esetben a gyakori mondatok – amelyek fordítása már bekerült a rendszerbe – jól interpretálódnak abban az értelemben, hogy nagy valószínűséggel a módszer kiválasztja a megfelelő interpretációt. A gond ott van, hogy a lehetséges mondatok száma sok nagyságrenddel meghaladja a nyelvben található szavak számát. Ha a Zipf törvényét nézzük, és szópárok, esetleg szóhármasok eloszlását vizsgáljuk egy nyelvben, akkor várhatóan az s konstans nagyon közel van az 1-hez, ennek következtében arra a kérdésre, hogy a gyakoribb osztályok mennyire fedik le a futó szövegben előforduló esetek felét, kétségbeejtő választ kapunk. A ritka elemek lesznek nagyobb hányadban. Ha ez igaz, márpedig így van, akkor a statisztikáink hibája nagyon nagy, akár a haleffektusról, akár a korpuszban fellelhető hibáról van szó. A fordítás ráadásul sohasem lehet tökéletes.

A fordítás minősítését is szeretnénk gépesíteni. Erre az általánosan használt módszer a statisztikai kiértékelés, pl. a BLEU [4]. A gond ezzel ott van, hogy a kiértékelés pont azokat a paramétereket nézi, amilyen alapon generálják a fordított szöveget, tehát részrehajló. Ha csak emberi fordításokat értékelnének ki ezzel a módszerrel, akkor persze objektívabb lenne, hisz a kiértékelés módja korrelálatlan a fordítás módszerétől. [5] Ezt igyekszik kikerülni az ITRANSLATE4 projekt. A módszer egyszerű. A felhasználók visszajelzése értékeli a fordítást. Ennél jobbat csak szakemberek bevonásával lehet elvégezni, ami költséges, és nagy mennyiség kiértékeléséhez kivihetetlen.

Bár a statisztikai fordítás hívei elméletileg támasztják alá, hogy a feldolgozott korpusz méretével egyre jobbak a fordítások, és elvileg – korlátlan forrást feltételezve – a minőség is tökéletesedik, a gyakorlat, és az általam vázolt képletek alapján ez nem igazolódik. Nem beszélve arról, hogy a felhasználható korpuszok mérete sem korlátlan, emiatt a (4)-es képletnél is rosszabb a helyzet. A különböző nyelvek közti fordítás minősége különbözik. A különbség alapvetően nem attól függ, melyik fordítóprogramot használjuk – a Google-ét a Bing-et, vagy bármi mást, hanem a nyelvpártól. Persze az is számít, mekkora anyag van a fordító tarsolyában. A statisztika általában annál pontosabb, minél több anyag van mögötte, de a következők sokkal inkább meghatározók:

1. Milyen nagy a szóformák készlete az adott nyelven
2. Milyen közel van a két nyelv nyelvtana
3. Vannak-e javító trükkök

Ebből a szempontból a magyar nehéz helyzetben van. A fordításnál a másik nyelvtől nagyon különbözik. A szóformák száma nagyságrenddel nagyobb, mint más nyelveknél. Meg is látszik az eredményen. Míg egy angol-francia fordításnál a statisztikai fordítók minőségével a hétköznapi felhasználók elégedettek, a magyar-angol esetben gyengébb minőséget kapunk.

3 Javítási lehetőségek

Ha ennek ellenére jobb minőséget akarunk elérni, akkor a következőket lehet tenni. A sokaság méretével csökken a Zipf-képletben szereplő s kitevő, vagyis a ritka elemek viszonylagosan kevesebben lesznek. Ha így csökkentjük az osztályok számát, akkor jobb eredményt tudunk elérni.

a. Ilyen lehet, ha szakszövegek, terminológiai kötések miatt az általános értelemben ritkább szavak egy részének gyakorisága megnő, de a ritka köznapi szavak nem kerülnek a feldolgozandók közé. Ekkor a szakszavakra összpontosítva bővíthetjük a szótárat. Persze ezzel csökkentjük a felhasználható korpuszok számát is.
b. Hasonló az eset az általános szóellenőrzőknél a sajátszótár használata esetén. Egy jobb tollú felhasználó is csak a tizedét, századát használja a nyelvnek, ezért, ha felmerül egy újabb helyes, de korábban nem szereplő szó, felveheti a szótárába. Egy ember nyelve mindig kisebb a teljes köznyelvnél, ezért az ő nyelvének lefedettségét így lehet biztonságosan bővíteni.
c. A módszerben az osztályok számát csökkentjük. Például egy helyesírás-ellenőrzőnél nem szóalakokat veszünk fel, hanem alapszavakat, melyekből szabályos módon generáljuk a szóalakokat. A ragozó nyelvek esetén e nélkül nem is lehetne jó helyesírás-ellenőrzőt készíteni, mert a szabályos szóalakok száma olyan sok, hogy a korábban említett kritikus mennyiség esetén a hibaszázalék meghaladja a helyesírás-ellenőrzőknél elvárt 5%-ot. Míg angol, francia, spanyol nyelveknél 100 000 – 200 000 szóalak felvételével a 97%-os lefedettség könnyen biztosítható, magyarban 1 000 000 000 szóalak esetén sem érhető el 90%-nál jobb lefedettség.

4 Mérési tapasztalatok

Megkíséreltem az elméletet összevetni a valósággal. Különböző nyelven írt azonos – a mai köznapi nyelvhez közel álló szöveget kerestem. Erre alkalmasnak tűnt Orwell 1984 című műve. Az *újbeszél* szavain kívül a nyelvezete megfelel. Többek között a MULTEXT [7] projekt korpuszának is fő darabja volt. A különböző fordítások elérhetőek voltak. Többségük PDF formátumban, mások MS Word dokumentumként.

Az anyagot saját eszközökkel csupaszítottam le tiszta szövegállományra. Kihagytam az oldalszámozást, fejezetjelöléseket, lábjegyzeteket, elő és utószavakat, mert azok tartalma eltért a kiadványokban. Arra voltam kíváncsi, milyen mérhető eltéréseket tapasztalok nyelvenként.

Az eredmény kicsit váratlan volt. Azt vártam, hogy egyszerű becslésnél már határozottan kimutatható a Zipf konstansaiban az eltérés. A különbség meg is mutatkozott (1. ábra), de nem volt annyira karakterisztikus. Az s értéke – bár 1 körüli volt, de általában kisebb a vártnál (1. táblázat). Utólag magyarázni is tudom.

1. A konstansok általam számított becslése bár megbízható, de nem torzításmentes. Sokkal nagyobb minta kell, hogy a becsült érték közelebb legyen a valósághoz, vagy meg kell találni a torzítatlan becslés képletét.
2. Az értékek erősen függtek a fordítás minőségétől, tehát nem csak a regényt jellemezték, hanem a fordító választékosságát is.
3. Az értékek függtek a szöveg kódolásának minőségétől. A hibás, trehány kódolás miatt több ritka szóalak szerepelt, mint amennyi valójában kellett volna lennie. Ez különösen a roszin nyelvű anyagon látszott.

Azt gondolná az ember, hogy a fordítás miatt határozott szinkronszövegeket talál. Ez bizonyos értelemben így is van, de nem teljesen. Ha például csak azt számoljuk, hogy a főhős, Winston neve hányszor szerepel a műben, akkor mindegyikben 500 körüli, de 10%-os eltérés is tapasztalható. A legkevesebb fajta szóalakot persze az angolban és a németben találtam, és az erősen ragozó nyelvekben többet. Korábban abban a hitben voltam, hogy a finnugor nyelveknél markánsabb lesz a fölény, de nem így volt. A szláv nyelveknél talán a cseh a vezető. Archaikusabb, mint az orosz. Ez nem abban nyilvánul meg, hogy nem hat névszói esete van, hanem hét, hanem abban, hogy a birtokviszonyt is melléknévi ragozott alakban fejezi ki, akár az ukrán.

A grafikon ritka szavaira vonatkozó görbeelnyúlásra jobban lehet következtetni azok arányából, mint a teljes görbéből. Itt határozottabb különbség mérhető a nyelvek között.

5 Összefoglalás

Az elmélet és a tapasztalat is alátámasztja, hogy a nyelvészeti eszközök minőségének objektív korlátai vannak, melyet lehet becsülni még az eszköz készítése előtt, esetleg közben. A becslés nem mindig egyszerű. Összetettebb esetekben a munka során ellenőrző méréseket kell végezni, érdemes-e folytatni a munkát, vagy le kell-e állni, esetleg módszert kell váltani. A minőségi korlát statisztikai alapon becsülhető, és ha nem vesszük figyelembe, a több munkával ronthatunk a készülő eszköz minőségén.

A méréseket érdemes nagyobb korpuszokra is kipróbálni, valamint nyelvi eszközökön, adatbázisokon megnézni, mik a számított korlátok.

A jövőben megkísérlem mérni, igaz-e a feltételezésem: a nagy memóriafordítók elérték a minőségi határukat.

Hivatkozások

1. Zipf, George K.: The Psychobiology of Language. Houghton-Mifflin (1935)
2. Cavnar, W. B., Trenkle, J. M.: N-Gram-Based Text Categorization. In: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (1994)
3. Füredi, M., Kelemen, J.: A mai magyar nyelv szépprózai gyakorisági szótára. Akadémiai Kiadó (1989)
4. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: a method for automatic evaluation of machine translation. In: ACL-2002: 40th Annual meeting of the Association for Computational Linguistics (2002) 311–318
5. Novák, A., Tihanyi, L., Prószéky, G.: The MetaMorpho translation system. In: Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio (2008) 111–114
6. Bach, I., Farkas, E., Naszódi, M.: A magyar nyelv elemzése számítógéppel. Tervek egy természetes nyelvű interfészhez. SzTAKI Tanulmányok, 199 (1987)
7. Erjavec, T.: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris (2004)
8. Seidl-Péch, O.: Autentikus magyar szövegek és fordítás eredményeként létrejött célnyelvi magyar szövegek lexikai kohéziós mintázatának összehasonlító elemzése. Tézisek (2011)

II. Szintaxis, szemantika

Konstituensfák automatikus átalakítása függőségi fákká vagy kézi annotáció?

Simkó Katalin Ilona[1], Vincze Veronika[1,2], Szántó Zsolt[1], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
kata.simko@gmail.com, {szantozs,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat A magyar azon ritka nyelvek egyike, ahol rendelkezésre áll kézzel annotált konstituens és függőségi annotáció is ugyanazon a szövegállományon, a Szeged (Dependencia) Treebanken. Ez lehetővé teszi, hogy megvizsgáljuk a szabályalapú automatikus átalakítás minőségét, valamint hogy összehasonlítsuk az etalon treebankeken tanított konstituens és függőségi elemzők kimenetét a konvertált mondatokon tanított elemzők kimenetével. Eredményeink szerint bár a különböző módszerek számszerűleg hasonló eredményeket érnek el, különböző hibákat ejtenek.
Kulcsszavak: szintaxis, konstituens, dependencia, statisztikai szintaktikai elemzés

1. Bevezetés

Manapság a statisztikai szintaktikai elemzésben legelterjedtebb megközelítések konstituens vagy függőségi nyelvtanon alapulnak. Konstituensalapú treebank számos nyelvre létezik, míg a függőségi treebankeket legtöbb esetben automatikusan, nyelvészeti szabályokon alapuló konverterekkel hozzák létre: például az SPMRL-2013 Shared Task [1] versenyen használt kilenc nyelv közül csak a baszkra [2] és a magyarra [3,4] létezik kézzel annotált függőségi treebank. A konvertált függőségi treebankek minősége ennek ellenére kevéssé vizsgált.

A magyar azon ritka nyelvek egyike, ahol rendelkezésre áll kézzel annotált konstituens és függőségi annotáció is ugyanazon a szövegállományon, a Szeged (Dependencia) Treebanken [3,4]. Ez lehetővé teszi, hogy megvizsgáljuk a szabályalapú automatikus átalakítás minőségét, valamint hogy összehasonlítsuk az etalon treebankeken tanított konstituens és függőségi elemzők kimenetét a konvertált mondatokon tanított elemzők kimenetével.

Az automatikus konverziónak az elemzésre gyakorolt hatását is megvizsgáljuk. Köztudott, hogy az angol esetén egy konstituenselemző kimenetét automatikusan függőségi elemzésre konvertálva hasonló eredmények érhetőek el, mint ha függőségi elemzőt tanítanánk konstituensből automatikusan konvertált fákon. Ennek egy lehetséges magyarázata, hogy az angol mint kötött szórendű nyelv

esetén a konstituenselemzők jobban működnek a függőségi elemzőknél. Megvizsgáljuk, hogy ez a megállapítás igaz-e a magyar esetén is, ami egy tipikus szabad szórendű nyelv.

Cikkünkben három pár függőségi elemzést hasonlítunk össze, hogy megvizsgáljuk a konvertált fák használhatóságát. Először magának a konvertálásnak a hibáit vizsgáljuk a konvertált fák és a kézzel annotált etalon függőségi elemzés összehasonlításával. Másodszor összehasonlítjuk a konvertált fákon tanított függőségi elemző kimenetét az etalon fákon tanított függőségi elemző kimenetével. Harmadszor megmutatjuk, hogy az angolhoz hasonlóan a magyar esetén is igaz az, hogy a konstituens treebanken tanított elemző kimenetét függőségi elemzésre konvertálva hasonló eredményeket érhetünk el, mint a függőségi elemzőt az automatikusan konvertált fákon tanítva, bár a tipikus hibák eltérőek a két módszer esetén.

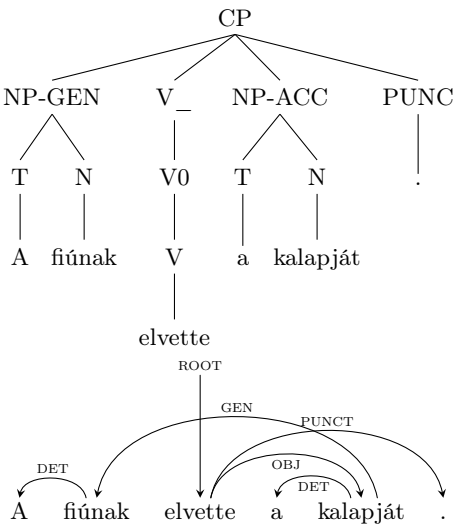
2. Magyar szintaxis a Szeged Treebankben

A magyar nyelv szintaktikai elemzése nagyban eltér az angolétól, főként a nyelv morfológiai gazdagsága és szabad szórendje miatt. Külön nehézséget okoznak a magyar esetén a távoli függőségek, például egyes birtokos szerkezetek (*A fiúnak vette el a kalapját.*), és azok a mondatok, amelyekben nem jelenik meg ige (*A kalap piros.*).

A Szeged Treebank [3] egy 82 000 mondatból álló, konstituensnyelvtan szerint kézzel annotált treebank magyarra. A frázisstruktúra mellett az igék argumentumainak nyelvtani szerepe és a szavakhoz tartozó morfológiai információ is annotálva van benne. Hat különböző doménből (üzleti rövidhírek, újságcikkek, irodalmi szövegek, fogalmazások és informatikai szövegek) származó szövegeket tartalmaz, ezek közül cikkünkben az üzleti rövidhírek alkorpuszt használtuk.

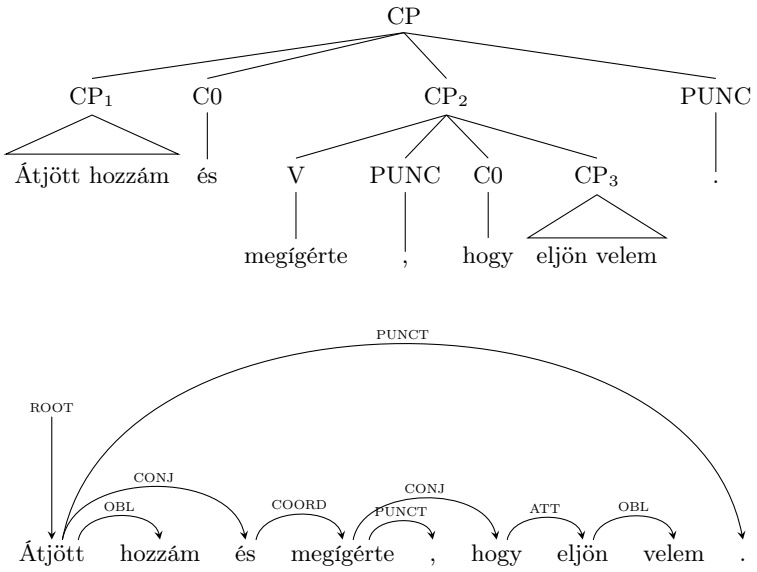
A Szeged Dependencia Treebank [4] kézi dependenciaannotációt tartalmaz ugyanezekre a szövegekre. Bizonyos nyelvtani összefüggések (például távoli függőségek) csak ebben a treebankben vannak jelölve, a konstituensben nem. A függőségi treebankben a birtokos a birtokhoz van kötve, míg a konstituenselemzés nem tartalmazza ezt az információt. A két szerkezet a 1. ábrán látható.

További különbség a két treebank között az összetett mondatok reprezentációja, ahogy a 2. ábrán látható. A függőségi treebankben az alá- és mellérendelések kezelése nagyon hasonló. Az egyik tagmondat feje (alárendelés esetén az alárendelt tagmondaté, mellérendelés esetén a második tagmondaté) a másik tagmondat fejéhez (alárendelés esetén a főmondatéhoz, mellérendelés esetén az első tagmondatéhoz) kapcsolódik, és csak a közöttük lévő reláció különbözteti meg a kétféle szerkezetet. A 2. ábrán látható mondat függőségi elemzésében a három tagmondat feje (*átjött*, *megígérte* és *eljön*) egymáshoz vannak kapcsolva a kötőszavakon keresztül ATT relációval alárendelés és COORD relációval mellérendelés esetén. A konstituens treebank eltérő szerkezetet rendel az alá- és mellérendelésekhez: alárendelés esetén az alárendelt tagmondat része a főmondatnak: a CP_3 a CP_2-n belül található a 2. ábrán. A mellérendelt tagmondatok



1. ábra: Távoli függőség kezelése konstituens- és függőségi elemzésben.

egy szinten vannak a struktúrában: az ábrán CP_1 és CP_2 mellérendelt tagmondatok.



2. ábra: A mellérendelés kezelése konstituens és függőségi elemzésben.

A kézzel annotált treebankek hasonlóságai megfelelővé teszik őket az automatikus függőségi átalakítással kapcsolatos hipotéziseink megvizsgálásához. Különbségeiket a konstituens és függőségi nyelvtanok alapvető eltérései okozzák.

3. Konstituensfák átalakítása függőségi fákká

Ebben a részben bemutatjuk a konstituensfák függőségi fákká alakításához alkalmazott módszerünket, valamint az átalakítás közben felmerült legtipikusabb hibákat.

3.1. Átalakítási szabályok

A konstituensfák függőségi fává alakításához egy szabályalapú rendszert használtunk. A virtuális csomópontokat tartalmazó mondatokat kihagytuk a vizsgálatból, mivel ezek a konstituens treebankben nincsenek külön jelölve, továbbá függőségi nyelvtanbeli kezelésük is problémás [5,6]. Így 7372 mondattal és 162960 tokennel dolgoztunk.

Első lépésben meghatároztuk a tagmondat (CP) fejét és a CP-k közötti kapcsolatokat az összetett mondatokban. A CP feje általában egy finit ige, ha a CP nem tartalmaz finit igét, akkor a fej egy főnévi igenév vagy határozói igenév, ha egyik sem található a CP-ben, akkor a fej egy névszói összetevő. A CP fejek közötti kapcsolatok alkotják a függőségi struktúra alapját: a főmondat feje ROOT relációval kapcsolódik egy absztrakt kiinduló csomóponthoz, a mellérendelt tagmondatok fejei COORD, az alárendelt tagmondatok fejei ATT relációval kapcsolódnak a főmondat fejéhez, esetleg a CP-k között lévő kötőszón keresztül, CONJ relációval.

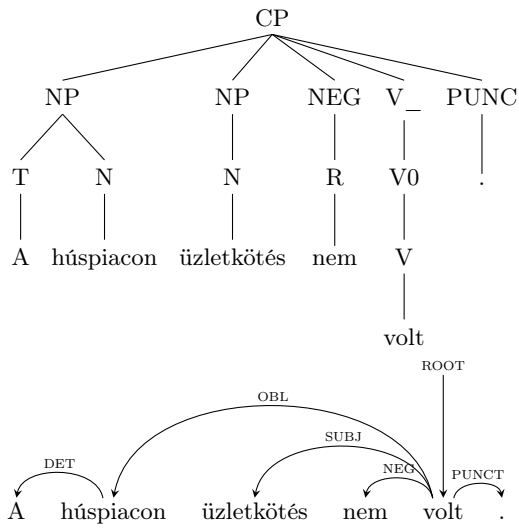
A Szeged Treebankben az igék, főnévi igenevek és határozói igenevek össze vannak kapcsolva az argumentumaikkal, azok nyelvtani szerepét is jelölve. Ezt az információt felhasználva állapítottuk meg a megfelelő függőségi relációt az igei kifejezések és argumentumaik között. A fő nyelvtani szerepek, azaz az alany, tárgy és részeshatározó, saját függőségi címkével rendelkeznek, míg az egyéb főnévi vonzatok egy összevont (OBL) relációt kapnak. Ezután az argumentumok módosítóit a fejhez vagy más módosítókhoz kapcsoltuk a frázisstruktúrájuknak és morfológiai kódjuknak megfelelően.

A távoli függőségek, mint a birtokos és birtok között lévő kapcsolat, nincsenek jelölve a konstituens treebankben. Ezekben az esetekben a morfológiai információt használtuk fel a megfelelő függőségi viszony megteremtéséhez.

A 3. ábrán egy mondat konstituensnyelvtanból függőségi nyelvtan szerinti átalakítása látható.

3.2. Hibaelemzés

A konstituens treebanket automatikusan függőségi fákká alakítottuk a honlapunkon leírt szabályoknak megfelelően (http://rgai.inf.u-szeged.hu/SzegedTreebank). A kiértékeléshez a labeled attachment score (LAS) és unlabeled attachment score



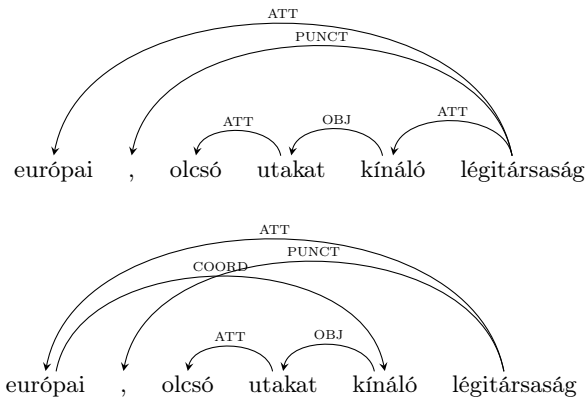
3. ábra: *A húspiacon üzletkötés nem volt* mondat konstituens annotációjának dependenciára alakítása.

(ULA) metrikákat alkalmaztunk, az írásjelek figyelembevétele nélkül. Az átalakítás pontossága 96,51 (ULA) és 93,85 (LAS). Az átalakítás hibáit az üzleti rövidhírek alkorpuszból véletlenszerűen kiválasztott 200 mondat kategorizációjával vizsgáltuk, a leggyakoribb hibák a 1. táblázat, *konvHiba* oszlopában láthatóak.

1. táblázat. Hibatípusok. konvHiba: konstituensfák függőségi fákká alakítása során vétett hibák. etalonTrain: a Bohnet parser etalon fákon tanított kimenetének hibái. silverTrain: a Bohnet parser silver standard fákon tanított kimenetének hibái. BerkKonv: etalon fákon tanított Berkeley parser kimenetének automatikus átalakítása során vétett hibák. KonvDep: függőségi címkék nélküli, silver standard fákon tanított Bohnet parser kimenetének hibái.

Hibatípus	**konvHiba**		**etalonTrain**		**silverTrain**		**BerkKonv**		**KonvDep**	
	#	%	#	%	#	%	#	%	#	%
Mellérendelés	26	13,00	39	13,22	59	14,82	55	16,37	64	19,57
Több módosító	26	13,00	30	10,17	49	12,31	52	15,48	47	14,37
Determináns	7	3,50	28	9,49	25	6,28	31	9,23	31	9,48
Kötőszó/határozószó kötés	33	16,50	23	7,80	45	11,31	39	11,61	42	12,84
Ige argumentuma	10	5,00	27	9,15	34	8,54	59	17,56	44	13,46
Alá- vagy mellérendelés	7	3,50	9	3,05	12	3,02	–	–	–	–
Birtokos	9	4,50	14	4,75	16	4,02	28	8,33	22	6,73
Rossz gyökérelem	14	7,00	17	5,76	23	5,78	35	10,42	27	8,26
Egymást követő főnevek	4	2,00	11	3,73	14	3,52	13	3,87	15	4,59
Többszavas NE	8	4,00	25	8,47	33	8,29	8	2,38	19	5,81
Rossz MOD címke	25	12,50	26	8,81	34	8,54	–	–	–	–
Egyéb rossz címke	17	8,50	33	11,19	30	7,54	–	–	–	–
Egyéb	14	7,00	13	4,41	24	6,03	16	4,76	16	4,89
Összesen	200	100	295	100	398	100	336	100	327	100

A leggyakoribb hibaforrás, ha egy frázisban egynél több módosító is volt, mint a 4. ábra mutatja. A következő ábrák mindegyikén bal oldalon, illetve felül látható az etalon elemzés, jobb oldalon, illetve alul pedig a hibás.

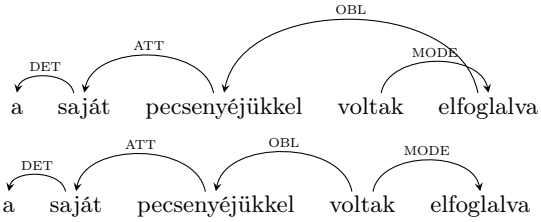


4. ábra: Több módosító miatti hiba.

Mellérendelési hibák akkor fordultak elő, amikor egy több tagból álló mellérendelés tagjai rosszul lettek összekötve. Másrészt a kötőszavak és néhány határozószó kapcsolása is problémás volt. Az 5. ábrán az *is* kötőszó az etalon elemzésben az igéhez van kötve, míg az átalakított változatban a főnévhez.

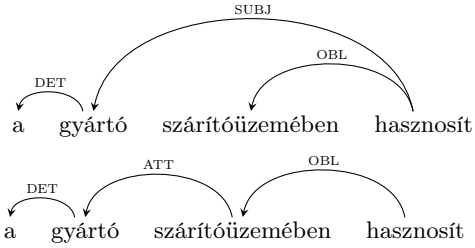
5. ábra: Kötőszó kapcsolásának hibája.

Bizonyos nyelvtani relációkat a konstituens treebank nem jelölt (például a számnevek és determinánsok egyszerűen csak az NP részei külön címkézés nélkül, mint *[NP az öt [ADJP fekete] kutya]*), de a dependencia reprezentációban ezekhez is szükséges volt szülőt és címkét rendelni. Ez nem minden esetben volt teljesen egyértelmű: például a *[NP nem [ADJP megfelelő] módszerek]* kifejezés konvertált reprezentációjában a tagadószó a melléknév helyett a főnévhez van kötve. A determináns hibák esetén a determináns rossz főnévhez lett kötve olyan NP-kben, ahol a fejet egy másik főnév módosítja. A több igei összetevőt is tartalmazó CP-k esetén (egy finit ige és egy főnévi vagy határozói igenév) az argumentumok néha rossz igei összetevőhöz kapcsolódtak, mint a 6. ábrán látható esetben.



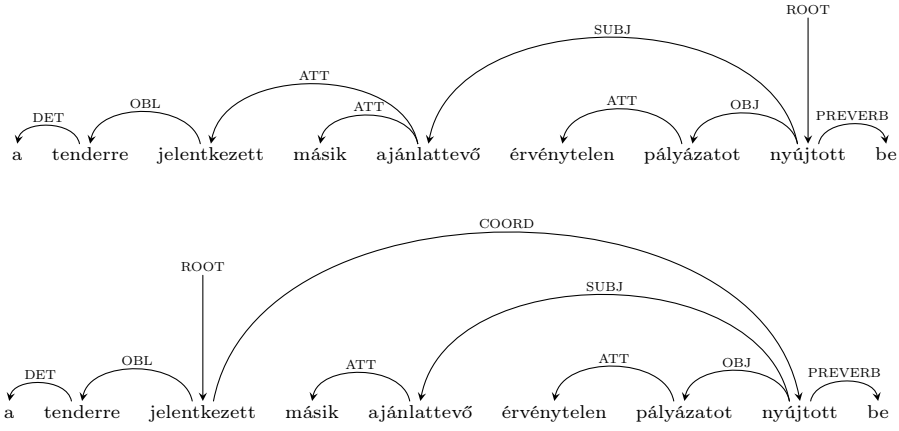
6. ábra: Ige argumentuma rossz helyre kapcsolva.

Mivel a konstituensannotációból ez hiányzik, így a birtokosok megtalálásában is előfordultak hibák, mint a 7. ábra esetén.



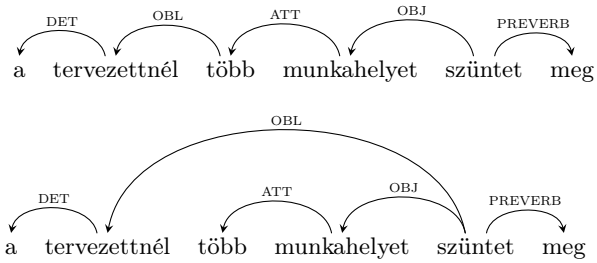
7. ábra: A birtokos kapcsolásának hibája.

Több igei összetevőt tartalmazó CP-kben nem mindig a megfelelő gyökérelem lett kiválasztva, mint a 8. ábrán.



8. ábra: Rossz gyökérelem.

Néhány esetben egymást követő, de különálló NP-k egy egységként lettek kezelve, mintha az egyik főnév a másikat módosítaná, mint a 9. ábrán.



9. ábra: Egymást követő főnevek kapcsolásának hibája.

A többszavas névelemek is okoztak átalakítási problémákat, mint a 10. ábrán látható.

10. ábra: Többszavas NE hiba.

Bizonyos esetekben a konstituens és dependencia treebankben előforduló annotációs hibák is okoztak eltéréseket az etalon és az átalakított fák között. Erre tipikus példa a rossz MODE címke hiba. A treebank a magyar határozószavakat tér- és időbeliség, valamint irányhármasság figyelembevételével megkülönbözteti, így hat külön címkével írja le ezeket a relációkat, a további határozószavak pedig egy összevont MODE relációval szerepelnek. Mivel ez a megkülönböztetés szemantikai jellegű, és gyakran hibásan lett annotálva a konstituens treebankben, ezek a hibák később a függőségi treebank annotációjában javítva lettek, így az átalakítás során hibákat okoztak, mint a 11. ábrán.

11. ábra: Rossz MODE címke.

Más hibák annyira ritkák voltak (például egy determináns hibásan az igéhez lett kötve), hogy egy kategóriába soroltuk őket, ezek a 1 táblázat „egyéb hiba" sorában láthatóak.

4. Tanítás etalon és silver standard fákon

Kísérleteztünk a Bohnet függőségi elemző [7] kézzel annotált (gold standard) és átalakított (silver standard) fákon tanításával. A Bohnet parser egy széles körben használt gráfalapú parser[1], amely perceptronra épülő online tanítást alkalmaz.

A korpuszunkból 5892 mondatot (130211 tokent) használtunk tanító adatbázisként, a megmaradt 1480 mondatot (32749 tokent) tesztelésre. A kiértékelés itt is a LAS és ULA metrikák alapján történt. Az eredmények a 2. táblázat *etalonTrain* és *silverTrain* sorában látható.

Látszik, hogy jobb eredmények érhetők el etalon adatokon tanítva, 1,6% (ULA) és 3,16% (LAS) különbséggel. A számszerű különbségeken túl itt is összehasonlítottuk a hibákat: az átalakítási hibák vizsgálatánál is használt, véletlenszerűen kiválasztott mondatokat felhasználva kézi hibaelemzést végeztünk az etalon annotációval összehasonlítva a két kapott kimenetet, l. 1. táblázat, *etalonTrain* és *silverTrain* sorok.

Bizonyos jelenségek mindkét esetben okoztak hibákat a parsernek. A mellérendelés és a több módosító például a leggyakoribb hibatípusok között van mindkét esetben, bár a hibák számát nézve észrevehetjük, hogy az etalonon tanítás mindkét esetben csökkenti a hibák számát. Emellett a kötőszó és határozószó szülő csomópontjának megtalálásában is sokat segít az etalon adat. Ennek oka valószínűleg az, hogy a konstituens treebank nem jelöli ezeket a nyelvtani kapcsolatokat, így a silver standard treebank ilyen szempontból nagyon zajos. Így összességében megállapíthatjuk, hogy bizonyos nyelvtani kapcsolatok (például a kötőszavak és határozószavak kapcsolása) kézi ellenőrzést igényel akkor is, ha automatikusan konvertálunk konstituens fákból függőségi fákat.

5. Tanítás előtt vagy tanítás után konvertáljunk?

Az angol nyelv esetén ismert, hogy egy konstituens parser kimenetét dependenciává alakítva hasonló ULA eredmények érhetőek el, mint egy átalakított fákon tanított függőségi parser használatával [8,9]. Ennek egyik lehetséges magyarázata, hogy mivel az angol kötött szórendű nyelv, így a konstituenselemzőktől jobb eredményre számíthatunk. Cikkünkben megvizsgáljuk, hogy igaz-e ez az állítás a magyar esetén, ami a morfológiailag gazdag, szabad szórendű nyelvek prototípusa.

Kísérleteinkhez a Berkeleyparser [10] nyelvtanok szorzatára épülő eljárását [11] alkalmaztuk. Ennek segítségével eltérő beállítások mellett ugyanazon tanító halmazon 8 különböző nyelvtant tanítottunk, és kiértékeléskor az egyes fák

[1] A Bohnet parsernek más parserekkel való összevető elemzéséről l. [5]. Az ott leírtak alapján a Bohnet parser érte el a legjobb eredményeket a treebanken, így jelen kísérleteinkben is ezt a parsert alkalmaztuk.

2. táblázat. A kísérletek eredménye. Konverzió: konstituensfák függőségi fákká alakítása. etalonTrain: a Bohnet parser etalon fákon tanítva. silverTrain: a Bohnet parser silver standard fákon tanítva. BerkeleyKonv: etalon fákon tanított Berkeley parser kimenetének automatikus dependenciává alakítása. KonvDep: függőségi címkék nélküli, silver standard fákon tanított Bohnet parser.

Kísérlet	**LAS**	**ULA**
Konverzió	93.85	96.51
etalonTrain	93.48	95.17
silverTrain	90.32	93.57
BerkeleyKonv	–	92.78
KonvDep	–	93.23

valószínűségét ezen nyelvtanok által adott valószínűségek szorzataként kaptuk meg. A parserek kimenetét ezután automatikusan átalakítottuk a 3. részben leírt szabályok használatával (*BerkKonv*). Emellett a silver standard függőségi treebanken a Bohnet parsert tanítottunk (*KonvDep*). Mivel a konstituensparserünk kimenetében nincsenek nyelvtani szerepek, így a megfelelő összehasonlíthatóság céljából a Bohnet parsert címkézetlen dependencia fákon tanítottuk (ez okozza a különbséget az 1. táblázat *BerkKonv* és *KonvDep* oszlopa között).

Láthatjuk, hogy a két módszer által elért eredmények hasonlóak, 92,78 és 93,23 ULA. Ez azt jelenti, hogy magyar nyelvre ugyanúgy igaz a megállapítás, mint angolra. Így – meglepő módon – azt mondhatjuk, hogy a tanítás utáni konverzió módszerével elérhető jó eredmény nem függ össze a kötött szórenddel.

Az előzőekhez hasonló kézi hibaelemzés alapján megállapítható, hogy ebben az esetben is előfordulnak hasonlóan gyakori hibatípusok, mint például a mellérendelés, a kötőszavak, módosítók és determinánsok kapcsolása. Másrészt a konstituensfákon tanítás egyéb hibákat is okozott. Egyrészt a birtokos szerkezetekben a birtokos ritkábban van helyesen a birtokhoz kötve, valószínűleg, mivel a konstituens treebankben nincs jelölve ez a kapcsolat, így a parser nem tudta megtanulni. Másrészt a több igei kifejezést tartalmazó tagmondatokban az argumentumok megfelelő helyre kapcsolása is több hibát okozott, főként határozói igenevek és főnévi igenevek esetén. A 6. ábrán látható, hogy a helyes elemzésben a *pecsenyéjükkel* főnév a határozói igenévhez van kötve, míg a másikban a főigéhez. Harmadrészt a CP fejének megtalálása is nehezebben ment ebben az esetben. A [9] cikkben a szerzők azt találták, hogy a CP fej megtalálása jobb eredménnyel történik tanítást megelőző konverzió esetén, így ez érdekes nyelvspecifikus különbségnek tűnik az angol és a magyar között. Emellett a konstituensfákon tanítás jó hatással van a többszavas NE-k felismerésére. Ezek alapján megállapíthatjuk, hogy bár számszerűen hasonló eredmények érhetők el a két módszerrel, ezek mégis különböznek egymástól a hibatípusok tekintetében.

6. Összegzés

Cikkünkben különböző módszerekkel nyert magyar nyelvű függőségi elemzéseket hasonlítottuk össze. Megállapítottuk, hogy bár a különböző módszerek számszerűleg hasonló eredményeket érnek el, különböző hibákat ejtenek. Másrészt bizonyos nyelvtani összefüggések (például mellérendelés, több módosító vagy a kötőszavak és határozószavak kötése) általában nehezek a függőségi elemző számára.

A konstituenskorpuszunk konvertálásával egy silver standard függőségi treebanken is tudtunk kísérleteket végezni. Az eredményeink azt mutatják, hogy jobb eredményeket érhetünk el az etalon annotáció használatával, ezért kívánatos a kézi függőségi annotáció. Másrészt ennek hiányában a konstituensparser kimenetének dependenciává konvertálásával vagy konstituensből átalakított dependencián tanítással az angolhoz hasonlóan jó eredmények érhetőek el, bár a hibák típusa itt is különbözik [12].

A jövőben szeretnénk megvizsgálni, hogy milyen további előnyei vannak a magyar konstituens és függőségi reprezentációjának, valamint mindkét elemző esetében szeretnénk „uptraining" kísérleteket folytatni.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J.D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Marton, Y., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A.: Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, Seattle, Washington, USA, Association for Computational Linguistics (2013) 146–182
2. Aduriz, I., Aranzabe, M.J., Arriola, J.M., Atutxa, A., Diaz de Ilarraza, A., Garmendia, A., Oronoz, M.: Construction of a Basque dependency treebank. In: Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT), Växjö, Sweden (2003) 201–204
3. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged TreeBank. In Matousek, V., Mautner, P., Pavelka, T., eds.: Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005. Lecture Notes in Computer Science, Berlin / Heidelberg, Springer (2005) 123–132
4. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)

5. Farkas, R., Vincze, V., Schmid, H.: Dependency Parsing of Hungarian: Baseline Results and Challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, Association for Computational Linguistics (2012) 55–65
6. Seeker, W., Farkas, R., Bohnet, B., Schmid, H., Kuhn, J.: Data-driven dependency parsing with empty heads. In: Proceedings of COLING 2012: Posters, Mumbai, India, The COLING 2012 Organizing Committee (2012) 1081–1090
7. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97
8. Petrov, S., Chang, P.C., Ringgaard, M., Alshawi, H.: Uptraining for accurate deterministic question parsing. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, Association for Computational Linguistics (2010) 705–713
9. Farkas, R., Bohnet, B.: Stacking of dependency and phrase structure parsers. In: Proceedings of COLING 2012, Mumbai, India, The COLING 2012 Organizing Committee (2012) 849–866
10. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. (2006) 433–440
11. Petrov, S.: Products of random latent variable grammars. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, Association for Computational Linguistics (2010) 19–27
12. Simkó, K.I., Vincze, V., Szántó, Zs., Farkas, R.: An Empirical Evaluation of Automatic Conversion from Constituency to Dependency in Hungarian. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, Dublin City University and Association for Computational Linguistics (2014) 1392–1401

Hungarian Data-Driven Syntactic Parsing in 2014

Zsolt Szántó[1], Richárd Farkas[1], Anders Björkelund[2], Özlem Çetinoğlu[2], Agnieszka Faleńska[2,3], Thomas Müller[4], Wolfgang Seeker[2]

[1] Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport, Szeged Árpád tér 2.
[2] Institute for Natural Language Processing, University of Stuttgart, Germany,
[3] Institute of Computer Science, University of Wrocław, Poland,
[4] Center for Information and Language Processing, University of Munich, Germany,
e-mail:{szantozs,rfarkas}@inf.u-szeged.hu,
{anders,ozlem,muellets,seeker}@ims.uni-stuttgart.de,
agnieszka.falenska@cs.uni.wroc.pl

1 Introduction

In prior work on data-driven syntactic parsing of Hungarian [1–4], it has been shown that parsers developed for English [5] struggle with the complexity introduced by morphologically rich languages (MRL). The Statistical Parsing of Morphologically Rich Languages (SPMRL) workshop series aims to foster the development of parsing techniques dedicated to morphologically rich languages.

In this year, the workshop hosted the SPMRL 2014 Shared Task, which was the second shared task on parsing morphologically rich languages. The challenge involves parsing both dependency and constituency representations of nine languages.

We present the contribution of the team IMS-Wrocław-Szeged-CIS, which was a joint effort of four universities. Our team achieved the best scores on all languages in the dependency track and on all languages (except for Polish) in the constituency track. In this paper, we introduce these dependency and constituency parsing systems and give extra analysis and discussions on the Hungarian treebanks.

2 The SPMRL Shared Tasks

In 2013, the organizers made the first shared task on parsing morphologically rich languages, which contains challenges in the two most commonly used syntactic frameworks (dependency and constituency) on nine morphologically rich languages (Arabic, Basque, French, German, Hebrew, Hungarian, Korean, Polish, and Swedish).

This year's shared task was an extension of the first challenge, where every annotated corpus from last year was extended with a large unlabeled data set

[3]. The system established by our team is also an extended version of the IMS-SZEGED-CIS [6] team's system, which managed to get the highest scores in every category last year.

The newspaper sub-corpora of the Szeged Treebank [7] and the Szeged Dependency Treebank [8] were employed as the Hungarian treebanks of the shared task as the organizers collected treebanks only from the newspaper domain for each language. The unlabeled data also contains newspaper articles, and came from the Hungarian National Corpus [9], which contains 1747239 sentences. We provided automatic POS-tagging and dependency parsing using magyarlanc [10] for the unlabeled data to the shared task organizers.

3 Preprocessing

The dependency parsers require POS/morphological tagging. To predict the data we use the language independent tool MarMoT[5] [11]. In Hungarian, we analyzed the word forms with the language-specific morphological analyzer of magyarlanc [10] and we use these information as features in MarMoT. We achieved 98.49 POS and 97.45 full morphological description accuracy on the development set.

4 Constituency Parsing

Our constituency parsing architecture consists of two steps. First, we deal with lexical sparsity and exploit product grammars. Second, we apply a reranker where we investigate new feature templates. In the following sections we focus on the methods to alleviate lexical sparsity and features we use in the reranker.

4.1 Lexical Sparsity

The out-of-vocabulary issue is a crucial problem in morphologically rich languages, as a word can have many different forms depending on its syntactic and semantic context. Last year, we replaced rare words by their morphological analysis produced by MarMoT [6] (similar to the strategy of backing off rare words to their POS tag in the CCG literature [12]). We call this strategy *Replace*.

This year, we experimented with an alternative approach, which exploits the available unlabeled data [2]. We followed [13] and enhanced a lexicon model trained on the treebank training data with frequency information about the possible morphological analyses of tokens (*ExtendLex*).

We note that the two strategies lead to fundamentally different representations. In the *Replace* version the output parses contain morphological descriptions instead of tokens and only main POS tags are used as preterminal labels while in the *ExtendLex* approach the tokens at the terminal level remain unchanged morphological analyses are employed as preterminal labels.

[5] `https://code.google.com/p/cistern/`

Table 1 shows the results achieved by the two strategies on the development sets. As our baselines we use the *Berkeley parser* [5] by removing morphological annotations and leaving only POS tags in preterminals (**mainPOS**), and by using full morphological descriptions (**fullMorph**). We adopt the products of respective grammars [14] as well (*ExtendLex Product* and *Replace Product*).

Table 1. PARSEVAL scores on the development set for the predicted setting.

	Hungarian
Berkeley Parser $_{mainPOS}$	83.84
Berkeley Parser $_{fullMorph}$	87.18
ExtendLex	88.99
Replace	89.59
ExtendLex Product	90.43
Replace Product	90.72

4.2 Reranker Features

The second step of our constituency pipeline is discriminative reranking. We conduct ranking experiments on the 50-best outputs of the product grammars. Like last year, we use a slightly modified version of the Mallet toolkit [15], where the reranker is trained for the maximum entropy objective function of [16] and uses the standard feature set from [16] and [17] (**dflt**). This year we investigated new feature templates exploiting automatic dependency parses of the sentence in question [18]; Brown clusters [19]; and atomic morphological feature values [2]. Our purpose here is to investigate the efficiency of these feature templates in Hungarian. For these studies we used the product grammar configuration.

The results of these feature template are shown in Table 2.

We create features from the full morphological description by using each morphological feature separately (**morph**). This approach allows us to combine a word with its morphological features (kutya-N-Cas=n). New features are established using constituency labels and morphological features of the word's head, as well as morphological features of the head and its dependent. As we only use the main POS tags in the case of the *Replace* method, these new features could only be applicable to *ExtendLex*. These new features yield a slight improvement over the *dflt* feature set (0.22 percentage points).

We also created features based on automatic dependency parsing (**dep**). These features are made from heads of constituents and their dependency relations. We used features describing relations between the same head-dependent pairs in both the constituency and dependency parses. The frequency of these relations was also used. These features are especially interesting for Hungarian because we have two manually annotated corpora in both representations as opposed to the other SPMRL languages. The results reveal that in spite of the

annotation differences, this feature template has a considerable added value. For *Replace*, the improvement is moderate, while for *ExtendLex* the result increases from 91.09 to 91.89.

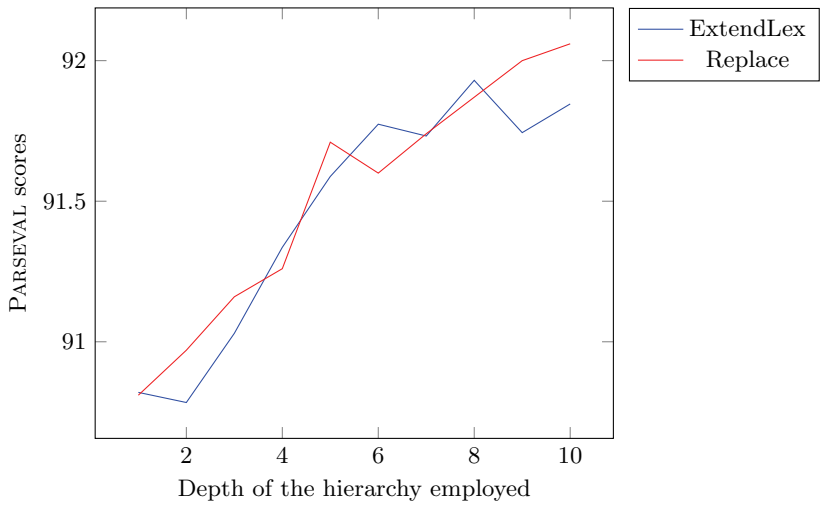


Fig. 1. Result of Brown cluster based feature templates.

We defined Brown cluster-based features (**Brown**). Brown clustering is a context-based hierarchical clustering over words. Utilizing these clusters we duplicate every other feature containing words and we replace words with their Brown clusterID (to a pre-set depth). The Brown cluster features improve our results in both representations. In the case of Brown clusters we investigated the effect of different levels of the hierarchical tree. The results achieved with *ExtendLex* are depicted on Figure 1.

Table 2 shows the final results of reranker on the development set. In the $ExtendLexReranked_{dflt+morph+Brown+dep}$ configuration we used the five level deep Brown clusters, because there was not enough time to calibrate this parameter. Reranking with default features improves the scores over product grammars both for *ExtendLex* and *Replace*. In the case of both representations the combination of feature templates slightly increases our scores.

5 Dependency Parsing

Like last year our dependency parsing system is based on two main steps. The first step is the parsing which creates the list of potential trees for each sentence and similar to constituency, the final step is reranking, which selects one of the possible trees. The full system is shown in Figure 2.

Table 2. PARSEVAL scores of the reranker on the development set for the predicted setting.

	Hungarian
ExtendLex Reranked $_{dflt}$	91.06
ExtendLex Reranked $_{dflt+morph}$	91.27
ExtendLex Reranked $_{dflt+dep}$	91.88
ExtendLex Reranked $_{dflt+Brown}$	91.93
ExtendLex Reranked $_{dflt+morph+Brown+dep}$	92.05
Replace Reranked $_{dflt}$	91.09
Replace Reranked $_{dflt+dep}$	91.89
Replace Reranked $_{dflt+brown}$	92.06
Replace Reranked $_{dflt+dep+brown}$	92.40

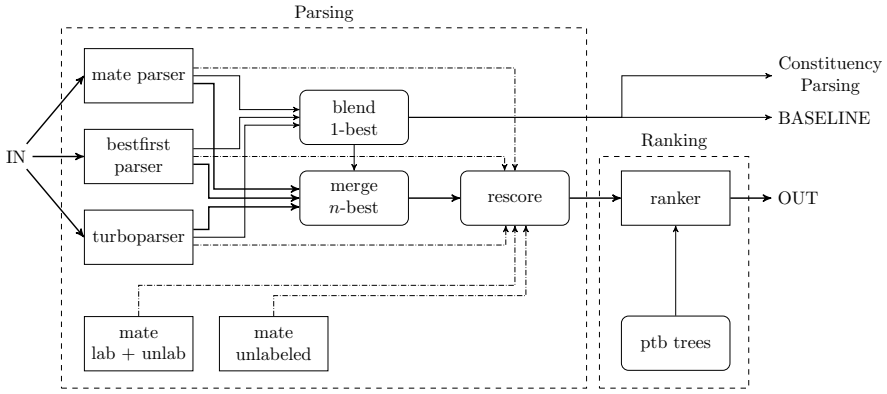


Fig. 2. The architecture of the 2014 dependency parsing system.

The parsing step is based on three different dependency parsers and a blender [20] which combines the results of the parsers. As baseline parsers with use the mate parser[6] [21], TurboParser[7] [22] and an internal implementation of the Easy-First parser [23].

In this year we used supertags which encode more syntactic information than standard POS tags. We used supertags following Ouchi et al. [24]. We made features from supertags for the mate parser and the TurboParser.

To make use of the unlabeled data we trained two self-trained models [25, 26], which are based on the mate parser. The first model was trained on unlabeled data only (**ulbl**) and in the second model we used also labeled and unlabeled data (**lbl+ulbl**).

[6] `https://code.google.com/p/mate-tools`
[7] `http://www.ark.cs.cmu.edu/TurboParser`

Table 3. UAS/LAS on Hungarian development set.

	UAS	LAS
mate	88.12	84.47
bestfirst	83.30	75.52
turbo	87.44	83.39
blend	88.09	84.24
mate_{ulbl}	86.17	82.26
$\text{mate}_{lbl+ulbl}$	88.07	84.38

Table 3 shows the result of the parsers, the last two lines show the result of the self-trained models. In Hungarian the mate parser got slightly better scores than the blending system while at other languages the blender gets great improvement compared to the standalone parsers (in Swedish more than 1%). The reason for this is the relatively bad performance of the Easy-First parser on Hungarian – in contrast to other languages.

The last step is the reranking, which chooses the best tree for each sentence from the output of the parsing step. In this step we optimized the feature sets for each language individually.

Table 4. UAS/LAS of the ranker on Hungarian development set for the predicted setting. Baseline denotes the blended trees.

	UAS	LAS
Baseline	88.09	84.24
Ranked_{dflt}	88.12	84.34
$\text{Ranked}_{no\text{-}ulbl}$	88.67	84.99
Ranked_{opt}	88.72	85.08
Oracle	91.91	8.37

Table 4 contains the results of the reranking system with different feature sets: default feature set ($Ranked_{dflt}$), optimized feature set ($Ranked_{opt}$), and optimized feature set but without features that are based on unlabeled data ($Ranked_{no\text{-}ulbl}$). The feature set optimization yields improvements, while the usage of unlabeled data only leads to minor improvements

To better understand what is special about Hungarian, we statistically analysed the output of the dependency parsing system on the development set.

From Table 5, it is striking that the labels which describe virtual nodes (containing VAN or ELL) get very low F-measures. These low scores might have two reasons, on the one hand, these relationships are relatively rare, so the parser cannot learn enough about them. On the other hand, these elements are not present in the surface structure, but they are present syntactically.

The accuracy of the parser is also poor on the FROM, TO, LOCY, TTO labels. These relations are also not too frequent and these labels contain not

Table 5. Precision, recall and F-measure of dependency relations.

Label	Recall	Prec.	F	Label	Recall	Prec.	F
APPEND	64.23	79.90	71.21	NE	92.54	90.07	91.29
ATT	93.32	94.07	93.69	NEG	94.88	93.41	94.14
ATT-VAN-CONJ	31.25	66.67	42.55	NUM	98.38	98.38	98.38
ATT-VAN-MODE	24.14	41.18	30.44	OBJ	97.64	96.13	96.88
ATT-VAN-OBL	25.93	63.64	36.85	OBL	94.91	92.07	93.47
ATT-VAN-PRED	51.35	62.30	56.30	PRED	62.61	80.90	70.59
ATT-VAN-PUNCT	37.29	61.11	46.32	PREVERB	98.44	97.83	98.13
ATT-VAN-SUBJ	44.62	53.70	48.74	PUNCT	99.03	96.24	97.62
AUX	100.00	100.00	100.00	QUE	93.33	87.50	90.32
CONJ	93.64	93.99	93.81	ROOT	85.00	88.24	86.59
COORD	75.17	79.75	77.39	ROOT-ELL-PUNCT	16.67	50.00	25.00
COORD-ELL-OBL	0.00	NaN	NaN	ROOT-VAN-ATT	16.13	35.71	22.22
COORD-ELL-PUNCT	21.05	50.00	29.63	ROOT-VAN-CONJ	76.85	74.11	75.46
COORD-VAN-CONJ	33.33	33.33	33.33	ROOT-VAN-COORD	36.36	26.67	30.77
COORD-VAN-MODE	18.75	23.08	20.69	ROOT-VAN-MODE	42.86	48.00	45.28
COORD-VAN-OBL	37.50	50.00	42.86	ROOT-VAN-NEG	60.00	46.15	52.17
COORD-VAN-PRED	46.15	48.65	47.37	ROOT-VAN-OBL	43.75	46.67	45.16
COORD-VAN-PUNCT	31.58	42.86	36.37	ROOT-VAN-PRED	69.91	63.20	66.39
COORD-VAN-SUBJ	59.46	57.89	58.66	ROOT-VAN-PUNCT	71.43	77.67	74.42
DAT	81.82	83.15	82.48	ROOT-VAN-SUBJ	60.44	56.70	58.51
DET	99.53	97.99	98.75	SUBJ	91.72	88.16	89.90
FROM	47.83	68.75	56.41	TFROM	72.73	80.00	76.19
INF	96.71	94.50	95.59	TLOCY	91.10	80.20	85.30
LOCY	50.62	73.21	59.85	TO	51.11	79.31	62.16
MODE	85.97	84.10	85.02	TTO	50.00	44.83	47.27

only syntactic but semantic information (e.g. they denote temporal or spatial dimensions) as well. Many of the phrases that get these labels are ambiguous between marking time or space, moreover, the tridirectionality in the Hungarian adverbial system may also lead to ambiguity, which makes it difficult for the parser to select the appropriate label.

Among the frequent labels ($freq > 1000$), the worst results were seen at *COORD* because coordination is a problematic phenomenon for dependency grammars in general.

Another interesting problem shows up on the POS tag level. Hungarian nouns in dative and genitive case have the same surface form, which makes POS tagging of these words difficult. The dative case usually marks an indirect object of verb, while the genitive case marks a possessive relation, and these syntactic roles are coded in different labels. The tokens with genitive case achieved an LAS of 77.08, with dative case an LAS of 78.21 while an LAS of 86,00 in case of all nouns. This example reveals a direct error propagation from the POS tagger to the dependency parser.

6 Summary

In this paper, we introduced the current state of the Hungarian data-driven syntactic parsing in dependency and constituency representations as well. We introduced the systems of the team IMS-Wrocław-Szeged-CIS, which achieved the highest scores in the SPMRL 2014 Shared Task. We also presented results on novel approaches for handling lexical sparsity in constituency parsers and we reported the added value of features in a constituency reranking framework. At the dependency parsing side, we presented a short error analysis in dependency results and highlighted Hungarian-specific challenges.

Acknowledgements

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

1. Farkas, R., Vincze, V., Schmid, H.: Dependency Parsing of Hungarian: Baseline Results and Challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. (2012) 55–65
2. Szántó, Zs., Farkas, R.: Special techniques for constituent parsing of morphologically rich languages. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden (2014) 135–144
3. Björkelund, A., Özlem Çetinoğlu, Faleńska, A., Farkas, R., Müller, T., Seeker, W., Szántó, Zs.: The IMS-Wrocław-Szeged-CIS entry at the SPMRL 2014 Shared Task: Reranking and Morphosyntax meet Unlabeled Data. In: Notes of the SPMRL 2014 Shared Task on Parsing Morphologically-Rich Languages, Dublin, Ireland (2014)
4. Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., Hajic, J.: Joint morphological and syntactic analysis for richly inflected languages. Transactions of the Association of Computational Linguistics **1** (2013) 415–428
5. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. (2006) 433–440
6. Björkelund, A., Çetinoğlu, O., Farkas, R., Müller, T., Seeker, W.: (re)ranking meets morphosyntax: State-of-the-art results from the SPMRL 2013 shared task. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, Seattle, Washington, USA (2013) 135–145
7. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In Matoušek, V., Mautner, P., Pavelka, T., eds.: Text, Speech and Dialogue: Proceedings of TSD 2005. Springer (2005)
8. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In: LREC. (2010)

9. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation. (2002) 385–389
10. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. (2013) 763–771
11. Müller, T., Schmid, H., Schütze, H.: Efficient Higher-Order CRFs for Morphological Tagging. In: Proceedings of EMNLP. (2013)
12. Clark, S., Curran, J.R.: Wide-coverage efficient statistical parsing with ccg and log-linear models. Computational Linguistics **33** (2007)
13. Goldberg, Y., Elhadad, M.: Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System. Computational Linguistics **39**(1) (2013) 121–160
14. Petrov, S.: Products of Random Latent Variable Grammars. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California (2010) 19–27
15. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
16. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05 (2005) 173–180
17. Collins, M.: Discriminative Reranking for Natural Language Parsing. In: Proceedings of the Seventeenth International Conference on Machine Learning. ICML '00 (2000) 175–182
18. Farkas, R., Bohnet, B.: Stacking of dependency and phrase structure parsers. In: Proceedings of COLING 2012, Mumbai, India (2012) 849–866
19. Brown, P.F., Della Pietra, V.J., deSouza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics **18**(4) (1992) 467–479
20. Sagae, K., Lavie, A.: Parser combination by reparsing. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York City, USA (2006) 129–132
21. Bohnet, B.: Top Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China (2010) 89–97
22. Martins, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Turbo Parsers: Dependency Parsing by Approximate Variational Inference. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA (2010) 34–44
23. Goldberg, Y., Elhadad, M.: An Efficient Algorithm for Easy-First Non-Directional Dependency Parsing. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California (2010) 742–750
24. Ouchi, H., Duh, K., Matsumoto, Y.: Improving dependency parsers with supertags. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, Gothenburg, Sweden (2014) 154–158
25. Charniak, E.: Statistical parsing with a context-free grammar and word statistics. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence. AAAI'97/IAAI'97 (1997) 598–603

26. McClosky, D., Charniak, E., Johnson, M.: Effective self-training for parsing. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, New York City, USA (2006) 152–159

Nyelvadaptáció a többszavas kifejezések automatikus azonosításában

Nagy T. István[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail: nistvan@inf.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: vinczev@inf.u-szeged.hu

Kivonat Ebben a munkában bemutatjuk gépi tanuláson alapuló nyelvfüggetlen rendszerünket különböző nyelvű félig kompozicionális szerkezeteknek automatikus azonosítására. A módszer szintaktikai elemzésen alapuló jelöltkiválasztó megközelítést használ, mely a lehetséges félig kompozicionális szerkezetekről egy gazdag jellemzőtérre támaszkodó gépi tanuló megközelítés segítségével hoz döntést. Az eredményekből kiderül, hogy a más nyelvekből származó adatok is pozitív hatással bírnak a szerkezetek azonosítására.
Kulcsszavak: többszavas kifejezések, többnyelvűség, nyelvadaptáció, 4FX

1. Bevezetés

Ebben a munkában bemutatjuk gépi tanuláson alapuló nyelvfüggetlen rendszerünket a többszavas kifejezések egy osztályának, a félig kompozicionális főnév–ige szerkezeteknek (FX-eknek) azonosítására. Megközelítésünk alapjául egy korábban ismertetett, angol és magyar nyelvre kifejlesztett kétlépéses módszerünk szolgál [1]. Első lépésben a lehetséges FX-jelölteket nyerjük ki a szövegből szintaktikai szabályok segítségével, majd a továbbiakban döntési fákon alapuló gépi tanuló módszerekkel határozzuk meg, hogy a jelölt ténylegesen többszavas kifejezés-e. Módszerünket a 4FX nevű párhuzamos korpuszon [2] teszteljük: a korpusz angol, spanyol, német és magyar nyelven tartalmaz szövegeket, melyekben kézzel be vannak jelölve a félig kompozicionális szerkezetek. A korpuszra támaszkodva bemutatjuk nyelvadaptációs kísérleteinket is, melyek igazolják, hogy az eltérő nyelvekből származó adatok is pozitívan hatnak a rendszer eredményességére.

2. Kapcsolódó munkák

Számos különböző módszer született már az FX-ek automatikus azonosítására különböző nyelveken. Az angol nyelvű kutatások [3, 4] során jellemzően csupán az adott szövegben előforduló ige–tárgy párokra fókuszálnak az FX-ek detektálása közben. Ugyanakkor a nem angol nyelvű kutatások során, mint például

holland [5], alapvetően ige-prepozíció-főnév szerkezeteket vizsgálták. Az alapvetően csak statisztikai jellemzőkre építő megközelítések [4,5] hatékonysága erősen korlátozott, hiszen ezen módszerek a ritkán előforduló FX-eket igen nehezen tudják detektálni. Ugyanakkor ezen szerkezetek nagy többsége meglehetősen ritkán fordul elő egy adott korpuszon belül [6].

A szabályalapú rendszerek [7, 8] jellemzően sekély nyelvi információkat felhasználva azonosították az FX-eket automatikusan, míg Vincze és társai [9] a SzegedParallelFX párhuzamos korpuszon mutatták be szabályalapú módszerüket magyar és angol nyelvű FX-ek azonosítására.

Statisztikai, valamint nyelvi információkat egyaránt felhasználó gépi tanuló megközelítéseket [3, 10, 11] szintén alkalmaztak FX-ek detektálására. Mindegyik módszer alapvetően ige + főnév párokat osztályoz aszerint, hogy az adott szerkezet FX-e vagy sem. Ugyanakkor a Nagy T. és társai [11] által bemutatott módszer már nem csak a szövegben előforduló ige-tárgy párokra fókuszált, hanem különböző szintaktikai jellemzők alapján automatikusan kinyert főnév + ige párokat osztályoz gazdag jellemzőtérre támaszkodó gépi tanuló módszerük alapján.

Az általunk bemutatott FX-azonosító megközelítés a [11] által demonstrált módszeren alapszik, melyet oly módon módosítottunk, hogy az képes legyen FX-ek automatikus azonosítására különböző nyelveken. Ismereteink szerint egyetlen korábban ismertett módszer sem tett kísérletet egy általános modell létrehozására, mely képes különböző nyelvű FX-ket automatikusan azonosítani.

3. Kísérletek

Célunk minden egyes FX azonosítása a 4FX korpusz [2] minden egyes nyelvén, gépi tanulási módszereket alkalmazva. Ehhez egy angol és magyar nyelvre már megalapozott módszerből [1] indulunk ki, melyet németre és spanyolra is átalakítunk az adott nyelv sajátságainak megfelelően.

Kísérleteink során azt is megvizsgáljuk, hogy a különféle nyelvek hogyan hatnak egymásra a tanítás során, így szükségesnek bizonyult egy nyelvfüggetlen reprezentáció kialakítása. Ezáltal a doménadaptációs eljárásokhoz hasonló nyelvadaptációs technikákat is ki tudunk próbálni a korpuszon.

3.1. A jelöltek kiválasztása

Az FX-jelöltek kiválasztásában [11] módszerét követjük, azaz a szintaktikailag elemzett szövegből kinyerjük az előre megadott függőségi kapcsolatok egyikét alkotó szavakat, majd ezt követően bináris osztályozás segítségével eldöntjük, hogy azok ténylegesen FX-ek-e.

Az angol, német és spanyol szövegek függőségi elemzéséhez a Bohnet parsert használtuk [12], az angol esetében a CoNLL-2008 korpuszon [13], a német esetében a TIGER treebanken [14], míg a spanyol esetében az IULA treebanken [15] tanítva. A magyar szövegek elemzéséhez a magyarlanc 2.0-t [16] alkalmaztuk a Szeged Dependencia Treebanken [17] tanítva.

Minthogy a különböző nyelvű treebankek eltérő függőségi címkéket használtak, így a 4FX korpusz különböző nyelvű változataiban is eltérő címkék szerepelnek ugyanannak a nyelvtani viszonynak a jelölésére, például az ige-tárgy kapcsolat jelölése az angolban a `dobj`, a németben az `OA`, a spanyolban a `DO` és a magyarban az `OBJ` címkével valósul meg. Így egységesítettük a nyelvtani viszonyok jelölését a nyelvek között, hasonlóan az univerzális dependenciaannotációhoz [18], azonban mi csupán az FX-ekre vonatkozó viszonyokkal foglalkoztunk.

A jelöltkinyerő fázisban FX-jelöltnek tekintettünk minden egyes ige–tárgy, ige–(passzív) alany, ige–adpozíciós frázis és főnév–igenévi módosító szókapcsolatokat.

3.2. Egységesített jellemzők

Gépi tanulási kísérleteinkhez [1] angolra és magyarra kifejlesztett módszereit vettük át, és adaptáltuk németre és spanyolra, bevezetve ezáltal néhány nyelvspecifikus jellemzőt. A nyelvfüggetlen jellemzők mellett, melyek az FX-ek nyelveken átívelő sajátságait tükrözik, az egyes nyelvekre saját jellemzőket is megadtunk, mivel az eltérő nyelvek eltérő nyelvtani sajátságokkal bírnak: például a főnevek nyelvtani neme jellemzőként szerepel a spanyol és a német nyelv esetében, ugyanakkor az angol és magyar esetében erre a jellemzőre nincs szükség.

Az egyes nyelvekre használt jellemzőket rendre megfeleltettük egymásnak, lehetővé téve ezzel a nyelvadaptációt. Például a leggyakoribb igei komponensek fordításait minden nyelvben megfeleltettük egymásnak, rendezett négyeseket képezve: *take – nehmen – tomar – vesz*. A lexikai jellemzőkhöz hasonlóan, a szintaktikai és morfológiai jegyeket is egységes nyelvfüggetlen alakra hoztuk.

Morfológiai jellemzők: megvizsgáltuk a főnevek szótövét, és bináris jellemzőként felvettük, hogy a főnév igéből képzett-e. Az FX-jelöltek tagjainak szófaját is felvettük mint jellemzőt, amennyiben az egyezett egy előre megadott lehetséges szófaji mintával, mint például `ige + főnév`.

Néhány nyelvfüggő jellemzőt is megadtunk minden egyes nyelv esetében. Az angol nyelvű morfológiai elemzés megkülönbözteti a főigéket és segédigéket, így tehát a *do* és *have* igék esetében azt is szerepeltettük, hogy azok főigei vagy segédigei használatban fordulnak elő az adott mondatban, mivel mindkét ige gyakran fordul elő FX-igeként is. A magyar nyelv morfológiailag gazdag lévén számos morfológiai jellemzőt vettünk fel a szavak morfológiai elemzése alapján mint például az igék módja, a főnevek esete, a birtokos száma és személye és a birtok száma. A nyelvtörténetileg igéből származtatott főneveket, melyeket a morfológiai elemző nem kezelt képzésként, szintén külön jelöltük.

A német és spanyol nyelv esetében a főnevek nyelvtani nemét is felvettük jellemzőként, mivel képzőiknek köszönhetően az FX-eket alkotó főnevek gyakran nőneműek ezekben a nyelvekben. Ezen túl, a német nyelvre felvettünk egy újabb jellemzőt, mely azt jelöli, hogy a főnév összetett szó-e vagy sem. A spanyol melléknévi igeneveket külön is megjelöltük végződésük alapján, mivel a morfológiai elemzés nem különbözteti meg a mellékneveket és a melléknévi igeneveket, azonban míg a melléknévi igenevek szerepelhetnek FX-ek részeként, addig a melléknevek nem.

Felszíni jellemzők: Mivel az FX-ek főnévi komponenseit gyakran képzik igéből, így a tipikus igeképzőket bi- és trigramként kezelve megvizsgáltuk, hogy az FX-jelölt főnévi komponense az adott bi- vagy trigramban végződik-e. Az FX-jelöltek szószámát szintén felvettük jellemzőként.

Statisztikai jellemzők: A jelöltkinyerő módszerrel kigyűjtöttük 10 000 angol Wikipedia-oldalból a lehetséges FX-eket, majd feljegyeztük ezek előfordulási gyakoriságait. Amennyiben az FX-jelölt megegyezett az egyik, listában szereplő egységgel, akkor jellemzőként felvettük a gyakoriságát is.

Lexikai jellemzők: Mivel általában a leggyakoribb igék fordulnak elő FX-igeként, ezért minden nyelvben kiválasztottunk 15 gyakori igét, és megvizsgáltuk, hogy az FX-jelölt igéje megegyezik-e velük. A nyelvközi méréseinkhz egyesítettük az egyes nyelvek FX-listáit, és minden egyes igét lefordítottunk mind a négy nyelvre, függetlenül attól, hogy az adott nyelvű ige éppen benne volt-e a leggyakoribb 15-ben. Így igenégyeseket kaptunk, mint például *make - machen - fazer - tesz*. Az így létrehozott lista összesen 29 igenégyest tartalmaz.

A főnevek lemmáját is jellemzőként hasznosítottuk. A parserek tanításához használt treebankekből gyűjtöttük össze az FX-ekben található főneveket.

A fentieken kívül lemmatizált FX-listákat is hasznosítottunk jellemzőként. Az angol és magyar esetében a SzegedParalellFX korpusz [19] megfelelő részéből kigyűjtött FX-eket használtuk, míg a német esetében a német PP-igei kollokációkat tartalmazó listából [20] szűrtük ki az FX-eket. A spanyol esetében az ige-főnév párokat lexikai függvények alapján kategorizáló szótár anyagából [21] indultunk ki.

Szintaktikai jellemzők: Jelöltkinyerő módszereink elsődlegesen a főnév és az ige közti szintaktikai kapcsolatra építenek, azonban a szintaktikai kapcsolatok a tényleges FX-ek kiválasztásában is hasznosíthatók. Így tehát a 3.1. részben bemutatott függőségi viszonyokat használtuk fel szintaktikai jellemzőként. Amennyiben a főnév rendelkezett névelővel, azt is jelöltük a jellemzők között.

A 1. táblázat mutatja, mely nyelvekre mely jellemzőket alkalmaztuk.

3.3. Gépi tanuláson alapuló osztályozás

[11] már korábban bemutatta, hogy ezen a feladaton a döntési fákon alapuló megközelítések teljesítenek a legjobban, ezért a WEKA gépi tanuló csomagban [22] található J48 döntési fa algoritmust tanítottuk a fentebb leírt jellemzőkészleten. Modelljeinket tízszeres keresztvalidációval értékeltük ki a korpusz minden részén.

Mivel a szintaktikai elemzésen alapuló jelöltkiválasztó megközelítés nem képes az összes manuálisan annotált FX-t kinyerni, ezért a kimaradt FX-eket téves negatívként kezeltük a kiértékelés során.

Mind a négy nyelven esetében egy kontextusfüggetlen szótárillesztési megközelítést is alkalmaztunk alapmódszernek, ahol a 3.2 fejezetben ismertetett FX-listákat alkalmaztuk. A szótárban található FX-eket abban az esetben jelöltük az adott szövegben, amennyiben azokat a szintaxisalapú jelöltkiválasztó megközelítés előzetesen kinyerte és a szövegben előfordultak.

1. táblázat. Nyelvfüggetlen és nyelvfüggő jellemzők

Jellemző	**Nyelvfüggetlen**	**Angol**	**Német**	**Spanyol**	**Magyar**
Felszíni	•	•	•	•	•
Szintaktikai	•	•	•	•	•
FX-listák	•	•	•	•	•
Igelisták	•	•	•	•	•
Főnévlisták	•	•	•	•	•
Szófaji minta	•	•	•	•	•
Igei szótő	•	•	•	•	•
Főnévképző	•	•	•	•	•
Statisztikai	–	•	–	–	–
Segédige	–	•	–	–	–
Összetett főnév	–	–	•	–	–
Nem	–	–	•	•	–
Melléknévi igenév	–	–	–	•	–
Agglutináló morfológia	–	–	–	–	•
Nyelvtörténeti képző	–	–	–	–	•

3.4. Nyelvadaptáció

Nyelvadaptációs vizsgálatainkban a doménadaptációhoz hasonló módszert használtunk. A doménadaptációs technikák alkalmazása leginkább akkor sikeres, ha egy adott doménből viszonylag kevés adat áll rendelkezésre, azonban egy másik doménből sok adathoz férünk hozzá. Esetünkben a különböző nyelveket tekintettük különböző doméneknek, így megvizsgálhattuk, hogy az eltérő nyelvekből származó adatok hogyan befolyásolják az FX-ek azonosításának eredményességét.

Többféle mérést is elvégeztünk a rendelkezésre álló korpuszon. Először mind a négy nyelven tízszeres keresztvalidációval tanítottuk és értékeltük ki a rendszert. Ezután minden egyes nyelvpár esetében keresztméréseket is alkalmaztunk, azaz a forrásnyelvet használtuk tanító adatbázisként, és a célnyelven értékeltük ki a rendszer teljesítményét. Végül nyelvadaptációs méréseket is végrehajtottunk minden egyes nyelvpár esetében, ahol a tanító adatbázis a forrásnyelvi adatok mellett a célnyelvből származó adatokat is tartalmazott kis mennyiségben, a kiértékelés pedig a többi célnyelvi adaton valósult meg. Összehasonlítási alapként szótárillesztéses méréseket is végeztünk minden egyes nyelvre.

A keresztmérésekhez elvégzéséhez az FX-jelöltek egységes reprezentációja szükséges. Ugyanakkor, ahogy a 1. táblázat is mutatja, az FX-ek különböző nyelveken való automatikus detektálásához nyelvspecifikus jellemzőket is definiáltunk, ezért az alap jellemzőkészletet kiegészítettük az összes nyelvspecifikus jellemzővel.

A nyelvadaptáció során egy egyszerű megközelítést alkalmaztunk (ADAPT): tízszeres keresztvalidációt alkalmaztunk, ahol a célnyelvből 10%-ot használtunk tesztelésre, míg a maradék 90%-t a tanítás során hozzáadtuk a forrásnyelv tanító

halmazához. Forrásnyelvnek a nyelvek összes lehetséges kombinációját alkalmaztuk, ami nem tartalmazta a célnyelvet.

A nyelvadaptáció kiértékelése során a gépi tanuló megközelítésünket a forrásnyelv és a célnyelv tesztelésre fel nem használt részének unióján tanítottuk, a kapott modelt pedig tízszeres keresztvalidációval értékelük ki a célnyelven. A keresztvalidáció során minden alkalommal a célnyelv 10%-át használtuk tesztelésre, míg a maradékot tanításra.

Az angol, német, spanyol és magyar nyelvekre végzett nyelvadaptáció eredményei a 2., 3., 4., illetve 5. táblázatokban találhatók.

4. Eredmények

Az alkalmazott módszerünk az indomén mérések során relatíve azonos eredményt ért el a korpusz magyar és angol részein 65 körüli F-mértékkel, míg spanyol és magyar nyelveken 50-es F-mértéket meghaladó eredményt ért el. A nyelvadaptáció eredményei minden korpuszon meghaladták a szótárillesztését, sőt, a spanyol nyelvet kivéve, a keresztmérések is hatékonyabbnak bizonyultak a szótárillesztésnél.

A korpusz angol részén elért eredményeket a 2. táblázat mutatja, ahol a nyelvadaptáció során mind a három másik nyelv képes volt javítani az eredményeken. A nyelvadaptáció 66,30-as F-mértéket elérve, abban az esetben bizonyult a leghatékonyabbnak, amikor a 4FX korpusz spanyol és német részének unióján tanítottuk, míg a keresztmérések esetében a német korpuszon tanított modell bizonyult a leghatékonyabbnak. A keresztmérések átlagos eredményei 7,99 F-mértékkel bizonyultak jobbnak a szótárillesztésnél, míg a nyelvadaptáció eredményei jelentősen meghaladták azt.

2. táblázat. Kísérleti eredmények a angol részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				**ADAPT**			**CROSS**		
EN	**DE**	**ES**	**HU**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				83,85	19,71	31,92	83,85	19,71	31,92
♠				**78,81**	**55,82**	**65,35**	–	–	–
♠			•	77,93	**56,60**	65,58	75,65	27,36	40,18
♠		•		79,18	56,13	65,69	67,84	21,23	32,34
♠	•			78,9	55,82	65,38	64,87	**36,01**	**46,31**
♠		•	•	81,65	54,72	65,52	81,7	28,77	42,56
♠	•		•	79,15	56,13	65,68	57,97	36,01	44,42
♠	•	•		**81,97**	55,66	**66,30**	**90,3**	23,43	37,2
♠	•	•	•	80,56	55,03	65,39	80,98	23,43	36,34
Átlag				79,91	55,73	65,65	74,19	28,03	39,91

A német részkorpuszon elért eredményeket a 3. táblázatban láthatjuk, ahol a legnagyobb volt a különbség az átlagos nyelvadaptáció eredményei és keresztmérések közt (33,11 F-mérték). Ebben az esetben akkor bizonyult legsikeresebbnek a rendszerünk, amikor a német tanítóhalmazt a 4FX korpusz spanyol részével egészítettük ki, hiszen ekkor 0,59 F-mértékkel jobb eredményt értünk el, mint az indomén eredmény. Az átlagos különbség a szótárillesztés és keresztmérések közt 3,91 volt, de a legjobb esetben a különbség 15,34 volt.

3. táblázat. Kísérleti eredmények a német részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				**ADAPT**			**CROSS**		
DE	**EN**	**ES**	**HU**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				85,71	7,45	13,71	85,71	7,45	13,71
♠				**64,52**	**41,68**	**50,64**	–	–	–
♠			•	64,4	42,44	51,17	**85,37**	**5,34**	10,06
♠		•		**65,68**	41,98	**51,23**	24,14	13,89	17,64
♠	•			64,48	41,83	50,74	23,26	**25,04**	24,12
♠		•	•	64,48	41,68	50,63	23,36	8,7	12,68
♠	•		•	62,22	41,83	50,03	37,62	23,66	**29,05**
♠	•	•		63,78	42,44	50,97	23,83	10,84	14,9
♠	•	•	•	59,93	**43,36**	50,31	22,93	10,99	14,86
Átlag				63,57	42,22	50,73	34,36	14,07	17,62

A 4. táblázat a spanyol nyelvű eredményeket mutatja. A nyelvadaptációs módszerünk akkor bizonyult a leghatékonyabbnak, amikor a tanítóhalmazt a német részkorpusszal egészítettük ki. Az eredmények azt mutatják, hogy a nyelvadaptáció elsősorban a pontosságra volt hatással, mivel ennek segítségével átlagosan 2,75 ponttal magasabb pontosságot értünk el az indomén mérésekhez képest. Ez a különbség akkor volt a legjelentősebb (4,60), amikor az angol és magyar részkorpuszok uniójával egészítettük ki a tanítóhalmazt. A nyelvadaptáció ebben az esetben is jelentősen meghaladta a szótárillesztés eredményeit, mivel annál 40,28 F-mértékkel jobb eredményt ért el, valamint az átlagos keresztmérések eredményeit is 19,05 F-mértékkel haladta meg.

A magyar részkorpuszon elért eredményeket a 5. táblázat mutatja be. A nyelvadaptáció ugyanazt a 65,25 F-mértéket érte el különböző pontosság és fedés mellett, amikor az angol részkorpusszal, valamint a német és spanyol részkorpuszok uniójával egészítettük ki. A nyelvadaptáció átlagos eredménye és az indomén átlagos eredmények különbsége 0,30 F-mérték, de módszerünk akkor tűnik hatékonyabbnak, ha a nyelvadaptáció során nem csak egy nyelvet használunk. Ugyanakkor a szótárillesztés érte el a legmagasabb pontosságértéket ezen a részkorpuszon 85,86% pontossággal, de a keresztmérések átlagosan 3,13 F-mértékkel magasabbak a szótárillesztésnél.

4. táblázat. Kísérleti eredmények a spanyol részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				ADAPT			CROSS		
ES	**EN**	**DE**	**HU**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				54,99	31,78	40,28	54,99	31,78	40,28
♠				**62,99**	**45,59**	**52,90**	–	–	–
♠			•	62,01	44,56	51,86	**51,41**	31,39	**38,98**
♠		•		65,32	45,36	**53,54**	32,47	30,13	31,25
♠	•			66,42	44,22	53,09	37,48	27,95	32,02
♠		•	•	65,21	45,02	53,26	34,62	31,84	33,17
♠	•		•	**67,59**	43,87	53,21	42,33	27,84	33,59
♠	•	•		66,77	43,87	52,95	36,32	**32,07**	34,06
♠	•	•	•	66,86	43,64	52,81	37,28	31,73	34,28
Átlag				65,74	44,36	52,96	38,84	30,42	33,91

5. táblázat. Kísérleti eredmények a magyar részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				ADAPT			CROSS		
HU	**EN**	**DE**	**ES**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				85,86	22,25	35,34	85,86	22,25	35,34
♠				**80,06**	**54,32**	**64,72**	–	–	–
♠			•	80,21	54,2	64,69	44,74	21,66	29,19
♠		•		79,38	54,44	64,58	45,67	**51,12**	48,24
♠	•			**80,64**	54,79	**65,25**	55,36	44,62	**49,41**
♠		•	•	80,13	**55,03**	**65,25**	46,12	22,49	30,23
♠	•		•	80,27	54,79	65,13	**67,17**	**21,07**	32,07
♠	•	•		80,16	54,91	65,18	49,48	44,73	46,99
♠	•	•	•	80,05	54,79	65,05	39,13	28,76	33,15
Átlag				80,12	54,71	65,02	49,67	33,49	38,47

5. Diszkusszió

Ebben a részben részletesen elemezzük a nyelven belüli és nyelvek közti, illetve nyelvadaptációval elért eredményeinket.

5.1. Nyelven belüli eredmények

Gépi tanuló megközelítésünk a 4FX korpusz minden egyes nyelve esetében jelentősen jobb eredményt ért el a szótárillesztési módszernél, ami igazolja, hogy a szintaxisra épülő gépi tanuló módszer hatékonyan működik az FX-ek automatikus azonosításában különféle nyelveken.

Módszerünk érdekes módon jobban teljesít az angol és magyar nyelveken, mint a német és spanyol nyelveken. Ennek valószínűleg a kevésbé hatékony függőségi elemzés lehet az oka, mivel a jelöltkinyerő módszer a lehetséges FX-eknek csak kisebb hányadát tudta azonosítani a németben és a spanyolban: a német FX-ek 29,01%-ánál és a spanyol FX-ek 25%-ánál a parser nem talált közvetlen szintaktikai kapcsolatot az FX-ige és főnév között. Ezzel szemben ez az arány a magyarban és angolban pusztán 10% körül található.

Általában véve is a német FX-ek azonosítása bizonyult a legnehezebb feladatnak, hiszen a szótárillesztés és a nyelvek közti mérések sem érték el a 20-as F-mértéket. Ezt feltehetőleg annak tudhatjuk be, hogy a németben viszonylag magas a csupán egyszer előforduló FX-ek aránya, ami a szótárillesztés által elért fedési értéket is erősen befolyásolja. Mindezek mellett a német FX-igék voltak a legváltozatosabbak a korpuszban, összesen 93 különböző FX-igét találhatunk a kézzel annotált FX-ekben. Ez a gépi tanulás eredményességére is kihatással volt, hiszen ezekben az esetekben kevés tanító példával találkozott a rendszer egy-egy adott szerkezetre vagy igére nézve.

5.2. Nyelvközi és nyelvadaptációs eredmények

A nyelvek közti mérések eredményei meghaladták a szótárillesztés által elért eredményeket, ami arra világít rá, hogy egy más nyelven tanított gépi tanuló modell hatékonyabbnak bizonyul, mint a célnyelvi szótárillesztés. Ez főleg annak köszönhető, hogy az elérhető szótárak mérete korlátozott volt, így alacsonyabb fedést láthattunk a szótárillesztés esetében, azonban a pontossági értékek kielégítők voltak. Ez alól egy kivételt találunk: a spanyol esetén a szótárillesztés jobban teljesített, mint a nyelvek közti mérések, főként a magas fedési értéknek köszönhetően. Ez egyrészt a nagyobb szótárméretnek tulajdonítható, másrészt pedig a szótárépítési elveknek, hiszen a szótár a lexikai függvények elméleti hátterén alapszik (vö. pl. [23]), és a 4FX korpusz annotációs elvei is részben a lexikai függvényekre hagyatkoznak.

A nyelvek közti eredmények részletesebb vizsgálata rámutat, hogy a magyar korpuszrészen tanított modell elsősorban a pontosságra volt jó hatással, míg a német modell a fedést javította. Ez valószínűleg annak köszönhető, hogy a német korpuszrészben szerepel a legtöbb fajta FX-ige, így a német adatok sokféle példát

tudnak nyújtani a lehetséges FX-ekre, és így a kevésbé gyakori célnyelvi FX-ek megtalálására is részben megoldást ad.

A 2. táblázat alapján elmondhatjuk, hogy a nyelvadaptáció minden esetben felülmúlta az angol nyelven belüli eredményeket. Mivel az angol korpuszrész tartalmazza a legkevesebb FX-et, nem meglepő, hogy a gépi tanuló modell jobb eredményt képes elérni, ha a tanító halmazba több példa kerül, még ha ezek más nyelvből származnak is.

Ha a különféle nyelveken elért eredményeket vetjük össze egymással, látszik, hogy a német alapvetően különbözik a többi nyelvtől. Itt a nyelvközi mérések nyújtották a legalacsonyabb teljesítményt, elsődlegesen a gyenge fedési értékeknek köszönhetően. Ez összefüggésben állhat a korábban már említett okokkal, nevezetesen, hogy a németben nagyon magas az egyszer előforduló FX-ek aránya, továbbá itt a legváltozatosabbak az FX-igék a négy vizsgált nyelv közül. Így tehát a más nyelvű adatokon tanított gépi tanuló modellek nem képesek megfelelő fedést elérni, mivel nincsenek olyan nagyon gyakori FX-ek, melyek lefednék az adatok jelentős hányadát. Az is látszik az adatokból, hogy az angol és magyar korpusz unióján tanított modell teljesít a legjobban a nyelvközi méréseket tekintve a német esetében. Ez a két nyelv jellegzetességeinek köszönhető: amikor csak a magyar adatokon tanítottunk, akkor értük el a legmagasabb pontosságot (85,37%), és a legjobb fedést (25,04%) akkor értük el, amikor csak angol adatokon tanítottunk.

A spanyol eredményeket tekintve észrevehetjük, hogy a legjobb fedési értéket a nyelven belüli mérés eredményezte, így a spanyol FX-ek megtalálása más nyelvű adatok alapján nehéznek bizonyul: csupán a pontossági értékek javulnak a más nyelvű tanító adatok használatával. Valószínűleg ebben a tekintetben a spanyol a némethez hasonlít: a spanyolban is viszonylag magas az egyszer előforduló FX-ek és FX-igék aránya, így a más nyelvű adatok nem tudták segíteni a gépi tanuló eljárást a ritka példák megtalálásában. Továbbá, a nyelvadaptáció eredményei is átlagosan csak 2,75 százalékponttal magasabbak, mint a nyelven belüli mérés esetében.

Ami a magyart illeti, a legsikeresebb nyelvközi kísérletnek az angol mint forrásnyelv alkalmazása bizonyult, míg az angol és spanyol adatok uniója adta a legmagasabb pontosságot. Ezt az magyarázza, hogy az angol modell is nagyon magas pontosságot ért el a nyelven belüli kísérlet során is, így az angol adatokból a modell meg tudja tanulni, hogyan válassza ki a jelöltekből a tényleges FX-eket. Mindemellett, a német adatokon tanított modell magas fedési értéket képes elérni, valószínűleg a korpuszban levő FX-ek változatossága miatt.

6. Összegzés

Ebben a munkában bemutattuk nyelvfüggetlen eljárásunkat félig kompozicionális szerkezetek azonosítására. Módszerünk első lépésben a lehetséges jelölteket nyeri ki a szövegekből szintaktikai jellemzőkre építve, majd egy gépi tanuló modell kiválasztja ezek közül a tényleges FX-eket. Eljárásunkat a 4FX korpuszon teszteltük.

A legtöbb esetben a gépi tanuláson alapuló keresztmérésekkel néhány százalékponttal jobb eredményt sikerült elérni, mint a célnyelvi szótárillesztés segítségével, például az angol nyelv esetében a különbség 8 százalékpontnyi az F-mértéket tekintve. Ez azt mutatja, hogy a gépi tanuló megközelítésünk még akkor is hatékonyabb az egyszerű szótárillesztésnél, ha a tanító halmaz és a teszthalmaz eltérő nyelvű. A nyelvadaptációval elért eredmények megközelítik, sőt bizonyos esetekben meg is haladják 0,5-1 százalékponttal a tízszeres keresztvalidációval elért eredményeket: például az angol nyelv esetében a legjobb eredményt a spanyol–német adathalmazról adaptálva értük el. Mindez arra utal, hogy a nyelvadaptációs technikák sikeresen alkalmazhatók a többszavas kifejezések automatikus azonosításában, különösen akkor, ha a célnyelven csak kis mennyiségű annotált adat áll rendelkezésre.

A későbbiekben szeretnénk az egyes nyelvek sajátságaira építve jellemzőinket bővíteni és a módszert más nyelvekre is kiterjeszteni.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Vincze, V., Nagy T., I., Farkas, R.: Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In: Proceedings of ACL 2013, Sofia, Bulgaria, ACL (2013) 255–261
2. Rácz, A., Nagy T., I., Vincze, V.: 4FX: Light Verb Constructions in a Multilingual Parallel Corpus. In: Proceedings of LREC'14, Reykjavik, Iceland, ELRA (2014)
3. Tan, Y.F., Kan, M.Y., Cui, H.: Extending corpus-based identification of light verb constructions using a supervised learning framework. In: Proceedings of MWE 2006, Trento, Italy, ACL (2006) 49–56
4. Stevenson, S., Fazly, A., North, R.: Statistical Measures of the Semi-Productivity of Light Verb Constructions. In: MWE 2004, Barcelona, Spain, ACL (2004) 1–8
5. Van de Cruys, T., Moirón, B.n.V.: Semantics-based multiword expression extraction. In: Proceedings of MWE 2007, Morristown, NJ, USA, ACL (2007) 25–32
6. Vincze, V.: Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses. PhD thesis, University of Szeged, Szeged, Hungary (2011)
7. Diab, M., Bhutada, P.: Verb Noun Construction MWE Token Classification. In: Proceedings of MWE 2009, Singapore, ACL (2009) 17–22
8. Nagy T., I., Vincze, V., Berend, G.: Domain-Dependent Identification of Multiword Expressions. In: Proceedings of RANLP 2011, Hissar, Bulgaria, RANLP 2011 Organising Committee (2011) 622–627
9. Vincze, V., Nagy T., I., Zsibrita, J.: Félig kompozicionális szerkezetek automatikus azonosítása magyar és angol nyelven. In Tanács, A., Vincze, V., eds.: VIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2011) 59–70

10. Tu, Y., Roth, D.: Learning English Light Verb Constructions: Contextual or Statistical. In: Proceedings of MWE 2011, Portland, Oregon, USA, ACL (2011) 31–39
11. Nagy T., I., Vincze, V., Farkas, R.: Full-coverage Identification of English Light Verb Constructions. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 329–337
12. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. (2010) 89–97
13. Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J.: The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning, Association for Computational Linguistics (2008) 159–177
14. Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic interpretation of a German corpus. Research on Language and Computation **2**(4) (2004) 597–620
15. Marimon, M., Fisas, B., Bel, N., Villegas, M., Vivaldi, J., Torner, S., Lorente, M., Vázquez, S., Villegas, M.: The IULA Treebank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, European Language Resources Association (ELRA) (2012) 1920–1926
16. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771
17. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010. (2010)
18. McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., Lee, J.: Universal dependency annotation for multilingual parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, Association for Computational Linguistics (2013) 92–97
19. Vincze, V.: Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In: Proceedings of LREC 2012, Istanbul, Turkey (2012)
20. Krenn, B.: Description of Evaluation Resource – German PP-verb data. In: Proceedings of MWE 2008, Marrakech, Morocco (2008) 7–10
21. Kolesnikova, O., Gelbukh, A.: Supervised machine learning for predicting the meaning of verb-noun combinations in Spanish. In: Advances in Soft Computing. Springer (2010) 196–207
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
23. Mel'čuk, I.: Esquisse d'un modèle linguistique du type "Sens<->Texte". In: Problèmes actuels en psycholinguistique. Colloques inter. du CNRS, no. 206, Paris, CNRS (1974) 291–317

Lexikális behelyettesítés magyarul

Takács Dávid[1], Gábor Kata[2]

[1]Prezi
takdavid@gmail.com
[2]INRIA
kata.gabor@inria.fr

Kivonat Cikkünkben a lexikális behelyettesítési feladat (lexical substitution) magyarra adaptálását és két különböző megoldásának tesztelését tárgyaljuk. A lexikális behelyettesítés célja olyan algoritmus megalkotása, mely képes egy lexikális egység egy-egy mondatbeli előfordulását másik egységgel helyettesíteni olyan módon, hogy a mondat eredeti jelentését a lehető legjobban megőrízze. A feladat általunk kipróbált változatában az algoritmusnak kell elvégeznie a behelyettesítésre javasolt jelöltek generálását, valamint a szövegkörnyezetbe legjobban illeszkedő lexikális egység kiválasztását. A kiértékelés során a rendszer által javasolt jelölteket annotátorok által adott válaszokkal vetjük össze. A behelyettesítési feladat magyarra alkalmazásának célja, hogy felmérjük a disztribúciós szemantikai módszerek működésének hatékonyságát, valamint - a más nyelveken végzett kísérletekkel összevetve - képet kapjunk az esetlegesen felmerülő magyar-specifikus kihívásokról: a rendelkezésre álló erőforrásokról, illetve a nyelvi jellegzetességekből adódó problémákról.

Kulcsszavak: lexikális behelyettesítés, lexikális szemantika, disztribúciós szemantika

1. Bevezetés

A lexikális szemantikai kutatások, ezen belül a disztribúciós szemantika egyre nagyobb teret nyer a számítógépes nyelvészet különböző ágaiban (pl. szinonimadetektálás, szemantikai relációk tanulása, ontológiák/lexikai adatbázisok automatikus építése, dokumentum-kategorizálás). A korpuszból kinyert vektoriális reprezentációk kiértékelésének egyik lehetséges módja az eredmények integrálása valamilyen nyelvtechnológiai alkalmazásba, ám erre nem minden esetben nyílik közvetlen lehetőség. Ennek megfelelően többféle kiértékelési feladat és gold standard létezik a témában (l. SemEval kampányok). A vektoros szemantikai reprezentációk lehetővé teszik, hogy a szavak jelentése/szemantikai tartalma közötti hasonlóságot, vagy éppen a szisztematikus eltéréseket számszerűsítsük. Egyes kiértékelési szabványok az annotátorok által megadott (szintén numerikus) szemantikai hasonlósági értékeket [21] vagy plauzibilitási ítéleteket [19] használnak. A lexikális behelyettesítés előnye az előbbi kiértékelési módszerekkel szemben, hogy az annotátorok számára természetesebb, a nyelvi tudást közvetlenebbül mozgósító feladatot jelent, és nem támaszkodik előre meghatározott jelentéstárakra vagy nyelvészeti definíciókra (szemben például a hagyományos WSD

feladattal).

A lexikális behelyettesítés [14,5] célja olyan algoritmus megalkotása, mely képes egy lexikális egység (egyszerű szó, többszavas kifejezés) egy-egy mondatbeli előfordulását másik egységgel helyettesíteni olyan módon, hogy a mondat eredeti jelentését a lehető legjobban megőrizze. A feladat általunk kipróbált változatában az algoritmusnak kell elvégeznie a behelyettesítésre javasolt jelöltek (elsősorban, de nem kizárólag szinonimák) generálását, valamint a szövegkörnyezetbe legjobban illeszkedő lexikális egység kiválasztását. A kiértékelés során a rendszer által javasolt jelölteket annotátorok által adott válaszokkal vetjük össze. A behelyettesítési feladat magyarra alkalmazásának célja, hogy felmérjük a lexikális/disztribúciós szemantikai módszerek működésének hatékonyságát, valamint a más nyelveken végzett kísérletekkel összevetve képet kapjunk az esetlegesen felmerülő magyar-specifikus kihívásokról: a rendelkezésre álló erőforrásokról, illetve a nyelvi jellegzetességekből adódó problémákról.

A lexikális behelyettesítés jellemzően két részfeladatra osztható. Az első lépés a jelöltek kinyerése egy erre alkalmas jelentés- vagy szinonima adatbázisból (általában WordNetből), illetve korpuszból disztribúciós módszerekkel, pl. vektoriális közelség szerint. Bár sok kritika fogalmazódott meg a WordNet alkalmasságát illetően (elsősorban jelentésegyértelműsítési kontextusban [25,10] illetve a magyarra [9]), az angol nyelvű lexikális behelyettesítési verseny (SemEval 2007) során a legjobbnak bizonyult módszerek mégis mind támaszkodnak a WordNetre [8,13]. A második lépés a jelöltek rangsorolása aszerint, hogy melyik illeszkedik legjobban az adott szövegkörnyezetbe. Ez a feladat közel áll a jelentésegyértelműsítéshez, ám annotált szinonima-tár hiányában nem támaszkodhatunk felügyelt tanítási módszerekre. Lesk szótári definíciókat [11], Aguirre és Rigau WordNet alapú távolsági mértékeket [1], Carrol és McCarthy szemantikai szelekciós információkat [4] használ az egyértelműsítéshez. A disztribúciós szemantikában használt vektoriális szó-reprezentációk is alkalmasak rá, hogy szavak vagy nagyobb szövegegységek közötti hasonlósági mértékeket számítsunk belőlük. Egyes kutatások látens szemantikai dimenziókat alkalmaznak a szójelentések automatikus elkülönítésére és kontextusbeli egyértelműsítésére [12,23]. A szavak elosztott reprezentációján (*distributed lexical representations* vagy *word embedding*) alapuló nyelvmodellek [16] által generált vektoriális reprezentációk is alkalmasak arra, hogy rajtuk értelmezhető közelségi metrikák alapján döntsünk a szavak szemantikai közelségéről. Ezek a módszerek több SemEval versenyen - szóhasonlósági és szóanalógiás feladatok esetében - jól teljesítettek (Semeval 2012, 2014). A word2vec [16] és a GloVe [20] módszerek a szavakhoz vagy tetszőleges nagyobb egységekhez egy valós vektortérbeli vektort rendelnek úgy, hogy az így létrejött reprezentációra két tulajdonság jellemző: egyrészt az egymáshoz közel eső szavak szemantikai, illetve morfológiai értelemben is közeliek, másrészt a vektorok közötti vektoriális különbségek konzisztensek, és egyik szópárról a másikra átvihetők. Jellegzetes példa a szópárok között kinyerhető analógiás hasonlóságra: $\mathrm{v}(king) - \mathrm{v}(queen) = \mathrm{v}(man) - \mathrm{v}(woman)$. Ez a két tulajdonság indokolja a

módszerek közvetlen használhatóságát a szószemantikai feladatokban. A behelyettesítéses feladaton legújabban Ferret [6] végzett kísérletet francia nyelvre a word2vec által generált reprezentáció felhasználásával.

Kísérletünkben létrehozunk egy ilyen vektoros reprezentációt magyar szavakra, és ennek használhatóságát mindkét részfeladatra kipróbáljuk. Másodsorban egy WordNet alapú módszerrel próbálkozunk [7], mely a WordNet-beli lemmákat, illetve a köztük definiált hierarchikus kapcsolatokból származó információt kombinálja a disztribúciós szemantika és a dokumentumkategorizálás területén használt eljárásokkal. A célszó különböző jelentéseit és az ezekhez tartozó lexikai egységeket a WordNetből nyerjük ki. A WordNet-jelentések klaszterezése után a jelentéseket körülvevő releváns csomópontok körbejárásával tematikus kategóriákat képezünk, melyekhez ezután a korpuszból gyűjtünk kategória-specifikus kontextusokat. Az egyértelműsítés során a jelöltek vektoros reprezentációját vetjük össze a kontextus szavaival. Végül egy hibrid módszert is kipróbálunk, mely a WordNetből kinyert jelölteket kizárólag korpusz alapú disztribúciós információ felhasználásával rangsorolja.

2. Erőforrások

2.1. Magyar Nemzeti Szövegtár

A disztribúciós információ kinyeréséhez a Magyar Nemzeti Szövegtár [24] (MNSZ, továbbiakban: korpusz) első, kibővített, elemzett változatát használtuk [24,18]. A korpusz ezen változata 260 millió szót tartalmaz, egyértelműsítése az MNSZ egyértelműsítő eszközlánc segítségével állt elő [17]. A kísérleteinkhez a korpusz lemmásított változatát használtuk.

2.2. WordNet

A WordNet olyan elektronikus lexikális szemantikai adatbázis, melyben a nyelvi fogalmak hálózatba szerveződnek. A fogalmakat szinonimahalmazok (synsetek), a közöttük lévő kapcsolatokat szemantikai relációk (hipernima, meronima, antonima stb.) reprezentálják. A WordNet alapegysége a szavakból álló szinonimaosztályok, úgynevezett synsetek.
A magyar WordNet [15] mintegy 40.000 synsetet tartalmaz, melyek nagy része meg van feleltetve ekvivalens angol WordNet synseteknek, így implicit módon más nyelvek wordneteinek is.

3. Jelöltek kinyerése a WordNet-ből

A célszó behelyettesítésére szánt jelölteket csoportosan nyerjük ki a WordNetből, ahol egy csoport a célszó egy jelentésének felel meg. Módszerünk célja egyfelől, hogy a különböző (és valóban megkülönböztetendő) jelentések mindegyikére találjunk jelöltet, másfelől, hogy az így kapott jelentések később hatékonyan felhasználhatók legyenek az egyértelműsítés során. Fontos azonban megjegyezni,

hogy nem közvetlen célunk a synsetek/jelentések közül választani a mondatbeli behelyettesítés során: a jelentések megkülönböztetése csak a jelölt-kinyerés és a jelentés-specifikus kontextusok kiválasztása során kerül előtérbe.

A szinonimák kinyerésének első lépéseként tehát azonosítjuk azokat a synseteket, melyek tartalmazzák a célszót. Mivel a korábbi, angol nyelvű kísérletekhez hasonlóan [8] a magyar WordNet esetében is előfordul, hogy a célszó az adott synsetből kinyerhető egyetlen szó, a keresést ilyen esetekben kiterjesztettük a hiperonimákra is[1].

Közismert, hogy a WordNet nagyon részletes jelentés-megkülönböztetésekkel operál [10]: számos olyan megkülönböztetést tartalmaz, mely az adott feladat kontextusában nem releváns, sőt, akár kifejezetten megnehezítheti a jelentés-egyértelműsítést [25]. Mivel a WordNet szerkezete önmagában nem feltétlenül nyújt információt a synsetek közti szemantikai távolságról, úgy döntöttünk, hogy a synsetek lexikális tartalmát felhasználva próbáljuk meg kiszűrni az irreleváns megkülönböztetéseket. Az egymással megegyező lexikális tartalmú synseteket tehát összevontuk, csakúgy, mint azokat a synset-párokat, melyek közül a kisebb synset részhalmazát alkotja a nagyobbnak. Az összevont synseteket a későbbiekben nem különböztetjük meg az eredeti synsetektől: valamennyit különböző jelentésként fogjuk kezelni.

4. Vektor alapú megközelítés

A word embedding előnye, hogy a szavak között többféle szemantikai viszonyt képezhetünk le egy valós vektortérben, és ezeket egyszerű numerikus, lineáris módszerekkel tárhatjuk fel. Megfigyelhető például, hogy a célszavaknak egy adott jelentéshez tartozó szinonimái többnyire egymás közelében fordulnak elő. Emiatt a tulajdonság miatt lehetőség van arra, hogy a lexikális behelyettesítési feladatot egy lépésben oldjuk meg: azokat a szavakat választjuk ki, amelyek tetszőlegesen választott szemantikai közelségmérték szerint a legközelebb esnek a célszóhoz, hiszen ezek a legalkalmasak jelöltnek. Ez a naiv megoldás egyszerű és intuitív, de nagyon jól szerepel a jelöltek generálásában - amint látni fogjuk az *oot* értékelésben (1 táblázat) - ezért baseline-nak tekinthetjük. Mi a skalárszorzatos megoldást implementáltuk *cosine* néven.

Megfigyelhetjük azt is, hogy az egy adott szemantikai mezőbe tartozó szavak szintén közelebb esnek egymáshoz. Mivel gyakran előfordul, hogy a célszóval egy mezőbe tartozó további szavak is vannak a kontextusban, ezért megkísérelhetünk a jelöltek közül aszerint választani, hogy mennyire esnek közel a kontextus szavaihoz. Precízebben, minden jelöltre kiszámolunk egy megfelelőségi mértéket úgy,

[1] A tesztadatok létrehozásakor a SemEval 2007 feladathoz hasonlóan a magyar behelyettesítési feladatban is megengedtük az általánosabb értelmű fogalommal való helyettesítést.

hogy összegezzük a jelölt és minden egyes kontextusszó közelségi mértékét, és eszerint rendezve a legjobb jelölteket adjuk vissza. Így [27] megközelítését implementáljuk. A közelségi mérték, többek között, lehet skalárszorzat és euklideszi távolság, ezeknek az alkalmazott mérték szerint *bestcosinecontext* és *bestl2context* a neve a kísérletünkben.

Lehetőségünk van arra is, hogy a célszóhoz közel eső jelöltek közül azokat részesítsük előnyben, amelyek a kontextushoz közelebb vannak, azaz amelyek a célszó vektorától a kontextus eredő vektora irányába esnek. Ez a módszer több paramétert is elfogad: a kontextus eredő vektorának kiszámításakor az egyes elemeket különféleképpen súlyozhatjuk, és megválaszthatjuk azt is, hogy milyen távolságban keressük a helyettesítő szavakat az eredetitől, azaz a kontextus és a célszó milyen lineáris kombinációját számítjuk ki. A kísérletezés során azt találtuk, hogy ha nagy súlyt adunk a kontextusnak, azaz távolabb keresgélünk az eredeti célszótól, rátalálhatunk egy-egy távolabbi szinonimára is, de nagyobb számban találunk használhatatlan (nem behelyettesíthető) szavakat is. Ezt a tulajdonságot a kiértékelő adatok is igazolják: az összes kísérleti konfigurációból az *averagecontext* szerepel az oot-kiértékelésben a legjobban (azaz nagyon jó jelölteket is talál), de a best-értékelése nem jó (azaz zajos: sok jelöltje nem megfelelő).

Ugyanezeket a számításokat elvégezhetjük tetszőleges szavakra, amelyeknek ismerjük a vektortérbeli reprezentációit. Ez lehetőséget ad egy egyszerű hibridizációra: más módszerek által generált jelölteket tudunk a fent leírt módszerekkel kiértékelni és sorbarendezni. A kiértékelésben látható, hogy a *hybrid bestcosinecontext* konfiguráció ötvözi a különböző megközelítések előnyeit, és összességében a legjobb eredményeket adja. Ebben a konfigurációban a WordNet-ből kinyert jelölteket rangsoroltuk a *bestcosinecontext* érték alapján.

5. WordNet alapú megközelítés

A WordNet alapú megközelítés motivációja a WordNet szerkezetében rejlő információ kihasználása és ötvözése a korpusz alapú megközelítés előnyeivel. A folyamat első lépésében a jelölteket a célszót tartalmazó synsetekből nyertük ki. Ezután a célszó synsetjeit csoportosítjuk, és az így kapott jelentésekhez a synsetek tartalmát felhasználva keresünk olyan kontextusokat a korpuszban, melyek az adott jelentésre specifikusak, és így megkülönböztető erővel bírnak az egyértelműsítés során. Ezek a kontextusok fogják képezni a célszó specifikus egyértelműsítő vektorterét, melyen valamennyi jelöltet elhelyezzük. Végül az utolsó lépésben a mondat szavait vetjük össze a jelölteknek az egyértelműsítő kontextusokra felvett értékeivel, és eszerint rangsoroljuk őket.

5.1. Jelentések megkülönböztetése

A WordNet nemcsak azt teszi lehetővé, hogy releváns behelyettesítési jelölteket generáljunk. A munkafolyamat következő lépéseinek célja, hogy a WordNet szerkezetét és lexikális tartalmát kihasználva összegyűjtsük azokat a kontextusokat,

melyek vélhetően releváns információt hordoznak a jelentések, és ezen keresztül a behelyettesítési jelöltek kontextusbeli egyértelműsítéséhez. Feltételezésünk szerint e megközelítés egyik előnye lehet, hogy egy tetszőleges korpuszban a szavak ritkább jelentései meglehetősen alulreprezentáltak, ezért a tisztán disztribúciós, illetve nyelvmodell alapú egyértelműsítési módszerek ezen jelentések előfordulásait nehezen tudják azonosítani. Ha azonban rendelkezésünkre áll egy lista a szó lehetséges jelentéseiről, valamint a különböző jelentésekkel kapcsolatba hozható szavakról - melyek a wordnet-hierarchiából könnyedén kinyerhetők - lehetőségünk nyílik arra, hogy olyan kontextusokat is figyelembe vegyünk az egyértelműsítés során, melyek egyébként a célszóval, illetve a behelyettesítési jelöltjeivel nem, vagy csak kevésszer fordulnak elő a korpuszban. A korábbiakban bemutatott tisztán disztribúciós alapú megközelítéssel szemben tehát a második kísérlet célja megvizsgálni, hatékony lehet-e egy külső erőforrás bevonása az egyértelműsítési folyamatba.

A jelentések megkülönböztetéséhez a 3 pontban bemutatott synset-klaszterekből indulunk ki. Célunk, hogy minden jelentést megfelelő mennyiségű lexikális elemmel tudjunk reprezentálni. Ezen lexikális elemek részben az előző pontban kinyert behelyettesítési jelöltek (szinonimák, illetve ezek hiányában hiperonimák). A következő lépésben azokhoz a jelentésekhez, melyek háromnál kevesebb lexikális elemmel hozhatók kapcsolatba, újabb szavakat kerestünk a WordNet környező synsetjeiben: minden olyan synsetben, mely közvetlenül hiperonim, category_domain, illetve mero_part relációban áll az adott synsettel (összevonás esetén a kiinduló synsetek valamelyikével). Ezen kapcsolódó synsetek lexikális tartalmát a jelentéshez csatoltuk [2].

A behelyettesítési jelölteket a későbbiekben úgy kívánjuk kiválasztani, hogy a mondatbeli kontextust összevetjük az egyes jelentésekre jellemző kontextusokkal. Ehhez létre kell hoznunk egy olyan egyértelműsítési vektorteret, mely a célszó összes különböző jelentésének leginkább jellemző kontextusait tartalmazza, és a lehető legkevésbé favorizálja a gyakori jelentéseket. A célszó jelentéseihez társított valamennyi szó minden előfordulását figyelembe véve kinyertük a korpuszból a szavak kontextusait. Szintaktikai elemzés híján kontextusként kezeltünk minden, a szó környezetében előforduló lemmát, kettő, illetve öt szóból álló kontextus-ablakok használatával (ezek a szó-ablakok bizonyultak a legeredményesebbnek [2] részletes összehasonlító vizsgálatában). Ezután minden S synset-klaszterre kikerestük azokat a w szavakat, melyek az adott synset-klaszter s szavaira leginkább jellemzőek az alábbi képlet szerint :

$$spec_{w,S} = \sum_{s \in S} weight_{s,w} \tag{1}$$

[2] Fontos megjegyezni, hogy az így kinyert új szavak már nem számítanak behelyettesítési jelöltnek, kizárólag a jelentések jellemző kontextusainak feltérképezéséhez használtuk őket.

ahol a *weight* súly a PMI (Pointwise Mutual Information) egy normalizált változata. A PMI kollokációk és jellemző kontextusok kinyerésének bevett módja:

$$PMI_{w_1,w_2} = log\frac{p(w_1,w_2)}{p(w_1)p(w_2)} \tag{2}$$

A PMI hátránya azonban, hogy erősen favorizálja a ritka kontextusokat. Esetünkben ez az egyértelműsítési feladatot megnehezíti, ezért két normalizálási eljárást is kipróbáltunk, hogy kivédjük a ritka kontextusok felülreprezentálását [3,22]:

$$NPMI_{w_1,w_2} = log\frac{p(w_1,w_2)}{p(w_1)p(w_2)}/-log\ \ p(w_1,w_2) \tag{3}$$

$$squaredPMI_{w_1,w_2} = log\frac{p^2(w_1,w_2)}{p(w_1)p(w_2)} \tag{4}$$

A jelentés-specifikus kontextusokat a fenti értékek szerint rangsoroltuk. A célszó egyértelműsítő vektortere a célszó jelentéseire jellemző kontextusok uniójából áll elő: jelentésenként a legjellemzőbb 200, illetve 500 kontextust tartottuk meg. Előfordulhat, hogy egy kontextus több jelentésre is jellemző, ezt is hasznos információnak tekintettük. A vektorterek mérete így a célszó jelentései számának, és a jelentések átfedésének függvényében változó.

Mivel feladatunk nem közvetlenül a jelentésegyértelműsítés, hanem a legmegfelelőbb jelentés kiválasztása, ezért a célszóhoz tartozó összes behelyettesítési jelöltet elhelyezzük a fenti vektortérben. Ehhez három különböző reprezentációt használtunk: a célszó és a kontextus együttes előfordulásainak számát ($freq(c,w)$), a célszó kontextus melletti relatív gyakoriságát:

$$\frac{freq(c,w)}{freq(c)} \tag{5}$$

illetve a relatív gyakoriság normalizált értékét:

$$\frac{\frac{freq(c,w)}{freq(c)} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \tag{6}$$

Az együttes előfordulásokat ugyanolyan kontextus-ablakot használva számoltuk, ahogyan az egyértelműsítő vektotereket előállítottuk.

5.2. Egyértelműsítés

A mondatbeli kontextusba illeszkedő jelölt kiválasztásakor a jelölteknek az egyértelműsítő vektortérben alkotott reprezentációját vetjük össze a mondat szavaival. Ehhez először is lemmatizáltuk a tesztmondatokat az MNSZ egyértelműsítő eszközlánc [17] segítségével. A mondat p vektora úgy áll elő, hogy a mondat i

szavait is ráképezzük a célszó egyértelműsítő vektorterére egy karakterisztikus függvénnyel:

$$p_i = \begin{cases} 1, & \text{ha i előfordul a mondatban} \\ 0 & \text{egyébként} \end{cases}$$

A jelölteket ezután a p mondat-vektor és a jelölt c egyértelműsítő vektora közti kompatibilitás szerint rangsoroljuk, melyet a következő képlet szerint számolunk ki:

$$compatibility(c, p) = c \cdot p = \sum_{i=1}^{n} c_i \times p_i \tag{7}$$

A mondat szavai közül tehát csak azokat vesszük figyelembe, melyek a célszó valamelyik jelentéséhez specifikus kontextusként lettek társítva, és azzal a súllyal esnek latba, amit az adott jelölt hozzájuk rendel a korpuszbeli előfordulásai alapján.

6. Kiértékelés és perspektívák

Az eredmények kiértékeléséhez használt adatok elkészítésekor a McCarthy és Navigli (Semeval 2007), illetve a Fabre és tsai (SemDis 2014) által követett módszert vettük alapul. Tíz poliszém főnevet választottunk, melyek minden jelentésükben rendelkeznek egytagú szinonimával. Feltétel volt továbbá, hogy maga a főnév, valamint szinonimái (jelentésenként legalább egy) is kellő mértékben reprezentálva legyenek a rendelkezésre álló korpuszban. Minden célszóhoz 10-10 példamondatot kerestünk oly módon, hogy minden szónak minden jelentése reprezentálva legyen. Az adatokat és az instrukciókat a Qualtrics online felmérés-készítő alkamazás segítségével osztottuk meg. A célszó mondatbeli előfordulásaihoz legalább 3-3 önkéntes annotátor javasolt mondatonként legfeljebb négy behelyettesíthető lexikai elemet. A rendszer által javasolt megoldásokat ezekkel a kiértékelési adatokkal vetettük össze, figyelembe véve azt is, hogy a rendszer megoldása hány annotátor javaslatai között szerepel.

Hasonlóan a korábbi lexikális behelyettesítési feladatokhoz, kétféle mértéket használtunk a gold sztenderddel való összevetéshez [14] : a *best* mérték a rendszer elsőnek rangsorolt javaslatát veszi figyelembe, míg az *oot* (*out of ten*) azt méri, hogy az első tíz javaslat között hány jó jelölt szerepel, a sorrendre való tekintet nélkül. Mivel a WordNet alapú módszerek gyakran ennél kevesebb (és csak nagyon ritkán több) jelöltből indulnak ki, ezért ebben az esetben az *oot* érték inkább a WordNet mint forrás lefedettségének indikátora. Az egy mondatra adott *best* érték azt mutatja meg, hogy a rendszer által javasolt legjobb jelölt hányszor szerepel az annotátorok megoldásai között (minél többen javasolták, annál valószínűbb, hogy erős jelölt), elosztva az annotátorok által javasolt összes megoldás számával. A rendszer *best* mutatója az összes mondatra kapott best értékek átlaga. Az *oot* érték számításakor a rendszer által javasolt első tíz jelölt

pontszámait (azaz szintén az egyes annotátorokéval egyező javaslatok pontszámát) osztjuk el az összes annotátor által tett javaslatok számával.

1. táblázat. Eredmények módszerenkénti bontásban

Módszer	**BEST**	**OOT**
veconly bestcosinecontext	**0.06702**	0.267739
veconly bestl2context	0.05913	0.25895
veconly averagecontext	0.02997	**0.29371**
veconly cosine	0.02806	0.28349
wnet.lemma2.size200.NPMI.rawcount.txt	**0.11064**	**0.23881**
wnet.lemma5.size200.NPMI.rawcount.txt	0.09560	0.23881
wnet.lemma2.size200.NPMI.relfreqnorm.txt	0.09451	0.22743
wnet.lemma2.size500.NPMI.rawcount.txt	0.09423	0.23881
wnet.lemma5.size500.NPMI.rawcount.txt	0.08731	0.23881
wnet.lemma5.size200.NPMI.relfreqnorm.txt	0.08717	0.22410
hybrid bestcosinecontext	**0.11029**	**0.24003**
hybrid bestl2context	0.07988	0.23741

Amint az 1. táblázat mutatja, az *oot* értékek elég konzisztensek a WordNet alapú jelölt-generálás esetében, ami nem meglepő, hiszen a WordNet csak kevés szó esetében adott tíznél több jelöltet. Érdekes azonban, hogy a vektor alapú megközelítések minden esetben túlszárnyalták a WordNet alapú jelölt-generálást, némileg több jó jelöltet állítva az első tízben. Összességében módszereink az esetek 40-45 százalékában képesek legalább egy jó jelöltet állítani, ami körülbelül megfelel a nemzetközi eredményeknek [5]. Ugyanakkor a tisztán vektor alapú módszerek a WordNetre támaszkodó megközelítésnél gyengébben teljesítettek a legjobb jelölt kiválasztásában, az átlagot és a legkedvezőbb beállításokat tekinve is.

A WordNet-alapú módszerek eredményei meglehetősen nagy szórást mutatnak. A figyelembe vett paraméterek közül a legnagyobb jelentősége a kontextusok kiválasztásához használt specifikussági mértéknek van: az NPMI sokkal jobban teljesít, mint a squaredPMI. A kontextusok fajtái közül a kétszavas ablak pontosabbnak bizonyult, mint az ötszavas, és az egyértelműsítő vektortér méretének növelése csökkentette az egyértelműsítés pontosságát. Összességében tehát a kevesebb, specifikusabb és közvetlenebb kontextusokból képzett információ bizonyult a leghasznosabbnak. A legjobb eredményt azonban, várakozásunknak megfelelően, a hibrid módszerrel értük el.

Eddigi munkánk természetes folytatása lehet az MNSZ2 teljes anyagának felhasználása a disztribúciós modellek számításakor. Igéretes lehetőség a hibrid megközelítés további kombinációinak kiértékelése, továbbá az optimális vektoros reprezentáció megkeresése a paraméterek finomhangolásával. További annotátori munkával lehetséges lenne a tesztanyag összekötése a már elérhető magyar jelentésegyértelműsítő korpusszal (hunwsd [26]), illetve az általunk gyűjtött goldstandard annotálása jelentésekkel.

A kísérletek során kézi és gépi munkával létrehozott adatokat szabadon elérhetővé tesszük.

Köszönetnyilvánítás

Ezúton köszönjük Oravecz Csabának az MNSZ-egyértelműsítő eszközlánc rendelkezésünkre bocsátását és a használatában nyújtott segítségét. Köszönetet mondunk továbbá minden önkéntesnek, akik közreműködtek a kiértékelési adatok létrehozásában.

Hivatkozások

1. Aguirre, E., Rigau, G.: Word Sense Disambiguation using Conceptual Density. In: Proceedings of COLING'96 (1996) 16–22
2. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the ACL Conference (2014)
3. Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference (2009) 31–40
4. Carroll, J., McCarthy, D.: Word Sense Disambiguation Using Automatically Acquired Verbal Preferences. In: Computers and the Humanities 34 (2000) 109–114
5. Fabre, C., Hathout, N., Ho-Dac, L., Morlane-Hondère, F., Muller, P., Sajous, F., Tanguy, L., Van de Cruys, T.: Présentation de l'atelier SemDis 2014: Sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In: Proceedings of the TALN 2014 Conference, Marseille, France (2014)
6. Ferret, O.: Using a generic neural model for lexical substitution (Utiliser un modèle neuronal générique pour la substitution lexicale) In: TALN-RECITAL 2014 Workshop SemDis 2014: Enjeux actuels de la sémantique distributionnelle (2014) 218–227
7. Gábor, K.: The WoDiS System - WOlf and DIStributions for Lexical Substitution (Le système WoDiS - WOLF et DIStributions pour la substitution lexicale) In: TALN-RECITAL 2014 Workshop SemDis 2014: Enjeux actuels de la sémantique distributionnelle (2014) 228–237
8. Hassan, S., Csomai, A., Banea, C., Sinha, R., Mihalcea, R.: Unt: Subfinder: Combining knowledge sources for automatic lexical substitution. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic: Association for Computational Linguistics (2007)

9. Héja, E., Kuti, J., Sass, B. Jelentésegyértelműsítés - egyértelmű jelentésítés? In: MSZNY 2009, VI. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged (2009) 348–352
10. Ide, N., Wilks, Y.: Making sense about sense. In: Word Sense Disambiguation: Algorithms and Applications, vol. 33 of Text, Speech and Language Technology. Dordrecht, The Netherlands: Springer (2006) 47–74
11. Lesk, M.: Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from a ice cream cone. In: Proceedings of SIGDOC-1986 (1986)
12. Lin, D., Pantel, P.: Concept discovery from text. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING) (2002)
13. Martinez, D., Kim, S. N., Baldwin, T.: Melb-mkb : Lexical substitution system based on relatives in context. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic: Association for Computational Linguistics (2007)
14. McCarthy, D., Navigli, R.: The English Lexical Substitution Task. Language Resources and Evaluation, 43/2 (2009) 139–159
15. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary (2008) 311–321
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Proceedings of NIPS (2013)
17. Oravecz, Cs., Dienes, P.: Efficient Stochastic Part-of-Speech tagging for Hungarian. In Proceedings of the Third International Conference on Language Resources and Evaluation (2002) 710–717
18. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) European Language Resources Association (2014)
19. Padó, S.: The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Sentence Processing. Dissertation, Saarland University, Saarbrücken (2007)
20. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation Empirical Methods in Natural Language Processing (EMNLP) (2014) (to appear)
21. Rubenstein, H., Goodenough, J.: Contextual correlates of synonymy. Communications of the ACM 8/10 (1965) 627–633
22. Thanopoulos, A., Fakotakis, N., Kokkinakis, G.: Comparative Evaluation of Collocation Extraction Metrics. In: Proceedings of the Third International Conference on Language Resources and Evaluation (2002)
23. van de Cruys, T., Poibeau, T., Korhonen, A.: Latent vector weighting for word meaning in context. In: Proceedings of the EMNLP 2011 Conference (2011) 1012–1022
24. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) European Language Resources Association (2002) 385–389
25. Véronis, J.: Sense tagging: does it make sense ? In: Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech. Frankfurt: Peter Lang (2003)
26. Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian Word-sense Disambiguated Corpus. In: Proceedings of

6th International Conference on Language Resources and Evaluation, Marrakech, Morocco (2008)
27. Zweig, G., Platt, J. C., Meek, C., Burges, C. J., Yessenalina, A., Liu, Q.: Computational approaches to sentence completion. In: 50th Annual Meeting of the Association for Computational Linguistics (ACL), Jeju Island, Korea (2012) 601–610

Szemantikus szerepek automatikus címkézése függőségi elemző alkalmazásával magyar nyelvű gazdasági szövegeken

Subecz Zoltán

Szolnoki Főiskola
5000 Szolnok, Tiszaligeti sétány 14.
subecz@szolf.hu

Kivonat: Jelen tanulmányunkban bemutatjuk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi elemző alkalmazásával. Munkánkban a vállalati vásárlások, tulajdonváltozások keretével foglalkoztunk. Jellemzőkészletünkben felszíni, morfológiai és a függőségi elemzés alapján kinyert jellemzőket használtunk fel. Ezen alapjellemzőket kiegészítettük a jellemzőkből számolt statisztikai arányokkal is. Megvizsgáltuk, hogy a modell hogyan teljesít egy gyakori célszóra önállóan, és a célszavak keretekbe összefoglalt csoportjára is.

1 Bevezetés

Az Információkinyerés egyik fontos feladata a névelemek felismerése mellett az események detektálása [15,16]. A szövegekben lévő események felismerése, analizá-lása, és hogy hogyan viszonyulnak egymáshoz időben, fontos a szöveg tartalmának megismerésében. Az események detektálása mellett fontos azok szemantikus kapcsolatainak, vagy szemantikus szerepeinek megtalálása is (szemantikus szerepek címkézése, Semantic Role Labeling, SRL). Az események és azok szemantikus szerepeinek detektálását a természetes nyelvfeldolgozás sok területén lehet hasznosítani. Például az összegzéskészítés, gépi fordítás és a válaszkeresés területein.

Munkánkban a *szemantikus szerepek címkézésével* foglalkoztunk. Ez a szemantikus kapcsolatok azonosítását jelenti egy *szemantikus kereten* belül (semantic frame). A *keretek* eseményeket írnak le azok szereplőinek szintaktikai és szemantikai megkötésein keresztül. Munkánkban a *vállalati vásárlások, tulajdonváltozások* keretével foglalkoztunk.

A *szemantikus szerepek címkézése* napjainkban a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe. Angol nyelvű szövegekre általában *konstituensfa alapú* szintaktikai elemzőt használnak az elő-feldolgozásnál az angol nyelv erősen konfiguratív tulajdonsága miatt, ahol is a legtöbb mondatszintű szintaktikai információt a szórenddel fejeznek ki. Ezzel ellentétben a magyar nyelv gazdag morfológiával és szabad szórenddel rendelkezik. A *függőségi fákkal* dolgozó elemzők különösen jól használhatóak szabad szórendű nyelvek elemzésére, így a magyarra is,

ezek ugyanis könnyebben teszik lehetővé az egymással nem szomszédos, de összetartozó szavak összekapcsolását is. Ezért mi a magyar nyelvű szövegeinkre *függőségi fákkal dolgozó elemzőt* használtunk a *magyarlanc* programcsomag segítségével [20]. A szövegek szavakra bontására, a szavak morfológiai elemzésére, szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére is ezt alkalmaztuk.

A *szerepek* a legegyszerűbb esetekben a célszó *szintaktikai kapcsolatai* voltak, de nem mindig. Sokszor a keresett szerep távol helyezkedett el a függőségi fában a célszótól, gyakran a mondat másik felében. És olyan is volt, hogy a szintaktikai kapcsolat alapján várt helyen nem a keresett szerep volt. Ez utóbbi gyakran a szintaktikai elemző hibájából adódott. Így a feladat a függőségi fában a célszótól távolabbi szerepek megkeresése és a közelebbi hamis pozitív jelöltek kiszűrése volt.

2 Kapcsolódó munkák

A *szemantikus szerepek címkézése* napjainkban a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe. Angol nyelvű szövegekre sok módszer született már, ezek általában konstituensfa alapú szintaktikailag elemzett mondatokat használnak, és mondat szinten vizsgálják az eseményeket.

Kezdetben az SRL munkákban csak igékkel foglalkoztak, az igéket önállóan vizsgálták és általános szerepeket kerestek (például Agent, Patient, Instrument). A PropBank korpusz [13] szövegeit használták fel, amiben angol nyelvű szövegekhez vannak kiemelt igék annotálva a hozzájuk tartozó szemantikus szerepekkel. A 2004-es és 2005-ös CoNLL feladatokban foglalkoztak ezzel a témával [2,4].

Később az igéket már nem önállóan vizsgálták, hanem tématerületenként csoportosították azokat (keretek). Az általános szerepek mellett már vizsgáltak doménspecifikus szerepeket is. Ehhez a FrameNet korpusz [7] szövegeit használták fel, amiben angol nyelvű szövegek vannak szemantikus szerepek szerint annotálva. Ezek is elsősorban igékkel foglalkoznak, de keresnek nem igei célszavakra is. Egy fontos alaptanulmányt készített D. Gildea és D. Jurafsky [8] az SRL témában. A Senseval-3 task [11] egyik része is a FrameNet-re alapozott SRL feladat volt. Az ACE program is más NLP feladatok mellett SRL témával is foglalkozik [1].

Xue és társa [19] a jelöltek számának csökkentésére mutatott be egy módszert. A jelöltek számát jelentősen csökkentették, miközben a fedést magasan tartották.

Koomen és társai [10] és Toutanova és társai [18] a szerepek azonosítása után a szerepek közötti kapcsolatokkal, függőségekkel foglakoztak. Azt vizsgálták, hogy a megtalált kifejezések hogyan lehetnek együtt a célszónak a szerepei.

Surdeanu és társai [17] és Pradhan és társai [14] számos SRL alapú rendszer kimenetét kombinálták egy rendszerbe.

Carreras és társai [2,3] és Surdeanu és társai [17] munkáiban, ha egy mondaton belül több célszó található, akkor ezeket nem csak egymástól függetlenül kezelték, hanem közös szerepeket is kerestek hozzájuk.

Johansson és társa [9] angol nyelvű szövegekre a konstituens alapú elemzés helyett függőségi elemzést használt.

Szemantikus szerepek címkézésére magyar nyelvű szövegekre is készültek már munkák. Farkas és társai [3] a szemantikuskeret-illesztésre *szabály alapú* módszert használtak. Mi gépi tanulásos módszert alkalmaztunk ugyanerre. A szabályalapú módszerrel ellentétben a gépi tanulásos módszer nem igényel annyi erőforrást és előfeldolgozást, és automatikusan alkalmazható más doménekre is. Ehmann és társai [4] pszichológiai témájú szövegeken szemantikus szerepek címkézésénél csak két általános szerepet keresnek: az ágens és a recipiens szerepeket (cselekvő, elszenvedő). Mi a vállalatfelvásárlások kereten belül nem csak a két általános szerepet, hanem több domén-specifikus szerepet is címkéztünk. A következő szerepeket vizsgáltuk: Vevő, Eladó, Áru, Ár, Idő. Csak az igei és főnévi igenévi célszavak szerepeit kerestük.

Az angol nyelvű szövegekre általában *konstituensfa* alapú szintaktikailag elemzett mondatokat használnak. Az előző pontban ismertetett okok miatt mi a magyar nyelvű szövegeinkre *függőségi fákkal dolgozó elemzőt* használtunk a *magyarlanc* programcsomag segítségével [20].

3 Szemantikus keretek és a szemantikus szerepek

Sok információkinyerő rendszer manapság *tárgykör (domén)* specifikus *keretekkel* dolgozik. Egy-egy tárgykör eseményeit célszerű egy *kereten* belül vizsgálni, hiszen ugyanazok a *szerepek* tartoznak minden eseményhez, ami egy adott csoporthoz tartozik. Például egy repülőjegy foglalásokat feldolgozó rendszer a következő *szerepeket* használhatja: indulási időpont, érkezési időpont, célállomás, indulási állomás, távolság, ár. Az előző rendszer *célszavai* lehetnek például: foglal, lefoglal, előjegyez, vált. Ha a célszavakat önállóan dolgozzuk fel, akkor csak kevés tanító adattal tudunk dolgozni. A célszavak *keretekben történő csoportosítása* jelentősen csökkenti ezt a problémát, hiszen a több célszó tanító adatai összeadódnak.

Munkánkban a *vállalati vásárlások, tulajdonváltozások keretével* foglalkoztunk, *igei és főnévi igenévi* célszavakhoz kerestük ki a szereplőket. A következő igei célszavakat vizsgáltuk meg az adott kereten belül: *vesz, vásárol, szerez, bekebelez, gyarapít, ad, átruház, értékesít, forgalmaz*. Valamint e célszavak minden igekötős, módbeli és időbeli változatát is. A célszavakhoz a mondatokon belül a következő szerepeket kerestük meg: *vevő, eladó*, *áru, ár, idő*.

Példák a szerepekre a vállalati vásárlások tárgykörben. A példákban vastag betűvel vannak kiemelve a *célszavak* és szögletes zárójelben a *szerepek* találhatóak. Alsóindexben szerepel az adott szerep típusa.

1. *[A svéd Electrolux]$_{Eladó}$ **eladja** [motorgyártó részlegét]$_{Áru}$ [az olasz Appliance Components Companies részvénytársaságnak]$_{Vevő}$ – tájékoztatott az Electrolux.*
2. *[A Deutsche Börse AG]$_{Vevő}$ pénteken bejelentette, hogy teljesen **megveszi** [a luxemburgi Clearstream elszámolóházat]$_{Áru}$.*
3. *[A Royal Dutch Shell csoport]$_{Vevő}$ [400 millió dollárért]$_{Ár}$ **megvenni** készül [a legnagyobb kínai offshore-földgáz- és olajmező 20 százalékát]$_{Áru}$.*
4. *[A svéd Ericsson]$_{Eladó}$ bejelentette, hogy [a német Infineonnak]$_{Vevő}$ **adja** el [chipgyártó részlegét]$_{Áru}$, [400 millió euróért]$_{Ár}$.*

5. *[A többnyire szárazföldi szállítással foglalkozó magyar tulajdonban lévő Cronus Kft.]$_{Vevő}$ [a közelmúltban]$_{Idő}$* ***megvásárolta*** *[a Magyar Államvasutak Rt.-től]$_{Eladó}$ [a debreceni székhelyű MÁV Hajdú Vasútépítő-mélyépítő Kft.-t]$_{Áru}$ – jelentette be szerdán Debrecenben a Cronus Kft. tulajdonosa.*

A példákban látszik, hogy egy szerep általában *több szóból* áll és a mondatok általában nem tartalmazzák mind az öt szerepet.

3.1 Felhasznált Korpusz

Az alkalmazásunk teszteléséhez a *Szeged Korpusznak* a rövidhírek csoportjának egy olyan változatát használtuk fel, amelyikben annotálva vannak a vállalati vásárlásokra a szemantikus szerepek. Ezek közül 1000 mondatot használtunk fel. A tanításhoz és kiértékeléshez 10-szeres keresztvalidációt alkalmaztunk.

3.2 Felhasznált programcsomagok

A feladatokat *bináris osztályozásra* vezettük vissza. Az osztályozáshoz a *Weka* programcsomagnak[1] a J48-as döntési fa elemzőjét használtuk fel. A Weka adatbányászati feladatokhoz készített gépi tanuló algoritmusok gyűjteménye. A feladathoz felhasználtuk még a magyarlanc 2.0 programcsomagot is. [20] A csomag magyar szövegek mondatra és szavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére alkalmazható.

4 A magyarlanc programcsomag elemzésének bemutatása

A magyarlanc a bemenetére érkező mondatoknak elkészíti az előző pontban leírt elemzését. A mondat minden szavához külön sorba elkészíti az elemzést (1. ábra). Minden szóról megadja a következő információkat: *sorszám, szó, lemma, szófaj, morfológiai kódok*. A sor végén megadja, hogy az adott szó melyik szóval van *szintaktikai kapcsolatban*, és hogy milyen a *kapcsolat típusa*. A szintaktikai kapcsolatok alapján a mondatok egy *függőségi fát* alkotnak.

Az elemzés után megjelenítettük *vizuális elemzővel* a szintaktikai kapcsolatok alapján a mondat függőségi fáját a program online elemzőjével[2] (2. ábra). Az elemzés és a vizuális ábrázolás egymásnak megfelelően megadja a szavak közötti *szintaktikai kapcsolatokat*.

Példa:

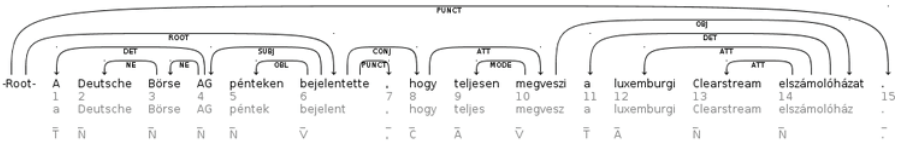
A Deutsche Börse AG pénteken bejelentette, hogy teljesen megveszi a luxemburgi Clearstream elszámolóházat.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.inf.u-szeged.hu/rgai/magyarlanc-service/

```
1       A       a       T       SubPOS=f        4       DET
2       Deutsche        Deutsche        N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     3       NE
3       Börse   Börse   N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     4       NE
4       AG      AG      N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     6       SUBJ
5       pénteken        péntek  N       SubPOS=c|Num=s|Cas=p|NumP=none|PerP=none|NumPd=none     6       OBL
6       bejelentette    bejelent        V       SubPOS=m|Mood=i|Tense=s|Per=3|Num=s|Def=y       0       ROOT
7       ,       ,       ,       _       6       PUNCT
8       hogy    hogy    C       SubPOS=s|Form=s|Coord=p 6       CONJ
9       teljesen        teljes  A       SubPOS=f|Deg=p|Num=s|Cas=w|NumP=none|PerP=none|NumPd=none       10      MODE
10      megveszi        megvesz V       SubPOS=m|Mood=i|Tense=p|Per=3|Num=s|Def=y       8       ATT
11      a       a       T       SubPOS=f        14      DET
12      luxemburgi      luxemburgi      A       SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none       14      ATT
13      Clearstream     Clearstream     N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     14      ATT
14      elszámolóházat  elszámolóház    N       SubPOS=c|Num=s|Cas=a|NumP=none|PerP=none|NumPd=none     10      OBJ
15      .       .       .       _       0       PUNCT
```

1. ábra.



2. ábra.

Az elemzésekből látszik, hogy a függőségi elemző egy *szabályos elemző fát* készít. A fa legfelső eleme a *Root*. A fa *csomópontjaiban* vannak a mondat szavai, az *ágak* a szavak közötti *szintaktikai kapcsolatokat* reprezentálják. A fában kiemelt szerepe van az *igéknek*. A *főige* (a példákban a *bejelentette*) általában a Root alatt helyezkedik el, a szintaktikai kapcsolatokon keresztül ehhez kapcsolódnak a többi elemek.

Ha a szerep *több szóból* áll, akkor ezek a szavak egy *részfát* alkotnak a mondat fáján belül. A részfa a *kiemelt szaván* (fejszó, headword) keresztül kapcsolódik a fa többi részéhez.

Van, amikor a szerep kiemelt szava (headword) a célszóhoz kapcsolódik közvetlenül. Ilyen esetben könnyebb megtalálni a szerepet. Van, amikor a szerep kiemelt szava nem a célszóhoz kapcsolódik közvetlenül. A példamondatnál a vevő kiemelt szava (AG) nem kapcsolódik szintaktikailag a *megveszi* célszóhoz az elemzőfában, hanem a *bejelentette* igén keresztül. Ilyenkor nehezebb megtalálni a szerepet. Minél közelebb van a szerep a célszótól a mondaton vagy az elemzőfán belül, annál nagyobb a valószínűsége a szerep azonosításának.

Bár a magyarlanc program elkészíti a mondatoknak a szintaktikai elemzését, de a példákon is láttuk, hogy a szintaktikai kapcsolat típusából nem következik a szemantikai szerep. Például a *vesz* célszónak az alanya a vevő, az *elad* célszónak az alanya az eladó. Így a szintaktikai kapcsolatok mellett több más tulajdonságot is meg kell figyelni a mondatban. A feladatot megnehezíti, hogy a magyarlanc elemző is hibával dolgozik, így ez a hiba és a hibákból eredő hamis döntések megjelennek a mi eredményeinkben is. Jobb eredményeket kaptunk volna, ha szövegeink kézzel lettek volna annotálva ezen szempontok szerint.

5 Az osztályozás bemutatása

A célszavakhoz a következő szerepeket vizsgáltuk: *vevő, eladó, áru, ár, idő*. Minden bemeneti mondatnál adott volt a *célszó*. A feladat az adott szerep megkeresése volt.

Az osztályozóknál a *jelöltek* a függőségi elemzőfa csomópontjai voltak. Egy mondaton belül általában egy csomópont a keresett szerep kiemelt szava (headword). Az osztályozásnál ezek a *true* esetek, a többi csomópont pedig a *false* eset.

Az osztályozáshoz *bináris osztályozót* használtunk. Az osztályozó az adott mondatnál bejelöli a keresett szerepet. Az osztályozónak *nem adtuk meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem. Voltak olyan mondatok is, amelyek nem tartalmazták a keresett szerepet. (1. táblázat)

A kiértékelésnél *szigorú szabályt* alkalmaztunk: csak azt a döntést fogadtuk el, amelyik pontosan az annotált szerepet jelöli meg. Sem az ezt tartalmazó fákat, sem ennek a részfáit nem fogadtuk el pozitív döntésnek. Ha ennél enyhébb szabályt alkalmaznánk, akkor magasabb eredményeket kapnánk.

5.1 Jellemzőkészlet

A tanító és a kiértékelő halmazon a *jelöltekhez jellemzőket* vettünk fel. Az SRL feladatokban használt általános jellemzőket [8] mi is alkalmaztuk. Ezeken kívül újakkal is kibővítettük a jellemzőkészletünket. Ehhez felhasználtuk a *függőségi elemzőfát* is, a jelölt és a célszó viszonyát a függőségi fában, mert ez gyakran egy fontos tulajdonsága az adott szerepnek.

A jelöltekhez a *következő jellemzőket* választottuk ki:

Felszíni jellemzők: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok. *Pozíció*: a jelölt a célszó előtt vagy után áll a mondatban. *Távolság-mondatban*: a jelölt és a célszó szótávolsága a mondaton belül.

Morfológiai jellemzők: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltunk. Jellemzőként definiáltuk az eseményjelöltek MSD-kódját felhasználva a következő morfológiai jegyeket: *típus*(SubPos), *mód*(Mood), *eset*(Cas), *idő*(Tense), *személy*(PerP), *szám*(Num), *határozottság*(Def). *Szófaj, Lemma*: a jelölt és a célszó szófaja és lemmája.

Jellemzők az elemzőfa alapján-1: Ide azokat a jellemzőket soroltuk, amelyeket az SRL feladatokhoz általában felhasználnak [8]. A jelölt és a célszó viszonyát vizsgáltuk a függőségi elemzőfában. Mindkettő egy-egy csomópont az elemzőfában. *Szófaj-útvonal*: Egymás után írtuk a jelölt és a célszó közötti csomópontok szófaját, feljegyezve azt is, hogy az elemzőfában felfelé, vagy lefelé haladtunk az adott kapcsolatnál. Például: C↑S↑V↑C↑V↑V↓V↓N↓N↓A. *Uralkodó-kategória-szófaja*: A jelölt és a célszó közötti útvonalon megkerestük a legmagasabban fekvő csomópontot, és feljegyeztük a hozzá tartozó szó szófaját.

Jellemzők az elemzőfa alapján-2: Itt az egyéni, új jellemzőket soroltuk fel. *Jelölt-célszó-távolság-elemzőfában*: A jelölt és a célszó csomópontjai közötti csomópontok száma az elemzőfában. *Lemma-útvonal*: Mint a Szófaj-útvonal, de itt a jelölt és a célszó között végigmenve a csomóponti szavak lemmáját jegyeztük fel. Például: Budapesti↑Értéktőzsde↑honlap↑közöl↓megvásárol. *Szintaktikai-kapcsolat-útvonal*: Az

előzőhöz hasonlóan itt azt vettük fel, hogy a jelölt és a célszó között az elemzőfában milyen szintaktikai kapcsolatokon keresztül tudunk eljutni. Például: ↑COORD*SUBJ↓ATT↓INF↓OBJ↓ATT. *Jelölt-alatti-részfában-van-e-névelem*: A magyarlanc program az elemzésében jelöli, ha talált névelemeket a mondatban. Mivel a vállalati tulajdonváltozások témakörében gyakran találkozunk vállalati névelemekkel, ezért felvettük, hogy a jelölt, vagy az alatta levő részfa tartalmaz-e névelemet? *Jelölt-alatti-részfában-névelem-távolság*: az előzőhöz hasonlóan megadtuk a részfában azt a mélységet, ahol először találtunk névelemet.

5.2 Statisztikai arány felhasználása az osztályozásnál

A jelöltekhez a jellemzőket *két módszer* alapján választottuk ki. *Első módszernél* az előző részben bemutatott alapjellemzőket használtuk fel. *Második módszernél* az alapjellemzők helyett a tanító adatokon a jellemzőkészletből számított statisztikai arányokat használtuk fel: a tanító halmaz alapján megszámoltuk minden jellemző esethez, hogy hány alkalommal fordult elő és ebből hányszor volt a jelölt *pozitív*. Ezek alapján kiszámítottuk a hozzá tartozó pozitív-arányt. Például ha a *Jelölt-lemma* jellemzőnél a *jelölt-lemma = Corp.* eset 11-szer fordult elő és ebből 7-szer volt *pozitív* eset (4-szer pedig negatív), akkor hozzá a 0,64-es pozitív-arány tartozott. Ebben az esetben az osztályozónak a jelöltekhez nem az alapjellemzőt, hanem a hozzá tartozó arányt adtuk meg. Az előző példánál *Jelölt-lemma-arány* = 0,64. Ezzel *jelentősen csökkentettük az osztályozó vektorterének méretét* az első módszerhez képest és így a futási időt is. Ez a kidolgozási időszakban hasznos volt. *Harmadik esetben* az előző két módszer jellemzőit együtt használtuk fel.

A statisztikai-arány jellemzők hatása az osztályozás eredményére. Megvizsgáltuk, hogy az előzőleg bemutatott *statisztika-arány jellemzők* hogyan befolyásolják az osztályozási eredményeinket. Először az osztályozást lefuttattuk csak a statisztikai-arány jellemzőkkel, majd csak az alapjellemzőkkel és végül a két jellemzőcsoporttal együtt. Azt tapasztaltuk, hogy az alapjellemzőkkel eset önállóan általában jobban teljesített, mint a statisztikai-arány eset önállóan. De a *legjobb eredményt* akkor kaptuk, amikor az alapjellemzőket és a statisztikai-arány jellemzőket együtt használtuk.

5.3 Vektortér méretének csökkentése

A *vektortér méretét csökkentettük* a következő módszerrel: csak azokat a *jellemző-előfordulásokat* vettük fel az osztályozáshoz, amelyek a tanító halmazon *legalább háromszor* szerepeltek. Ezzel *jelentősen csökkentettük a futási időt* és csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytuk ki.

5.4 Célszavak csoportosítása a kereten belül

Először a modell viselkedését egy gyakori célszóra önállóan néztük meg. Ehhez a *vásárol* célszót választottuk ki.

Majd a célszavakat csoportosítottuk. A vásárlásokkal kapcsolatos mondatoknál a *vevő* és az *eladó* szerepek viselkedését meghatározza, hogy az adott célszónál az alany általában vevő vagy eladó. Ezért a célszavakat két csoportra bontottuk a következő egyszerű módszerrel. A *vevő-centrikus* csoportba azok a szavak kerültek, amelyeknél az alany általában a *vevő*: vesz, vásárol, szerez, bekebelez, gyarapít. Az *eladó-centrikus* csoportba pedig azok, amelyiknél az alany általában az *eladó*: ad, átruház, értékesít, forgalmaz. Ez a felosztás segítette a *vevő* és az *eladó* szerepek megtalálását. Egy harmadik esetben pedig nem végeztünk csoportosítást.

5.5 Baseline mérések

A Baseline módszereket a *döntési fa legfontosabb feltételei alapján* állítottuk össze.

Azokat a jelölteket vettük pozitívnak, amelyekre teljesül:

Az *Áru szerepnél* azokat, amelyek tárgy (OBJ) szintaktikai kapcsolatban vannak a célszóval.

Az *Ár szerepnél* azokat, amelyeket egy előre elkészített pénznemek lista tartalmazott.

Az *Idő szerepnél* azokat, amelyeket a következő lista tartalmazott: évszámok 1990-2014-ig, hónapnevek, napnevek, sorszámok 1-31-ig.

A *vevő-centrikus célszavaknál a Vevő szerepnél* és az *eladó-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek alany (SUBJ) kapcsolatban vannak a célszóval.

A *vevő-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek végén a következő trigramok állnak: tól, től, ból, ből.

Az *eladó-centrikus célszavaknál a Vevő szerepnél* azokat, amelyek részes eset (DAT) kapcsolatban vannak a célszóval.

A következő eredményeken látni fogjuk, hogy gépi tanulási modell jóval *felülteljesítette a Baseline modellünket*.

5.6 Statisztikai adatok

Mondatok száma összesen: 1000 db. Azon mondatok száma, amelyek tartalmazzák az adott szerepet:

1. táblázat. Statisztikai adatok (db).

Célszavak	**Mondatok száma**	**Vevő**	**Eladó**	**Áru**	**Ár**	**Idő**
kiemelt: *vásárol*	265	263	107	276	104	99
Vevő-centrikus	548	531	222	573	214	208
Eladó-centrikus	452	261	374	459	82	115
csoportosítás nélkül	1000	783	579	1025	299	312

Az osztályozónak *nem adtuk meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem. (Az Áru szerep azért nagyobb, mint a mondatok száma, mert volt olyan mondat, ahol több áru szerepelt.)

6 Eredmények

6.1 Baseline mérések eredményei

2. táblázat. Baseline mérések eredményei.

Szerep	**Pontosság**	**Fedés**	**F-mérték**
Vevő-centrikus célszavak			
Vevő	48,24	59,73	**53,37**
Eladó	54,77	72,13	**62,26**
Áru	73,25	73,25	**73,25**
Ár	67,33	96,02	**79,16**
Idő	34,74	57,89	**43,42**
Eladó-centrikus célszavak			
Vevő	78,18	44,10	**56,39**
Eladó	42,63	47,50	**44,93**
Áru	77,47	72,97	**75,15**
Ár	62,64	93,44	**75,00**
Idő	23,95	46,51	**31,62**

6.2 Eredmények a *vásárol* kiemelt célszóra

3. táblázat. Eredmények a *vásárol* kiemelt célszóra (%).

Szerep	**Pontosság**	**Fedés**	**F-mérték**
Vevő	69,88	49,77	**57,63**
Eladó	82,10	60,30	**68,70**
Áru	80,72	77,11	**78,70**
Ár	90,38	83,02	**85,78**
Idő	78,75	52,82	**61,27**
Átlag:	**80,37**	**64,60**	**70,71**

6.3 Eredmények a vevő-centrikus célszavakra

4. táblázat. Eredmények a vevő-centrikus célszavakra (%).

Szerep	Pontosság	Fedés	F-mérték
Vevő	76,01	57,33	**65,09**
Eladó	79,57	66,15	**71,58**
Áru	79,18	80,93	**79,94**
Ár	87,78	80,07	**82,47**
Idő	83,13	63,89	**71,26**
Átlag	**81,13**	**69,67**	**74,07**

A 3. és a 4. táblázat eredményeit összehasonlítva látható, hogy ha a hasonló viselkedésű célszavakat *egy csoportban kezeltük*, akkor majdnem minden esetben jobb eredményeket értünk el, mint ha a célszavakat önállóan vizsgálnánk. Ennek oka, hogy a több célszó több mondatot és jelöltet ad meg és a több jelölt jellemzőiből általánosabb szabályokat tudott készíteni az osztályozó. A modell legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig a *Vevő* szerepre teljesített.

6.4 Eredmények az eladó-centrikus célszavakra

5. táblázat. Eredmények az eladó-centrikus célszavakra (%).

Szerep	Pontosság	Fedés	F-mérték
Vevő	74,59	66,82	**70,13**
Eladó	68,97	48,51	**56,35**
Áru	85,92	82,16	**83,64**
Ár	83,64	63,87	**71,58**
Idő	76,38	51,78	**59,86**
Átlag	**77,90**	**62,63**	**68,31**

Az eladó-centrikus esetben a modell legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig az *Eladó* szerepre *teljesített*.

6.5 Eredmények a célszavak csoportosítása nélkül

6. táblázat. Eredmények a célszavak csoportosítása nélkül (%).

Szerep	Pontosság	Fedés	F-mérték
Vevő	76,93	60,11	**67,05**
Eladó	72,04	50,39	**59,13**
Áru	83,62	80,24	**82,01**
Ár	88,47	76,26	**81,77**
Idő	85,44	64,53	**73,14**
Átlag	**81,30**	**66,31**	**72,62**

A *célszavak csoportosításától* azt vártuk volna, hogy a Vevő-centrikus célszavaknál a Vevő szerepre, az Eladó-centrikus célszavaknál pedig az Eladó szerepre jobb eredményt kapunk, mint a csoportosítás nélküli esetben. Ez az Eladó szerepre nem teljesült. Ennek egyik oka, hogy az Eladó-centrikus mondatokban az Eladó szerep sokszor távolabb volt a célszótól az elemzőfában. Másik oka, hogy a Vevő-centrikus célszavaknál az Eladó szerepre jó eredményt kaptunk (*71,58-es F-mérték*) a *JeloltVegenBigram* és a *JeloltVegenTrigram* jellemzők hatására. Ez a jó eredmény javította erősen a csoportosítás nélküli esetben is az Eladó szerep eredményét. Így a jobb eredményt a csoportosítás nélküli esetre kaptunk (*72,62-es F-mérték*).

Az eredmények összehasonlítás a kapcsolódó munkákkal. Angol nyelvű szövegekre Gildea és társa [8] sok keretre és azokon belül sok szerepre végezték el a feladatot. Elsősorban igékkel foglalkoznak, de keresnek nem igei célszavakra is. Ezek átlagolt eredményére 63%-os F mértéket kaptak. Eredményeink (*72,62% F-mérték átlag*) jónak számítanak annak ellenére, hogy mi csak egy keretet és ahhoz csak öt főszerepet vizsgáltunk, és csak igei és főnévi igenevekhez kerestünk szerepeket.

7 Összegzés

Munkánkban bemutattunk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi elemző alkalmazásával. A *vállalati vásárlások, tulajdonváltozások* keretével foglalkoztunk. Ezen a kereten belül 1000 annotált mondatot dolgoztunk fel és a következő szerepeket kerestük: *Vevő, Eladó, Áru, Ár, Idő*. *Jellemzőkészletünkben* felszíni, morfológiai és a függőségi elemzés alapján kinyert jellemzőket használtunk fel. Ezen alapjellemzőket kiegészítettük a jellemzőkből számolt *statisztikai arányokkal* is. Megvizsgáltuk, hogy a statisztikai jellemzők hogyan befolyásolják a modell hatékonyságát. Megvizsgáltuk, hogy a modell hogyan teljesít *egy gyakori célszóra önállóan*, és a *célszavak keretekbe összefoglalt csoportjára* is. A mérésekhez célszavainkat csoportosítottuk több szempont szerint. Bár munkánkban a vizsgált szövegek kevesebb témát fedtek le, mint az angol nyelvű szövegekre bemutatott munkák, de eredményeink jónak számítanak a bemutatott angol munkák eredményeivel összehasonlítva.

Hivatkozások

1. Ahn, D.: The stages of event extraction. In: Proceedings of the Workshop on Annotating and Reasoning about Time and Events (ARTE) (2006) 1–8
2. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL) (2004) 152–164
3. Carreras, X., Marquez, L., Chrupała, G.: Hierarchical recognition of propositional arguments with perceptrons. In: Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL), Boston, MA (2004) 106–109

4. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2005 shared task: semantic role labeling. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL) (2005) 89–97
5. Ehmann, B., Lendvai, P., Mihăltz, M., Vincze, O., László, J.: Szemantikus szerepek a narratív kategoriális elemzés (NARRCAT) rendszerében. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 121–123
6. Farkas, R., Konczer, K., Szarvas, Gy.: Szemantikus keret illesztés és az IE-rendszer automatikus kiértékelése. In: II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2004) 49–53
7. Fillmore, C.L., Ruppenhofer, J., Baker, C.F.: Framenet and representing the link between semantic and syntactic relations. In: Huang, Ch. and Lenders, W, eds.: Frontiers in Linguistics, volume I of Language and Linguistics Monograph Series B, Institute of Linguistics, Academia Sinica, Taipei (2004) 19–59
8. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics Journal 28/3, (2002) 245–288
9. Johansson, R., Nugues, P.: Semantic structure extraction using nonprojective dependency trees. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), Prague, Czech Republic (2007) 227–230
10. Koomen, P., Punyakanok, V., Roth, D., Yih, W.: Generalized inference with multiple semantic role labeling systems. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), Ann Arbor, MI. (2005) 181–184
11. Litkowski, K. C.: SENSEVAL-3 TASK: Automatic Labeling of Semantic Roles. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (2004)
12. Márquez, L., Carreras, X., Litkowski, K.C., Stevenson, S.: Semantic Role Labeling: An Introduction to the Special Issue. Computational Linguistics 34/2 (2009) 145–159.
13. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics 31/1 (2005) 71–105
14. Pradhan, S., Hacioglu, K., Ward, W., Martin, J.H., Jurafsky, D.: Semantic role chunking combining complementary syntactic views. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), Ann Arbor, MI. (2005) 217–220
15. Subecz, Z.: Detection and Classification of Events in Hungarian Natural Language Texts. Proceedings of the 17th International Conference, TSD 2014, Brno, Czech Republic (2014), Springer Lecture Notes in Computer Science 8655 (2014) 68–75
16. Subecz, Z, Nagyné, Cs.É.: Igei események detektálása és osztályozása magyar nyelvű szövegekben. In: Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2014) 237–247
17. Surdeanu, M., Marquez, L., Carreras, X., Comas, P.R.: Combination strategies for semantic role labeling. Journal of Artificial Intelligence Research (JAIR) 29 (2007) 105–151
18. Toutanova, K., Haghighi, A., Manning, C.: Joint learning improves semantic role labeling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI (2005) 589–596
19. Xue, N., Palmer, M.: Calibrating features for semantic role labeling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain (2004) 88–94
20. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 368-374

III. Morfológia, korpusz

Mennyiségből minőséget: Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében

Oravecz Csaba, Sass Bálint, Váradi Tamás

MTA Nyelvtudományi Intézet
e-mail:{oravecz.csaba,sass.balint,varadi.tamas}@nytud.mta.hu

Kivonat A Magyar Nemzeti Szövegtár egymilliárd szavas új változatának fejlesztése során egyrészt a szövegek mennyiségéből, másrészt a nyelvi elemzés minőségével kapcsolatos elvárásokból adódóan számos olyan feldolgozási kérdés merült fel, melyekre nem lehetett a jelenleg hozzáférhető nyelvi elemző eszközök „polcról levett" alkalmazásával kielégítő választ adni. A tanulmány azokat a megoldásokat és javaslatokat mutatja be, melyek hozzájárulnak ahhoz, hogy az olyan jelentős méretű korpuszokban is, ahol a manuális hibajavítás nem lehetséges, az annotáció minősége megfeleljen a felhasználói elvárásoknak.

Kulcsszavak: korpusz-előfeldolgozás, tokenizálás, morfológiai elemzés, szófaji egyértelműsítés

1. Bevezetés

Nagy méretű szövegkorpuszok előállításakor két kritikus dimenzió határozza meg a fejlesztés körülményeit: a mennyiség és a minőség. Az utóbbi időben mindkét irányban jelentős előrelépések történtek, egyrészt részletes, „mély" elemzést tartalmazó szöveggyűjtemények jelentek meg, másrészt szinte mindennapossá vált a milliárd szavas méret [1,2,3,4]. A két követelmény között nem kézenfekvő az ideális kompromisszum, amit egy további fontos tényező is nagyban befolyásol, a korpusz majdani felhasználóinak igényei. Az MNSz éppen ebből a szempontból speciális helyzetű, egyszerre kívánja kiszolgálni a számítógépes alkalmazásokat, a nyelvészeti kutatásokat és a nyelv iránt érdeklődő nagyközönséget is. Ennek következtében az új változat elkészítésekor számos kihívással kellett szembenézni, amire a peremfeltételekhez legjobban illeszkedő megoldásra volt szükség. A dolgozat azokat a fejlesztés során felmerült problémákat és javasolt megoldásokat ismerteti részletesen, melyek véleményünk szerint tanulságosak és hasznos információval szolgálnak azok számára, akik magyar nyelvi szövegeket nagy mennyiségben kívánnak nyelvtechnológia eszközökkel feldolgozni. A Szövegtár fejlesztésének általános kérdéseit megelőzően már tárgyalja [5] és [6] is. A jelen tanulmány azonban ezeken túlmutatva azokat a problémákat fejti ki részletesebben és más hangsúlyokat meghatározva, melyek leginkább érdeklődésre számítanak a magyar kutatóközösség körében.

2. Előfeldolgozás, normalizálás

A korpuszépítés egyik fontos lépése a beszerzett forrásszövegek előszűrése, feldolgozása addig a célszerűen sztenderd formátumig, amely alkalmas arra, hogy a nyelvi elemző rendszer bemeneteként szolgáljon. Az új MNSz ezzel kapcsolatos munkálatai alapvetően a szokásos, a régi szövegtárban is (részben) elvégzett feladatokat jelentették (forrásszöveg kivonása, nyelvazonosítás, duplikátumok eltávolítása, bekezdés szintig kódolt XML formátumra alakítás stb.), ezért itt csupán azt a problémát tárgyaljuk részletesebben, amely részben gyakorlati, kényszerű szempontok miatt új kihívást jelentett.

Az elektronikus szövegek túlnyomó része manapság UTF-8 karakterkódolású. Az Unicode szabvány által rendelkezésre bocsájtott szimbólumhalmaz tág teret ad azonban azoknak a típusú „visszaéléseknek", ahol egyes szövegek, szövegrészek valamilyen megjelenítési, formázási okból nem kanonikus karaktereket használnak. Ez a típusú információ, gyakorlatilag procedurális *markup*, ideális esetben természetesen leválasztandó, elkülönítendő a kanonikus tartalomtól. Az MNSz esetében ezt a fajta megjelenítési információt nem őrizzük meg. Nem egyértelmű azonban, hogy az Unicode szabvány által meghatározott normalizáló algoritmusok közül bármelyik is közvetlenül alkalmazható lenne. A szövegek változatossága és mennyisége nem teszi lehetővé, hogy minden egyes szövegegységre megvizsgáljuk, hogy vajon elegendő-e, ha a kanonikus ekvivalencia elvének megfelelő normalizáló formát választunk, vagy fennállnak-e a feltételei annak, hogy az esetenként a karakterek szemantikáját is befolyásoló, enyhébb kompatibilitás ekvivalenciát biztosító formát alkalmazzuk [7].

Az alkalmazott adatvezérelt megoldást végül az is jelentős mértékben meghatározta, hogy a 3.1. részben részletezett okok miatt a szövegek további átalakítására, egy ISO-8859-2 kódolásra történő konverzióra is szükség volt. Ennek folyamán az ISO-8859-2 kódtáblán kívüli karakterek szabványos XML numerikus entitásokra képződtek le. Ezek gyakorisági listáját vetettük alá egy manuális vizsgálatnak, melynek segítségével kialakítottunk egy olyan egyedi leképezést, amely a benne szereplő entitásokat, amennyiben lehetséges, a nyelvi szempontból ekvivalens absztrakt karakter ISO-8859-2 kanonikus alakjára képezi le. Ez a leképezés magában foglalja az Unicode kódtábla azon teljes tartományait, melyek a vizsgálat alapján előszeretettel használatosak procedurális markupként (pl. teljes- vagy félszélességű karakterek), függetlenül attól, hogy minden elemük konkrétan előfordult-e a szövegekben, illetve a leképezhető XML néventitásokat is (1. ábra). Az ezen kívül eső karakterek maradtak XML numerikus entitások, ily módon az eredeti UTF-8 szövegek a korpusz szempontjából releváns információ elvesztése nélkül voltak konvertálhatók.[1]

[1] Ez véleményünk szerint célravezetőbb, mint valamely sztenderd transzliterációs megoldás (például az *iconv* segédprogram TRANSLIT opcióval) alkalmazása, amely esetében lényegi információ (pl. ékezet) veszhet el az átalakítás során.

```
Entitás    Latin2 karakter
&#x01c4;        DŽ
&#x01c5;        Dž
&#x01c6;        dž
&#x01ca;        NJ
&#x01cb;        Nj
&#x01cc;        nj
...
&Uacute;        Ú
&Ucirc;         Ű
&Udblac;        Ű
&Uuml;          Ü
&Yacute;        Ý
```

1. ábra. Normalizáló leképezés részlet.

3. Elemzés és annotáció

Mind a korpusz méretéből, mind a leginkább a felhasználói igények által képviselt minőségi kényszerből természetszerűleg adódnak feldolgozási nehézségek a nyelvi elemzés minden szintjén. Ez megköveteli olyan elemző eszközök használatát, melyek rugalmasak, robusztusak és jól testre szabhatók, egyedi igényekre alakíthatók. Az annotált korpuszokban az elemzés magáért az elemzésért van, ez alapvetően különbözik attól, amikor valamilyen további alkalmazásban van az eszközök kimenete beágyazva. Mások a követelmények, a kiértékelés alapja pedig az elemzés minősége, és nem egy befogadó alkalmazás teljesítménye. Itt elkerülhetetlenül merül fel az a kérdés, hogy van-e készen kapható, „polcról levehető" és a célnak megfelelő magyar nyelvtechnológiai elemzőkészlet, illetve milyen mértékben használhatók egyes komponensek az adott feladat elvárt szintű megoldására.

A továbbiakban három szokványos alapvető elemzési lépést vizsgálunk: a tokenizálást/szegmentálást, a morfológiai elemzést és a szófaji egyértelműsítést. Mindhárom esetben meghatározzuk azokat a kívánalmakat, amelyeket a bevezetőben említett peremfeltételek mellett a felhasználandó eszköznek teljesítenie kell, megvizsgáljuk, hogy a rendelkezésre álló eszközök mennyiben felelnek meg ezeknek a kívánalmaknak, ismertetjük az általunk alkalmazott, néhány esetben a praktikus kényszer által is vezérelt megoldást, illetve esetenként javaslatot teszünk olyan fejlesztési lépésekre, amelyeket feltétlen szükségesnek tartunk ahhoz, hogy egy adott elemzési feladatot magas minőségben megoldani képes, konfigurálható és tárgykörre adaptálható eszköz jöjjön létre. Mivel a jelen dolgozat leginkább egy helyzetjelentés, és az ezzel a helyzettel szembesülve, jelentős részben pragmatikus szempontok által indokolt megoldások kiválasztási módszereinek a leírása, vagyis semmiképp sem tekinthető a tárgyalt eszközök sztenderd környezetben történő minőségi kiértékelési jelentésének.

3.1. Mondatszegmentálás, tokenizálás

A tokenizálás és mondatszegmentálás esetében a minőségi kényszer a magas *pontosság* mellett magas *fedést* is megkíván, vagyis meglehetősen változatos (szépirodalomtól a webes blogokig) szövegtípusokban kell jó eredményt elérni, nemcsak a rendkívül változatos konfigurációban előforduló mondatok határainak megállapításában, hanem speciális szövegelemek (nyílt tokenosztályok, például URL-ek, email címek stb.) felismerésében is. Minőségi magasabb szintű nyelvi annotációhoz elengedhetetlenül szükséges a pontos tokenizálás, amely messze túlmutat az egyszerű reguláris kifejezésekre támaszkodó, alkalmi szkriptek által nyújtott megoldási lehetőségeken [8].

Mint sok más nyelvfeldolgozó alkalmazásban, itt is két megközelítés szokásos, statisztikai modell [9,10] alapú illetve szimbolikus, szabályalapú [11]. Az előbbi típusban elérhető megoldások a szövegmennyiség és változatosság miatt széles körű tanítás, tesztelés és finomhangolás nélkül nem teljesíthetnek kielégítően, az ehhez szükséges idő és erőforrás viszont a projekt során nem állt rendelkezésre, így az elterjedtebb és számos készen használható eszközt kínáló utóbbi megközelítés volt kézenfekvő.

1. táblázat. Tokenizálók tulajdonságai.

	MtSeg[2]	Europarl[3]	huntoken[4]	Nagel[5]	FreeLing[6]	NYTI lánc[7]
1.	+	+	+	+	+	+
2.	C	perl	flex/C	C	C++	perl
3.	latin1/2	UTF-8	latin1/2	latin/UTF-8	UTF-8	latin2
4.	minimális	minimális	jó	minimális	kevés	nincs
5.	kérdéses	jó	jó	jó	közepes	jó
6.	+ + + + +	+ + +	+	++	+ + ++	+ + +
7.	nyelvi modulok konfigurálhatóság	egyszerű	nyelvi tudás	fejleszthető	gyors	egyszerű
8.	elavult kód, lassú	nincs nyelvi tudás	tokenizálási hibák	nincs nyelvi tudás	nincs nyelvi tudás	nincs nyelvi tudás

1.: forrás elérhető 2.: implementáció 3.: kezelt karakterkódolás 4.: dokumentáció 5.: fejleszthetőség 6.: becsült fejlesztési igény 7.: előnyök 8.: hátrányok

[2] Már nem elérhető, saját példány.
[3] http://search.cpan.org/~achimru/Lingua-Sentence-1.03/lib/Lingua/Sentence.pm
[4] http://mokk.bme.hu/resources/huntoken
[5] http://www.cis.uni-muenchen.de/~wastl/misc/tokenizer.tgz
[6] http://www.lsi.upc.edu/~nlp/freeling
[7] Marcia Munoz mondatrabontója (http://aye.comp.nus.edu.sg/~forecite/services/uiuc-srl/srl-demo2/bin/sentence-boundary2.pl) és a Grefenstette-féle tokenizáló szkript (http://nora.hd.uib.no/corpora/1999-3/0348.html) házi használatra módosított változatban

A MNSz tekintetében a legfontosabb szempontok az adatmennyiség miatt a gyorsaság, a komplex nyelvi elemek (nyílt tokenosztályok) felismerhetősége miatt a nyelvi tudás, a szövegek változatossága miatt pedig a doménilleszthetőség lehetősége voltak. Az 1. táblázat foglalja össze egy informális vizsgálat alapján néhány tipikus eszköz jellemző tulajdonságait főként az említett szempontok alapján. Az „egyszerűség" annyiban előny, hogy gyors és robusztus működést eredményez, abban pedig természetesen hátrány, hogy a (teljesen) hiányzó nyelvi tudás beépítése rendkívül erőforrásigényes. A vizsgálat egyértelmű tanulsága, hogy azon túl, hogy több eszköz csak szerény mértékben képes a triviális szegmentálási/tokenizálási megoldásnál[8] többet nyújtani, nincs minden szempontnak megfelelő, készen használható eszköz magyar nyelvre, amely az UTF kódolást is képes lenne kezelni. Ennek következtében olyan kompromisszumot kellett kialakítani, amely a leggazdaságosabb eredményt adja a befektetett fejlesztés – kimeneti minőség tekintetében.

A választás a komplex tokenek beépített kezelési képessége és a nyelvi illesztés miatt a *huntoken* eszközre esett, és ez a projekt keretében visszafordíthatatlan elkötelezettséget jelentett, az elemzés alatt felmerülő problémákkal szembesülve a kiindulópontra visszatérni és egy újabb eszközzel ismét elölről kezdeni a szükséges fejlesztést és kiegészítést nem volt lehetséges. Első lépésben a karakterkódolási problémát kellett megoldani, a 2. részben tárgyalt módon. Ezen kívül a szövegek sokfélesége felszínre hozott számos implementációs hibát, szabályhiányosságot, ezeket javítani kellett. A módosítás több száz sornyi új kódot, többek között a rövidítések kezelési módjának teljes átdolgozását, és a kimeneti formátum egyszerűsítését eredményezte, és végül messze túlment az eredetileg becsült minimális fejlesztési igényen. Eredményül viszont a működési körülményekhez képest a legjobb minőségű elemzést adó eszköz jött létre. Ez azonban mégsem tekinthető egy magyar nyelvi tokenizáló/szegmentáló eszközre adott optimális megoldásnak. Nem teljesülnek ugyanis azok a feltételek, melyek ehhez szükségesek lennének.

Az UTF kódolás natív kezelése természetes követelmény, de ennél alapvetőbb, architekturális hiányosságok is felmerülnek, nemcsak ennek, hanem sok más eszköznek az esetében is. Az egy lépésben történő elemzés veszélye, hogy a fedés növelésével együttjáró egyre összetettebb szabályrendszert rendkívül nehezen lehet karbantartani, a komplex kifejezések által meghatározott halmaz elemei a humán fejlesztő számára már nem láthatók át teljes körűen, így a halmaz tartalmazhat olyan elemeket, melyek más kifejezések nyelvébe is beletartoznak. Ha ilyenkor a szabályok közötti hierarchia nincs egyértelműen és konfigurálhatóan meghatározva, akkor az alkalmazott implementáció belső szabályrendezése ér-

[8] Triviális megoldásnak tekintjük a szóköz(értékű) karakterek és központozás mentén történő tokenizálást és az általában mondatzáró központozás utáni nagybetűs elem által meghatározott mondathatár bejelölését, esetenként kiegészítve segédlexikon alapú rövidítésdetektálással.

vényesül, ami hibához vezethet.[9] Erre a problémára jó megoldás a tokenizálási feladat többlépcsős megközelítése. A mondatra bontás és a tokenizálás elkülönítése gyakorlatilag magától értetődő, de az utóbbit is célszerű felbontani: az első lépésben azonosított elemi egységek további lépésekben alkothatnak összetett egységeket, és minden lépésért különálló, egyedileg konfigurálható modul felel.[10] A sebességcsökkenés megtérül a pontosabb működésen. Ennek a felépítésnek a legjobb példája a Multext Segmenter [13,14], ez azonban az elavult implementáció miatt nem használható, az alkalmazott architektúra és az elérhető funkcionalitás viszont jó kiindulópont.

A doménilleszthetőség olyan funkció, ami elengedhetetlenül szükséges egy robusztus és változatos bemenetet kezelni képes eszköz esetén. Egyszerű illusztrációja ennek például a szóközhiány mondatvégi központozásjelek után és nagybetű előtt, hiszen előfordulnak olyan szövegtípusok, ahol ez az esetek nagy részében hiba (mindennapi prózai szövegek), de olyanok is, ahol viszont általában nem hiba (számítógépes szakszövegek). Ezért az adott szövegtípushoz kell illeszteni bizonyos szabályok alkalmazhatóságát, és ezt egyszerűen konfigurálhatóvá kell tenni.

Összegzésképpen legcélszerűbbnek látjuk az alapoktól felépíteni a fenti követelményeknek megfelelő rendszert, szintetizálva mindazt a felhalmozott tudást és előnyös tulajdonságot, ami a jelenleg rendelkezésre álló rendszerekben megtalálható.

3.2. Morfológiai elemzés

A magyar nyelvi korpuszokban szokásos gyakorlat, hogy a morfológiai annotáció az utolsó képzőt magában foglaló és ez által meghatározott szófajú tövet és az ehhez járuló inflexiós toldalékolást tartalmazza. Felmerült az igény azonban a morfofonológiai kutatások kiszolgálása érdekében, hogy a morfológiai elemző kimenetéből további hasznos és kutatási kérdésekhez könnyen lekérdezhető információt is szolgáltassunk, és most először a morféma (és fonéma) szintű elemzés és annotáció is hozzáférhető legyen a korpuszban[11]. Ez mind az alkalmazott eszközzel, mind a kapott elemzés feldolgozásával kapcsolatban teljesen új problémákat vetett fel, különösen a morfémaszintű strukturális többértelműségek és szóösszetételi anomáliák feloldásában. Ez a követelmény az eszközválasztást is jelentősen befolyásolta, tekintve, hogy azon magyar nyelvi elemzők közül, amelyek a morfémákra történő felbontást is visszaadják kimenetként az egyik (*Xerox*) teljesen zárt és jelen körülmények között megváltoztathatatlan rendszer, így sem hibajavításra sem bővítésre nem ad lehetőséget, a másik (*ocamorph*) túlgenerálása olyan mértékű, amit rendkívül körülményes kezelni, és az ilyen részletes

[9] Esetünkben például a `flex` belső szabályhierarchiáját felülírni csak rendkívül körülményesen lehet [12], ami az alkalmazott szabálymennyiség mellett karbantarthatatlan.

[10] Az általunk használt eszköz alapvetően csak a rövidítéseket próbálja így kezelni.

[11] Ezen túlmenően egyes képzők jelenléte a nyelvi elemzés magasabb szintjein is releváns információ lehet.

elemzést kiadó üzemmódban sebességben is elmarad a végül általunk felhasznált harmadik (*HUMOR*) eszköztől.

A morfémákra bontás strukturális többértelműségeinek feloldására az [5]-ben említett egyszerű heurisztika (a legrészletesebb felbontás a választott elemzés) alkalmazásának érdekében minden a fent említett módon számított tő+inflexió alakhoz hozzárendeljük a lehetséges derivációs elemzések halmazát (2. ábra). Az

```
[kenyérkalaposság/FN.PSt2.SUB]:{
kenyérkalap[FN]+os[_SKEP]+ság[_PROP]+otok[PSt2]+ra[SUB];
kenyérkalap[FN]+os[_SFN]+ság[_COL]+otok[PSt2]+ra[SUB];
kenyér[MN]+kalap[FN]+os[_SKEP]+ság[_PROP]+otok[PSt2]+ra[SUB];
kenyér[MN]+kalap[FN]+os[_SFN]+ság[_COL]+otok[PSt2]+ra[SUB]
}
```

2. ábra. Elemzési alternatívák.

egyes halmazokon belül tehát vesszük a legrészletesebb (legtöbb morfémát tartalmazó) elemzési uta(ka)t, és minden egyes morfémára eltároljuk a lehetséges elemzéseit, az esetleges többértelműségeket megőrizve. Ugyancsak tároljuk az összetételek minden elemét (3. ábra). Azt a meglehetős kihívást jelentő kérdést, hogy az elemzés ezen szintjén keletkező többértelműségek automatikus feloldása miként lenne lehetséges, a projekt keretében nem vizsgáltuk. A végső annotáció tartalmaz még egy további egyszerű reprezentációt a morfémahatárok megjelölésére (`kenyér+kalap+os+ság+otok+ra, bokr+os+od+ás`).

```
[kenyérkalaposság/FN.PSt2.SUB] -> {
 stem  => {[kenyér],[kalap]},
 'os'  => {SKEP,SFN},
 'ság' => {PROP,COL},
 'otok'=> {PSt2},
 'ra'  => {SUB}
}
```

3. ábra. Morfémaszintű reprezentáció.

A szóösszetételek nagyfokú produktivitását kezelni képes elemző elkerülhetetlenül túlgenerál, ennek automatikus szűrése olyan eddig nem kielégítően kezelt probléma [15], aminek a megoldását a projekt nem tudta felvállalni. Erre az esetre ismét egy adatvezérelt megközelítés volt az alkalmazott eljárás alapja, ahol egy gyakorisági lista morfológiai elemzése után az összetételnek elemzett alakok közül a leggyakoribbak, illetve bizonyos tipikus utótagra[12] végződők teljes körű

[12] Pl. *ad, árok, dám, dia, est, jak, kád, kan, kos, lak, tat, tó, velő.*

manuális vizsgálat alá estek. Ennek eredménye egy mintaillesztő szűrő erőforrás lett, jelenleg 112 mintát és több ezer szűrt szóalakot tartalmazva (4. ábra)

```
(ár|borz|fog|hal|láz|mar|rag|szak|tag)\[FN\]\+ad\[IGE\]
(ár|borz|fog|hal|láz|mar|rag|szak|tag)\[FN\]\+ad(ó|ás)\[FN\]
...
(Balla|bor|Bor|cella|Cella| ... szó|téboly|vaj|Ver|zár)\[FN\]\+dám\[FN\]
(abszorber|adapter|áttétel ... törvény|zsilip|zsindely)\[FN\]\+est\[FN\]
```

4. ábra. Hamis összetételeket (pl. *láz+adó, áttétel+est*) szűrő minták.

Mindezen újdonságnak számító hozzáadott információ ellenére a morfológiai elemzés kérdése az MNSz-ben sincs teljes körűen megoldva. Hiányzik a tokenizáló és a morfológiai elemző gördülékeny együttműködése például a tokenizáló által felismert komplex alakulatok toldalékolásának felismerésében, és magának az elemzőnek is maradtak hibái. Az eddig a morfológiai elemzés területén befektetett hazai erőfeszítések igazán hatékony kihasználását egyértelműen akadályozza egy közös, harmonizált (vagy legalább harmonizálható) kimeneti kódkészlet és reprezentáció hiánya a különböző morfológiai elemzőkben. Szükség lenne egységes és nyíltan hozzáférhető segéderrőforrások, lexikonok közösségi fejlesztésére is.[13] Összességében, ahogy a tokenizálásnál, itt is úgy látjuk, hogy az eddigi eszközök tudását szintetizáló szabadon elérhető, megfelelően gyors és testre szabható elemző létrehozása lenne a kívánatos megoldás.

3.3. Szófaji egyértelműsítés

A szófaji, morfoszintaktikai egyértelműsítés, mint a nyelvi elemzés egyik sarokpontja, már magyar nyelvre is viszonylag alaposan feltárt területnek számít, a nagyméretű korpusz és a szövegek változatossága azonban jelenthet még kihívást [1]. Egy ennél sokkal alapvetőbb problémát is érdemes azonban felvetni. Valóban teljes körűek az ismereteink az egyes eszközök teljesítményéről? Összehasonlíthatók-e egyáltalán a különböző rendszerek eredményei? Az eddig megjelent rendszerekről (lásd pl. többek között [16,17,18,19]) közölt teljesítményadatok alapján lényegében lehetetlen kijelenteni, hogy az egyik módszer jobb a másiknál, annyira eltérőek a kiértékelési környezetek, a korpusz kódkészlete, a felhasznált külső erőforrás (például a morfológiai elemző). Nem kizárható, hogy némi változás a kézi tanító korpuszban vagy a tagkészletben már nagyobb hatással van a végeredményre, mint az alkalmazott rendszer lecserélése egyikről a másikra [20]. Rendkívül fontos tehát egy sztenderdizált kiértékelési környezet és protokoll meghatározása, az egyértelműsítés hibáinak megalapozott vizsgálata és értékelése. Ez mindezidáig hiányzik a magyar szakirodalomból. Mivel az nehezen vitatható, hogy egy morfológiai elemző kimenetének integrálása az egyértelműsítési folyamatba jelentős javulást eredményez, először ezt az információforrást

[13] Egy ilyenre példa lehet az itt említett összetételi szűrő.

kellene egységesíteni, ennek hiányában nincs lehetőség objektívan értékelni. További problémát okoz a korpuszkód-készletek, illetve a bennük tárolt információ különbözősége. Hiába mérjük az egyes egyértelműsítő eszközök teljesítményét egységes kódkészlettel, részletesen vizsgálni kell azt is, hogy a kódkészlet nem elfogult-e valamelyik eszközzel szemben, vagyis pont azokat a jellemzőket tartalmazza, amik az egyik eszköznek hasznosak, míg olyan jellemzőket, amiket esetleg másik eszköz tudna felhasználni, nem tartalmaz, neutralizál. Ez a típusú kiértékelés igen munkaigényes, így az MNSz készítésének keretében erre nem volt lehetőség. Ezért a továbbiakban olyan általános alkalmi vizsgálatok eredményét mutatjuk be, amelyek ettől függetlenül is tudnak tanulsággal szolgálni.

Három egyértelműsítő eszköz [16,18,19] kimenetét vizsgáltunk, ebből kettő [16,19] azonos kódkészlettel tanítható volt, a harmadik [18] előre megépített modellel és kódkészlettel rendelkezett. A tévesztési mátrixok alapján jellemző idioszinkratikus és paradigmatikus hibákat lehetett azonosítani, az eszközök közötti minőségi különbség viszont a fenti problémák miatt megállapíthatatlan volt. A 2. táblázat olyan mindegyik eszközre jellemző[14] tévesztéseket mutat be, melyek egy-egy tipikus problémát és egyben megoldási lehetőséget is illusztrálnak.

2. táblázat. Jellemző tévesztések (aszimmetrikus mátrixból).

Helyes kód	Összes előf.	Ebből hibás	(%)	Hibás kód	Előfordulása	(%)
1. AS_V	446	80	(17.94%)	VS3PI	46	(57.50%)
2. AS_A	2655	74	(2.79%)	NS3NN	42	(56.76%)
3. NS3PN	350	48	(13.71%)	D__D	25	(52.08%)
4. R__P	306	24	(7.84%)	C	14	(58.33%)
5. VS3SD	24	11	(45.83%)	VS3RD	11	(100.00%)

AS_[AV]: mn/ige tövű mn.; VS3PI: ige, múlt i.;
NS3NN: fn. egyes szám nom.; NS3PN: 3.szem. egyes sz. fn.-i névmás;
D__D: névelő; R__P: hsz.; C: kötőszó;
VS3SD: ige, felszólító m.; VS3RD: ige, kijelentő m.

1. A melléknév(i igenév) és a múlt idejű ige megkülönböztetése olyan összetett információt igényel, ami nem érhető el az eszközök által épített modelleken maradéktalanul. Ilyen esetben célszerű külön modellt használni a feladatra.
2. A főnév és melléknév homonímák megkülönböztetése bizonyos esetekben a humán annotátorok számára sem egyértelmű, és elméleti nyelvészeti szempontból sem teljesen tisztázott terület. Meghatározott eseteket lehet automatikusan javítani[15], de ennél több nem nagyon várható.
3. Az ebben a típusban található hibák (*az* mint névmás illetve névelő) legegyszerűbb megoldása célzott modellel, akár külső szabállyal a legegyszerűbb.

[14] A mért értékek minimális eltéréssel megegyeznek mindhárom eszközre.
[15] Pl. névelő előtt legyen mindenképpen főnév a megoldás, de itt is gondot okozhatnak az elliptikus szerkezetek.

4. A kötőszó és határozószó (*így, amikor* stb.) elemzés megkülönböztetésének nehézsége hasonlít az 1. esethez, ezen túl bizonyos esetekben az is célravezető lehet, vagyis kevesebb hibát eredményez, ha a ritkább elemzést egyszerűen figyelmen kívül hagyjuk a modellben, tehát meg sem próbálunk egyszerű automatikus megoldást alkalmazni arra a problémára, ahol több hibát okoz a megoldás alkalmazása, mint amennyit meghagy a nem alkalmazása.
5. Ez a típus az 1. eset szófajon belüli megjelenése, hasonló konklúzióval.

Az MNSz-nek ebben a feldolgozási lépésében a fenti eredmények figyelembe vételével igazi opportunista döntés született és lényegében az eddig is használt saját célra alakított feldolgozó láncot használtuk, amely egy szabályokat használó előszűrőből és a morfológiai elemző kimenetével megszorított nagyon gyors HMM alapú egyértelműsítőből áll [21]. Nem volt egyértelmű bizonyíték arra, hogy létezik jelentősen jobb és hasonlóan gyors megoldás, ezért nem volt indokolt egy jól működő eszközlánc lecserélése.

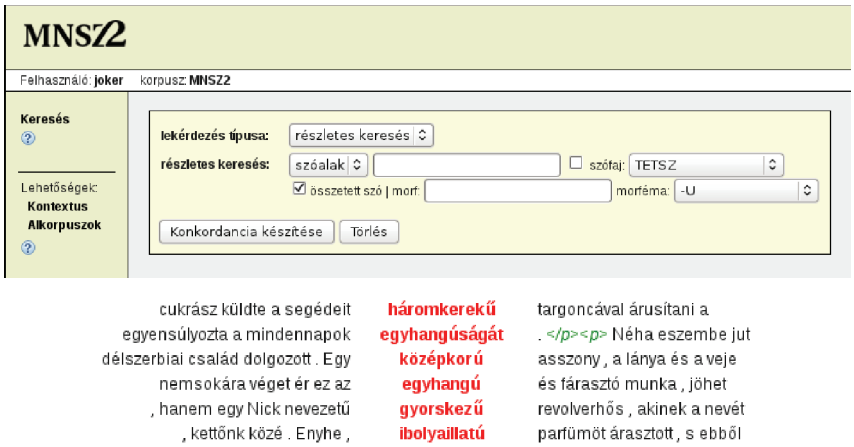
4. Korpuszkezelés és megjelenítés

Az MNSz új változata a megszokott `http://mnsz.nytud.hu` címen érhető el. A korpusz mögött egy korszerű, megbízható korpuszkezelő motor működik [22]. Sebessége a milliárd szavas méret mellett is megfelelő, a lekérdezőfelület válaszideje rövid.

A motorhoz tartozó felület eleve számos hasznos beépített funkciót tartalmaz, melyek újdonságot jelentenek az MNSz régi változatához képest. Nagy mennyiségű találat esetén is lekérhetjük az összes találatot, és kényelmes formátumban elmenthetjük további feldolgozásra. Egy gombnyomással testre szabhatjuk a megjelenítést, rendezhetjük a konkordanciát. A kapott találatokat újabb lekérdezéssel szűrhetjük, több lépésben is. Különféle gyakorisági listákat készíthetünk, és kollokációs vizsgálatot is végezhetünk.

Annak érdekében, hogy az egyes nyelvi szinteken megjelenő igen részletes elemzési információt felhasználóbarát módon hozzáférhetővé tegyük, az eredeti felületet számos ponton bővítettük. Egyrészt kiegészítettük a magyar inflexiós morfológia jelenségeit lefedő menürendszerrel, mely funkcionalitásában megfelel a régi MNSz-felület hasonló részének. Újdonság, hogy fonológiai jegyek, fonémaosztályok alapján is kereshetünk. Például a részletes keresőben beállított `{pal,aff}u.*{son}` kereséssel a palatális affrikátával kezdődő, *u*-val folytatódó és szonoránsra végződő szavakat keressük, és ennek megfelelően a találatokból képzett gyakorisági lista a *csupán*, *dzsungel*, *csupor* szavakkal kezdődik. Szintén újdonság, hogy a korpuszban meglévő morfémaszintű elemzésnek köszönhetően vizsgálhatók az összetett szavak, illetve hozzáférünk a derivatív morfológiához: az egyes morfémákhoz és konkrét morfmegvalósulásokhoz is (5. ábra).

A fentiek mellett arra is lehetőség van, hogy a rendszer belső korpuszlekérdező nyelvét – a CQL-t – közvetlenül használjuk, és általa rugalmasan hozzáférjünk a korpuszban rejlő információ egészéhez.



5. ábra. Az *-ú* képzőt tartalmazó összetett szavak lekérdezése és a válaszkonkordancia egy részlete.

5. Összefoglalás

A MNSz munkálatai során egyértelműen igazolódott, hogy a nyelvi elemzés egyes szintjein nincs, és persze nem is nagyon lehetséges olyan széles alkalmazhatóságú, elegendően gyors kész eszköz, amely magas minőségű megoldást képest adni a korpusz teljes szövegspektrumán. Ezért rendkívül fontos az eszközök konfigurálhatósága, a hatékony doménillesztés lehetősége. Ez viszont gyakorlatilag hiányzik minden jelenleg hozzáférhető szimbolikus alapú eszközből, a sztochasztikus megoldásokban pedig sztenderdizált nyelvi erőforrások híján jelentős befektetést kíván. A statisztikai modelleket alkalmazó eljárások hiába taníthatók az adott doménen, ha az elvárt pontosságú annotációhoz szükséges tanító adat nem áll rendelkezésre, és előállítása jelentős befektetést igényel, így nem lehet kijelenteni, hogy egyértelmű előnyben lennének a szimbolikus megoldásokkal szemben (lásd például a mondatszegmentálás és tokenizálás problémáját).

A minőségi igény következtében nagyon fontos szempont, hogy a nagy mennyiségű szöveg feldolgozása elkerülhetetlenül számos hibát derít fel az alkalmazott eszközben, és ezek folyamatos javításához elengedhetetlen az eszköz erőforrásaihoz történő átlátható hozzáférés, az alkalmazott modell(ek) gyors és rugalmas újraépítésének lehetősége.

Nagy szükség lenne végre egy közös fejlesztés eredményeként előálló magyar BLARKra [23], az eszközök, a hozzájuk szükséges erőforrások és az objektív, sztenderdizált tesztkörnyezet tekintetében is, ahol az egyes feldolgozási lépések akár többféle modullal is elvégezhetők, ezek azonban jól definiált API-n keresztül kommunikálhatnak egymással.

Hivatkozások

1. Wei-yun, M., Huang, C.R.: Uniform and effective tagging of a heterogeneous gigaword corpus. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy (2006) 2182–2185
2. Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K.: English Gigaword Fifth Edition. Linguistic Data Consortium (2011)
3. Halácsy, P., Kornai, A., Németh, P., Varga, D.: Parallel creation of gigaword corpora for medium density languages – an interim report. In: Proceedings of the International Conference on Language Resource and Evaluation (LREC08). (2008)
4. Ferraresi, A., Zanchetta, E., Baroni, M., Bernardini, S.: Introducing and evaluating ukWaC, a very large web-derived corpus of English. In: Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google. (2008) 47–54
5. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., szerk.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014)
6. Váradi, T., Oravecz, Cs.: A Magyar Nemzeti Szövegtár egymilliárd szavas új változata. Magyar Tudomány **175** (2014) 1054–1061
7. The Unicode Consortium: The Unicode Standard, version 7.0.0 (2014)
8. Forst, M., Kaplan, R.M.: The importance of precise tokenizing for deep grammars. In: Proceedings of Fifth Language Resource and Evaluation Conference (LREC2006), Genoa, Italy (2006)
9. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Computational Linguistics **32** (2006) 485–525
10. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: Proceedings of ANLP-97, Washington, USA (1997)
11. Grefenstette, G., Tapanainen, P.: What is a word, what is a sentence? problems of tokenization. In: Papers in Computational Lexicography. COMPLEX'94, Budapest, Research Institute for Linguistics (1994) 79–87
12. Flex Frequently Asked Questions: (Flex is not matching my patterns in the same order that i defined them) http://flex.sourceforge.net/manual/FAQ.html.
13. Ide, N., Véronis, J.: MULTEXT: Multilingual Text Tools and Corpora. In: Proceedings of the 15th Conference on Computational Linguistics. (1994) 588–592
14. Erjavec, T., Ide, N., Petkevic, V., Véronis, J., Schuman, A.R.: Multext-east: Multilingual text tools and corpora for central and eastern european languages. In: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar. (1995) 87–97
15. Novák, A., Pintér, T.: Milyen a még jobb Humor? In: Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006) 60–69
16. Oravecz, Cs., Dienes, P.: Efficient stochastic part of speech tagging for Hungarian. In: Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas (2002) 710–717
17. Halácsy, P., Kornai, A., Csaba Oravecz: Hunpos – an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, ACL (2007)
18. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771

19. Orosz, Gy., Novák, A.: Purepos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria (2013) 539–545
20. Manning, C.D.: Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I. CICLing'11, Berlin, Heidelberg, Springer-Verlag (2011) 171–189
21. Oravecz, Cs., Dienes, P.: Large scale morphosyntactic annotation of the Hungarian National Corpus. In Hollósi, B., Kiss-Gulyás, J., szerk.: Studies in Linguistics. Volume VI., Debrecen, Institute of English and American Studies, University of Debrecen (2002) 277–298
22. Rychlý, P.: Manatee/Bonito – a modular corpus manager. In: Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing, Brno: Masaryk University (2007) 65–70
23. Krauwer, S.: The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of SPECOM, Moscow (2003)

Magyar nyelvű webes szövegek morfológiai és szintaktikai annotációja

Vincze Veronika[1,2], Varga Viktor[1], Papp Petra Anna[1],
Simkó Katalin Ilona[1], Zsibrita János[1], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{vinczev,zsibrita,rfarkas}@inf.u-szeged.hu,
{viktor.varga.1991,papp.petra.anna,kata.simko}@gmail.com
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.

Kivonat Cikkünkben bemutatjuk az első magyar, kézzel annotált, webes szövegeket tartalmazó korpuszt, melyet tesztadatbázisnak szánunk a webes szövegekre optimalizált nyelvi elemzőink fejlesztéséhez. A korpusz morfológiai és (összetevős és függőségi szemléletű) szintaktikai elemzést, valamint szemantikai és diskurzusbeli bizonytalan kifejezések annotációját tartalmazza. Beszámolunk a magyarlanc elemző webes szövegekre történő adaptálási kísérleteiről is.

1. Bevezetés

Az interneten fellelhető szövegek mint könnyen hozzáférhető és már-már kifogyhatatlan adatforrás jelentősége már régóta a köztudatban van, sok kutatás irányította rá a tárgyát. Az utóbbi években a webes szövegek legnagyobb részét már maguk a felhasználók adják közre, legyen az a produktum blogbejegyzés, fórumokon való beszélgetés, bejegyzésekhez tartozó véleménynyilvánítás (komment) vagy mikroblogbejegyzés (pl. Twitter). Egyértelmű tehát, hogy a „webes szöveg" korántsem homogén szövegtípus, felhasználónként, csoportonként és műfajonként változik a stílus, a megszerkesztettségre való igény és ebből adódóan – a számítógépes nyelvészet szempontjából nézve – a feldolgozás nehézsége is. A már létező nyelvi elemzők viszonylag jól működnek a sztenderdhez közelebb álló szövegek (pl. blogbejegyzések) esetén, a kommentek és mikroblogbejegyzések nyelvi feldolgozása viszont zajossága és nem sztenderd formájuk miatt nehezen megvalósítható a meglevő eszközök adaptálása nélkül.

Más nyelvekre már készültek internetes szövegeket tartalmazó annotált korpuszok, l. például az angolra [1] és a franciára [2]. Célunk az volt, hogy webes szövegekből (nyilvános kérdés–válasz párokból és bejegyzésekre érkezett kommentekből) olyan kézzel ellenőrzött, helyes morfológiai és szintaktikai annotációval ellátott magyar nyelvű korpuszt hozzunk létre, amelyen a jövőben természetesnyelv-feldolgozásra kifejlesztett elemzőket lehetséges tesztelni és segítségével optimalizálni. Cikkünkben e korpusz létrehozásának folyamatát ismertetjük, az

annotáció szintjeinek bemutatásával, továbbá ismertetjük a magyarlanc elemzővel [3] elért morfológiai és szintaktikai elemzés végeredményét is. A korpuszt oktatási és kutatási célokra szabadon elérhetővé kívánjuk tenni.

2. A korpusz összetétele

A szövegek összeválogatása során szempont volt, hogy a kommunikáció szóbeli jellegzetességeit mutassák, mind stílusban, mind formában (kérdés–válasz párok, beszélgetések). A korpusz 2014 augusztusában gyűjtött nyilvános Facebook-kommentekből, valamint a gyakorikerdesek.hu oldalon feltett kérdésekből és válaszokból áll. A gyűjtés során a teljesség és visszakereshetőség követelményének megfelelően egész thread-eket mentettünk el, elérési útvonallal és időbélyeggel együtt. A bejegyzések nagyrészt hobbival, személyes érdeklődési körökkel és életmóddal kapcsolatosak.Megemlítendő, hogy a Szeged Korpuszban is találhatóak többé-kevésbé hasonló hangvételű és stílusú írások (szépirodalmi művek és általános iskolai fogalmazások formájában), bár esetükben a zajosság nem (illetve jóval kevésbé) jelenik meg.

A korpusz főbb adatait az 1. táblázat foglalja össze. A korpuszt – méreteinél is fogva – elsődlegesen benchmark adatbázisként kívánjuk hasznosítani, mely a különféle nyelvi elemzők webes doménre történő adaptálását teszi lehetővé.

1. táblázat. A korpusz mérete.

Típus	**Facebook**	**Gyakorikérdések**	**Összesen**
Bejegyzések	879	258	1 137
Mondatok	1 208	728	1 936
Tokenek	8 739	9 880	18 620
szó	7 171	8 236	15 106
írásjel	904	1 551	2 455
emotikon	674	91	756

3. Morfológia

A webes szövegek elemzésének nehézségei már az előfeldolgozás, azaz a mondatszegmentálás és a tokenizálás során jelentkeztek. A sztenderd szövegen tanult magyarlanc hangulatjelekkel, extrém írásjelhasználattal és helytelen egybe- és különírással nem találkozott, így – mint az várható is volt az előzetes kutatás [4] után – a folyamat nem volt megoldható automatikusan. Sok esetben a tagmondat–mondat viszony és különbség nem volt tökéletesen megállapítható. Ezek a problémák a központozás elhagyása vagy szokatlan használata miatt adódtak (pl. *Péter én meg a gépem xD majdnem szét vertem, hogy amikkor oda*

ülnék tankolni persze akkor fagyok ki:D). Megjegyzendő, hogy a szóbeli kommunikációban sokszor előforduló néma szünet gyakran megjelenik mintegy gondolategységeket elválasztó írásjelként (többnyire „..." formában), azonban használata korántsem következetes.

A szegmentálás után a korpuszban előforduló szóalakokat összegyűjtöttük, majd a magyarlanc felhasználásával és kézi kiegészítéssel megadtuk az összes lehetséges elemzésüket. Következő lépésben előállítottunk egy, a Szeged Korpusz formátumához hasonló szerkezetű szövegkorpuszt, amelyben az annotátorok egy erre a célra kifejlesztett szoftverrel kézileg egyértelműsítették a szóalakokat.

A webes szövegekben előforduló morfológiai jellegű hibákat, pontosabban a sztenderd nyelvhasználattól való eltéréseket egy korábbi kutatásban [4] már felvázoltuk, a típushibák és a lehetséges automatikus megoldások gyűjtése a jelenlegi folyamatnak is részét képezte. A legtöbb tévesztés ékezetekkel, egybeírással és egyéb helyesírási hibákkal kapcsolatos. A morfológiai annotáció során megtartottuk az eredeti – hibásan írt – szóalakot, és ehhez a kontextusnak megfelelő helyes elemzést (lemmát és morfológiai kódot) rendeltük hozzá, a Szeged Korpusz 2.5-ben [5] használt eljáráshoz hasonlóan. Speciális esetekben, például ékezetek vagy betűk elhagyása miatt felmerülő poliszémia esetén (pl. *joban – jóban* vagy *jobban*; *tok – tök* vagy *tudok*) minden lehetséges (ill. gyakori, a szövegben előforduló) helyes kódot és lemmát felvettünk, míg a tévesen egybeírt alakok a szövegben előforduló, esetleg ékezetesített lemmát kapták meg, konszenzus alapján eldöntött kóddal ellátva (pl. *jolesz* esetében igeivel).

A 2. táblázatban a tartalmas (azaz nem írásjel és hangulatjel) tokenek mondatbeli eloszlása látható. A legszembetűnőbb eltérés, hogy míg a Gyakorikérdések alkorpusznak majdnem fele több mint 10 tokenből áll, a Facebook esetén ez a nagyságrend a 3-6 tokenes kategóriában jelenik meg. A két domén átlagos mondatonkénti tokenszáma is hasonló arányokat mutat, az előző sorrendet követve 11,07 és 5,84 szóalak/mondat (összes tokenre nézve 11,19 és 6,39).

2. táblázat. Tokenszám mondatonként.

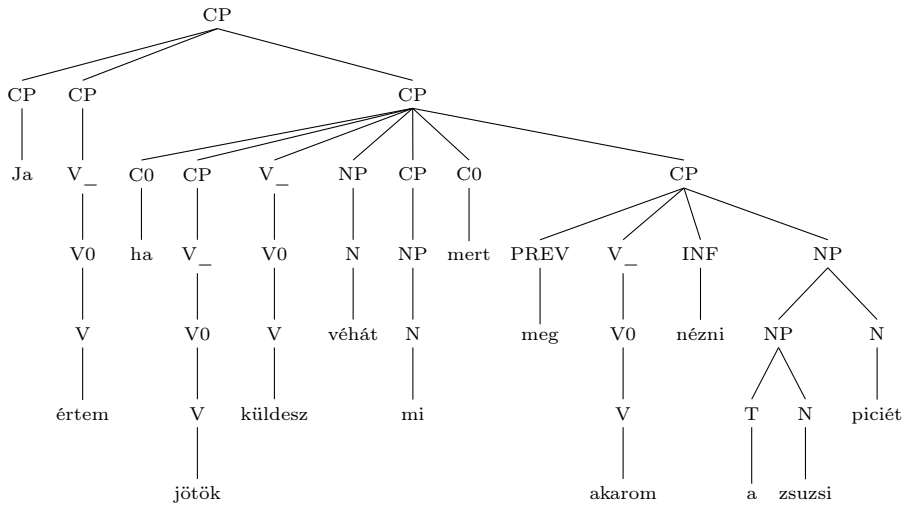
Tokenszám	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
1-2 token	33	4,53	94	7,78	127	6,56
3-6 token	233	32,01	685	56,71	918	47,42
7-10 token	128	17,58	199	16,47	327	16,89
10+ token	334	45,88	230	19,04	564	29,13
Összesen	728	100,00	1208	100,00	1936	100,00

4. Szintaxis

A korpusz szövegeit konstituens- és függőségi elemzéssel is elláttuk. A következőkben ezeket a munkafolyamatokat részletezzük.

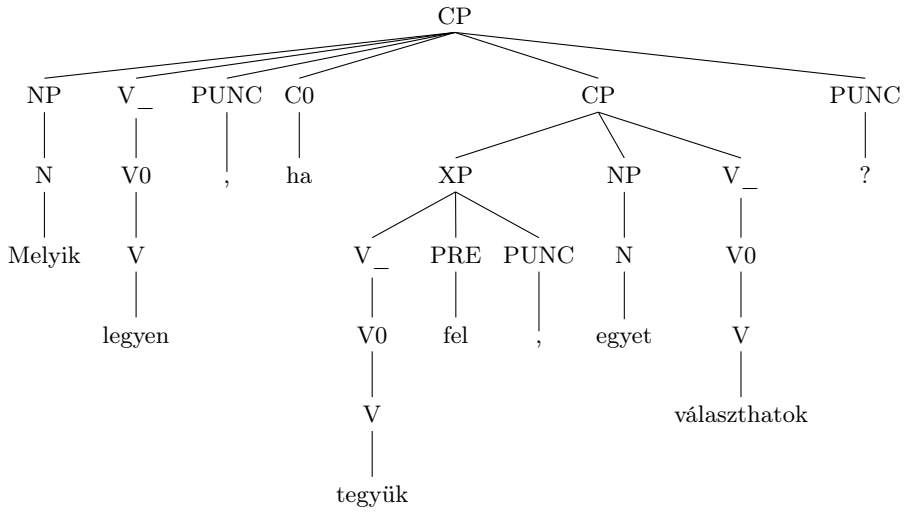
4.1. Összetevős elemzés

A mondatokat a morfológiai annotálás és hibaelemzés után a nyelvész szakértők először összetevős (azaz konstituens-) elemzéssel látták el, a morfológiához hasonlóan kézzel. Az annotáció során törekedtünk arra, hogy a Szeged Treebankhez hasonló módon történjen, kiegészítve a webes szövegek annotációja során felmerült megoldásokkal. Ilyen módosítás például, hogy a hibásan különírt szavak és szócsoportok (tipikusan toldalékolt szavak és szóösszetételek) egy összetevőbe kerültek. Az emotikonok a mondathoz szorosan nem tartozó összetevőként (azaz XP-ként) lettek felvéve.



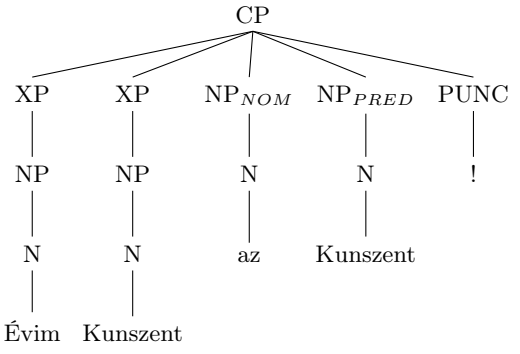
1. ábra: Központozás hiánya.

Az annotáció során felmerült további problémákat a szövegek beszélt nyelvhez közeli stílusából adódó nyelvhasználati jellemzők okozták. A kérdés–válasz struktúrának megfelelően jelentős számban fordultak elő hiányos mondatok. Gyakran okozott problémát a mondat- és tagmondathatárok megállapítása, ebben a jelentésbeli összefüggéstelenség és a központozás nem rendeltetésszerű használata is közrejátszottak (főként a Facebook alkorpuszban, 1. ábra).



2. ábra: Közbevetés.

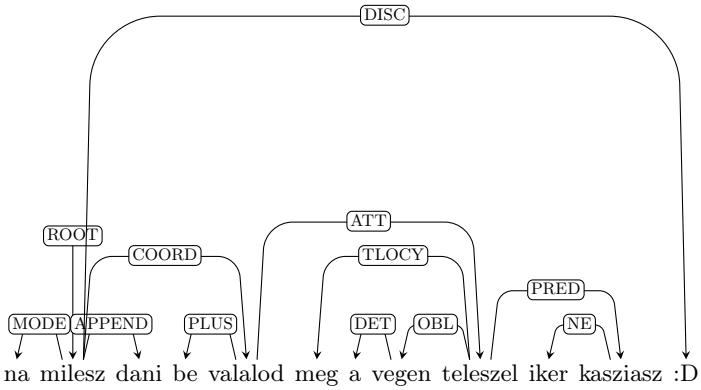
Ehhez a problémakörhöz kapcsolódik, hogy előfordulnak a mondatot megszakító, beágyazott mondatok, amelyek a mondatnak a beszélő által szükségesnek vélt kiegészítései, pl. hangsúlyozzák a szubjektivitást vagy modalitást közölnek. Ezeknek a státusza nyelvészeti szempontból sem teljesen tisztázott, közbevetésként (azaz a mondat szerkezetébe nem tartozó összetevőként) elemeztük őket (2. ábra). Egyelőre hasonlóan jártunk el a megszólítások és a nyelvészeti szakirodalomban kontrasztív topiknak nevezett jelenség esetében is (3. ábra).



3. ábra: Megszólítás és kontrasztív topik.

4.2. Függőségi annotáció

A korpusz mondataihoz kézzel hozzárendeltük azok függőségi ábrázolását is. A mondatokat a magyarlanc [3] függőségi elemző moduljával előelemeztük, majd az így kapott automatikus annotációt nyelvész szakértők kézzel kijavították. A munkálatok során alapvetően a Szeged Dependencia Treebank [6] létrehozása során alkalmazott elveket követtük, néhány változtatásra azonban szükség volt a webes szövegek sajátságainak megfelelően. Két új függőségi viszonyt vezettünk be az eddig is használatosak mellé: a DISC relációval jelöltük a szövegben előforduló diskurzusjelölőket és emotikonokat, a PLUS reláció pedig a hibásan különírt szavakat vagy betűkapcsolatokat köti össze (vö. 4. ábra).



4. ábra: Függőségi fa.

A korpuszban – a Szeged Dependencia Treebankhez hasonlóan – jelöltük az ún. virtuális csomópontokat is. Összesen 299 virtuális kopulát (kijelentő módú, jelen idejű harmadik személyű, fonológiailag üres létige) annotáltunk a korpuszban, ebből 119 a Facebook-, 180 pedig a Gyakorikérdések alkorpuszban található. Összesen 10 ellipszist találtunk a korpuszban, ami feltehetőleg annak köszönhető, hogy viszonylag kevés összetett mondat szerepel az anyagban, így a tagmondatok között ismétlődő elemek nem törlődnek.

5. Bizonytalansági annotáció

A korpusz szövegeiben megjelöltük a bizonytalanságot jelző nyelvi elemeket is. Az annotálás során a [7] és [8] által kidolgozott, majd a [9] által magyarra alkalmazott annotációs elveket követtük, azaz szemantikai és diskurzusszintű bizonytalanságot egyaránt annotáltunk. A szövegekben található bizonytalansági kategóriák megoszlását a 3. táblázat mutatja.

Az adatokból látszik, hogy a szemantikai bizonytalanság jóval gyakoribb a Gyakorikérdések alkorpuszban, ami valószínűleg annak köszönhető, hogy itt a

3. táblázat. Bizonytalansági annotáció a korpuszban.

	Típus	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
Szemantika	Episztemikus	7	3,45	39	10,32	46	7,92
	Doxasztikus	30	14,78	59	15,61	89	15,32
	Feltételes	18	8,87	64	16,93	82	14,11
Diskurzus	Weasel	23	11,33	31	8,20	54	9,29
	Hedge	58	28,57	117	30,95	175	30,12
	Peacock	67	33,00	68	17,99	135	23,24
	Összesen	203	100	378	100	581	100

felhasználók többsége nem ismeri egymást, ezért számos feltételezéssel élnek beszélgetéseik során, amit ki is fejeznek nyelvi eszközökkel. Ezzel szemben a Facebookon az ismerősök között zajló beszélgetéseket dolgoztuk fel elsősorban, és ebben a körben a felhasználók előszeretettel állítják be tényként a nem feltétlenül objektív állításaikat, azaz a diskurzusszintű bizonytalanság elemei lesznek gyakoribbak.

6. Statisztikai adatok

A 4. táblázatban láthatjuk a szófajok eloszlását, az 5. táblázat pedig a függőségi viszonyok eloszlását mutatja a korpuszban. Látható, hogy elsősorban a DISC reláció használata, illetve a mondatszavak, indulatszavak gyakorisága mutat nagy különbséget a domének között: a Facebookról származó szövegekben sokkal gyakoribb a használatuk, mint a Gyakorikérdések oldalon. Ez valószínűleg annak köszönhető, hogy a Facebookon a felhasználók egymás közti beszélgetéseikben sokkal inkább kimutatják érzelmeiket, illetve az adott beszélgetéshez való viszonyulásukat, mint a valamivel formálisabb Gyakorikérdések oldalon, ahol többnyire ismeretlen emberekkel társalognak.

Sajátos jelenség még a központozás hiánya vagy megléte. A Facebookról származó szövegekben arányaiban jóval kevesebb írásjelet találunk, mint a Gyakorikérdések alkorpuszban. Ez a mondatok átlagos hosszával függ össze: míg a Facebookon a felhasználók általában rövid, akár 1-2 szavas megjegyzéseket írnak, addig a Gyakorikérdések oldalon a hozzászólások többnyire hosszabbak, bővebben kifejtettek.

Az összetevők eloszlása a 6. táblázatban látható. Szembetűnő, hogy a két alkorpusz közötti különbség az XP kifejezésekben, azaz az emotikonok és egyéb, mondatszerkezetbe nem beleszóló összetevőkben mutatkozik meg (az egymáshoz viszonyított arány 82 és 17%). Ez annak köszönhető, hogy a facebookos szövegekben az emotikonok használata jóval elterjedtebb, mint a sztenderd nyelvhasználathoz közelebbi fórumhozzászólásokban, ahol a kategória főként zárójeles közbevetéseket jelölt. Érthető is, hiszen a kommentek élőbeszédszerű, azaz online-ságot és gyorsaságot megkövetelő használata megkívánja a kommunikációt kisegítő multimodális eszközök (pl mimika, gesztusok) pótlását.

4. táblázat. Szófajok (POS) eloszlása.

Szófaj	Facebook	%	Gyakorikérdések	%	Összesen	%
Főnév	1401	18,17	1700	20,88	3102	19,56
Ige	1470	19,06	1510	18,54	2980	18,79
Határozószó	1364	17,69	1321	16,22	2685	16,93
Névmás	746	9,67	834	10,24	1580	9,96
Kötőszó	565	7,33	898	11,03	1463	9,23
Névelő	530	6,87	785	9,64	1315	8,29
Melléknév	467	6,06	644	7,91	1111	7,01
Indulatszó	944	12,24	153	1,88	1097	6,92
Számnév	153	1,98	206	2,53	359	2,26
Névutó	38	0,49	78	0,96	116	0,73
Nyílt osztály	15	0,19	14	0,17	29	0,18
Ismeretlen	19	0,25	0	0,00	19	0,12
Összesen	7712	100,00	8143	100,00	15856	100,00

Különbségek láthatóak továbbá a C (kötőszó), ADJP (melléknévi frázis) és a PP (névutós kifejezés) kategóriában, amelyek a Gyakorikérdések alkorpuszban fordultak elő gyakrabban. Ezek az adatok valószínűleg a mondatok összetettségében és hosszában lévő különbségekkel magyarázhatók.

7. Automatikus szófaji egyértelműsítés és függőségi elemzés

A létrejött korpusz lehetővé tette azt is, hogy kísérleteket végezzünk a magyarlanc 2.0 [3] szófaji egyértelműsítő és függőségi elemző moduljával is. Első lépésben megnéztük, hogy a Szeged Korpusz 2.5-ön [5] tanított szófaji egyértelműsítő és függőségi modell milyen eredményeket képes elérni a webes szövegeken. A kísérleteket a Facebook és Gyakorikérdések alkorpuszokon külön-külön is, továbbá a teljes szövegállományon is elvégeztük. A függőségi elemzéshez az etalon (kézzel annotált) morfológiai kódokat használtuk fel. A kiértékeléshez a szófaji egyértelműsítés esetén a pontosság (accuracy) metrikát, míg a függőségi elemzés esetén a Labeled Accuracy Score (LAS) és Unlabeled Accuracy Score (ULA) metrikákat alkalmaztuk.

Az eredményekből azt láttuk, hogy a sztenderd szövegen tanított modellek szerényebb eredményt képesek csak elérni a webes szövegeken. Ezért megnéztük azt is, hogy webes (azaz hibával terhelt) szövegen tanítva milyen eredményt tudunk elérni. Ehhez a korpusz mondatait véletlenszerűen osztottuk fel tanító és teszthalmazra: a mondatok 20%-a került a teszthalmazba, 80%-a pedig a tanító halmazba. Az összehasonlíthatóság kedvéért azt is megvizsgáltuk, hogy ugyanezen a teszthalmazon a sztenderd szövegen tanított magyarlanc milyen eredményt képes elérni. Az így kapott eredmények a 7. táblázatban láthatók.

Az eredmények azt mutatják, hogy – nem meglepő módon – a webes szövegek nyelvi elemzése nehezebb a sztenderd szövegekénél. Ugyanakkor, ha már a tanító

5. táblázat. Függőségi viszonyok eloszlása.

Függőségi viszony	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
PUNCT	898	10,28	1541	15,60	2439	13,10
ATT	785	8,98	1279	12,95	2064	11,08
ROOT	1208	13,82	727	7,36	1936	10,40
CONJ	559	6,40	894	9,05	1453	7,80
MODE	632	7,23	735	7,44	1367	7,34
DET	540	6,18	801	8,11	1341	7,20
SUBJ	570	6,52	708	7,17	1278	6,86
COORD	524	6,00	675	6,83	1199	6,44
OBL	395	4,52	550	5,57	945	5,08
DISC	689	7,88	91	0,92	780	4,19
OBJ	306	3,50	378	3,83	684	3,67
TLOCY	332	3,80	283	2,86	615	3,30
NEG	243	2,78	249	2,52	492	2,64
PRED	219	2,51	270	2,73	489	2,63
INF	134	1,53	181	1,83	315	1,69
PREVERB	133	1,52	141	1,43	274	1,47
APPEND	154	1,76	67	0,68	221	1,19
NE	110	1,26	79	0,80	189	1,02
PLUS	125	1,43	45	0,46	170	0,91
DAT	75	0,86	74	0,75	149	0,80
LOCY	47	0,54	62	0,63	109	0,59
TTO	14	0,16	16	0,16	30	0,16
TO	11	0,13	15	0,15	26	0,14
AUX	15	0,17	7	0,07	22	0,12
TFROM	9	0,10	4	0,04	13	0,07
QUE	8	0,09	3	0,03	11	0,06
FROM	3	0,03	5	0,05	8	0,04
NUM	1	0,01	0	0,00	1	0,01
Összesen	8739	100,00	9880	100,00	18620	100,00

6. táblázat. Összetevők eloszlása.

Összetevők	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
NP	2125	23,23	2634	28,70	4759	25,97
CP	2271	24,83	1995	21,74	4266	23,28
V	1306	14,28	1302	14,19	2608	14,23
ADVP	979	10,70	1011	11,02	1990	10,86
C	568	6,21	898	9,78	1466	8,00
XP	1011	11,05	210	2,29	1221	6,66
ADJP	249	2,72	439	4,78	688	3,75
NEG	246	2,69	248	2,70	494	2,70
INF	148	1,62	204	2,22	352	1,92
PREVERB	187	2,04	143	1,56	330	1,80
PP	40	0,44	81	0,88	121	0,66
PA	16	0,17	13	0,14	29	0,16
Összesen	9146	100,00	9178	100,00	18324	100,00

7. táblázat. Szófaji egyértelműsítés és függőségi elemzés a magyarlanccal.

Tanító halmaz	**Teszthalmaz**	**Pontosság**$_{POS}$	**LAS**	**ULA**
Szeged Korpusz 2.5	100% Facebook	66,41	69,88	76,03
Szeged Korpusz 2.5	20% Facebook	64,32	75,03	80,39
80% Facebook	20% Facebook	75,11	75,43	81,89
Szeged Korpusz 2.5	100% Gyakorikérdések	79,24	80,17	82,91
Szeged Korpusz 2.5	20% Gyakorikérdések	79,18	85,11	88,36
80% Gyakorikérdések	20% Gyakorikérdések	79,54	79,57	84,96
Szeged Korpusz 2.5	100% webes szöveg	74,63	75,34	79,66
Szeged Korpusz 2.5	20% webes szöveg	71,68	80,77	84,65
80% webes szöveg	20% webes szöveg	78,97	79,9	85,01

halmazban is hibával terhelt szövegek szerepelnek, akkor sokkal jobb eredményeket tudunk elérni szófaji egyértelműsítésben, a teljes szövegállományon több mint 7 százalékpontnyi a javulás, a Facebook-szövegek esetében pedig több mint 10 százalékpont. Ez arra utal, hogy a Facebook-szövegek esetében különösen nagy jelentőséggel bír a doménadaptáció, hiszen nyelvezetük távolabb esik a sztenderd nyelvhasználattól, mint azt a Gyakorikérdések esetében láthatjuk, ahol nem számottevő a különbség a sztenderd modell és a doménen belüli modell által elért eredmények különbsége.

A függőségi viszonyokat vizsgálva valamivel árnyaltabb képet kapunk. A Facebook-szövegeken ismét látszik a tanítóhalmaz doménjének fontossága: a Facebookon tanult modell jobb eredményt ér el, mint a Szeged Korpuszon tanított modell. Ezzel szemben a Gyakorikérdések esetében számottevően jobb eredményt ér el a sztenderd szövegeken tanult modell, mint a saját doménbeli adatokon tanult modell. A különbség magyarázata feltehetőleg az, hogy a Gyakorikérdések alkorpusz – a mondatok szintaktikai szerkezetét tekintve – közelebb áll a sztenderd szövegekhez, mint a Facebook-szövegek, így a nagyságrendekkel nagyobb tanító adathalmazon (kb. 60 000 mondaton) tanított modell 3-4 százalékponttal nagyobb pontosságot képes elérni, mint a saját doménen (kb. 600 mondaton) tanított modell. A jövőben tervezett doménadaptációs kísérleteink remélhetőleg pontosabb képet nyújtanak majd a magyarlanc webes szövegekre történő adaptálási lehetőségeiről.

8. Összegzés

Cikkünkben bemutattuk az első magyar, kézzel annotált, webes szövegeket tartalmazó korpuszt, melyet morfológiai és (összetevős és függőségi szemléletű) szintaktikai elemzést, valamint szemantikai és diskurzusbeli bizonytalan kifejezések annotációját tartalmazza. Ismertettük az annotáció folyamatát, illetve beszámoltunk a magyarlanc elemző webes szövegekre történő adaptálási kísérleteiről.

A korpusz méreteinél fogva nem alkalmas statisztikai elemzők tanítására, célunk egy benchmark adatbázis előállítása volt. Véleményünk szerint mivel a webes szövegek témában és műfajban is igen változatosak, nem is lenne célravezető

a felügyelt gépi tanulás paradigmáját követni, hanem doménadaptációs megoldások jelenthetik a megoldást. A jövőben tovább kívánunk foglalkozni a webes szövegekre adaptált elemzők továbbfejlesztésével, illetve terveink között szerepel a korpusz újabb annotációs rétegekkel (névelemek, többszavas kifejezések) való kézi bővítése is.

A korpusz oktatási és kutatási célokra ingyenesen elérhető a http://rgai.u-szeged.hu/SzegedTreebank oldalon.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében valósult meg. Vincze Veronika kutatásait a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosítószámú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt támogatta. Mindkét projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Mott, J., Bies, A., Laury, J., Warner, C.: Bracketing Webtext: An Addendum to Penn Treebank II Guidelines. Linguistic Data Consortium (2012)
2. Seddah, D., Sagot, B., Candito, M., Mouilleron, V., Combet, V.: The French Social Media Bank: a treebank of noisy user generated content. In: Proceedings of COLING 2012, Mumbai, India, The COLING 2012 Organizing Committee (2012) 2441–2458
3. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013) 763–771
4. Varga, V., Wieszner, V., Hangya, V., Vincze, V., Farkas, R.: Magyar nyelvű webes szövegek számítógépes feldolgozása. In Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2014) 327–331
5. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC'14, Reykjavik, Iceland (2014) 1074–1078
6. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
7. Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. Computational Linguistics **38** (2012) 335–367
8. Vincze, V.: Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 383–391
9. Vincze, V., Simkó, K.I., Varga, V.: Annotating Uncertainty in Hungarian Webtext. In: Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop, Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 64–69

Finnugor nyelvű közösségek nyelvtechnológiai támogatása online tartalmak létrehozásában

Benyeda Ivett, Koczka Péter, Ludányi Zsófia, Simon Eszter, Váradi Tamás

MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33.
e-mail:
{benyeda.ivett,koczka.peter,ludanyi.zsofia,simon.eszter,varadi.tamas}@nytud.mta.hu

Kivonat A cikkben bemutatott folyamatban levő projekt célja, hogy kisebb finnugor nyelvekre állítson elő nyelvi erőforrásokat, amelyekkel revitalizálni lehet ezeket a veszélyeztetett nyelvi közösségeket. A projekt során párhuzamos és összevethető korpuszokból kétnyelvű protoszótárakat állítunk elő, melyeket anyanyelvi beszélők fognak ellenőrizni. A különböző nyelvű, egymásnak megfeleltetett szóalakok morfológiai, lexikai, etimológiai információkkal kibővítve kerülnek majd feltöltésre a Wiktionarybe. A projekt során számolnunk kell azzal a nehézséggel, hogy nyelvtechnológiai erőforrások a kisebb finnugor nyelvekre kevéssé állnak rendelkezésre, ezért a szövegfeldolgozás során nyelvfüggetlen gépi tanulási módszereket alkalmazunk. A projekt összes melléktermékét (modellek, korpuszok, szövegfeldolgozó eszközláncok, elemzett szövegek) nyilvánosan elérhetővé tesszük.

Kulcsszavak: párhuzamos korpusz, összevethető korpusz, automatikus szótárgenerálás, finnugor nyelvek, veszélyeztetett nyelvek

1. Bevezetés

Hagyományosan veszélyeztetett nyelvnek azokat a nyelveket szokták nevezni, amelyeknek kevés a beszélője, azok is az idősebb generációba tartoznak, a beszélők száma egyértelműen csökken, és a nyelvhasználat területe határozottan az informális, családi keretek felé tolódik. Kornai [1] a fenti tényezők mellett – többek között – a nyelvtechnológiai (és tágabban az infokommunikációs) eszközök használatát és a webes tartalmak előállításának ütemét is beleveszi a nyelvek állapotának kiértékelésébe. Az újfajta szempontrendszer alapján az is fontos kritérium, hogy az egyes nyelvek mennyire vannak jelen a digitális térben: milyen mennyiségben születnek az interneten nyilvánosan hozzáférhető szövegek az adott nyelven. Kornai szerint a digitális nyelvhalál a következőket foglalja magában: funkció- és ezzel együtt presztízsvesztés, és végül a nyelvi kompetencia elvesztése.

A nyelvtechnológia ebben a kontextusban egyfajta támogató technológiaként tud működni: a szabad és nyilvános nyelvhasználatot támogatva, a nyelvi határokat ledöntve segíti a kommunikációt [2]. Nyelvtechnológiai alkalmazások és

erőforrások viszont leginkább a széleskörűen használt nyelvekre léteznek – ezeket nevezi Kornai [1] ún. viruló nyelveknek. Ennek legfőbb oka az, hogy ezeken a nyelveken érhető el digitális szöveges tartalom. A kisebb, veszélyeztetett nyelvek ebből a szempontból is hátrányban vannak, hiszen hozzáférhető digitális tartalom híján nyelvtechnológiai eszközöket is sokkal nehezebb rájuk fejleszteni.

Cikkünkben egy olyan folyamatban lévő projektet mutatunk be, amelynek célja, hogy segítse a veszélyeztetett finnugor nyelvű közösségeket nyelvük felvirágoztatásában azáltal, hogy online tartalmakat hoz létre az adott nyelveken. A projekt során több veszélyeztetett finnugor és néhány széles körben használt viruló nyelvre állítunk elő protoszótárakat, amelyeket lexikai információkkal gazdagítva feltöltünk a Wiktionarybe.

A munkafolyamat első lépése a szöveggyűjtés, valamint párhuzamos és összevethető korpuszok építése (l. 3.1. fejezet). Az alábbi finnugor nyelvekre gyűjtöttünk szövegeket: komi-zürjén, komi-permják, udmurt, mezei és hegyi mari, valamint északi számi. A kis finnugor nyelvek mellett azokra a viruló nyelvekre is kerestünk szövegeket, amelyek a finnugrisztikában fontos szerepet töltenek be, ezek: angol, orosz, finn és magyar.

A párhuzamos és összevethető korpuszok előfeldolgozása automatikusan történik. Tekintve, hogy kifejezetten a szóban forgó kis finnugor nyelvekre fejlesztett tokenizáló és mondatra bontó eszközök nem léteznek, többféle gépi tanulási módszerrel kísérleteztünk. A nyelvi erőforrások hiánya a morfológiai elemzés szintjén több problémát is felvet: felügyelt gépi tanulási módszerek alkalmazása nem lehetséges, mivel a tanításhoz és teszteléshez morfológiailag annotált szövegekre lenne szükség, ezek viszont nem állnak rendelkezésre (l. 3.2. fejezet).

Az összegyűjtött párhuzamos és összevethető szövegeket felhasználva, többféle szótárépítési metódust követve ún. protoszótárakat állítunk elő (l. 4. fejezet), amelyek jelenleg nyelvpáronként néhány száz fordítási jelöltet tartalmaznak. A későbbiekben ezek alapján készülnek majd el azok a szótárak, amelyeket a vizsgált finnugor nyelvek anyanyelvi beszélői fognak kézzel ellenőrizni és javítani. Ezekben a végső szótárakban lesznek azok a szótári elemek, amelyek bizonyos morfológiai információkkal (szófaj, ragozási paradigma), etimológiai, illetve lexikai-szemantikai adatokkal (szinonimák, antonimák) kibővítve kerülnek feltöltésre a Wiktionarybe.

2. Kitekintés

A kétnyelvű szótáraknak nem csupán a gépi fordításban [3] és a nyelvközi információkinyerésben [4] van kritikus szerepük, hanem egyéb nyelvtechnológiai alkalmazásokban is, például a nyelvtanulásban [5], a számítógépes szemantikában, továbbá számos olyan feladatban, ahol megbízható lexikai-szemantikai információra van szükség [6]. Tekintve, hogy a kézzel történő szótárkészítés rendkívül erőforrásigényes, meglehetősen ritkák a szabadon hozzáférhető hagyományos kétnyelvű szótárak. Komplett kétnyelvű szótárak teljesen automatikus módon történő előállítását a jelenlegi technológia nem teszi lehetővé, de a protoszótárak támogatást tudnak nyújtani a lexikográfiai munkához.

A sztenderd szótárépítési módszerek alapjául párhuzamos korpuszok szolgálnak, amelyek az eredeti nyelvű szöveget és annak fordítását tartalmazzák, jellemzően mondatszinten párhuzamosítva. Viszont, ahogy Rapp [7] fogalmaz: mindig kivételes esetnek számít, ha egy adott doménre és adott nyelvpárra elégséges méretű párhuzamos korpusz áll rendelkezésre; általánosnak inkább az tekinthető, ha nincs ilyen. Ilyen korpuszok ugyanis jobbára csupán a legtöbb erőforrással rendelkező nyelvpárokra léteznek. Ez az egyik oka annak, hogy egyre nő az összevethető (nem párhuzamos) korpuszok előállítása iránti érdeklődés.

Az összevethető korpuszokból történő szótárépítési metodológia sztenderd megközelítése kontextusvektorok hasonlóságát méri a két vizsgált nyelvre [8,7]. Ennek lépései a következők: kontextusvektorok létrehozása és fordítása, a forrás- és a célnyelvi vektorok összehasonlítása és a fordítási jelöltek rangsorolása valamilyen hasonlósági metrika alapján. Ehhez a módszerhez szükség van egy ún. magszótárra, amelynek használatával újabb fordítási párokat nyerhetünk ki a szövegekből. A módszer hátránya, hogy a teljesítménye erősen függ a magszótár, a kontextus és a korpusz méretétől, valamint a választott hasonlósági metrikától is. Mivel az általunk vizsgált finnugor nyelvekre nem áll rendelkezésre megfelelő méretű korpusz és szótár, alternatív módszerekkel kell kísérleteznünk. Számos újabb módszert alkalmaztak nem párhuzamos korpuszokból történő fordítási párok kinyerésére, például [9,10,11]. Mivel az idézett cikkekben leírt módszerekhez tartozó forráskódok nem publikusak, az eredmények nem reprodukálhatók. Az egyik legújabb trend a nyelvtechnológiában a neurális hálón alapuló vektoros nyelvmodellek használata, amelyet többek között kétnyelvű szótárak előállítására is alkalmaznak (pl. [12]). Ennek a módszernek az általunk vizsgált nyelvpárokra való adaptálását tervezzük elvégezni a projekt során.

3. Párhuzamos és összevethető korpuszok építése

A célkitűzések megvalósításához első lépésként az interneten elérhető párhuzamos és összevethető szövegeket gyűjtöttünk az általunk vizsgált nyelvpárokra. A szótárépítéshez elengedhetetlenül szükséges a gyűjtött szövegek alapszintű nyelvi feldolgozása (tokenizálás, mondatra bontás, lemmatizálás, morfológiai elemzés és egyértelműsítés). Ebben a fejezetben a korpuszépítési munkafolyamat lépéseit ismertetjük.

3.1. Szöveggyűjtés

McEnery és Xiao [13] definícióját követve akkor beszélhetünk párhuzamos korpuszról, ha a korpuszt felépítő szövegek egy az egyben egymás fordításai. Ha a korpusz különböző nyelvű részei nem pontos fordításai egymásnak, de a mintavétel módját tekintve megegyeznek, akkor összevethető korpuszról beszélünk. Időnként azonban nem teljesen egyértelmű, hogy egy többnyelvű szöveghalmaz párhuzamos vagy inkább összevethető korpuszként kezelendő. Szigorúan véve csak a biblia- és regényfordítások, a szoftverdokumentációk és az olyan hivatalos dokumentumok, mint például az Egyetemes Emberi Jogok Nyilatkozata, tekinthetők valódi párhuzamos szövegeknek. A Wikipédia-szócikkek többé-kevésbé

egymás fordításai, de a szócikkek különböző nyelvbeli megfelelői között igen jelentős különbségek lehetnek, így ezek nem tekinthetők párhuzamos szövegeknek, de összevethető korpuszok építésére jól hasznosíthatók.

Párhuzamos korpuszok. A Bibliának párhuzamos szövegként történő felhasználása régi hagyománnyal bír a szótárépítésben [14], így mi is alkalmaztuk ezt a módszert. A Parallel Bible Corpus [15], a Bible.is és a The Unbound Bible oldalakról letöltöttük az Újszövetség fordításait a szóban forgó nyelvekre. Hogy a készülő szótárak ne tartalmazzanak archaikus és kihalt szavakat, mindig a legújabb bibliafordítást választottuk. A fordítások versszinten párhuzamosítva vannak, így a szövegek további feldolgozása könnyűszerrel megoldható. Nehézséget jelent azonban, hogy bizonyos nyelvekre (udmurt, hegyi mari, komi-permják) nem találtunk elektronikus szöveges formátumban elérhető bibliafordításokat.

A Biblián kívül további párhuzamos korpuszként használható az OPUS korpusz [16], amelyben találtunk északi számi szoftverdokumentációt párhuzamosítva mind a négy viruló nyelvi megfelelőjével. Párhuzamos korpuszok forrásaként használhatók továbbá Finnország és Norvégia egyes hivatalosan kétnyelvű régióinak weboldalai is. Olyan párhuzamos szövegeket, ahol az egyik nyelv komi-permják, hegyi mari, illetve udmurt, sajnos nem találtunk.

Összevethető korpuszok. Az összevethető korpuszok előállításához az egyik leggyakrabban használt forrás a Wikipédia. A munka első fázisaként minden általunk vizsgált nyelvre letöltöttük a Wikipédia dump fájlokat. Ezt követően a nyelvközi linkek segítségével összepárosítottuk az azonos témájú, különböző nyelveken íródott cikkeket. A szövegek kinyeréséhez a Wikipedia Extractort[1] használtuk, annyi módosítással, hogy a cikkszövegek mellett megtartottunk néhány egyéb metaadatot is (szövegben előforduló nyelvközi linkek, Wikidata-azonosítók), amelyek a további feldolgozást segítik. A Wikidata a Wikipédia testvérprojektje: egy ingyenes, közösség által szerkesztett, többnyelvű tudásbázis, ahol a nyelvközi linkek által összekapcsolt Wikipédia-címszavak egy és ugyanazon entitáshoz tartoznak, egyetlen Wikidata-azonosítóval. Ezek az azonosítók alkalmasak arra, hogy megtaláljuk a szövegekben lévő azonos névelemeket – a cikk nyelvétől függetlenül –, valamint horgonyként használva segítséget nyújthatnak a szövegek párhuzamosításában is.

Míg a nagyméretű, aktív digitális közösség által szerkesztett és karbantartott Wikipédia-cikkek elég terjedelmesek, a kis nyelvi közösséggel rendelkező Wikipédiák esetében a cikkek mennyisége és terjedelme jóval kisebb. Ebből kifolyólag az egymásnak megfelelő különböző nyelvű cikkek hossza általában meglehetősen eltérő. Feltételezve, hogy a cikkek első, definíciós része minden nyelven nagyjából megfelel egymásnak, minden cikkpárnál a szöveg első x mondatát tartottuk meg, ahol x egyenlő a finnugor nyelvű cikk mondatainak számával.

A sztenderd megközelítés szerint (pl. [8]) azok a szövegek is összevethető korpuszokként kezelhetők, amelyek két vagy több nyelvű újságcikkeket, híreket

[1] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

tartalmaznak ugyanabból az időintervallumból és ugyanarról a helyről. Ez utóbbi konstelláció miatt feltételezhetjük, hogy a cikkek ugyanazokról a helyi vagy fontosabb globális történésekről számolnak be, így ha nem is egymás fordításai, témában nagyon közel állnak egymáshoz. Erre alapozva gyűjtöttünk cikkeket finnországi online újságok honlapjairól északi számi–{finn, angol, orosz} nyelvpárokra.

További módszer összevethető korpuszok építésére az azonos téma köré szerveződő alkorpuszok felhasználása, vagyis olyan egynyelvű szövegek letöltése különböző nyelveken, amelyek azonos tárgykörhöz tartoznak [8]. E korpuszok létrehozásához olyan szövegeket töltöttünk le északi számi és angol nyelven, amelyek a számi kultúráról, oktatásról és társadalomról szólnak.

Egynyelvű szövegek. Az előbbiek mellett minden kis finnugor nyelvre létrehoztunk egynyelvű korpuszokat is. Míg a párhuzamos és összevethető korpuszokat szótárépítésre használjuk, az egynyelvű korpuszok tanítóanyagként funkcionálnak a tokenizáló és mondatra bontó alkalmazások számára. Az egynyelvű korpuszokat felépítő anyagokat különféle weboldalakról töltöttük le, így ezek témában igen változatosak (pl. irodalmi szövegek, hírek, személyes blogok, hivatalos szövegek).

Az 1. táblázat a párhuzamos, az összevethető és az egynyelvű korpuszok tokenszámát mutatja. Az egynyelvű szövegek tokenszámába a párhuzamos és az összevethető korpuszok adott nyelvű részei is bele vannak számolva. Az összevethető korpuszok számadatai a csökkentett méretű Wikipédia-cikkek szövegeire vonatkozóan értelmezendők, vagyis a cikkeknek kizárólag az első x mondatát tartalmazzák (lásd feljebb). Az összevethető korpuszok közül az időintervallum-alapúak évenkénti bontásban lettek számolva, vagyis nem számoltuk bele azokat a szövegeket, amelyeknek nem volt ugyanazon évből származó másik nyelvű megfelelője. A táblázat adatai a szöveggyűjtés jelenlegi állapotát mutatják; a számok a projekt előrehaladtával természetesen változnak.

3.2. Szövegfeldolgozás

A szótárelőállítás további lépéseihez elengedhetetlenül szükséges az összegyűjtött szövegek minél pontosabb alapszintű nyelvi feldolgozása, vagyis a tokenizálás, a mondatra bontás, a morfológiai elemzés és egyértelműsítés, mivel az ezen feldolgozási szakaszokban bekövetkezett hibák jelentős problémákat okozhatnak a magasabb feldolgozási szinteken, illetve a szótárépítésben. Sajnos a cirill betűs finnugor nyelvekre nem találtunk tokenizáló és mondatra bontó eszközöket. Az egyetlen kis finnugor nyelv, amely nyelvtechnológiai eszközökkel kellően támogatott, az a latin ábécés északi számi. Ez nem okozhat különösebb meglepetést, hiszen az északi számi rendelkezik a legjelentősebb mértékű online forrásokkal, beleértve ebbe a Tromsői Egyetemen fejlesztett eszközöket[2] és az online elérhető tartalmakat.

[2] http://giellatekno.uit.no/cgi/index.sme.eng.html

1. táblázat. Az egynyelvű, párhuzamos és összevethető korpuszok tokenszámai. A táblázatban szereplő ISO 639-3 nyelvkódok: sme – északi számi, kpv – komi-zürjén, koi – komi-permják, mhr – mezei mari, mrj – hegyi mari, udm – udmurt; eng – angol, fin – finn, rus – orosz, hun – magyar.

nyelv	**egynyelvű**	**nyelvpár**	**párhuzamos**		**összevethető**	
			L1	**L2**	**L1**	**L2**
sme	1.364.254	sme-eng	691.260	724.750	253.930	1.754.968
		sme-fin	245.440	273.973	239.651	5.259.591
		sme-rus	173.179	220.790	212.332	233.748
		sme-hun	171.668	224.014	86.244	106.391
kpv	480.609	kpv-eng	121.108	174.742	89.580	183.602
		kpv-fin	121.120	133.715	88.507	80.797
		kpv-rus	117.903	125.085	108.013	141.369
		kpv-hun	121.319	134.344	68.179	74.274
koi	719.325	koi-eng	0	0	257.871	194.784
		koi-fin	0	0	137.578	77.696
		koi-rus	0	0	188.334	139.976
		koi-hun	0	0	95.120	64.794
mhr	1.335.457	mhr-eng	128.316	175.075	121.588	250.583
		mhr-fin	128.328	133.965	118.120	115.028
		mhr-rus	109.449	109.818	158.977	215.724
		mhr-hun	128.565	134.618	106.813	121.453
mrj	366.964	mrj-eng	0	0	137.088	306.465
		mrj-fin	0	0	85.134	93.622
		mrj-rus	0	0	124.289	187.687
		mrj-hun	0	0	77.855	90.168
		mrj-hun	0	0	77.855	90.168
udm	584.113	udm-eng	0	0	67.306	135.450
		udm-fin	0	0	56.222	49.961
		udm-rus	0	0	80.800	129.293
		udm-hun	0	0	41.883	48.736

Előfeldolgozás. Ahogy a 3.1. fejezetben szó volt róla, viszonylag nagy mennyiségű egynyelvű szöveget gyűjtöttünk minden nyelvre. Nehézséget jelent azonban, hogy az egyes szövegek több helyütt tartalmaznak nem odavaló szövegrészeket.

Az első nehézség, hogy a cirill betűs finnugor nyelvek az ábécé módosított verzióit használják, amelyekben sok a különféle diakritikus jelekkel ellátott cirill karakter. Ezek a speciális karakterek gyakran az alapkarakter és a diakritikus jel kombinációjából állnak elő, amelyek így a későbbi lépésekben használt eszközök számára külön karakterekként értelmeződnek. Ennek elkerülése érdekében a további szövegfeldolgozás előtt karakternormalizálást kell végezni minden forráson.

A második, szintén jelentős probléma a cirillel írt nyelvek esetében, hogy az egymással közeli rokonságban álló nyelvek (komi-permják és zürjén, mezei és hegyi mari) gyakran egyszerre is megjelenhetnek egy dokumentumon belül, ezért az ilyen részeket el kell különítenünk egymástól. A nyelvek megkülönböztetésére a Blacklist Classifiert[3] használtuk, amely 97,47%-os pontossággal szűri a komi-zürjén és komi-permják, 96,77%-os pontossággal pedig a mezei és hegyi mari nyelveket.

A harmadik probléma, hogy bizonyos finnugor nyelvű személyes blogok angol vagy orosz nyelvű blogszolgáltatókon találhatók, így az egyébként egynyelvű blogbejegyzések nyelvileg kevertek, mivel számos hasznosítható információ, a dátumok és a weblap bizonyos elemei nem a kívánt finnugor nyelven szerepelnek a szövegben. Az idegen nyelvű részek kiszűrésére egy trigramstatisztikát és Katz–Backoff-simítást alkalmazó nyelvfelismerő szkriptet, a Langid-t[4] használtuk. A szükséges modellek kézzel válogatott szövegek felhasználásával készültek. A dátumokat minden esetben megtartottuk, mivel ez alapján az információ alapján tudjuk előállítani az időintervallum-alapú összevethető korpuszokat.

Mondatra bontás, tokenizálás. A mondatra bontáshoz és tokenizáláshoz az Apache OpenNLP[5] mondatra bontó és tokenizáló moduljait használtuk. Ahogy már említettük, az északi számi igen jól támogatott NLP-eszközöket illetően, ezért csak a cirill ábécét használó finnugor nyelvekre építettünk modelleket. Az OpenNLP-nek mind a tokenizáló, mind a mondatra bontó eszköze 98% feletti F-mértékkel teljesített. Ez a teljesítmény részben annak köszönhető, hogy a mondatra bontó rövidítésszótár használatát is lehetővé teszi, így ennek segítségével elkerülhető az, hogy az eszköz tévesen mondathatárként ismerje fel a rövidítések utáni pontot. A mondatra bontásban bevett szokás a rövidítéslisták használata, bár nyilvánvalóan nem lehetséges egy teljeskörű lista létrehozása, különösen a kutatásunkban releváns finnugor nyelvek esetében. Az általunk használt rövidítésszótár a Wiktionary orosz rövidítéslistáján alapul; ezt a későbbiekben kibővítjük a kis finnugor nyelvekben előforduló rövidítésekkel. Mindazonáltal az orosz rövidítéseket tartalmazó szótár használata jelenleg elégségesnek bizonyult, hiszen a cirillel írt finnugor nyelvekben használt rövidítések gyakran azonosak az orosz rövidítésekkel.

[3] https://bitbucket.org/tiedemann/blacklist-classifier/wiki/Home
[4] https://github.com/juditacs/langid
[5] https://opennlp.apache.org/

Morfológiai elemzés és egyértelműsítés. Online hozzáférhető morfológiai elemző és/vagy egyértelműsítő elérhető északi számi[6], udmurt és komi-zürjén[7], valamint hegyi mari[8] nyelvekre. Legjobb tudomásunk szerint azonban mezei mari és komi-permják nyelvekre nem létezik morfológiai elemző. Ezen a feldolgozási szinten az erőforrások hiánya még komolyabb problémát okoz, hiszen ezekre a nyelvekre még morfológiai annotációval ellátott szövegek sincsenek, amiken tanítani lehetne egy felügyelt gépi tanuló rendszert.

A morfológiai elemzővel nem rendelkező nyelvek esetében több lehetőség közül választhatunk. Az egyik opció, ha egy félig felügyelt vagy felügyelet nélküli morfológiai szegmentáló eszközt használunk (pl. Morfessor[9]), annak működését igényeink szerint kibővítve. Végeztünk ilyen irányú kísérleteket a Morfessorral: egy udmurt szólistán betanítottuk, és a szegmentált kimenetet összehasonlítottuk egy udmurt morfológiai elemző [17] kimenetével. A biztató eredmények ellenére azonban jelentős munkát igényelne olyan további eszközök fejlesztése, amelyek lemmákat és morfológiai címkéket adnának a Morfessor kimenetéhez.

Egy másik lehetőség az, hogy a közeli rokonságban álló nyelvekre alkalmazunk már létező eszközöket. A legegyszerűbb megoldás az, ha egy eszközt közvetlenül a rokon nyelvre alkalmazunk, vagyis pl. a komi-zürjénre kifejlesztett modellt közvetlenül használjuk a komi-permják adatokon. Reményeink szerint a komi-zürjén morfológiai jegyek közvetlenül átvihetők az azonos szöveg komi-permják változatára, valamint ugyanígy járnánk el a hegyi és mezei mari nyelvek esetében. Mivel általánosságban véve a nyelvek nagy részére nem áll rendelkezésre elégséges mennyiségű tanítóanyag, a közeli rokonságban álló nyelvek közötti annotációátvitelre épülő különböző kísérletek az utóbbi időben kitüntetett szerepet élveznek az NLP-kutatásokban (pl. [18,19]). Terveink között szerepel annak további vizsgálata, hogy hogyan lehet átültetni az annotációkat egy nyelvtechnológiailag jobban támogatott nyelv kevésbé támogatott közeli rokon nyelvére.

4. Protoszótárak építése

A jelenlegi módszerek nem teszik lehetővé kétnyelvű szótárak teljesen automatikusan történő létrehozását, de kivitelezhető bizonyos lexikai erőforrások, ún. protoszótárak gépi előállítása, amelyek nagy segítséget jelenthetnek a szótárépítési munkálatokban. A hagyományos, manuálisan készített szótáraknál a protoszótárak általában nagyobb méretűek, és a lexikai elemek nagyobb lefedettségét biztosítják, ugyanakkor jóval több nem megfelelő fordítási jelöltet tartalmaznak. A protoszótárak méretének kiválasztása ezért nagyban függ a későbbi felhasználás módjától.

Többféle szótárépítési módszerrel kísérleteztünk, melyeket az alábbiakban ismertetünk. Mindegyik módszerrel több száz fordítási jelöltet tartalmazó protoszótárat hoztunk létre majdnem minden nyelvpárra. Ezek a szótárfájlok a

[6] http://giellatekno.uit.no/cgi/index.sme.eng.html
[7] http://www.morphologic.hu/urali/
[8] http://www.univie.ac.at/maridict/site-2014/morph.php
[9] http://morfessor.readthedocs.org/en/latest/general.html#techrep

későbbi végleges szótárak kiindulási pontjaként fognak szolgálni, melyekbe csak a legvalószínűbb fordítási párok fognak bekerülni. A fordítások kézi ellenőrzését anyanyelvi beszélők végzik majd a projekt utolsó szakaszában.

4.1. Létező szótárak felhasználása

Online szótárakat minden finnugor nyelv legalább egy nyelvpárjára találtunk. Ezek többségében online használhatók, néhány azonban letölthető változatban is hozzáférhető volt. A HTML-fájlok feldolgozásával megtörtént a szópárok kinyerése, aminek eredményeképpen kétnyelvű szótárak jöttek létre néhány nyelvpárra, átlagosan néhány száz szópár méretben. Ezek a szótárak az automatikus szótárépítési munkák során segítséget nyújthatnak egyrészt a párhuzamos vagy összevethető cikkpárok mondat- és frázisszintű illesztésénél, másrészt a szótárgeneráló algoritmusok számára kiinduló, ún. magszótárakként szolgálhatnak.

4.2. Wikipédia-címszavak

A Wikipédia nemcsak a legnagyobb, szabadon elérhető adatbázis, amely összevethető szövegeket tartalmaz, de többféle módon is felhasználható kétnyelvű szótárak létrehozására. Erdmann et al. [20] címszavakból készített kétnyelvű szótárakat használt a cikkek szövegeiből történő további fordítási párok kinyeréséhez. Mohammadi és Quasim Aghaee [21] az angol és a perzsa Wikipédiából párhuzamos mondatpárokat nyert ki, szintén Wikipédia-címszavakból épített szótárakat felhasználva. Ezt a módszert követve kétnyelvű szótárakat hoztunk létre Wikipédia-címpárokból a nyelvközi linkek segítségével, amely nyelvpáronként további néhány száz elemű szótárat eredményezett.

4.3. Wiktionaryre épülő módszerek

A Wikipédia mellett a Wiktionary egy másik, szintén nyílt, közösség által szerkesztett tudásbázis, amely kiváló forrásul szolgálhat kétnyelvű szótárak létrehozásához. Bár a Wiktionary elsősorban emberi felhasználásra készült, a benne található adatok kinyerése bizonyos fokig automatizálható. Ács et al. [22] minden Wictionary-cikkhez tartozó fordítási elemet kinyert a cikkekben található fordítási táblákból. Mivel az általuk fejlesztett Wikt2dict eszköz[10] szabadon elérhető, fel tudtuk használni a projektünkben szerepet kapó nyelvpárokra. Az angol, finn, orosz és magyar Wiktionary oldalak parszolásával szinte minden szóban forgó nyelvpárra számos fordítási szópárt sikerült kinyernünk.

Ács [23] a szópárok halmazát újabbakkal bővítette, oly módon, hogy új kapcsolatokat hozott létre a már meglévő fordítási párokból egy ún. háromszögelési módszerrel. A háromszögelés azon a feltételezésen alapul, hogy két elem nagy valószínűséggel fordításpár abban az esetben, ha mindkettő egy harmadik nyelv szavának fordításpárja. A Wikt2dict háromszögelési technikájával szótárainkat további néhány száz elemmel tudtuk bővíteni.

[10] https://github.com/juditacs/wikt2dict

4.4. Hundict

A Hundict[11] egy kísérleti projekt kétnyelvű szótárak párhuzamos szövegekből való előállítására. A program az egymásnak megfeleltetett szövegekből nyeri ki a gyakran együtt előforduló szópárokat a Sørensen–Dice-együttható alapján. Az eszköz hatékonysága növelhető gold standard szótárak és stopszólisták hozzáadásával. Végeztünk néhány kísérletet bibliafordításokkal északi számi–finn és komi-zürjén–angol nyelvpárokra, amelyek eredményképpen olyan fordítási párokat kaptunk, ahol a szópárok mellett a fordítások konfidenciaértéke is szerepel. A rendszer több olyan paramétert is tartalmaz, melyek további finomhangolása szükséges az elérhető legjobb eredmény érdekében. Terveink közt szerepel ezek kimérése és további nyelvpárokra való felhasználása is, bár az eszköz lemmatizált szövegeket kíván bemenetként, így előbb a párhuzamos korpuszok lemmatizált változatát kell elkészítenünk.

5. Összegzés, további feladatok

Cikkünkben egy olyan folyamatban lévő projektet mutattunk be, amelynek forrásául a webről letöltött párhuzamos és összevethető korpuszok szolgálnak. A projekt fő célja, hogy olyan protoszótárakat hozzunk létre automatikus módszerekkel, amelyeknek egyik nyelve az alábbi kis finnugor nyelvek egyike: komi-zürjén, komi-permják, mezei és hegyi mari, udmurt, valamint északi számi. Ezekre a nyelvekre kevés nyelvtechnológiai eszköz készült, sőt bizonyos nyelvek esetében még digitális szöveges tartalmak is csupán igen kis számban lelhetők fel, ebből kifolyólag mind a szöveggyűjtés, mind a szövegfeldolgozás nagy kihívásokat jelent. A szövegfeldolgozás alsóbb szintjein problémát jelent az, hogy kifejezetten a szóban forgó kis finnugor nyelvekre nem létezik tokenizáló és mondatra bontó eszköz. A morfológiai elemzés szintjén további nehézségekkel kell számolnunk: nem alkalmazhatók felügyelt gépi tanulási módszerek, mivel nem áll rendelkezésünkre tanításhoz és teszteléshez szükséges morfológiai annotációval ellátott szöveg.

A leírt nehézségek ellenére kellő méretű egy- és többnyelvű szöveget gyűjtöttünk az általunk vizsgált nyelvpárokra, melyeket felhasználva nyelvfüggetlen módszerekkel protoszótárakat állítottunk elő. A létrehozott szótárakat bizonyos morfológiai, etimológiai, szemantikai információkkal és többnyelvű fordítási megfelelőikkel kibővítve feltöltjük a Wiktionarybe. Mind a szótárakat, mind a nyelvészeti információkat a lehetőségekhez mérten automatikusan állítjuk elő. Természetesen nem nélkülözhető a kézi kiértékelés és javítás, ezt anyanyelvi beszélők fogják végezni a projekt utolsó szakaszában.

A Wiktionary rendszerét felhasználva a szótári elemek a Wiktionary különböző nyelvű változataiban összekapcsolhatók. Ez lehetővé teszi, hogy a közösség gazdag lexikai anyagokhoz férjen hozzá, mindemellett olyan új adatok is elérhetők lesznek a lexikai elemekhez, mint például az etimológiai adatok vagy a fordítási megfelelők. A Wiktionary sajátos markup formátumot használ, amit

[11] https://github.com/zseder/hundict

mi XML-formátumra alakítunk a további feldolgozás érdekében. A jogi kérdések tisztázása után az összes létrehozott anyagot: a korpuszokat, a szótárakat, illetve a nyelvmodelleket publikusan elérhetővé tesszük.

Köszönetnyilvánítás

A projektet az Országos Tudományos Kutatási Alapprogram támogatja (szerződésszám: 107885).

Hivatkozások

1. Kornai, A.: Digital Language Death. PLoS ONE **8**(10) (2013)
2. Simon, E., Lendvai, P., Németh, G., Olaszy, G., Vicsi, K.: A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age. Georg Rehm and Hans Uszkoreit (Series Editors): META-NET White Paper Series. Springer (2012)
3. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
4. Grefenstette, G.: The Problem of Cross-Language Information Retrieval. In Grefenstette, G., ed.: Cross-Language Information Retrieval. Kluwer Academic Publishers (1998) 1–9
5. Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J.B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., Volodina, E.: Corpus-Based Vocabulary lists for Language Learners for Nine Languages. Language Resources and Evaluation **48**(1) (2013) 121–163
6. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08), Marrakech, Morocco, ELRA (2008)
7. Rapp, R.: Identifying word translations in non-parallel texts. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics. ACL '95, Stroudsburg, PA, USA, Association for Computational Linguistics (1995) 320–322
8. Fung, P., Yee, L.Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: Proceedings of the 17th International Conference on Computational Linguistics – Volume 1. COLING '98, Stroudsburg, PA, USA, Association for Computational Linguistics (1998) 414–420
9. Hazem, A., Morin, E.: ICA for Bilingual Lexicon Extraction from Comparable Corpora. In: The 5th Workshop on Building and Using Comparable Corpora, Istanbul, Turkey (2012) 126–133
10. Tamura, A., Watanabe, T., Sumita, E.: Bilingual lexicon extraction from comparable corpora using label propagation. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. EMNLP-CoNLL '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 24–36
11. Vulić, I., Moens, M.F.: Detecting highly confident word translations from comparable corpora without any prior knowledge. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. EACL '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 449–459

12. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed Word Representations for Multilingual NLP. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, Association for Computational Linguistics (2013) 183–192
13. McEnery, A., Xiao, R.: Parallel and comparable corpora: What are they up to? In James, G., Anderman, G., eds.: Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters (2007)
14. Resnik, P., Olsen, M.B., Diab, M.: The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. Computers and the Humanities **33**(1–2) (1999) 129–153
15. Mayer, T., Cysouw, M.: Creating a massively parallel Bible corpus. In: Proceedings of LREC '14, Reykjavik, Iceland, ELRA (2014)
16. Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, ELRA (2012)
17. Novák, A.: Morphological Tools for Six Small Uralic Languages. In: Proceedings of LREC '06, ELRA (2006)
18. Scherrer, Y., Sagot, B.: A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, ELRA (2014) 502–508
19. Ingason, A.K., Loftsson, H., Rögnvaldsson, E., Sigurdsson, E.F., Wallenberg, J.C.: Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. In: Proceedings of LREC'14, Reykjavik, Iceland, ELRA (2014) 91–95
20. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia. ACM Transactions on Multimedia Computing, Communications, and Applications **5**(4) (2009) 31:1–31:17
21. Mohammadi, M., GhasemAghaee, N.: Building Bilingual Parallel Corpora Based on Wikipedia. In: 2010 Second International Conference on Computer Engineering and Applications (ICCEA). Volume 2. (2010) 264–268
22. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
23. Ács, J.: Pivot-based multilingual dictionary building using Wiktionary. In: Proceedings of LREC '14, Reykjavik, Iceland, ELRA (2014)

„Olcsó” morfológia

Novák Attila[1,2]

[1] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport ,
[2] Pázmány Péter Katolikus Egyetem Információtechnológiai és Bionikai Kar,
1083 Budapest, Práter utca 50/a
{novak.attila}@itk.ppke.hu

Kivonat A számítógépes morfológiai leírások egy része a lexikon mellett szabálykomponenst is tartalmaz. Ez utóbbi biztosítja egyrészt a morfológiai leírás konzisztenciáját, másrészt megkönnyíti a morfológia új lexikai elemekkel való bővítését. Azonban egy ilyen típusú leírás elkészítése komoly erőfeszítést és különféle kompetenciákat igényel. A legtöbb szabadon elérhető morfológiai leírás viszont nem tartalmaz szabályokat. Ezek általában egy alaktani szótáron alapulnak, és a szavak lemmája és esetleg ettől eltérő töve mellett valamilyen a szó ragozási paradigmáját leíró információt tartalmaznak, gyakran valamiféle paradigmaazonosító címke formájában. Ezt esetleg még egyéb lexikai–szintaktikai–szemantikai információ egészítheti ki. Az ebben a cikkben bemutatott kutatás célja egy olyan algoritmus kidolgozása volt, amely lehetővé teszi, hogy a szabályalapú morfológiákhoz hasonlóan egyszerű módon lehessen az ilyen szótáralapú morfológiai leírásokba is új lexikai tételeket felvenni. A felügyelt tanításon alapuló algoritmus a szótárból hiányzó szavak helyes ragozási paradigmáját próbálja meg megjósolni a leghosszabb illeszkedő végződések és lexikai gyakorisági adatok felhasználásával. Az algoritmust orosz nyelvű adatokon mutatjuk be és értékeljük ki.

Keywords: morfológia, ragozási paradigma azonosítása, orosz

1. Bevezetés

A morfológiai elemzés a legtöbb természetesnyelv-feldolgozó rendszer fontos alapfeladata, amelyre sok más feldolgozási szint épül. Az információ-visszakeresési és szövegindexelési algoritmusok többsége is alkalmaz valamiféle morfológiai feldolgozást, mert a szöveget alkotó szavak lemmájának azonosítására szükség van a valóban használható kereséshez. Az utóbbi feladatok esetében ugyanakkor általában nincs szükség arra a morfoszintaktikai információra, amely az adott szóalak paradigmabeli helyét azonosítja, és amely a teljes körű morfológiai elemzés esetén a lemma és a szófaj mellett az elemzés részét képezi.

Az igényesen kidolgozott számítógépes morfológiákat általában olyan formalizmus használatával készítik el, amely a szavak morfológiai viselkedésének valamiféle szabályalapú leírását felhasználva minimalizálja az egyes lexikai tételekről a lexikonba felveendő információ mennyiségét. Ez egyrészt megkönnyíti az új lexikai tételek helyes felvételét a lexikonba, ugyanakkor lehetővé teszi, hogy a morfológia készítője teljesen ellenőrzése alatt tarthassa az általa létrehozott nyelvi

erőforrás minőségét. A szabályalapú morfológiai nyelvtanok létrehozása ugyanakkor többféle kompetenciát igényel: ismerni kell a formalizmust, az adott nyelv morfológiáját, helyesírását, morfofonológiáját, és kiterjedt lexikai ismeretekre van szükség. Sok számítógépes morfológiai adatbázis ugyanakkor nem tartalmaz külön szabálykomponenst. Ezeket az adatbázisokat általában valamilyen ragozási szótárban szereplő információ konverziójával hozzák létre. A szavak lemmája (és esetleg ettől eltérő töve) mellett valamilyen a szó ragozási paradigmáját leíró információt tartalmaznak (gyakran valamiféle paradigmaazonosító címke formájában), ezt esetleg még valamiféle egyéb lexikai–szintaktikai–szemantikai információval kiegészítve. Szabályok híján azonban az ilyen erőforrások új szavakkal való kiegészítése nem olyan egyszerű, mint a szabályalapú morfológiák bővítése. A gépi tanulás alkalmazása azonban lehetővé teheti, hogy a más morfológiákban a szabálykomponensben leírt tudást magából az adatbázisból kinyerve azt új szavak ragozási paradigmájának azonosításához használjuk. Módszerünk a tő különböző hosszúságú végződéseit és egyéb lexikai jellemzőit használja jellemzőkként a megfelelő ragozási paradigma kiválasztásához. Általában a leghosszabb illeszkedő végződésre leginkább jellemző morfológiai viselkedést veszi a legnagyobb súllyal figyelembe. Működését egy nyílt forráskódú orosz morfológiai lexikonon mutatjuk be és értékeljük ki.

Az automatikus paradigmaazonosítás lehetőségét a következő feladat megoldásával kapcsolatban vizsgáltuk meg. Egy szótárprogramot kellett képessé tennünk arra, hogy egy konkrét orosz-angol szótár szóanyagának összes ragozott alakját felismerje és helyesen lemmatizálja. A program a megszorításalapú morfológiai modellt használó Humor morfológiai elemzőt [1] használja a többi nyelv kezeléséhez, ezért az oroszhoz is ilyen elemzőt kellett készítenünk. Ahelyett azonban, hogy a semmiből hoztunk volna létre egy új orosz morfológiai adatbázist, a www.aot.ru címről letölthető LGPL-licenszű orosz morfológia [2] adaptálása mellett döntöttünk. Az erőforrást a Humor elemző által megkövetelt formátumra konvertáltuk. Ezután a szókincsét ki kellett egészítenünk a szótár azon a szavaival, amelyek az eredeti morfológiából hiányoztak.

Cikkünk felépítése a következő: a kapcsolódó munkák áttekintése után a 3. részben bemutatjuk a tanító- és tesztanyagként használt adatbázist. Ezt követően leírjuk azokat a jellemzőket, amelyeket az orosz esetében a szavak ragozási paradigmájának megjósolásához használtunk, majd részletesen bemutatjuk a tővégződéseket használó modellt és a paradigmajelöltek rangsorolását végző algoritmust. Végül a 7. részben kiértékeljük a rendszer teljesítményét, és áttekintjük a rendszer által elkövetett osztályozási hibák típusait.

2. Kapcsolódó munkák

A ragozási paradigmák automatikus azonosításával számos kutatás foglalkozott. Néhány tanulmány keretében a ragozási paradigmákat teljesen automatikusan nyers korpuszból próbálták megtanulni. A szóalakokat automatikusan csoportosították (clustering), és az így létrejött csoportokat elemezték [3,4,5]. További felügyelet nélküli morfológiatanuló rendszereket ír le Wicentowski [6], Hammar-

ström [7] és Goldsmith [8]. Az utóbbi munkában az azonos toldalékhalmazok szignatúráknak nevezett struktúrákba szerveződnek, amelyek ragozási paradigmákat reprezentálnak. A felügyelet nélküli módszerek teljesítménye mindazonáltal messze elmarad a felügyelt tanítást alkalmazóké mögött, dolgozzanak azok akár lexikai adatbázisok, akár elemzett korpuszok alapján.

Egy másik megközelítést alkalmazó kutatók ugyancsak használnak nyers korpuszokat. Ezekben a munkákban az adott szó számára megtippelt lehetséges ragozási paradigmák elemeit szóalak-gyakorisági adatokkal vetik össze az adott paradigmajelölt érvényességének ellenőrzéséhez [9,10]. Amennyiben az adott paradigma által jósolt szóalakok nem fordulnak elő a korpuszban, a paradigmajelöltet érvénytelenként elveti az algoritmus. Hasonló rendszert ír le Lindén [11], amely lexikai jellemzőkre és korpuszgyakorisági adatokra is támaszkodik a ragozási viselkedés analógiás meghatározásához. SVM-alapú osztályozót tanítanak be a Šnajder [12] által leírt alaki és gyakorisági jellemzőket egyaránt használó rendszerben annak a döntésnek a meghozatalára, hogy egy lehetséges paradigmajelölt elfogadható-e vagy sem.

Megközelítésünk a legtöbb itt leírt korábbi kutatástól különbözik abban, hogy kizárólag egy morfológiai lexikont használunk a ragozási paradigma automatikus megállapítását végző rendszerünk betanításához. Célunk a lemma és néhány egyszerű szótárakban is szereplő lexikai tulajdonság alapján a legvalószínűbb ragozási paradigma meghatározása. A szótárból származó információk alapján megbízhatóbban meg tudjuk jósolni a szótárba felveendő új szavak ragozási mintázatát, mint ha csak nyers korpuszadatok állnának rendelkezésünkre és mind a lemmát, mind az egyéb lexikai tulajdonságokat (pl. a szófajt), valamint a ragozási paradigmát is kizárólag ezek alapján kellene megjósolnunk.

3. A tanító- és a tesztanyag

Az itt leírt kísérletekhez az www.aot.ru [2] webhelyről letölthető LGPL-licenszű nyílt forráskódú morfológiai lexikont használtuk. A lexikon alapszókincse Zaliznyák ragozási szótárán alapul [13]. 174 785 lexikai tételt tartalmaz, amelyek mindegyike 2 767 ragozási paradigma valamelyikébe van besorolva. A paradigmaazonosító algoritmus kiértékeléséhez egy Serge Sharoff által készített orosz lemmagyakorisági adatbázist is használtunk.[3]

A morfológiai lexikont a lexikai tételeket a korpuszbeli lemmagyakoriságuk alapján csoportosítva háromféleképpen osztottuk tanító- és tesztanyagra. Az egyik csoportba a viszonylag ritka szavak kerültek. Ebbe a csoportba azok a lemmák kerültek, amelyek gyakorisága elérte a 8-at, de nem haladta meg a 10-et az internetes korpuszban (3970 szó). A második csoportba közepesen gyakori szavak kerültek, olyanok, amelyeknek lemmagyakorisága nem haladta meg a 100-at (36917 szó). A harmadik csoportba a nagyon gyakori szavak kerültek, 1000-nél nem kisebb gyakorisággal (9633 szó). Teszthalmazként ezeket a szóhalmazokat használtuk, tanítóhalmazként minden esetben az adott teszthalmaz teljes lexikonra vett komplementuma szolgált.

[3] http://corpus.leeds.ac.uk/frqc/internet-ru.num

4. Az orosz szavak paradigmatikus viselkedését befolyásoló tényezők

Amikor orosz szavak ragozási paradigmájának megjósolására teszünk kísérletet, bizonyos grammatikai jegyek ismeretére szükség van ahhoz, hogy helyesen tippelhessünk. A lemma és a szófaj egyértelműen ilyen jellemzők, bár a szófaj melléknevek és igék esetében általában igen jól megjósolható a lemma alakja alapján. Mindazonáltal ezeket a jellemzőket ismertnek tekintettük, hiszen minden szótárban szerepelnek.

A főnevek esetében emellett számos más lexikai/szemantikai jellemző is szerepet játszik annak meghatározásában, hogy milyen morfoszintaktikai jegyegyüttesek fordulnak elő egyáltalán az adott szó ragozási paradigmájában. Ilyen jegy a grammatikai nem, a megszámlálhatóság és az élőség. Emellett vannak ragozhatatlan főnevek. Ezek közül a jegyek közül a grammatikai nem minden szótárban szerepel. Emellett általában a ragozhatatlan főneveket is megjelölik. Bizonyos absztrakt, kollektív és anyagnév jelentésű szavaknak nincsenek többes számú alakjai. Másrészt csak többes számú alakokkal rendelkező főnevek is vannak. Az utóbbiak egy része az alakjuk alapján felismerhető: a lemmájuk tipikus a többes számú alakokra jellemző végződést visel.

Az élőség olyan formában befolyásolja a ragozási paradigmát, amely az adott főnév által felvett lehetséges alakok halmazára nincs hatással. A szó élő vagy élettelen mivoltától függően azonban a szó alakjai közül különbözőek esnek egybe. Az élők esetében a tárgyeset a birtokos esettel esik egybe (többes számban minden nemben, egyes számban csak hímnemben), a nem élők esetében a tárgyesetű alak az alanyesetűvel azonos. Ez a különbség élő–élettelen homonimapárok esetében is érvényesül. Ezt a jelenséget mutatjuk be a 1. ábrán a *ёж* sün: állat', és cseh sün: tankelhárító akadály' szavak ragozási paradigmájának összevetésével. Ugyanakkor, mivel az élőség jegy, bár az aot lexikonban szerepel, más szótárakban azonban általában nem jelölik,[4] ezért mi sem használtuk.

Hasonlóképp az igék esetében az, hogy a morfoszintaktikai jegyegyüttesek mely kombinációi érvényesek, az ige aspektusától és tranzitivitásától, illetve visszaható voltától függ. Például a nem tranzitív igéknek nincsenek passzív melléknévi igenévi alakjai, a befejezett aspektusú igéknek nincsenek jelen idejű melléknévi igenévi alakjai, és a folyamatos aspektusú igék legnagyobb részének nincsenek múlt idejű (különösképp passzív) melléknévi igenévi alakjai. Emellett az, hogy milyen határozói igenévi alakjai vannak egy igének, szintén az aspektuson, illetve egyéb idioszinkratikus lexikai tulajdonságokon múlik. Ezért ezeket a tulajdonságokat ismerni kell, és ezek az információk valóban szerepelnek a szótárakban.

A melléknév-ragozási paradigma defektivitásai, pl. a rövid predikatív alakok és a szintetikus közép- és felsőfokú alakok megléte különböző szemantikai és

[4] A szótár használójának a megadott jelentés alapján kell azonosítania ezt a tulajdonságot, ami általában sikerül is neki néhány szokatlan esetet kivéve: pl. a *мертвец* 'hulla' szó grammatikai szempontból élő (az ugyancsak 'hulla' jelentésű *труп* ugyanakkor élettelen).

ёж[num:Sg.cas:Nom]	ёж[num:Sg.cas:Nom]
ежа[num:Sg.cas:Gen]	ежа[num:Sg.cas:Gen]
ежу[num:Sg.cas:Dat]	ежу[num:Sg.cas:Dat]
ежа[num:Sg.cas:Acc]	ёж[num:Sg.cas:Acc]
ежом[num:Sg.cas:Ins]	ежом[num:Sg.cas:Ins]
еже[num:Sg.cas:Prp]	еже[num:Sg.cas:Prp]
ежи[num:Pl.cas:Nom]	ежи[num:Pl.cas:Nom]
ежей[num:Pl.cas:Gen]	ежей[num:Pl.cas:Gen]
ежам[num:Pl.cas:Dat]	ежам[num:Pl.cas:Dat]
ежей[num:Pl.cas:Acc]	ежи[num:Pl.cas:Acc]
ежами[num:Pl.cas:Ins]	ежами[num:Pl.cas:Ins]
ежах[num:Pl.cas:Prp]	ежах[num:Pl.cas:Prp]
(a) ёж[N.gnd:Mas.ani:**Ani**][:8];	(b) ёж[N.gnd:Mas.ani:**Ina**][:9];

1. ábra. A *ёж* sün' szó élő (a) és élettelen (b) jelentésű változatának ragozási paradigmája.

más látszólag idioszinkratikus jegyektől függenek. Például a relációs melléknéveknek általában nincsenek ilyen alakjai. Ezek a tulajdonságok ugyanakkor nem szerepelnek explicit módon az aot lexikonban, és általában a hagyományos szótárakban sem tüntetik fel őket, ezért a melléknevek esetében mi sem használtunk semmilyen lexikai jegyet a szófajon kívül.

Az adott lexikai tételhez ragozási paradigmát rendelő algoritmusunk számára tehát a lemma mellett a fent megadott lexikai tulajdonságokat (szófaj, nem, igeaspektus stb.) is hozzáférhetővé tettük. Ugyanakkor azok az információk, amelyek sem a hagyományos szótárakban nem szerepelnek, sem a szó alakjából nem jósolhatóak meg, nem szerepeltek az algoritmusunk számára ismert jellemzők között. Ilyenek voltak többek között, hogy egy főnév élő vagy élettelen dolgot jelent-e, hogy egy adott melléknévnek bizonyos alakjai léteznek-e, illetve hogy az adott szó ragozási paradigmájában bizonyos hangsúlyingadozások, helyesírási vagy egyéb rendhagyóságok szerepelnek-e.

A fent említett lexikai jegyek mellett algoritmusunk az adott lemma n karakter hosszú végződéseit használta jellemzőkként különböző n értékekre. A maximális n végződéshossz paramétere az algoritmusnak. Kísérleteinkben ezt a paramétert 10-re állítottunk. A végződések és a lexikai jegyek által hordozott információ ábrázolásához toldalékmodellt építettünk a lexikon tanítóanyagként használt része alapján. Hogy ez a modell hogyan épül fel, azt a 2. ábra mutatja.

5. A végződésmodell létrehozása

A lemmavégződéseket és a lexikai tulajdonságokat a tanulóalgoritmus a 2. ábra jobb oszlopában bemutatott módon szófa adatszerkezetbe gyűjti. A lemmához az alábbi tulajdonságokat kódoló stringeket fűzi hozzá (jobbról balra haladva):

– A szögletes zárójelben álló címke két részből áll: a szófajból (és az alábbi példákban emellett a nemből), amelyet egy kötőjelet követően az adott szó ragozási paradigmájának az aot adatbázisban használt numerikus azonosítója követ. Ez az az információ, amelyet az algoritmusnak egy ismeretlen szóra meg kell tippelnie. A tanítóanyag feldolgozása során felépített szófa adatszerkezet végcsomópontjaiból kiinduló élek egy olyan adatszerkezetre mutatnak, ami az adott végződésű és lexikai tulajdonságokkal bíró szavakra a tanítóanyagban található [szófaj-paradigmaazonosító] párokból álló címkék eloszlását (relatív gyakoriságát) tartalmazza.
– A lemma végéhez attól egy függőleges vonallal elválasztva az adott lexikai elem ismert lexikai tulajdonságait kódoló stringet illesztünk.[5]
– Bizonyos paradigmákba tartozó lemmáknak egy adott végződést kell viselniük. Ilyen esetben a lemmában kettős kereszt jelöli annak a végződésnek a kezdetét, amely az adott paradigmaazonosító által jelölt paradigma alkalmazhatóságának a feltétele. Az adott paradigma biztosan nem jön szóba érvényes jelöltként olyan szavakra, amelyek nem az adott karaktersorra végződnek. Pl. minden a 1433-es számú paradigmához tartozó szó *ъё*-ra kell, hogy végződjön.

`мумиё[N.n.*.-];prd:25`	`мумиё|n*[N.n-25]`
`остриё[N.n.-];sfx:ё;prd:1709`	`остри#ё|n[N.n-1709]`
`бабьё[N.n-];sfx:ё;prd:210`	`бабь#ё|ns[N.n-210]`
`дубьё[N.n-];sfx:ё;prd:210`	`дубь#ё|ns[N.n-210]`
`свежевьё[N.n-];sfx:ё;prd:210`	`свежевь#ё|ns[N.n-210]`
`цевьё[N.n.-];sfx:ьё;prd:1433`	`цев#ьё|n[N.n-1433]`
`жнивьё[N.n];sfx:ё;prd:1103`	`жнивь#ё|n[N.n-1103]`
`суровьё[N.n];sfx:ё;prd:210`	`суровь#ё|ns[N.n-210]`
`мостовьё[N.n];sfx:ё;prd:210`	`мостовь#ё|ns[N.n-210]`

2. ábra. A végződésmodell egy részlete. A jobb oldali oszlopban álló elemek szerkezete: `lem#ma|lex-jegyek[Szófaj-ParadigmaID.]`, ahol a `ma` a lemma kötelező végződése minden olyan szó esetén, amely az adott numerikus `ParadigmaID` által azonosított paradigmába tartozik.

6. Rangsorolás

Az általunk használt toldalékszófa alapú rangsorolási algoritmust a Thorsten Brants TnT taggerében ([14]) a tanítóanyagban nem látott ismeretlen szavak lexikai valószínűségének becslésére használt toldalékguesser algoritmus ihlette. Azonban a Brants-féle algoritmus nem nyújtott kielégítő teljesítményt az általunk megoldani kívánt feladat esetében. Ezért addig módosítottuk az algoritmust, míg végül az eredetinél egyszerűbb, de annál lényegesen jobb eredménnyel

[5] n: semlegesnemű főnév, *: ragozhatatlan, s: csak egyes számú

működő modellt nem kaptunk. A paradigmabecslési algoritmus az adott szóhoz szóba jöhető paradigmák mindegyikéhez pontszámot rendel. Ez alapján a pontszám alapján rangsoroljuk a paradigma-jelölteket, és a legmagasabb pontszámút választjuk.

A pontszámot iteratív módon számítjuk ki az adott lemma ismert lexikai tulajdonságokkal bővített változatának végződésein végighaladva a legrövidebbtől a leghosszabb végződésig. Az iteratív pontszámszámítási algoritmus a 1. képlet szerint megadott módon módosítja az adott címkéhez tartozó pontszámot minden lépésben.

$$rank^{i+1}[tag] = sign \times len\_sfx \times rel\_freq + rank^{i}[tag] \tag{1}$$

ahol

a $sign$	negatív, ha a végződés rövidebb, mint az adott paradigma által megkövetelt minimális végződés
len_sfx	a végződés hossza a lexikai tulajdonságok nélkül
rel_freq	a tag címke relatív gyakorisága az adott végződésre
$rank^{i}[tag]$	-t elosztjuk len_sfx-szal, ha $len\_sfx > 1$ negáljuk, ha $sign > 0$ és $rank^{i}[tag] < 0$ mielőtt a $rank^{i+1}[tag]$-t kiszámítanánk

Ez a rangsorolási eljárás általában a leghosszabb illeszkedő végződéshez tartozó leggyakoribb paradigmát részesíti előnyben. A 3. ábrán néhány példát mutatunk be az algoritmus által rangsorolt paradigmajelöltekre.

```
рурба|f [N.f]    [N.f:50]#2.857270 [N.f:175]#0.756756 [N.f:48]#0.293840
                 [N.f:105]#0.175658 [N.f:88]#0.098045 [N.f:103]#0.051742
                 [N.f:396]#0.03995 [N.f:611]#0.039730 [N.f:69]#0.029693
                 [N.f:121]#0.021167

дурака|f [N.f]     [N.f:88]#4.466005 [N.f:15]#1.341181 [N.f:273]#0.904291
                   [N.f:36]#0.738748 [N.f:50]#0.467147  [N.f:16]#0.443249
                   [N.f:39]#0.300179 [N.f:105]#0.175658 [N.f:96]#0.155983
                   [N.f:103]#0.051742
```

3. ábra. A *рурба|f* és *дурака|f* inputra adott tíz legjobb paradigmajelölt. A jelölteket a pontszámuk alapján sorrendezi a rendszer. A pontszámot # jel választja el a javasolt címkétől.

7. Kiértékelés

A rangsoroló algoritmust a 3 részben leírt tesztanyagokon értékeltük ki. Ezeket a következőképpen jelöltük: ritka szavak (LT10), közepes gyakoriságú szavak (LT100), és gyakori szavak (MT1000). Módszerünk teljesítményének kiértékeléséhez a szokásos kiértékelési metrikákat alkalmaztuk.*A nyertes jelölt pontossága* azt adja meg, hogy az esetek mekkora részében rangsorolta az algoritmus a helyes paradigmát legelőre. Ez azt mutatja meg, hogy a rendszer mennyire jól rendel automatikusan paradigmaazonosítót egy adott szóhoz. Emellett a 2.–9. helyre rangsorolt azonosítók pontosságát is megmértük. A *fedés* azoknak a szavaknak az aránya, ahol a helyes paradigma benne van az első tíz helyre rangsorolt azonosítók halmazában. Lindén [11] megfontolásai alapján a hagyományos értelemben alkalmazott pontosság helyett a *maximális fedés melletti átlagos pontosság* mértékét használtuk. Ennek számítási módja $1/(1+n)$ minden szóra, ahol n a helyes paradigma rangja a javaslatok között. Így a rangsoroló algoritmus minősége is mérhető. Mivel lehetséges olyan alkalmazás, ahol a paradigmaazonosítók automatikus meghatározása az emberi besorolást segítő alkalmazás csupán, ezért ez a metrika mutatja a zaj mértékét, amit az emberi validálás során ki kell szűrni. A fenti két metrika alapján meghatároztuk továbbá az *f-mértéket*, ami a pontosság és fedés harmonikus közepe.

Az algoritmus hatékonyságának meghatározásához két baseline rendszert állítottunk össze. Az első Brants toldalékguesser modelljét használja ([14]) a leghosszabban illeszkedő végződések helyett. Ebben a modellben szerepel egy θ tényező, amit a végződések alapján meghatározott címkevalószínűségek becslésének simításához használ. A θ tényező értékét a címkék valószínűségeloszlásának szórásával megegyező értékre állítja be. Először meghatározza az összes toldalék valószínűségi eloszlását a tanítóhalmaz alapján, majd a 2. képlet alapján szukcesszív approximációval simítja a modellt.

$$P(t|l_{n-i+1},...l_n) = \frac{\hat{P}(t|l_{n-i+1},...l_n) + \theta_i P(t|l_{n-i},...l_n)}{1+\theta_i} \tag{2}$$

minden $i = m...0$-ra, amelynek kezdeti értéke $P(t) = \hat{P}$, ahol

$\hat{P}$ a maximum likelihood becslések a lexikonbeli gyakoriság alapján

θ_i súlyok a címkék tanítóhalmazbeli feltétel nélküli maximum likelihood valószínűségének szórása minden i-re

A másik alkalmazott baseline rendszer minden szóhoz annak szófajcímkéje alapján a leggyakoribb paradigmaazonosítót rendeli hozzá. A két baseline rendszer teljesítményét hasonlítottuk össze a saját rendszerünkkel. Az összehasonítás eredménye a 1. táblázatban látható. Ahogy várható volt, a második baseline (amely egyszerűen a leggyakoribb paradigmát választja) nagyon alacsony pontosságot ért el. Saját módszerünk azonban az első baseline algoritmusnál is lényegesen jobban teljesít. Az utóbbi kettő közötti teljesítménykülönbséget az

magyarázza, hogy Brants modellje a feltétel nélküli, illetve a rövidebb végződések eloszlásának nagyobb súlyt ad, mint a hosszabbaknak. Ezzel szemben a hosszabb végződések alapján működő rendszerünk éppen fordítva működik, és így jobban figyelembe veszi az adott szóosztály viselkedését.

1. táblázat. A nyertes jelölt pontossága a paradigmaazonosítók esetén a leghosszabb toldalékillesztés módszere, Brants modellje és a leggyakoribb paradigma hozzárendelése esetén.

	Leghosszabb toldalék	Brants modellje	Leggyakoribb paradigma
LT10	0.9166	0.6191	0.3464
LT100	0.9039	0.6062	0.3403
MT1000	0.7679	0.5372	0.3291

A paradigmákat és szintaktikai tulajdonságokat meghatározó címkék a szavak nagyon nagy felbontású osztályozását határozzák meg. Vannak olyan tulajdonságok, amik két paradigmát ugyan megkülönböztetnek egymástól, azonban a szavak ragozása szempontjából nem relevánsak. Például a nem ragozható határozószók számos alcsoportra oszlanak, de mindegyiknek csak egy alakja van. Ezek közül a tulajdonságok közül ráadásul a legtöbb nem is megjósolható. Sok esetben csupán a különböző hangsúlyingadozások tesznek különbséget két paradigma között, ami szintén nincs hatással a szavak leírt alakjára, azonban ilyenkor is más-más a helyes paradigmaazonosító. Végül, vannak olyan paradigmabeli különbségek is, amik az általunk megcélzott szótár-kiegészítési feladat szempontjából irrelevánsak, mert nincsenek hatással a paradigmában szereplő szóalakok halmazára. Ilyen például az azonos tövű élő és élettelen főnevek esete. Ezért, hogy a rendszerünk teljesítményét az eredetileg meghatározott tövesítési feladat szempontjából is kiértékelhessük, meghatároztuk a paradigmák ekvivalenciaosztályait. Ebben az esetben az automatikusan meghatározott paradigmaazonosítót helyesnek tekintettük, ha meghatározott paradigma által generált szóalakok halmaza megegyezett a helyes paradigma által meghatározott szóalakok halmazával. A 2 767 különböző közül a 921 nem egyedi paradigma 283 ekvivalenciaosztályba volt összevonható. A 2. táblázatban láthatóak az így mérhető eredmények, ahol a teljes' és 'equip' oszlopok a teljes paradigmaazonosító-egyezést megkövetelve, illetve az azonos ekvivalenciaosztályba tartozó paradigmák meg nem különböztetésével kapott eredmények. Megjegyzendő, hogy az 'equip' oszlopokban szereplő értékek összege nem 1, hiszen számos olyan eset fordul elő, ahol kettő vagy több olyan paradigma is szerepel az első helyekre soroltak között, amelyek a helyes paradigmával azonos szóalakhalmazt generál az adott lexikai tételhez.

Ahogy a számokból is látszik, a rendszerünk a ritka szavak esetében (LT10) teljesít legjobban, míg a gyakori szavak esetén mérhettük a legalacsonyabb teljesítményt (MT1000). Ez nem meglepő, hiszen a rendhagyó szavak a gyakoribbak

2. táblázat. A teljes címkeegyezés és az ekvivalenciaosztályok alapján elért eredmények

	LT10		LT100		MT1000	
	teljes	equip	teljes	equip	teljes	equip
#1	0.8924	**0.9274**	0.8750	**0.9174**	0.7416	**0.8087**
#2	0.0614	0.2322	0.0685	0.2278	0.0684	0.2371
#3	0.0168	0.2090	0.0223	0.2201	0.0314	0.2435
#4	0.0057	0.1518	0.0078	0.1452	0.0168	0.1900
#5	0.0035	0.1692	0.0037	0.1723	0.0090	0.2165
#6	0.0015	0.1884	0.0019	0.1683	0.0083	0.1697
#7	0.0000	0.1871	0.0012	0.1836	0.0032	0.1562
#8	0.0005	0.1400	0.0011	0.1496	0.0043	0.1418
#9	0.0010	0.1095	0.0007	0.1573	0.0017	0.1078
pontosság	0.9329	0.9538	0.92195	0.9481	0.8067	0.8550
fedés	0.9841	0.9876	0.9832	0.9875	0.8872	0.9158
f-mértéke	0.9578	**0.9704**	0.9516	0.9674	0.8450	0.8843

között fordulnak leginkább elő, míg a ritka szavak viselkedése ritkán rendhagyó, tehát jobban megjósolható. Nem meglepő, hogy a tanítóanyagban nem szereplő névmások vagy rendhagyó igék paradigmájának helyes meghatározása nem sikerül olyan jól. Ezen túl, mivel a célunk egy meglévő morfológiai lexikon kiegészítése, az ilyen lexikonok pedig a leggyakoribb szavakat eleve tartalmazzák, ezért ebben a feladatban éppen a ritka szavak helyes besorolása a fontosabb cél.

Szintén látszik az eredményekből, hogy hasonló fedésértékek mellett a pontosság és a nyertes jelölt pontossága szignifikánsan magasabb lett, amikor az ekvivalenciaosztályokat összevontuk. Az algoritmus tehát jól használható olyan erőforrások kiegészítésére, amelyeket teljes morfológiai elemzést nem igénylő, például információkinyerési vagy szótári kereséssel kapcsolatos feladatok megoldására használunk.

A 3. táblázatban a nyertes jelöltek pontossága látható szófajonkénti bontásban: az összes szóra, főnevekre, igékre és melléknevekre. Ebben az esetben a teljes paradigmacímkének való megfelelés helyett csak a paradigmaazonosítót vettük figyelembe. Így például a `[N.n._nam:Org.--49]`, `[N.n.--49]` és `[N.n--49]` javaslatok azonosnak tekinthetők. Az igék és melléknevek pontos paradigmájának meghatározása nehezebbnek bizonyult, mint a főneveké. Ennek okát elsősorban a következő fejezetben részletesebben tárgyalt szemantikai tényezők és hangsúlybeli különbségek között kereshetjük.

8. Hibaelemzés

A leggyakoribb tévesztések okai a leghosszabb végződést alkalmazó algoritmusunk esetén a ritka szavakra a következők: a rendszer nem tudja helyesen megjósolni, hogy

3. táblázat. A nyertes jelöltek pontossága minden szóra, illetve főnevekre, igékre és melléknevekre.

	ÖSSZES	FŐNÉV	IGE	MELLÉKNÉV
LT10	0.9166	0.9547	0.8158	0.8665
LT100	0.9039	0.9489	0.8114	0.8381
MT1000	0.7679	0.8594	0.6884	0.5991

- egy melléknévnek vannak-e szintetikus középfokú alakjai
- a *-ние* végű elvont főneveknek van-e *-нье* alakú alternatív alakjuk
- egy főnévnek van-e második birtokos alakja (amelyet a partitívuszi szerkezetekben használnak)
- bizonyos igeosztályokban a múlt idejű melléknévi igenevek hangsúlya hogyan alakul (ez *e*∼*ё* váltakozáshoz vezet ezekben az alakokban – ugyanakkor a hétköznapi helyesírási gyakorlat ezt általában nem jelöli, tehát valójában ez nem vezet hibához)
- egy melléknévnek vannak-e szintetikus felsőfokú alakjai
- bizonyos melléknevek rövid és középfokú alakjainak hangsúlya hogyan alakul (ez szintén *e*∼*ё* váltakozáshoz vezet ezekben az alakokban)
- egy nem ragozódó főnév értelmezhető-e többes számúként
- egy folyamatos igének vannak-e múlt idejű melléknévi igenévi alakjai
- vannak-e a paradigmában szabad hangsúlyingadozások
- egy melléknévnek vannak-e rövid predikatív alakjai

A hangsúlyozással kapcsolatos és a szemantikából fakadó alakváltozatok hiányával járó esetek kivételével az algoritmus ritkán javasol helytelen paradigmákat. Az ismeretlen szavak esetén emberek is hasonló hibákat követnének el, különösen akkor, ha a szó jelentését sem ismerik. Továbbá rendszerünk az eredeti aot lexikonban szereplő inkonzisztenciákat is kimutatott, amelyek komolyabb orosz nyelvtudás híján is egyértelműen felismerhető hibák. Például, míg az *Кубаньэнерго* energiacég neve olyan megjelöléssel szerepel, hogy nincs többes száma, addig a hasonló *Сахалинэнерго* szónak nincs ez a tulajdonsága.

Az algoritmus által a gyakori szavak esetén vétett hibákat vizsgálva azok típusa hasonló. Mindazonáltal, ebben az adathalmazban a közép- és felsőfok, a második birtokos vagy második helyhatározói esetek helytelen meghatározása sokkal gyakoribb hibák, hiszen a gyakori szavak között sokkal nagyobb az ilyen „rendhagyó” alakok aránya.

Ugyanakkor a Brants-féle modellt használó algoritmus leggyakoribb hibái között olyan alapvető tévesztések szerepelnek, amiket kezdő orosz nyelvtanulók sem követnének el. Ez a rövid végződésekhez tartozó eloszlások túl nagy súllyal való figyelembevételéből ered, a hosszabbakkal szemben. Az első helyre rangsorolt paradigmajelölt gyakran egyáltalán nem alkalmazható az adott végződésű szavakra. A ritka szavakból álló tesztkorpuszon ennél a rendszernél a leggyakoribb

hiba például az, hogy *-ный* végű melléknevekre a *-кий* végűek paradigmáját javasolja alkalmazni.

9. Konklúzió

Jelen cikkünkben bemutattunk és kiértékeltünk egy toldalékszófa alapú felügyelt tanulási módszert alkalmazó algoritmust, ami ismeretlen szavak ragozási paradigmájának meghatározására alkalmas, azok lemmájának végződése és néhány lexikai tulajdonságuk alapján. A módszer a morfológiai szótárak alapján készített, és ezért szabálykomponenst nem tartalmazó számítógépes morfológiák lexikonának automatikus kiegészítésére használható. A módszer alkalmazhatóságát orosz nyelvre mutattuk be, azonban minimális adaptáció után az eszköz bármilyen más nyelvre alkalmazható, amihez rendelkezésre áll a megfelelő morfológiai erőforrás. Emellett éltünk azzal a feltételezéssel is, hogy a morfológia mellett létezik olyan elérhető szótár, amiben bizonyos lexikai tulajdonságok is megtalálhatóak, így ezeket a paradigmajavaslatok egyértelműsítése során figyelembe lehet venni. Módszerünk jól teljesít minden teszteset során, legjobban azonban a ritkán előforduló szavak esetén működik. Éppen ezek hiányoznak az eredeti lexikonból a legnagyobb valószínűséggel.

Az eredményekből az is kiviláglik, hogy a hosszabb szóvégződések előnyben részesítése a rövidebbekkel szemben lényegesen jobb teljesítményhez vezet. Ez akkor is világosan látszik, ha pusztán azt tekintjük, hogy milyen gyakran találja el az algoritmus pontosan a helyes paradigmát. Az elkövetett hibák elemzése azonban még inkább rávilágít a javasolt megoldás erősségeire. Míg a baseline toldalékguesser algoritmus az adott szóra egyáltalán nem alkalmazható paradigmákat javasol, mikor hibázik, addig az általunk bemutatott módszer csupán a szemantikai ismeret hiányából fakadó tévesztéseket követ el. Ezek azonban olyan hibák, amiket emberek ugyanígy elkövetnének.

Hivatkozások

1. Novák, A.: What is good Humor like? [Milyen a jó Humor?]. In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
2. Sokirko, A.V.: Morphological modules at the site www.aot.ru. (2004)
3. Nakov, P., Bonev, Y., Angelova, G., Gius, E., von Hahn, W.: Guessing morphological classes of unknown German nouns. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R., eds.: RANLP. Volume 260 of Current Issues in Linguistic Theory (CILT)., John Benjamins, Amsterdam/Philadelphia (2003) 347–356
4. Monson, C., Carbonell, J.G., Lavie, A., Levin, L.S.: Paramor: Finding paradigms across morphology. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D., eds.: CLEF. Volume 5152 of Lecture Notes in Computer Science., Springer (2007) 900–907
5. Dreyer, M., Eisner, J.: Discovering morphological paradigms from plain text using a dirichlet process mixture model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 616–627

6. Wicentowski, R.: Modeling and learning multilingual inflectional morphology in a minimally supervised framework. Technical report (2002)
7. Hammarström, H., Borin, L.: Unsupervised learning of morphology. Comput. Linguist. **37**(2) (2011) 309–350
8. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Comput. Linguist. **27**(2) (2001) 153–198
9. Forsberg, M., Hammarström, H., Ranta, A.: Morphological lexicon extraction from raw text data. In: Proceedings of the 5th International Conference on Advances in Natural Language Processing. FinTAL'06, Berlin, Heidelberg, Springer-Verlag (2006) 488–499
10. Oliver, A., Tadic, M.: Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: LREC, European Language Resources Association (2004)
11. Linden, K.: Entry generation by analogy – encoding new words for morphological lexicons. In: Journal Northern European Journal of Language Technology. (2009) 1–25
12. Šnajder, J.: Models for predicting the inflectional paradigm of Croatian words. In: Slovenščina 2.0. (2013) 1–34
13. Zaliznyak, A.A.: Russian grammatical dictionary – Inflection. Russkij Jazyk, Moskva (1980)
14. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA (2000)

IV. Beszédtechnológia

Kétszintű algoritmus spontán beszéd prozódiaalapú szegmentálására

Beke András[1], Markó Alexandra[2], Szaszák György[3], Váradi Viola[2]

[1] MTA Nyelvtudományi Intézet
[2] ELTE BTK Fonetikai Tanszék
[3] BME Távközlési és Médiainformatikai Tanszék
e-mail:andras.beke@mta.nytud.hu

Kivonat Cikkünkben egy kétlépcsős automatikus szupraszegmentális frázisszegmentálót mutatunk be spontán beszédre. Miután az olvasott beszédre kidolgozott eljárások spontán beszédre nem működnek megfelelően, illetve a felügyelet nélküli kontúrklaszterezés sem hozott kielégítő eredményt, részletesebben is áttekintjük a vonatkozó irodalmat, ami alapján a szünetekkel határolt intonációs frázist tekintjük modellezési alapegységnek. A második szinten beágyazott fonológiai frázisok detektálását vizsgáljuk alapfrekvencia, energiamenet és magánhangzóidőtartamok alapján. A fonológiai frázisok detektálása az akusztikai jellemzők időbeli változását megragadó, szimmetrikus Kullback–Leibler-távolság alapú metrikával képzett különbségjelére meghatározott adaptív küszöbértékkel történik. A fonológiai frázisok detektálási pontossága spontán beszédben 80% körül adódik, percepciós szempontú címkézéssel összevetve. A kétszintű detektálási feladatban az energiaszint és a szünetek az intonációs frázisokban, az alapfrekvencia pedig a fonológiai frázisokban tűnik meghatározónak. Az időtartamok hozzájárulása a frázishatárok észleléséhez – az olvasott beszédhez hasonlóan – spontán beszédben sem igazolható.

Kulcsszavak: prozódia, spontán beszéd, automatikus klaszterezés

1. Bevezetés

A prozódiai frázis és/vagy prozódiai egységek határainak detektálása jelentős és fontos kutatási kérdés. Az elmúlt évtizedek során kifejlesztettek néhány olyan rendszert, amely olvasott, illetve félspontán beszédben képes automatikusan jelölni prozódiai egységek határait [8]. A magyar nyelvű fejlesztésekről az MSZNY konferenciákon is rendszeresen beszámoltunk [26,21]. Spontán beszédben ezek a felügyelt gépi tanuláson alapuló eljárások sokkal kisebb hatékonysággal és pontossággal használhatók [1], emellett jelentős problémaként merül fel a modellezési alapegységnek a megválasztása.

1.1. A prozódiai alapegységről

A beszédprozódia tudományos vizsgálatának elengedhetetlen feltétele annak tisztázása, hogy mit tekintsünk a prozódia egységeinek. A szegmentális fonetikai

megközelítésben vitán felül áll a beszédhang mint alapegység (annak ellenére, hogy a hangátmenet-hanghatár problematikája természeténél fogva megoldatlan marad), a szupraszegmentális megközelítésben azonban az alapegységet illetően sem elnevezésében, sem terjedelmében/tartalmában/definíciójában nincs konszenzus a kutatók között. A spontán beszéd vizsgálatában alapvető problémát jelent a hangfolyam tagolása, annak a közlésegységnek a kijelölése, amely „méreténél" és információtartalmánál fogva alkalmas arra, hogy a beszéd prozódiai alapegységének tekintsük. A legelterjedtebb elképzelés mind a hazai, mind a nemzetközi szakirodalomban a hierarchikus felépítés [19,23,14,12]. Felülről lefelé haladva a következő szintek különíthetők el: megnyilatkozás, intonációs frázis, fonológiai frázis, fonológiai szó, láb, szótag. Felmerül a kérdés, hogy spontán beszédben is alkalmazható-e ez az elkülönítés.

A fentieknek megfelelően jellegzetes különbségeket mutatnak azok a szakirodalmi források, amelyek felolvasott szövegek vagy hipotetikus közlések kapcsán foglalkoznak a tagolással, és azok, amelyek spontán beszéd prozódiai elemzése alapján kísérlik meg megállapítani az egységhatárokat. Az előbbi csoportba tartozó tanulmányok egyértelmű megoldásokat kínálnak. Például Elekfi 1962-es tanulmányában [7] azt mondja, hogy „nagyobb értelmi és kritikai egységeket kell keresni, melyekből a mondat már viszonylag közvetlenül felépíthető. Alkalmas egységnek látszott erre a beszédütem", amely nagyjából a klitikumos egységnek felel meg. Bolla Kálmán a szupraszegmentális alapegység meghatározásakor [3] ezt írja: „A szegmentális szerkezet struktúraképző alapeleme a beszédhang, míg a szupraszegmentális hangszövet legkisebb szerkezeti építőblokkját szupraszegmentális hangszerkezetnek nevezzük", majd megjegyzi, hogy hasonló értelemben használják még az intonációs szerkezet kifejezést is. A bollai hangszerkezet ez alapján tehát analóg a más szerzők által intonációs frázisként tárgyalt alapegységgel. Alapegységnek tehát a szupraszegmentális hangszerkezetet (intonációs frázist) tekinti, de ugyanakkor bizonyos értelemben alapegységként kezeli az ezt hordozó beszédszakaszt is, amelynek mibenlétét, definícióját azonban nem adja meg. A beszédszakasz a fentiek alapján a szintagma és a megnyilatkozás közötti egységet jelentheti, tehát valószínűsíthetjük, hogy a wachai megnyilatkozásegységnek feleltethető meg. Bollánál azonban nem találunk olyan kritériumrendszert, amely alapján meghatározhatnánk az alapegység határait.

Olaszy Gábor rádiós (felolvasott) szövegműfajok – hírek, novella, mese, reklám – komplex akusztikai elemzését végezte el [20]. A legnagyobb szövegegység, amelyen a prozódiát vizsgálta, a mondat volt. Ezeken belül úgynevezett prozódiai egységeket (PrE) határozott meg, amelyeket a szünettartás alapján határozott meg (tehát a mondat eleje-vége, illetve a mondat belsejében tartott szünet jelölte ki a PrE-k határát). A PrE-ket az alapfrekvencia alapján bontotta tovább hangsúlyközi szakaszokra, ezeket intonációs frázisoknak (IF) nevezte.

Élőbeszéddel dolgozott többek között Wacha [27], aki a megnyilatkozást tekinti alapegységnek, s a következőképpen definiálja: „Megnyilatkozáson az élőszóbeli, a (spontán) beszélt nyelvi közlésegésznek (szövegnek, szövegegésznek, beszédműnek) azt a – pontosan többé-kevésbé – elkülöníthető kisebb részét/egységét értem, melyet az írott nyelvhasználatról szólva a mondat, szövegmondat

terminussal szokás megnevezni. A megnyilatkozás a beszélt nyelvnek-nyelvhasználatnak olyan mondat értékű része tehát, melynek határait (kezdetét és végét) utólag – az elhangzó szöveg lejegyzésekor (átírásakor) – állapítottuk meg és jelöltük meg írásjelekkel, figyelembe véve az írásbeliség alapján kialakult mondatfelfogást (konvenciót) is." A megnyilatkozások megnyilatkozásegységekből állnak, így ezek tekintendők alapegységeknek, ugyanakkor Wacha használja a fonemikus frázis kifejezést is „hangképzési szünettől hangképzési szünetig tartó, csöndekkel határolt beszédszakasz" értelemben. Wacha problematikusnak tartja a megnyilatkozáshatárok megállapítását, végül az alábbi öt tényező közül valamely kettőnek az együttes jelenléte esetén tekint befejezettnek egy megnyilatkozást: „1. akusztikus zár, illetőleg ennek hiánya: a közlés dallama úgynevezett pont-hanglejtéssel zárul-e vagy sem, azaz a beszéd dallama mélyre szálló hanglejtéssel jelzi-e a megnyilatkozás befejezését, vagy esetleg nyitva tartott dallammal a folytatást ígéri (a beszédét vagy a megnyilatkozását); [...] 2. grammatikai zár: a megnyilatkozás grammatikailag befejezettnek, esetleg kereknek tekinthető-e vagy sem; [...] 3. követi-e az intonációs vagy grammatikai zárat új megnyilatkozást (új megnyilatkozás kezdetét) jelző intonációs indítás; [...] 4. követi-e vagy sem a megnyilatkozást valamiféle kiegészítő hozzátoldás (pl. értelmező, valamilyen »mondatrész«); [...] 5. a megnyilatkozás-sorozatot megtöri, megszakítja-e szünet vagy sem?" Wacha szerint a megnyilatkozáshatár legbiztosabb jelzője az új megnyilatkozás kezdetét jelző intonáció megjelenése.

Németh T. Enikő [16] a szóbeli diskurzusok megnyilatkozáspéldányokra tagolásakor arra a következtetésre jut, hogy figyelembe kell venni a dallam- és szünetprozodémák mellett a mondattal való kapcsolatba hozhatóságot, valamint pragmatikai szempontokat is. Ez utóbbiak között tekintetbe veszi az interperszonális funkciót, a beszédtettsorozatok összetételét és a pragmatikai kötőszókat, valamint az attitudinális funkciót (és az azt mutató nyelvi eszközöket).

Varga László [24,25] meghatározása fonológiai szempontú, s mint ilyen, a fonológiai szabályrendszer működési hatóköréül szolgáló nyelvi egység mibenlétét határozza meg, de komplex megközelítése figyelembe veszi a spontán beszéd sajátosságait is. Varga szerint a magyar intonációs frázis minimáldefiníciója: olyan szótagsorozat, amely vagy egy a) függelékdallammal, vagy egy b) (előkészítő dallam +) karakterdallammal realizálódik [24]. Ugyanakkor az intonációs frázist olyan egységként is meghatározhatjuk, amely egynél több dallamhangsúlyt fog át, ezt nevezi Varga maximáldefiníciónak, amely szerint az intonációs frázis kétféle lehet: a) függelék típusú, amely csak egy függelékdallamot tartalmaz; illetve (b) hangsúlyos, amely legalább egy karakterdallamból áll (előtte állhat egy vagy több szünet nélküli féleső karakter, s ez(eke)t megelőzheti egy főhangsúlyt nem tartalmazó előkészítő dallam), és szünet követi [25]. Varga tehát egyszerre alkalmazza a minimális (Elekfihez hasonlóan) és a maximális (ahogy Wacha és Németh T. Varga ismeretében teszi) szekvenciákra tagolás elvét, s e kettő közül az utóbbit preferálja, mert az a szintaktikai szerkezetet is tükrözi, ugyanakkor elismeri, hogy bizonyos spontán közlések esetében csak a minimáldefiníció alkalmazható.

Levelt [15] a beszédprodukciós folyamatot középpontba állító munkájában ugyancsak intonációs frázisnak nevezi az alapegységet. Ez megfogalmazása szerint szünettől szünetig tart: a beszélő azzal, hogy (általában több mint 200 ms-os) szünetet tart, befejez egy intonációs frázist. Az intonációs frázis leírásában Levelt öt összetevőt mutat be. A szünettartás bizonyos mértékig a beszélő döntésén múlik, ha jól érthető kíván lenni a hallgató(k) számára, lassan és rövid, akusztikai kulcsokban gazdag intonációs frázisokban beszél. A másik fontos tényező tehát a beszédtempó, amely természetesen befolyásol(hat)ja az intonációs frázisok tartamát is. Az intonációs frázisok megközelítőleg azonos időtartamban valósulnak meg a spontán beszédben (izokrónia), amit az artikulációs program végrehajthatósága indokol (az artikulációs tároló mérete és/vagy a rendelkezésre álló levegő mennyisége). A szintaktikai és a szemantikai szerkezet, valamint végül a beszédtervezés működési feltételei, minősége is nagymértékben hatással vannak az intonációs frázisok egymásra következésére.

Az eddig részleteiben áttekintett elméleti háttérirodalom alapján spontán beszédben is az intonációs frázis kínálkozik egy lehetséges alapelemként a prozódia modellezésében. Jelen munkában az intonációs frázis domináns akusztikai markereként a szünetet jelöljük ki, és kísérletet teszünk az intonációs frázisba mélyebb szinten beágyazott fonológiai frázisok detektálására is, a modellezésben is megtartva ezt a kétszintű hierarchiát. A fonológiai frázisok detektálásában prozódiai-akusztikai jellemzők követésére koncentrálunk eseménydetekciós attítűddel.

1.2. Modellezési megfontolások

A korábbi kísérletekben alkalmazott felügyelt tanulási módszerekkel a gépi tanulást címkézett adatokon végeztük el, amelynek a végén kialakult az osztályozó vagy detektáló, amely képes előre jelezni a prozódiai határokat a prozódiai egységekből származó akusztikai-prozódiai jellemzők alapján. Ez a megközelítés azt is feltételezi, hogy a detektálni kívánt egységek jól definiáltak (határaik és típusaik), és a priori ismertek, amelyek pontosan jelölve vannak a tanító korpuszban.

Újabban jelentős figyelem összpontosul a prozódiai egységek felügyelet nélküli modellezésére [5], vagy a már meglévő, felügyelt tanításból származó prozódiai modellek felügyelet nélküli adaptálására [22]. Ezek a megközelítések akkor is hatékonynak bizonyulhatnak, ha a modellezendő alapegység kérdéses.

Jelen kutatásban arra összpontosítunk, hogy nem felügyelt megközelítésben vizsgáljuk a spontán beszéd prozódiaalapú tagolhatóságát, majd ezt összevetjük intonációs, illetve fonológiai frázisokra való címkézéssel.

2. Anyag és módszer

A kutatáshoz a spontán beszédmintákat a BEA (BEszélt nyelvi Adatbázis [11]) szolgáltatta. 8 beszélő spontán narratíváit választottunk ki, 4 férfi és 4 női beszélőtől. Az így nyert korpuszt 2 fonetikus szakember felcímkézte, ezt a címkézést

kizárólag a kiértékelésnél, referenciaként használjuk. A címkézést 3 szinten végezték: intonációs frázisra (IF), fonológiai frázisra (FF), illetve szószintű átiratra. A vizsgálatokhoz használt korpusz összesen 398 IF-t és 751 FF-t tartalmazott.

2.1. Szegmentálás intonációs frázisokra

A BEA-ban a beszédfordulók szegmentáltak, a beszédfordulók megnyilatkozásokra bontása azonban nem egyértelmű, ahogyan – mint arra korábban már utaltunk – a megnyilatkozás definíciója sem. A beszéd intonációs frázisokra tagolásának irodalmát áttekintve kiolvasható, hogy annyiban viszonylagos konszenzus mutatkozik a kérdéskört vizsgáló kutatásokban, hogy az IF-ok határait, illetve lehetséges határait szünetek, esetenként időtartambeli nyúlás, F0 és energia jelzik. E jellemzők közül a szünetek szerepét több további kutatás is kiemeli [13], magyarra [10] az IF-ok percepciójában.

A fenti megfontolások fényében az IF detektálását energia, spektrális súlypont (centroid) és alapfrekvencia (F0) jellemzők alapján végezzük [9]. A jellemzőket 50 ms ablakkal számítjuk, melyek közül az energia nyilván kisebb lesz szünetekben, a spektrális súlypont pedig robusztusabbá teszi a beszéd/nem-beszéd elválasztását: a magasabb spektrális súlypont általában a beszédszegmensekre jellemző. Az F0-t azért vonjuk be a vizsgálatba, hogy egyrészt robusztusabbá tegyük a beszéd/nem-beszéd detektálását, másrészt lehetőségünk legyen tipikus fráziszáró intonációs markerek követésére is. A jellemzőket normalizáljuk, majd küszöbértéket számolunk, amelynek meghatározására hisztogramanalízist végzünk: a hisztogramokban a leggyakoribb értékeket k-közép klaszterezéssel szeparáljuk, lényegében súlypontokat határozunk meg. Két klaszterközéppontot alapul véve (M_1 és M_2) a küszöbértéket (T) az alábbi összefüggéssel számíthatjuk:

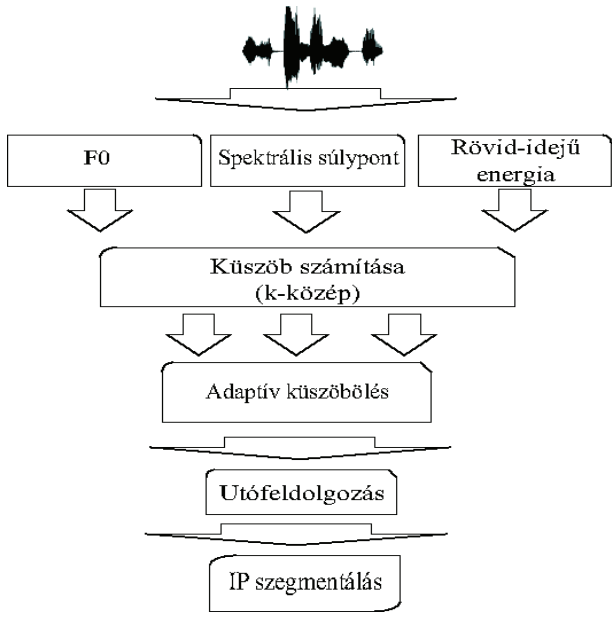
$$T = \frac{W * M_1 + M_2}{W + 1}, \tag{1}$$

ahol W szabadon választható paraméter (értéke kísérleteinkben $W = 0.5$). A küszöbértékkel a jelfolyamon csúcskeresést végzünk, végezetül pedig a csúcskeresés eredményeképpen kapott, keretekre értelmezett detekciós pontokat nagy ablakkal (250 ms) simítjuk, a folyamatot az 1. ábrán illusztráltuk.

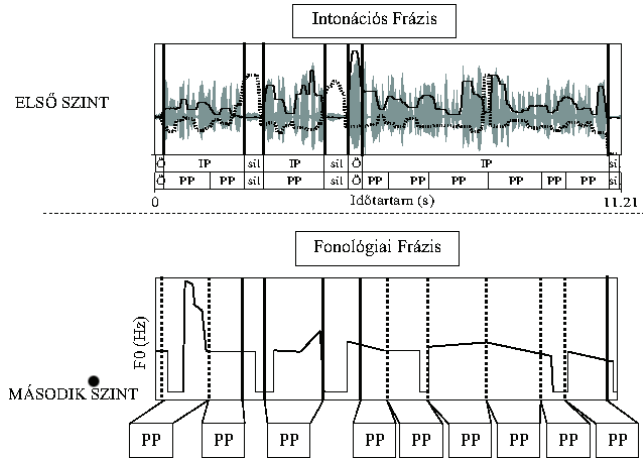
2.2. Szegmentálás fonológiai frázisokra

A FF-ra történő szegmentálás nagyobb kihívás a folyamatos, szünetekkel sem tagolt jelen. Mivel a FF-t az IF-ba beágazottként tekintjük, (2 ábra) a FF-szegmentálás bemenete az IF szegmentálás eredménye. Ily módon a hierarchiát is tükröző eljárást kapunk.

Jellemzőkinyerés. A FF-ok detektálásához a három alapvető prozódiai jellemzőt – alapfrekvencia (F0), energia és időtartam – vizsgáljuk szeparáltan és kombináltan spontán beszédben.



1. ábra. Az IF-szegmentálás vázlata



2. ábra. FF-ok azonosítása a prozódiai hierarchiában

A jellemzőkinyerést az olvasott beszédre is ismertetett módon végezzük [21] F0-ra és energiára, beleértve a részleges extrapolációt is a zöngétlen szakaszokon (kivéve, ha a zöngétlen szakasz hosszabb, mint 150 ms, vagy a zöngétlen szakasz után az F0 értéke eléri a korábbi F0-érték 110%-át).

Az időtartamokat magánhangzók hosszára és magánhangzók közötti távolságra automatikusan számítjuk, e két adatból lényegében a szótaghosszt is megkaphatjuk. Az időtartammérés alapja egy HMM-GMM alapú beszédhang-osztályozó, amely alapvető beszédhang-kategóriákat illeszt a folytonos beszédjelen: magánhangzókat, nazálisokat és approximánsokat, felpattanó zárhangokat, affrikátákat és frikatívákat. A beszédhang-osztályozó front-endje MFCC-jellemzőket számít, első és második deriváltakkal (MFCC39). A back-enden a beszédhang osztálya, kezdő- és végidőpontja jelenik meg.

A kapott magánhangzóhosszakat beszélőnként normalizáljuk, majd időben folytonos (pontosabban 10 ms keretidejű diszkrét) kontúrt állítunk elő simítással, a továbbiakban ezt értjük a tempó alatt.

A beszédhang-osztályozót azért választottuk a tempó kinyerésében, hogy komplett beszédfelismerés végrehajtása nélkül, illetve átíratlan anyagon is működjön a jellemzőkinyerés. Jóllehet maga a beszédhang-osztályozó sem működik hibamentesen, konzekvens hibázása esetén még ki is emeli a fontos jellemzőket – például megnyilatkozás végi irreguláris, kisebb energiájú beszédrészleteken a magánhangzókat nyújtja a többi beszédhang rovására, ez a működési jellegzetesség jelen alkalmazásban még előnyös is lehet.

Az F0, energia és tempó jellemzők mindegyikére első és második deriváltakat is számítunk, majd 0 és 1 közé normalizálunk.

Szegmentálás szimmetrikus Kullback–Leibler-távolsággal. A Kullback–Leibler-távolság (KL) az egyik leggyakrabban használt algoritmus arra, hogy hogyan mérhető a különbözőség két eloszlás között [2]. A KL-távolságot számos területen alkalmazzák mint beszélődetektálás, beszédfelismerés, beszélőfelismerés vagy beszéd/nem beszéd detektálás, stb. Ezek mellett igen népszerű a KL-távolság algoritmus használata szegmentálási problémák megoldására is, mint a beszéd vagy a zene szegmentálásra. A jelen tanulmányban a KL-távolságot a FF-ok határainak meghatározására alkalmazzuk. Matthew és munkatársai [18] kimutatták, hogy a szimmetrikus Kullback–Leibler-távolság egy olyan hatékony távolságmérő eljárás, amely könnyen mérhetővé teszi statisztikailag a különbözőség mértékének kifejezését két beszédjel között. A matematikai hátterét a következőkben ismertetjük: legyen X és Y két random eloszlás, és KL pedig a különbözőség mértéke e két eloszlás között. A távolság $KL(X;Y)$ az X és az Y között a következőképpen számolható:

$$KL(X;Y) = E_X(log(\frac{P_X}{P_Y}), \tag{2}$$

ahol E_X jelöli az X valószínűségi sűrűség függvény várható értékét. Ha az eloszlások Gauss-eloszlással modellezhetők, akkor a fenti egyenlet a következőképpen

alakul:

$$KL(X;Y) = \frac{1}{2}tr[(\Sigma_X - \Sigma_Y)(\Sigma_Y^- 1 - \Sigma_X - 1)] + \\ + \frac{1}{2}tr[(\Sigma_Y^- 1 - \Sigma_X^- 1)(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T], \tag{3}$$

ahol Σ a kovariancia mátrixra, míg a μ a középérték vektorra utal az adott eloszlásban. A Kullback–Leibler-távolság ugyan nemnegatív, de nem valódi metrika, mivel nem szimmetrikus, azaz megkülönböztetheti a modellt és modellezett eloszlást. A KL aszimmetrikus távolságot szimmetrikussá lehet tenni a következő lépéssel:

$$KL2(X;Y) = KL(X;Y) + KL(Y;X). \tag{4}$$

Mint korábban írtuk, ha a két eloszlás Gauss-eloszlással közelíthető, akkor a szimmetrikussá tett formában is létezik $KL2$ szimmetrikus KL távolság.

Jelen munka során a $KL2$-távolságot a beszédjelben egymást követő részek között számoltuk, amely részek 4 keret hosszúságúnak felel meg, vagyis 40 ms időtartamúak. Az ablakhossz 1 keretnyi volt, ami 10 ms-os időtartam. Minden egymást követő beszédrész között számolt $KL2$ érték egy folytonos görbét adott, amely után a következő feladat az volt, hogy ebben a folytonos jelben megtaláljuk a csúcsokat, amelyek a jelen esetben azt jelezték, ahol a két beszédszegmens között a legnagyobb eltérés jelentkezett. A magas $KL2$ érték tehát azt feltételezi, hogy a két beszédszegmens között jelentős az eltérés, míg az alacsony $KL2$ érték az azonosságot feltételezi. A csúcsdetektálás szempontjából igen fontos, hogy milyen ablakhosszban keressük az adott csúcsot. Ezért különböző ablakhosszokat alkalmaztunk, amelyet 10 kerettől (100 ms) 40 keretig (400 ms) növeltük a $KL2$ folytonos görbén. A csúcsdetektálás szempontjából igen fontos feladat a küszöbérték megválasztása is, mivel ettől függ, hogy az adott $KL2$ értéket váltási pontnak fogadjuk el, vagy sem. Ennek megválasztására két adaptív küszöbölési technikát alkalmaztunk (thr_A and thr_B). Az első adaptív küszöbértéket úgy számoljuk, hogy az adott keretben található érték középértékét vesszük, majd megszorozzuk egy konstanssal:

$$thr_A = \alpha \frac{1}{2N_1} \sum(F), \tag{5}$$

ahol F a jellemzővektor, N_1 az ablak hossza, és α a konstans.

Annak érdekében azonban, hogy FF határait detektáljuk, az adott értéknek nagyobbnak kell lennie, mint thr_B, amelyet a következőképpen számolhatunk:

$$thr_B = \sigma_F + \beta \frac{1}{2N_1} \sum(F), \tag{6}$$

ahol σ_F az adott ablakhosszban lévő értékek átlagos eltérése, β az ablak hossza. Az első küszöbérték azt biztosítja, hogy az adott érték nagyobb, mint a környező területen számított értékek, amelyet egy rövid idejű ablakra számolandó. A második küszöbérték, amelyet egy hosszabb ablakra számolunk, azt

biztosítja, hogy a változás figyelembe veszi az általános tendenciákat az ablakon kívüli adatok változásának figyelembevételével. Az ablakok méretét 3 és 4 másodpercre állítottuk. Ezen küszöbölési technika használata biztosította, hogy a téves elfogadások száma csökkenjen, és csak a valóban magas $KL2$ értéket fogadja el, amelyek a FF-ok határait jelentették.

3. A rendszer kiértékelése

A jelen munka során ugyanazon kiértékelési eljárásokat használtunk, mint ahogyan azt más szegmentálási feladatokban szokás. Az egyik szokásos kiértékelési eljárás a Brandt által kidolgozott GLR módszer [4]. Ez a módszer három gyakori mutatót javasol, amelyek az automatikus szegmentálási teljesítményt mutatják. Az első a beszúrás (Insertion Ins), amely azt jelenti, hogy az automatikus szegmentációban extra határok (események) vannak a referencia annotáláshoz képest. A második mérőszám a törlés (Omission Oms), amely azt jelenti, hogy az automatikus címkesorból hiányoznak események a referencia címkesorhoz képest. A harmadik mérőszám a helyes detektálások száma (Accuracy Acc), amelyet úgy számolunk, hogy a helyesen felismert határok számából – ha az időbeli eltérés az automatikus címkesor és a referencia címkesor között egy előre definiált tolerancia időkeretet nem lép túl – kivonjuk a törlések és beszúrások számát, majd elosztjuk az összes szegmenshatár számával (All):

$$Acc = \frac{Corr - (Ins + Oms)}{All} \tag{7}$$

A helyes detektálások száma fogja legjobban jellemezni a rendszer működését. A rendszer kiértékelése azonban függ a tolerancia időkeret hosszától. A jelen kutatás során több tolerancia értéket vizsgáltunk 25 ms és 100 ms között.

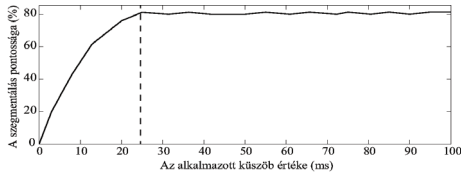
4. Eredmények

Elsőként az IF határok automatikus detektálásának eredményeit ismertetjük. Az eredmények azt mutatják, hogy az általunk kialakított rendszer, amelyben az akusztikai jellemző a beszédjel energiája, a spektrális súlypont és az F0 volt, az IF-ok 83,1%-át jelölte helyesen a spontán beszédben. A legtöbb IF szegmentálási hibát a kitöltött szünetek okozták. A második lépésben az IF-okon belül automatikusan jelöljük a FF-okat. A FF-ok határainak detektálására különböző akusztikai jellemzőket és azok kombinációit teszteltük. Összességében a három akusztikai jellemzővel öt kombinációt teszteltünk. Mindezek mellett változtattuk a $KL2$ érték csúcskeresésében alkalmazott időkeretek hosszát is (100 ms-tól 400 ms-ig). Az eredmények azt mutatták, hogy a helyes detektálások száma jelentős mértékben függ a $KL2$ értékek csúcskeresésében használt időkeret hosszától, illetve az akusztikai jellemzőtől is (1. táblázat). A legjobb eredményt akkor kaptuk, ha csak az alapfrekvenciát használtuk mint akusztikai jellemzőt, illetve a $KL2$ csúcsdetektáláskor használt időkeret hosszát 400 ms-ra állítottuk.

1. táblázat. *FF szegmentálás pontossága a csúcsdetektáláskor használt ablak hosszának függvényében.*

Ablakhossz (ms)	100	200	300	400
F0	68.71	75.83	76.33	**80.18**
Tempó	55.33	56.67	56.91	57.38
F0+energia	68.45	74.75	74.25	79.04
F0+tempó	68.79	73.15	72.33	78.22
F0+energia+tempó	68.86	72.60	71.84	77.05

A FF-ok határainak detektálásakor kiemelten vizsgáltuk a tempó jellemzőt a spontán beszédben, mivel korábbi szakirodalomban bizonyítottan nem volt alkalmazható jellemző a magyar felolvasásokban az FF-ek detektálására olvasott beszédben [21]. Az eredmények szinkronban az előző vizsgálatokkal azt mutatták, hogy ha a tempó jellemzőt önállóan alkalmazzuk a FF-ok határainak detektálására, akkor igen alacsony helyes detektálási arányt kapunk. Ugyanakkor, ha kombináljuk az F0-val, akkor az eredmények javulnak, de még akkor sem érik el azt a helyes detektálási arányt, mint amikor csak az F0 a bemenő jellemző. A helyes detektálási arány abban az esetben is csökken, ha a tempót az alapfrekvenciával és az energiával kombináljuk. Összességében a tempó törlése a jellemzőkészletből a helyes FF-határok detektálását növeli. Ezt az is alátámasztja, hogy a tempó sem az F0-val (R=-0,06), sem az energiával (R=-0,04) nem korrelál. A következő vizsgálatunk a teszteléskor alkalmazott tolerancia időtartamának hatását elemezte a szegmentálás eredményére. Ennek empirikus tesztelésére különböző toleranciatartományokat használtunk (3. ábra). Az eredmények szerint, ha a toleranciatartomány 25 ms, akkor a helyes szegmentálási arány 80,2%. A tolerancia időtartam további emelése már nem hoz eredményjavulást. Mindez azt jelenti, hogy a szegmentáló időben kifejezetten precíznek mondható (hasonló feladatban, igaz, más algoritmussal és más szempontú referenciaképzéssel, de olvasott beszédben 100 ms körül alakult az optimális toleranciatartomány [21]).



3. ábra. A szegmentálás eredménye a tolerancia időtartam függvényében

Az eredményekből arra következtethetünk, hogy az F0 az alapvető akusztikai jellemző a FF-ok detektálásban, az energia pedig a felsőbb szinten, az IF detektálásában fontos jellemző a spontán beszéd esetén.

5. Összegzés

A kutatásban spontán beszéd automatikus prozódiai szegmentálását vizsgáltuk hierarchikus struktúrában, szofisztikált csúcsdetektálási megközelítésben. Első lépésben az intonációs frázist két szünet közötti alapegységnek tekintettük, és k-közép klaszterezéssel detektáltuk. Az IF-okon belül második lépésben FF-ok izolálását kísérletük meg. Az alap prozódiai jellemzőkre (F0, energiaszint és magánhangzó-időtartamok) a korábbi jelértékekhez képesti szimmetrikus Kullback–Leibler-távolságokat számítottuk jeleltolással. A csúcsdetektálást ezzel távolságmetrikával képzett jeleken hajtottuk végre. A három alapjellemző, illetve kombinációik használatával nyert detektálási pontosságot összevetve a legnagyobb pontosságot FF-okra, 80,2%-ot egyedül az F0 használatával kaptuk. Az eredmények alapján az akusztikai markerek tekintetében az energiaszint az IF szintjéhez, az alapfrekvencia pedig a FF szintjéhez sorolható spontán beszédben. Olvasott beszédben a FF-ok detektálásában az F0 domináns, az energiaszint pedig járulékos szerepet játszott [21], spontán beszédben a kísérletek ezt nem mutatták. Az időtartam – legalábbis a magánhangzók hossza – az implementált megközelítésben a FF szintjén nem befolyásolta az eredményeket – ez egybevág az olvasott beszéd esetében tapasztaltakkal. Fontos különbség, hogy a FF-határok a spontán beszédben időben pontosabban detektálhatók, mint felolvasásban (a találati pontosságok felolvasásra ±100 ms, spontán beszédre mindössze ±25 ms időbeli pontosságnál vethetők össze), ennek oka az algoritmikus különbségeken túl a spontán beszéd gyakoribb megszakadása és nagyobb prozódiai dinamikája. Az eredményeket az is befolyásolja, hogy a kiértékelés spontán beszéd esetén percepciós alapú, míg felolvasásra szigorú szintaxis alapon címkézett referenciákkal összevetve történt.

Köszönetnyilvánítás

A kutatás az Országos Tudományos Kutatási Alapprogramok PD112598 számon, "Automatikus fonológiai frázis és prozódiai eseménydetektálás szintaktikai, szemantikai és pragmatikai információk közvetlen kinyerésére a beszédből" címmel támogatott projektje keretében készült.

Hivatkozások

1. Beke A., Szaszák Gy., Váradi V.: Automatic phrase segmentation and clustering in spontaneous speech. In: IEEE 4th International Conference on Cognitive Infocommunications: CogInfoCom 2013, Budapest (2013) 459-462
2. Couvreur, L. Boite, J-M.: Speaker tracking in broadcast audio material in the frame work of the THISL project (1999) 84–89
3. Bolla K.: Szupraszegmentális elemzések. Egyetemi Fonetikai Füzetek 7. ELTE Fonetikai Tanszék, Budapest (1992)
4. Jarifi, S., Pastor, D., Rosec, O.: Brandt's GLR method and refined HMM segmentation for TTS synthesis application (2005) 23–33

5. Chiang, C., Chen, S., Yu, H., Wang Y.: Unsupervised joint prosody labeling and modeling for mandarin speech. Journal of the Acoustical Society of America, Vol. 125, No. 2 (2009) 1164–1183
6. Cruttenden, A.: Intonation. Cambridge University Press (1997)
7. Elekfi L.: Vizsgálatok a hanglejtés megfigyelésének módjaihoz. Nyelvtudományi Értekezések 34. Akadémiai Kiadó, Budapest (1962)
8. Gallwitz F., H. Niemann, E. Nöth, Warnke, W.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36 (2002) 81–95
9. Giannakopoulos, T.: Study and application of acoustic information for the detection of harmful content, and fusion with visual information. Ph.D. dissertation, Dpt of Informatics and Telecommunications, University of Athens, Greece (2009)
10. Gósy, M.: Virtual sentences in spontaneous speech. In: Speech Research 2003, Budapest, Hungary (2003) 19–43
11. Gósy, M. „BEA - A multifunctional Hungarian spoken language database," PHONETICIAN 105-106, 2012 (2008) 50–61
12. Gussenhoven, C.: The phonology of tone and intonation. Cambridge University Press, Cambridge (2004)
13. Hansson, P.: Prosodic phrasing in spontaneous Swedish. Lund University (2003)
14. Hunyadi, L.: Hungarian sentence prosody and universal grammar. On the phonology – syntax interface. Metalinguistica 13. Peter Lang, Frankfurt/M., Berlin, Bern, Bruxelles, New York, Oxford, Wien (2002)
15. Levelt, W. J. M.: Speaking: From Intention to Articulation. A Bradford Book. The MIT Press, Cambridge, London (1989)
16. Németh T. E.: A szóbeli diskurzusok megnyilatkozáspéldányokra tagolása. Nyelvtudományi Értekezések 142. Akadémiai Kiadó, Budapest (1996)
17. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press (1967) 281–297
18. Siegler, M.A., Jain, U., Raj, B., Stern, R.M.: Automatic segmentation, classification, and clustering of broadcast news audio. In: Proc. of the Speech Recognition Workshop (1997) 97–99
19. Nespor, M., Vogel, I.: Prosodic phonology. Foris Publications, Dordrecht (1986)
20. Olaszy G.: Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella és a reklámok felolvasásában. Beszédkutatás 2005 (2005) 21–50
21. Szaszák Gy., Beke A.: Szintaktikai szerkezet automatikus feltérképezése a beszédjel prozódiai elemzése alapján. In: Tanács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2011 (2011) 178–189
22. Yildirim, S., Narayanan, S.: Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 1 (2009) 138–149
23. Varga L.: A hanglejtés. In: Kiefer F. (szerk.): Strukturális magyar nyelvtan 2. Fonológia. Akadémiai Kiadó, Budapest (1994) 468–549
24. Varga L.: The unit of the Hungarian intonation. In: Szathmári, I. (szerk.): Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös nominatae. Sectio Linguistica tomus XXIV. ELTE Eötvös Kiadó, Budapest (1999–2001) 5–13
25. Varga L.: Intonation and stress. Evidence from Hungarian. Palgrave Macmillan, Houndmills, Basingstoke (2002)
26. Vicsi K., Szaszák Gy.: Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján: II. rész: Statisztikai eljárás, finn-magyar nyelvű összehasonlító vizsgálat. In: Alexin Z., Csendes D. (szerk.): III. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2005 (2005) 360–370

27. Wacha I.: Élő nyelvi (spontán) szövegek megnyilatkozásainak (szintaktikai) vizsgálati szempontjaihoz (a gazdagréti kábeltelevízió élő nyelvi felvételei alapján). In: Kontra M. (szerk.): Beszélt nyelvi tanulmányok. Linguistica, Series A, Studia et Dissertationes 1. MTA Nyelvtudományi Intézet, Budapest (1988) 102–158

Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler–divergencia alapú klaszterezéssel

Grósz Tamás, Gosztolya Gábor, Tóth László

MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103.
{ groszt, ggabor, tothl } @ inf.u-szeged.hu

Kivonat Az elmúlt néhány év során a beszédfelismerésben a rejtett Markov modellek Gauss keverékmodelljeit (Gaussian Mixture Models, GMM) háttérbe szorították a mély neuronhálók (Deep Neural Networks, DNN). Ugyanakkor a neuronhálókra épülő felismerők számos olyan tanítási algoritmust megörököltek (változatlan formában vagy apróbb változtatásokkal), melyeket eredetileg HMM/GMM rendszerekhez fejlesztettek ki; ezek optimalitása az új környezetben egyáltalán nem garantált. Ilyen tanítási lépés a környezetfüggő fonémaállapot-halmaz meghatározása is, amire az általánosan elfogadott megoldás egy döntésifa-alapú algoritmus. Ez az eljárás arra törekszik, hogy az előálló állapotokhoz tartozó példák Gauss-görbékkel optimálisan modellezhetőek legyenek. Jelen cikkünkben egy alternatív eljárást vizsgálunk meg, mely a döntési fát egy Kullback-Leibler–divergencia alapú döntési kritériumra támaszkodva építi fel. Feltételezésünk szerint ez a kritérium alkalmasabb a neuronháló kimeneteinek leírására, mint a gaussos modellezés. A módszert korábban már sikeresen alkalmazták egy KL-HMM rendszerben, most pedig megmutatjuk, hogy egy HMM/DNN hibrid rendszerben is működőképes. Alkalmazásával 4%-os relatív hibacsökkenést értünk el egy nagyszótáras szófelismerési feladaton. [1]

Kulcsszavak: beszédfelismerés, környezetfüggő fonémaállapotok, mély neuronhálók, Kullback-Leibler divergencia

1. Bevezetés

Az utóbbi pár évben a hagyományos Gauss keverékmodelleket (Gaussian Mixture Models, GMMs) alkalmazó beszédfelismerő rendszerek helyét átvették a mély neuronhálókra (Deep Neural Networks, DNN) épülő HMM/DNN hibridek. A rejtett Markov-modellek (Hidden Markov Models, HMM) megjelenése

[1] Jelen kutatási eredmények megjelenését a „Telemedicina-fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken” című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatja. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

óta azonban elég sok eljárást fejlesztettek ki, melyeket a DNN-t használó keretrendszerek is átvettek, holott ezek az algoritmusok DNN-ek használata esetén nem feltétlenül működnek optimálisan. Talán a legismertebb ilyen a „flat start" indítás (melyben a hangfelvételből és annak átiratából iterálva találjuk meg az egyes beszédhangok helyét), valamint a környezetfüggő (context-dependent, CD) fonémamodellek kialakítása.

Habár a HMM/ANN hibrid modellekben sokáig csak környezetfüggetlen modelleket alkalmaztak (azaz egy-egy beszédhangot önmagában, az azt megelőző és követő fonémák ignorálásával modelleztek), mostanra nyilvánvalóvá vált, hogy a nagypontosságú beszédfelismeréshez hibrid modellek esetében is célszerű környezetfüggő (trifón) beszédhangmodelleket alkalmazni. Az összes trifónt külön modellezni azonban nem hatékony, érdemes ehelyett az egymáshoz valamilyen szempontból hasonlóakat összevontan kezelni.

Erre a feladatra már megjelenése óta Young [1] és Odell [2] döntésifa-alapú klaszterezési módszerét szokás alkalmazni. Ez az eljárás a faépítés során egyetlen normális eloszlással modellezi az egy állapothalmazhoz tartozó összes példát, és arra törekedve osztja ketté a halmazt, hogy a két nem átfedő részhalmaz külön-külön optimálisan legyen modellezhető. Ez egy igen gyors eljárás, azonban, bár kapcsolódása egy HMM/GMM alapon nyugvó rendszerhez nyilvánvaló, egy neuronháló-alapú beszédfelismerő rendszer esetén optimalitása több okból is megkérdőjelezhető.

Az egyik ilyen ok, hogy a GMM-alapú eljárások feltételezik, hogy a jellemzők kovarianciamátrixa diagonális, azaz dekorrelált jellemzőkészletet (pl. MFCC) várnak el. Ugyanakkor a HMM/DNN hibrid rendszerek általában jobban teljesítenek egyszerűbb jellemzőkön (pl. Mel szűrősorok). Mivel a hagyományos HMM/GMM rendszerek ilyen jellemzővektorokra nem taníthatóak hatékonyan, először egy HMM/GMM rendszert kell tanítanunk hagyományos jellemzőkön, ennek segítségével elkészíteni a környezetfüggő állapotok összevont halmazait és a fonémák keretszintű illesztését, majd eldobni a már leszámolt jellemzővektorokat. Ehelyett logikusabbnak tűnik az állapotok összevonását egy neuronháló kimenete alapján végezni (Senior et al. [3]). Ennek finomított változata az utolsó rejtett réteg (Bacchiani et al. [4]) értékeit használni, esetleg ezen réteg kimeneteit normális eloszlású valószínűségi eloszlásokká konvertálni (Zhang et al. [5]).

Bár a felsorolt kutatók a neuronháló kimenetét igyekeztek az eljáráshoz igazítani, maga az állapotok összevonására szolgáló algoritmus minden esetben változatlan maradt, csupán annak bemenete változott meg. Ugyanakkor jogosnak tűnő ellenvetés, hogy az eljárásnak olyan környezetfüggő állapotokat kellene különválasztania, melyek külön kezelése az adott beszédfelismerő rendszerben alkalmazott eljárás (GMM vs. DNN) számára kedvezőbb. Mivel egy GMM és egy DNN tanítása során alapvetően más jellegű döntési függvényt optimalizálunk, annak vizsgálata, hogy egy normális eloszlással hogyan tudjuk modellezni az egyes állapotokhoz tartozó példákat, akár teljesen független is lehet attól, hogy egy mély neuronháló hogyan tudja modellezni az adott osztályt. Akkor viszont a normális eloszláson alapuló döntési kritérium helyett érdemesebb lenne valamilyen másfajta építési kritériumot alkalmazni.

A közelmúltban Imseng et al. a döntésifa-alapú eljárás olyan változatát dolgozta ki, mely közvetlenül a neuronhálók kimenetét használja [6]. A korábban felsorolt művekkel ellentétben, melyek a neuronháló-kimeneteket feltételes osztályvalószínűséggé konvertálták és normális eloszlással modellezték, ez az eljárás kihasználja, hogy egy neuronháló kimenetvektora diszkrét valószínűségi eloszlás. Ezek különbözőségének mérésére kézenfekvő választás a Kullback-Leibler–divergencia [7], így az állapothalmazokat meghatározó eljárás döntési kritériumában érdemesebb ezt használni a normális eloszlásra épülő, hagyományos döntési függvény helyett. Imseng et al. sikerrel alkalmazta ezt az algoritmust Kullback-Leibler–divergenciára épülő rendszerükben (KL-HMM) [8].

Jelen cikkünkben ezt az eljárást egy HMM/DNN hibrid beszédfelismerő rendszerben értékeljük ki. A teszteket egy 28 órányi magyar nyelvű híradófelvételt tartalmazó adatbázison [9] végezzük; viszonyítási alapnak egy HMM/DNN hibrid rendszert veszünk, melynek környezetfüggő fonémamodell-halmazait a bevett GMM-alapú eljárással állítjuk elő.

2. Döntésifa-alapú modellösszevonás

A döntésifa-alapú fonémamodell-összevonási algoritmus húsz évvel ezelőtti bevezetése óta [1] a nagyszótáras beszédfelismerő rendszerek tanításának elhagyhatatlan részévé vált. Alapötlete, hogy egy (környezetfüggetlen) állapot összes előfordulását összevonja egy $\mathcal{S}$ halmazba, majd ezen halmaz lépésenkénti kettéosztásával egy döntési fát épít. Az algoritmus minden lépésben kiválaszt egyet az előre definiált kérdések közül annak alapján, hogy az így előálló két nem átfedő részhalmaz elemei a lehető legjobban különbözzenek egymástól. Ezt a különbözőséget egy valószínűség-alapú döntési kritérium méri. Ez az eljárás annyira sikeresnek bizonyult, hogy kisebb javításokat (pl. a kérdések automatikus előállítását [10]) leszámítva azóta is változatlan formában használják.

2.1. Valószínűség-alapú döntési kritérium

Odell [2] megfogalmazott egy maximum likelihood-alapú döntési kritériumot, és adott is egy hatékony algoritmust a kiszámítására, a szétválasztási kritériumot a következő képlettel becsülve:

$$L(\mathcal{S}) \simeq -\frac{1}{2}\big(\log[(2\pi)^K|\Sigma(\mathcal{S})|] + K\big)\sum_{s\in\mathcal{S}} N(s), \tag{1}$$

ahol $s \in \mathcal{S}$ jelöli az egyes állapotokat, $\Sigma(\mathcal{S})$ az $\mathcal{S}$-ba tartozó példák szórása, míg $N(s)$ az s állapothoz tartozó példák száma a tanítóhalmazban. Így azt a q kérdést kell választanunk a példák kettéválasztására, melyre a $\Delta L(q|\mathcal{S})$ valószínűségkülönbség maximális, ahol

$$\Delta L(q|\mathcal{S}) = \big(L(\mathcal{S}_y(q)) + L(\mathcal{S}_n(q))\big) + L(\mathcal{S}), \tag{2}$$

és $\mathcal{S}_y(q)$ és $\mathcal{S}_n(q)$ az $\mathcal{S}$ halmaz két nem átfedő részhalmaza a q kérdésre adott válasznak megfelelően. Látható, hogy a valószínűség-értékek nem függnek a tanítópéldáktól, csupán azok szórásától és az egyes állapotokhoz tartozó tanítópéldák (keretek) számától. Ez a feltevés tökéletesen illeszkedik egy GMM-alapú beszédfelismerő rendszerhez, ugyanakkor egy HMM/DNN hibridben valamely más döntési kritérium használata a mély neuronhálókhoz jobban illeszkedő állapothalmazhoz is vezethet.

2.2. Kullback-Leibler–divergencia alapú döntési kritérium

Ezt a kritériumot Imseng et al. vezette be [11], és sikeresen alkalmazták KL-HMM rendszerükben. A következőkben [6] és [8] alapján röviden ismertetjük az eljárást.

Habár a Kullback-Leibler–divergencia nem távolságfüggvény (például nem szimmetrikus), a szimmetrikus KL-divergenciára épülő költségfüggvény kiszámítására nincs zárt formula. Emiatt az aszimmetrikus KL-divergenciát fogjuk alkalmazni, mely két K-dimenziós posterior-vektorra (z_t és y_s) a következő alakot veszi fel [7]:

$$D_{KL}(y_s||z_t) = \sum_{k=1}^{K} y_s(k) \log \frac{y_s(k)}{z_t(k)}. \tag{3}$$

A KL-divergencia mindig nemnegatív, és pontosan akkor nulla, ha a két eloszlásvektor megegyezik. Így a faépítés során a likelihood maximalizálása helyett minimalizáljuk a KL-divergenciát:

$$D_{KL}(\mathcal{S}) = \sum_{s \in S} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log \frac{y_{\mathcal{S}}}{z_f(k)}, \tag{4}$$

ahol $\mathcal{S}$ állapotok egy halmaza, és $F(s)$ az s állapothoz tartozó tanítóminták halmaza. Az $\mathcal{S}$ halmazhoz tartozó $y_{\mathcal{S}}$ posterior valószínűségi vektor $\mathcal{S}$ elemeinek mértani közepeként számítható, azaz

$$y_{\mathcal{S}}(k) = \frac{\left(\prod_{s \in \mathcal{S}} \prod_{f \in F(s)} z_f(k)\right)^{\frac{1}{N(S)}}}{\sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k)}. \tag{5}$$

Néhány behelyettesítő és egyszerűsítő lépés után a következőt kapjuk [6]:

$$D_{KL}(\mathcal{S}) = -\sum_{s \in \mathcal{S}} N(s) \log \sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k), \tag{6}$$

tehát a $\mathcal{S}$ állapothalmazhoz tartozó KL-divergenciát kiszámíthatjuk az egyes állapotok y_s és $N(s)$ értékei alapján.

Egy $\mathcal{S}$ állapothalmaz kettéosztása során kézenfekvő azt a kérdést választani, amely maximalizálja a KL-divergencia különbségét ($\Delta D_{KL}(q|\mathcal{S})$):

$$\Delta D_{KL}(q|\mathcal{S}) = D_{KL}(\mathcal{S}) - \big(D_{KL}(\mathcal{S}_y(q)) + D_{KL}(\mathcal{S}_n(q))\big). \tag{7}$$

3. KL-alapú állapotösszevonás HMM/DNN hibrid rendszerekben

Viszonyítási alapként a hagyományos tanítási utat követtük: első lépésben környezetfüggő HMM/GMM fonémamodelleket tanítottunk, majd ezeket felhasználva kényszerített illesztéssel állítottuk elő a tanító címkéket a DNN számára. Ez a módszer MFCC jellemzőkészletet használ, megvalósításához a HTK [12] programcsomagot használtuk. A HMM/GMM környezetfüggő fonémamodellek tanítása során a hagyományos normáliseloszlás-alapú állapotösszevonást alkalmaztuk, majd miután megkaptuk a klaszterezett állapotokat, felhasználásukkal egy mély neuronhálót tanítottunk. Az így tanított DNN-t használtuk a dekódolás során akusztikus modellként, a HTK módosított Hdecode rutinja segítségével.

A KL-alapú klaszterező algoritmus bemenetként környezetfüggetlen állapotok posterior valószínűségeit várja. Ezen értékek előállításához egy környezetfüggetlen *segéd neuronhálót* használtunk (a keretszintű címkézést a fönti HMM/GMM rendszer szolgáltatta). Ezután alkalmaztuk a KL-alapú klaszterező algoritmust a segédháló kimenetére, és a környezetfüggő mély neuronhálót az így kapott összevont állapotokat címkeként használva tanítottuk be. A viszonyítási alapként szolgáló módszerhez hasonlóan itt is a klaszterezés után tanított mély hálót használtuk a felismerés során.

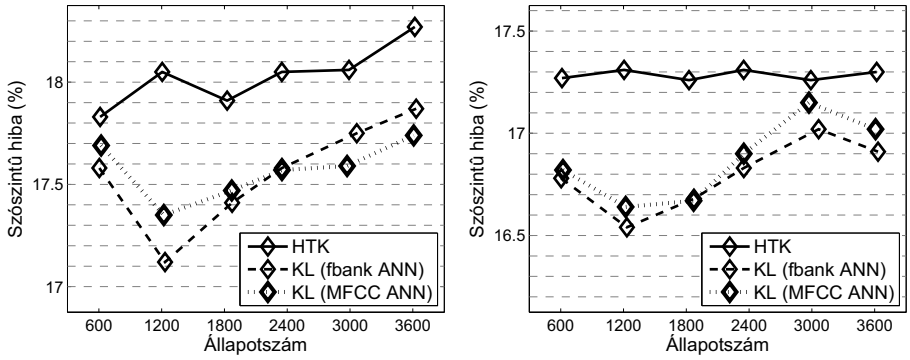
4. A kísérletek technikai jellemzői

A hibrid rendszerünk DNN komponenseként egy mély rectifier hálót [13] alkalmaztunk, amelynek fő előnye, hogy körülményes előtanítási módszerek nélkül, hagyományos backpropagation algoritmussal is hatékonyan tanítható [14]. Saját implementációnkat használtuk, amellyel a TIMIT adatbázison az általunk ismert legjobb eredményt, 16, 7%-os fonémaszintű hibát tudtunk elérni [15].

Az akusztikus modellezésre használt mély rectifier hálónk 5 rejtett rétegből állt, mindegyikben 1000 neuronnal, míg a kimeneti rétegben a softmax aktivációs függvényt alkalmaztuk. Bemenetként az ún. FBANK jellemzőkészletet használtuk [12], amely 40 mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból állt.

Kísérleteinket híradófelvételeken végeztük [9]. Az adatbázis összesen 28 órányi hangzóanyagot tartalmaz, melyet a szokásos felosztásban használtunk: 22 órányi anyag volt a betanítási rész, 2 órányi a fejlesztési halmaz, a maradék 4 órányi hanganyag pedig a tesztelésre szolgáló blokk. Az adatbázisban összesen 13 467 különböző trifón fordult elő, ami összesen 40 401 kiindulási fonémaállapotot eredményezett.

A segéd-neuronhálók inputjaként először, a HMM/GMM rendszerrel megegyezően, MFCC jellemzőkészletet használtunk, majd kipróbáltuk az FBANK jellemzőkészletet is. A viszonyítási alapként szolgáló módszer esetében eltérő jellemzőkészletet kellett használnunk a klaszterezéshez és az akusztikus modell tanításához (MFCC vs. FBANK), mivel az FBANK jellemzőkön tanított GMM-ek használhatatlan eredményt adtak volna. A KL-klaszterezés hátránya, hogy



1. ábra. Az elért szószintű hibaarányok az állapotok számának függvényében a fejlesztési (balra) és a teszthalmazon (jobbra)

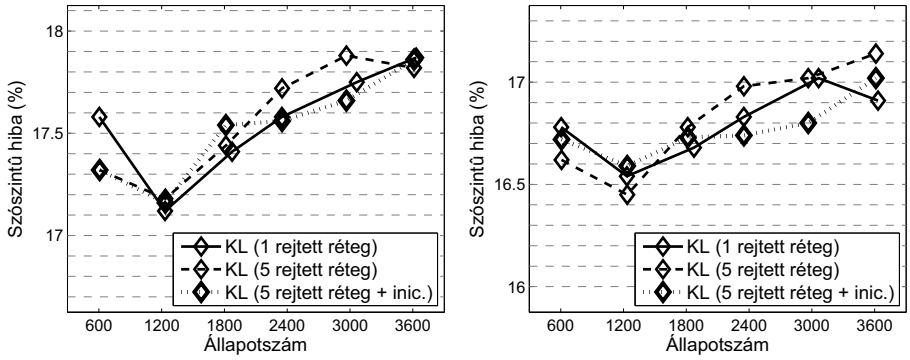
két neuronhálót kell tanítanunk; ennek csökkentése céljából kísérletet tettünk a segéd-neuronháló „újrahasznosíthatóságára" a második háló tanítása során. A klaszterezési eljárások küszöbértékeit úgy választottuk meg, hogy végül körülbelül 600, 1200, 1800, 2400, 3000 és 3600 összevont állapotot kapjunk.

5. Eredmények

Ahogy az az 1. ábrán megfigyelhető, a KL-divergencia–alapú klaszterezési algoritmus mindkét halmazon következetesen és szignifikánsan jobban teljesített, mint a hagyományos GMM/HMM alapú eljárás. A hagyományos módszer optimuma 600 környezetfüggő állapot körül van, habár a szószintű pontosságok minden kipróbált állapotszám esetén nagyon hasonlóan alakulnak. A KL-alapú algoritmus optimuma 1200 összevont állapotnál van: itt mintegy 4%-os relatív hibaarány-csökkenést hoz az alkalmazása a standard eljárás legjobb eredményéhez viszonyítva. A segéd-neuronháló két kipróbált változata közül a mel szűrősorokat használó bizonyult valamivel jobbnak (ezt a jellemzőkészletet használtuk a mély neuronháló tanításánál is), bár a különbség nem jelentős.

A KL-divergenciát használó klaszterezési eljárás alapvetően a segéd-neuronháló kimenete alapján dönt, így annak pontossága triviális módon meghatározza az állapothalmaz minőségét; ugyanakkor ennek mértéke egyáltalán nem nyilvánvaló, és mivel utána ezt a hálót eldobjuk, nem biztos, hogy megéri nagy pontosságú (és nagyméretű) segédhálót használni. Ennek kiderítésére további kísérleteket végeztünk: az eddigi egyrétegű háló helyett próbát tettünk egy mély (5 rejtett rétegű) neuronháló alkalmazásával is.

Egy másik lehetőség a segédháló felhasználása a végső DNN súlyainak inicializálásához, mely egyrészt csökkentheti a tanítási időt, másrészt pontosabb akusztikus modellhez vezethet. Természetesen ez csak akkor megvalósítható, ha mindkét neuronháló azonos számú neuront használ a rejtett rétegeiben, továbbá azonos jellemzőkészleten dolgozik; ugyanakkor korábban azt tapasztaltuk, hogy szűrősorok használatával nem kapunk rosszabb eredményeket, mint MFCC-vel,



2. ábra. Az elért szószintű hibaarányok az állapotok számának függvényében a fejlesztési (balra) és a teszthalmazon (jobbra)

így ez nem nagy megkötés. Emiatt a továbbiakban minden segédhálót FBANK szűrősorokra tanítottuk. Ezt az inicializálási stratégiát kipróbáltuk az egy és az öt rejtett réteggel rendelkező segédhálók alkalmazása során is.

Az eredmények (ld. 2. ábra és 1. táblázat) alapján annak, hogy a segédháló egy vagy öt rejtett réteget tartalmaz, nincs különösebb jelentősége. Hasonlóképpen, bár az akusztikus mély neuronháló inicializálása a segédháló megfelelő súlyainak felhasználásával 2-3 iterációval csökkentette a tanítás időigényét, a betanított háló pontossága enyhén romlott.

1. táblázat. A különböző állapotklaszterezési eljárások használatával elért szószintű hibaarányok

Klaszterezési eljárás	Szószintű hiba (%)	
	Fejl. halmaz	Teszthalmaz
KL (MFCC ANN)	17.35%	16.64%
KL (fbank ANN)	17.12%	16.54%
KL (fbank ANN) + ANN inic.	17.38%	16.79%
KL (fbank ANN, 5 rejtett réteg)	17.18%	16.45%
KL (fbank ANN, 5 rejtett réteg) + ANN inic.	17.16%	16.59%
GMM/HMM	17.83%	17.26%

Az eredményeket összegezve kijelenthetjük, hogy a Kullback-Leibler–divergenciára épülő döntési kritérium használata a környezetfüggő állapothalmazok kialakítása során szignifikánsan csökkentette a felismerés szószintű hibáját. Következtetésünk megerősítése érdekében a közeljövőben tervezzük, hogy a módszert más adatbázisokon is kiértékeljük.

6. Konklúzió

Jelen cikkben egy olyan eljárás hatékonyságát vizsgáltuk meg, amely a környezetfüggő fonémamodellek halmazát egy Kullback-Leibler–divergenciára épülő kritérium használatával határozza meg. Azt feltételeztük, hogy ez a kritérium alkalmasabb a neuronháló kimeneteinek leírására, mint a gaussos modellezés. Az algoritmus környezetfüggetlen állapotok valószínűség-eloszlását várja bemenetként; erre a célra egy segéd neuronhálót tanítottunk. A módszert egy nagyszótáras beszédfelismerési feladaton teszteltük, és használatával szignifikánsan tudtuk csökkenteni a szószintű hibát a hagyományos normális eloszlásra alapuló döntési kritériumhoz képest, több mint 4%-os relatív hibaarány-csökkenést elérve.

Hivatkozások

1. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. In: Proceedings of HLT. (1994) 307–312
2. Odell, J.: The Use of Context in Large Vocabulary Speech Recognition. PhD thesis, University of Cambridge (1995)
3. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN training. In: Proceedings of ICASSP. (2014)
4. Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In: Proceedings of ICASSP. (2014) 230–234
5. Zhang, C., Woodland, P.: Standalone training of context-dependent Deep Neural Network acoustic models. In: Proceedings of ICASSP. (2014) 5597–5601
6. Imseng, D., Dines, J.: Decision tree clustering for KL-HMM. Technical Report Idiap-Com-01-2012, Idiap Research Institute (2012)
7. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Statist. **22**(1) (1951) 79–86
8. Imseng, D., Dines, J., Motlicek, P., Garner, P., Bourlard, H.: Comparing different acoustic modeling techniques for multilingual boosting. In: Proceedings of Interspeech. (2012)
9. Grósz, T., Kovács, G., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvü beszédfelismerésben. In: Proceedings of MSZNY. (2014) 3–13
10. Beulen, K., Ney, H.: Automatic question generation for decision tree based state tying. In: Proceedings of ICASSP. (1998) 805–808
11. Razavi, M., Rasipuram, R., Magimai-Doss, M.: On modeling context-dependent clustered states: Comparing HMM/GMM, hybrid HMM/ANN and KL-HMM approaches. In: Proceedings of ICASSP. (2014)
12. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)
13. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier networks. In: Proceedings of AISTATS. (2011) 315–323
14. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proceedings of ICASSP. (2013) 6985–6989
15. Tóth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. In: Proceedings of ICASSP. (2014) 190–194

Hibajavítási idő csökkentése magyar nyelvű diktálórendszerben

Szabó Lili[1], Tarján Balázs[1], Mihajlik Péter[1,2], Fegyó Tibor[1,3]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék,
{lili,tarjanb}@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.,
mihajlik@thinktech.hu
[3] SpeechTex Kft.
tfegyo@speechtex.com

Kivonat A gépi beszédfelismerésen alapuló diktálórendszerek természetes velejárója a felismerési hiba, melyet tipikusan a szófelismerési hibaaránnyal jellemzünk. A felhasználó számára azonban nem a klasszikus szóhibaarány a meghatározó mint használhatósági metrika, hanem sokkal inkább a hibajavítási idő. Cikkünkben azt vizsgáljuk, hogy valós, magyar nyelvű, relatíve kötött témájú (EU-s joganyagok) diktálási feladat esetén milyen faktorok befolyásolják elsődlegesen a hibajavítási időt, azt hogyan lehet csökkenteni. A saját rendszerünket összevetettük a piacon ingyenesen hozzáférhető magyar nyelvű diktálórendszerekkel. Megmutattuk, hogy a beszédfelismerési modellek feladatra szabásán túl az írásjelek, speciális rövidítések és egyéb szövegszerkesztési parancsok diktálhatóságának lehetővé tétele jelentősen csökkentheti a diktálásra fordított időt és energiát, így növelve a felhasználói elégedettséget.

1. Bevezetés

Cikkünk témája egy korábbi nagyszótáras, magyar nyelvre fejlesztett, folyamatos beszédfelismerőn alapuló e-mail diktálórendszer [1] továbbfejlesztése, valamint a diktálási feladat felhasználó számára történő megkönnyítése. A legfejlettebb technológiát alkalmazó, beszélőre adaptált, célfeladatra tanított automatikus beszédfelismerő rendszerek angol nyelvre, 90-95%-os felismerési pontossággal működnek. Egy, a beszédfelismerésen alapuló diktálást napi szinten, professzionális célokra használó felhasználó számára akár ennél magasabb felismerési pontosság is elégedetlenséghez vezethet, tekintve, hogy egy felismerési hiba észlelése és javítása akár 15-30 másodpercet is igénybe vehet [2]. Mivel a diktálási feladat természetes velejárója a felismerési hibák javítása, ezért a hibajavítás támogatása elengedhetetlen komponense egy diktálórendszernek. Jelen tanulmányunkban hibajavításon mind a szöveg utólagos formázását, mind a felismerő által ejtett hibák utólagos korrigálását értjük. Feltételezésünk szerint a felismerési kimenetben szereplő hibák észlelését nagymértékben könnyíti a szöveg jólformázottsága. Egy felismerési kimenetben alapértelmezésben nem szerepelnek sem

írásjelek, sem nagy kezdőbetűs alakok, illetve szövegszakaszokat határoló sortörések, és szükséges ezek helyreállítása ahhoz, hogy a felhasználó egy formázott szöveget tudjon létrehozni.

Természetesen már önmagában a felismerési hibák csökkentése is nagy szerepet játszhat a hibajavítási idő csökkentésében. Ennek, és a piaci termékekkel való összevethetőségnek az érdekében egy kötött témájú (*closed domain*), jogitörvénykezési diktálásra optimalizált rendszert építettünk, ami azért előnyös, mert a felismerési hiba csökkenését eredményezi, ezáltal lehetőséget teremtve a hibajavítási folyamatot támogató új módszerek kipróbálására. Magyar nyelvre ingyenesen hozzáférhető, beszédfelismerésen alapuló diktálórendszer a *Google* webalapú[4], és a *Nuance* okostelefonra készített szövegbeviteli alkalmazásába[5] integrált beszédfelismerési szolgáltatásaként érhető el. Tanulmányunkban ezeket vetjük össze rendszerünkkel, leginkább a hibajavítási folyamat szempontjából.

A 2. fejezetben beszédfelismerési kimenetek szerkesztésével foglalkozó legújabb kutatásokat tekintjük át. A 3. fejezet a diktálórendszerek kiértékelését végző metrikákat, és a SpeechTex rendszer felépítését írja le. A 4. fejezet a diktálórendszerek összehasonlításához végrehajtott kísérletek felépítését ismerteti. Az 5. és 6. fejezetek a kísérletek eredményeit és tanulmányunkban levont következtetéseket tartalmazzák.

2. Irodalmi áttekintés

A mondathatárok automatikus detektálása, nagy kezdőbetű- és írásjel-visszaállítás a beszédfelismerési kimenetben egy gyakran kutatott téma a szakirodalomban [3]. Gépi tanulásos algoritmusokat használó módszerekkel ezen feladatokon elért pontosság 30-50% körül mozog [4,3]. A feladatot nehezítik a felismerési kimenetben különböző arányban előforduló hibák. Magyar nyelvre hasonló megoldást [5]-ben találhatunk. Ebben a kutatásban különböző modalitású tagmondattípusokra HMM modelleket építettek, amelyek segítségével a tagmondatfajtákat felismerték. A felismeréshez felhasználták egy, a tagmondatok egymás utáni sorrendjét figyelembe vevő szöveg szintű prozódiai modellt is. 6 tagmondattípus és egy szünetmodell, mellett 50%-os pontosságot értek el (úgy, hogy a helyesen felismert írásjelek aránya 70% körül mozgott). Az automatikus írásjelezésre alternatív megoldás, az írásjelek diktálhatóságának lehetővé tétele, melyet jelen fejlesztés során alkalmaztunk, igen magas pontossággal működik.

A hibajavítási folyamat támogatása ehhez szorosan kapcsolódó téma. Az eddigi kutatások azt tükrözik, hogy csupán a felismerési pontosság javítása nem elegendő, hiszen hibák mindig lesznek a felismerési kimenetben, ezért magát a hibajavítási folyamatot kell meggyorsítani és megkönnyíteni a felhasználó számára. Az egyik megközelítés a jelenség kezelésére a beszédfelismerő adaptálása a felhasználók javításait visszacsatolva: szótárban nem szereplő szavak hozzáadása a nyelvi modellhez, a nyelvi modell újrasúlyozása, valamint kiejtési alternatívák

[4] https://www.google.com/intl/en/chrome/demos/speech.html
[5] http://www.swype.com

generálása [6]. A másik módszer a felismerési kimenet utógondozása. A hagyományos helyesírás-ellenőrzéstől abban lényegesen különbözik ez a feladat, hogy a beszédfelismerési kimenetben kizárólag olyan szavak fordulhatnak elő, amelyek szerepeltek a nyelvi modell tanításához használt korpuszban. Ebből következik, hogy a kimenetben előforduló hibák „valódi szavas" (*real word*) hibák, melyek kezelése egy, a kontextust is figyelembe vevő eljárást igényel. Számos módszer született már a probléma megoldására; a hagyományosnak tekinthető *noisy channel* [8] modellben egy mondat összes szavától adott szerkesztési távolságra lévő szavak potenciálisan helyes szavak, a javítás a legvalószínűbb szósorozat kiválasztásával történik, tetszőleges n-gram alapon.

Az automatikus beszédfelismerés felhasználó-központú, illetve a hibajavítási folyamat szempontjából történő kiértékelése egy aránylag kevés figyelmet kapó terület, [7] tartalmaz egy körültekintő tanulmányt különböző diktálási tapasztalattal rendelkező felhasználók újonnan elsajátított hibajavítási szokásaival, a hangsúly itt inkább az egyének közötti változatosságon van, mintsem a hibajavítási folyamat kvantitatív értékelésén.

3. Módszer

3.1. Kiértékelés

Szóhibaarány. A szóhibaarány (*word error rate - WER*) az automatikus beszédfelismerésen alapuló rendszerek egyik legnépszerűbb kiértékelési módszere. A szavak szintjén méri a hibás behelyettesítések (S), törlések (D) és beillesztések (I) számát a felismerési kimenetben, és ezek arányát a referenciában előforduló szavak számához (N) képest.

$$\text{Szóhibaarány} = \frac{S + D + I}{N} \tag{1}$$

Új metrikák. Egy diktálórendszer teljeskörű kiértékelése csak úgy lehetséges, ha az a felhasználó nézőpontját is figyelembe veszi. Ennek érdekében három új mérőszámot/metrikát vezettünk be, amik a hibajavítási folyamatot hivatottak kiértékelni:

1. **Szerkesztési Idő**: mennyi időt vesz igénybe a felhasználónak a felismerési kimenetben a hibákat megtalálni és javítani, valamint a szöveget jólformázott alakra hozni.
2. **Sikerességi Ráta**: milyen mértékben sikerül a felhasználónak a felismerési kimenetet a kívánt/eredeti szöveg alakjára hozni. A szóhibaarányhoz hasonlóan a behelyettesítéses (S), törléses (D) és beillesztéses (I) hibák karakterszintű számolása a már szerkesztett kimenetben (lényegében Levenshtein-távolság az eredeti szövegtől) elosztva/normalizálva az eredeti szövegben előforduló szavak számával (N), az írásjeleket is figyelembe véve.
3. **Gépelési Idő**: mennyi időt vesz igénybe ugyanazon eredeti szöveg legépelése másodpercben.

3.2. Rendszerek

A 1. táblázat összefoglalja a három diktálórendszer jellemzőit. Megjegyzendő, hogy míg a Google rendszer magyar nyelvre nem rendelkezik sem az írásjelek diktálhatóságának, sem az írásjelek automatikus helyreállításának funkciójával, a Nuance rendszer az automatikus helyreállítás jeleit mutatja, noha ez becslésünk szerint az esetek kevesebb, mint 10%-ban fordul elő.

1. táblázat. A három rendszer funkcióinak összehasonlítása.

Funkciók	**Google**	**Nuance**	**SpeechTex**
Írásjelek	–	automatikus	diktálva
Nagybetűsítés	–	automatikus	diktálva
Ütemezés	valós idejű	késleltetett	valós idejű
Domain	nyitott	nyitott	törvénykezés

3.3. Korpusz és normalizálás

A nyelvi modell építéséhez használt korpusz [9] egy többnyelvű adatbázis az európai parlamenti ülések leiratainak hivatalos fordításaiból, amiből a magyar ún. *fordítási egységeket* használtuk fel. A korpusz adatait a 2. táblázat foglalja össze.

2. táblázat. DGT-TM korpusz adatok

	normalizálás előtt		**normalizálás után**	
korpusz rész	**token**	**type**	**token**	**type**
tanító	35.3 M	1.3 M	43.3 M	645 K
dev	129 K	27 K	145 K	18 K
eval	94 K	21 K	114 K	15 K

A normalizálás első lépése a mondathatárok helyreállítása volt. Ez a mondatvégi pont és a rövidítések, valamint a mondatkezdő nagybetűs szó és a tulajdonnevek egymástól való elválasztásával történt, a korpuszban előforduló gyakoriságok alapján. Az ezt követő *tokenizálás* során a következő *token* típusokat különböztettük meg: szavak, tulajdonnevek, mozaikszavak, rövidítések, URL-ek, email-címek, számok, dátumok, jogi jelölések, speciális szimbólumok, egyéb nem nyelvi elemek. Ezek átalakítása szöveges alakra, valamint a beszélt formára nem alakítható egyéb nem nyelvi elemek eltávolítása reguláris kifejezések segítségével történt.

Duplikációk detektálása és eltávolítása a nyelvi modell simításához fontos, hogy megtörténjen, mert a simítási eljárás során használt *counts-of-counts*-ok

eloszlását zavarja, ha páros számú *count*-ok kiugróan magasabbak, mint a páratlanok, és a *count*-ok nem egyenletesen csökkenő eloszlást követnek.

3.4. Nyelvi modell

A nyelvi modell módosított Kneser-Ney simítás használatával készült az SRI Language Modeling Toolkit (SRILM) [11] segítségével. A létrehozott trigram (3-gram), szóalapú modellekben entrópiaalapú metszést egyetlen esetben sem alkalmaztuk.

3.5. Akusztikus modell

Az Egri Katolikus Rádió (EKR) beszélgetéseiből válogatott, összesen 43 óra hanganyagon tanított, környezetfüggő akusztikus modell a HTK [10] eszközeinek segítségével készült, ami összesen 6121 egyenként 13 Gauss-függvényből álló állapotot tartalmaz. A 16 kHz-en mintavételezett felvételek lényegkiemeléséhez 39 dimenziós, delta és delta-delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseken alapuló jellemzővektorokat hoztunk létre, és vak csatornakiegyenlítő eljárást is alkalmaztunk.

3.6. Hálózatépítés és dekódolás

A legvalószínűbb illeszkedés kereséséhez használt dekódolási folyamat szerver-kliens-architektúra alapján működik. A rendszer által használt beszédfelismerő kliens (VOXclient) végzi a beszédalapú információ lényegkiemelését és a 3.5 fejezetben vázolt jellemzővektorokká való alakítását. A jellemzővektorokat ezután a kliens továbbküldi a szerveroldali alkalmazásnak (VOXerver), ahol a tényleges dekódolási lépések megtörténnek. A legvalószínűbb illeszkedés megtalálásához a beszédfelismerési modelleket ún. súlyozott, véges állapotú átalakítókban (*Weighted Finite State Transducer* - WFST) [12] egyesítjük. A szerver a kliensoldal felé végül visszaküldi a megtalált legvalószínűbb felismerési kimenetet; minden frissítés 250 ms-onként zajlik. A normalizálás során átalakított nem verbális nyelvi elemek (számok, URL-ek) írott formára való visszaalakítása szintén a kliens oldalon történik a már visszaküldött legvalószínűbb felismerési kimeneten.

4. Kísérletek

A magyar nyelvre ingyenesen elérhető gépi beszédfelismerő rendszerek teljesítményének felhasználó szempontú összehasonlítását egy 6 résztvevős (3-3 férfi/nő, életkor: 22-38 év) kísérletben végeztük, amiben a résztvevők

1. egy rövid (7 mondatból álló) jogi szöveget olvastak fel 2 módban:
   (a) normál olvasási mód és
   (b) az írásjelek hangalakjának diktálásával,
   majd a hanganyagokat rögzítettük.

2. A felismerés
   (a) a Google és Nuance rendszerek esetében a normál olvasási módban,
   (b) SpeechTex rendszer esetében pedig az írásjelek hangalakjának diktálásával
   készült változatokon történt.
3. Ezeken a kimeneti szövegeken zajlott aztán a hibajavítási feladat, ami a felismerési hibák detektálását, javítását és egyéb szövegszerkesztési műveleteket foglalja magába.
4. A beszédfelismerésen alapuló diktálási tapasztalatokról végül egy kérdőívben kérdeztük a résztvevőket, amiben egy Likert-alapú skálán (az adott állítással való egyetértés erősségének kifejezése egy 1-5-ig terjedő intervallumban) kellett értékelniük a diktálási feladatot és a kísérletben szereplő diktálórendszerek teljesítményét.

5. Eredmények

Nem parametrikus páros, egymintás Mann-Whitney-Wilcoxon tesztekkel ellenőriztük, hogy a rendszerek közti szóhibaarányok szignifikánsan különböznek-e. A SpeechTex rendszer szóhibaaránya szignifikánsan ($p < 0.01$) alacsonyabb volt, mint Google-é és Nuance-é, ami nem meglepő annak fényében, hogy a SpeechTex rendszer *in-domain* nyelvi adaton lett tanítva. Részletes eredmények a 3. táblázatban találhatók, ebben fel vannak tüntetve mindhárom rendszer mindkét olvasási módon (normál, illetve az írásjelek diktálásával) elért szóhibaarányai. Jól látható, hogy a Google és Nuance rendszerek az írásjelek diktálása módban magasabb szóhibaaránnyal dolgoznak, ezek nyelvi modelljei noha tartalmazzák az írásjelek (pont, vessző, stb.) kiejtett alakjait mint homofónokat, tehát nem abban funkcióban és sorrendiségben mint ahogy azok az írásjeles diktálási módban történő diktáláskor szerepelnek. A SpeechTex rendszernél a normál olvasási mód eredményez magasabb szóhibaarányt, hiszen a nyelvi modell tartalmazza az írásjelek kiejtett alakját, és ezek elég gyakran fordultak elő a korpuszban ahhoz, hogy el nem hangzásuk rontsa a felismerési pontosságot.

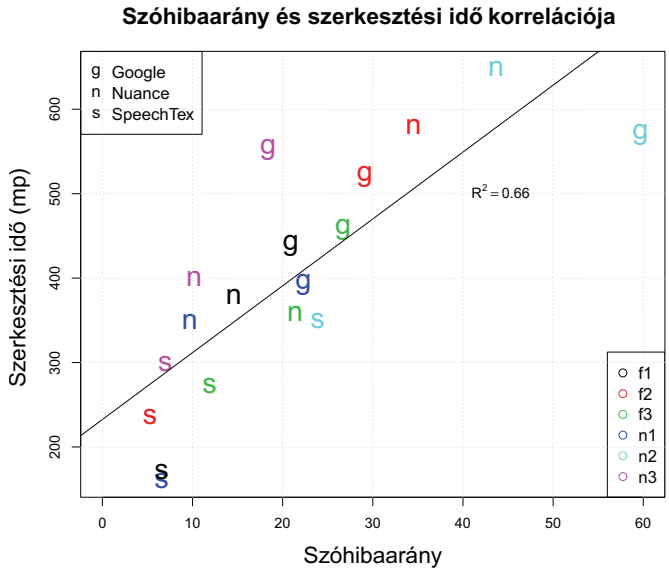
Szerkesztési időt tekintve azonban csak a SpeechTex rendszeré szignifikánsan ($p < 0.01$) alacsonyabb, mint Google-é és Nuance-é, ez utóbbi kettő közt nincs szignifikáns különbség.

Mint azt a 1. ábra mutatja, hogy a szóhibaarány és szerkesztési idő közötti korreláció szignifikáns és erős ($R^2 = 0.66$). Ezt árnyalja, hogyha az írásjelek diktálhatóságát mint faktort tekintjük; a 2. ábrán jól látható, hogy ugyanolyan szóhibaarány mellett az írásjelek diktálhatóságát lehetővé tevő rendszer alacsonyabb szerkesztési időt eredményez.

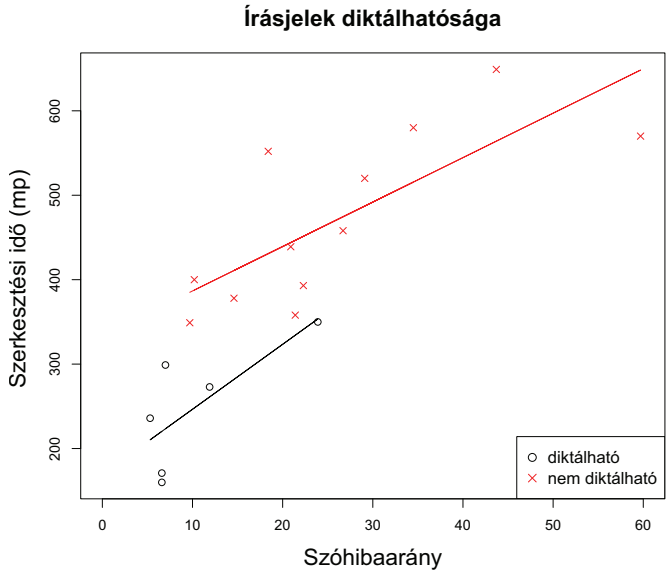
A 3.1. fejezetben leírt sikerességi ráta átlagosan 97,5% a Google, 98,9% a Nuance és 99,4% a SpeechTex rendszernél, és kizárólag a SpeechTex és Google közti különbség szignifikáns. Érdekes a sikerességi ráta összefüggése a szerkesztési idővel; azt találtuk, hogy fordított kapcsolat áll fenn: minél hosszabb a szerkesztési idő ($R^2 = -0,47079$), annál alacsonyabb a sikerességi ráta. Ez azt jelenti, hogy átlagosan több hiba marad egy több ideig szerkesztett kimeneti szövegben.

3. táblázat. Szóhibaarányok.

	Normál			Írásjeles		
Résztvevő	**Google**	**Nuance**	**SpeechTex**	**Google**	**Nuance**	**SpeechTex**
n1	22.3	9.7	8.3	38.5	23.4	6.6
n2	59.7	43.7	40.8	51.4	33.3	23.9
n3	18.4	10.2	8.3	34.2	na	7.0
f1	20.9	14.6	11.2	36.2	19.3	6.6
f2	29.1	34.5	10.7	42.4	32.9	5.3
f3	26.7	21.4	14.1	40.7	21.0	11.9
Átlag	**29.5**	**22.3**	**15.5**	**40.5**	**25.9**	**10.2**
Szórás	15.2	13.9	12.5	6.0	6.6	7.0



1. ábra. Szóhibaarány és szerkesztési idő korrelációja.

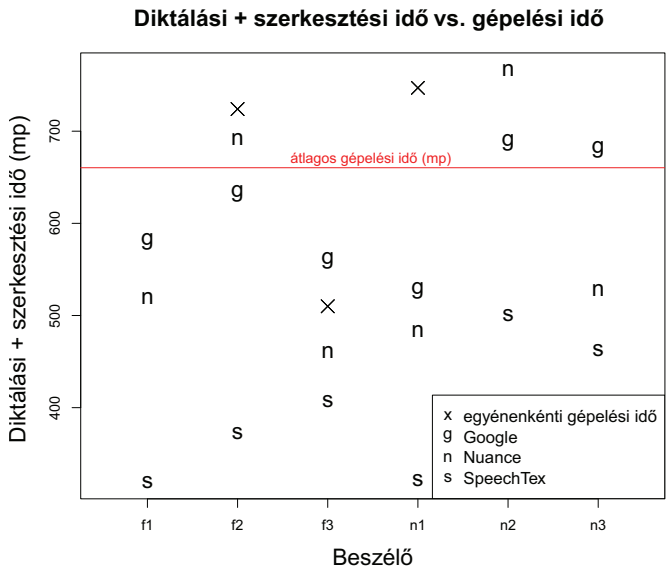


2. ábra. Szóhibaarány és szerkesztési idő az írásjelek diktálhatóságának függvényében.

4. táblázat. Diktálási tapasztalat és szerkesztési vs. gépelési idő becslése az automatikus beszédfelismerőn alapuló diktálás függvényében (nincs', 'van', 'rendszeres')

a diktálás és szerkesztés ... mint a gépelés	**nincs**	**van**	**rendszeres**
sokkal lassabb	1		
lassabb	1		
ugyanannyi		2	
gyorsabb			
sokkal gyorsabb			2

Végül a diktálással és szerkesztéssel eltöltött időt mértük össze ugyanazon szöveg begépelésének idejével - ilyen adat csak 3 résztvevőtől állt rendelkezésünkre. A 3. ábra mutatja, hogy néhány kiugróan magas együttes szerkesztési és diktálási időtől eltekintve, az együttes diktálási és szerkesztési idő rövidebb, mint a gépelési idő átlagosan. Az utólagos kérdőívből azonban az derült ki, hogy a diktálási tapasztalattal nem rendelkező résztvevők hosszabbnak érzékelték a diktálással és szerkesztéssel együttesen eltöltött időt, mint ugyanezen szövegbeviteli feladat gépeléssel való végrehajtását (l. 4. táblázat).



3. ábra. Diktálási és szerkesztési idő összevetése a gépelési időtartammal.

6. Összefoglalás

A beszédfelismerési kutatások középpontjában tipikusan a szófelismerési hiba csökkentése áll. Azonban az egyes speciális alkalmazásoknál, mint például a diktálás, a felhasználó számára közvetlenül nem a szóhibaarány, hanem elsősorban a diktálásra és javításra fordított idő csökkentése releváns. Tanulmányunkban az utóbbi célt tűztük ki. Egyrészt a számunkra közvetlenül hozzáférhető SpeechTex beszédfelismerési motor feladatra szabását végeztük el, másrészt a megoldásunkat összehasonlítottuk a lehetséges piaci alternatívákkal. A kimeneti szöveg jólformázottsága érdekében a fejlesztés során kiemelt hangsúlyt kapott az írásjelek diktálhatósága. Ezzel a funkcióval ismereteink szerint a vizsgálatok végzésekor nem bírtak a magyar nyelven ingyenesen hozzáférhető piaci termékek. Az összehasonlítást kontrollált körülmények között végeztük, kitüntetett figyelemmel a hibajavítási folyamatra. Az eredmények igazolták, hogy a lecsökkent szóhibaarány gyorsabb hibajavítással jár együtt. Ugyanakkor, tapasztalataink szerint az írásjelek diktálás során történő elhelyezése magát a hibajavítási folyamatot is gyorsította azáltal, hogy a felismerési kimenetben előforduló hibák detektálását megkönnyítette. Vagyis, az írásjelek diktálását lehetővé tevő megközelítés ugyanolyan szóhibaarány mellett alacsonyabb szerkesztési időt eredményezett. A kísérleteinkben résztvevő diktálási tapasztalattal nem rendelkező felhasználók azonban így is hosszabbnak érzékelték a diktálással és hibajavítással eltöltött időt, mint a szöveg begépelésének időtartama. Ennek egyik lehetséges oka, hogy az írásjelek diktálása szokatlan a felhasználó számára, hiszen a beszélt nyelvre ez nem jellemző. Megoldás lehetne az automatikus írásjelezés, de a legújabb kuta-

tásokban elért 50% körüli pontossága általános témakörben egyelőre nem valós alternatíva. Az automatikus írásjelezés pontossága kötött témakörnél alkalmazott gépi tanulási eljárással várhatóan jelentősen fokozható, ahogy a beszédfelismerési pontosság is magasabb kötött témájú korpuszon való tanítás esetén. Végül a hibajavítási felület ergonomikussá tétele és a felismerési hibák automatikus detektálása is lehetőségek a diktálás megkönnyítésében - ebben az irányban további kutatásokat tervezünk.

Köszönetnyilvánítás

Kutatásunkat a PIAC_13-1-2013-0234 (Patimedia) és KMR_12-1-2012-0207 (DIANA) projektek támogatták.

Hivatkozások

1. Tarján B., Nagy T., Mihajlik P., Fegyó T.: Magyar nyelvű, kísérleti e-mail diktálórendszer. In: IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013), Szeged, Magyarország (2013) 21–28
2. Désilets, A., Stojanovic, M., Lapointe, J.-F., Rose, R., Reddy, A.: Evaluating Productivity Gains of Hybrid ASR-MT Systems for Translation Dictation. In: Proc. of IWSLT 2008, Hawaii, USA (2008) 158–166
3. Kolar, J., Lamel, L.: Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text. In: Proc. of Interspeech 2012, Portland, Oregon, USA (2012) 1374–1377
4. Batista, F., Caseiro, D., Mamede, N. and Trancoso, I.: Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. Speech Communication, Vol. 50, No. 10 (2008) 847-862
5. Vicsi K., Szaszák Gy., Németh Zs.: Prozódiai információ használata az automatikus felismerésben; mondatmondalitás felismerése. In: V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007), Szeged, Magyarország (2007) 69-80
6. D. Yu, M.-Y. Hwang, P. Mau, A. Acero, and L. Deng: Unsupervised learning from users' error correction in speech dictation. In: Proc. of Interspeech 2004, Jeju Island, Korea (2004) 1969–1972
7. Leijten, D.J.M., van Waes, L.: Error correction strategies of professional speech recognition users: Three profiles. Computers in Human Behavior, Vol. 26 (2010) 964—975
8. Jurafsky, D. and Martin, J.H.: Speech and language processing. An introduction to NLP, computational linguistics, and speech recognition, Englewood Cliffs, NJ: Prentice Hall (2000)
9. Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P.: Dgttm: A freely available translation memory in 22 languages. In: Proc. of LREC 2012 (2012) 454—459
10. G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, The HTK Book, version 3.4. Cambridge, UK: Cambridge University Engineering Department (2006)
11. A. Stolcke, Srilm – an extensible language modeling toolkit. In: Proceedings International Conference on Spoken Language Processing 2002, Denver, USA (2002) 901–904

12. M. Mohri, F. Pereira, and M. Riley: Weighted finite-state transducers in speech recognition. Computer Speech and Language, Vol. 16, No. 1 (2002) 69–88

V. Véleménykinyerés

TrendMiner: politikai témájú Facebook-üzenetek feldolgozása és szociálpszichológiai elemzése

Miháltz Márton, Váradi Tamás

MTA Nyelvtudományi Intézet Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály, Nyelvtechnológiai Kutatócsoport
1068 Budapest, Benczúr utca 33.
{mihaltz.marton, varadi.tamas}@nytud.mta.hu

Kivonat

Az előadásban bemutatjuk a *Trendminer* projekt[1] módszereit és eredményeit. Az FP7-es finanszírozású, európai kooperációban megvalósuló munkálatok célja közösségi média stream-ek valós idejű figyelése, trendkövetése és összegzése. Ezen belül az MTA NYTI által fejlesztett, MTA TTK KPI Narratív Pszichológiai Kutatócsoportjának munkatársaival együttműködésben készített eszközök célkitűzése a magyar Facebook-felhasználók politikai tartalmú posztokra adott nyilvános kommentjeinek nagymennyiségű gyűjtése, elemzése és szociálpszichológiai aspektusokból történő automatikus kiértékelése volt. Ezzel a kutatással reményeink szerint támogatást nyújtunk annak vizsgálatához, hogy milyen érzelmi és társas pszichológiai jelenségek, trendek figyelhetők meg a magyar közösségimédia-hozzászólók politikai témákra reagáló üzeneteiben.

A projektben a Facebook Graph API segítségével 12 hónap alatt mintegy 140 ezer publikus posztot és az ezekre érkezett mintegy 2 millió publikus kommentet gyűjtöttünk össze több mint 1300 Facebook-oldalról, melyek Magyarországon bejegyzett politikai pártokhoz, azok tagszervezeteihez, képviselőihez és -jelöltjeihez kötődnek. Figyelmet fordítottunk a 2014-es évben lezajlott országgyűlési és európai parlamenti választások jelöltjeinek és tisztséget elnyerő képviselőinek oldalain megjelent üzenetek gyűjtésére is.

A begyűjtött üzeneteket adatbázisban tárolás után az alábbi pipeline segítségével dolgoztuk fel: mondatszegmentálás, tokenizálás (*huntoken* eszköz), morfológiai elemzés és szófaji egyértelműsítés (*hunpos* és *hunmorph* eszközök), szótő és morfológiai elemzés egyértelműsítése (saját eszköz). Ezt követte a domain számára releváns entitások azonosítása, illetve a tartalomelemzés a *NooJ* eszköz nyílt forrású változatával, melyhez parancssori változatot készítettünk[2]. A tartalomelemzés az üzenetek érzelmi polaritásának (sentiment) vizsgálatán túl további 5 pszichológiai dimenziót érintett [2] (elsődleges-másodlagos gondolkodási szint, közösségiség-ágencia, optimizmus-pesszimizmus, individualizmus-kollektivizmus), melyek ilyen célú felhasználására tudomásunk szerint ez az első példa.

[1] http://www.trendminer-project.eu

[2] https://bitbucket.org/tkb-/nooj-cmd

Mivel az általunk alkalmazott *hun** nyelvi feldolgozó eszközöket[3] a sztenderd nyelvváltozatot reprezentáló, többnyire híroldalak anyagát tartalmazó szövegek felhasználásával fejlesztették ki, teljesítményük elmaradt a várt szinttől a projektben vizsgált közösségimédia-domain speciális nyelvezetű szövegein. Emiatt szükség volt az eszközök vizsgált nyelvterülethez való adaptációjára. Ehhez készítettünk egy 1,25 millió Facebook üzenetet (29M token) tartalmazó korpuszt és az ismeretlen szavak gyakorisági listáján 15-ször vagy annál gyakrabban szereplő tételeket vizsgáltuk manuálisan. Ennek segítségével azonosítottuk az ismétlődően előforduló reguláris problémákat, valamint a gyakori és releváns ismeretlen szavakat, rövidítéseket, szleng kifejezéseket, emotikonokat stb. Ezek alapján a tokenizáló eszközt elő- és utófeldolgozó szintekkel egészítettük ki a gyakori írásjel-használati és egyéb reguláris problémák kezelésére. Az ismeretlen szavak és rövidítések egy részét sztenderd nyelvi (az elemzők számára ismert) alakra javítottuk a szövegben, másik részüket pedig analóg ismert szavak felhasználásával hozzáadtuk a morfológiai elemző szótárának megfelelő paradigmáihoz annak érdekében, hogy ezután azok különböző inflexiós alakjaikat is fel tudjuk ismertetni.

Az üzenetekben szereplő névelemek azonosítására a *hunNER* statisztikai gépi tanuló eszközzel [3] – a vizsgált nyelvi domain sajátosságai miatt – szerzett kedvezőtlen tapasztalatok után lexikon alapú felismerőt fejlesztettünk a *Java NooJ* eszköz segítségével. Az 5 szociálpszichológiai dimenzió nyelvi kifejezéseinek felismerését – a szótő, szófaj és morfológiai annotáció eredményeire támaszkodva – szintén *NooJ* nyelvtanok fejlesztése segítségével végeztük el. Az érzelmi valencia és a közösségiségágencia dimenziók lexikonjainak fejlesztői felhasználtak egy 176K kommentből (4,9M token) álló korpuszt, melynek 100-szor vagy gyakrabban szereplő kifejezéseit 7 független humán annotátor kódolta a megfelelő kategóriákra.

Az eredmények kiértékelésére három, az egyes pártok oldalain érkezett hozzászólásokat a teljes 2M szavas korpusz arányában tartalmazó gold standard korpuszt állítottunk össze, melyek 337, 336 és 672 kommentet tartalmaztak (7,6K, 7,5K, 17,9K token). A korpuszokat egy egyszerű algoritmussal tagmondatokra bontottuk, majd ezeket 3-3 független annotátor látta el jelölésekkel a különböző annotációs kategóriákra. Az egyes modulok annotációs pontossága (precision) a gold standardhoz képest 65,75%–98,36% között változott, fedése (recall) 13,8%–82,05% között. Az egyes kommentek érzelmi (sentiment) polaritásának felismerési pontossága (accuracy) 84,63% volt [1].

A projekt során elkészítettük a magyar politikai élet vizsgált időszakban releváns entitásait, fogalmait és azok tulajdonságait modellező, OWL nyelven írt politikai ontológiát is. Az ontológiát felhasználtuk többek között a szövegekben azonosított entitások linkelésére a projekt számára készült, időszakok, kulcsszavak, entitások, együttelőfordulások, érzelmi polaritás grafikonok stb. megjelenítésére alkalmas vizualizációs felületben is.

A projekt zárultával szabadon hozzáférhetővé tettük az előállított eszközök és erőforrások egy részét: a *hun** eszközök domain-adaptációját megvalósító kiegészítéseink és a Java NooJ eszköz parancssori változatának forráskódját, a politikai ontológiát,

[3] http://mokk.bme.hu/en/eszkozok/

valamint a projektben vizsgált 2M kommentet (az összes metaadattal, NLP- és tartalomelemzési annotációival együtt) tartalmazó korpuszt is[4].

Hivatkozások

1. Miháltz, M.: Socio-psychological Analysis of Social Media Messages in Politics. In: Martínez Fernández, J. L., Martínez, P., Ogrodniczuk, M., Miháltz, M.: Newly generated domain-specific language data and tools. TrendMiner Project Public Deliverable D10.1., (2014) 46–63 http://www.trendminer-project.eu/images/d10.1_final_version.pdf
2. Pólya, T., Csertő, I., Fülöp, É., Kővágó, P., Miháltz, M., Váradi, T.: A véleményváltozás azonosítása politikai témájú közösségi médiában megjelenő szövegekben. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (2015), ld. jelen kötetben
3. Simon, E.: Approaches to Hungarian Named Entity Recognition. PhD dissertation. Budapest University of Technology and Economics, Budapest (2013)

[4] Letöltések, további információ: http://corpus.nytud.hu/trendminer/

A véleményváltozás azonosítása politikai témájú közösségi médiában megjelenő szövegekben

Pólya Tibor[1], Csertő István[1], Fülöp Éva[1], Kővágó Pál[2],
Miháltz Márton[3], Váradi Tamás[3]

[1] Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.mta.hu
cserto.istvan@ttk.mta.hu
fulop.eva@ttk.mta.hu
[2] Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
kovago.pal@ttk.mta.hu
[3] Magyar Tudományos Akadémia, Nyelvtudományi Intézet,
1068 Budapest, Benczúr utca 33.
mihaltz.marton@nytud.mta.hu
varadi.tamas@nytud.mta.hu

Kivonat: A tanulmány a Trendminer projekt keretében a politikai témájú közösségi médiában tetten érhető véleményváltozások automatikus felismerésére kidolgozott elemzési eszközöket mutatja be. A projekt keretében a következő öt modult dolgoztuk ki: Individualizmus-kollektivizmus, Optimizmus-pesszimizmus, Közösségiség és ágencia, Érzelmi polaritás és Politikai szereplők. A tanulmányban ismertetjük a modulok bemérésének eredményeit és a modulok pszichológiai jelentésének ellenőrzésére végzett vizsgálatok eredményeit.

1 A véleményváltozás felismerésére fejlesztett modulok

1.1 Individualizmus versus kollektivizmus modul

Az individualizmus versus kollektivizmus fogalompár jelentése arra vonatkozik, hogy az egyes társadalmak tagjai hogyan gondolkodnak az egyén és a csoport viszonyáról. Individualista társadalmakban az egyén, kollektivista társadalmakban a csoport cselekvése áll a figyelem középpontjában (például [9]). Az individualizmus és kollektivizmus újabb vizsgálatai szoros összefüggést tártak fel az adott társadalomra jellemző figyelmi fókusz és aközött, hogy az adott társadalom által használt nyelv megköveteli a személyes névmások használatát, vagy pedig megengedi személyes névmások elhagyását [10]. Az angol kifejezést használva az utóbbi csoportba sorolt nyelveket hívjuk ’pronoun-dropping’ nyelveknek. Az individualizmus és a névmáselhagyás közötti kapcsolatot vizsgáló kutatások azt állapították meg, hogy az individualizmus magas szintjével bíró társadalmak esetében a nyelvhasználat rendszerint megköveteli a sze-

mélyes névmások használatát, míg az individualizmus alacsonyabb szintjével rendelkező társadalmak által használt nyelv megnyilatkozásaiból kihagyhatók a személyes névmások. Kashima és Kashima [10] például egyik vizsgálatukban azt találták, hogy az individualizmus szintje és a 'pronoun drop' jelenség közötti összefüggés mértéke korrelációs együtthatóval kifejezve r=,75 erősségű. Az individualizmus általunk végzett vizsgálata arra a feltevésre épít, hogy az individualizmus és a személyes névmások elhagyása közötti összefüggés nemcsak társadalmak közötti összehasonlításban értelmezhető, hanem egy-egy társadalom csoportjai közötti összehasonlításban is.

Az individualizmus-kollektivizmus modul a személyes névmásokat, a személyragos és személyjeles szóalakokat ismeri fel a szövegekben. Az individualizmus szintjét ezen kategóriák előfordulásából számolt hányadossal fejezzük ki. A hányados a szövegben ténylegesen és potenciálisan előforduló személyes névmások számát veti össze egymással. A potenciálisan előforduló személyes névmások számát a személyragos vagy személyjeles szóalakok száma adja meg, mivel mindkét esetben lehetséges lenne a személyes névmások használata. Az individualizmus szintjét jelző mutató két változatát definiáljuk. Az első mutató esetében bármilyen számú és személyű lehet a személyes névmás, a személyrag és a személyjel. A második mutató esetében csak az egyes szám első személyű személyes névmásokat, személyragokat, és személyjeleket vesszük figyelembe.

Individualizmus mutató 1 = személyes névmások előfordulásának száma / (személyes névmások előfordulásának száma + személyraggal vagy birtokos személyjellel ellátott szóalakok előfordulásának száma)

Individualizmus mutató 2 = egyes szám első személyű személyes névmások előfordulásának száma / (egyes szám első személyű személyes névmások előfordulásának száma + egyes szám első személyű személyraggal vagy személyjellel ellátott szóalakok előfordulásának száma)

1.2 Optimizmus versus pesszimizmus modul

Az optimizmus a helyzetek és lehetőségek jó oldalának kiemelését jelenti (például [18]. Feltevésünk szerint az optimizmus magában foglalja a cselekvési lehetőségek nyitottságát is, ami szisztematikusan összefügg azzal, hogy a személy megnyilatkozása az idő mely tartományára vonatkozik. A múlt idejű tartomány esetében a cselekvési lehetőségek bezárultak. A jelen idejű tartomány esetében nyitottak a cselekvési lehetőségek, de a cselekvés végrehajtását rendszerint nehezítő körülményeket is figyelembe kell venni. A jövő idejű tartomány esetében a cselekvési lehetőségek nagy mértékben nyitottak. Mindezek alapján azt feltételezzük, hogy az optimista személyek többet beszélnek a jövőről és kevesebbet beszélnek a múltról, szemben a pesszimista személyekkel, akik többet beszélnek a múltról és kevesebbet beszélnek a jelenről. Ezt a feltevést olyan vizsgálatok eredményei támogatják, amelyek kimutatták, hogy a pesszimizmus magas fokával rendelkező depressziós személyek jellemzően az idő múltbeli tartományára fókuszálnak [8]. A feltevésünket támogatja az a megállapítás is, amely szerint azok a kognitív torzítások, amelyek a jövőre vonatkoznak, adaptívak is lehetnek, mivel energetizálják a viselkedést [11]. A viselkedés energetizáltsága pedig fontos jellemzője az optimizmus magas szintjével rendelkező személy pszichológiai álla-

potának. Feltevésünknek értelemszerűen azokban az esetekben van jelentősége, amikor a személy az idő mindhárom tartományára fókuszálhat. Ilyen például egy döntés meghozatala, amikor a személy egyaránt fókuszálhat a jelentést megalapozó előzményekre, a döntés kivitelezésére vagy a döntés következményeire.

Az optimizmus versus pesszimizmus modul azokat a nyelvi markereket azonosítja, amelyek jelzik azt, hogy a megnyilatkozás tartalma az idő mely tartományára vonatkozik. A múlt és jelen idő tartományának azonosítása az igeidő, a jövő idő tartományának azonosítása a jövőre utaló igei szerkezetek, időhatározószavak és a jelentésükben jövő időre utaló aspektuális igék alapján történik. Az optimizmus versus pesszimizmus szintjének mérését szintén két mutató kiszámításával végezzük el. Az első mutató a nem múlt idejű és múlt idejű tartományok arányát, a második mutató a jövő idejű és múlt idejű tartományok arányát fejezi ki.

Optimizmus mutató 1 = nem múlt idejű igék előfordulásának száma / (nem múlt idejű igék előfordulásának száma + múlt idejű igék előfordulásának száma)

Optimizmus mutató 2 = jövő időt kifejező nyelvi elemek előfordulásának száma / (jövő időt kifejező nyelvi elemek előfordulásának száma + múlt idejű igék előfordulásának száma)

1.3 Közösségiség és ágencia modul

A közösségiség-ágencia modul azáltal haladja meg a politikai közszereplők iránti attitűd pozitív-negatív (támogatás-elutasítás / szimpátia-antipátia) dimenzióját, hogy a közvéleményben megjelenő nyelvi értékeléseket a valencia mellett tartalmuk szerint is differenciálja. A tartalmi kategorizáció alapja a szociális kogníció szakirodalmában elterjedt kétdimenziós modell. A közösségiség (communion) és az ágencia (agency) két olyan fő jelentésdimenzió avagy tartalmi kategória, amelyek a társas ítéletalkotás (személyészlelés, önészlelés, benyomásformálás, sztereotípiák, előítéletek, csoportközi attitűdök) általánosan érvényes alapjait képezik [1,2]. A közösségiség az egyén / csoporttag más egyénekhez / csoporttagokhoz fűződő viszonyának minőségét jellemzi a társakkal való együttműködés és a csoportcéloknak való alárendelődés erkölcsi normái szerint (például becsületes, hűséges, barátságos, tisztelettudó). Az ágencia a célkövető viselkedés hatékonysága szempontjából jellemzi az egyént (például céltudatos, hozzáértő, önérvényesítő, sikeres; [2]). A szociálpszichológián belül több különböző kutatási területen olyan dimenziópárokat azonosítottak, amelyek a közösségiség és az ágencia dimenzióinak feleltethetők meg (barátságosság/erkölcsösség és kompetencia [3,7,27,28]; társas hatékonyság és intellektuális hatékonyság [20]; társas haszon és egyéni haszon [17]; társas kívánatosság és társas haszon [5]). Az empirikus vizsgálatok kétféle funkcionális aszimmetriát állapítottak meg a közösségiség és az ágencia társas ítéletalkotásban betöltött szerepét illetően. Egyrészt a közösségiség meghatározóbb az ítéletalkotásban: különböző személyiségvonások megítélésének egyéni különbségeit tekintve például a közösségiség faktora több mint kétszer annyi varianciát magyaráz, mint az ágencia. Másrészt, míg a másokra vonatkozó ítéletekben a közösségiség elsődleges, addig a szelfre vonatkozó ítéletekben az ágencia [1,2].

A két jelentésdimenzió megkülönböztetése a politikai közvéleménykutatásban lehetővé teszi az egyes dimenziók relatív súlyának vizsgálatát a politikai szereplők iránti

attitűdök alakulásában. A közösségi médiában megjelenő közvélemény releváns tartalmainak kvantitatív elemzésével feltárható, hogy a közösségiség és ágencia szempontjából pozitív és negatív tartalmak gyakorisága milyen összefüggést mutat a politikai szereplők közmegítélésével, van-e különbség a prediktív erejüket tekintve, ha igen, melyikük jobb prediktora a közvélemény alakulásának, s számos további, gyakorlati jelentőségű kérdés vizsgálható empirikusan (például politikai orientáció, pártpreferencia, szocioökonómiai státusz stb. szerint különböző társadalmi csoportok politikai ítéletalkotásában milyen súllyal esnek latba az egyes dimenziók). A kvantitatív elemzés elméleti jelentősége, hogy a közösségiség és ágencia szociális kognícióban betöltött szerepével kapcsolatos hipotézisek és kísérleti eredmények a való életben is tesztelhetők, bármilyen elektronikus korpuszon.

Magyar nyelven már létezik két olyan számítógépes tartalomelemző eszköz, amely a közösségiséggel és ágenciával összefüggő tartalmakat azonosít, és amelyek szintén a NooJ környezetben lettek kifejlesztve. A narratív pszichológiai tartalomelemzésben alkalmazott NarrCat rendszer két moduljáról van szó (Narrative Categorial Content Analytical Tool; [13]). Az értékelés modul az explicit pozitív és negatív társas ítéletek nyelvi kategóriáit azonosítja [4], míg az ágencia modul a történet szereplőinek tulajdonított, aktív és passzív igékben, valamint a szándék és a kényszer intencionális állapotaiban kifejeződő cselekvőképességet méri [6]. Az értékelés modulhoz képest a közösségiség-ágencia modul újdonsága, hogy a nyelvi értékeléseket a valencia mellett a két tartalmi kategória szerint is differenciálja, ebből fakadóan ugyanakkor az értékelések szűkebb tartományát fedi le. Az ágencia modultól abban különbözik, hogy az aktivitás és a szándékteliség mértéke helyett a motivációra, kompetenciára, produktivitásra és kontrollra vonatkozó gyakori pozitív és negatív tartalmakat azonosítja. Összességében, míg a NarrCat modulok az elbeszélő szöveg kompozíciós összetevőit kvantifikálják, addig a közösségiség-ágencia modul hagyományos értelemben vett tartalmi kategóriák nyelvi markereit azonosítja [12]. Ezek olyan szavak, amelyek az ágencia és közösségiség szempontjából pozitív és negatív ítéleteket fejeznek ki, vagy ilyen ítéletek részei. Szófajilag lehetnek a személy vagy viselkedés minőségét leíró melléknevek (tisztességes/gerinctelen; intelligens/buta), viselkedést leíró igék (összefog/átver; győz/megbukik), vagy a viselkedés minőségét leíró, melléknevekből és igékből képzett határozószók, ítélet részei pedig ezek főnévi alakjai vagy absztrakt fogalmi kategóriák.

A közösségiség és ágencia nyelvi markereit független kódolók ítéletei alapján gyűjtöttük ki a szótárfejlesztésre használt korpuszból. A korpuszban leggyakrabban előforduló lemmákat ($n \geq 100$), összesen 3108 szót tartalmi kategória (közösségiség/ágencia) és valencia (pozitív/negatív) szerint osztályozták a kódolók írott instrukció alapján. Tartalmi kategóriánként három-három független kódoló döntött a szavak besorolásáról, a kategóriába sorolt szavakat valencia (pozitív/negatív) szerint elkülönítve. A nem száz százalékos konszenzus alá eső besorolások esetében egy negyedik kódoló, a modul fejlesztője (Cs. I.) döntött a szavak kategóriatagságáról. A tartalom és valencia szerint kialakított négy kategóriát négy szótárba rendeztük (közösségiség-pozitív, $n = 65$; közösségiség-negatív, $n = 88$; ágencia-pozitív, $n = 88$; ágencia-negatív; $n = 41$). Az egyes szótárakon belül szófaj szerint további alszótárakat különítettünk el (melléknevek, igék, határozók, főnevek). A szótárakat a NooJ környezetben létrehozott lokális nyelvtanokba építettük, amelyek az igék, főnevek és melléknevek

első és második személyű alakjait, vagyis a szelfre és a kommunikációs partnerre referáló tartalmakat kizárják a találatok köréből.

1.4 Érzelmi polaritás modul

Egy szöveg pszichológiai vonatkozásainak legfőbb tényezői közé tartoznak az érzelmek, az értékelő megnyilvánulások, melyek az elbeszélő sajátos perspektívájának létrehozásához járulnak hozzá. A szövegek valenciával rendelkező elemei valójában arról tudósítanak, hogy az elbeszélő hogyan viszonyul az őt körülvevő világhoz vagy akár önmagához, milyen szubjektív értékeléssel látja el azokat. A számítógépes tartalomelemzésben ennek mérésében nagyon gyakran alkalmazott módszer az ún. sentiment analysis, vagyis érzelemelemzés [16,26], melynek segítségével feltárják a természetes nyelvhasználatban előforduló attitűdöket bizonyos témákkal kapcsolatban. A különböző internetes portálokon, fórumokon, közösségi oldalakon ezt az ott előforduló pozitív és negatív szavak automatizált detektálásával valósítják meg. A Trendminer projekt keretében létrehozott érzelem/értékelés elemző specifikusan adott szövegbázisra készült az egyes politikai pártok facebook kommentjeiben és posztjaiban előforduló szavak felhasználásával. Első lépésben nyolc független kódoló sorolta be a gyakorisági sorrendbe állított szavak közül a legalább tízszer előfordulókat a pozitív-negatív dimenzió mentén, így létrejött az érzelmi valencia modul két pólusán elhelyezkedő szavak gyűjteménye. A bemérési szakaszban a szótári elemek szövegbeli előfordulásának elemzése történt, mellyel ellenőrizni lehetett a szavak kontextusának figyelembevételével a kategorizáció érvényességét. Az érzelmi polaritás elemző számára a legnagyobb kihívást a politikai fórumokon nagyon gyakran előforduló irónia, tudatosan kommunikációs maximát sértő („átvitt értelmű") közlések, szleng és vulgáris nyelvezet figyelembevétele jelenti. A modult a szövegben előforduló szereplők azonosításával együtt alkalmazva megtudhatjuk, hogy kire vonatkoznak a szöveg szerzőjének érzelmi, értékelésbeli véleménynyilvánításai.

1.5 Politikai szereplők modul

A politikai szereplők modul azonosítja a politikai pártokra vagy politikusokra vonatkozó utalásokat a szövegben, a személyneveket a megfelelő pártokhoz rendeli. Megtalálja továbbá azokat az utalásokat, melyek a 2014-es országgyűlési választásokon induló pártokkal kapcsolatosak. A szótár 277 elemét 500, véletlenszerűen kiválasztott Facebook-komment leggyakoribb szavaiból válogattuk ki, kiegészítve a 2014-es országgyűlési választásokon induló pártok és politikusok listájával. Külön figyelmet kaptak a helytelenül írott nevek (pl. „órbán"), illetve a gúnynevek (pl. „libás"). Az ilyen esetek akkor kerültek a szótárba, ha legalább kétszer előfordultak a mintában. A politikai szereplők modul információt szolgáltat arról, hogy adott politikai szervezet közösségi oldalain mely politikai pártokról folyik a diskurzus. A jelen vizsgálatban bemutatott többi modullal együtt lehetővé teszi, hogy a többi modul által megtalált tartalmakat összekapcsolja a referált entitással

2 A modulok statisztikai megbízhatósága

A modulok megbízhatóságának vizsgálatához az elemzett korpuszból vett kisebb mintán végzett gépi és manuális elemzés eredményeit vetettük össze, a manuális elemzés alapján kapott gold standard-eket véve referenciának. Az öt modul teszteléséhez három különböző szövegmintát használtunk: (1) Politikai szereplők: 337 komment, 7615 token; (2) Érzelmi polaritás: 336 komment, 7540 token, 1295 tagmondat; (3) Közösségiség-ágencia, Optimizmus-pesszimizmus, Individualizmus-kollektivizmus: 672 komment, 17 924 token, 3188 tagmondat. A szövegmintába került kommentek pártok szerinti eloszlása megegyezett a teljes korpusz eloszlásával. A Politikai szereplők modul szövegmintáját kommentekre, a másik két mintát tagmondatokra szegmentáltuk, az így kapott elemzési egységeket vizsgáltuk a gépi és manuális kódolás egyezése szempontjából. A minta manuális elemzését minden modul, illetve elemzési kategória esetében két vagy három független kódoló végezte írott kódolási instrukció alapján, míg egy további kódoló hozott végső döntést azokban az esetekben, ahol a többi kódoló nem értett egyet az elemzési egység besorolását illetően. A kódolás során az egyes elemzési kategóriákat bináris változókként reprezentáltuk, ahol az 1 kód találatot jelzett az adott elemzési egység esetében, a találatok számától függetlenül, a 0 kód pedig a találat hiányát jelezte, kivéve az érzelmi polaritást, ahol egy-egy folytonos változóban a pozitív és negatív kifejezések számát jelölték a kódolók elemzési egységenként. A Politikai szereplők esetében hét bináris változót hoztunk létre a hét vizsgált pártnak megfelelően; a közösségiség és ágencia kategóriáin belül két-két változót a pozitív és negatív valenciának megfelelően (összesen tehát négy változót); az optimizmus-pesszizmus esetében hármat a három igeidőnek megfelelően; az individualizmus-kollektivizmus esetében pedig egyet-egyet a személyes névmások és a személyragos, illetve személyjeles szóalakok számára. A manuális elemzés végeredményeként kapott gold standard-et vetettük össze elemzési kategóriánként és elemzési egységenként az analóg módon kvantifikált gépi elemzési eredményekkel. A kapott eredmények az 1. táblázatban láthatók.

3 Kísérleti alkalmazások

A kidolgozott modulokhoz társított pszichológiai jelentések érvényességének ellenőrzéséhez politikai témájú közösségi médiában megjelent tartalmak szövegét elemeztük. A szövegkorpuszt 1341 Facebook oldal szövegeiből állítottuk össze, amely a pártokhoz és politikusokhoz köthető oldalak mellett a pártokat támogató személyek és csoportok Facebook oldalait is tartalmazta. A Facebook oldalakon megjelenő szövegeket 2013. október 1. és 2014. szeptember 21. között gyűjtöttük napi rendszerességgel.

1. táblázat: A modulok megbízhatóságának mutatói.

Modul	Nyelvi jegy	Találat	Pontosság	F1 érték
Individulizmus-kollektivizmus	Személyes névmás	0,6563	0,3520	0,4582
	Személyrag/-jel	0,9474	0,7727	0,8512
Optimizmus-pesszimizmus	Ige: múlt idő	0,9397	0,7890	0,8578
	Ige: jelen idő	0,9254	0,3140	0,4688
	Ige: jövő idő	0,6703	0,3280	0,4404
Közösségiség-ágencia	Közösségiség poz.	0,6575	0,3840	0,4848
	Közösségiség neg.	0,9639	0,4145	0,5797
	Közösségiség össz.	0,8205	0,4025	0,5401
	Ágencia pozitív	0,6943	0,7059	0,5283
	Ágencia negatív	0,2551	0,6579	0,3676
	Ágencia összes	0,3670	0,6943	0,4802
Érzelmi polaritás	Pozitív	0,7450	0,8256	0,7738
	Negatív	0,5368	0,6703	0,5962
	Összes	0,6244	0,7451	0,6794
Politikai szereplők	-	0,5714	0,9836	0,7229

A gyűjtés eredményeképpen 141,825 poszt és ezekhez kapcsolódóan 1 939 356 komment került be a szövegkorpuszba. A kommentek teljes terjedelme 46 211 723 token volt. A szövegkorpusz a 7 legnagyobb párthoz kapcsolódó szövegeket tartalmazza a következő arányban: FIDESZ-KDNP 25,2%, EGYÜTT-2014 19,3%, JOBBIK 19,2%, MSZP 16,6%, DK 12,5%, PM 4,2%, LMP 2,9%. A szövegelemzés lépéseinek leírását lásd Miháltz és munkatársai jelen kötetben szereplő tanulmányában [15]. Az elemzés során a modulok futtatásainak eredményeit havonta összegeztük és így vetettük össze a Tárki által mért közvéleménykutatási adatokkal [25], amelyek szintén havi bontásban álltak rendelkezésünkre. A közvéleménykutatási adatok összevontan tartalmazzák az Együtt-PM támogatására vonatkozó adatokat, így e két párt esetében a szövegelemzési adatokat is összevontuk. A közvéleménykutatási adatokból a biztos szavazó pártválasztó személyek adataival számoltunk, mivel azt feltételezzük, hogy a politikai pártok közösségi médiafelületein elsősorban az elkötelezett pártszimpatizánsok kommunikálnak. A pártpreferenciát a párt népszerűségét mutató adatként értelmeztük, erre a továbbiakban röviden népszerűségként utalunk.

Az Individualizmus és Optimizmus modulok pszichológiai érvényességét kvantitatív és kvalitatív módon is megvizsgáltuk. A kvantitatív elemzés során hipotéziseket fogalmaztunk meg az Individualizmus és az Optimizmus mutatók illetve a pártok népszerűségére vonatkozó adatok között. Mindkét esetben az a hipotézisünk, hogy pozitív összefüggést találunk. Az Individualizmus mutató esetében hipotézisünk alapja az, hogy az individualizmus magasabb szintje esetén a személyek nagyobb felelősséget

tulajdonítanak a pártválasztásnak, ami összességben nagyobb népszerűséget eredményez. Az Optimizmus mutató esetében hipotézisünk alapja az, hogy az optimizmus magasabb szintje esetén a személyek könnyebben hoznak döntést arról, hogy melyik pártot támogatják, ami szintén nagyobb népszerűséget eredményez. A hipotézisek teszteléséhez korrelációs elemzést végeztünk. Az individualizmus hipotézist támogatják az eredmények, mivel közel szignifikáns pozitív korreláció jelentkezik az Individualizmus 1 mutatóval (r=.22; p=.052). A tanulmány megírásakor nem rendelkeztünk adattal az Individualizmus 2 mutatóra vonatkozóan. Az Optimizmus 1 mutató esetében azonban bár szintén közel szignifikáns erősségű korreláció jelentkezik, ennek előjele negatív (r=.22; p=.055). Az Optimizmus 2 mutató esetén nem szignifikáns a korreláció. A negatív korrelációt az magyarázhatja, hogy a pártválasztásban a személyek múltbeli tapasztalatai is fontos szerepet kapnak.

Az elemzés kvalitatív részében azt vizsgáltuk meg, hogy a parlamenti választás előtt és után hogyan változik az Individualizmus 1 és az Optimizmus 1 mutató értéke. Mindkét változó esetében közvetlenül a választást követő időszakban láthatunk jelentős változást. Az Individualizmus 1 mutató értéke megemelkedik 2014 áprilisában. Ezt a változást az magyarázhatja, hogy a választással csökken az összefogás jelentősége a politikai csoportokba szerveződő személyeknél. Az Optimizmus 1 mutató értéke szintén áprilisban mutat változást. A FIDESZ-KDNP-hez köthető honlapokon növekszik a mutató értéke, a többi párt esetében azonban csökken. A változást a siker és kudarc megtapasztalása magyarázhatja: siker esetén nő az optimizmus, kudarc esetén csökken. Az elemzés eredményei az mutatják, hogy az Individualizmus 1 és az Optimizmus 1 mutatók valóban a politikai témában véleményt formáló személyek gondolkodásának jellemzőit mérik: a gondolkodás individualista fókuszának mértékét és a gondolkodás optimista voltát.

A közösségiség és ágencia mutatóinak népszerűséggel (párttámogatottsággal) mutatott összefüggéseire vonatkozó hipotéziseinket a vonatkozó szociálpszichológiai szakirodalom két megfigyelésére alapoztuk. A szociálpszichológiában klasszikus jelenség a (1) csoportközi elfogultság, mely a pozitívan értékelt csoportidentitás igényéből fakad: a csoporttagok hajlamosak, különösen csoportközi konfliktus vagy versengés helyzeteiben a saját csoportjukat elfogult módon felülértékelni, míg a külső csoportokat leértékelik. Ezt az aszimmetriát számos terep- és laboratóriumi kísérletben demonstrálták [14,19,21,22,23,24]. A társas ítéletalkotás kutatói leírták, hogy míg (2) a társakat, illetve a külső csoportok tagjait elsősorban közösségiség szempontjából értékeljük, addig a szelfet és a saját csoport tagjait elsődlegesen ágencia szempontjából ítéljük meg [1,2]. A két megfigyelést integrálva a vizsgált nyelvi mutatókra vonatkozóan azt vártuk, hogy az ágencia pozitív tartalmai és a közösségiség negatív tartalmai jelentős negatív korrelációt mutatnak a népszerűséggel: a csökkenő vagy alacsony támogatottság olyan fenyegető helyzet a csoportidentitás szempontjából, melynek kompenzálására a pártszimpatizánsok körében felerősödik a saját csoport felértékelése (több pozitív ágencia tartalom) és a külső csoportok leértékelése (több negatív közösségiség tartalom).

Az elemzett minta összesített havi adataira vonatkozóan hat mutatót alakítottunk ki: négy a pozitív és negatív közösségiség (K+%, K-%) és ágencia (Á+%, Á-%) százalékos arányait mutatja (kategória nyers gyakorisága / tokenek száma × 100), egy-egy ezekre épülő további mutató pedig a közösségiség és az ágencia tartalmainak összesí-

tett valenciáját (V_K, $V_Á$), amely egy 1 és -1 közötti arányszám, és azt mutatja, hogy adott hónapban adott pártra vonatkozóan mennyire dominálnak a vizsgált kategória pozitív vagy negatív tartalmai. A közösségiségre vonatkozóan $V_K = (K+\%-K-\%)/(K+\%+K-\%)$. Ezzel analóg kalkulussal számítottuk ki az ágencia összesített valenciáját ($V_Á$) is.

A 2. táblázatban láthatók a nyelvi mutatók átlagos értékei a teljes mintára, valamint a választások előtti és utáni hat-hat hónapra vonatkozóan. Mindhárom mintára érvényes, hogy a közösségiség esetében a negatív tartalmak (K-%) dominálnak, ezzel szemben az ágencia esetében a pozitív tartalmak (Á+%) túlsúlya figyelhető meg. Ez a csoportközi elfogultság feltevése alapján arra utal, hogy közösségiség szempontjából valóban elsősorban más pártokat (külső csoportokat) értékelnek az egyes pártok szimpatizánsai, míg ágencia szempontjából inkább a saját pártjukat (csoportjukat). A választások előtti és utáni időszakok között csak a pozitív ágencia átlaga mutat jelentős változást, negatív irányban, ami valószínűleg azzal függ össze, hogy a választások előtti kiélezett versenyhelyzetben fontosabb téma a hatalom megszerzése, illetve megtartása, mint azt követően. Megjegyzendő, hogy pozitív ágencia nem kizárólag értékelő jellegű megállapításokban fordulhat elő, hanem pl. célt vagy várakozást megfogalmazó közlésekben is.

2. táblázat. A pozitív és negatív közösségiség (K+%, K-%) és ágencia (Á+%, Á-%) százalékos arányainak átlagai, valamint a tartalom szerint összetartozó átlagok közötti különbségek (K+%-K-%, Á+%-Á-%) a teljes mintában, a választások előtti és utáni hat hónapban, valamint a két időszak átlagainak különbségei.

	K+%	K-%	K+%-K-%	Á+%	Á-%	Á+%-Á-%
Teljes minta (n=72)	.3507	.4291	-.0784***	.6966	.2531	.4434***
Választások előtt (n=36)	.3581	.4463	-.0883***	.7349	.2458	.4891***
Választások után (n=36)	.3433	.4118	-.0686**	.6582	.2605	.3978***
Vál. előtt-után (n=36)	.0148	.0345	-	.0767**	-.0147	-

Megjegyzés: ** $p < 0,01$; *** $p < 0,001$ a t-próba eredménye szerint (összefüggő mintás próbát alkalmaztunk a pozitív és negatív mutatók közti különbség tesztelésére, független mintás próbát a választások előtti és utáni értékek összehasonlítására).

A népszerűség és a közösségiség/ágencia mutatói közötti korrelációk a 3. táblázatban láthatók. A teljes mintát tekintve csak a pozitív ágencia mutat jelentős együttjárást, amely várakozásunknak megfelelően fordított irányú: alacsonyabb népszerűség mellett magasabb a pozitív ágencia százalékos aránya (Á+%) az adott hónapban adott párthoz tartozó Facebook oldalakra írt kommentekben, és vice versa. Csak a választások előtti időszakot vizsgálva ugyanez az összefüggés jelentkezik, valamint az ágencia összesített valenciája ($V_Á$) igen magas negatív korrelációt mutat a népszerűséggel. Az elfogult ítéletek hatása mellett elképzelhető, hogy a motivációval, kompetenciával, produktivitással és kontrollal kapcsolatos pozitív tartalmak a na-

gyobb támogatottság igényét is jelzik a pártszimpatizánsok diskurzusában. E feltevés ellenőrzéséhez a leggyakrabban előforduló tartalmak és referenciáik, illetve kontextusuk elemzése szükséges. A választások utáni időszakban a népszerűség összefüggése az ágenciával nem mutatkozik, ugyanakkor a közösségiség összesített valenciájával (V_K) és különösen negatív tartalmaival (K-%) magas negatív korrelációt mutat. Ez az eredmény szintén megfelel előzetes feltevésünknek: minél kisebb egy párt népszerűsége, annál erőteljesebben jelentkezik más pártok (külső csoportok) leértékelése a közösségiség negatív tartalmaiban, amely a vártnál kevésbé sikeres csoport (párt) fenyegetett identitásának védelmét és a csoportkohéziót szolgálja.

3. táblázat: A népszerűség korrelációja a közösségiség és ágencia mutatóival. (K+%, K-%, Á+%, Á-%: pozitív/negatív közösségiség/ágencia százalékos aránya; V_K, $V_Á$: közösségiség/ágencia összesített valenciája)

	K+%	K-%	V_K	Á+%	Á-%	$V_Á$
Teljes minta (n=54)	-.015	-.186	.079	-.328*	-.050	-.228
Választások előtt (n=30)	-.143	.219	-.281	-.429*	.259	-.677**
Választások után (n=24)	.115	-.574**	.454*	-.288	-.305	.146

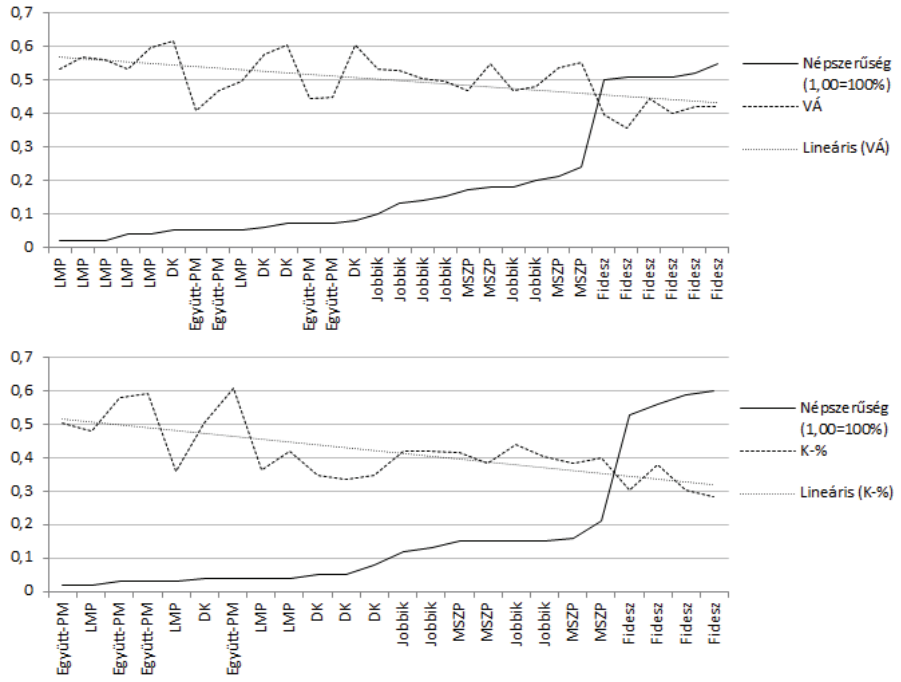
Megjegyzés: * $p < 0,05$; ** $p < 0,01$

Az 1. ábra korrelációs görbéi a népszerűség együttjárását mutatják az ágencia összesített valenciájával ($V_Á$) a választások előtti hat hónapban (fent) és a negatív közösségiséggel (K-%) a választások utáni hat hónapban (lent). Az adatokat népszerűség szerint növekvő sorba rendeztük. A vízszintes tengely azt jelzi, hogy a népszerűség és a nyelvi mutatók egyes havi adatai melyik párthoz tartoznak. A hiányzó havi adatok oka, hogy a Tárki kimutatásából bizonyos havi adatok hiányoznak. A nyelvi mutatókon áthaladó egyenesek a görbék lineáris trendvonalai. Az ábrákon látszik, hogy nem az egyes pártok egymástól eltérő, ugyanakkor konstans népszerűsége, illetve nyelvi jellemzői határozzák meg a magas korrelációkat, mivel a népszerűségi adatok nem pártok szerint állnak sorba, eltekintve a Fidesz kiugró népszerűségétől. Ugyanakkor az is látszik, hogy ez a kiugró népszerűség valójában rontja a korrelációt: kisebb emelkedés esetén nagyobb lenne a trendvonalakkal mutatott szimmetria. Mindez alátámasztja a népszerűség és a nyelvi mutatók közötti összefüggést meghatározó dinamikára vonatkozó feltevéseinket.

Eredményeink szerint az Individualizmus, Optimizmus, Közösségiség és Ágencia modulok pszichológiai szempontból valid elemzőeszközök, felhasználhatók politikai csoportok közösségi médiafelületein megjelenő véleményváltozás felismerésére.

Köszönetnyilvánítás

A fejlesztés a Trendminer Project támogatásával valósult meg.



1. ábra. A népszerűség korrelációja az ágencia összesített valenciájával ($V_{Á}$) a választások előtti hat hónapban és a negatív közösségiséggel (K-%) a választások után.

Hivatkozások

1. Abele, A. E., Wojciszke, B.: Agency and communion from the perspective of self versus others. Journal of Personality and Social Psychology 93/5 (2007) 751–763
2. Abele, A. E., Cuddy, A. J. C., Judd, C. M., Yzerbyt, V. Y.: Fundamental dimensions of social judgment. European Journal of Social Psychology 38/7 (2008) 1063–1065
3. Cuddy, A. J. C., Fiske, S., Glick, P.: Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS Map. Advances in Experimental Social Psychology 40 (2008) 61–149
4. Csertő, I., László, J.: Intergroup evaluation as an indicator of emotional elaboration of collective traumas in national historical narratives. Sociology Study 3/3 (2013) 207–224
5. Dubois, N., Beauvois, J. L.: Normativeness and individualism. European Journal of Social Psychology 35/1 (2005) 123–146
6. Ferenczhalmy, R., Szalai, K., László, J: Az ágencia szerepe történelmi szövegekben a nemzeti identitás szempontjából. Pszichológia 31/1 (2011) 35–46
7. Fiske, S. T., Cuddy, A., Glick, P.: Universal dimensions of social cognition: Warmth and competence. Trends in Cognitive Sciences 11/2 (2007) 77–83
8. Habermas, T., Ott, L. M., Schubert, M., Schneider, B., Pate, A.: Stuck in the Past: Negative Bias, Explanatory Style, Temporal Order, and Evaluative Perspective in Life Narratives of Clinically Depressed Individuals. Depression and Anxiety 25/11 (2008) 1091–4269

9. Hofstede, G.: Culture's consequences. Beverly Hills, CA, Sage (1980)
10. Kashima, E. S., Kashima,Y.: Culture and language: The case of cultural dimensions and personal pronoun use. Journal of Cross-Cultural Psychology 29 (1998) 461–486
11. Kunda, Z.: Social cognition. Making sense of people. Cambridge, MA, MIT Press (1999)
12. László, J.: Történelemtörténetek. Bevezetés a narratív szociálpszichológiába. Budapest, Akadémiai Kiadó (2012)
13. László, J., Csertő, I., Fülöp, É., Ferenczhalmy, R., Hargitai, R., Lendvai, P., Péley, B., Pólya, T., Szalai, K., Vincze, O., Ehmann, B.: Narrative Language as an Expression of Individual and Group Identity: The Narrative Categorical Content Analysis. SAGE Open 3/2 (2013) 1–12
14. Maass, A., Salvi, D., Arcuri, L., Semin, G.: Language use in intergroup contexts: the linguistic intergroup bias. Journal of Personality and Social Psychology 57/6 (1989) 981–993
15. Miháltz, M., Váradi, T.: Trendminer: politikai témájú közösségimédia-üzenetek feldolgozása és szociálpszichológiai elemzése. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (2015), ld. jelen kötetben
16. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol. 10 (2002) 79–86
17. Peeters, G.: Evaluative meanings of adjectives in vitro and in context: Some theoretical implications and practical consequences of positive-negative asymmetry and behavioral-adaptive concepts of evaluation. Psychologica Belgica 32/2 (1992) 211–231
18. Petrides, K. V., Furnham, A.: The role of trait emotional intelligence in a gender-specific model of organizational variables. Journal of Applied Social Psychology 36 (2006) 552–569
19. Pettigrew, F. T.: The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice. Personality and Social Psychology Bulletin 5/4 (1979) 461–476
20. Rosenberg, S., Nelson, C., Vivekananthan, P. S.: A multidimensional approach to the structure of personality impressions. Journal of Personality and Social Psychology 9/4 (1968) 283–294
21. Sherif, M., Harvey, O. J., White, J., Hood, W., Sherif, C.: Intergroup Conflict and Cooperation: The Robber's Cave Experiment. Norman, University of Oklahoma, Institute of Social Relations (1961)
22. Sherif, M.: In Common Predicament: Social Psychology of Intergroup Conflict and Cooperation. Boston, Houghton Mifflin (1966)
23. Szabó, Zs. P., Banga, Cs., Ferenczhalmy, R., Fülöp, É., Szalai, K., László, J.: A nyelvbe kódolt társas viszonyok. Az implicit szemantika szociálpszichológiai kutatása. Pszichológia 30/1 (2010) 1–16
24. Tajfel, H.: Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations. New York, NY, Academic Press (1978)
25. Tárki közvéleménykutatási adatok: http://www.tarki.hu/hu/research/elect/gppref_table_03.html. Letöltve: 2014.10.24.
26. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proc. of the ACL (2002)
27. Wojciszke, B.: Morality and competence in person and self perception. European Review of Social Psychology. 16/1 (2005) 155–188
28. Ybarra, O., Chan, E., & Park, D.: Young and old adults' concerns about morality and competence. Motivation and Emotion. 25/2 (2001) 85–100

Doménspecifikus polaritáslexikonok automatikus előállítása magyar nyelvre

Hangya Viktor, Farkas Richárd

Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{hangyav,rfarkas}@inf.u-szeged.hu

Kivonat Napjainkban a közösségi média jelentős népszerűségre tett szert, szinte bármilyen témakörben nagy mennyiségű szöveg érhető el. Ennek köszönhetően nagy figyelmet kaptak a különböző véleménydetekciós módszerek, melyek feladata szövegek osztályozása azok tartalmának polaritása alapján. A feladat megoldása során segítséget nyújtanak az ún. polaritáslexikonok, melyek az egyes szavak polaritására nézve hordoznak információkat. Munkánkban bemutatunk különböző módszereket lexikonok előállítására, valamint azok kiegészítésére és adaptálására más doménekre. Vizsgálatainkat kifejezetten számítástechnikai eszközökkel kapcsolatos véleményeken és általános hírekből származó szövegeken végeztük el, melyekből kiderül, hogy az osztályozás pontosságára nézve a megfelelő lexikon kiválasztása meghatározó.

Kulcsszavak: véleménydetekció, lexikon, természetesnyelv-feldolgozás

1. Bevezető

A közösségi média elterjedésével az embereknek lehetőségük nyílt különböző tartalmak megosztására más internetfelhasználókkal. Ennek köszönhetően nagy mennyiségben érhetőek el olyan elektronikus tartalmak, melyekben az egyes felhasználók véleményeiket fejezik ki bizonyos termékekről, ismert személyiségekről vagy cégekről. Az utóbbi években nagy figyelmet kaptak a különböző automatikus véleménydetekciós módszerek [1,2], melyeknek feladata dokumentumok polaritásának meghatározása.

A feladatra tekinthetünk úgy, mint egy osztályozási probléma, mely során egy dokumentumot pozitív vagy negatív osztályba kell sorolni. Nagy mennyiségű jelölt adat esetén egy jól bevált módszer a szövegekben előforduló szavak, esetleg szópárosok alapján történő felügyelt gépi tanulás alapú osztályozás. Ilyenkor az egyes szavak polaritását a tanító adat segítségével határozzuk meg. A módszer azonban nem kellően pontos abban az esetben, ha kevés számú tanító adat áll rendelkezésünkre, mivel az egyes szavak polaritásának megbecslése pontatlan lehet. Másik probléma, hogy sok szó nem fordul elő a tanító szövegekben, így azok polaritásáról nem tudunk mondani semmit. Az ún. polaritáslexikonok nyújtanak segítséget ilyen esetekben, melyek az egyes szavak polaritását előre adott módon tartalmazzák, így a tanítás során nem látott szavak polaritását tekintve is vannak információink. Angol nyelvre számos általános célú lexikon érhető el, magyar nyelv esetén azonban nem ilyen kedvező a helyzet.

Egyik kézenfekvő lehetőség a már létező, idegen nyelvű lexikonok lefordítása. A módszernek azonban több hátránya is van. Egyrészt a folyamat időigényes és költséges. A többjelentésű szavak különböző jelentései más-más polaritással rendelkezhetnek az egyes doménekben, ezért szükség van doménspecifikus lexikonokra, melyek a kérdéses szövegek esetén jól használhatóak. Az idegennyelvű lexikonok lefordításának másik nehézsége az, hogy az elérhető lexikonok általános célúak. Munkánkban megmutatjuk, hogy az ilyen általános célú lexikonok nem teljesítenek jól speciális doménekben. Olyan módszereket dolgoztunk ki, melyek lehetőséget nyújtanak doménspecifikus lexikonok létrehozására. A javasolt módszerek egy része kicsi, kézzel összeállított lexikon kiegészítésére alkalmas, ahol a kulcs hasonló jelentéssel bíró szavak automatikus gyűjtése. Ezen felül adott doménből származó szöveges adatok segítségével tejesen új lexikont hoztunk létre automatikus módszerrel.

Vizsgálatainkat számítástechnikai eszközökkel kapcsolatos véleményeken és általános hírekből [3] származó szövegeken is elvégeztük. Az eredményekből kiderül, hogy az osztályozás pontossága nagyban függ attól, hogy a felhasznált lexikon mely doménből származik. Az általános lexikonokkal ellentétben, a doménspecifikusak használata szignifikáns hibacsökkenést eredményez.

2. Polaritáslexikon

Nagy jelölt adathalmaz esetén lehetőség van az egyes szavak polaritásának meghatározására gépi tanulási módszerek segítségével. Ezt a tudást felhasználva korábban még nem látott szövegek esetén eldönthető azok véleménytartalma. Abban az esetben azonban, ha a jelölt adatok száma kicsi, ezek a módszerek pontatlanabbá válnak. Ilyenkor érdemes használni polaritáslexikonokat, melyek tartalmazzák a polaritással rendelkező szavakat. Ez az előtudás felhasználható az osztályozás során például arra, hogy megtudjuk, hány darab pozitív, illetve negatív szó szerepel az adott dokumentumban.

Figyelembe kell azonban venni azt a tényt, hogy az egyes szöveghalmazok különböző doménekből származhatnak. Bizonyos szavak doménenként eltérő polaritással rendelkeznek. Tekintsük a következő példákat:

– A mixer használata egyszerű és **halk** működés közben.
– Ennyi pénzért nekem túl **halk**, nagyobb hangerőre számítottam.

Az első mondat konyhai eszközökkel kapcsolatos és a *halk* szó pozitív töltetű. Ezzel szemben, a második mondat egy hangszóró értékelése, mely során a szó már negatív polaritású. Fontos tehát, hogy a véleménydetekciós feladat elvégzéséhez a megfelelő doménből származó lexikont használjuk. Mivel legtöbb esetben nem áll rendelkezésre a megfelelő lexikon és azok manuális előállítása költséges, statisztikai módszerekkel újakat kell létrehozni vagy meglévőeket kell adaptálni.

A következőkben bemutatásra kerülnek különböző megoldások lexikon létrehozására, illetve kiegészítésére, valamint ezek problémáira is rávilágítunk.

2.1. Idegen nyelvű lexikon lefordítása

Sajnos magyar nyelvre nem áll rendelkezésre szabadon elérhető (referencia) polaritáslexikon. Angol nyelvben azonban több általános célú is megtalálható [4,5]. Így egy értelemszerű megoldás egy meglevő lexikon lefordítása. Ehhez egy már angol nyelven működő véleménydetekciós rendszerben használt lexikont vettünk alapul, mely 3322 szót tartalmaz. A lexikonban az egyes szavak a $[-5, 5]$ intervallumon vett értékkel vannak jellemezve polaritásuktól függően. Az angol szavak fordítását kézzel végeztük és az adott szó összes magyar megfelelőjét felvettük a **fordított** lexikonba.

A módszernek azonban több hátránya is van. Először is a fordítás költséges és időigényes, főként abban az esetben, ha a lefordítandó lexikon mérete nagy. Többjelentésű szavak esetében sokszor nem egyértelmű, hogy melyik jelentés használandó. Például a *terrific* szó (*nagyszerű*, *szörnyű*) két ellentétes polaritású jelentéssel rendelkezik. Az eredeti lexikonban lévő érték segítségével kiválasztható a megfelelő jelentés. A *cool* szó (*hűvös*, *menő*) esetében azonban már nehezebb a döntés, mivel mindkettő jelentés lehet pozitív töltetű. Másik probléma, hogy az elérhető idegen nyelvű lexikonok általános célúak, melyek a kívánt doménben nem megfelelően használhatóak a már ismertetett okok miatt. Így fordítás során figyelembe kell venni ezt a tényt is, bizonyos esetekben szükség van az eredeti polaritási értékek megváltoztatására.

2.2. Polaritáslexikon bootstrapping módszerrel

A fent említett problémák megoldására több automatikus módszert dolgoztunk ki. Az első lexikonok létrehozására alkalmas jelölt szöveghalmaz segítségével. Egy adott w szó polaritása a következő módon számítható [6]:

$$pol(w) = PMI(w, pozitív) - PMI(w, negatív) \tag{1}$$

ahol PMI a páronkénti kölcsönös információt jelenti az adott polaritásra nézve. Számítása a következő képlettel adható meg:

$$PMI(w, p) = log_2 \frac{freq(w, p) * N}{freq(w) * freq(p)} \tag{2}$$

ahol $p \in \{pozitív, negatív\}$ a kérdéses polaritás, $freq(w, p)$ megadja a w szó p polaritású szövegekben való előfordulásainak számát, $freq(w)$ és $freq(p)$ a w szó összes előfordulásainak száma, illetve a p polaritású szövegek száma a korpuszban, N pedig a különböző szóalakok száma. Ezzel a módszerrel a korpuszban előforduló összes szóalakra kapunk egy polaritási értéket, mely a szavak adott doménen belüli megfelelő polaritását fogják tükrözni. Az értékeket úgy skáláztuk, hogy azok a $[-5, 5]$ intervallumba essenek. A továbbiakban ezzel a módszerrel előállított lexikonokra a **pmi** névvel fogunk hivatkozni.

Szükség van azonban a kérdéses doménből vagy egy ahhoz hasonlóból származó jelölt adathalmazra, melyen a statisztikai számítások elvégezhetőek. Abban az esetben, ha csak jelöletlen adathalmaz áll rendelkezésre, egy pontatlanabb

módszerrel osztályozhatjuk azt, majd ezen az automatikusan jelölt adathalmazon számíthatunk polaritáslexikont. Munkánk során végrehajtottunk egy ilyen kísérletet is, melynek eredményeit a 4. fejezetben részletezünk.

2.3. Lexikon kiterjesztése

A következőkben olyan módszereket mutatunk be, melyek kis számú szóval rendelkező lexikonokból kiindulva, bizonyos összefüggéseket felhasználva egy kiterjesztett lexikont adnak eredményül. A módszerek alapja, hogy az egyes elemekhez hasonlóan viselkedő szavakat gyűjtünk. Ily módon egyrészt a lexikonban levő szavak számát tudjuk növelni, másrészt pedig lehetőség van a lexikon domén adaptálására.

A módszerek bemenetként egy kiindulási lexikont kapnak. Mivel ilyen doménspecifikus lexikon nem létezik, ezért ennek előállítására egy félautomata módszert alkottunk meg. Első lépésként egy szóelőfordulás alapú maximum entrópia osztályozót tanítottunk a szöveges adathalmazon. A tanult modellből lehetőség van kinyerni, hogy az egyes szavak mennyire jellemzőek a pozitív, illetve negatív polaritású szövegekre. Ezt felhasználva kézzel kigyűjtöttük a legjellemzőbbeket (kb. 20 darab szó polaritásonként már elegendő). Az így gyűjtött szavakat használtuk kiindulási (**seed**) lexikonként (5, illetve -5 értékekkel), mely ily módon a kérdéses doménre jellemző szavakat tartalmazza.

WordNet. A *wordnet* egy olyan lexikai adatbázis, mely a szavakat szinonimahalmazokba sorolja, valamint ezek kapcsolatait is tartalmazza. Első módszerünk a magyar wordnetben [7] található információk alapján egészíti ki a megadott lexikont. Feltételezhetjük, hogy a kiindulási lexikonban található egyes szavak és azok szinonimáinak polaritása megegyezik. Így a kiterjesztett lexikonba felvettük ezeket a szinonimákat a megfelelő polaritási értékekkel. Előfordulhat azonban, hogy egy szó több szinonimahalmazba is tartozhat, így az többször is belekerül a kiterjesztett lexikonba, akár különböző polaritási értékekkel. Ezt úgy kezeltük, hogy az adott szó polaritási értékeinek átlagát számítottuk ki. A szinonimákon felül felhasználtuk azt az információt is, hogy mely szinonimahalmazok jelentése hasonló vagy ellentétes. Felvettük a kiterjesztett lexikonba az összes olyan szót, amely egy adott, a kiindulási lexikonban levő szó szinonimahalmazához hasonló (*similar_to*, *hyponym*[1]) vagy ellentétes (*near_antonym*) jelentéssel bíró halmaz eleme, mégpedig a hasonló jelentéssel bíró szavakat az eredeti polaritási értékkel, míg az ellentétes jelentéssel bíróakat negált értékkel. A módszer iteratívan futtatható, azaz az egyik lépésben eredménye a következő bemeneteleként használható, így tovább bővítve a lexikont. Figyelembe kell azonban venni azt a tényt, hogy az egyes kiterjesztési lépések során egyes szavak rossz polaritással is bekerülhetnek. Példa erre a számítástechnikai doménben pozitív szónak számító *csendes* alapján bekerülő *elzárkózott* szó, melyet negatív polaritással kellene felvenni. Mivel a wordnet egy általános nyelvi erőforrás, a benne előforduló szóhasonlóságok

[1] A hyponym kapcsolat nincs felvéve a wordnetben, a hypernym kapcsolat megfordításával határozható meg.

nem doménspecifikusak. Példa erre a *hangos* és az *erős* szavak, melyek nem szinonimák a wordnetben, de annak tekinthetők hangszórók véleményezése során. Emiatt a módszer egy más doménből származó kiindulási lexikont nem képes adaptálni. Abban az esetben azonban, ha a kiindulási lexikon doménspecifikus, a kiterjesztés során az adott doménre jellemző szavak és azok értékei alapján fog megtörténni.

Szavak klaszterezése. Hasonló szavak nem csak wordnet segítségével határozhatóak meg, hanem szövegek statisztikai elemzésével is. A következő módszerünkben az ún. Brown klaszterező eljárást alkalmaztuk [8], mely egy hierarchikus klaszterező eljárás. Az előző módszerhez hasonlóan, feltettük, hogy az egy klaszterbe tartozó szavak polaritása megegyező. Ennél fogva a kiterjesztett lexikonba felvettünk minden olyan szót, mely egy klaszterbe tartozik valamely a kiindulási lexikonba lévő szóval, annak megfelelő polaritási értékkel. Ennél a módszernél is bekerülhetnek szavak a lexikonba rossz polaritási értékkel, abban az esetben, ha egy klaszterbe kerülnek nem hasonlóan viselkedő szavak. Ennek egyik oka, ha túl kevés klasztert definiálunk, így sok szó kerül egy csoportba. Ellentétben a wordnettel, ennél a módszernél az egyes szavak hasonlósága a domén jellemzői alapján vannak meghatározva. Így a kiindulási lexikont (legyen szó doménspecifikusról vagy nem doménspecifikusról) kiterjesztés során az adott doménhez adaptáljuk.

3. Korpuszok

Ebben a fejezetben bemutatásra kerülnek a munkánk során felhasznált szöveges adatbázisok. Kísérleteinket általánosabbnak mondható és doménspecifikus szöveghalmazokon is elvégeztük.

3.1. Általános szövegek

Kifejezetten véleménydetekciós feladatok elvégzésére létrehozott szöveges adatbázis az *OpinHuBank* [3], mely tartalma különböző magyar nyelvű híroldalak, blogok és fórumok alapján lett összeállítva. Minden egyes szövegpéldány legalább hét token hosszú mondat, melyet öt annotátor pozitív, negatív vagy semleges kategóriába sorolt. Kísérleteinkhez csak a pozitív és negatív véleménytartalmú szövegekre volt szükségünk, ezért kiválogattuk azokat, melyek legalább három pozitív vagy negatív jelölést kaptak. Az így kapott **opinhu** adatbázis 882 pozitív és 1629 negatív mondatot tartalmazott. Az egyes mondatok esetében további információ az, hogy azok véleménytartalma kire vonatkozik, ezt azonban mi nem használtuk fel.

3.2. Doménspecifikus szövegek

Módszereinket doménspecifikus szöveghalmazon is elvégeztük. Ehhez létrehoztunk egy adatbázist, melyek számítástechnikával kapcsolatosak. Ezt az áruke-

reső[2] oldal tartalma alapján állítottuk össze. Az oldalon számos termékről találhatók értékelések, melyek közül a *számítógép* és *műszaki cikk* kategóriákba esőeket töltöttük le. Az egyes vélemények írása során a véleményezőnek meg kell adni szövegesen az adott termék előnyeit és hátrányait, melyeket rendre pozitív és negatív szövegeknek tekintettünk. Ezek a vélemények eltérő hosszúságúak, több mondatból, de akár pár szóból is állhatnak. Tesztelésünkhöz a letöltött vélemények közül kiválogattuk azokat, melyek egy mondatból állnak, azaz legalább négy token hosszúak és megfelelő mondatzáró jelet tartalmaznak. Az így összeálló **árukereső** adatbázis 3573 pozitív és 3149 negatív mondatot tartalmaz.

4. Eredmények

Cikkünk fő célja különböző lexikonok építése és azok használhatóságának összehasonlítása véleménydetekciós problémákon. Ehhez egy kétosztályos osztályozási feladatot definiáltunk, mely során a szövegeket pozitív vagy negatív címkével láttunk el. Az osztályozást maximum entrópia osztályozó segítségével végeztük el. Az egyes tanító és tesztpéldák esetében jellemzőként a bennük előforduló szavak szótöveit és a polaritási lexikonok alapján kinyert információkat használtuk. Egy lexikont annál jobbnak tekintünk egy adott szöveghalmazra nézve, minél nagyobb pontosságot érünk el segítségével a mondatok polaritásalapú osztályozása során. A következőkben polaritásos szavaknak tekintjük azokat a szavakat, melyek szerepelnek az adott lexikonban és a hozzájuk tartozó érték abszolútértéke legalább egy. Az adott polaritásos szó értékének előjelétől függően pozitív, illetve negatív a szó. A lexikonok alapján kinyerhető jellemzők a következők, melyekre egy példa látható az 1. táblázatban:

- szövegben szereplő polaritásos szavak (eredeti alakban)
- szövegben szereplő pozitív, illetve negatív szavak értékeinek összege különkülön
- szövegben szereplő polaritásos szavak értékeinek összege
- polaritásos szavak polaritásából és azokat megelőző, illetve követő szavak szótöveiből képzett párosok

1. táblázat. Példamondat és az abból kinyert jellemzők. A mondatban szereplő polaritásos szó a *jobb*, mely 5.0 értékkel szerepel a lexikonban.

Mondat:	A laptop kijelzőjének **jobb** paraméterei vannak!
Szótövek:	a, laptop, kijelző, jó, paraméter, van, !
Polaritásos szavak:	jobb
összértékek:	POZITÍV=5.0, NEGATÍV=0.0, POLARITÁSOS=5.0
szomszédság:	kijelző_POZITÍV, POZITÍV_paraméter

[2] `www.arukereso.hu`

2. táblázat. Seed lexikonok kiegészítése wordnet (wn) és klaszter alapú módszerekkel. Az első táblázat az opinhu, a második pedig az árukereső adatbázisokon mért pontosság, tízszeres keresztvalidációval.

opinhu-seed	86.2
opinhu-seed-wn-1	86.4
opinhu-seed-wn-2	85.9
opinhu-seed-wn-3	86.3
opinhu-seed-wn-4	86.0
opinhu-seed-klaszter-15	86.7
opinhu-seed-klaszter-15-t3	86.8

árukereső-seed	90.7
árukereső-seed-wn-1	90.8
árukereső-seed-wn-2	90.5
árukereső-seed-wn-3	90.8
árukereső-seed-wn-4	90.9
árukereső-seed-klaszter-18	90.8
árukereső-seed-klaszter-19-t3	90.8

A 2. táblázatokban láthatóak a lexikonkiegészítő módszerek eredményei az opinhu és árukereső adathalmazokra. Mindkettő esetében az adott adatbázis alapján kinyert seed lexikont egészítettük ki. Az egyes értékek a helyesen eltalált osztálycímkék százalékos arányát adják meg tízszeres keresztvalidációval mérve. A táblázatokban a *wn* jelölés a wordnet alapú kiterjesztésre utal, az utána szereplő szám pedig az iteráció számot jelöli. Az opinhu adatbázis esetében legnagyobb hibacsökkenést az első iteráció eredményezte, míg az árukereső esetében a negyedik. Az ötödik iterációtól kezdődően az eredmények folyamatosan romlanak, ami annak köszönhető, hogy a lexikonok egyre zajosabbá válnak. A táblázatok utolsó két sorában a klaszterezésen alapuló lexikonkiegészítés eredményei láthatók. Az egyes számok a hierarchikus klaszterezés vágási szintjét adják meg, míg a *t3* jelölés megléte esetén a klaszterekből kiszűrtünk minden olyan szót, aminek előfordulása legfeljebb három. Az opinhu adatbázis esetén a 15 mélységben történő, míg az árukereső esetében a 18 és 19 mélységben történő vágás eredményezte a legnagyobb hibacsökkenést.

3. táblázat. Különböző lexikonok segítségével elért pontosság *opinhu* és *árukereső* adatbázisokon, tízszeres keresztvalidációval.

	opinhu	árukereső
unigram	86.1	90.0
opinhu-seed-klaszter-15-t3	86.8	90.1
árukereső-seed-wn-4	86.2	90.9
fordított	88.4	90.2
opinhu-pmi	96.3	90.0
árukereső-pmi	84.3	91.9
árukereső2-pmi	-	91.0

A 3. táblázatban láthatók a kiegészítés, fordítás és bootstrapping eljárások segítségével létrehozott lexikonok eredményei. Referencia-rendszer az *unigram*, mely esetében az osztályozáshoz nem használtunk polaritáslexikont. Látható, hogy a lexikonkiegészítéssel szignifikáns hibacsökkenés érhető el, abban az esetben, ha a megfelelő doménből származó lexikon került alkalmazásra. Ellenkező

esetben is sikerült javulást elérni, viszont csak kis mértékben. Az angolról magyarra fordított általános lexikon hatása látható a táblázat *fordított* sorában. Mivel az opinhu különböző újságcikkeket tartalmaz, melyek tartalma nem csak egy adott doménnek kapcsolatosak, így az általános célú lexikon nagy (2,3%) javulást eredményezett. Ezzel szemben az árukereső adathalmaz esetében a javulás sokkal kisebb mértékű, még a 2. táblázatban látható kis elemszámú *árukereső-seed* lexikon eredményeitől is elmarad.

A 3. táblázat *pmi* végződésű soraiban láthatók az eredmények, melyek során a bootstrapping eljárással összeállított lexikonokat használtuk. Azokban az esetekben, amikor a lexikont ugyanazon szöveghalmaz segítségével állítottuk össze, mint amelyiken később az osztályozást végeztük, a kapott lexikon az adatokra nézve legpontosabbnak tekinthető, azaz egyfajta elméleti maximumot adnak meg. Fontos, hogy ilyen lexikon csak tesztkörnyezetben állítható elő. Kísérletet tettünk azonos doménből származó jelöletlen adatok felhasználására. Ehhez az árukereső oldalról letöltött olyan szövegeket, melyek egy mondatnál rövidebbek vagy hosszabbak, az egymondatos árukereső adatbázison tanult modellel automatikusan címkéztük. Az így kapott *árukereső2* alapján egy újabb lexikont hoztunk létre, mely segítségével tovább növeltük a pontosságot. Mivel az árukereső2 szöveghalmaz méretben nagyobb, mint az árukereső, ezért a lexikonban levő szavak polaritása pontosabban lettek meghatározva.

5. Összefoglalás

Munkánk során szövegek osztályozását végeztük el pozitív és negatív osztályokba azok véleménytartalma alapján. Az egyszerű szóalapú osztályozók pontosságát polaritáslexikonok felhasználásával javítottuk. Célunk olyan módszerek kidolgozása volt, melyek lehetőséget nyújtanak lexikonok automatikus létrehozására és kiegészítésére (doménadaptálására). Megmutattuk, hogy megfelelő mennyiségű, a kérdéses doménből származó jelölt vagy akár jelöletlen adat segítségével létre tudunk hozni lexikonokat automatikus módszerekkel vagy kis elemszámú kiindulási lexikont tudunk kiegészíteni szavak hasonlósága alapján. Utóbbi módszer esetén figyelembe kell venni azonban azt a tényt, hogy a kiegészítés során zajossá válhat a lexikon, ezért egy további feladat lehet az irreleváns szavak szűrése. Eredményeinkből kiderül, hogy fontos a megfelelő doménből származó lexikon használata. Általános szövegeken automatikus módszerekkel létrehozott lexikonok segítségével javítottunk az eredményeken, a legnagyobb javulást azonban a szintén általános célú fordított lexikonnal sikerült elérni. Ebből látható, hogy a legpontosabb eredményeket manuálisan összeállított lexikonok segítségével érhetjük el. Ennek oka, hogy ezek a lexikonok egyrészt kevésbé zajosak, másrészt olyan többletinformációval rendelkeznek, melyek a szövegek alapján statisztikai módszerekkel nem határozhatóak meg. A kézi összeállítás azonban hosszadalmas, költséges és a kérdéses domén ismeretét igényli. Számítástechnikával kapcsolatos domén esetében az általános célú lexikon elenyésző javulást hozott, mivel ez a lexikon nem vagy más polaritással tartalmazza a doménre jellemző véleményt kifejező szavakat. Az automatikus módszerek képesek meg-

határozni a domén jellemzőit, így ezekkel a módszerekkel létrehozott lexikonok a hibák 10%-os csökkenését eredményezték.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Hangya, V., Berend, G., Farkas, R.: SZTE-NLP: Sentiment Detection on Twitter Messages. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). (2013) 549–553
2. Hangya, V., Berend, G., Varga, I., Farkas, R.: SZTE-NLP: Aspect level opinion mining exploiting syntactic cues. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 610–614
3. Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 343–345
4. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). (2010)
5. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics (2005) 347–354
6. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) **21**(4) (2003) 315–346
7. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of the Fourth Global WordNet Conference (GWC-2008). (2008)
8. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational linguistics **18**(4) (1992) 467–479

Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai

Szabó Martina Katalin[1,2], Vincze Veronika[3,4]

[1] Precognox Informatikai Kft.
[2] Szegedi Tudományegyetem, Orosz Filológiai Tanszék
mszabo@precognox.com; szabomartinakatalin@gmail.com
[3] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[4] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
vinczev@inf.u-szeged.hu

Kivonat: A jelen dolgozat egy magyar nyelvű kézzel annotált szentimentkorpusz létrehozásáról számol be. A korpusz építésének célja, hogy megfelelő segédletet teremtsünk a magyar nyelvű szövegek véleménykivonatolásával kapcsolatos nyelvtechnológiai feladatok, köztük a szentimentlexikonunk és az automatikus szentimentelemző rendszerünk hatékonyságának teszteléséhez és fejlesztéséhez. A korpusz emellett lehetőséget kíván nyújtani a magyar nyelvű szövegek szentimentelemzését érintő elméleti nyelvészeti problémák feltárására is, amely nélkülözhetetlen a szentimentelemző rendszer hatékony működésének biztosításához.

1 Bevezetés

A jelen dolgozatban a magyar nyelvű szövegek automatikus szentimentelemzését célzó kutatómunkánk egyik részfeladatáról, egy szentimentekre annotált korpusz létrehozásáról számolunk be.

A *szentimentelemzés* vagy *véleménykivonatolás* (*sentiment analysis* vagy *opinion mining*) a természetesnyelv-feldolgozás részterülete, amely a szerzői attitűdöt tükröző nyelvi elemek detektálására, valamint értékének (*sentiment orientation*) és tárgyának (*target*) a megállapítására törekszik automatikus megoldások segítségével.

A szentimentelemzés a nemzetközi kutatásban és fejlesztésben egyre nagyobb figyelmet kap, amelynek oka egyrészt a feladat elméleti nyelvészeti, valamint nyelvtechnológiai kihívásaiban, másrészt az eredmények gazdasági hasznosítási lehetőségeiben keresendő (pl. a tőzsdeindex mozgásának előrejelzése; a fogyasztói csoport benyomásai, tapasztalatai bizonyos termékek és szolgáltatások vonatkozásában; politikusokkal, politikai eseményekkel kapcsolatos attitűdök felmérése; választási előrejelzések stb.). Ugyanakkor, e növekvő nemzetközi figyelem ellenére a magyar nyelvű szövegek véleménykivonatolási feladatával csupán rendkívül csekély számú dolgozat foglalkozik. Emeljük ki közülük Berend és Farkas [1] dolgozatát, amely a kettős állampolgárság témájához kapcsolódó szövegek gépi tanuláson alapuló feldolgozását célozza, valamint az *Opinhu* rendszert [2], illetve az *OpinHuBank* projektet [3], amely

az internetes hírportálokon, blogokon és közösségi oldalakon publikált szövegek szentimentszintű annotálásának megoldására törekszik automatikus és manuális megoldások segítségével.

Ami a magyar nyelvű szövegek szentimentannotálását illeti, jelenleg egyetlen magyar nyelvű korpuszról van tudomásunk, az *OpinHuBank*ról [3], amelyben a korpusz építői a munka során a szentimentek annotálását célozták. Ugyanakkor, az elkészült korpusz több lényegi sajátsága okán elemzési és tesztelési célokra csupán korlátozottan alkalmazható. Egyrészt, a szövegekben a szentimentkifejezéseket egyenként nem annotálták a korpusz építői, a szentimentértékeket (pozitív vagy negatív) ugyanis magasabb, a mondatok vagy a tagmondatok szintjén határozták meg, az azon belüli további elemzés nélkül. Másrészt, az annotátoroknak az aktuális mondat szentimentértékének pozitív vagy negatív voltáról a mondatban szereplő tulajdonnévi entitás viszonylatában kellett döntést hozniuk, azaz arra kérték őket, hogy ítéljék meg, vajon pozitív vagy negatív ítéletet fejez-e ki az elemzett mondat a bennfoglalt PERSON (személynév) típusú entitás vonatkozásában. Mindez azért is problematikus, mert a szentiment targetjének szerepét a mondatban a személynéven kívül számtalan elem (pl. egy hely, egy esemény, egy termék vagy akár a termék egy aspektusa is) betöltheti. Az a sajátság tehát, miszerint a korpuszban kizárólag személynév tölti be a target szerepét, nyilvánvalóan jelentősen korlátozza az eszköz alkalmazhatóságát. Ugyanakkor, a legnagyobb problémát nem is ez a korlátozás jelenti. Bár a korpusz készítői hangsúlyozzák, hogy automatikus, majd kézi módszerrel kiszűrték azokat az eseteket, ahol a PERSON típusú entitás nem az adott mondat targetje, hanem a mondatban megfogalmazott vélemény forrása volt, a korpusz sajnálatos módon számos ilyen esetet tartalmaz; pl.

(1) Martonyi János leszögezte: noha a jelenlegi szlovák kormánykoalíció egyik pártjának vezetői gyakran elfogadhatatlan kijelentéseket tesznek, a magyar kormány nem ilyen stílusban fog reagálni (…)
[http://www.belfoldihirek.com/belfold/martonyi-janos-szlovakiaba-latogat]

A korpuszból idézett példa beláthatóan értékítéletet fogalmaz meg, azonban azt nem a mondat tulajdonnévvel jelölt entitásának viszonylatában teszi.

A fentebb leírt sajátságokat és problémákat megfontolva úgy döntöttünk, hogy szentimentelemző rendszerünk teszteléséhez és fejlesztéséhez, valamint a szentimentelemzés problémaköréhez kapcsolódó elméleti nyelvészeti és nyelvtechnológiai kutatások támogatása céljából létrehozunk egy olyan manuálisan annotált korpuszt, amely képes a magyar nyelvű szövegek véleménykivonatolásával kapcsolatos kutatói és fejlesztői feladatok hatékony támogatására.

2 A korpuszannotálás alapelvei és eszközei

A korpusz szöveganyagát a [http://divany.hu/] honlap termékvéleményeiből állítottuk össze. A honlap készítői időközönként bizonyos termékcsoportokat tesztelnek, s közzéteszik a tesztelők véleményét. A honlap szövegeiből 111-et gyűjtöttünk össze. A

nyers korpusz jelenleg összesen mintegy 13 000 mondatot és 190 000 tokent tartalmaz.

A manuális annotálás keretében a teljes értékelő kifejezést, azon belül pedig a pozitív és negatív polaritású szentimentkifejezéseket, azok targetjeit, valamint esetleges siftereit jelöltük be a korpuszban [4,5]. Szentimentkifejezésnek olyan egy szóból álló, vagy állandósult többszavas szókapcsolatokat tekintettünk, amelyek lexikai szinten értékítéletet hordoznak valamely target vonatkozásában [6,7]. Azokat a nyelvi elemeket, amelyek valamilyen módon hatást gyakorolnak a szövegekben megfogalmazott értékelő tartalmakra, az angol nyelvű terminológia alapján *szentimentsifterek*nek nevezzük, és külön taggel látjuk el a korpuszban [8,9].

2.1 A szentimentsifterek annotálása

A szentimentsiftereken belül két alapvető csoportot különböztethetünk meg. Az egyikbe azok az elemek tartoznak, amelyek a szentimentkifejezések szintaktikai kontextusában befolyásolják azok lexikális szintű, prior szentimentértékét, a másikba azok, amelyek a prior szentimentértékeket nem változtatják meg ugyan, azonban lehetetlenné teszik az értékelést megfogalmazó szövegrész faktív olvasatát. Az alábbiakban rövid áttekintést adunk e két átfogó kategóriáról.

Az első típusba az ún. negáló és az intenzifikáló elemek tartoznak. A szentimentértékek negálói a következő közös sajátsággal bírnak: vagy az ellenkezőjére változtatják a kifejezés prior értékét (2a), vagy pedig törlik azt (2b); pl.

(2) a. Mari nem szép. (’Mari csúnya’)
b. A béka nem gusztustalan. (nem jelenti azt, hogy ’gusztusos, tetszetős’)

A szentimentértékek negálói többek között lehetnek tagadószók (pl. *ne, sem, dehogy*), a létige tagadó alakjával (*nincs, nincsen, sincs, sincsen*), tagadó névutóval (pl. *hiányában, nélkül*) és egyéb módosítószók (pl. *aligha, látszatra*) [10].

A szentimentértékek ún. intenzifikáló elemei közé soroljuk azokat a nyelvi elemeket, amelyek a közös jellemzője, hogy a prior szentimentértéket egy bizonyos mértékben, valamilyen irányban módosítják, mégpedig úgy, hogy azt vagy erősítik (3a), vagy ellenkezőleg, csökkentik (3b); pl.

(3) a. A hangminőség nagyon jó.
b. A hangminőség aránylag jó.

A szentimentértékek intenzitásának befolyásolására számtalan elem alkalmas lehet, pl. rendkívül, rendkívüli módon, borzasztóan, elképesztően, valamennyire, valamelyest, feliből-nagyjából, részben, kevésbé stb. [11,12].

Ugyanakkor jegyezzük meg, hogy egy adott szentimentkifejezés prior értékére egy negáló és egy intenzifikáló elem is hatást gyakorolhat egyszerre; pl.

(4) A hangminőség nem nagyon jó.

A szentimentsifterek másik nagy kategóriájának elemeit irreálóknak nevezzük, és közéjük tartozónak tekintünk minden olyan nyelvi eszközt, amely lehetetlenné teszik

az értékelést megfogalmazó szövegrész faktív olvasatát. Másképpen, az irreálók megakadályozzák, hogy az adott szentimentet a megfogalmazó által tényként kezelt információként fogadjuk el. Vessük össze az (5) alatti, faktív olvasatú példát a (6) alatti, nem faktív olvasatú példákkal!

(5) A hangminőség jó.
(6) a. A hangminőség valószínűleg jó.
b. Lehet, hogy a hangminőség jó.
c. Jó a hangminőség?
d. Nem tudom, hogy a hangminőség jó-e.
e. A hangminőség jó lehet.

Amint látjuk, amíg az (5) alatti példában az értékelés megfogalmazója elkötelezi magát a propozíció igazsága iránt, addig a (6) alatti példák esetében nem, ennek következtében azok értékelő tartalmát nem is kezelhetjük a szentimentelemzés során teljes értékű adatként. Minden olyan elemet tehát, amely azt jelöli, hogy az értékelés propozíciós tartalmát a beszélő nem tényként tekinti, külön taggel láttuk el a korpuszban.

2.2 Az annotáció bemutatása

A feldolgozott szövegek sajátsága okán úgy döntöttünk, hogy a tesztelt termékek címbeli elnevezéseit *topic* címkével látjuk el, míg az egyes szentimentekhez kapcsolódó targetek *target* címkét kapnak.

A topikok és a targetek annotációszintű elkülönítése indokolható, hiszen a szentimentelemzés egy fontos része abban áll, hogy meg kell tudnunk különböztetnünk egymástól az entitásokat (*entity*), valamint azok aspektusait (*aspect*) [9]. Ennek a különbségtételnek a szentimentértékek súlyozásában jelentős szerepe van; egy adott szentiment ugyanis mind egy adott entitáshoz, mind annak csupán egy adott aspektusához is kapcsolódhat. Például, egy fényképezőgép mint entitás többek között a képminőség, a szín és az ár aspektusokkal rendelkezik. Az, hogy az értékelő az entitás, illetve az egyes aspektusok vonatkozásában milyen értékítéleteket közöl, nyilvánvalóan nagy jelentőséggel bír annak szempontjából, hogy magát az entitást hogyan értékeli; pl.

(7) Bár az ára nem volt alacsony, nagyon megérte ez a fényképezőgép.

Amint azt a fentebbi példa is mutatja, egy adott entitás egy adott aspektusáról tett negatív értékítélet nem jelent feltétlenül negatív értékítéletet a teljes entitás vonatkozásában. Ily módon az entitás–aspektus-kettősség az egyes szentimentértékek súlyozásában, ezáltal az aktuálisan elemzett szöveg összesített szentimentértékének a kiszámításában lényegi szereppel bír.

A korpuszban alkalmazott annotációt, miszerint a topikot megkülönböztetjük a targettől, a jövőben az entitás−aspektus-kettősség automatikus feldolgozásában is ki szeretnénk aknázni.

A korpusz annotációs megoldását az alábbi példával szemléltetjük:

```
(8) Negyedik helyezett: <topic>Kolios goat's
cheese</topic>
„<SentNeg> <target>Állagra</target> olyan, mint a
<SentiWordNeg>gumi</SentiWordNeg> </SentNeg>, <SentNeg>
<target>ízre</target> pedig
<SentiWordNeg>fanyar</SentiWordNeg> </SentNeg>.
<SentNeg> Nekem <ShiftNeg>nem</ShiftNeg>
<SentiWordPos>jön be</SentiWordPos> </SentNeg>.”
```

A szentimentsifterek e kezelési megoldásával alapot kívánunk teremteni egy magyar nyelvű szövegekre alkalmazható szentimentérték-kalkulátor, a *SOCal-Hun* létrehozásához [5,13].

3 A korpusz adatai

Az annotálás során a nyers szövegkorpuszból 15 szöveget dolgoztunk fel, ami összesen 1834 mondatot és 26 503 tokent tartalmaz.

Az annotáció egyetértési adatait az alábbi táblázat foglalja össze:

1. táblázat. Az annotáció egyetértési adatai

az annotált tag	F-mérték
PosSentiment	0,36
NegSentiment	0,40
SentiWordPos	0,68
SentiWordNeg	0,60
Topic	0,99
Target	0,53
Negation	0,68
IntensifierPlus	0,57
IntensifierMinus	0,63
Irreal	0,17
OtherShifter	0,30

Amint az a táblázat statisztikái alapján látható, a legnagyobb egyetértési arányt a topikok annotálásában értük el. Ez nem meglepő, hiszen topic címkével – a már említetteknek megfelelően (l. fentebb) – a tesztelt termékek tulajdonnévi jelölőit láttuk el, amelyek megtalálása és terjedelmének megállapítása nem okozhatott különösebb nehézséget az annotátorok számára. Megfelelő eredményességet produkáltunk továbbá a negáló kifejezések (Negation), az intenzifikáló sifterek (elsősorban az IntensifierMinus tag esetében), valamint a szentimentkifejezések (SentiWordPos és SentiWordNeg) annotálásában.

A targetek annotálásában már kevesebb eredményességgel dolgoztunk. Az annotáció kézi ellenőrzése arra mutatott rá, hogy az eltérés alapvetően a feldolgozott szöve-

gek domén-sajátságára vezethető vissza. Mivel az annotált korpusz termékvéleményeket tartalmaz, a tesztelők által megfogalmazott értékelések rendre a tesztelt termékek különböző aspektusaira irányulnak, azokat minősítik. Ennek köszönhetően a feldolgozott szövegek rendkívüli mennyiségű targetet tartalmaznak, amelyből számos példány elsikkad a feldolgozási munka során.

Még kisebb egyetértést mértünk a teljes szentimentegységek annotálását illetően, amelynek oka – a kézi ellenőrzés tapasztalatai alapján – egyértelműen abban keresendő, hogy a korpusz feldolgozását végző két annotátor eltérően kezelte a többszörös mellérendelő szerkezeteket: amíg az egyik annotátor azok tagjait rendre külön-külön egységekként annotálta, addig a másik gyakorta egyetlen szentimentként jelölte őket. Ez alapján feltétlenül szükségesnek tartjuk az erre vonatkozó annotálási alapelvek pontosabb rögzítését.

A legkisebb hatékonyságot az ún. irreáló elemek taggelésében értük el. Ennek valószínű oka az, hogy az irreálás jelensége, ahogyan azt már korábban a (6) alatti példákkal is igyekeztünk megmutatni (l. fentebb), számos formában jelenhet meg a szövegekben, és e sokféleségnek az egységes kezelése nehézséget okozhatott az annotátorok számára.

Az alábbi táblázat összefoglalja az annotált korpuszrész statisztikai adatait:

2. táblázat. Az annotáció statisztikai adatai

annotált tag	**darabszám**
PosSentiment	603
NegSentiment	743
SentiWordPos	708
SentiWordNeg	827
Topic	169
Target	528
Negation	316
IntensifierPlus	332
IntensifierMinus	68
Irreal	66
OtherShifter	30
ÖSSZESEN:	4390

Az annotáció fentebbi statisztikai adatai alapján a következő megállapításokat tehetjük:

A negatív véleményt megfogalmazó kifejezések (NegSentiment) többségben vannak a pozitív véleményt megfogalmazó kifejezésekkel (PosSentiment) szemben. Hasonló megoszlást találunk a szentimentkifejezések között is, ami azonban nem következik szükségszerűen az előbbi megállapításunkból, hiszen negatív vélemény pozitív szentimentkifejezéssel, illetve pozitív vélemény negatív szentimentkifejezéssel is megfogalmazható, amennyiben a kifejezés lexikai szintű polaritását egy sifter segítségével megváltoztatjuk. Ennek ellenére a táblázat adatai alapján azt látjuk, hogy a lexi-

kai szinten negatív polaritással rendelkező kifejezések fordulnak elő nagyobb számban a korpusz általunk feldolgozott részében. Az annotáció tapasztalatai meglepőek az ún. Pollyanna-hipotézis tükrében, amely nyelvi univerzáléként tételezi a pozitív töltetű kifejezések magasabb használati arányát a negatív töltetű nyelvi elemekkel szemben [14]. Mindezek alapján a megfigyelt jelenséget szeretnénk nagyobb mennyiségű annotált szöveganyagon behatóbb vizsgálat tárgyává tenni a jövőben.

Ugyancsak szembeötlő eltérés mutatkozik az intenzifikáló elemek gyakorisági megoszlásában, hiszen a fokozó típusúak (IntensifierPlus) túlnyomó többségben szerepelnek a mérséklő típusú elemekkel (IntensifierMinus) szemben. Valószínűsíthető, hogy a mért adatok összhangban állnak Székely megállapításával, miszerint a magyar nyelvben (s talán nem csak a magyar nyelvben) a mérséklés eszközrendszere szegényesebb a fokozás eszközrendszerénél [12].

Végezetül emeljük ki, hogy az annotált korpuszrész 316 negáló kifejezést (Negation) tartalmaz (ebből 140 pozitív és 176 negatív polaritású véleményben szerepel), ami jelentős előfordulási aránynak tekinthető annak fényében, hogy összesen 1346 szentimentet azonosítottunk a munka során. Az eredmény arra mutat, hogy a negáció feltétlen megoldást sürget a szentimentelemzés feladatában, hiszen figyelembe nem vételük jelentős torzulást okozhat az elemzés során kapott szentimentértékeket tekintve.

4 A korpusz felhasználási lehetőségei

Az annotált korpusz nyelvtechnológiai feladatokban és elméleti nyelvészeti kutatásokban − így tesztelési és fejlesztési célokra − egyaránt alkalmazható.

A kutatómunka következő lépéseként szeretnénk az annotációt nagyobb mennyiségű szövegre kiterjeszteni, majd az annotált korpuszt beható empirikus vizsgálat tárgyává tenni. Terveink szerint a korpuszban alkalmazott annotációra támaszkodva sikerül kialakítanunk egy olyan automatikus szentimentelemző rendszert, amely képes a szentimentkifejezéseket azok targetjeivel és siftereivel összefüggésben hatékonyan kezelni a jövőben.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Berend, G., Farkas, R.: Opinion Mining in Hungarian based on textual and graphical clues. In: Proceedings of the 8th conference on Simulation, modelling and optimization. Stevens

Point, Wisconsin, USA, World Scientific and Engineering Academy and Society (WSEAS) (2008) 408–412
2. Miháltz, M.: OpinHu: online szövegek többnyelvű véleményelemzése. In: Tanács, A., Vincze, V., eds.: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010), Szegedi Tudományegyetem, Szeged (2010) 14–23
3. Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013), Szegedi Tudományegyetem, Szeged (2013) 343–345
4. Ding, X., Liu, B., Yu S., Ph.: A holistic lexicon-based approach to opinion mining. In: Najork, M., Broder, A. Z., Chakrabarti, S. eds.: Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), New York, NY, USA (2008) 231–240
5. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37/2, Association for Computational Linguistics, MA, USA, MIT Press Cambridge (2011) 267–307 http://dl.acm.org/citation.cfm?id=2000518
6. Vincze, V.: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. In: Gecső, T., Sárdi, Cs., eds.: Új módszerek az alkalmazott nyelvészeti kutatásban, Budapest, Tinta (2010) 327–332
7. Szabó, M. K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai. In: Nyelv, kultúra, társadalom konferencia konferenciakötete (2014) (megjelenés előtt)
8. Szabó, M. K.: A magyar nyelvű szövegek szentimentelemzésének dilemmái, különös tekintettel a szentimentsifterek kezelésére. LingDok 18. Nyelvészdoktoranduszok 18. országos konferenciája, Szeged (2014)
9. Liu, B.: Sentiment Analysis and Opinion Mining. Draft (2012) http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
10. Pete, I.: Az állító és tagadó mondatok szinonimiája a magyarban. Magyar Nyelv 95/3. (1999) 305–312
11. Moilanen, K., Pulman, S.: Sentiment Composition. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007) (2007) 378–382
12. Székely, G.: Egy sajátos nyelvi jelenség, a fokozás. In: Segédkönyvek a nyelvészet tanulmányozásához 66. Budapest, Tinta (2007)
13. Brooke, J., Tofiloski, M., Taboada, M.: Cross-linguistic sentiment analysis: From English to Spanish. In: Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets (2009) 50–54
14. Boucher, J., Osgood, C.: The Pollyanna hypothesis. Journal of Verbal and Learning Behavior 8/1 (1969) 1–8

Entitásorientált véleménydetekció webes híranyagokból

Hangya Viktor, Farkas Richárd, Berend Gábor

Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{hangyav,rfarkas,berendg}@inf.u-szeged.hu

Kivonat Napjainkban a hírközlés jelentős hányada digitális formában történik, a híranyagokban említett entitásokra vonatkozó vélemények polaritásának automatikus meghatározása pedig komoly előnyökkel járhat. Éppen ezért munkánk során az OpinHuBank adatbázisban található entitásokra vonatkozó vélemények bekategorizálását tűztük ki feladatunkul. A javasolt megoldásunk többek között a szövegegységek dependenciaelemzésére is támaszkodva képes az entitások mondatbeli szerepének figyelembevételével pontosabb képet adni a rájuk vonatkozó véleményekről.

1. Bevezetés

A hírportálok egyre növekvő száma és kibocsátási aktivitása mind nagyobb mértékben teszi lehetővé számunkra, hogy a híranyagokban közölt entitásokkal kapcsolatos véleményeket megfigyeljük. Az ezen entitásokra (például cégekre vagy politikai szereplőkre) vonatkozó vélemények monitorozása hasznos segítséget nyújthat azok pozitív reputációjának megtartásában, felépítésében [1].

Mivel az online tartalmak száma egyre gyarapszik, így ezen feladat hatékony elvégzésére csupán automatikus rendszereket használva nyílik lehetőségünk. Munkánk során éppen ezért az OpinHuBank adatbázisát fölhasználva építettünk az entitásokkal kapcsolatos hírek polaritását detektáló rendszert. Rendszerünket úgy terveztük, hogy az a célentitásokra nézve képes legyen – az azt tartalmazó mondat mélyebb elemzésének végrehajtása által – az entitásra vonatkozó polaritás minél pontosabb meghatározására.

Az általunk kitűzött feladat nehézségei közül kiemelendő, hogy amennyiben egy mondat több entitást is tartalmaz, úgy könnyen megeshet, hogy azok egy része pozitív, míg másik részük negatív (vagy akár semleges) kontextusban kerül említésre, vagyis a mondatban található pozitív, illetve negatív konnotációjú szavak nem egyforma mértékben képesek befolyásolni egy-egy entitás véleménytöltetének végső polaritását. Példaképp tekintsük az OpinHuBank-ban található következő mondatot:

> „A befutó *Tóth Adrienn* az első sorozatban többet is hibázott, a németek utolsó versenyzője, az olimpiai bajnok *Lena Schöneborn* viszont jól célzott, s 10 mp-re csökkentette a különbséget."

Jól látható, hogy míg a mondat által közvetített vélemény Tóth Adriennre vonatkozóan negatív, addig Lena Schönebornra nézve ezzel éppen ellentétes. Az előzőekkel összhangban elmondható, hogy a *hibázott* szó jelenléte nem bír befolyásoló erővel a mondat Lena Schönebornnal kapcsolatban megfogalmazott állítás polaritására nézve. A fenti jelenség kezelésére a mondatok dependenciaelemzésének a modellünkbe történő beépítése mellett döntöttünk.

2. Korábbi munkák

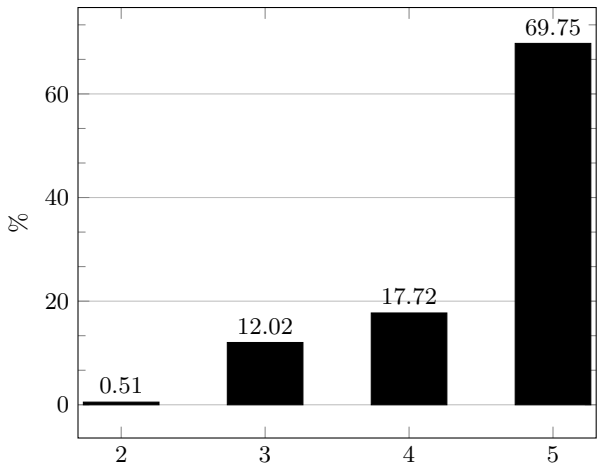
Az elmúlt években a véleménydetekció feladatát nagyfokú tudományos érdeklődés övezte, számos versenyfeladatot tűztek ki témájában [1,2]. Az angol nyelvű szövegek feldolgozására vonatkozó munkák [3,4,5] mellett a magyar nyelvű szövegekből történő véleménydetekcióra is egyre nagyobb figyelem irányul, köszönve a benne rejlő gazdasági lehetőségeknek. Cikkünkben olyan módszert mutatunk be, mely lehetőséget nyújt az egyes entitásokra irányuló vélemények beazonosítására, illetve monitorozására, azaz a velük kapcsolatos hírekben található vélemények polaritásuk szerinti kategorizálására.

Munkánk a korábbi versenykiírások közül a 2014-es SemEval aspektus-alapú véleményanalízisre vonatkozó feladatával rokonítható leginkább. A versenyfeladat kapcsán éttermekről és laptopokról szóló termékvéleményezések mondataiból az azokban említett főnévi csoportok formájában jelentkező aspektusokra (pl. kiszolgálás minősége) vonatkozó vélemények polaritását kellett meghatározni. A cél ebben az esetben tehát nem egyszerűen dokumentum- vagy mondatszintű véleményértékek meghatározása volt, hanem aspektus-mondatpárokra vonatkozóan kellett a megfelelő döntéseket meghozni. Utóbbi feladat kétségkívül nehezebb, hiszen egy mondat nyilatkozhat teljesen tárgyilagosan (vagy akár pozitívan is) a kiszolgálás minőségéről, miközben adott esetben negatív véleményeket fogalmazhat meg a termék vagy szolgáltatás árára vonatkozóan. Az ilyen és ehhez hasonló esetek kezelése a mondatok szerkezeti elemzése nélkül nem vagy csak korlátozott mértékben képzelhető el. Munkánk során mi is a verseny szervezői által megfogalmazott célhoz hasonlót tűztünk ki magunk elé, azzal a különbséggel, hogy esetünkben az OpinHuBank adatbázisban található entitások képezték a mondatok azon elemeit, melyekkel kapcsolatosan a véleménytöltet polaritását meg kívántuk határozni, nem pedig termékvéleményezések – jellemzően köznévi – aspektusai.

Módszerünk kidolgozásához és validálásához a 2013-ban létrehozott OpinHuBank [6] szöveges adatbázist használtuk fel, mely egy szabadon hozzáférhető annotált magyar nyelvű korpusz véleményelemzéshez. Az adatbázisban olyan mondatok találhatók, melyek mindegyike egy előre megadott entitással kapcsolatosak. Feladatunk tehát az volt, hogy ezen mondat-entitás párokra meghatározzuk, hogy az adott mondat tartalmaz-e az entitásra nézve pozitív vagy negatív információt. Az adatbázis közel háromnegyede hírportálok, illetve hírügynökségek oldalairól lett összegyűjtve, a fennmaradó dokumentumok forrásai blogok voltak.

3. Módszerek

Különféle modelljeinket az OpinHuBank adatbázison értékeltük ki. Az adatbázis 10006 entitás-mondatpárra tartalmazza az egyes mondatokban az entitásokra vonatkozó vélemény polaritását, melyek pozitív, negatív avagy semleges besorolásba eshetnek. Az entitás-mondatpárokat öt független annotátor sorolta be a három kategória valamelyikébe, mely feladatra vonatkozó egyezés mértékét az 1. ábra foglalja össze. Az ábrából kitűnik, hogy az entitás-mondatpárok közel 70%-ára tökéletes egyezés mutatkozott mind az öt annotátor között. A fennmaradó entitás-mondatpárok kapcsán az osztályozó modellünk tanítása és kiértékelése során osztálycímkékül az annotátorok leggyakoribb döntését vettük. Ahogy az az 1. ábrából kitűnik, az esetek több, mint 29%-ában egyértelműen meghatározható volt az annotátorok által választott osztálycímkék leggyakrabbika, és mindössze 0,5%-ban alakult ki holtverseny az annotátorok döntései kapcsán. Ezekre az esetekre – ellentmondásos jelölésükből adódóan – a továbbiakban semleges egyedekként tekintettünk. Ilyen módon az OpinHuBank adatbázisban található entitás-mondatpároknak rendre 74,9, 16,3, valamint 8,8%-ára tekintettünk semleges, negatív, valamint pozitív címkéjű példaként további vizsgálataink során. A valamilyen polaritással bíró 2511 entitás-mondatpár esetében pedig 64,9, valamint 35,1% mutatkozott negatív, illetőleg pozitív osztálybelinek. Modelljeinket egyaránt kiértékeltük azokban az esetekben, ahol feladatunkul az entitás-mondatpárok három, illetve két osztályba sorolását tűztük ki célunkul. Utóbbi esetben az adatbázisban semlegesnek mutatkozó egyedek kihagyásával tanítottuk, illetve értékeltük ki modelljeinket, melyek eredményeit a három-, illetve kétosztályos tanítás kapcsán egyaránt tízszeres keresztvalidáció eredményeként közöljük.



1. ábra. Entitás-mondatpáronként a leggyakoribb címkét választó annotátorok számának eloszlása

Munkánk során a véleményérték meghatározásának feladatára felügyelt osztályozási problémaként tekintettünk, melyben a szövegek (legfeljebb[1]) három osztályba tartozhattak annak függvényében, hogy azok egy célentitásra nézve pozitív, negatív vagy semleges információt hordoztak. Maximum entrópia alapú modelljeink paramétereinek meghatározásához a MALLET gépi tanulási programcsomagot [7] használtuk.

Az entitások kontextusának kategorizálása során nem csupán a környező szavak, szókapcsolatok jelenlétét vettük figyelembe, hanem azoknak az entitásokhoz vett relatív pozíciójának alapján történő **súlyozás**t is alkalmaztunk a modellalkotás során. Egy n hosszúságú – a célentitást az i. tokenpozíción tartalmazó – mondat j. tokenjéhez rendelt jellemző értékét a

$$\frac{1}{e^{\frac{1}{n}|i-j|}}$$

formula által határoztuk meg, így juttatva nagyobb fontossághoz a célentitás környezetében fellelhető szavakat.

Az osztályozás során az entitások kontextusának vizsgálata alkalmával a környező és kapcsolódó szavak pozitív, illetve negatív voltát is figyelembe vettük. Ennek elvégzéséhez azonban pozitív és negatív szavakat tartalmazó listákra volt szükségünk, melyekből angol nyelven több lexikon (pl. [8]) is rendelkezésre áll, ugyanakkor a magyar nyelvű szövegek esetében saját **polaritáslexikon**ok előállítására volt szükség [9]. A lexikon létrehozása során egy 2676 negatív, valamint 646 pozitív szót tartalmazó lista lefordítására került sor. A lista segítségével az egyes mondatokban található pozitív, illetve negatív szavak számából, illetve a mondatban szereplő listaelemekből külön is hoztunk létre jellemzőket ennek a jellemzőcsoportnak a kapcsán.

Mivel egyes mondatok több, adott esetben eltérő polaritású véleménnyel illetett entitást is tartalmazhatnak, ezért olyan módszerekre is támaszkodtunk, melyek képesek szeparálni a különböző entitásokra vonatkozó véleményeket, majd ezek közül kiválasztani a vizsgált entitásra nézve relevánsakat. A szövegek **dependenciaelemzés**ére támaszkodva lehetőségünk nyílt az egyes szavak pozícióin túlmenően azok nyelvtani szerepének figyelembevételére is. A dependenciaelemzés végrehajtására a `magyarlanc` [10] eszközt használtuk, egy célentitás kapcsán pedig a mondat dependenciagráfjának gyökerétől az adott entitásig elhelyezkedő szavak képezték a jellemzőcsoport által bevezetett jellemzők halmazát. Az előzőekben bemutatott, és az eredmények ismertetése során hivatkozott jellemzőcsoportokra vonatkozó rövidítéseinket a 3. táblázat foglalja össze.

4. Diszkusszió

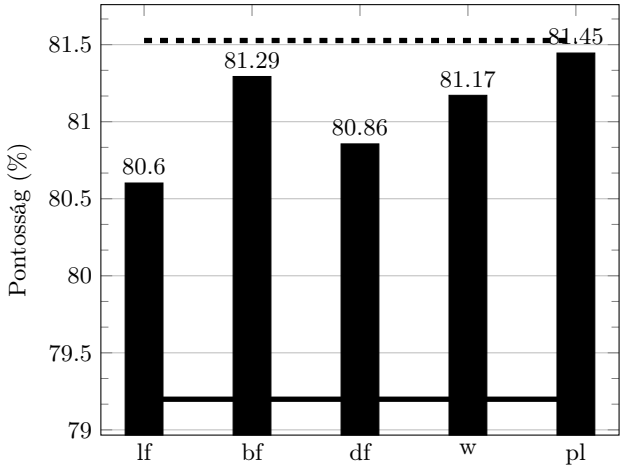
Ablációs kísérleteinket két-, illetve háromosztályos osztályozás kapcsán is kiértékeltük, melyek eredményeit a 2., illetve a 3. ábra tartalmazza. A 2. ábrán a

[1] Kétosztályos modelljeink esetében csak a pozitív, illetve negatív osztályokba történő besorolás volt a célunk.

1. táblázat. A jellemzőcsoportokra vonatkozó rövidítések

Rövidítés	Jelentés
lf	lexikai jellemzők (uni,-és bigramok)
bf	bigram jellemzők
df	dependencia alapú jellemzők
w	távolság alapú súlyozás
pl	polaritáslexikonból származó jellemzők

háromosztályos feladat megoldása során mért eredményeink láthatók, melyeket a 3. ábrán látható eredményekkel összevetve megállapíthatjuk, hogy – ahogy a várakozásoknak megfelelően – a semleges vélemények felismerését is magában foglaló feladat némileg nehezebbnek tekinthető a csupán a pozitív, illetve negatív vélemények elkülönítését megcélzó feladattól. Szaggatott vonallal azon rendszerek pontossága látható, melyek a 3. fejezetben bemutatott jellemzőtípusok mindegyikét egyidejűleg használták, folytonos vonallal pedig a – csak unigram jellemzőket használó – baseline rendszerünk eredményessége látható. Az egyes oszlopokhoz társított értékek pedig azt mutatják, hogy mennyiben változik az osztályozási feladat során elért eredményességünk, amennyiben egy-egy jellemzőcsoportot nem építünk be a jellemzőterünkbe.

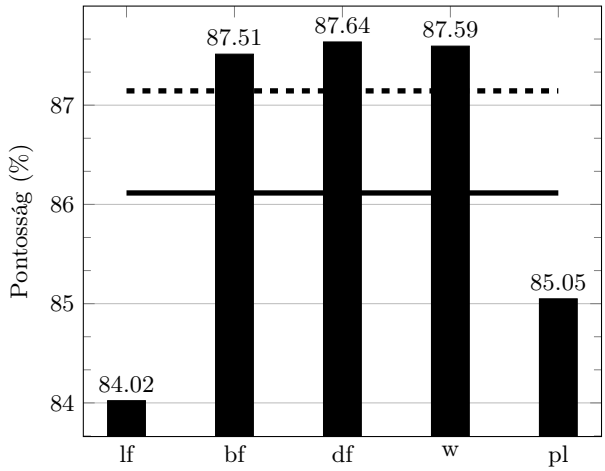


2. ábra. Ablációs kísérleteink pontosságértékei a háromosztályos feladat kapcsán

Az ábrák alapján mindkét feladat kapcsán kijelenthető, hogy a modellekből az lf jelzéssel ellátott – lexikális (uni- és bigram) – jellemzőket elhagyva figyelhető meg az osztályozási pontosság legnagyobb mértékű csökkenése. Az is elmondható, hogy ez a csökkenés javarészt az unigram jellemzők elhagyásának számlájára írható, ha ugyanis az eredményesség változásának mértékét össze-

vetjük a bf jelzésű – kizárólag a bigram jellemzőket mellőző – oszlopokéval, az eredmények hasonló fokú romlása nem volt tapasztalható.

A 2. ábra további vizsgálatából kitűnik, hogy a lexikális jellemzőket követően a dependenciaelemzésből előálló jellemzők elhagyása okozta a teljesítmény legnagyobb fokú romlását, így kijelenthető, hogy ezek alkalmazása a végső rendszer eredményességéhez nagyban hozzájárult. A dependenciaelemzésből származtatott jellemzőket követően leginkább a távolságalapú súlyozás tudott hozzájárulni a végső eredményességhez, a legkisebb hozzáadott értéke pedig a polaritásszótár használatának volt a háromosztályos feladat megoldása során az ablációs kísérleteink tanúsága szerint.

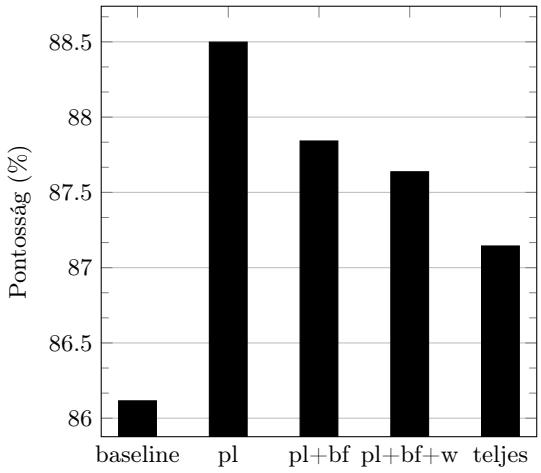


3. ábra. Ablációs kísérleteinek pontosságértékei a kétosztályos feladat kapcsán

A 2. ábrának a 3. ábrával való összevetése rámutat a két különböző – a három-, illetve kétosztályos (a neutrális osztályt tartalmazó, illetve mellőző) – tanulási feladat nagyfokú különbözőségére. Ugyan a leghasznosabbnak mindkét esetben a lexikális (főképp az unigram) jellemzők mutatkoztak, a további jellemzőcsoportok relatív hasznosságának sorrendje jelentősen eltér. Míg a háromosztályos esetben például a polaritáslexikonnak volt tulajdonítható a legkisebb szerep, addig a kétosztályos esetben a teljes rendszer szempontjából ezen jellemzők szerepe elsődlegesnek tekinthető. Ez persze nem meglepő, ha figyelembe vesszük, hogy a háromosztályos esetben a példányok közel háromnegyede a neutrális osztályba tartozó volt, a kétosztályos esetben pedig mindezen példáktól eltekintettünk, így ott kizárólag valamilyen polaritással bíró egyedek szerepeltek már csupán. Meglepő eredményként azt tapasztaltuk, hogy a bigramok figyelembevételéből, a dependenciaelemzésből, valamint a távolságalapú súlyozásból származtatott jellemzők használata – noha a háromosztályos feladat esetében mind kivétel nélkül hozzá tudtak járulni a végső rendszer eredményességéhez – nem bizonyultak

elég hatékonynak a kétosztályos feladat megoldása során, hiszen az bármelyikét elhagyva javulást tapasztaltunk ahhoz a rendszerhez képest, amelyik az összes jellemzőcsoportot egyidejűleg használta. Ezek alapján úgy tűnik, hogy alkalmazott jellemzőink egy része alkalmasnak mutatkozik annak megítélésére, hogy egy adott entitás neutrális kontextusban kerül-e említésre vagy sem, ugyanakkor kevésbé alkalmasak – a polaritás megléte esetén – annak pozitív vagy negatív voltának meghatározására.

Előző megfigyeléseinkből kiindulva a kétosztályos feladat kapcsán a jellemzőcsoportokat relatív hasznosságuk szerint sorrendbe állítva külön modelleket hoztunk létre. Ennek során megvizsgáltuk, hogy a baseline rendszer jellemzőterének – az egyre csökkenőnek mutatkozó hasznosságú jellemzőcsoportokkal történő – fokozatos bővítése milyen módon befolyásolja az eredményeket, melynek eredménye a 4. ábrán látható. A kétosztályos tanulásra vonatkozó ábrák összevetéséből magyarázatot kaphatunk arra a kontraintuitív jelenségre is, hogy a teljes jellemzőtérből csupán a polaritáslexikonra támaszkodó jellemzők elhagyásával hogyan kaphattunk a baseline rendszernél gyengébb teljesítményt (3. ábra folytonos egyenese, illetve 5. oszlopa). Ebben az esetben ugyanis az unigramokból származó jegykészleten túl csupa olyan jellemzőcsoport által került kialakításra a jellemzőkészletünk, amelyek a 4. ábra tanúsága szerint nem képesek javítani az osztályozás pontosságán.



4. ábra. A kétosztályos kiértékelés kísérleteinek pontossága

5. Konklúzió

Megközelítésünk alapvetően a szövegekben előforduló szavakon és szópárosokon alapul, azonban a pontosabb eredmények eléréséhez további – többek között a szövegek dependenciaelemzéséből, illetve polaritáslexikon alapján meghatározott – információk vizsgálatát is végrehajtottuk. A kidolgozott módszerek segítségével

a 80%-os pontosságértéket meghaladva sikerült javítani az egyszerű unigram alapú véleménykinyerésre vonatkozó osztályozás eredményein. Ezen felül olyan módszereket dolgoztunk ki, melyek hasznosak lehetnek más véleménydetekciós feladatok megoldása során is.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In: Proceedings of Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative (CLEF 2013). (2013) 333–352
2. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 27–35
3. Hangya, V., Berend, G., Varga, I., Farkas, R.: SZTE-NLP: aspect level opinion mining exploiting syntactic cues. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 610–614
4. Hangya, V., Farkas, R.: Filtering and polarity detection for reputation management on tweets. In: Working Notes of CLEF 2013 Evaluation Labs and Workshop. (2013)
5. Hangya, V., Farkas, R.: Target-oriented opinion mining from tweets. In: Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, IEEE (2013) 251–254
6. Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 343–345
7. McCallum, A.K.: MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
8. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). (2010)
9. Hangya, V., Farkas, R.: Doménspecifikus polaritáslexikonok automatikus előállítása magyar nyelvre. In: MSzNy 2015 – XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2015)
10. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 368–374

VI. Alkalmazások

Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban

Siklósi Borbála[1], Novák Attila[1,2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
e-mail:{siklosi.borbala, novak.attila}@itk.ppke.hu

Kivonat A kórházi körülmények között létrejövő klinikai dokumentumok feldolgozása a nyelvtechnológia egyik központi kutatási területévé vált az utóbbi időben. A más jellegű, általános nyelvezetű szövegek feldolgozására használt kész eszközök azonban nem alkalmazhatóak, illetve gyengén teljesítenek a speciális orvosi szövegek esetén. Továbbá számos olyan feladat van, amelyek során a szakkifejezések azonosítása és a közöttük lévő kapcsolatok meghatározása nagyon fontos lépés, azonban csak külső lexikai erőforrások, tezauruszok és ontológiák segítségével oldhatók meg. Az olyan kisebb nyelvek esetén, mint a magyar, ilyen tudásbázisok nem állnak rendelkezésre. Ezért a szövegekben lévő információk annotálása és rendszerezése emberi szakértői munkát igényel. Ebben a cikkben bemutatjuk, hogy statisztikai módszerekkel milyen módon alakíthatók át a nyers dokumentumok egy olyan előfeldolgozott, részben strukturált formára, ami ezt az emberi munkát könnyebbé teszi. A csupán a korpusz felhasználásával alkalmazott modulok felismerik és feloldják a rövidítéseket, azonosítják a többszavas kifejezéseket és meghatározzák azok hasonlóságát. Végül létrehoztuk a szövegek egy magasabb szintű reprezentációját, ahol az egyes kifejezések helyére a hasonlóságuk alapján kialakított klaszterek azonosítóját helyettesítve a szövegek egyszerűsíthetőek, a gyakran ismétlődő mintázatok általános alakja meghatározható.

1. Bevezetés

A klinikai rekordok olyan dokumentumok, amelyek kórházi körülmények között jönnek létre a mindennapi esetek, kezelések dokumentálása céljából. Ezeknek a szövegeknek a minősége messze elmarad az orvosbiológiai szövegekétől, amelyek feldolgozásával jóval több tanulmány foglalkozik éppen könnyebb kezelhetőségük miatt. Az orvosbiológiai szövegek általában angol nyelvűek, és tudományos folyóiratokban, könyvekben, kiadványokban jelennek meg, nyelvezetük követi a nyelvi és helyesírási normákat [1,2]. Ezzel szemben, a klinikai beteglapok olyan strukturálatlan szövegek, amelyek minden ellenőrzés nélkül jönnek létre. Ezért gyakran fordulnak elő bennük helyesírási hibák, elírások, és az egyedi szóalakok használata is igen jellemző ezekre a dokumentumokra. Nyelvezetük pedig

gyakran a helyi nyelv (a mi esetünkben a magyar) és a latin sajátos keveréke [3,4]. Jellemző rájuk továbbá a gyakran egyedi módon is használt rövidítések magas aránya, olyannyira, hogy akár teljes állítások, mondatok is csupán rövidített alakokból állnak.

A klinikai dokumentumok egy további jellemzője, hogy olvasóik általában maguk a lejegyzést készítő vagy diktáló orvosok, ezért az egyedi nyelvhasználat és rövidítési szokások nem okoznak információveszteséget, amikor az orvos újra elolvassa ezeket. Azonban az egyes betegek kórtörténetének tárolása mellett az ezekből a dokumentumokból kinyerhető információk az orvostudomány más területein is felhasználhatóak lehetnének. Ahhoz, hogy hozzáférjünk ezekhez a rejtett adatokhoz, a szövegekben található tények és állítások hatékony reprezentációjára van szükség.

Több kísérlet irányult már általános szövegekre működő eszközök klinikai dokumentumokra való alkalmazására, azonban ilyenkor teljesítményük jóval gyengébb, mint általános, jól formált szövegek esetén (pl. [5]). Azok az eszközök, amik pedig doménspecifikus szövegek feldolgozására alkalmazhatóak, általában valamilyen külső, kézzel készített lexikai erőforrásokat, ontológiákat használnak. Azon nyelvek esetén azonban, amelyekre kevés ilyen erőforrás áll rendelkezésre, ezek a módszerek nem alkalmazhatóak, az erőforrások létrehozása pedig jelentős emberi munkát igényel. Egy további lehetőség a klinikai dokumentumokban használt nyelv alnyelvként való kezelése [6]. Azonban ehhez is szükség van az alnyelv kifejezéseinek doménspecifikus kategorizációjára, ami szintén nem oldható meg teljesen automatikus módszerekkel [7,8,1].

A már meglévő eszközök adaptálhatósága és a strukturált erőforrások építésének támogatása céljából egy magyar szemészeti korpuszt vizsgáltunk meg. Majd különböző, nem felügyelt statisztikai módszerek alkalmazásával tettünk kísérletet többféle információ felfedezésére a nyers korpuszban. Bár az egyes modulok által létrejött eredmény önmagában nem tekinthető a dokumentumok információtartalmát teljesen lefedő reprezentációnak, azonban az ezekből a félig strukturált adatokból létrejövő csoportok felhasználhatóak a későbbi konstrukciók megalkotása során. Mindegyik modul magja a nyers korpuszból kinyert statisztika, csupán néhány ponton volt szükség alapvető nyelvi szabályok, illetve erőforrások bevonására.

2. A korpusz

Vizsgálataink során anonimizált, szemészeti osztályon keletkezett magyar nyelvű dokumentumokat használtunk. A rendelkezésünkre álló korpusz mérete 334.546 token (34.432 mondat). A korpusz a feldolgozás előtt [3]-ban ismertetett struktúrájú magas szintű xml formátumban volt. Ebben jelölve voltak a mondat- és tokenhatárok, illetve a szófaji egyértelműsítés eredménye is. A szegmentálás és szófaji egyértelműsítés automatikusan történt [9]-ben és [5]-ben bemutatott módszerekkel. Bár ezt a két előfeldolgozási lépést a legtöbb nyelv esetében általános nyelvű szövegekre megoldották már automatikus módszerekkel, az előbb említett tanulmányok kitérnek arra is, hogy esetünkben a teljesítmény jelentősen

elmarad az elvárthoz képest a klinikai szövegeken. Jelen munkánkban azonban, mivel ezek az előfeldolgozási lépések elengedhetetlenek, elfogadottnak tekintettük a korpusz ilyen minőségű kiindulási állapotát.

Egy általános nyelvű magyar korpuszhoz hasonlítva, számos jelentős különbség fedezhető fel a két domén között, ami magyarázatot ad arra, hogy az általános szövegeken akár bonyolultabb feladatok esetén is jól teljesítő eszközök miért nem alkalmazhatóak a klinikai dokumentumokra. A különbség a két szövegtípus között nem csak azok tartalmában nyilvánul meg, hanem már a nyelvtani szerkezetek és a szövegekben előforduló szóalakokban is. A két domén részletes összehasonlítása megtalálható [10]-ben és [11]-ben.

3. Az alkalmazott módszerek

Ebben a fejezetben négy módszert mutatunk be, amelyeket később egymással kombinálva alkalmaztunk a klinikai szövegekre, ami az egyes dokumentumok félig strukturált kivonatát eredményezte. Az első modul a rövidítések feloldásáért felelős, a második összetett szakkifejezések felismerését végzi, illetve rangsorolja ezeket, a harmadik szó-, illetve jelen esetben kifejezéspárok hasonlóságát határozza meg, a negyedik pedig fogalmi klasztereket hoz létre, illetve helyettesíti be ezeket az eredeti szövegekbe. Mind a négy modul működése során a korpusz statisztikai jellemzői a meghatározók.

3.1. Rövidítések feloldása

A rövidítésfeloldás feladatát többen a jelentés-egyértelműsítés (word sense disambiguation, WSD) egy speciális eseteként kezelik ([12]). A jelentés-egyértelműsítés megoldása során legjobb eredményt elérő módszerek felügyelt gépi tanulási algoritmusokat alkalmaznak. A magyar orvosi nyelvhez viszont sem kézzel előre annotált adatok, sem a lehetséges jelentések, azaz feloldási javaslatok adatbázisa nem áll rendelkezésre, ami a felügyelt tanulási módszerek alkalmazhatóságának előfeltétele ([13]). A szintén WSD feladatokra alkalmazott nem felügyelt módszerek két fázisból állnak. Az egyértelműsítés előtt meg kell határozni a lehetséges jelentések halmazát is (word sense induction, WSI). Kontextuális jellemzők alapján ugyan meghatározhatóak a lehetséges jelentések egy adott korpuszból, azonban ehhez nagy méretű korpuszra van szükség, különösen akkor, ha a többértelmű kifejezések (rövidítések) aránya olyan nagy, mint a klinikai szövegekben. Mivel kellően nagy méretű korpusz nem állt rendelkezésünkre, ezért ezt a megközelítést sem követhettük.

Így egy olyan korpuszalapú megoldást dolgoztunk ki a rövidítések automatikus feloldására, ami csupán kiegészítésként fordul a néhány, magyar nyelven elérhető, klinikai kifejezéseket is tartalmazó erőforrásokhoz. Mivel a csupán a korpuszra építő módszer nem teljesített kielégítően, ezért szükséges volt egy doménspecifikus lexikon létrehozása is. Az orvosi, klinikai kifejezéseket teljesen lefedő adatbázis helyett [14]-ben megmutattuk, hogy egy kisebb, doménspecifikus lexikon is elégséges, az ebben definiálandó rövidítések pedig a korpuszból

közvetlenül kinyerhetők. Miután ez a lexikon rendelkezésre áll, valamint a rövidítések azonosítása is megtörtént, a feloldást rövidítések sorozatára végeztük. Erre azért volt szükség, mert annak ellenére, hogy az egyes rövidítések önmagukban állva erősen többértelműek, gyakran fordulnak elő rövidítéssorozatok részeként, ahol biztosabban meghatározható az egyértelmű jelentésük. Például, az "o." rövidítés bármely o-val kezdődő magyar vagy latin szó rövidítése is lehet. Még az orvosi szaknyelvre szűkítve is igen nagy a lehetőségek száma. Az általunk vizsgált klinikai korpusz szemészeti részében azonban az "o." rövidítés csak elvétve fordul elő önmagában, sokkal inkább olyan szerkezetekben, mint például "o. s.", "o. d.", vagy "o. u.", melyek jelentése *oculus sinister* (bal szem), *oculus dexter* (jobb szem), illetve *oculi utriusque* (mindkét szem). Az ilyen összetételekben az "o." jelentése már egyértelműen meghatározható. Természetesen az ugyanazzal a jelentéssel bíró rövidített alakoknak is számos variációja előfordulhat, így az "o.s." gyakori változatai például az "o. sin.", "os", "OS" stb. A rövidítések feloldása során tehát olyan rövidítéssorozatokat tekintettünk kiindulási egységnek, melyek a szövegben egymást folytonosan, megszakítás nélkül követő rövidített alakok vagy rövidítések sorozata.

Továbbá, az olyan vegyes kifejezések egyben tartása végett, melyeknek nem minden tagja rövidített, a rövidítéssorozatokhoz azok adott méretű környezetét is csatoltuk. A feloldás során ilyen módon elérhető szövegkörnyezet a rövidítések jelentésének egyértelműsítésében is szerepet játszik, hiszen egy konkrét felszíni alakkal rendelkező rövidítés jelentése (feloldása) a környezetétől függően változhat.

Az így automatikusan felismert, majd a szövegkörnyezettel kiegészített rövidítéssorozatok feloldására először a korpuszból nyertünk ki lehetséges feloldásjelölteket, majd csupán az ebben a lépésben nem, vagy csak részlegesen feloldott rövidítéssorozatok feloldása során fordultunk a külső lexikonhoz. Az algoritmus részletei és eredményei megtalálhatóak [14]-ben és [15]-ben. Az eredmények alapján kimutatható, hogy bár a korpusz önmagában nem elégséges minden rövidítés feloldásához, ennek használatával pontosabb és helyesebb feloldások nyerhetők ki, mint csupán külső lexikon alkalmazásával.

A 1. táblázatban látható néhány automatikusan felismert rövidítés a hozzájuk tartozó feloldásokkal. A feloldásokhoz a magyar és latin változatot is megjelenítjük, ahol ezek a változatok elérhetőek és relevánsak (pl. *mindkét szem*; *oculi utriusque*), illetve több latin, vagy több magyar feloldás is szerepelhet (pl. *szemfenék*; *fundus oculi*; *fundus*).

3.2. Többszavas szakkifejezések azonosítása

A klinikai nyelvben (bármely más szaknyelvhez hasonlóan) gyakoriak az olyan többszavas kifejezések, melyek együtt jelölnek egy fogalmat. Mivel olyan releváns információk jelenhetnek meg ilyen formában, mint a betegségek, kezelések, testrészek neve, ezért fontos ezek azonosítása. Az ilyen kifejezések azonosítására nem elegendő egy általános lexikon, hiszen vannak olyan kifejezések, melyek az általános nyelvben nem feltétlenül tartoznak össze. Például a *szem* szó, mint testrész, a szemészeti szövegekben önmagában nem sok információt tartalmaz, viszont a

1. táblázat. Néhány rövidítés és a hozzá tartozó feloldások

Rövidítés	Feloldások
mydr	mydrum
mksz	mindkét szem; oculi utriusque
V	visus
D	dioptria
mou	méterről olvas ujjat
ünj	üveg nem javít
o. u	oculi utriusque; mindkét szem
F	szemfenék; fundus oculi; fundus
j.o.	jobb oldal

bal szem, *jobb szem*, *mindkét szem* kifejezések már konkrétan meghatározzák a dokumentumban leírt jelenségek pontos helyét. Éppen emiatt a szemészeti korpuszban a szem szó önmagában nem is gyakran fordul elő. Az ilyen kifejezések felismerésére tehát jól alkalmazható a kölcsönös információ (mutual information) és a kollokációk vizsgálatán alapuló módszerek, melyek éppen a korpuszbeli előfordulások alapján definiálhatóak. Ezeknek a módszereknek a többszavas szakkifejezések felismerésére való alkalmazását [16] foglalja össze, majd az egymásba ágyazott kifejezések problémájára is megoldást nyújtó c-value módszert ismerteti.

Ennek a c-value algoritmusnak egy módosított változatát használtuk. Először egy nyelvi szűrőt alkalmaztunk annak érdekében, hogy a kifejezésjelöltek listáján csak nyelvtani szempontból is helyes kifejezések szerepeljenek. A megengedett kifejezések formája a következő:

$$(FN|ADJ|IGE\_OKEP|IGE\_MIB)^{+}FN$$

Ez a minta biztosítja, hogy egyrészt csak főnévi csoportok legyenek a jelöltek között, másrészt kizárja a gyakori kifejezéstöredékeket is. Természetesen más jellegű kifejezések, mint például igei csoportok, is relevánsak lehetnek. Azonban, ahogy a 2. fejezetben bemutattuk, az igék gyakorisága alacsony a klinikai szövegekben. Ezért egy viszonylag kis méretű korpuszból nem építhetőek pontos statisztikai modellek az ilyen, ritkábban előforduló kifejezésekre.

Miután az összes, a fenti mintára illeszkedő n-gramot kigyűjtöttünk ($1 < n < 10$), mindegyikre meghatároztuk a hozzá tartozó c-value-t, ami az adott n-gram kifejezés voltára utaló mérőszám. Ez az érték négy komponens alapján határozható meg:

- a kifejezésjelölt gyakorisága;
- annak gyakorisága, hogy hányszor fordul elő hosszabb kifejezés részeként;
- az ilyen, hosszabb kifejezések száma; és
- a kifejezés hossza.

Ezeket a statisztikákat a korpusz alapján lehet meghatározni. A c-value számítását végző algoritmus részletei [16]-ban találhatóak meg. A 2. táblázatban látható néhány többszavas kifejezés, amit egy dokumentumból nyertünk ki, a hozzájuk tartozó c-value értékkel.

2. táblázat. Egy dokumentumból kinyert többszavas kifejezések a hozzájuk tartozó c-value értékkel

Kifejezés	c-value
bal szem	2431.708
ép papilla	1172.0
tiszta törőközeg	373.0
békés elülső szegmentum	160.08
hátsó polus	47.5
tompa sérülés	12.0

3.3. Disztribúciós szemantikai modellek

A releváns kifejezések csoportosításához szükség van egy hasonlósági metrikára is, ami két kifejezés jelentésbeli távolságát határozza meg. Erre szintén olyan nem felügyelt módszert alkalmaztunk, amely a hasonlóságokat nem egy külső erőforrás, ontológia alapján határozza meg, hanem a kifejezések korpuszbeli előfordulásai, az adott korpuszban való használatuk alapján.

A disztribúciós szemantika lényege, hogy a szemantikailag hasonló szavak hasonló környezetben fordulnak elő. Tehát két szó jelentésének hasonlósága meghatározható a környezetük hasonlósága alapján. A szavak környezetét olyan jellemzőhalmazokkal reprezentáltuk, ahol minden jellemző egy relációból (r) és az adott reláció által meghatározott szóból (w') áll. Ezek a relációk más alkalmazásokban általában függőségi relációk, azonban a klinikai szövegekre ilyen elemzés a zajos mivoltuk miatt nem végezhető el kellően jó eredménnyel. Carrol et al. [17] szintén klinikai szövegekre alkalmazva csupán a vizsgált szó meghatározott méretű környezetében előforduló szavak lexikai alakjának felhasználásával építettek ilyen szemantikai modellt. Mivel a mi esetünkben a morfológiai elemzés is rendelkezésre állt, ezért a következő jellemzőket vettük figyelembe:

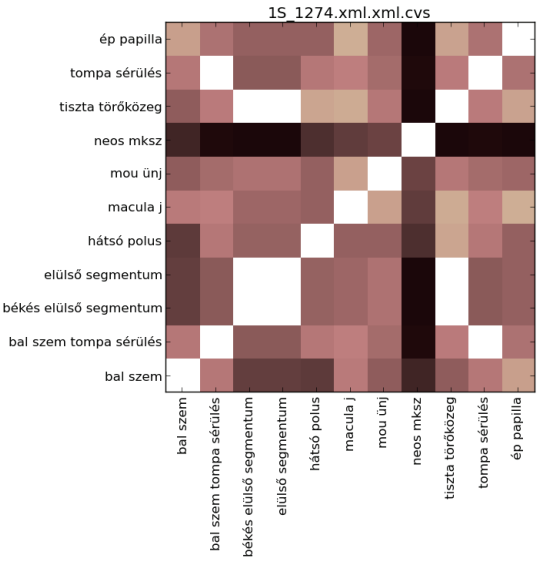
- prev_1: a szót megelőző szó lemmája
- prev_w: a szó előtt 2-4 távolságon belül eső szavak lemmái
- next_1: a rákövetkező szó lemmája
- next_w: a szó után 2-4 távolságon belül eső szavak lemmái
- pos: a szó szófaja
- prev_pos: a szót megelőző szó szófaja
- next_pos: a szót követő szó szófaja

Minden egyes jellemzőhöz meghatároztuk a korpuszbeli gyakoriságát. Ezekből a gyakoriságokból határozható meg a (w,r,w') hármas információtartalma

(I(w,r,w')) maximum likelihood becsléssel. Ezután a két szó (w_1 és w_2) közötti hasonlóságot a következő metrikával számoltuk [18] alapján:

$$SIM(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w)\in T(w_1)} I(w_1, r, w) + \sum_{(r,w)\in T(w_2)} I(w_2, r, w)}$$

ahol T(w) azoknak az (r,w') pároknak a halmaza, ahol az I(w,r,w') pozitív.



1. ábra. Egy dokumentumhoz tartozó kifejezések páronkénti hasonlóságát megjelenítő hőtérkép

Ezzel a metrikával bármely két kifejezés közötti disztribúciós hasonlóság meghatározható. Bár elvileg két bármilyen típusú szóra alkalmazhatjuk, egy ige és egy főnév összehasonlításának gyakorlati szempontból nem sok haszna lenne. Ezért a metrika többszavas kifejezésekre való alkalmazásából fakadó komplexitás elkerülése érdekében a többszavas kifejezéseket összevonva és az [FN][MT] címkével ellátva vettük figyelembe. Mivel minden kifejezésjelölt, az előző alfejezetben leírtaknak megfelelően, főnévi csoport, ezért ez a hasonlóság megfelel a két kifejezés közötti hasonlóságnak. A 1. ábra egy hőtérképen jeleníti meg az egy dokumentumhoz tartozó kifejezések páronkénti hasonlóságát. Minél világosabb egy négyzet, a hozzá tartozó kifejezések annál hasonlóbbak. Látható, hogy a *"tiszta törőközeg"* és a *"békés elülső segmentum"* hasonló viselkedést mutatnak, míg például a *"neos mksz"* az aktuális dokumentumból kinyert kifejezések közül egyikhez sem igazán hasonlít.

3.4. Fogalmi klaszterek és mintázatok

A szavak és kifejezések páronkénti hasonlóságából kiindulva fogalmi hierarchiát határozhatunk meg. Ehhez a leggyakoribb kifejezések és szavak csoportján agglomeratív klaszterezést hajtottunk végre. Bár adná magát, hogy a klaszterezés során szükséges távolságmetrikának a fent definiált disztribúciós hasonlóságot használjuk, ez önmagában nem bizonyult elégségesnek, illetve a klaszterezési algoritmusok rugalmassága is csökkent volna, ha csupán ezt a metrikát használjuk. Ezért az egyes kifejezéseket a többi kifejezéshez való hasonlóságukból álló jellemzővektorokkal ábrázoltuk. Így az egy kifejezéshez tartozó $c(w)$ vektor c_i eleme $SIM(w, w_i)$. Az egyes kifejezésekhez így létrehozott jellemzővektorokat klaszteresztük, ahol a klaszterek távolságát Ward ([19]) módszere alapján határoztuk meg. Ennek köszönhetően a kapott dendrogram alsó szintjein tömör, egymáshoz közel álló kifejezésekből álló csoportok jöttek létre.

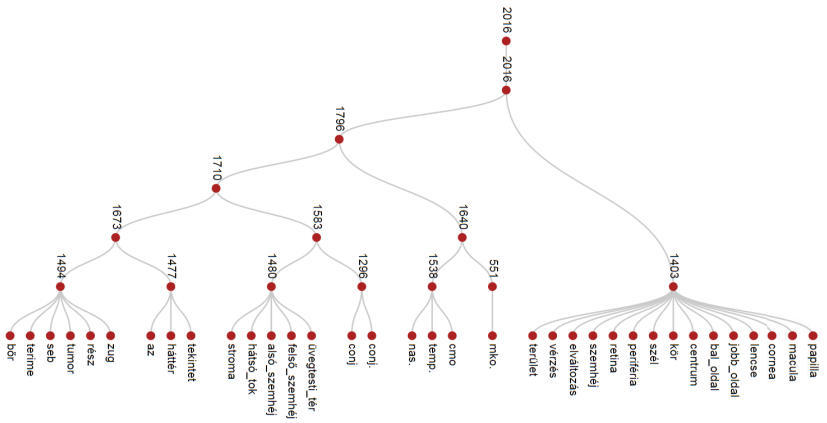
Ezeket egy, a dendrogramról szabad szemmel is jól leolvasható küszöbérték alatt összevontuk. Így a 3. táblázatban példaként felsoroltakhoz hasonló csoportok jöttek létre.

3. táblázat. Klaszterezés által létrejött levél csomópontok

ID:1403	ID:1636	ID:1549	ID:1551	ID:2045
papilla	hely	tbl	folyamat	ép papilla
macula	kh	medrol	kivizsgálás	halvány papilla
cornea	kötőhártya	üveg	érték	jó színű papilla
lencse	szaru	szemcsepp	idegentest	szűk ér
jobb oldal	conjunctiva	gyógyszer	gyulladás	ép macula
bal oldal	szemrés		retinaleválás	fénytelen macula
centrum	szempilla		látásromlás	kör fekvő retina
kör	pilla			fekvő retina
szél	könnypont			rb.
periféria				tiszta törőközeg
retina				bes
szemhéj				békés elülső szegmentum
elváltozás				békés es
vérzés				
terület				

A létrejött csoportok vagy rokonértelmű kifejezéseket, vagy szemantikai szempontból azonos szerepet betöltő kifejezéseket (pl. hét napjai, gyógyszernevek, stb.) tartalmaznak, akár rövidített alakokkal együtt (pl. *"bes"*, *"békés elülső szegmens"*, *"békés elülső szegmentum"*, *"békés es"*). Ezek mellett létrejöttek azonban olyan absztrakt csoportok is, amelyek az orvosi eljárás egyes fázisaihoz kapcsolódó kifejezések csoportjai (pl. az *"éhgyomor"* az időpontokhoz kapcsolódó kifejezésekkel került egy klaszterbe, illetve a *"strab"* és a *"párhuzamos szem"* kifejezések is az orvosi jelentőségük miatt kerültek egy csoportba). Ezeket a levél

csomópontokat összevonva, a magasabb szintű hierarchia természetesen megmarad, illetve a létrejött fa bármilyen küszöbérték mentén tovább vágható. Ezt kihasználva olyan részfákat vágtunk ki a teljes hierarchiából, amelyek a korábban kinyert kifejezéscsoportok közül az egymáshoz közel állókat fogja össze, megjelenítve azok hierarchiáját is. Egy ilyen részfa látható a 2. ábrán.



2. ábra. A teljes hierarchiából kivágott részfa, a leveleken összevont kifejezéscsoportokkal

Az orvosi szövegekben található kifejezések ilyen módon létrejött csoportosítása és rendszerezése önmagában is felhasználható lehet egy, az orvosi gyakorlatban használatos kifejezésekből álló ontológia kiindulásaként. Mivel azonban minden csoportot (illetve a létrejött hierarchia minden csomópontját) egy egyedi azonosítóval láttunk el, ezért ezek az eredeti szövegekbe visszahelyettesíthetők. Így egy tetszőleges szintű, de az eredetinél absztraktabb reprezentációját hoztuk létre az egyes dokumentumoknak. Ebből a reprezentációból könnyen azonosíthatók a dokumentumokban gyakran előforduló mintázatok, függetlenül attól, hogy a konkrét alakja egy-egy állításnak mennyire ritka, vagy gyakori kifejezést tartalmaz. A 4. táblázatban látható néhány mondat ilyen módon behelyettesített változata. A behelyettesítés következtében a mondatok nem csak egyszerűbbé válnak, hanem gyakran ismétlődő mintázatok is kiemelhetőek lesznek. A példában is megjelenő *"1889 1706 1706"* minta arról szól, hogy valamilyen állapotában valamelyik szemről milyen adatokat jegyeztek fel. Ennek a mintának a néhány leggyakoribb megjelenési formája a szövegekben: *"st. o. u."*, *"st. o. s."*, *"st. o. d."*, *"moct o. d."*, *rl. o. u."*, *"rl. o. sin."*, *"status o. s."*, *"távozáskor o. d."*, *"b-scan o. d."*, stb. Jellemző továbbá erre a mintára, hogy sor elején jelenik meg. A példamondatokban látható az is, hogy hol jelennek meg a 2. ábrán látható *2016*-os azonosítóval ellátott részfa egyes kifejezései, amik a mondatban leírt állítás helyét jelölik meg.

Bár vannak esetek, amikor az ugyanazzal az azonosítóval ellátott kifejezések nem mind tekinthetőek egy csoportba sorolhatónak, a közöttük fennálló kapcsolat hierarchiáját a részfákból meg lehet határozni, illetve a csoportokat meghatározó küszöbérték megfelelő hangolásával állítható a klaszterezés finomsága. Tapasztalataink szerint nem érdemes nagyon szűk csoportokat definiálni, hiszen sok esetben az egymással lazább kapcsolatban álló kifejezések is jól illeszthetőek az eredményül kapott mintázatokba. Jó példa erre a *NUMx* (a NUM valamilyen számot jelöl) és a *th.*, illetve ezek változatainak csoportja, ahol az egy gyógyszer adagolására vonatkozó meghatározás tulajdonképpen egy terápia. Ezzel szemben a *NUMán* sokkal inkább a vizsgálatokhoz, időpontokhoz besorolt kifejezés. Ha a csoportosításnál használt küszöbértéket alacsonyabbra állítjuk, akkor az ilyen tágabb asszociációkat elveszítjük. Az egyensúly megtalálása további kutatás témája, melyhez orvosszakértők bevonására is szükség van.

4. táblázat. Néhány példamondat, ahol a szavakat és kifejezéseket a hozzájuk tartozó klaszterazonosítóval helyettesítettük

1518 1706 1706 : ***2016*** tiszta üti *2007* , *2045* , szemfenék-szerte *2007* , *1956* , a macula_kemény_exsudatum , *2007* . *fu. o. u* : ***mko.*** tiszta üti *tér* , *ép_papilla* , szemfenék-szerte *ma-k* , *pontszerű_vérzés* , a macula_kemény_exsudatum , *oedema* .
2071 1706 1706 : ***2016*** felett és nasalisan szivárgó , ischaemiás_terület , kis neovasc._burjánzás , *2049* *flag o. d.* : ***macula*** felett és nasalisan szivárgó , ischaemiás_terület , kis neovasc._burjánzás , *macula_oedema*
2016 sima , csillogó , *2007* és a *1789* tiszta . ***cornea*** sima , csillogó , *állománya* és a *hátlapja* tiszta .
2016 tizsta . *1706* friss *1884* nem látható . ***lencse*** tizsta . *funduson* friss *kóros* nem látható .
2016 tiszta , *1789* tiszta , *1789* tiszta , *1789* békés , *1789* jól reagál . ***stroma*** tiszta , *hátlap* tiszta , *csarnok* tiszta , *iris* békés , *pupilla* jól reagál .
2016 nem vizsgálható erős_fénykerülés miatt . ***periféria*** nem vizsgálható erős_fénykerülés miatt .
1889 1706 1706 : *1998* , halvány_conjunctiva , ***2016*** epithelialis pontszerű_kiemelkedő_szürkésfehér_laesi_(j»b, balon csak NUM-NUM laesio) , a cornea_mély_rész épek ,transparens , *1812* mély , tiszta , *1789* békés , *1812* tág , kerek , centrális , *2007* jól reagál . *st. o. u* : *ép_védőszerv* , halvány_conjunctiva , ***corneán*** epithelialis pontszerű_kiemelkedő_szürkésfehér_laesi_(j»b, balon csak NUM-NUM laesio) , a cornea_mély_rész épek , transparens , *csarnok_kp* mély , tiszta , *iris* békés , *pupilla_kp* tág , kerek , centrális , *fényre* jól reagál .

4. Konklúzió

A klinikai dokumentumok mind a tartalmuk, mind a nyelvhasználatuk miatt egy doménspecifikus alnyelvet reprezentálnak. Ezeknek a szövegeknek a fő tulajdonsága a nagy mennyiségű zaj jelenléte, ami a helyesírási hibákból, a rövidítésekből és a hiányos szintaktikai szerkezetekből adódik. A szövegekben található információk kinyerését tovább nehezíti, hogy az olyan kis nyelvekre, mint a magyar, nincsenek kész lexikai erőforrások, amiket más nyelvek esetén a szövegekben szereplő kifejezések és a közöttük fennálló kapcsolatok azonosítására gyakran használnak. Ezért az ilyen lexikonok előállítása orvosszakértői feladat. A nyers dokumentumok kezdetleges előfeldolgozással történő átalakítása azonban jelentősen megkönnyítheti és hatékonyabbá teheti ezt a munkát. Cikkünkben olyan korpuszalapú módszereket mutattunk be, amik jól teljesítenek rövidítések automatikus feloldására, többszavas kifejezések felismerésére és ezek hasonlóságának meghatározására. A dokumentumoknak egy olyan félig strukturált reprezentációja jött létre ezeknek a moduloknak az alkalmazásával, ami az emberi feldolgozást támogatja, hatékonyabbá teszi. Továbbá a hasonlósági metrikák által definiált szorosabb vagy lazább kapcsolatok egy relációs tezaurusz építése során annak kezdeti állapotát képző információk is lehetnek.

Hivatkozások

1. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J.: Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association **1**(2) (1994)
2. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform **35** (2008) 128–44
3. Siklósi, B., Orosz, Gy., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.M., Forcada, M., M. Tyers, F., Waiganjo Wagacha, P., eds.: 8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages. (2012) 29.–34.
4. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. In Dediu, A.H., Martin-Vide, C., Mitkov, R., Truthe, B., eds.: Statistical Language and Speech Processing. Volume LNAI 7978., Springer Verlag (2013) 248–259
5. Orosz, Gy., Novák, A., Prószéky, G.: Lessons learned from tagging clinical Hungarian. International Journal of Computational Linguistics and Applications **5** (2014)
6. Harris, Z.S.: The structure of science information. J. of Biomedical Informatics **35**(4) (2002) 215–221
7. Friedman, C., Kra, P., Rzhetsky, A.: Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics **35**(4) (2002) 222–235
8. Kate, R.J.: Unsupervised grammar induction of clinical report sublanguage. J. Biomedical Semantics **3**(S-3) (2012) S4

9. Orosz, Gy., Novák, A., Prószéky, G. In: Hybrid text segmentation for Hungarian clinical records. Volume 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg (2013)
10. Siklósi, B., Novák, A.: A magyar beteg. In: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2014) 188–198
11. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. Computer Speech and Language (2014)
12. Navigli, R.: A quick tour of word sense disambiguation, induction and related approaches. In: Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM). (2012) 115–129
13. Nasiruddin, M.: A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. CoRR **abs/1310.1425** (2013)
14. Siklósi, B., Novák, A., Prószéky, G.: Resolving abbreviations in clinical texts without pre-existing structured resources. In: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, LREC 2014. (2014)
15. Siklósi, B., Novák, A.: Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: MICAI 2013: 12th Mexican International Conference on Artificial Intelligence. Volume 8265 of Lecture Notes in Artificial Intelligence., Heidelberg, Springer-Verlag (2013) 318–328
16. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms:. the C-value/NC-value method. International Journal on Digital Libraries **3**(2) (2000) 115–130
17. Carroll, J., Koeling, R., Puri, S.: Lexical acquisition for clinical text mining using distributional similarity. In: Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II. CICLing'12, Berlin, Heidelberg, Springer-Verlag (2012) 232–246
18. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics - Volume 2. COLING '98, Stroudsburg, PA, USA, Association for Computational Linguistics (1998) 768–774
19. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301) (1963) 236–244

Az enyhe kognitív zavar automatikus azonosítása beszédátiratok alapján

Vincze Veronika[1,2], Hoffmann Ildikó[3,4], Szatlóczki Gréta[4], Bíró Edit[5], Gosztolya Gábor[2], Tóth László[2], Pákáski Magdolna[5], Kálmán János[5]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2., e-mail: vinczev@inf.u-szeged.hu
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: {ggabor,tothl}@inf.u-szeged.hu
[3]Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék,
Szeged, Egyetem u. 2., e-mail: {hoffmannildi,szatloczkigreti}@gmail.com
[4]MTA Nyelvtudományi Intézet
Budapest, Benczúr u. 33.
[5]Szegedi Tudományegyetem, Pszichiátriai Klinika,
Szeged, Kálvária sugárút 57. e-mail: edit17@gmail.com,
magdolna.pakaski@gmail.com, kalman.janos@med.u-szeged.hu

Kivonat Ebben a munkában az enyhe kognitív zavarban szenvedő páciensek automatikus azonosítására törekszünk beszédátirataik alapján. A rendszer elsődlegesen beszélt nyelvi sajátosságokra, illetve a beszédátiratok automatikus morfológiai és szintaktikai elemzésén alapuló jellemzőkre épül. Cikkünkben elemezzük az egyes jellemzők megkülönböztető szerepét mind statisztikai, mind gépi tanulási szempontból. Eredményeink alapján elsődlegesen a morfológiai jellemzők és a beszédjellemzők bírnak fontos szereppel a páciensek státuszának automatikus megállapításában.
Kulcsszavak: enyhe kognitív zavar, demencia, gépi tanulás, beszédátirat

1. Bevezetés

Az enyhe kognitív zavar olyan tünetegyüttes, melynek fontos szerepe van az Alzheimer-kór korai felismerésében [1]. Tünetei már akár kilenc évvel a tényleges diagnózis előtt jelentkezhetnek, például nyelvi zavarok formájában [2]. Így tehát sok esetben a páciensek nyelvhasználata alapján már a demencia tényleges klinikai megjelenése előtti fázisban is megállapíthatók az enyhe kognitív zavar jelei.

Nagyon sok beteg esetében az enyhe kognitív zavart egyáltalán nem diagnosztizálják, mivel a kognitív képességek károsodásának felismerése a betegség korai szakaszában még a szakértők számára sem triviális, egyes becslések szerint [3] a demenciában szenvedő betegek akár 50%-a sem részesül a megfelelő diagnózisban. Ugyan léteznek a Mini Mental Teszthez hasonló, széles körben használt szűrővizsgálatok, ezek azonban többnyire nem elég érzékenyek ahhoz, hogy megbízhatóan kimutassák az enyhe kognitív zavart annak korai szakaszában. A nyelvi

memória tesztelésére irányuló szűrővizsgálatok hatékonyabbak az enyhe kognitív zavar felismerésében, azonban sok esetben tévesen betegnek diagnosztizálják az egyébként egészséges pácienst [4].

Mind a mai napig kutatás tárgyát képezi, hogy milyen tesztek és vizsgálatok képesek a legérzékenyebben kimutatni a korai Alzheimer-kórban és egyéb demenciákban megjelenő kognitív és nyelvi változásokat [5]. Noha a nyelvi képességek károsodása már a betegség igen korai szakaszában is jelentkezik, a nyelvi képességek értékelésére mégsem fordítottak kellő figyelmet az Alzheimer-kór diagnosztizálása során [6]. A demencia korai felismerésének és pontos diagnosztizálásának igen fontos szerepe van abban, hogy a szakszerű kezelés megindításával a betegség előrehaladása lelassuljon, illetve az új tünetek megjelenése minél később következzen be [7].

Ebben a munkában az enyhe kognitív zavarban szenvedő páciensek automatikus azonosítására törekszünk beszédátirataik alapján. A rendszer elsődlegesen beszélt nyelvi sajátosságokra, illetve a beszédátiratok automatikus morfológiai és szintaktikai elemzésén alapuló jellemzőkre épül.

Távlati célunk egy olyan automatikus rendszer kifejlesztése, amely képes az enyhe kognitív zavarra jellemző nyelvi tünetek időben történő detektálására, így a személy még időben megfelelő kezelésben tud részesülni. Fontosnak tartjuk azonban elmondani, hogy semmi esetre sem kívánjuk a pácienseket automatikusan diagnosztizálni, hiszen ez orvosi szakértelmet és gyakorlatot kívánó feladat. A mesterséges intelligencia eszközeivel azonban egyfajta szűrővizsgálatot tudunk létrehozni, melynek során kiszűrjük a rizikócsoportba sorolt pácienseket, akiket a későbbiek folyamán szakorvosok vizsgálnak meg, felállítva a tényleges diagnózist.

2. Anyagok

Vizsgálatainkban 69 személy beszédátiratait használtuk fel. A vizsgáltak mindegyike azonos feladatot kapott: spontán beszéd keretében fel kellett idézniük két rövid történetet, illetve a tegnapi napjukat. A vizsgálati személyek teljes nyelvi produkciójáról hangfelvétel készült. Az adatfelvételre minden esetben a szegedi memóriaambulancián került sor.

A fentebb ismertetett felvételek hanganyagához nyelvész szakértők kézzel készítettek beszédátiratokat. Jelenlegi kutatásainkban e kézi leiratok képezték nyelvtechnológiai vizsgálataink alapját, vagyis csakis írásbeli jellemzőkkel dolgoztunk, azonban a hanganyagok beszédtechnológiai vizsgálata is zajlik kutatásainkkal egyidejűleg.

Minden vizsgált személy esetében rendelkezésünkre állt a pontos orvosi diagnózis, azaz ismert volt, hogy az illető szenved-e enyhe kognitív zavarban vagy más demenciában. Ezen információk alapján két csoportba soroltuk a vizsgált személyeket: enyhe kognitív zavarban szenvedők (39 személy tartozott ide) és kontrollcsoport (30 személlyel). A nemek, illetve diagnózis szerinti megoszlást az 1. táblázat mutatja.

Vizsgálataink során a betegek személyes adatait teljes mértékben bizalmasan és az adatvédelmi előírásoknak megfelelően kezeltük.

1. táblázat. A betegek adatai.

	Enyhe kognitív zavar	**Kontroll**	**Összesen**
Férfi	14	11	25
Nő	25	19	44
Összesen	39	30	69

3. Nyelvi sajátosságok a beszédátiratokban

A beszédátiratokra jellemző nyelvi sajátságokat az alábbi példa segítségével mutatjuk be.

> *Tegnap* ***ö .. hát általába*** *én nyolc órakor kelek* ***...*** *fél kilenc körül szedem be a gyóccereimet tehát közbe eszek* ***ö*** *a gyóccerekre tehát mire az utolsó cukorgyógyszer is bekerül* ***... öhhöhh mm*** *kávét iszok* ***ö ... utánna .. feltkáva*** *ha van főzve akkor* ***ö*** *megiszom a gyerekektől maratt megiszom ha nem akkor teszek föl* ***...*** *De tegnap* ***ö*** *volt és* ***utánna*** *megittam és tettem föl* ***... utánna*** *vo elötte már bekapcsoltam a számítógépet mivel* ***öüöü hát nem könyvelés hanem tehát adatrögzitést*** *csinálok* ***... és akkor ö*** *tehát az küldöm be a cégnek* ***...*** *és* ***ö ...*** *eszt* ***ö*** *mekcsináltam* ***..*** *Közben az internet elment nálam és* ***ö akkor*** *mégegyszer megcsináltam a műveletet* ***... és akkor*** *kérdesztem a titkárnőt hogy* ***ö .. ement*** *az első üzenet, tehát* ***aa*** *fájlátvitel* ***... éss*** *ily* ***ez eszt*** *már délután kérdesztem* ***mer*** *egész nap ott volt* ***... höhö ...*** *nem egész nap hanem* ***olyan*** *jó délig ott volt mondom bisztos nem ér rá* ***mer*** *láttam hogy ott van a* ***szkájtnál ... és akkor öö ö ...***

A fenti beszédátirat jól tükrözi az élőbeszéd sajátosságait. Egyrészt számtalan, hezitációt, illetve néma szünetet jelölő formát tartalmaz (*ö*, *höhö*, *...*), másrészt mivel a beszédátiratok a kiejtést híven követik, találhatunk bennük fonológiai törléseket (*mer*, *ement*) és nyújtásokat is (*utánna*). Kettős szóindítások is előfordulnak (*ez ezt*) különféle szótévesztések mellett (*hát nem könyvelés hanem tehát adatrögzítést*), ezeken felül pedig a vizsgálati személyek által újonnan alkotott, és ily módon a nyelvhasználatban nem elterjedt egységeket is találhatunk (*feltkáva*).

A beszédátiratok vizsgálata arra is rámutatott, hogy érdemes figyelmet fordítani a töltelékszavakra is. Többek között a következő szavakat és kifejezéseket soroltuk ebbe a kategóriába: *ilyen*, *olyan*, *izé*, *és aztán*, *és akkor*, illetve a határozatlan névmásokat, úgymint *valamilyen*, *valahogy*, *valamerre*[1]. Úgy tűnik, hogy élőbeszédben az enyhe kognitív zavarban szenvedők gyakran helyettesítenek szavakat határozatlan névmásokkal vagy valamilyen töltelékszóval. Melléknevek helyett pedig előszeretettel használnak parafrázisokat. Ennek megfelelően nem ritkák az *egy ilyen bagolyszerűség* vagy *az olyan délelőtt volt* körülíró, bizonytalanságra utaló kifejezések.

[1] E szavak hasonlítanak a bizonytalanságot jelző ún. weasel és hedge szavakra [8].

4. Módszerek

A vizsgálati személyek státuszának automatikus megállapítására gépi tanulási kísérleteket végeztünk. A feladatra bináris osztályozásként tekintettünk: a vizsgálati személyt az enyhe kognitív zavarban szenvedő, illetve az egészséges csoportok valamelyikébe soroltuk be a rendelkezésre álló beszédátirataik alapján.

Első lépésben a beszédátiratokat automatikus nyelvi előelemzésnek vetettük alá a magyarlanc elemző [9] segítségével. Az elemzés eredményeképpen a szövegeket mondatra, illetve szavakra bontottuk, a szavakhoz morfológiai elemzést rendeltünk, illetve a mondatokhoz szintaktikai (függőségi) elemzést is társítottunk. Az osztályozáshoz többek között felhasználtuk a beszédátiratok automatikus elemzéséből gyűjtött morfológiai, szintaktikai és szemantikai jellemzőket is.

Minden egyes vizsgált személy három felidézési feladatot kapott. Mivel úgy gondoljuk, hogy memóriazavarról lévén szó maguknak a feladatoknak a sorrendje is hasznos információt hordozhat a személy státuszának megállapításában, az egyes feladatokhoz tartozó beszédátiratokat külön-külön dolgoztuk fel, azaz egy-egy beteg esetében három szöveggel dolgoztunk, és ezekben külön-külön vizsgáltuk az alább részletezendő nyelvi jellemzőket.

4.1. Felhasznált jellemzők

Vizsgálataink során számos, a beszédátiratokból, illetve azok automatikus nyelvi elemzéséből származó jellemzőt használtunk fel, melyek között találhatunk beszélt nyelvi, morfológiai és szemantikai jellemzőket is. Az alkalmazott jellemzőtér a következő volt:

- **Beszédjellemzők:**
  - kitöltött szünetek száma;
  - néma szünetek száma;
  - hezitációk száma;
  - hezitációk aránya;
  - névelőt követő szünetek száma;
  - nyújtások száma.
- **Morfológiai jellemzők:**
  - szavak száma;
  - írásjelek száma;
  - főnevek száma;
  - igék száma;
  - ismeretlen szavak száma;
  - ismeretlen szavak aránya.
- **Szemantikai jellemzők:**
  - bizonytalan szavak száma;
  - bizonytalan szavak aránya;
  - emlékezetre utaló kifejezések száma;
  - emlékezetre utaló kifejezések aránya.
- **Demográfiai jellemzők:**
  - nem;
  - születési év.

4.2. A jellemzők statisztikai elemzése

Statisztikai vizsgálatokat is végeztünk annak érdekében, hogy kiderítsük, mely jellemzők bírnak a legnagyobb megkülönböztető erővel. Ennek érdekében minden egyes jellemzőre és szövegre lebontva kétmintás t-próbát végeztünk az adott jellemző szerepét vizsgálva az enyhe kognitív zavarban szenvedők és a kontrollcsoport tagjainak elkülönítésében. Azt találtuk, hogy a jellemzők nagy része statisztikailag szignifikáns különbségeket mutat a két csoport között, az ezekhez tartozó szignifikanciaszinteket (p-értékeket) részletesen a 2. táblázat ismerteti.

2. táblázat. Statisztikailag szignifikáns jellemzők.

Jellemző	**1. szöveg**	**2. szöveg**	**3. szöveg**
szavak száma		0,0028	
hezitációk száma	0,0083	0,0019	0,0012
bizonytalan szavak száma	0,0188	0,0006	
ismeretlen szavak száma		0,0354	
hezitációk aránya	0,0033		0,0012
bizonytalan szavak aránya	0,0216	0,0007	
mondatbeli szavak száma	0,0133	0,0435	0,0404
néma szünetek száma	0,0073	0,0011	0,0024
nyújtások száma		0,0031	
főnevek száma			0,0331
írásjelek száma		0,0187	

A táblázatból kitűnik, hogy a hezitációk száma, mondatbeli szavak száma és a néma szünetek száma mindhárom szövegtípus esetében szignifikáns eltéréseket mutat a két csoport között. A bizonytalan szavak szintén fontos indikátornak tűnnek. Ezek alapján arra következtethetünk, hogy minél több hezitáció, illetve szünet található a beszédátiratban, illetve minél rövidebbek a mondatok és minél több a bizonytalan szó, annál nagyobb a valószínűsége, hogy a beteg enyhe kognitív zavarban szenved. A születési év, pontosabban az életkor is szignifikáns különbséget mutat: az 1943 előtt született személyek (vagyis akik a vizsgálat idején legalább 71 évesek voltak) nagyobb valószínűséggel szenvednek enyhe kognitív zavarban, mint az ennél fiatalabbak ($p < 0{,}0309$).

4.3. Gépi tanulási kísérletek

Az enyhe kognitív zavarban szenvedő személyek automatikus azonosítására gépi tanulási kísérleteket is végeztünk a beszédátiratokon. A Weka szoftver [10] segítségével több gépi tanuló algoritmust is kipróbáltunk, és az előzetes mérések alapján a legeredményesebbnek a döntési fa (C4.5) algoritmus [11] tűnt, valamivel meghaladva az SVM-mel [12] elért eredményeket, így a továbbiakban döntési fákat alkalmaztunk.

Méréseikben a fenti jellemzőket vettük alapul. 69 személy adataival dolgoztunk leave-one-out módszerrel, azaz 68 személy adatain tanítottuk a rendszert,

majd az így felépített modell alapján jósoltuk meg a hiányzó 1 státuszát. Ezt a folyamatot 69-szer ismételtük meg, amíg minden egyes személy státuszára kaptunk egy predikciót. A kiértékeléshez a pontosság (accuracy) metrikát alkalmaztuk, emellett a pontosság, fedés, F-mérték metrikákat is használtuk, osztályokra is kivetítve. Az eredmények a 3. táblázatban láthatók.

3. táblázat. Eredmények. EKZ: enyhe kognitív zavar, SVM: Support Vector Machine, C4.5: döntési fák, P: pontosság (precision), R: fedés, F: F-mérték, %: pontosság (accuracy).

	EKZ			**Kontroll**			**Teljes**			
Módszer	P	R	F	P	R	F	P	R	F	%
SVM	0,721	0,795	0,756	0,692	0,600	0,643	0,708	0,710	0,707	71,01
C4.5	0,794	0,692	0,740	0,657	0,767	0,708	0,735	0,725	0,726	72,46

Az egyes jellemzőcsoportok hozzáadott értékét is szerettük volna megvizsgálni. Ennek érdekében porlasztásos méréseket is végeztünk, melyek során egy adott jellemzőcsoportot kivettünk a gépi tanuló által használt adatok közül. Eredményeinket a 4. táblázat szemlélteti.

4. táblázat. Porlasztásos eredmények. EKZ: enyhe kognitív zavar, SVM: Support Vector Machine, C4.5: döntési fák, P: pontosság (precision), R: fedés, F: F-mérték, %: pontosság (accuracy).

Hiányzó	**EKZ**			**Kontroll**			**Teljes**				
jellemzők	P	R	F	P	R	F	P	R	F	%	kül.
beszéd	0,629	0,564	0,595	0,500	0,567	0,531	0,573	0,565	0,567	56,52	-15,94
morfológia	0,550	0,564	0,557	0,414	0,400	0,407	0,491	0,493	0,492	49,28	-23,18
szemantika	0,703	0,667	0,684	0,594	0,633	0,613	0,655	0,652	0,653	65,22	-7,24
demográfia	0,765	0,667	0,712	0,629	0,733	0,677	0,706	0,696	0,697	69,57	-2,89

5. Eredmények

A teljes jellemzőkészlet használatával 72,46%-os pontosságot értünk el a C4.5 algoritmussal, azaz a 69 esetből 50-szer állapított meg a rendszer pontos diagnózist. Az eredmények alapján van néhány olyan jellemző, amely igen fontosnak bizonyul a páciensek státuszának automatikus megállapításában. A legfontosabb jellemzőknek a következők bizonyultak: bizonytalanságot jelző szavak száma; hezitációk száma; szünetek száma; ismeretlen szavak aránya; főnevek száma.

Amennyiben összevetjük a két gépi tanuló által használt eredményeket, érdekes különbségeket láthatunk a két osztályt nézve. Az enyhe kognitív zavarban szenvedő páciensek megtalálásában jobban teljesít az SVM, mint a C4.5

algoritmus (0,795 fedési értékkel, szemben a 0,692-vel), a pontosság viszont alacsonyabb; a kontrollcsoport esetében viszont fordított a helyzet. Ha tehát az a célunk, hogy a lehetséges betegeknek minél nagyobb arányát fedjük le az automatikus szűrővizsgálattal (akiket aztán tovább lehet irányítani orvosi konzultációra), akkor talán célravezetőbb az SVM használata, ez a feltevés azonban további vizsgálatokat igényel.

Az egyes jellemzőcsoportok hozzáadott értékét megvizsgálandó porlasztásos méréseket is végeztünk a C4.5 algoritmussal. Ezek alapján a legtöbb hozzáadott értéke a morfológiai, illetve a beszédjellemzőknek van, ugyanakkor mindegyik jellemzőcsoport hozzájárult a rendszer pontosságának növeléséhez.

6. Összegzés

Ebben a munkában bemutattuk az enyhe kognitív zavarban szenvedő személyek automatikus azonosítását beszédátirataik alapján megcélző rendszerünket. A rendszer elsődlegesen beszélt nyelvi sajátosságokra, illetve a beszédátiratok automatikus morfológiai és szintaktikai elemzésén alapuló jellemzőkre épül. Megvizsgáltuk az egyes jellemzők megkülönböztető szerepét mind statisztikai, mind gépi tanulási szempontból. Az eredmények azt igazolják, hogy elsődlegesen a morfológiai jellemzők és a beszédjellemzők bírnak fontos szereppel a vizsgálati személyek státuszának automatikus megállapításában.

A későbbiekben szeretnénk adatbázisunkat újabb személyek beszédátirataival bővíteni, illetve gépi tanuló rendszerünket is továbbfejleszteni a minél nagyobb pontosság elérése érdekében. További terveink közé tartozik, hogy a hanganyagok beszédtechnológiai vizsgálatával és részletes elemzésével szerzett jellemzőkkel is kiterjesszük rendszerünket, ezáltal beszéd- és nyelvtechnológiai eszközök egyaránt hasznosulhatnak az enyhe kognitív zavar automatikus felismerésében.

Köszönetnyilvánítás

Jelen kutatást a Telemedicina fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatta, valamint a Bolyai János Kutatói Ösztöndíj. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Negash, S., Petersen, L.E., Geda, Y.E., Knopman, D.S., Boeve, B.F., Smith, G.E., Ivnik, R.J., Howard, D.V., Howard Jr, J.H., Petersen, R.C.: Effects of ApoE genotype and Mild Cognitive Impairment on implicit learning. Neurobiology of Aging **28**(6) (2007) 885–893
2. APA: DSM-IV-TR. American Psychiatric Association (2000)

3. Boise, L., Neal, M.B., Kaye, J.: Dementia assessment in primary care: Results from a study in three managed care systems. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences **59**(6) (2004) M621–M626
4. Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K., Kaye, J.: Spoken language derived measures for detecting mild cognitive impairment. Audio, Speech, and Language Processing, IEEE Transactions on **19**(7) (2011) 2081–2090
5. Chapman, S.B., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W., Burns, M.H.: Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. Alzheimer Disease & Associated Disorders **16**(3) (2002) 177–186
6. Bayles, K.A.: Language function in senile dementia. Brain and Language **16**(2) (1982) 265–280
7. Kálmán, J., Pákáski, M., Hoffmann, I., Drótos, G., Darvas, G., Boda, K., Bencsik, T., Gyimesi, A., Gulyás, Z., Bálint, M., et al.: Early mental test – developing a screening test for mild cognitive impairment. Ideggyógyászati szemle **66**(1-2) (2013) 43–52
8. Vincze, V.: Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 383–391
9. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
11. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
12. Cortes, C., Vapnik, V.: Support-vector networks. Volume 20. Kluwer Academic Publishers (1995)

Beszéd-zene lejátszási listák nyelvtechnológiai vonatkozása

Benyeda Ivett[1], Jani Mátyás[2], Lukács Gergely[2]

[1] MTA Nyelvtudományi Intézet,
1068, Budapest, Benczúr utca 33.
benyeda.ivett@nytud.mta.hu
[2] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
{jani.matyas, lukacs}@itk.ppke.hu

Kivonat: Az internetes és okostelefonos médiafogyasztás lehetővé és szükségessé teszi a tartalom személyre szabását. Hangalapú média esetén ezzel a lejátszási lista (playlist generation) témakör foglalkozik. A korábbi munkák a területen kizárólag a zene alapú lejátszási listákkal foglalkoztak, a beszéd-zene lejátszási listákkal foglalkozó első kutatások is az akusztikai oldalt vizsgálták. Jelen munka, úttörő módon, a beszéd-zene lejátszási listák készítésének nyelvtechnológiai oldalával foglalkozik. Az előzetes vizsgálatok alapján javaslatot tesz a beszéd-zene lejátszási lista készítésének vázára. A nyelvtechnológiai feldolgozásnál különösen a hangulati, érzelmi vonatkozásnak, ezek dalszövegekből, interjúátiratokból és hangzó beszédből való hangulatkinyerésének van jelentősége. Ehhez hangulati szótárakat használunk fel, hangulati szavak dalszövegekben és interjúátiratokban való előfordulását vizsgáljuk. A beszédet tartalmazó hanganyagok esetén a szöveg előállításához automatikus beszédfelismerést is végzünk, kétféle módon: a teljes hanganyag felismerésével, ill. a hangulati szavakra való fókuszálással. Vizsgáljuk, hogy a hangulati szavak előfordulását hogyan változtatja meg a beszédfelismerés korlátozott minősége. A munkát angol nyelvű szótárakkal és BBC anyagokon végeztük.

1 Bevezetés

Az Internet alapvetően megváltoztatta a médiafogyasztási lehetőségeket és szokásokat, mivel a felhasználók számára elérhető médiatartalom robbanásszerűen növekedett. Az egyéni válogatás időigénye és nehézkessége miatt szükségessé vált a médiatartalmat személyre szabni. A személyre szabott tartalom függhet az egyéni érdeklődéstől, az aktuális szituációtól, de a felhasználó (médiafogyasztó) közösségi kapcsolatoktól is. Az okostelefonok és a mobilinternet jelenlegi terjedése ezeket a trendeket – médiatartalmak elérhetőségét, az egyszerű kezelés iránti igényt és a személyre szabás lehetőségét – tovább erősíti.

A hangalapú médiát több szempont miatt is kiemelt jelentőségűnek látjuk: jóval nagyobb az információtartalma az írásos anyagokénál, a befogadáshoz szükséges erőfeszítés ugyanakkor jóval kisebb [1] és megengedi az egyidejű fizikai aktivitást, pl. a közlekedést, a házi vagy a házkörüli fizikai munkát. Ezt látszik alátámasztani a

személyre szabott online zenei rádiók, pl. Pandora, Spotify vagy last.fm megjelenése és gyors növekedése. A szokásos közösségi médiához képest érdekes alternatívát jelent a hangalapú közösségi média is. Az előzménynek tekinthető kisközösségi rádiók (angol: small community radio, low power radio station) közösségépítő hatására számos példa van a világban, az okostelefonos hangalapú közösségi média prototípusa egy kontrollált kísérletben 30-60%-kal növelte az emberi kapcsolatok számát [2].

A tartalom személyre szabása a hang- (vagy videó-) alapú folyamatos tartalomszolgáltatásnál nagyobb kihívást jelent, mint szöveges tartalom esetén. Nem elég a potenciálisan érdekes anyagok kiválasztása, hanem azok sorba rendezése is szükséges. A jó befogadási, hallgatási élmény alapvető egy ilyen tartalomszolgáltatás élvezhetőségéhez, elfogadottságához. A munka motivációját jelentő hangalapú kisközösségi média területén (1) beszédet és (2) zenét tartalmazó hanganyagokat is kezelni kell, ezek együtteséből kell a személyre szabott tartalmat kiválasztani.

A kihívás megoldásához a hang akusztikai és szöveges dimenzióját is érdemes vizsgálni a multimédiával kapcsolatos kutatásoknak megfelelően. Jelen munka az utóbbira vonatkozó első vizsgálatokat írja le, a beszédet tartalmazó hanganyagok leiratának, a dalszövegek és a hangzó beszédből való hangulatkinyerés vizsgálatával.

A munka felépítése a következő. A 2. fejezetben a tartalom személyre szabására vonatkozó szakirodalmat mutatjuk be. A 3. fejezetben a rádiós zenei szerkesztés gyakorlatának néhány kérdését vizsgáljuk. Ez alapján a nyelvi aspektusnál az érzelmek kezelése kulcsfontosságú, ezt a 4. fejezet tárgyalja. Az 5. fejezet a beszédfelismeréssel, a nyelvi reprezentáció szükségszerű minőségromlásával és ennek a hangulatfelismerésre való következményével foglalkozik, ezt követően összegezzük az eddigi eredményeket és tapasztalatokat, majd kitérünk a kutatás további terveire.

2 Kapcsolódó munkák

A tartalom személyre szabásának egyik alapeszköze az ún. ajánlórendszerek (recommender systems). Ezek célja olyan személyre szabott javaslatok készítése, amelyek az adott felhasználónak várhatóan tetszeni fognak. Ez történhet a tartalom- és az egyénprofil összehasonlításával, de akár – és sokszor ez a célszerűbb – kizárólag a többi felhasználó visszajelzései alapján az ún. collaborative filtering segítségével, pl. hasonló ízlésű felhasználók keresésével. Az ajánlórendszerek kutatási területén jelentős és gyakorlatreleváns eredményeket értek el, számos megközelítéssel [3]. Az ajánlórendszerek kimenete ugyanakkor nem rendezett úgy, hogy az lineáris médiához felhasználható legyen, erre a területen csak első próbálkozások ismertek [4,5].

A hanganyagok összeállítása folyamatos lejátszáshoz a szakirodalomban a lejátszási lista készítés (playlist generation) kifejezéshez kapcsolódik, egy aktuális áttekintés: [6]. A lejátszási listák készítésénél fegyelembe veendő tulajdonságokat többféleképpen csoportosítják, a szerzők meglátásaitól függően. Ezeket felhasználva [7] három szintet különböztet meg, és összegzi az egyes szinteken lényeges tulajdonságokat. Alulról felfelé haladva a szintek és tulajdonságok:

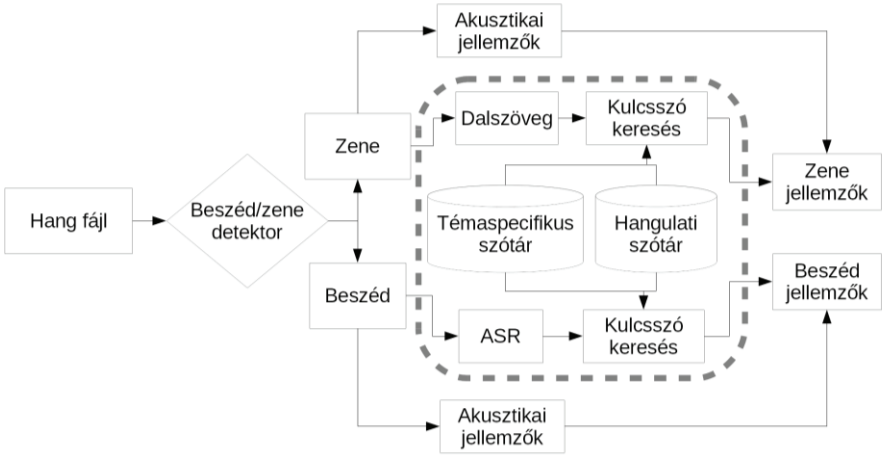
1. egyes dalok kiválasztása: frissesség-ismertség;
2. egymás utáni dalok kiválasztása: rendezettség-váratlanság;
3. a lejátszási lista egésze: koherencia-változatosság.

A fenti, lejátszási lista készítésére vonatkozó irodalom szinte kizárólag zenei lejátszási listákat kezel. A kevert beszéd-zene lejátszási listák készítése nyilvánvalóan különbözik ettől. Az utóbbiakra vonatkozó első kutatások [8] akusztikai szempontból vizsgálják az egymást követő beszéd-zene párokat. A [9]-ben leírt szabadalom sokkal átfogóbban, de a jelen munka szempontjából csak érintőlegesen foglalkozik a műsorválasztással, elsősorban források közötti átkapcsolással.

3 Megoldás tervezett felépítése

A rádiószerkesztés gyakorlatának vizsgálatához gyakorló zenei szerkesztőket kérdeztünk és a szöveges tartalom vizsgálatára egy, korábban a szakértői rendszerek felépítéséhez használt ún. korlátozott információs kísérletet (limited information experiment) [10] végeztünk. Ennek keretében előkészítettünk öt beszédet tartalmazó hanganyagot, majd elkészítettük ezek szöveges leiratát. A zenei szerkesztőnek elsőként csak a leiratot adtuk, és kértük, javasoljon zenét. Ezt követően az eredeti hanganyagot is lejátszottuk, azzal a kérdéssel, hogy mennyiben módosul a javasolt zene.

A tapasztalatok alapján csak a szöveg felhasználásával is jó minőségben tudott a szakember zenét ajánlani. A meghallgatás után ez csak kis mértékben módosult, pl. a beszéd tempója, vagy a beszélő kora, neme alapján. Az adatok kis száma ellenére is megerősödött így az a feltevés, hogy a beszédet tartalmazó hanganyagok szövegének jelentősége van a zenei szerkesztés szempontjából.



1. ábra. Hanganyagok jellemzőkinyerése automatikus beszéd-zene lejátszási lista készítéshez (ASR: beszédfelismerő).

A szakértőkkel való beszélgetésekből és a tesztekből kiderült, hogy szöveg-zene kapcsolódás esetén a legfontosabb illesztési szempont a hangulati jellemzők harmóniája, bizonyos speciális esetekben ezen kívül szerepet kap a témabeli illeszkedés. Ez többségében az ünnepkörök témáinál kívánatos. A cikkben leginkább a szöveg alapú

hangulatkinyeréssel foglalkozunk, a végső felhasználásban azonban a témabeli jellemzők is szerepet fognak játszani a feldolgozásban. Az 1. ábrán a hanganyagok jellemzőinek kinyeréséhez készített feldolgozási terv látható, kiemelve a cikk fókuszában található nyelvi feldolgozó részt.

4 Hangulatkinyerés írott szövegből

4.1 Dalszövegek és interjúátiratok előfeldolgozása

Bár a zenék hangulatának kinyeréséhez eredetileg kizárólag az akusztikai tulajdonságokat használták fel, a frissebb kutatások [11,12], illetve [13] kimutatták, hogy a dalszövegek felhasználása nagyban javítja a kategorizálás minőségét, sőt, sok esetben felülmúlja az akusztikai alapon végzett besorolás teljesítményét. Természetesen a dalszövegek felhasználását kívánatos ötvözni az akusztikai tulajdonságok feldolgozásával, a 3. fejezetnek megfelelően. Nem kizárólag azért, mert így érhető el a legjobb minőség, hanem azért is, mert nem minden zeneanyag szöveges, illetve nem minden anyag szövege hozzáférhető, nem megemlítve azt a tényt, hogy bizonyos esetekben a szöveg tartalmilag vagy terjedelmileg nem megfelelő a feldolgozásra.

Szintén a dalszövegek felhasználása mellett szól, hogy beszerzésük egyszerű és olcsó, szövegeik pedig többségében jól felhasználhatók a kívánt célra. Előbbi sajnos nem jellemző a beszédes tartalmakra. Bár megtalálhatók és hozzáférhetők interjúleiratok, ezek többségében nem elegendőek ahhoz, hogy az adatbázis legalább jó részét lefedő leiratokat szerezzünk, hiszen a tárolni kívánt beszédfelvételekhez aktualitásukból adódóan ritkán érhetőek el leiratok. Az ennek megoldására való kísérletek és a probléma részletezése az 5. fejezetben található. Ezen részben azokkal az esetekkel foglalkozunk, ahol – valamilyen szerencse folytán – rendelkezésre áll hivatalos leirat.

A dalszövegek feldolgozása bizonyos szempontból sajátos előkészítést igényelhet. Talán a legszembetűnőbb különbség az átlagos szövegek és a dalszövegek között (a sorok rövid volta mellett) az ismétlések jelzése. Ha olyan formában kívánjuk feldolgozni a szöveget, ahogyan az a dalban megjelenik, az ismétléseknek a szövegekben újra szerepelniük kell. Bár ezek a szövegek tulajdonképpen hivatalos leiratok és a különböző oldalakról kinyert anyagok egészen megegyezőek, az ismétlések jelzésére nincs kialakult módszer (‘repeat’, ‘4x’, ‘ref.’, ‘ref x2’), így nem is könnyű kezelni. Amellett, hogy többféle módon jelzik, meg kell állapítani az ismétlés hatókörét. Ez általában egy strófa vagy egy sor, attól függően, hol helyezkedik el a jelzés. Többségében ezek a jelzések az ismételni kívánt sor vagy sorok után helyezkednek el, azonban néha előttük. Előfordul, hogy a refrént sem írják ki, csak ennyivel jelzik: ‘ref.’ Ezek mind figyelmet érdemelnek, ha a feldolgozást valóban a teljes hangzó szövegen akarjuk végezni. Emellett azokat a részeket, amik nyilvánvalóan nem a dal szövegéhez tartoznak, mint az említett jelzések, sok esetben érdemes kiszűrni. Bár a dalszövegeket feldolgozó tanulmányok többségében nem foglalkoznak ezekkel a kérdésekkel, érdemes megfontolnunk, hogyan kezeljük az említett sajátosságokat. [14] és [15] normalizálja mindezeket és csak ezután következnek a feldolgozás további lépései.

Beszédleiratokban, legfőként interjúátiratok esetében meglehetősen sok kiszűrendő adat található, elég, ha csak arra gondolunk, hogy minden beszélőváltásnál kiírják az

éppen megszólaló nevét. Ezen kívül többnyire az átiratok elején szerepel egy kisebb bevezető, hogy mikor milyen műsorban hangzott el a beszélgetés egyéb kapcsolódó információk mellett.

Nagyobb anyag feldolgozásánál sok esetben szükséges, hogy felismerjük, milyen nyelvű a szöveg, legyen az akár dalszöveg, akár valamilyen beszédleirat. A kísérleteinkhez ezt a dalszövegeken el kellett végeznünk, mivel a használt adatbázisban különböző nyelven íródott dalok találhatók meg, a mérésekhez kizárólag az angol nyelvűeket használtuk fel. Az interjúátiratok esetében ezt nem tettük meg, mivel a BBC interjúátirataival dolgoztunk, melyek mind angol nyelvűek. Később szükséges lesz más nyelvű adatokkal is foglalkozni, így a nyelvfelismerés különösen fontos, leginkább az olyan további lépések esetében, ahol nyelvfüggő eszközt használunk (tipikusan ilyen a tokenizálás és lemmatizálás - amik szinte minden esetben szükségesek).

A dalszövegek felépítésének sajátosságához tartozik, hogy általában inkább csak mondattöredékeket tartalmaznak, nem pedig teljes mondatokat, írásjeleket ritkán találni, akkor is inkább figyelemfelhívó szerepben jelennek meg, mint mondathatárként. Emellett a mondatok, ha vannak is, sokszor sorokon átívelnek, így a mondatra bontás szinte lehetetlen, a Part of Speech taggelés igen nagy kihívás és kétséges, hogy megéri-e az eredmény az idő- és energiabefektetést.

4.2 A szövegekből való hangulatkinyerés módszerei

A hangulatkinyerés módszerét meglehetősen meghatározza, milyen típusú feldolgozást kívánunk végezni. Kategorizálni szeretnénk az anyagokat hangulati szempontból vagy egyszerűen metaadatként hozzáadjuk a hangulati jellemzőiket. Utóbbi többségében többdimenziós hangulati modell esetében jellemző. Ilyenkor a különböző dimenziókhoz tartozó értékeket rendelik az anyaghoz (ezek kétdimenziós modell esetén: pozitív-negatív érték, illetve az aktivitás mértéke). Jellemzőbb azonban a zenei hangulatkinyerésre, a hangulati kategóriába való besorolás. A használt kategóriarendszerek változatossága azonban igen nagy. A kategóriák számát illetően talán a legrobosztusabb a hat Ekman-féle [16] hangulati alapkategóriát használó, klasszikusnak mondható klasszifikálás. Ezek a következők: düh, undor, félelem, öröm, szomorúság, meglepettség. [15] 18 hangulati kategóriát különít el a last.fm címkéi alapján. Talán a legjellemzőbb továbbra is az Ekman-féle rendszer, valószínűleg részben azért, mert ez a pszichológiában klasszikusnak tekintett hangulati felosztás, illetve mert a jól felhasználható hangulati szótár, a WordNetAffect szintén ezt követi.

A dalszövegek hangulati feldolgozásában alapvetően kétféle módszer mérvadó leginkább. Az egyik a hangulati szótárakon alapuló, a másik pedig a gépi tanulásos módszer. Mindkettő esetében a meglévő és felhasználható eszközök igen függnek attól, milyen kategóriarendszert szeretnénk használni. Jónéhány hangulati szótár szabadon elérhető. Ilyenek a WordNetAffect (WNA) [17], a Linguistic Inquiry and Word Count (LIWC)[18], Affective Norms for English Words (ANEW) [19], illetve elérhető néhány más nyelvű szótár is, bár sok esetben ezeket az angol forrásokból fordítják.

A felügyelt gépi tanulás szintén jó eredményeket tud adni, hátránya azonban, hogy meglehetősen nagy méretű gold standard korpusz szükséges a tanuláshoz, amit igen ritkán adnak közre, ez a szükséges korpuszméret csökkenthető a hangulati szótárak felhasználásával. Felügyelt gépi tanulásnál többféle jellemzőt is figyelembe vesznek.

[14] elsősorban a TF*IDF módszert használja fel a tanuláshoz, de olyan globális jellemzőket is használnak, mint a szövegek és sorok hossza karakterben számolva.

A vizsgálatokban az Ekman-féle kategóriarendszert használtuk és a WordNetAffect hangulati szótárával dolgoztunk. A későbbiekben lehetséges, hogy az applikációban gyűjtünk mintát a hangulati kategorizálásra, így létrehozva egy gold standard korpuszt a későbbi módszerek kipróbálásához.

A hangulati szótárak használata bizonyítottan használható a dalszövegek hangulati kategorizálására, interjúátiratokra azonban eddig nem voltak kísérletek. Mivel a feldolgozás szempontjából a legegyszerűbb, ha az interjúszövegek és dalszövegek kategorizálása a lehető leghasonlóbban történik, arra voltunk kíváncsiak, hogy az interjúszövegek esetében is található-e annyi hangulati szó, hogy a találatok alapján bekategorizálhatók legyenek.

1. táblázat. A hangulati szavak gyakorisága dalszövegeknél (D: düh, U: undor, F: félelem, Ö: öröm, Sz: szomorúság, M: meglepettség).

id	D	U	F	Ö	Sz	M	összes hangulati szó (db)	lemmák száma (db)	hangulati szavak aránya a szövegben (%)
	hang. szavak aránya (%) a szövegben								
5710	0	0	0	8	0	92	13	194	6,70
…	…	…	…	…	…	…	…	…	…
5978	0	0	100	0	0	0	1	135	0,74
5980	0	0	0	36	27	36	11	368	2,99
						átl.:	**9**	**283**	**3,17**

Az 1. táblázat a dalszövegekben talált hangulati szavakat, a hangulati megoszlások arányában, illetve az összes hangulati szó számát a dalszövegben, a dalszöveg szószámát, illetve annak értékét, hogy a dalszöveg hány százalékát teszik ki hangulati szavak (utolsó oszlop). Látható, hogy ez az érték a vizsgált mintában 3,17%-os átlaggal szerepel. Ez azon okból fontos, hogy a hangulati kategorizálás általában kizárólag a talált hangulati szavak számából (illetve azok arányából) számítódik, azonban az, hogy milyen mértékben megbízható a történt besorolás, az függ attól, hogy milyen gyakorisággal szerepelnek hangulati szavak a szövegben. A kísérlet egy 200 elemű mintán folyt, a szövegeken nyelvfelismerést hajtottunk végre a Python langid.py [20] nyelvfelismerőjével és csak az angol szövegűeket dolgoztuk fel (174 szöveg). Ezeket a szövegeket a Python NLTK moduljának felhasználásával tokenizáltuk, majd lemmatizáltuk (WordNetLemmatizer). A szövegeket nem normalizáltuk, a refrének és ismétlések úgy szerepelnek, ahogy eredetileg voltak, ezt követően a szózsákokban kerestettük a hangulati szótárak elemeit. Az előfeldolgozásnál kipróbáltuk a stopszavak kiszűrését, de rontotta az eredményeket, minden valószínűség szerint a többszavas hangulati kifejezések esetében okozhatott ez hibát.

A 2. táblázat az előző táblázattal azonosan épül fel, a szövegfeldolgozás is ugyanazokat a lépéseket tartalmazta, azzal a különbséggel, hogy kiszűrtük a szövegből az olyan sorokat, melyek nem az interjú szövegéhez tartoztak (beszélők megjelölése,

interjú elejét és lezárását jelző címkék stb.). A BBC ‘Andrew Marr show’[1] műsoraiban lévő interjúk átiratait dolgoztuk fel, 282 átiratot elemeztünk a hangulati szavak megjelenésének szempontjából. Látható, hogy viszonylag sok hangulati szó szerepel az egyes átiratokban, ami a hangulati besorolás szempontjából nagyon jó, azonban látható, hogy a szövegek itt sokkal hosszabbak, mint a dalok esetében.

2. táblázat. A hangulati szavak gyakorisága interjúátiratoknál.

id	D	U	F	Ö	Sz	M	összes hangulati szó (db)	lemmák száma (db)	hangulati szavak aránya a szövegben (%)
	hang. szavak aránya (%) a szövegben								
1	0	0	8	29	14	49	49	1752	2,80
…	…	…	…	…	…	…	…	…	…
281	3	0	10	27	30	30	30	1997	1,50
282	0	0	0	36	27	36	30	1291	2,32
						átl.:	**28**	**2239**	**1,26**

Bár a hangulati kategorizálásnál nem veszik figyelembe a szövegek hosszát, csak a hangulati szavak megoszlásának arányát, a besorolás megbízhatósága függ attól, hogy az egész szövegben mekkora a hangulati szavak gyakorisága. A mérés itt igen nagy különbséget mutat a dalszövegek eredményeivel összehasonlítva, míg a dalszövegeknél a hangulati szavak százalékos aránya 3,17%, az interjúátiratoknál csak 1,26%. Ezek az értékek azt mutatják, hogy az interjúátiratoknál is használhatók a hangulati szótárak, azonban az eredmények kevésbé megbízhatók.

Érdekes még a feldolgozás szempontjából, hogy bár a kisebb gyakoriság miatt talán kevésbé megbízható az interjúátiratokra történő hangulati kategorizálás, nem volt olyan interjú, ahol ne találtunk volna hangulati szót. Tehát itt elvileg minden esetben lehetséges a hangulati szótár használata, a dalszövegek esetén azonban több alkalommal előfordult, hogy a szövegben egy hangulati szó sem szerepelt, így a csak hangulati szótár alapján történő besorolás ezen esetekben nem lehetséges. Lehetséges, hogy a hangulati szavak kis gyakorisága az interjúkban azok formális voltára vezethető vissza. Az interjúszövegek általában kifejtettebbek és sok olyan elemet tartalmaznak, amik kizárólag a társalgás fenntartására szolgálnak, pl.: szó átadása, köszönés stb., mely részek többnyire nem tartalmaznak hangulati szót.

5 Hangulatkinyerés hangzó szövegből

5.1 Bevezetés

A beszéd hangulatának tartalom alapján történő felismeréséhez a beszéd leiratára van szükség. Ez a legtöbb esetben nem áll rendelkezésre, ennek automatikus elkészítéséhez beszédfelismerő rendszert alkalmaznak. Az automatikus beszédfelismerő rendsze-

[1] http://www.bbc.co.uk/programmes/articles/3hshxFhHM4dKd3px6Q3NzRF/transcripts

rek a tanítóhalmaz és a kísérlethez használt felvétel paramétereitől függően (beszélők, akusztikai paraméterek, stb.) nagyon változatos eredményeket produkálhatnak a felismerési hibák szempontjából. A tanítóhalmazhoz hasonló felvételeken kevesebb hibát produkál, mint az attól jobban eltérőkön.

A hibák számszerűsítésére a szófelismerési hiba (word error rate) mértéket szokták használni. Minél kevesebb hibát okoz a beszédfelismerő, annál jobb eredményt kaphatunk a szótár alapú hangulatfelismerés szempontjából. Érdekes jelenség, hogy a hangulati szavak felismerésének növelésével nem lineárisan javul a hangulatfelismerés eredménye [21].

Kísérleteinkben a BBC 'In Touch' nevű műsorának fölvételeit és leiratait használtuk. Két módszert vizsgáltunk meg a hangulati szavak megtalálásához: a beszédfelismerő által felismert (legvalószínűbbnek számolt) szöveget használtuk a korábban ismertetett módon, illetve kulcsszókeresési eljárást alkalmaztunk a beszédfelismerő által generált, több valószínű útvonalat is tartalmazó hipotézis gráfon (lattice).

5.2 Felvételek

A felvételeket és a leiratokat a BBC In Touch[2] műsorának honlapjáról töltöttük le. Itt az utolsó öt adás mp3 formátumú hanganyaga és leirata érhető el. A hangfelvételek nem igényeltek különösebb előfeldolgozást, csak az elejéről kellett levágni a szerzői jogi információkat, valamint a beszédfelismerő tanulóhalmazához illeszkedve 16 KHz-es egycsatornás formátumba kellett konvertálni.

A leiratok általában pdf vagy rtf formában érhetőek el. Ezekben a szövegen kívül jelölve vannak a beszélőváltások is. A leiratozás nem gépi beszédfelismerést szem előtt tartva készült, ezért például a számok, a pénznemek nincsenek kiírva, az egyszerre beszélés, érthetetlen részek, stb. nincsenek jelölve, illetve ahol jelölve vannak ott sem egységes módon. Mivel nem volt kapacitás az ilyen jellegű hibák kijavítására, ezért minimális kézi előfeldolgozás után a problémák többségét (pl. számok átírása) automatikusan oldottuk meg.

5.3 Beszédfelismerés, kulcsszókeresés

A kísérletekhez a Kaldi beszédfelismerő rendszert használtuk [22]. A triphone modelleket változtatás nélkül a Kaldiban található előre elkészített TEDLIUM példakódokkal tanítottuk. Az akusztikai modell tanító adatbázisa a TEDLIUM[3] első verziója volt [23], nyelvi modellnek a CMUSphinx projekt által készített amerikai angol n-gram nyelvi modellt[4] használtuk.

A kulcsszókeresés és a beszédfelismerő által visszaadott legvalószínűbb útvonal esetén is ugyanazt a dekódolt hipotézis gráfot használtuk. Mindkét esetben tízszeres súlya volt a nyelvi modellnek az akusztikai modellhez képest. A beszédfelismerőnél a szófelismerési hibára (WER) 47,2%-ot kaptunk.

[2] http://www.bbc.co.uk/podcasts/series/intouch

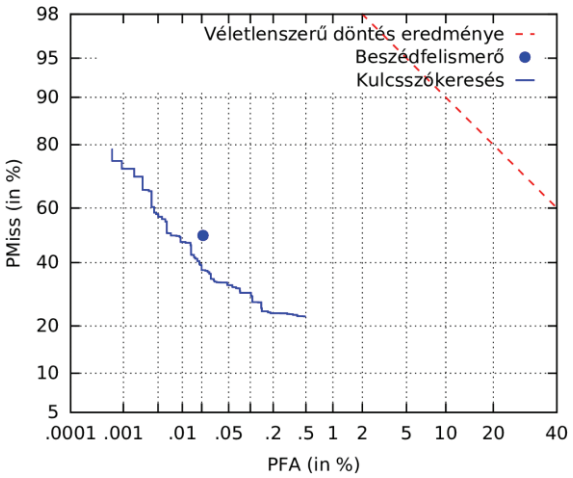
[3] http://www.openslr.org/resources/7/TEDLIUM_release1.tar.gz

[4] http://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/US%20English%20Generic%20Language%20Model/cmusphinx-5.0-en-us.lm.gz/download

A kulcsszókeresésnél a Kaldiba beépített eljárást használtuk, ami a hipotézis gráfot indexeli a hivatkozott cikkben leírt módon [24]. Eredményül egy listát készít a megtalált kulcsszavakból, az elhangzás időpontjával és hosszával, valamint egy nulla és egy közötti értékkel, ami a találat bizonyosságát jelzi. Ez utóbbi alapján egy adott küszöbérték megadásával lehet eldönteni, hogy melyik találatokat tekintjük érvényesnek.

5.4 Eredmények

A két módszer – felismert szöveg és kulcsszókeresés – eredményét a nyers szövegen végzett kísérlet referenciaeredményével hasonlítottuk össze.



2. ábra. DET (Detection Error Tradeoff) görbe; a kulcsszókeresési módszer görbéje jó küszöbérték mellett kevesebb hibásan megtalált (PFA) és kevesebb nem megtalált (PMiss) kulcsszót eredményez (folytonos görbe), mint a beszédfelismerő által felismert szövegben való keresés (görbe feletti pont).

A módszerek eredményeit kulcsszókeresési algoritmusok kiértékeléséhez használt módszerek segítségével értékeltük ki [25], ezt az F4DE[5] eszközzel készített diagram (2. ábra) szemlélteti. Jó küszöbérték megválasztásával kevesebb hibásan megtalált (false alert) és kevesebb nem megtalált (miss) kulcsszót kaptunk a kulcsszókeresési módszerrel, mint a beszédfelismerő által felismert szöveg használatának esetén.

Mindkét módszernél kiszámoltuk a megtalált kulcsszavak érzelmi kategóriák szerinti eloszlását felvételenként, ezt hasonlítottuk a nyers szövegben történő keresésnél kapott eloszláshoz. Annak ellenére, hogy az előző kiértékelés alapján a kulcsszókeresést találtuk pontosabbnak, a hangulati szavak kategóriánkénti megoszlását vizsgálva a beszédfelismerő által felismert szövegben keresve az arányok jobban hasonlítottak a referenciaarányokhoz, mint bármelyik küszöbérték mellett a kulcsszókeresés esetén (3. táblázat).

[5] http://www.nist.gov/itl/iad/mig/tools.cfm

3. táblázat. a hangulati szavak eloszlása felvételenként a referencia, a beszédfelismerő által felismert szöveg és kulcsszókeresés esetén.

	Referencia (leirat)							Beszédfelismerő							Kulcsszókeresés						
Id	D	U	F	Ö	Sz	M	db	D	U	F	Ö	Sz	M	db	D	U	F	Ö	Sz	M	db
	%							%							%						
1021	3	0	14	63	6	14	35	4	0	16	60	9	11	45	0	0	22	61	9	9	23
1028	6	0	0	49	14	31	49	8	0	0	60	2	30	53	4	0	0	68	4	24	25
1104	19	0	0	44	6	31	48	12	0	0	47	17	25	60	19	0	0	51	11	19	37
1111	8	0	3	63	8	18	38	4	0	4	56	16	20	45	5	0	5	53	11	26	19
1118	4	0	7	52	9	28	46	0	0	8	46	12	33	48	0	0	7	52	11	30	27

A kulcsszókeresés rosszabb eredménye az eloszlás vizsgálatánál vagy a kevés adaton végzett vizsgálatnak köszönhető, vagy pedig annak, hogy a beszédfelismerés konzekvensen azonos arányú hibát okoz az egyes érzelmi kategóriák esetén. Annak eldöntéséhez, hogy melyik esetről van szó, nagyobb adathalmazon végzett további vizsgálatok szükségesek.

6 Összegzés és további tervek

A kutatás során kiderült, milyen módon képzelhető el az adatbázisban való hanganyagok összeillesztése és milyen szempontok játszanak szerepet a zene-beszéd listák létrehozásában. Kiderült, hogy a zene-beszéd illesztésnél a hangulati jellemzők a leginkább fontosak, megvizsgáltuk a hangulatkinyerés lehetséges módszereit és felmértük, mennyire lehet hatékony a hangulati szótár alapú kategorizálás interjúátiratok esetében. Ez alapján úgy tűnik, hogy bár interjúátiratok esetében kisebb a hangulati szavak gyakorisága (valószínűleg formális nyelvezetéből adódóan), mint az a dalszövegeknél szerepel, elég hangulati szót találni bennük a kategorizálásra, a dalszövegekkel ellentétben, ahol nem ritka, hogy átlagos hosszúságú szövegben egy hangulati szó sem szerepel, így ez alapján nem kategorizálható be.

Később azon általános esettel foglalkozunk, ahol a hangzó szövegből kell kinyernünk a hangulati szavakat, átirat hiányában. Az általunk vizsgált két módszer – a beszédfelismerő által felismert szövegben történő keresés és a kulcsszókeresés – jól közelíti a leiratban (referencia) való keresés eredményét. A kapott eredmények alapján nem lehet egyértelműen megállapítani, hogy melyik a jobb módszer. Ezt további, több adaton végzett vizsgálat segítségével lehetne kideríteni.

A továbbiakban figyelembe kívánjuk venni a szavak TF*IDF-értékét a hangulati szavaknál, így tehát azok a szavak nagyobb hangulati súllyal szerepelnének, melyek megkülönböztető szerepe nagyobb a szövegben, emellett részletes kidolgozásra kerül majd a hangulat akusztikai alapú felmérése, hiszen a végső felhasználásban a szöveges és akusztikai feldolgozás együtt szerepel majd.

Mivel szöveg-zene összekapcsolódásnál, bár csak bizonyos esetekben, de fontos a specifikus témák felismerése, ezt is kezelni kívánjuk. A megoldás a terv szerint hasonlóan működne, mint a hangulati szótárak alapján való besorolás, a különböző témákra, melyek fontosnak bizonyulnak (karácsony, újév stb.) egy-egy szótárat készítenénk, ezek alapján felismerve, hogy az anyag a témák valamelyikébe tartozik-e.

Hivatkozások

1. Kock, N.: The evolution of costly traits through selection and the importance of oral speech in e-collaboration. Electron. Mark. 19 (2009) 221–232
2. Lukacs, G., Pethesné, D. B., Madocsai, B.: Impact of Personalized Audio Social Media on Social Networks. In: XXXIII. Sunbelt Social Networks Conference of the International Network for Social Network Analysis Abstract Proceedings, Hamburg, Germany (2013) 210
3. Ricci, F., Shapira, B.: Recommender Systems Handbook. Springer (2011)
4. Zibriczky, D., Hidasi, B., Petres, Z., Tikk, D.: Personalized recommendation of linear content on interactive TV platforms: beating the cold start and noisy implicit user feedback. In: Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP), Montreal (2012)
5. Felfernig, A., Friedrich, G., Gula, B., Hitz, M., Kruggel, T., Leitner, G., Melcher, R., Riepan, D., Strauss, S., Teppan, E., Vitouch, O.: Persuasive Recommendation: Serial Position Effects in Knowledge-Based Recommender Systems. Persuasive Technology, Springer, Berlin / Heidelberg (2007) 283–294
6. Fields, B., Lamere, P.: Finding A Path Through The Jukebox – The Playlist Tutorial, ISMIR. Utrecht (2010)
7. Benyeda, I.: Zene–beszéd lejátszási listák készítésének nyelvtechnológiai vonatkozása. Pázmány Péter Katolikus Egyetem (2014)
8. Jani, M., Lukács, G., Takács, Gy.: Experimental Investigation of Transitions for Mixed Speech and Music Playlist Generation. In: Proceedings of ACM International Conference on Multimedia Retrieval, Glasgow, United Kingdom (2014) 392–398
9. Bull, W., Rottler, B.: Auto-station tuning, US Patent 8634944B2, Apple Inc, Cupertino, CA, US (2008)
10. Hoffmann, R.R.: The Problem of Extracting the Knowledge of Experts from the Perspective of Experimental Psychology 8 (1987) 53–67
11. Hu, X., Downie, J. S.: When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In: Proceedings of the 10th International Conference on Music Information Retrieval, Utrecht, The Netherlands (2010) 619–624
12. Hu, X., Chen, X., Yang, D.: Lyric-based Song Emotion Detection with Affective lexicon and Fuzzy Clustering Method. In: 10th International Society for Music Information Retrieval Conference (ISMIR 2009) (2009)
13. Mihalcea, R., Strapparava, C.: Lyrics, Music, and Emotions. In: Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn (2012) 590–599
14. van Zaanen, M., Kanters, P.: Automatic Mood Classification Using TF*IDF Based on Lyrics. In: Proceedings of the 11th International Society for Music Information Retrieval Conference (2010)
15. Hu, X., Downie, J. S., Ehmann, A. F.: Lyric Text Mining in Music Mood Classification. (2009) 411–416
16. Ekman, P.: Facial expression of emotion. Am. Psychol. 48 (1993) 384–392
17. Strapparava, C., Valitutti, A.: WordNet-Affect: an Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (2004) 1083–1086
18. Pennebaker, J.W., Martha, E.F., Booth, R.J.: Linguistic Inquiry and Word Count. Mahwah, NJ (2001)
19. Bradley, M. M., Lang, P. J.: Affective Norms for English Words (ANEW): Instruction manual and affective ratings, http://www.uvm.edu/~pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf (1999)

20. Lui, M., Baldwin, T.: langid.py: An Off-the-shelf Language Identification Tool. In: 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea (2012)
21. Chuang, Z.-J., Wu, C.-H.: Multi-Modal Emotion Recognition from Speech and Text. Comput. Linguist. Chin. Lang. Process. 9 (2004) 45–46
22. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding , Hilton Waikoloa Village, Big Island, Hawaii, US (2011)
23. Rousseau, A., Deléglise, P., Estève, Y.: TED-LIUM: an Automatic Speech Recognition dedicated corpus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey (2012)
24. Can, D., Saraclar, M.: Lattice Indexing for Spoken Term Detection. IEEE Trans. Audio Speech Lang. Process. 19 (2011) 2338–2347
25. Fiscus, J. G., Ajot, J., Garofolo, J. S., Doddingtion, G.: Results of the 2006 spoken term detection evaluation. In: Proc. SIGIR (2007) 51–57

VII. Poszterbemutatók

Gyógyszermellékhatások kinyerése magyar nyelvű orvosi szaklapok szövegeiből

Farkas Richárd[1], Miklós István[1], Tímár György[2], Zsibrita János[1]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
Szeged, Árpád tér 2.
{rfarkas,mikist,zsibrita}@inf.u-szeged.hu
[2] Comfit Kft.
gyorgy.timar@comfit.hu

1 Kivonat

A magyar nyelvű orvosi szaklapok gyakran közölnek ún. esettanulmányokat, melyben leírják, hogy bizonyos hatóanyagok (ágensek), például gyógyszerek hatására pácienseknek milyen tünetei (szimptómák) voltak megfigyelhetőek. A Comfit Kft. és a Szegedi Tudományegyetem együttműködésében megvalósuló projekt azt tűzte ki célul, hogy megvizsgálja a nyelvtechnológia lehetőségeit és a jelenleg kézi átolvasással történő elemzése helyett egy automatikus megoldást szolgáltasson.

A projekthez rendelkezésre állt 4600 magyar nyelvű orvosi szakcikk és a bennük megtalálható ágens-tünet összerendelések, melyet a Comfit Kft. munkatársai az elmúlt években manuálisan gyűjtöttek.

Első lépésben a PDF formátumban lévő újságcikkekből kellett a szöveges tartalmakat kinyerni. Itt komoly gondot okozott a többhasábos szerkesztés és a grafikonok, hirdetések nagy száma. A hasábok azonosítására egy, a szóközök sűrűségét figyelő algoritmussal találtunk megoldást. Egy következő lépésben a dokumentumokat megszűrjük. Például kidolgoztunk egy gépi tanulási módszereken alapuló bibliográfiablokk-azonosító modult, amire azért volt szükség, mert a hivatkozások címei gyakran tartalmaztak ágensmegnevezéseket, ami félrevezette az információkinyerő rendszert.

Maga az információkinyerő rendszer három nagy részből áll össze. Először a szövegben azonosítjuk a potenciális hatóanyag- és tünetemlítéseket, majd megállapítjuk, hogy melyik tünet mely hatóanyagokra vonatkozik és végül az azonosított említéseket az adatbázis azonosítóira kell leképeznünk.

Az egyik legnagyobb problémát az okozta, hogy a tanító adatbázis egy szövegbeli előfordulástól független adattáblaként állt rendelkezésre, míg az információkinyerő rendszernek szövegbeli említések pontos helyének (mondatkörnyezet stb.) ismerete szükséges. Hogy ezeket az említéseket azonosítsuk szótárakat és rövidítéslistákat illesztettünk a szövegekre [1]. Ezeket használtuk utána a gépi tanuló algoritmusok tanító adatbázisaként. A tanulás során keletkezett szekvenciajelölő modell használatával még nem látott újságcikkekben is felismerhetjük a hatóanyagokat és szimptómákat [2]. A hatóanyagok és tünetek közti relációk azonosításánál azzal a problémával szembesültünk, hogy ismét csak azonosító alapján voltak a párok jelölve, így egy párt a szövegben a két fél minden előfordulása reprezentálta. Ezért nem hagyatkozhattunk

csupán a párok egymástól független osztályozására, hanem figyelembe kellett venni minden egyedet. Erre globális optimalizálásban használt és különböző gépi tanulással segített eljárásokkal kísérleteztünk [3].

Hivatkozások

1. Farkas, R., Dobó, A., Kurai, Z., Miklós, I., Miszori, A., Nagy, Á., Vincze, V., Zsibrita, J.: Információkinyerés magyar nyelvű önéletrajzokból a nexum Karrierportálhoz. In: X. Magyar Számítógépes Nyelvészeti Konferencia (2014)
2. Móra, Gy., Farkas, R.: Szótáralapú, névelem-felismerés szóhatárainak javítása gépi tanulási módszerrel. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (2010) 317–324
3. Björne, J., Ginter, F., Salakoski, T.: University of Turku in the BioNLP'11 Shared Task. BMC BioInformatics Volume 13 Supplement 11 (2011)

Elliptikus listák jogszabályszövegekben

Hamp Gábor[1], Syi[1], Markovich Réka[2,3]

[1] BME Szociológia és Kommunikáció Tanszék 1111 Budapest, Egry József u. 1.
hampg@eik.bme.hu, i@syi.hu
[2] ELTE Filozófiatudományi Doktori Isk., Logika Tanszék 1088 Budapest, Múzeum krt. 4/I
[3] BME Üzleti Jogi Tanszék, 1111 Budapest, Magyar Tudósok körútja 2.
markovich@phil.elte.hu

Kivonat: Kutatásunk jogszabályszövegek gépi feldolgozására irányul. A jogszabályokat adatbázisba töltve tízezres korpuszt hoztunk létre, majd kidolgoztunk egy olyan algoritmust, amely – döntően reguláris kifejezések segítségével – képes a szerkezeti egységek azonosítására, típusba sorolására, a jogszabály szerkezeti struktúrájának feltárására, a grafoelliptikus felsorolások (Goody-listák) megtalálására és az ellipszisek feloldására. Az ilyen elliptikus listák jelentőségét jelzi, hogy a korpuszunk közel harmadát teszik ki. A jogszabályszerkesztésre vonatkozó jogszabály megadja, hogy a Goody-listáknak négy típusa lehet, ezért megvizsgáltuk, hogy milyen nyelvi eszközök állnak rendelkezésre a típusok azonosításához. Azt találtuk, hogy a természetes nyelv felszíni szerkezetének elemzése könnyen hibás értelmezésekhez vezethet a tipizálás során. A problémás esetek helyes megoldásához deontikus és kijelentéslogikai eszközöket is igénybe vettünk, hogy – olykor a hétköznapi intuícióinkkal ellentétes – logikailag érvényes végeredményt kaphassunk. Kutatásunk tanulsága, hogy a jogszabályok elemzése során a nyelvtechnológiai eszközök mellett szükség van logikai apparátus használatára is.

Jack Goody, a torontói kommunikációelméleti iskola tagja az írásbeliség elemzése során megemlíti, hogy az írásnak vannak olyan technikai lehetőségei, amelyek a szóban kifejezett nyelvi kommunikáció számára nem állnak rendelkezésre [5]. Példaként hivatkozik a **listákra**, **táblázatokra** vagy a **keresztrejtvények** sajátos elrendezéseire. Bennük is a természetes nyelv mondatai reprezentálódnak, de úgy, hogy közben a beszédhez képest többre vagyunk képesek. Az írásbeliségre jellemző megoldásokat Goody **grafonyelvi technikáknak** nevezi. Közös jellemzőjük, hogy segítségükkel az írás kétdimenziós, vizuális terében olyan információt is reprezentálni lehet, amelyet a beszélt nyelv linearitásába nem tudunk megragadni, kifejezni. Goody grafonyelvi technikái közül számunkra a listák lesznek érdekesek a következőkben, mivel ezek értelmezhetők speciális ellipszistechnikaként is. [14].

> *3. § E törvény alkalmazásában*
> *a) dohánytermék: bármilyen módon fogyasztásra szánt, részben vagy egészben dohányból készült termék, […]*
> *c) fiatalkorú: aki a tizennegyedik életévét betöltötte, de a tizennyolcadikat még nem…* [2]

Ez a lista(részlet) azért mondható ellipszisnek, mert a két listatétel önmagában nem értelmezhető teljes mondatként. Csak akkor érthetjük meg a jogalkotói szándékot, ha a két listapontot egyenként összeolvassuk az idézet elején álló, bevezető szöveggel. A

lista bevezető része („E törvény alkalmazásában”) csak egyszer szerepel, de ezt a lista értelmezése során többször „használjuk” egymás után. A jogszabályszövegekben nagyon nagy arányban fordulnak elő ilyen listák.

A jogszabályok a hétköznapi nyelvhasználathoz képest formálisabb természetes nyelvű szövegek. Ez a formalizáltság különösen megnyilatkozik a jogszabályok szerkezeti tagolásában. Rendelet szabályozza a jogszabályszerkesztés szigorú rendhez kötött menetét. [1]

A jogszabályokat pontosan definiált szerkezeti egységekre kell felosztani, és rögzítve van az is, hogy a jogszabály „… szerkezeti egységeit folyamatos sorszámozással vagy a latin ábécé betűivel meg kell jelölni. Jogszabály tervezete jelöletlen szerkezeti egységet nem tartalmazhat” [1 (37. § (1))]. Ezt a minősítést értelmezhetjük úgy is, hogy a jogszabályok **jelölt szerkezeti egységekből** állnak. A vonatkozó rendelet az alábbi szerkezeti egységeket nevezi meg.

36. § (2) Jogszabály tervezetében alkalmazható szerkezeti egység…
a) az alpont,
b) a pont,
c) a bekezdés,
d) a szakasz,
e) az alcím,
f) a fejezet,
g) a rész és
h) a könyv. [1]

A jogszabályok szerkezeti egységei között meglepően nagy számban vannak **jelölt listák**, olyan elliptikus felsorolások, amelyek elemei önmagukban nem alkotnak teljes mondatot, így a pontos jelentésük kibontásához szükség van az ellipszisek feloldására. A jogszabályszerkesztésre vonatkozó normaszöveg előírásai miatt bizonyos szerkezeti egységek esetén nem lehetséges az ellipszistechnikák alkalmazása, ugyanis a **bekezdések** (és az ennél „magasabb szintű” szerkezeti egységek) esetében nem megengedett a bekezdésen túlnyúló hatókör. [1 (46.§)] Ez a tilalom nem engedi meg a bekezdések szintjén, tehát a bekezdések közti ellipszis alkalmazását. Ebből következően csak a pontok és alpontok esetében lehetséges az ellipszistechnika és a jelölt listák igénybevétele.

Ezeket a jelölt listákat **Goody-listáknak** nevezzük. Mivel a Goody-listák strukturált ellipszistechnikának minősíthetőek, ezért a leírásukban érdemes megjeleníteni azokat a strukturális jellemzőket, amelyek segíthetnek minket a listákban rejlő teljes mondatok felépítésében. A Goody-listáknak mindig van egy (és csak egy) **fejtétele**, amely felvezeti a lista többi elemét. Ezt követik a **listatételek**, amelyekből legalább kettőnek léteznie kell ahhoz, hogy listáról beszélhessünk. Nem mindig fordul elő, de elképzelhető, hogy a listatételeket egy önálló sorral zárják le, amely viszont ugyanolyan szereppel rendelkezik, mint a lista fejtétele. A lista elején álló elemet nevezhetjük **nyitófejtételnek**, az egész szerkezetet lezáró komponenst pedig a lista **zárófejtételének**. A vonatkozó jogszabály az általunk nyitófejtételnek nevezett konstrukciót „nyitó szövegrész”-nek, míg a zárófejtételt „záró szövegrész”-nek nevezi [1 (47.§ (1))]. Amennyiben a jogszabályok szövegében be tudjuk azonosítani a nyitófejtételt (és, ha van, akkor a zárófejtételt), valamint a felsorolás tételeit, akkor a

fejtétel(eke)t és az egyes listatételeket sorban összetéve (konkatenálva) teljes mondatokat kaphatunk.

Kutatásunkban arra a feladatra vállalkoztunk, hogy a jogszabályszövegek gépi elemzésével megtaláljuk a Goody-listákat, és az elliptikus elemeket feloldjuk. Az elemzés első lépéseként létrehoztunk egy jogszabálykorpuszt úgy, hogy egy PostgreSQL adatbázisba betöltöttük hat jogszabály teljes szövegét. Minden szerkezeti egység egy-egy rekordba került. A több, mint tízezres rekordszám azért jött létre, mert a hat jogszabály között volt a polgári törvénykönyv is, amely az átlagosnál jóval hosszabb törvénynek számít. A korpuszban reguláris kifejezések segítségével tipizálást végeztünk a rekordokon, amelynek eredményeként az egyes rekordokat szerkezetiegység-típusokba soroltuk be. A tipizálás során a reguláris kifejezéseken keresztül megragadható morfológiai jellemzőkön túl olykor figyelembe kellett venni strukturális információt is, ú. m. a szerkezeti egységek sorrendiségét, a rekordok egymáshoz való viszonyát is. A következő lépésben azonosítottuk a korpusz Goody-listáit és a listák szerkezeti elemeit – ismételten reguláris kifejezések segítségével (néha strukturális információ igénybevételével). A 10 500 rekordnyi korpuszban 540 Goody-listát találtunk úgy, hogy ebben összesen 3000 rekord volt érintett. Tehát a jogszabályok közel harmada Goody-listába rendezett (tehát elliptikus) szöveg. Az elliptikus fejtételeket és listatételeket egymáshoz illesztésével teljes mondatokat képeztünk (reguláris kifejezésekre támaszkodó SQL-parancsok segítségével), így a 3000 Goody-lista rekordból 2200 szemantikailag is érvényes állítást vagy másként: Goody-mondatot kapunk. Egy-egy Goody-listából annyi szemantikailag teljes mondatot „állíthattunk össze", ahány listatétel szerepel benne összesen.

A listatételek és a fejtételek egymáshoz való kapcsolódását két szinten is elemezhetjük. Egyfelől megvizsgálhatjuk, hogy milyen a viszony a fejtétel és a listatételek között, ha utóbbiakat együttesen, egyetlen összetett állításként vesszük figyelembe. Nevezzük ezt **külső kapcsolatnak**, amivel azt fejezhetjük ki, hogy a listatételek milyen logikai művelettel kapcsolódnak (egyenként is, egészében is) a hozzájuk képest „külső" fejtételhez. A listatételek közti viszonyt viszont nevezzük **belső kapcsolatnak**, amivel az „egyenrangú" listatételek egymáshoz kapcsolódásának minőségét jellemezhetjük. Elsőként foglalkozzunk a belső kapcsolatokkal!

Egy adott Goody-listából jól képzett Goody-mondatok valamilyen logikai művelettel kapcsolódnak egymáshoz, és együttesen lefedik az eredeti Goody-lista teljes szemantikai tartalmát. A megvizsgálandó kérdés itt az lehet, hogy milyen logikai kapcsolatok léteznek a Goody-listákon belüli listatételek (illetve a Goody-mondatok) között. A jogszabályszerkesztésről szóló jogszabály – a felsorolás tételei közti logikai kapcsolat szerint – négyféle típust különít el.

> *7. § Felsorolás alkalmazása esetén egyértelművé kell tenni, hogy a felsorolás elemei közül*
> *a) valamennyinek teljesülnie kell,*
> *b) egyik sem teljesülhet,*
> *c) pontosan egynek kell teljesülnie vagy*
> *d) legalább egynek teljesülnie kell … a joghatás kiváltásához.* [1]

A fenti típusokat úgy értelmezhetjük, hogy azok a felsorolt tételek mint kijelentések közt érvényesített logikai műveletek minőségében térnek el egymástól. Amikor valamennyi felsorolt elemnek teljesülnie kell (*a)* pont), akkor logikai **konjunkcióról**

(„és”-kapcsolatról) van szó. A *d)* pont esetében logikai **diszjunkció** („megengedő vagy”-kapcsolat) köti össze az elemeket, míg a *c)* pont a **biszubtrakcióra** (a „kizáró vagy”-kapcsolatra) utal. Végül: amikor egyik elem sem teljesülhet (ez a *b)* pont elvárása), akkor a **konnegáció** műveletéről beszélhetünk. [13] A pontos és egyértelmű értelmezés érdekében nézzünk meg pár példát mindegyik típusra! A legkevesebb problémát a konjunkcióval összekötött mondatok adják. A következő idézet erre szolgáltat példát.

6. § (5) A felsőoktatási intézmény a működését akkor kezdheti meg, ha
a) a fenntartó kérelmére a felsőoktatási intézmények nyilvántartását vezető szervtől (a továbbiakban: oktatási hivatal) megkapta a működési engedélyt, nyilvántartásba vették és
b) az Országgyűlés döntött az állami elismeréséről. [3]

Ez az eset azért tartozik a **konjunktív** felsorolások közé, mert a teljes lista igazságához az kell, hogy a felsorolás mindegyik eleme teljesüljön. Ennek a típusnak fontos minősége az, hogy a következmény (joghatás) nem jön létre, amennyiben bármelyik „feltétel” nem teljesül. Ezzel – bizonyos értelemben – ellentétes a **konnegációs lista** logikája, amikor a felsorolás egésze akkor igaz, ha a benne szereplő tételek mindegyike hamis. A következő két listában a listaelemek konnegációs viszonyban állnak.

4:108. § Nem lehet az apaság vélelmét megtámadni, ha
a) a származás reprodukciós eljárás következménye …; vagy
b) az apaságot bíróság állapította meg. [4]

A konnegációs összetételt úgy lehet a „legkönnyebben” felismerni, ha a listatételekre alkalmazható érezzük a ‘sem-sem’ formula. A fenti idézetben szereplő fő állítás („az apaság vélelmének megtámadhatósága”) akkor lehet igaz, ha sem egyik, sem másik listatétel nem igaz (vagyis a származás nem reprodukciós eljárás következménye és az apaságot nem állapította meg bíróság). Ugyanezen logikával értelmezhetjük a következő idézet tartalmát is.

4:127. § (2) Nincs szükség az örökbefogadáshoz az örökbefogadó házastársának hozzájárulására, ha
a) a házastárs cselekvőképtelen vagy ismeretlen helyen tartózkodik; vagy
b) a házastársak között az életközösség megszűnt. [4]

Amennyiben a joghatás kiváltásának lehetséges feltételei „kizáró vagyos” viszonyban vannak, amikor a lehetséges feltételek közül mindig pontosan egy (egy és csak egy) feltétel érvényesülését „várjuk el”, akkor a feltétel közti logikai kapcsolat a **biszubtrakció**.

106. § (4) Egyazon módosító rendelkezéssel kell az egymást közvetlenül követő, azonos időpontban módosuló,
a) egyazon magasabb szintű szerkezeti egységbe tartozó vagy
b) egyetlen magasabb szintű szerkezeti egységbe sem tartozó, azonos szintű szerkezeti egységeket újraszabályozni. [1]

A két feltétel egyszerre sosem teljesülhet, ezért ez a – két elemű – felsorolás mint egész akkor lesz igaz, ha valamelyik listatétel igaz, a másik pedig nem. Ez a konstrukció alkalmazható akkor is, ha több, mint két tételt kapcsolunk össze egymással.

Ekkor is csak egy tétel lehet igaz úgy, hogy eközben az összes többi elemnek hamisnak kell lenni ahhoz, hogy a felsorolás egésze igaz lehessen.

Az egyik leggyakoribb és talán legegyszerűbbnek tűnő logikai művelet a **diszjunkció**, ám jogi környezetben meglepő állítás fogalmazható meg ezzel kapcsolatban. A belső kapcsolatokban gyakran előfordulnak vagylagos felsorolások, ám amikor a listatételek diszjunkcióját a külső kapcsolat mentén fel akarjuk bontani, akkor meglepő jelenségeket tapasztalhatunk. Egyfelől nem (feltétlen) segítenek a felszíni jegyek az elemzésben, másfelől furcsa paradoxonok adódnak bizonyos kapcsolatokban. A probléma érzékeltetésére nézzünk meg két jogszabályrészletet. Mindkettő ugyanabból a törvényből származik, a jogi tételek címzettjei, előírásai is hasonlóak egymáshoz, ám a felsorolás elemei közti logikai kapcsolat jelzésére más kifejezést alkalmaznak a két listán belül. Az első listában az ‘és’ terminus használata a konjunkció mondatkapcsolódási típusát „sejteti”.

2:28. § (1) A gondnokság alá helyezést a bíróságtól
a) a nagykorú együtt élő házastársa, élettársa, egyenesági rokona, testvére;
b) a kiskorú törvényes képviselője;
c) a gyámhatóság; és
d) az ügyész
kérheti. [4]

Ha a jogszabályrészlet valós helyzetben történő alkalmazását képzeljük el, akkor tűnhetne úgy is, hogy a fenti passzust – az ‘és’ konjunkciós kapcsoló jelenléte ellenére is – inkább diszjunkcióként, azaz vagylagos lapcsolatként kellene értelmeznünk. A következő, alakilag a fentihez rendkívül hasonló részlet is erősítheti ezt az interpretációs lehetőséget, hiszen ebben még a feltüntetett kapcsolóelem is a diszjunkciót sugalló ‘vagy’ terminus.

2:30. § (2) A gondnokság alá helyezés megszüntetését a bíróságtól
a) a gondnokolt;
b) a gondnokolt együtt élő házastársa, élettársa, egyenesági rokona, testvére;
c) a gondnok;
d) a gyámhatóság; vagy
e) az ügyész
kérheti. [4]

Ha példákat keresünk a mindennapi életből olyan helyzetekre, amelyekben a fenti két jogi passzus alapján kell eljárni, akkor mindkét esetben mondhatjuk azt, hogy akárki kéri is a felsoroltak közül akár a gondnokság alá helyezést, akár annak megszüntetését, a listán szereplők közül bárki élhet ezzel a jogával, tehát egyáltalán nem szükséges az, hogy minden felsorolt szereplő egyszerre nyújtson be kérelmet a szabályozandó kérdésben. Ez a gyakorlati megfontolás újra csak azt erősíti, hogy ezekben az esetekben diszjunkcióról kellene beszélnünk. Azonban minden „sugallat” és a gyakorlati példákra való hivatkozás ellenére ezekben az idézetekben konjunkcióként kell értelmeznünk a listatételek közötti kapcsolatokat. Állításunk igazolásának első lépésként fel kell hívnunk a figyelmet a jogszabályok ama minőségére, hogy az ezekben szereplő mondatok mind normatív jellegűek, amit legtöbbször deontikus operátorok alkalmazásával fejezünk ki. A fenti két idézetben a kulcsmozzanat a zárófejtételekben szereplő ‘kérheti’ predikátum, amelyben „tetten érhető” a megenge-

dés minősége (tehát egy deontikus operátor jelenléte). A deontikus logikában már évtizedek óta ismert az a tétel, ami a most tárgyalt jelenséget is megmagyarázza, hogy ti. a diszjunkcióval összekapcsolt propozíciókra együttesen érvényesített megengedő operátort úgy bonthatjuk fel összetevőkre, hogy a megengedő operátor hatókörében levő diszjunkció helyett konjunkciót kell írnunk. [15], [7], [8], [6], [9] Idézzük fel itt Jennings egyik példáját! Ha vesszük a következő mondatot:

'Fred vagy Bill jöhet.'

akkor a mondatban a két szereplőre egyszerre vonatkozó megengedés ('jöhet') úgy fejezhető ki a két szereplőre külön-külön, ha azt mondjuk, hogy:

'Fred jöhet és Bill jöhet'

nem pedig úgy, hogy:

'Fred jöhet vagy Bill jöhet'

mert ez utóbbi összetett mondat akkor is igaz lehet, amikor valamelyik összetevője hamis (tehát akár Fred, akár Bill nem jöhet), viszont ezt a értelmezést az első – belső összetételű – mondat még nem tette lehetővé. Ebből pedig azt a következtetést kell levonnunk, hogy nem a diszjunkciós, hanem a konjunkciós szétbontás a megfelelő megoldás. A korábban idézett deontikus logikai cikkekben ezt nevezik 'free choice permission' jelenségnek. A tételt a következő formulával reprezentálhatjuk:

$$\mathbf{P}(A \vee B) \Leftrightarrow \mathbf{P}(A) \wedge \mathbf{P}(B) \qquad (\mathbf{1})$$

ahol a **P**-vel jelöljük a megengedés deontikus operátorát, A-val és B-val a propozíciókat (a jog által szabályozandó cselekvéseket leíró kijelentéseket), illetve a szokásos módon hivatkozunk a két, szóban forgó logikai kapcsolóra (illetve a modális kijelentések közti ekvivalenciarelációra).

Csak akkor érthetjük meg ezt a furcsa jelenséget, ha tekintetbe vesszük a jogrendszernek azt a minőségét, hogy a jogszabályokban megfogalmazott előírások címzettjei az állampolgárok és a jogalkalmazók (ítélkezők, bírók) egyaránt. A gondnoksággal kapcsolatos két fenti jogszabályrészletet úgy kell értelmeznünk, hogy az az ítélkező személynek szól, és amikor a bírónak döntenie kell, hogy megfelelő ember kéri-e a gondnokság alá helyezést (vagy a gondnokság megszüntetését), akkor neki a felsorolás(ok) mindegyik elemének igazságát kell feltételeznie. Bárki jelentkezhet a felsoroltak közül, a bírónak az idézetben rögzített módon kell eljárnia, és ezt csak a konjunkciós művelet tételezésével tudjuk megfelelőképpen leírni. Mivel ebben a jelenségben fontos, hogy a felszíni szerkezet diszjunkciót sugall, miközben ténylegesen konjunkcióról kell beszélnünk, ezért elnevezhetjük „**diszjunktív konjunkciónak**", hogy jelezni tudjuk azt, hogy itt nem a hagyományos értelemben vett, „tiszta" konjunkcióról van szó. Mindez azt jelenti, hogy az ilyen, első látásra diszjunktív szerkezetű listák valójában konjunkciók, és eltekinthetünk (sőt, el kell tekintenünk) attól, hogy a formális jellemzőik (például a tételek közé betett kapcsolószavak, mint a 'vagy' vagy az 'és' használata) mit sugallnak önmagukban.

A jogi korpusz elemzésekor találhatunk olyan példákat is, amelyek újra csak diszjunkciónak látszó konjunkcióként értelmezhetőek, miközben nincs a szerkezeten belül megengedő operátor, tehát más típusként vagy altípusként kell minősítenünk a fent definiált diszjunktív konjunkcióhoz képest. Előfordulnak például olyan listák, amelyekben nincs semmilyen felszíni utalás a deontikus jellegre (nincs a mondatban deontikus operátor), de azért tudjuk, hogy kötelezésről van szó. Ilyenkor **rejtett** vagy **látens kötelezésről** beszélhetünk. Vegyük a következő példát.

2:33. § (2) A gyámhatóság a gondnokot a tisztségéből elmozdítja, ha a gondnok
a) a kötelezettségét nem teljesíti;
b) nem az előzetes jognyilatkozatban foglaltak szerint jár el; vagy
c) egyéb olyan cselekményt követ el, amellyel a gondnokolt érdekeit súlyosan sérti vagy veszélyezteti. [4]

Az idézetben szereplő ‘elmozdítja’ kifejezést úgy kell értelmeznünk, hogy az ítélkezőnek (itt: a gyámhatóságnak) ‘el kell mozdítania’ a gondnokot, ha bármelyik feltétel teljesül a felsorolásból. Itt újra csak arról van szó, hogy a felsorolás egésze csak akkor igaz, ha az összetétel felbontása során a listatételek közti kapcsolatot konjunkcióként értelmezzük (hiába vannak diszjunkcióra utaló felszíni jelek). A konjunkciót azonban már mással magyarázhatjuk az előző típushoz képest. Itt ugyanis egy bikondicionális külső kapcsolat van a fejtétel és a listatételek között, amelyben a listatételek a fejtétel érvényességi feltételeinek számítanak. E kapcsolatnak az alábbi a logikai szerkezete:

$$H \leftrightarrow (A \vee B) \Leftrightarrow (A \rightarrow H) \wedge (B \rightarrow H) \wedge ((A \vee B) \leftarrow H) \qquad (2)$$

ahol H a fejtétel, míg A, B és C a listatételekben megfogalmazott feltételeket jelenti. Ez az ekvivalencia mutatja meg, hogy az ilyen típusú feltételfelsorolások a jogszabályokban a szükséges-elégséges feltételek megjelölését jelenti. Azt, hogy ez a számítógépes szövegfeldolgozás szempontjából mit jelent, az ekvivalencia jobb oldalából olvashatjuk ki. Ezen az oldalon a nyilak irányára az a magyarázat, hogy míg a listaelemek egyenként elégséges feltételül szolgálnak ahhoz, hogy a fejtételben jelölt kötelezés beálljon, a teljes képhez számításba kell vennünk, hogy valamelyik listatétel bekövetkezése szükséges feltétel (ott ezért nem a feltételek felől megy a nyíl, mint az elégséges feltételek esetében a kondicionális logikai kapcsolónál, hanem a másik irányba, retrokondicionális viszonyt jelölve) [10] (12)]. Ahhoz, hogy erre a fejtétel-feltétel viszonyra ekként tekintsünk, használjuk azt a előfeltevést, hogy egy adott jogkövetkezmény beálltához szükséges valamennyi feltételt egyazon listában, kimerítően felsorol a jogalkotó (taxatív felsorolást végez). A felbontás után az egyes – újonnan megkapott – egész mondatoknál figyelnünk kell arra, hogy a felszíni szerkezetükben retrokondicionálist fognak ugyan mutatni, kondicionálisok lesznek, ahogyan azt az ekvivalencia jobb oldalának első két tagjánál látjuk. Nem kevés ilyen bikondicionális diszjunkció szerepel a Goody-listák között. Sok esetben a fejtétel végén szereplő ‘ha’ kötőszóval egyértelműen jelzik is ezt a feltételes kapcsolatot, de előfordul olyan eset is, amikor ez elmarad.

3:68. § (1) A tagsági jogviszony megszűnik
a) a tag kilépésével;
b) a tagsági jogviszony egyesület általi felmondásával;
c) a tag kizárásával;
d) a tag halálával vagy jogutód nélküli megszűnésével. [4]

Ezt a felsorolást úgy is értelmezhetjük, hogy a listaelemek külön-külön, mind a fejtételben megragadott jogi konstrukció, jogi fogalom (itt: a ‘tagsági jogviszony megszűnése’) érvényességének feltételeit jelentik. Ismét kiemeljük, hogy pusztán a ‘ha’ kötőszó szerepeltetésével egy retrokondicionális viszonyt jelölünk, amennyiben a feltételek hátra vannak vetve: ahhoz, hogy tudjuk, bikondicionálisról van szó, hasz-

nálnunk kell azt az előfeltevést, hogy kimerítő a felsorolás. Gyakran előforduló jogalkotói megoldás azonban a példálózó felsorolás, melyet rendszerint a 'különösen' szó közbeiktatásával jelölnek a fejtételben. Ilyen esetekben bizonyosan nem beszélhetünk retrokondicionálisról annak ellenére sem, hogy hátra vannak vetve a feltételek. Kondicionális viszony áll fenn az egyes feltételek és a fejtételben elhelyezett jogkövetkezmény között, hiszen nem kell valamelyik feltételnek teljesülnie ahhoz, hogy beálljon a jelzett jogkövetkezmény, viszont ha teljesül valamelyik feltétel, biztosan számolhatunk a jogkövetkezménnyel. Így például:

2:43. § A személyiségi jogok sérelmét jelenti különösen
a) az élet, a testi épség és az egészség megsértése;
b) a személyes szabadság, a magánélet, a magánlakás megsértése;
c) a személy hátrányos megkülönböztetése;
d) a becsület és a jóhírnév megsértése;
e) a magántitokhoz és a személyes adatok védelméhez való jog megsértése;
f) a névviseléshez való jog megsértése;
g) a képmáshoz és a hangfelvételhez való jog megsértése. [4]

Bármelyik feltétel bekövetkezése a személyiségi jogok sérelmét jelenti, de a felsoroltakon kívül még számos egyéb lehetőség is van, ami azt jelenti, így ezen feltételek megvalósulása nem szükséges, csak elégséges, vagyis kondicionális viszonnyal van dolgunk. Az ilyen eseteket az alábbi formula írja le.

$$(H \leftarrow (A \vee B \vee C)) \Leftrightarrow ((H \leftarrow A) \wedge (H \leftarrow B) \wedge (H \leftarrow C)) \quad (3)$$

Az IRM rendelet [1 (7. § (2))] előírja, hogy a különböző felsoroláselemek (a mi listatételeink) közötti viszony vonatkozásában az utolsó előtti listaelem után kell szerepelnie a kötőszónak. Így a logikai kapcsolat megtalálásához elegendő lenne elvileg ezt megnéznünk a vonatkozó rekordban. Azonban, mint fentebb láttuk, korántsem az, hiszen a tényleges logikai kapcsolatot befolyásolja, hogy a lista fejtételében milyen deontikus operátor van, illetve az, hogy kondicionális normáról van-e szó. Vagyis az automatizált feldolgozáskor ezeket a tényezőket is számításba kell vennünk.

*

A tanulmány a 83887. sz. OTKA kutatás keretén belül készült.

Hivatkozások

1. A jogszabály szerkesztésről szóló 61/2009. (XII. 14.) IRM rendelet
2. A gazdasági reklámtevékenység alapvető feltételeiről és egyes korlátairól szóló 2008. évi XLVIII. törvény
3. A nemzeti felsőoktatásról szóló 2011. évi CCIV. törvény
4. A Polgári Törvénykönyvről szóló 2013. évi V. törvény
5. Goody, J.: Nyelv és írás. In: Nyíri, K., Szécsi, G. (szerk.): Szóbeliség és írásbeliség. Áron, Budapest (1998) 189–221
6. Jennings, R. E.: Can There Be a Natural Deontic Logic? Synthese 65/2 (1985) 257–273

7. Kamp, H.: Free Choice Permission. Proceedings of the Aristotelian Society, New Series 74 (1973–1974) (1974) 57–74
8. Makinson, D.: Stenius' Approach to Disjunctive Permission. Theoria 50/2–3 (1984) 138–147
9. Makinson, D.: On a Fundamental Problem of Deontic Logic. In: McNamara, P., Prakken, H., eds.: Norms, Logics and Information Systems. New Studies in Deontic Logic and Computer Science, IOS Press (1999) 29–53
10. Markovich, R., Hamp, G., Syi: A kondicionálisok problémája jogszabályszövegekben. In: Tanács, A., Varga, V., Vincze, V., eds.: X. MSZNY Konferencia, Szeged (2014) 295–302
11. Markovich, R., Hamp, G., Syi: Errata parva. Jogszabályszövegek gépi elemzésének tanulságai (kutatási műhelytanulmány) (2014)
12. Madarászné, Zs. A., Pólos, L., Ruzsa, I.: A logika elemei. Budapest, Osiris (2006)
13. Syi: syi.hu/cse. L'Harmattan – Könyvpont Kiadó (2014)
14. Syi, Hamp, G., Markovich, R.: Goody-listák jogszabályszövegekben (műhelytanulmány) (2014)
15. von Wright, G. H.: An Essay on Deontic Logic and the Theory of Action, Amsterdam (1969)

FinUgRevita: nyelvtechnológiai eszközök fejlesztése kisebbségi finnugor nyelvekre

Horváth Csilla[1], Kozmács István[2], Szilágyi Norbert[2], Vincze Veronika[3], Nagy Ágoston[1], Bogár Edit[1], Fenyvesi Anna[1]

[1]Szegedi Tudományegyetem, Angol-Amerikai Intézet
[2]Szegedi Tudományegyetem, Finnugor Tanszék
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
e-mail:finugrevita@gmail.com

Kivonat A jelenleg is futó nemzetközi FinUgRevita projekt (2013-2017) keretében olyan nyelvtechnológiai eszközöket fejlesztünk, melyek a kis finnugor népek közülük a manysik (vogulok) és udmurtok (votjákok) nyelvének digitális és online jelenlétét teszi lehetővé, és segíti az anyanyelvi beszélőket és a tanulni vágyókat a nyelvi kommunikáció különféle színterein. A kezdeti fázisban a két nyelv jelenkori leíró nyelvtanai alapján nyelvtani kivonatok készülnek, melyek a készülő online morfológiai elemző szabályrendszerét adják, míg az eddig megjelent nyomtatott szótárak szkennelésével, OCR-es elemzésével és manuális javítással az udmurt esetében 13000, míg a manysi esetében előreláthatólag 10-15000 szavas elektronikusan felhasználható szótár készül. A morfológiai elemző futtatásához és további nyelvtechnológiai eszközök fejlesztéséhez az interneten szabadon elérhető udmurt és manysi nyelvű tartalmakból nagy tokenszámú korpuszt építünk. A projekt célja, hogy a készülő eszközök online szabadon hozzáférhetőek legyenek az udmurt és manysi nyelvek beszélőinek és tanulóinak számára, és nem utolsó sorban kutatási célokra is alkalmazhatóak legyenek.

Kulcsszavak: udmurt, manysi, nyelvtechnológiai eszközök, veszélyeztetett nyelvek

1. Bevezetés

A modern technológia fejlődése, az internet és okostelefonok elterjedése lehetővé teszi azt, hogy az emberek a világ minden tájján valós időben kommunikáljanak egymással. Az emberek közti kommunikáció, illetve a gép–ember kommunikáció elősegítését szolgálják a nyelvtechnológiai eszközök és alkalmazások, mint például helyesírás-ellenőrzők, gépi fordítóoldalak vagy keresőprogramok, a digitális világban történő kommunikációt pedig különféle online erőforrások és alkalmazások segítik elő. Problémát jelent azonban az, hogy míg a világ nagy nyelveire jelenleg is számos nyelvtechnológiai eszköz létezik, addig a kisebbségi nyelvekre sokszor még a legalapvetőbb digitális nyelvi eszközök sem léteznek. A projekt elsődleges célja, hogy olyan nyelvtechnológiai eszközöket készítsünk finnugor kisebbségi

nyelvek beszélőinek számára, amelyek megkönnyítik számukra a digitális világban való anyanyelvi kommunikációt.

A kisebbségi nyelvek nemcsak beszélőik számában különböznek más nyelvektől, hanem legfőképpen abban, hogy esetükben leginkább olyan nyelvekről van szó, amelyek nem hivatalosak országukban (hanem egy nagy, hivatalos státusszal rendelkező nyelv mellett, annak árnyékában léteznek), és beszélőik is ezért legtöbbször olyan kétnyelvűek, akik a hivatalos/többségi nyelven (végzik vagy) végezték iskolai tanulmányaikat, hivatalos és írott funkciókban, a munkahelyen leginkább azt használják. Ily módon a kisebbségi nyelv a privát szférára (családon belüli, barátok közötti stb.) és azon belül is a szóbeli kommunikációra korlátozódik, írásban kevéssé használatos lesz.

Napjainkban a digitális (azaz számítógépes közegű) nyelvhasználat (pl. e-mail írás és olvasás, chatelés, fórumozás, kommentelés, blogírás és -olvasás) megnövelte a nyelvhasználó írott nyelvhasználatát. Kétnyelvű beszélők esetében ezért elsőrendűen fontos kérdés, hogy tudják-e kisebbségi nyelvüket digitálisan használni [1].

A felhasználói oldalról is hasznos nyelvtechnológiai alkalmazások létrehozásához, mint például a fentebb is említett helyesírás-ellenőrző vagy gépi fordító, elengedhetetlen, hogy rendelkezésre álljanak az alapszintű nyelvfeldolgozó technológiák az adott nyelvre. A kisebbségi nyelvek esetében a szövegfeldolgozás akár a karakterkódolás szintjén is problematikus lehet, amennyiben nem létezik egy egységesített (sztenderdizált), széles körben elterjedt karakterkészlet. A nyelvtechnológiai alkalmazások létrehozásához szükséges továbbá egy szegmentáló (mondatra, illetve azokat szavakra bontó) eszköz, morfológiai elemző és szófaji egyértelműsítő, a szövegek jelentésének megértésében pedig a szintaktikai és szemantikai mélyelemzők játszanak fontos szerepet. Ezen alaptechnológiák kifejlesztése egymásra épül: például a szegmentáló kimenetéhez, azaz az egyes szavakhoz rendel elemzést a morfológiai elemző, majd a szintaktikai elemző a szófaji kódokat is figyelembe véve elemzi a mondatokat stb.

A projekt elsődleges céljának eléréséhez, azaz a finnugor kisebbségi nyelvekre történő, felhasználói szintű nyelvtechnológiai eszközök létrehozásához így tehát szükség van az adott nyelvű szegmentálók és morfológiai elemző eszközök, továbbá szótári adatbázisok létrehozására.

A projektben elsődlegesen az udmurt és manysi nyelvekre összpontosítunk.A jelenlegi szakaszban az udmurt és manysi nyelvű korpuszok létrehozása zajlik, ezzel párhuzamosan az adott nyelvű digitális szótárak fejlesztése is folyamatban van. E szótárak szóanyagát a későbbiekben nyelvtani (morfológiai) információval is ellátjuk, így a szótárként való hasznosítás mellett a morfológiai elemzők alapjául is szolgálhatnak, melyek létrehozása szintén elkezdődött. A későbbiekben a szótári adatbázisra, korpuszainkra és a morfológiai elemzőkre építve különféle nyelvtechnológiai alkalmazásokat, például interneten szabadon elérhető szótárakat és nyelvoktató játékokat, illetve helyesírás-ellenőrzőt szeretnénk létrehozni, figyelembe véve a lehetséges jövőbeli felhasználók igényeit is.

2. A FinUgRevita projekt

A projekt célja, hogy veszélyeztetett oroszországi finnugor nyelvek beszélőit támogassuk számítógépes nyelvi eszközökkel, amelyek a felhasználók/beszélők kisebbségi nyelvi nyelvhasználatát segítik a digitális térben, valamint hogy szociolingvisztikai eszközökkel lemérjük ezen számítógépes nyelvi eszközök sikerességét. Kutatásunkkal annak a kérdésnek a praktikus megválaszolásához kívánunk hozzájárulni, hogy mivel lehet aktívan támogatni a veszélyeztetett finnugor kisebbségi nyelveket, megerősíteni a beszélő közösségeket és ilyen módon szolgálni a nyelvi revitalizációt.

A projekt nyelvtechnológiai komponenseként fel kívánjuk használni a veszélyeztetett kisebbségi finnugor nyelveken már létező nyelvi forrásokat (szótárakat és morfológiákat), hogy azokat felhasználva számítógépes eszközöket (tanulást és szövegalkotást segítő eszközöket) hozzunk létre. Ezen eszközök lehetővé teszik majd, hogy a beszélők modernizált populáris beszédmódokban használják anyanyelvüket. Úgy gondoljuk, hogy ezek az eszközök pozitív hatással lesznek a beszélők írott nyelvi tudására, anyanyelvükhöz kapcsolódó nyelvi attitűdjeikre, és végső soron a revitalizációs folyamatot segítik elő.

A projekt szociolingvisztikai komponenseként fel készülünk mérni a kifejlesztett és használatra bocsátott számítógépes eszközök sikerességét. A szociolingvisztikai méréseket az eszközök kifejlesztése és használatra bocsátása előtt és után is elvégezzük az eredmények összehasonlíthatósága érdekében.

3. Nyelvek

Az udmurt – régebbi elnevezés szerint: votják – az uráli nyelvcsaládba tartozó, kevéssé veszélyeztetett őshonos nyelv. Udmurtok nagyobb számban élnek Kazahsztánban, és szórványban Oroszország számos városában, kerületében. A legfrissebb, 2010-es népszámlálási adatok szerint az udmurtok lélekszáma 552 299 fő, az udmurt nyelvet saját bevallása szerint 324 338 fő beszéli. (Mindkét szám népszámlálásról népszámlálásra csökken.)

A manysi – régebbi elnevezés szerint: vogul – egy az uráli nyelvcsaládba tartozó, erősen veszélyeztetett őshonos nyelv. A manysi nyelvet elsősorban Nyugat-Szibériában, a Hanti-Manysi Autonóm Körzet területén beszélik. A legfrissebb, 2010-es népszámlálási adatok szerint a manysik lélekszáma 12 269 fő, a manysi nyelvet saját bevallása szerint 938 fő beszéli. (Előbbi szám népszámlálásról népszámlálásra nő, utóbbi szám folyamatosan csökken.)

A manysi és az udmurt nyelv elsősorban a családi, baráti érintkezések során használatos, nem hivatalos nyelv, nem rendelkezik gazdasági jelentőséggel, nem játszik szerepet a törvényhozásban és a politikában sem. Ugyanakkor jelen van a sajtó mellett a manysi nyelv a médiában, a kulturális életben, az interneten és az oktatásban is.

4. A FinUgRevita projekt számítógépes vonatkozásai

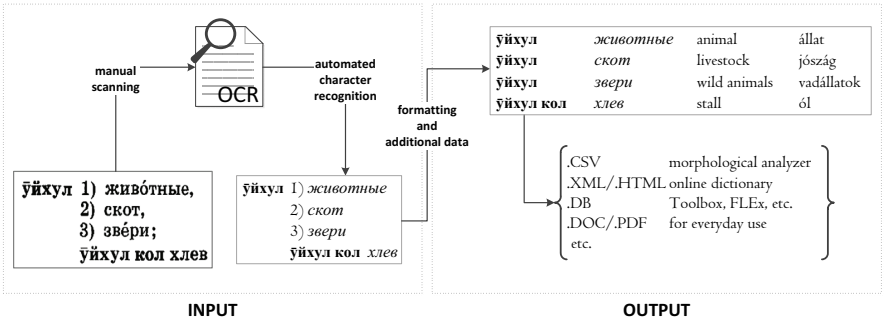
4.1. Online szótárak létrehozása

A projekt két nyelvére, az udmurtra és a manysira jelenleg folyamatban van online elektronikus szótárak készítése.

Az udmurt elektronikus szótárhoz a Kozmács István által létrehozott és szerkesztett Udmurt-magyar szótárból indulunk ki [2]. Egészen pontosan az elektronikus szótárt ez utóbbinak elektronikus verziójából (Microsoft Word dokumentum) alakítjuk át, félig automatikus módon: a dokumentum formázása alapján strukturált (CSV) formátumra konvertáljuk az anyagot, majd ezt kézzel ellenőrizve javítjuk. Az automatikus konverzió megtörtént, jelenleg a kézi javítás zajlik. A szótár kb. 13000 címszót tartalmaz.

A projekt online manysi szótára nagyobbrészt a már létező manysi–orosz és orosz–manysi szótárak [3,4,5,6] alapján készül. Az online manysi szótár tartalmazza Rombandeeva-Kuzakova 4000 szócikkes kétirányú [3] és Rombandeeva 11000 szócikkes orosz–manysi szárának [4] teljes anyagát, valamint Balandin-Vakhrusheva manysi–orosz szótárának [6] egyes válogatott szócikkeit. Ezen felül a lexikon anyagát törekszünk bővíteni a jelenkori nyelvhasználat neologizmusaival (mint például a városi környezet, életmód szavaival, az olajbányászat, jogi és közigazgatási terminusaival), melyek főként a manysi sajtóban, a *Luima Seripos*ban jelennek és honosodnak meg.

Az online manysi szótár megközelítőleg 10000 elemet tartalmaz majd. A manysi lexémákhoz angol, orosz és magyar fordítást, szófaji címkét, valamint szükség esetén ragozási paradigmát is rendelünk. A manysi lexémákat orosz fordításukkal a szótárak PDF-formátumából nyerjük ki OCR-karakterfelismeréssel. Az angol és orosz fordításokat nyelvészeink biztosítják. A 1. ábra bemutatja, hogyan készül a szótár: az OCR-t kézi javítás, kézi fordítás követi, aminek eredménye egy kereshető, digitális szótár [7].



1. ábra. A szótárépítés folyamata

4.2. Morfológiai elemzők fejlesztése

A projekt egyik legfontosabb feladata a morfológiai elemzők készítése. Első lépésben feltérképeztük a finnugor nyelvekre kifejlesztett morfológiai elemzőket és felmértük hozzáférhetőségüket.

A manysi esetében már létezik egy morfológiai elemző [8], melyet a Morpho-Logic Kft.[1] fejlesztett ki. Ennek az elemzőnek az alkalmazási és felhasználási köre viszont több szempontból eltért a FinUgRevita irányvonalaitól. Az első szempont a latin betűs átiratú folklórgyűjtések helyett napjaink cirill ortográfiájú szövegeinek elemzése (lásd Korpuszépítés 4.3 bekezdést). Másodsorban az elemző lexikona Munkácsi *Wogulisches Wörterbuch*-jára [5] épül, és Kálmán két könyvének, a *Chrestomathia Vogulica* [9] és *Wogulische Texte* [10] szövegeire lett optimalizálva. Munkácsi szótára 19. század végi szövegeken alapul, és a Kálmán köteteiben fellelhető szövegeket a 20. század első felében gyűjtötték, így az elemző lexikonából hiányzik a jelenkori 20. és 21. századi szókincs, mely az újabb szövegek elemzéséhez elengedhetetlen. Végezetül, a már létező morfológiai elemző nem szabad felhasználású. Mindezen szempontok figyelembe vételével döntöttünk egy teljesen új morfológiai elemző létrehozásáról, melyhez az Online szótárak létrehozása 4.1 bekezdésben említettek szolgálnak alapul. A szótárak feldolgozásának pillanatnyi állásában az épülő lexikon egyes szavai, szócikkei már rendelkeznek a jelentésen túl morfológiai kategorizációs jegyekkel, melyekkel a szócikkeket különböző ige- és névszóragozási paradigmákba sorolhatjuk. Az elemzés nyelvtani alapjául szolgál több manysi leíró nyelvtani munka [11,12], és anyanyelvi adatközlők segítségére is számítunk.

Az udmurt nyelv esetében a már létező udmurt elemző fejlesztőivel felvettük a kapcsolatot, velük együttműködve átalakítjuk és tovább bővítjük az általuk létrehozott elemző[2] mögött álló szótári adatbázist és nyelvtani szabályokat. Hosszabb távú terveink között szerepel az előző fejezetben is említett, készülőben lévő udmurt elektronikus szótár anyagának integrálása a már elindított online morfológiai elemzőbe.

4.3. Korpuszépítés

A projektbeli munkálatokat segítendő manysi és udmurt nyelvű korpuszokat hozunk létre. A korpuszba elsődlegesen újságcikkeket, szépirodalmi anyagokat építünk be, de más típusú szövegek felvételét is tervezzük. Jelenleg a nyers szövegek beszerzése zajlik, később ezek egységes formátumra hozatala, illetve annotációja fog megtörténni.

Az 1. táblázat összefoglalja az udmurt korpusz szavainak és karaktereinek számát diskurzustípusonként. Mint látható, a legjobban képviselt szövegtípus az újságcikkek, amelyeket az udmurt nyelvű újság, az *Udmurt Dunne*[3] biztosít számunkra, de ezen kívül található még itt gyermekek számára írt folyóiratok is,

[1] `http://www.morphologic.hu/urali/index.php?lang=hungarian&a_lang=chv`
[2] `http://giellatekno.uit.no/cgi/d-udm.eng.html`
[3] `http://udmdunne.ru/`

pl. *Kizili* és *Zechbur*. A cikkek témája elég változatos: található köztük interjú, sport- vagy kulturális hírek is.

Az internetről is töltöttünk le anyagokat, pl. a Wikipédia oldalairól és udmurt nyelvű blogokról. A legtöbb anyag már digitalizálva volt, ami megkönnyítette azok feldolgozását. A korpusz most körülbelül 70,000 tokent tartalmaz.

1. táblázat. Az udmurt korpusz diskurzustípusonkénti karakter- és szószáma

Diskurzustípus	Karakterszám	Szószám
Blogok	26615	3969
Wikipédia	32110	4293
Szépirodalom	142272	20899
Sajtó	216740	30664
Oktatás	49294	6897
Esszék	25388	3255

A manysi korpusz magját a manysi nyelven kiadott *Luima Seripos* szolgáltatja, amely 1989 óta jelenik meg. A *Luima Seripos* online archívuma[4] 46 példányt tartalmaz. Ezek a kiadások, valamint a korábbi megjelenések, együttesen 5200 cikket tartalmaznak, a korpusz mérete így több mint egymillió token.

Mivel ez a manysi újság az egyetlen stabil forrása a manysi szövegeknek, ezért ennek van legnagyobb hatása a nyelv használóira is. Abból adódóan, hogy a *Luima Seripos* képezi korpuszunk alapját, a preferált manysi ortográfiát is megválasztjuk. A 19. század végén és 20. század első évtizedeiben a manysi nyelv különböző latin betűs lejegyzései maradtak fenn nyelvészek gyűjtéseiből. A manysi írásbeliség első éveiben ugyancsak a latin betűs lejegyzést használták, bár nem központosított módon, így mindenki saját intuíciói alapján írt. A Szovjetunió nyelvi politikájának következtében 1937-től cirill alapú ortográfiára kellett váltania az itt honos kis népeknek, így a manysiul íróknak is. Ekkor ugyancsak nem alakult ki központosított változat, így a manysi helyesírás jószerével egyéni és kisebb alkotócsoportok konvencióin alapul mindmáig. Ilyen meghatározó csoport a *Luima Seripos* mindenkori szerkesztői köre és tudományos intézetek, iskolák. 1937 óta több manysi ortográfia is kialakult, kezdve a manysi morfofonológiájához nem igazodó csak orosz karaktereket alkalmazó helyesírástól, eljutva egészen napjaink ortográfiájáig, mely speciális karaktereket is szerepeltet a csak a manysiban fellelhető fonémák jelölésére (így a magánhangzóhosszúságot jelölő macron például a *а̄* „folyó" szóban, vagy *ӈ* a veláris nazális mássalhangzó [ŋ] fonéma jelölésére). A magánhangzók hosszúságának jelentésmegkülönböztető szerepe van (*ос* „felület" és *ӧс* „bárány"), ennek ellenére a jelölése csak a 20. század vége felé válik általánossá. A legutolsó ortográfiai változtatás a 2000-es évek folyamán vált konvencionálisan elfogadottá, melynek során a palatalizált zöngétlen szibiláns [s^j] jelölésére a korábbi *с* + {lágy magánhangzó: *е, ё, и, ю, я*; lágyjel: *ь*} helyett a legújabb kiadványokban a *щ* betű

[4] `http://www.khanty-yasang.ru/luima-seripos/archive`

szolgál a fonéma jelölésére. Ezen szabályokat követve a projektben is a legutolsó elfogadott ortográfiát használjuk alapértelmezettként, de a készülő morfológiai elemző a maitól eltérő transzliterációjú szövegek elemzésére is képes lesz.

5. Összegzés

Cikkünkben bemutattuk a FinUgRevita projektet, melynek célja nyelvtechnológiai eszközök létrehozása két oroszországi kisebbségi státuszú finnugor nyelvnek, az udmurtnak és manysinak. A projekt jelenlegi fázisában a két nyelv elektronikus szótárának felépítése és bővítése, valamint a világhálón fellelhető irodalmi szövegek, újságcikkek és a közösségi médiában létrejött blogok, bejegyzések nyelvi anyagából nagy tokenszámú korpuszok építése folyik. Az általunk szerkesztett szótárak lexémáin alapuló új morfológiai elemzőt fejlesztünk a korpuszok feldolgozására.

A jövőbeni terveinket tekintve elsőképpen szeretnénk a készülő szótárakat és morfológiai elemzőinket szabadon hozzáférhetővé tenni az udmurt és manysi nyelv beszélőinek, valamint a nyelveket tanulni és kutatni vágyóknak. További célunk a korpuszok morfológiai és lehetőség szerinti szintaktikai annotálása, mely alapul szolgálhat egy statisztikai szófaji egyértelműsítő és szintaktikai elemző létrehozásához.

Végül célunk olyan online nyelvi játékok tervezése és megalkotása, melyek segíthetik a nyelvtanulás folyamatát. Reményeink szerint a FinUgRevita projekt által elért eredmények szerepet játszanak majd az udmurt és manysi nyelv revitalizálásában, és a kifejlesztett nyelvtechnológiai eszközök segítséget nyújtanak, hogy az általunk támogatni szándékozott nyelvek meghonosodjanak a digitális térben is.

Köszönetnyilvánítás

A kutatás a Számítógépes eszközök a veszélyeztetett finnugor nyelvek nyelvi revitalizációjáért (FinUgRevita) nevű, FNN 107883 azonosítószámú projekt keretében valósult meg, az OTKA támogatásával.

Hivatkozások

1. Kornai, A.: Digital language death. PLoS ONE **8**(10) (2013) e77056
2. Kozmács, I.: Udmurt-magyar szótár. Savaria University Press (2002)
3. Rombandeeva, E.I., Kuzakova, E.A.: Slovar' mansijsko-russkij i russko-mansijskij. Prosvešenie, Leningrad (1982)
4. Rombandeeva, E.I.: Russko-mansijskij slovar'. Mirall, Sankt-Peterburg (2005)
5. Munkácsi, B., Kálmán, B.: Wogulisches Wörterbuch. Akadémiai Kiadó, Budapest (1986)
6. Balandin, A.N., Vahruševa, M.I.: Mansijsko-russkij slovar' s leksičeskimi paralellami iz južno-mansijskogo (kondinskogo) dialekta. Prosvešenie, Leningrad (1958)

7. Thieberger, N., Berez, A.L.: Linguistic data management. In Thieberger, N., ed.: The Oxford Handbook of Linguistic Fieldwork. Oxford University Press, Oxford (2012) 90–118
8. Prószéky, G.: Endangered uralic languages and language technologies. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria (2011) 1–2
9. Kálmán, B.: Chrestomathia Vogulica. Tankönyvkiadó, Budapest (1963)
10. Kálmán, B.: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest (1976)
11. Riese, T.: Vogul. Number 158 in Languages of the World/Materials. Lincom Europa, München - New Castle (2001)
12. Rombandeeva, E.I.: Mansijskij (vogul'skij) jazyk. Nauka, Moskva (1973)

Az automatikus irreguláriszönge-detekció sikeressége az irregularitás mintázatának függvényében magyar (spontán és olvasott) beszédben

Markó Alexandra[1], Csapó Tamás Gábor[2]

[1] Eötvös Loránd Tudományegyetem, Fonetikai Tanszék
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
marko.alexandra@btk.elte.hu, csapot@tmit.bme.hu

Kivonat: Az automatikus irreguláriszönge-detekció problémája előtérbe került az utóbbi évtizedekben. A jelen kutatásban a [10] algoritmust futtattuk le az irreguláris zönge előfordulásaira manuálisan felcímkézett, magyar nyelvű spontán és felolvasott beszédkorpuszokon, és azt vizsgáltuk, hogy 1. Mennyire pontos a pusztán akusztikai kulcsokon alapuló gépi detekció, és mennyire pontos az akusztikai és percepciós paramétereket egyaránt figyelembe vevő humán annotáció? 2. Milyen tényezők befolyásolják (rontják) az irreguláris zönge detekciójának sikerességét a gépi és a humán annotációkban? Eredményeink szerint az irregularitás általános célú annotációjára folyamatos szövegekben az automata algoritmus nagy hatásfokkal alkalmazható, mivel az előfordulások több mint 90%-át pontosan jelöli. A magánhangzók vizsgálatára létrehozott korpuszokban a gégezárhang figyelmen kívül hagyása miatt kevésbé pontos az automata detekció. Összességében az irregularitás alkalmazott definíciójától függ a gépi detektáló módszer hatásfoka.

1 Bevezetés

Az emberi beszédben reguláris (más néven modális) zöngeképzés esetén a hangszalagok kváziperiodikusan rezegnek. A gégében azonban hosszabb-rövidebb időtartamra instabilitás léphet fel, ami a hangszalagok irreguláris rezgését okozza. Ez eltér a modális zöngeképzéstől, és irreguláris fonációnak, glottalizációnak, érdes zöngének vagy recsegő beszédnek nevezik [5]. Az irreguláris fonáció a magyar nyelvben általában szakaszhatárokon (pl. mondat végén) [15] vagy magánhangzó–magánhangzó kapcsolatokban fordul elő [16]. Gyakran kíséri extrém alacsony alapfrekvencia és a glottális pulzusok gyors lecsökkenése [2]. Érzetileg recsegő, érdes jellegű beszédet jelent [3].

Az automatikus irreguláriszönge-detekció problémája előtérbe került az utóbbi évtizedekben, mivel mind a beszéddel kapcsolatos alapkutatásokban (fonetika), mind az alkalmazott kutatásokban (beszédtechnológia) szembesültek vele a kutatók, hogy az irreguláris zönge viszonylag gyakori jelenség. Egy 30 beszélős korpusz felolvasásaiban a szótagok 4,9–44,7%-a valósult meg részben vagy egészben irreguláris zöngével, a spontán beszédben az arányok 6,0 és 47,4% között szóródtak [18]. Az átlag 21,3%

volt az olvasott, 25,6% a spontán beszédben. A jelenség gyakoriságából adódóan a fonetikában leginkább az irreguláris zönge funkciói és más beszédparaméterekkel való összefüggésének kérdései állnak a kutatások középpontjában; míg a mesterséges beszéd-előállításban azért került az érdeklődés homlokterébe, mert a hangadatbázis minőségét több szempontból is befolyásolja.

Mivel a beszédtechnológia alapmódszereit idealizált beszédre dolgozták ki, az irreguláris zöngével képzett szakaszokon hibák léphetnek fel az automatikus F0-detekcióban, illetve spektrális elemzésben. Ugyanakkor a gépi szövegfelolvasó rendszerek minőségét javíthatja, ha a természetes beszédhez hasonlóan bizonyos pozíciókban (pl. szakaszhatárokon) modellezzük az irreguláris zöngét [8].

1.1 Automatikus irreguláriszönge-detekció

A zöngeminőség-osztályozók általában néhány, a beszédjelen mért akusztikai paraméter alapján hoznak döntést arról, hogy a zöngét reguláris vagy irreguláris zöngével képezték-e. Surana szupport vektor gép alapú osztályozást alkalmaz négy akusztikai jegyen [20]. Ishi és társai három másik jegy bevezetését javasolják, amelyek a beszédjel nagyon rövid szakaszában számolt teljesítményén alapulnak, és egyszerű küszöbértéket használnak a döntéshez [13]. Bőhm egyesíti az előző két osztályozót, és algoritmikus finomhangolással, valamint SVM alapú osztályozással javítja a pontosságot [5,6]. Kane és munkatársai 2013-ban publikálták szabadon elérhető automatikus irreguláriszönge-detektáló algoritmusukat [14], amely a beszéd lineáris predikció alapú maradékjelében méri a másodlagos csúcsok előfordulását és a kiugró, impulzusszerű csúcsokat. Az eljárást többféle irreguláris mintázaton, illetve több nyelven tesztelték, és bemutatták, hogy jobb eredmények érhetők el a korábbi irreguláriszönge-detektoroknál [10]. A fenti automatikus osztályozó eljárásokkal a reguláris és az irreguláris zöngével képzett beszéd közel tökéletesen elkülníthető egymástól.

1.2 A kutatás célja

A jelen kutatásban a [10] algoritmust futtattuk le az irreguláris zönge előfordulásaira manuálisan felcímkézett, magyar nyelvű spontán és felolvasott beszédkorpuszokon (interjúk, szövegfelolvasások, mondatolvasások, tipikus és diszfóniás beszélők). A kutatás célja annak megállapítása, hogy milyen mértékben illeszkedik egymáshoz a manuálisan és az automatikusan előállított címkesor. A kutatás kérdései: 1. Mennyire pontos a pusztán akusztikai kulcsokon alapuló gépi detekció, és mennyire pontos az akusztikai és percepciós paramétereket egyaránt figyelembe vevő humán annotáció? 2. Milyen tényezők befolyásolják (rontják) az irreguláris zönge detekciójának sikerességét a gépi és a humán annotációkban?

Hipotézisünk szerint a humán annotáció kevésbé pontos rövid időtartamú irreguláriszönge-előfordulások, valamint igen alacsony átlagos alaphangmagasság esetén, továbbá azokban az esetekben, amikor a zöngeminőség nemcsak az irregularitásban tér el a modálistól (pl. leheletes zönge). Az automatikus detektáló kapcsán

feltételezzük, hogy eredményességét befolyásolják a felvételi körülmények, a hangfelvétel minősége.

2 Módszerek

2.1 Hanganyag

A kutatásban többféle beszédkorpuszból válogattunk az irreguláris zöngére felcímkézett mintákat: felnőtt beszélők (egy 44 éves nő és egy 39 éves férfi) szövegfelolvasásait és interjúrészleteit a BEA adatbázisból [12]; diszfóniás (öt beszélő 21–38 év között) és kontroll (öt beszélő 20–24 év között) női beszélők mondatfelolvasásait [19]; valamint felnőtt beszélők (két – 25 és 31 éves – nő és két – 31 és 37 éves – férfi) mondatfelolvasásait a magánhangzó-kapcsolatok megvalósulásának elemzésére [17].

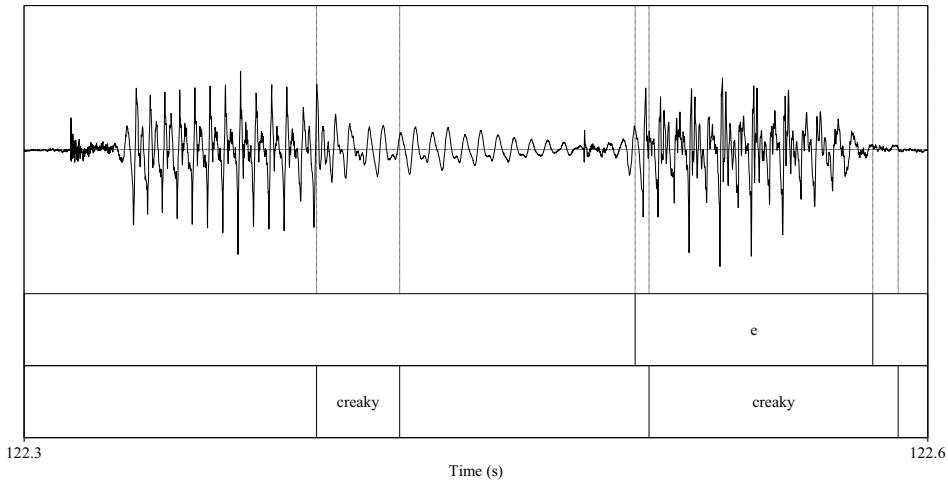
2.2 Irreguláriszönge-detekció

A beszédmintákon alkalmaztuk a CreakyDetection [10] irreguláriszönge-detektort, majd megvizsgáltuk, hogy a 10 ms-onként meghozott irreguláris/reguláris döntés milyen arányban felel meg a manuális irreguláriszönge-címkéknek.

2.3 Pontosság számítása

A célunk az volt, hogy egy-egy hangmintához meghatározzuk a manuális és automatikus címkézés pontosságát. Mivel azt feltételeztük, hogy mind a manuális, mind az automatikus címkézésben előfordulhat tévedés, ezért referenciának az összes irreguláris zönge címkével ellátott szakaszt vettük (azaz a manuális és automatikus címkék unióját). A manuális címkék határai tetszőlegesek lehetnek, míg az automatikus címkék 10 ms-os pontosságúak – emiatt az átfedő manuális/automatikus címkéket egyezőnek vettük. Az 1. ábrán egy példát mutatunk erre az esetre (az „e” és az alatta lévő „creaky” címke ugyanarra a magánhangzóra vonatkozik).

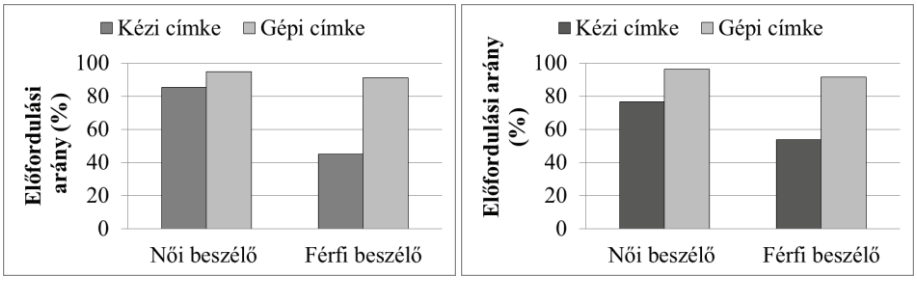
Hangmintánként kiszámítottuk a referenciához képest a manuális és automatikus címkék számát, majd ezt százalékos formába váltva kaptuk meg a pontosságot. Ez természetesen nem jelenti azt, hogy abszolút értelemben valóban a pontosságot határoztuk meg, hiszen az irregularitás meghatározása igen eltérő lehet (vö. pl. [1]).



1. ábra. Példa manuális (felső) és automatikus (alsó) irreguláris zönge címkére.

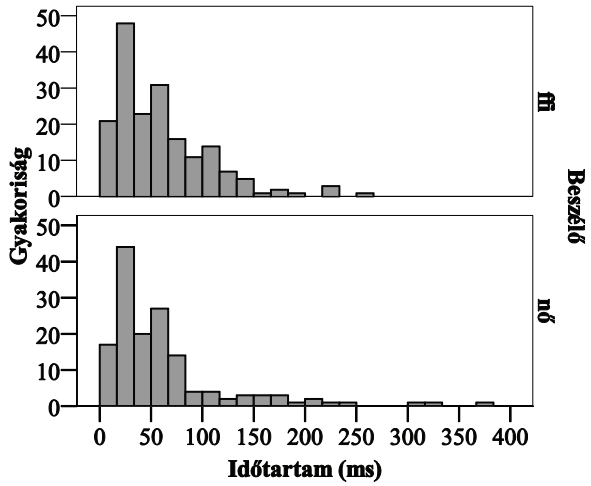
3 Eredmények

A BEA adatbázisból kiválasztott beszélők mindegyike valamilyen szempontból eltér az átlagostól, ezért beszédük címkézése az irreguláris zönge tekintetében nehezítettnek mondható. A női beszélő sokat glottalizálónak számít, felolvasásban a szótagjainak közel 30%-át, spontán beszédben 36%-át találtuk részben vagy egészben irregulárisnak. A férfi beszélő alaphangmagassága gyakran olyan mély tartományban valósul meg, hogy ez nehezíti a zöngeminőség manuális címkézését (az ő átlagos alaphangmagassága felolvasásban 100, spontán beszédben 88 Hz, de a 60–70 Hz közötti adatok sem ritkák a beszédében). A 2. ábrán látható, hogy ebben a két anyagban minden esetben az automatikus címkéző a pontosabb, kevés kivételtől eltekintve minden irreguláris szakaszt megtalált. A kézi címkézésnek a gépitől való eltérése több okra vezethető vissza (lásd alább), míg azoknak a manuálisan címkézett szakaszoknak a megjelenését, amelyeket az automata nem jelölt, általában az irregularitás értelmezésének különbségével magyarázhatjuk. Jelentős eltérés van a női és a férfi beszélő hanganyagának manuális címkézési eredményei között. Az előbbi, bár sokat glottalizál, az irreguláris szakaszok egyértelműen elkülöníthetők a beszédében, míg az utóbbi esetében gyakran nehéz megkülönböztetni az irreguláris zöngét az extrém alacsony alaphangmagasságtól.



2. ábra. A címkézés pontossága a BEA adatbázis két beszélője esetében: balra az olvasott, jobbra a spontán hanganyag eredményei.

Érdemes megvizsgálni azokat az eseteket, amelyeket a gépi annotáció jelölt, a kézi azonban nem. A manuális jelölésekre egyértelműen hatással van az irregularitás időtartama. Mintegy 50 ms-ban határozható meg a humán percepció küszöbértéke: a kézzel nem jelölt irreguláris szakaszok nagy része ennél rövidebb időtartamú (vö. 3. ábra). Ez természetesen nem jelenti azt, hogy a rövidebb irreguláris szakaszokat biztosan nem észleljük, csak azt, hogy ha valamilyen pozicionális marker nem teszi prominenssé, hajlamosak lehetünk elsiklani felettük.



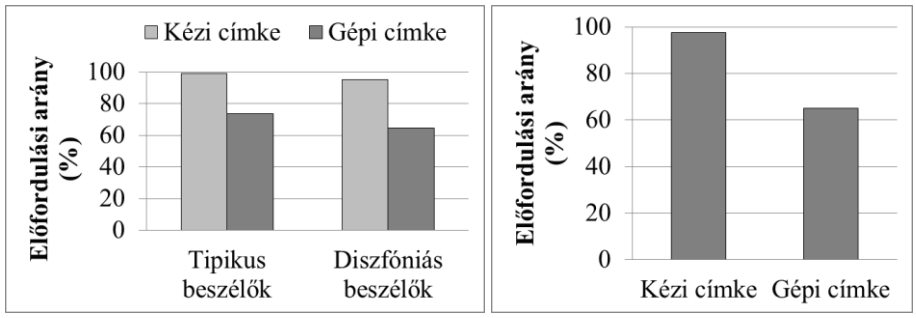
3. ábra. A kézi annotációból kimaradt gépi címkék gyakorisági eloszlása az irregularitás időtartama függvényében

Az eltérés tipikus oka még az, hogy az automatikus annotálás a zöngés obstruensek zárszakaszában, illetve a nem teljes zárfelpattanás esetén is irregularitást jelöl, akárcsak a pergőhang realizációiban. Ezek egy része ugyancsak rövid időtartamú címke. Egyes beszélőkre jellemző, hogy fojtott zöngét produkálnak, miközben nem artikulálnak – e felett a humán percepció elsiklik, az automata azonban ezekben is jelöli az

irregularitást. Találtunk néhány olyan esetet is, amikor az irregularitás látható ugyan a hullámformán, de a hallási percepció alapján nem észlelünk eltérést. Mivel [9] és [3] alapján akkor címkéztünk glottalizáltnak egy beszédrészletet, ha az akusztikai lenyomaton szemmel és auditív úton füllel egyaránt észlelhető volt az irregularitás, ezeket az eseteket a gépi annotációval való összevetés alapján sem tartanánk glottalizáltnak. Végül, különösen a hosszabb időtartamú címkék esetén, természetesen felmerül a humán címkéző figyelmetlensége, fáradása is.

Vizsgáltuk azon manuális címkék sajátosságait is, amelyekhez nem volt megfelelő gépi címke. Előfordult néhány esetben, hogy az utólagos ellenőrzés alapján ezeket nem tartanánk glottalizáltnak.

Egészen más eredményeket kaptunk a másik két korpuszból származó hanganyagok elemzésekor (4. ábra). Ezeknél nem minden egyes irreguláris szakasz címkézése volt a cél, hanem a magánhangzók, illetve magánhangzó-kapcsolatok irregularitásának címkézése. Tekintettel arra, hogy a felvételeken előre be volt jelölve ezeknek a szakaszoknak a pozíciója, a címkéző feladata az volt, hogy eldöntse, az adott magánhangzó(k)ban van-e irregularitás. Ennek alapján azt vártuk, hogy a manuális címkék aránya jobban meg fogja közelíteni a gépi címkékét, azonban ebben a két korpuszban a kézi címkék még felül is múlják számosságban az automatikus címkéket. Tekintettel arra, hogy a vizsgált esetek nem kis hányadában magánhangzós szókezdetek, illetve magánhangzó-kapcsolatok realizációi alkották az elemzés anyagát, a gégezárhang jelensége igen nagy arányban jelent meg az anyagban, ezeket azonban az automatikus címkézés egyetlen egy esetben sem jelölte, és az eljárást bemutató szakirodalmi forrás [10,14] sem utal arra, hogy ezeket az eseteket hogyan kezeli. Dilley és társai ugyanakkor egyértelműen az irregularitás altípusának veszik a gégezárhangot [9], sőt az újabb források szerint a gégezárhang és a glottalizáció valójában egyazon fiziológiai mechanizmussal jön létre [11]. A szerzők szerint gégezárhangot észlelünk, ha a csak egy periódusnyi tartamú, és glottalizációt hallunk, ha sorozatban több egymást követő periódusra kiterjed az ezeket létrehozó laringális konfiguráció.



4. ábra. A címkézés pontossága balra a patológiásbeszéd-korpuszban (magánhangzók jelölésében), jobbra a magánhangzó-kapcsolatok vizsgálatára létrehozott korpuszban.

Ugyanakkor ezekben a korpuszokban is megfigyelhető volt a humán percepció „túlműködése”, ugyanis néhány olyan esetben, amikor a hangzásbeli eltérés oka nem

az akusztikai jelben mérhető nagymértékű ingadozás, hanem például jellegzetes hangszínezet vagy leheletes zöngeképzés volt, a címkéző ezeket is jelölte.

4 Összefoglalás, következtetések

A vizsgálatunk arra irányult, hogy megállapítsuk, különböző magyar nyelvű korpuszokon milyen hatásfokkal alkalmazható a [10,14] által kifejlesztett automatikus irreguláriszönge-detektáló. Természetesen mindig az adott alkalmazástól és az irregularitás definíciójától függ, hogy a gépi elemző mennyire hatékony. Eredményeink alapján az irregularitás általános célú annotációjára folyamatos szövegekben az automata algoritmus nagy hatásfokkal alkalmazható, mivel az előfordulások több mint 90%-át pontosan jelöli. A humán percepció számára nehézséget okozó hangminőség (pl. alacsony alaphangmagasság) esetén is hatékony, és alkalmas az emberi tényező (figyelmetlenség, fáradás) ellensúlyozására. A gépi annotáció manuális ellenőrzésére ugyanakkor feltétlenül szükség van, mivel bizonyos beszédhangtípusok (pl. pergőhang, zöngés obstruensek) esetében szükségtelenül is jelöl, ugyanakkor a gégezárhangot – az algoritmus sajátosságaiból adódóan – nem ismeri fel. A tapasztalatok alapján az általános irreguláriszönge-címkézéshez érdemes 40-50 ms-os küszöbértéket beállítani. Így egyrészt a detektálónak az említett beszédhangtípusokra való érzékenysége is csökken, másrészt a címkék nagyobb mértékben korrelálnak a humán percepció által irregulárisnak minősített szakaszokkal.

A magánhangzók vizsgálatára létrehozott korpuszokban éppen a gégezárhang figyelmen kívül hagyása miatt kevésbé pontos az automata detektor, ugyanakkor ellenőrzésképpen való használata elkerülhetővé teszi, hogy a címkézést a humán percepciót zavaró tényezők (pl. leheletes zönge, feszített zöngeképzés) negatívan befolyásolják.

Köszönetnyilvánítás

A kutatást részben az SP2 Scopes Project on Speech Prosody támogatta.

Hivatkozások

1. Batliner, A., Burger, S., Johne, B., Kiessling, A.: MÜSLI: A classification scheme for laryngealizations. In: Working Papers, Prosody Workshop, Lund, Schweden (1993) 176–179
2. Blomgren, M., Chen, Y., Ng, M. L., Gilbert, H. R.: Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. Journal of the Acoustical Society of America 103 (1998) 2649–2658
3. Bőhm, T., Ujváry, I.: Az irreguláris fonáció mint egyéni hangjellemző a magyar beszédben. Beszédkutatás 2008 (2008) 108–120

4. Bőhm, T., Audibert, N., Shattuck-Hufnagel, S., Németh, G., Aubergé, V.: Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. In: Acoustics'08 (2008) 6141–6146
5. Bőhm, T.: Analysis and modeling of speech produced with irregular phonation. PhD disszertáció, BME TMIT (2009)
6. Bőhm, T., Both, Z., Németh, G.: Automatic Classification of Regular vs. Irregular Phonation Types. In: NOLISP, (2009) 43–50
7. Collins, B., Mees, I. M.: Practical phonetics and phonology: A resource book for students. Routledge, New York (2008)
8. Csapó, T. G., Németh, G.: Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. IEEE Journal on Selected Topics in Signal Processing 8 (2014) 209-220
9. Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics 24 (1996) 423–444
10. Drugman, T., Kane, J., Gobl, C.: Data-driven Detection and Analysis of the Patterns of Creaky Voice. Computer Speech and Language 28 (2014) 1233–1253
11. Esling, J. H., Harris, J. G.: States of the glottis: an articulatory phonetic model based on laryngoscopic observations. In: Hardcastle, W. J. – Mackenzie Beck, J., eds.: A figure of speech: A festschrift for John Laver. Lawrence Erlbaum Association, Mahwah (2005) 347–383
12. Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T. E., Beke, A., Neuberger, T., Nikléczy, P.: BEA: beszélt nyelvi adatbázis. In: Gósy, M., ed.: Beszéd, adatbázis, kutatások. Budapest, Akadémiai Kiadó (2012) 9–24
13. Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., Hagita, N.: A Method for Automatic Detection of Vocal Fry. IEEE Transactions on Audio, Speech and Language Processing 16 (2008) 47–56
14. Kane, J., Drugman, T., Gobl, C.: Improved automatic detection of creak. Computer Speech and Language 27 (2013) 1028–1047
15. Markó, A.: A glottalizáció határjelző szerepe a felolvasásban. Beszédkutatás 2011 (2011) 31–45
16. Markó, A.: Az irreguláris zönge szerepe a magánhangzók határának jelölésében V(#)V kapcsolatokban. Beszédkutatás 2012 (2012) 5–29
17. Markó, A.: Boundary marking in Hungarian V(#)V clusters with special regard to the role of irregular phonation. The Phonetician 103 (2012) 7–26
18. Markó, A.: Az irreguláris zönge funkciói a magyar beszédben. Budapest: ELTE Eötvös Kiadó, (2013)
19. Markó, A.: Glottalizáció és diszfónia. Gyógypedagógiai Szemle XLII/1 (2014) 23–36
20. Surana, K.: Classification of vocal fold vibration as regular or irregular in normal voiced speech. MS thesis, MIT, USA, (2006)

Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű szövegelemzőben

Miháltz Márton[1], Indig Balázs[2], Prószéky Gábor[1,2,3]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport, 1083 Budapest, Práter utca 50/a
[2] PPKE Információs Technológiai és Bionikai Kar, 1083 Budapest, Práter utca 50/a
[3] MorphoLogic, 1122 Budapest, Ráth György u. 36.
mmihaltz@gmail.com, indig.balazs@itk.ppke.hu,
proszeky@morphologic.hu

1 Bevezetés

Az MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport célkitűzései közé tartozik egy olyan, pszicholingvisztikailag motivált nyelvfeldolgozó rendszer kifejlesztése, amely nyers magyar nyelvű szövegből képes szintaktikai és szemantikai reprezentációt felépíteni [6]. Ennek egyik kulcsfontosságú lépése a természetes nyelvi mondatokban található ige-vonzat viszonyok felismerése és osztályozása (nyelvtani szerepek, tematikus szerepek). Ebben a cikkben bemutatjuk jelenleg zajló munkánkat, melynek célja az MTA-PPKE szövegelemző konstrukciós tudástárának bővítése. A munkához igyekszünk felhasználni és újrahasznosítani több, korábban kifejlesztett nyelvi erőforrás anyagát is.

Az elemző egyik alapvető felépítési elve a párhuzamosság: több különböző „erőforrásszál" egymást felülbírálva-javítva dolgozik. Ezek az erőforrásszálak a kategoriális nyelvtanokéra [3] emlékeztető mechanizmussal működnek: a nyelvi egységek közötti viszonyokat „ajánlások" és „elvárások" párhuzamos, ún. „strukturális szálai" közötti megfeleltetések teremtik meg. Ebben a paradigmában a mondatbeli potenciális vonzatok (névszói csoportok) „felajánlásokat" fogalmaznak meg (lexikai, morfológiai és szemantikai tulajdonságokat), melyek a mondatbeli igék által előhívott vonzatkeretek vonzatpozíciónak „elvárásaival" (megkötések az előbbi típusú tulajdonságokra) tudnak összekapcsolódni [8].

A továbbiakban először röviden bemutatjuk nyelvi elemzőnk igei vonzatkereteket kezelő mechanizmusát, majd felvázoljuk annak lehetőségét, hogyan lehet az elemzőben használt vonzatkeret-adatbázist tematikusszerep-leírásokkal kibővíteni az angol VerbNet erőforrás felhasználása segítségével.

2 Igei szerkezetek elemzése

Az igei vonzatszerkezetek elemzésének támogatására felhasználjuk a *MetaMorpho* magyar–angol (és angol–magyar) fordítóprogram [7] névszói és igei adatbázisának környezetfüggetlen, jegystruktúrás újraíró szabályait. A szabályok egy részére jel-

lemző, hogy bennük *minden* jobboldali szimbólum tartalmaz lexikális megkötést. Ezek a szabályok írják le a lexikont, vagyis a főnévi, melléknévi és határozószói egy- vagy többszavas kifejezéseket és ezek szemantikai és egyéb tulajdonságait [5,10], pl. „házőrző kutya”: <főnév, megszámlálható, élőlény>, „tegnapelőtt”: <időhatározó>, „hindi”: <főnév/melléknév, nyelv>.

A számunkra érdekes MetaMorpho szabályok másik csoportja nem minden (de legalább egy) jobboldali szimbólumra tartalmaz lexikális megkötést, míg a többi összetevőkre csak szófaji, morfológiai, szemantikai stb. feltételeket. Ezek közé tartoznak az igei vonzatkereteket leíró szabályok, melyeknél legalább az igei pozíció kötött lexikálisan, a vonzatpozíciók közül legfeljebb csak az idiomatikus igemódosítók (pl. „<valaki> jövendöl <valamit> <valakinek>” vs. „*szó* esik <valamiről>”).

Elemzőnk működése során szigorúan balról jobbra halad. Minden újabb mondatbeli token megjelenése a már ismert tokenek kapcsolati rendszerének (az elemzést reprezentáló gráf) időszerűsítéséhez (kiegészítéshez, és amennyiben szükséges, módosításához) vezet. A HuMor elemző felhasználásával minden tokent ellátunk morfológiai annotációval, a lehetséges elemzési szekvenciák közül egy, a PurePoS tagger [4] elvén működő, csak baloldali kontextust figyelő komponensünk választ, amely csak lokális, pontozott elemzési lehetőségeket produkál, Viterbi beam search nélkül. Az N legjobb elemzésen ezután párhuzamosan futtatjuk nyelvi elemzőnk központi elemző komponensét [8].

Az ige–vonzat viszonyok azonosítása a mondat feldolgozása során a keresletkínálat elv szerint történik. Névszói tokenek megjelenésekor ezekre szófaji és morfológiai tulajdonságaik segítségével igyekszünk új lexikális szabályokat illeszteni, vagy ezekkel korábban megkezdett többszavas mintákat folytatni. Ha egy egy- vagy többszavas névszói kifejezést teljesen felismertünk, megkíséreljük vele a mondatban korábban már szerepelt igék betöltetlen vonzatpozícióit „kielégíteni”. Igei tokenek megjelenése az igetőhöz tartozó lehetséges vonzatkeretek betöltésével jár, melyek vonzatpozícióit az elemző ekkor megkísérli „kielégíteni” az addig a pozícióig a mondatban már szerepelt (és teljesen befejezett) névszói elemek „felajánlásaival”.

3 Tematikus szerepek azonosítása

Az ige-vonzat viszonyok azonosításán túl a vonzatkeretek alkalmasak azok jellemzésére is. A szemantikai reprezentáció kialakítása szempontjából hasznos **tematikusszerep-leírások** azonban csupán a MetaMorpho magyar vonzatkeret-leírások mintegy 10%-ához állnak rendelkezésre (ezeket egy, a MetaMorpho fejlesztéséhez kapcsolódó független projekt, történelmi szövegek narratív pszichológiai elemzésének támogatásához készítettek [9]. Ugyanakkor a vonzatkeret-leírások fontos tulajdonsága, hogy angolul és magyarul is tartalmazzák e szabályok megfelelőit: minden forrásnyelvi (magyar) elemző szabályhoz rendelkezésre áll egy célnyelvi (angol) generáló szabály is, amely az adott nyelvi konstrukció fordítása. Ezért lehetőség van arra, hogy szabadon hozzáférhető, angol nyelvű erőforrásokra támaszkodva először a szabályok angol, majd a későbbiekben a szabályok magyar oldalát is kibővítsük az angol erőforrás MetaMorpho szabálypárokhoz kapcsolásával és az angol–magyar megfeleltetések helyes kezelésével (linked resource).

Ilyen erőforrás például a SemLink projekt [2] termékeként előállt egységes angol lexikai adatbázis, ami a *VerbNet* igei szótár , a PropBank szintaktikai és szemantikai jegyekkel annotált korpusz és a FrameNet szemantikaikeret-adatbázis összekapcsolása, amelyben az angol igék vonzatkeretei, szintaktikai és szemantikai információi is elérhetők megfelelő minőségben. Célunk a MetaMorpho szabályminták ezen egységes erőforráshoz kapcsolása és a MetaMorpho angol nyelvű, igei vonzatkereteket leíró szabályainak tematikus szerepekkel való minél teljesebb automatikus annotálása. Ezt követően megvalósíthatjuk az így nyert információk hatékony átvitelét a magyar nyelvű vonzatkeret-elemekre, kibővítve a magyar nyelvi elemzéshez rendelkezésre álló erőforrást. Fontos, hogy ez olyan módon történjen, hogy minél inkább megkönnyítse a magyar nyelvű szabályok emberi javítását, hogy a tisztán emberi módszerekkel létrehozott erőforrás minősége ne romoljon.

A következő részben bemutatjuk a MetaMorpho és a VerbNet összekapcsolásának és a tematikus szerepek gépi úton történő átvitelének kezdeti gyakorlati problémáit.

4 MetaMorpho és VerbNet vonzatkeretek megfeleltetése

Az összekapcsolás megvalósítása során figyelembe kellett venni, hogy az erőforrások számtalan ponton különböznek és ezeket az eltéréseket egységesíteni kell. Mindkét erőforrás tartalmazhat hibákat, amik befolyásolhatják a párhuzamosságok megtalálásának sikerességét. A MetaMorpho rendszer szabályainak készítői csak lazán voltak szabványokhoz és konvenciókhoz kötve, részben saját elképzeléseik alapján is dolgoztak, írásos dokumentáció a fejlesztési elvekről nem maradt fenn. Vizsgálataink során kiderült, hogy jópár elütésből és emberi hibából származó elem is található a szabályok között. Az elírási hibák egy része természetesen helyesírás-ellenőrzővel javítható automatikusan, de az előzetes tesztek azt mutatták, hogy sokszor ritka szavakról van szó, amit a helyesírás ellenőrző sem ismer.

Egy másik lényeges probléma, ami megnehezíti a két erőforrás harmonizálását az amerikai és a brit nyelvváltozatok írásmódjának kérdése: míg a MetaMorpho-t az eredeti szándékok szerint a brit angol ortográfiájának megfelelően fejlesztették, ezzel szemben a VerbNet az amerikai angol írásmódját követi.

A VerbNet összesen 6343 igét tartalmaz, ebből 2057 ige csak felsorolásszinten van jelen, mivel a többi, VerbNettel összekapcsolt erőforrásban előfordul. Az ilyen igéknek nem volt vonzatkeret-információja, ezért nem tudtuk őket a kutatás jelen állapotában hasznosítani. A maradék 4286 ige közül, melyekhez van vonzatkeret-információ, 2957 ige csak egy vonzatosztályban szerepel.

További problémát okozott az összetett igék kezelése. Az angol WordNetben összesen 7440 igéből 1410 *phrasal verb*, ami 549 igetövet érint. A VerbNetben 404 darab többszavas kifejezés található, melyekben 223 ige szerepel. A *MetaMorpho* 30 292 igei vonzatkeretes szabályába 3505 egyedi angol igető tartozik, amiből 920 darab nincs benne a VerbNetben (ebből 143 a helyesírás-ellenőrző számára hibás, illetve feltehetőleg ismeretlen szó). A vonzatkeretek és az angol igetövek száma közötti közel 10-szeres MetaMorpho-beli különbség egyik oka az, hogy kicsit több mint a szabályok egyharmadában idiomatikus vagy más lexikális megszorítás található (10 694 angol, 8347 magyar vonzatkeretben). Másfelől a magyar–angol

fordítórendszer fejlesztésekor nem volt cél, hogy a célnyelvi (angol) igék fedése jó legyen, elég volt a célnyelvi nyelvi hűség, jó fedésre a forrásnyelven (a magyar oldalon) volt inkább szükség. Szem előtt kell tartanunk, hogy ez a tulajdonság később még okozhat problémákat.

Az eddigieket figyelembe véve 2600 olyan egyedi ige van, ami mindkét erőforrásban szerepel és a VerbNetben osztályozva van. Ezekből 1545 csak egy, 622 kettő és 246 három VerbNet osztályba is beletartozik, továbbá van 10 darab olyan ige, ami 7 osztályba is be van sorolva. Ebből látszik, hogy az igék mintegy 42%-a esetén még az osztályok egyértelműsítésére is szükség van az ige minden egyes MetaMorpho-beli előfordulásánál.

A VerbNetben minden igeosztály tartalmaz egy vagy több, a VerbNet formalizmusában megadott szintaktikai vonzatkeret-leírást. Fő célunk ezeknek a VerbNetes vonzatkereteknek az egyértelmű azonosítása a MetaMorpho szabályok szintjén és az argumentumok kölcsönös leképezése a két erőforrás között.

Az egyértelműsítés során segítségünkre voltak a két erőforrás által definiált jeghalmazok, amik leírják az egyes argumentumok megszorításait a szintaxis és a szemantika oldaláról egyaránt. Míg a VerbNet a COMLEX formalizmust [1] alkalmazza, addig a MetaMorpho egy saját szempontrendszert [5,7,10]. Ezeket a leíró rendszereket kellett feltérképeznünk, illetve egymásra leképeznünk a különféle előfordulásaikat.

A MetaMorpho angol szabályok jobboldalának formája a legegyszerűbb esetben a SUBJ TV[1] [OBJ] mintát (egyszerű intranzitív/tranzitív szerkezet) követi. Az ilyen típusú szabályok az összes szabály kétharmadát teszik ki (kb. 20 000 szabály). Ezek az esetek képezték vizsgálódásunk első lépését. Az ilyen típusú szabályok esetén csak a sorrend és a típus figyelembevételével sikerült 1658 egyértelmű és 2908 többértelmű párosítást találni. A többértelmű párosítások feloldhatónak látszanak, amennyiben figyelembe vesszük az argumentumokra előírt megszorításokat mindkét oldalon. A vizsgált alakú szabályok és a közülük megtalált párosítások közötti több mint 4-szeres különbség a prepozíciók eltérő kezelésének tudható be, ugyanis a *MetaMorpho* a prepozíciókat egy egységként kezeli az őket követő szerkezetekkel, míg a VerbNetben a prepozíció önálló egységet alkot. Az ilyen különbségek helyes kezelése még folyamatban van.

Vannak olyan esetek is, amikor maga a vonzatkeret nem ad egyértelmű leképezést, annak ellenére, hogy a tematikus szerepek egyértelműek, mivel a VerbNet-ben a szemantikai információk is fel vannak tüntetve (és a szintaktikai szerkezettel együtt adják a rendezés kulcsát) és bár a szintaxis szintjén megegyeznek, a szemantika szintjén többértelműség keletkezik, amit fontosnak tartottak jelölni. Ilyen például a *meet* ige: a VerbNet különbséget tesz szemantikai szinten, amikor két ember találkozik, illetve amikor egy csoport tagjai találkoznak, míg szintaktikai szinten két azonos vonzat keret áll rendelkezésre. Ezen két eset között géppel a rendelkezésre álló információk segítségével nem lehet dönteni.

[1] TV (transitive verb): az igének megfelelő szimbólum a vonzatkeretben.

5 Összefoglalás

Jelen cikkben bemutattuk kereslet-kínálat elvű nyelvi elemzőnk igei vonzatkereteket azonosító működését, valamint megvizsgáltuk a *MetaMorpho* és a *VerbNet* összekapcsolásának és a tematikus szerepek átvitelének lehetőségét gépi úton. Fontos, hogy ez olyan módon történjen, hogy minél inkább megkönnyítse a magyar nyelvű szabályok emberi javítását, hogy a tisztán emberi módszerekkel létrehozott erőforrás minősége ne romoljon. A lexikai információk és nyelvi tulajdonságok további kétirányú megosztását tervezzük az összekapcsolt erőforrások között, amennyiben valamelyikben változás (verzióváltás, formátumváltozás stb.) állna be. Az ehhez szükséges lépéseket a tervezés során figyelembe vesszük. A jövőben megvizsgáljuk továbbá a lehetőségét annak, hogy a tematikus szerepekkel bővített igei konstrukciók segítségével hogyan lehet pszicholingvisztikailag reális módon szemantikai struktúrát kapcsolni a szintaktikai elemzéshez.

Hivatkozások

1. Grishman, R., Macleod, C., Meyers, A.: COMLEX Syntax: Building a Computational Lexicon. In: Proceedings of Coling, Kyoto (1994)
2. Loper, E., Yi, Sz.-t., Palmer, M.: Combining lexical resources: Mapping between PropBank and VerbNet. In: Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg (2007)
3. Morrill, G.V.: Categorial Grammar: Logical Syntax, Semantics, and Processing. Oxford University Press (2010)
4. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (2013)
5. Orosz, K.: Főnevek szemantikai jegyei és kódolásuk a MetaMorpho projektben. In: Alexin, Z., Csendes, D., eds.: IV. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged, (2006) 157–166
6. Prószéky, G., Indig, B., Mihältz, M., Sass, B.: Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In: Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged, (2014) 79–90
7. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta: Foundation for International Studies (2004) 138–142
8. Sass, B.: Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (2015), ld. jelen kötetben.
9. Vincze, O., Gábor, K., Ehmann, B., László, J.: Technológiai fejlesztések a NooJ pszichológiai alkalmazásában. In: VI. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged (2009) 285–294
10. Vincze, V., Lucza, M., Csendes, D., Kiss, G.: Szótárazási dilemmák a MetaMorpho magyar–angol névszói adatbázisának építésében. In: Alexin, Z., Csendes, D., eds.: IV. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged (2006) 180–189

28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet

Sass Bálint

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

Kivonat Két nagy méretű, magyar nyelvi erőforrást teszünk közzé. Az egyik a régi MNSZ [1] tagmondatainak sekély szintaktikai elemzéssel ellátott változata, mely a Mazsola [2] lekérdező adatbázisaként szolgál; a másik pedig az ebből az adatbázisból automatikusan származtatott igeiszerkezet-lista, melyből a Magyar Igei Szerkezetek című szótár [3] is született. Az erőforrások elérhetők a `http://corpus.nytud.hu/isz` címen.

Kulcsszavak: nyelvi erőforrás, szintaktikai elemzés, igei szerkezetek, Mazsola, méret

1. Az erőforrások létrehozása

A *Mazsola adatbázis* a Magyar Nemzeti Szövegtár 187 millió szavas régi változatának teljes anyagát tartalmazza, melyet a feldolgozás során tagmondatokra bontottunk és részleges szintaktikai elemzésnek vetettünk alá. Utóbbi során (1) megállapítottuk a tagmondat igéjét (főnévi igeneves szerkezet esetén a főnévi igenév a tagmondat igéje), és hozzákapcsoltuk az igéhez az esetleges odatartozó elváló igekötőt; (2) számba vettük az ige mellett felső szinten megjelenő névszói és névutói csoport bővítményeket (tehát a határozószói csoportokat például nem), ezeket a fej szótövével és esetével (értsd: esetragjával vagy névutójával) reprezentáltuk. A részleteket lásd [4] 2.2. fejezetében. A Mazsola [2] felületén keresztül lekérdezhető adatbázishoz képest a jelen adatbázis tartalmaz bizonyos javításokat, továbbfejlesztéseket: (1) a birtokos szerkezetek jobb kezelésének, valamint a főnévi igenév mellett *-nAk* raggal megjelenő alany funkciójú bővítmény alanyként való reprezentálásának köszönhetően csökkent a helytelen *-nAk* esetű bővítmények száma; (2) a *maga mögött* és a *mögöttem* típusú szerkezetek helyesen névutós névmásként elemződnek; valamint (3) szerepel egy további információ is az annotációban: hogy az adott bővítmény birtokos személyjeles-e.

Az *igeiszerkezet-lista* a fenti adatbázis alapján egy speciális igeiszerkezet-kinyerő algoritmussal automatikusan meghatározott igei szerkezeteket tartalmazza. Az algoritmus lényege, hogy a fent leírt reprezentáció szerinti mondatvázakat, igei kereteket azáltal összesíti, hogy a ritka (legfeljebb 5-ször előforduló) mondatvázakat egy rövidebb, illeszkedő mondatvázhoz rendeli hozzá; majd az eljárás végén lévő ellenőrző lépésben a túl általános mondatvázhoz került mondatokat a lehető legspecifikusabb meglévő mondatvázhoz helyezi át. A módszer

képes feltárni, hogy az adott esetragos bővítmény általában jellegzetes-e, illetve ezen túl azt is, hogy a bővítményi helyen megjelenő egyes konkrét tartalmas szavak tipikusak-e. Ennek megfelelően vonzatokat (*hisz vmiben*), kollokatív igei szerkezeteket (*süt (a) nap*, *döntés születik*), illetve a két eset kombinációjaként vonzatos komplex igéket (*szó van vmiről*, *igényt tart vmire*) egyaránt eredményez. Az igeiszerkezet-kinyerő módszer részletes bemutatása és kiértékelése [4] 3.3. fejezetében olvasható.

2. Az erőforrások formai leírása

A Mazsola adatbázis egy egyszerű szöveges fájl, sorainak formátumát az 1. ábra mutatja be.

```
engem meg sem hallgattak . stem@@meghallgat ACC@@én
A hasmenéstől szenvedő betegeknek sokat kell inniuk , stem@@iszik ACC@@sok NOM@@beteg
A Profi egyik támadójátékosa elhúzta mellettem a labdát , stem@@elhúz ACC@@labda mellett@@én ...
                                                    ... NOM@@támadójátékosPOSS
```

1. ábra. A Mazsola adatbázis sorainak felépítése

A tagmondat után következik a fentiek szerint elvégzett sekély szintaktikai elemzés eredményeként kapott reprezentáció: először – stem@@ után – az ige, majd eset@@fej_szótöve formában a névszói és névutói csoport bővítmények eset szerinti ábécésorrendben. Az igét nem tartalmazó tagmondatokban stem@@NULL, a határozott ragozású igét, de explicit tárgyat nem tartalmazó tagmondatokban pedig ACC@NULL szerepel. Látjuk az *engem*, *mellettem* elemzését, az igekötő igéhez kapcsolását, a főnévi igenév főigekénti kezelését, a főnévi igenév melletti -nAk ragos szó alanyként való értelmezését, az igeneves (*A hasmenéstől szenvedő betegeknek*) és a birtokos szerkezet (*A Profi egyik támadójátékosa*) egy egységként való kezelését (a POSS a birtokos személyjelet kódolja).

Az igeiszerkezet-lista szintén egy egyszerű szöveges fájl, soronként egy szerkezetet tartalmaz a 2. ábrán látható, szemléletesebb, ember számára jobban olvasható formában: a Mazsola adatbázisban szereplő szokásos hárombetűs esetrövidítések helyett itt az esetragok szerepelnek (önállóan vagy a tartalmas szavak végéhez kapcsolva); a névutókat egyenlőségjel jelzi; a birtokos személyjelet pedig -A. A két formátum szükség esetén egyszerűen átalakítható egymásba. A 2. ábrán a *csap* karaktersorozatot tartalmazó néhány példát látunk.

Minden sor egy igei szerkezetet és egy gyakorisági mérőszámot tartalmaz. Az első elem mindig az igető, utána következnek a névszói és névutói csoport bővítmények. A fent leírt kinyerési módszernek köszönhetően a bővítmények között egyaránt megjelennek a szabad esetrag/névutó által képviseltek és a konkrét szóval kitöltöttek is. A fenti mér csapás-t -rA szerkezet mindkét esetet példázza: a mér komplex igét alkot a konkrét szóval kitöltött tárggyal, és ehhez a kéttagú szerkezethez járul még egy -rA ragos vonzat. A kitöltött alanyi bővítménnyel nem bíró szerkezetekhez az alanyi bővítményt implicite mindig odaértjük. A kinyerő algoritmus által szolgáltatott gyakorisági mérőszám

becsap -t 1248
lecsap -rA 620
mér csapás-t -rA 360
átcsap -bA 345
megcsappan 217
lesz csapadék 205
csap -t hón-A=alá 80
becsap ajtó-t maga=mögött 28
átcsap =fölött 20

2. ábra. Az igeiszerkezet-lista sorainak felépítése

jelentése: ennyi olyan mondat volt a korpuszban, ami megfelel az adott szerkezetnek, és nincs olyan specifikusabb szerkezet a listán, aminek megfelelne. Következésképpen ha azon mondatok számára vagyunk kíváncsiak, amikben például a *becsap* ige mellett van tárgy, akkor össze kell számolni a lista összes olyan bejegyzését, amiben ez a két elem (*becsap* + tárgy) szerepel.

3. Mennyiség és minőség

Ahogy a cím is kiemeli, igen jelentős méretű erőforrásokról van szó: ez pontosan 27970403 sekély elemzéssel ellátott tagmondatot és 535609 igei szerkezetet jelent. Mindkét mennyiség egyedülállónak mondható a magyar nyelv tekintetében. A szótárral [3] összevetve azt látjuk, hogy az igeiszerkezet-lista két nagyságrenddel bővebb anyag (a szótár csak a 250-nél nagyobb gyakorisági mérőszámmal bíró 6266 szerkezetet tartalmazza), ugyanakkor tisztítatlan, nyers adat, érvényesek rá a szótár bevezetőjében említett korlátok [3, 9-17. oldal] és természetesen nélkülözi a szótári példamondatokat, illetve mutatókat. Összevetettük az igeiszerkezet-listát egy kézzel annotált, gold sztenderd korpuszból származó félig kompozicionális szerkezeteket tartalmazó listával[1] [5] is. Azt látjuk, az igeiszerkezet-lista (a más típusú, illetve kompozicionális szerkezetek mellett) nagy mennyiségű félig kompozicionális szerkezetet tartalmaz. A nagyobb korpuszméret a gyakoriságok jobb becslésére ad lehetőséget. Kiemelendő, hogy az igeiszerkezet-listán a teljes szerkezetek (is) szerepelnek, azaz nemcsak a komplex igék, hanem a hozzájuk tartozó vonzatok is megjelennek: a *zsebre vág* szerkezetet vág -t zseb-rA formában, azaz a tárggyal együtt találjuk meg.

Tudni kell, hogy a Mazsola adatbázis bemutatott sekély szintaktikai elemzése részletesség és hibamentesség tekintetében nem közelíti meg a kézzel készített elemzések minőségét [6], ugyanakkor az erőforrás a nagy méret miatt fontos előnyös tulajdonsággal bír: a nagy korpusz lehetőséget ad a ritka jelenségek, szerkezetek jellemzésére [7, 323. oldal]. Emiatt és a kinyerő módszernek köszönhetően, a nem hibátlan elemzés ellenére van lehetőség olyan ritkább szerkezetek felfedezésére, azonosítására és gyakoriságának becslésére, mint a visz prím-t -bAn, ter-

[1] http://rgai.inf.u-szeged.hu/project/nlp/research/mwe/fx_list_hu.txt

jeszt rémhír-t, telik erő-A-bÓl -rA vagy tapos -t sár-bA. Az igeiszerkezet-lista a Mazsola adatbázis elemzési hibái ellenére képes megbízható adatokat szolgáltatni az igei szerkezetekről. A Mazsola adatbázis alapvetően az igeiszerkezet-lista elkészítése érdekében jött létre, ugyanakkor hasznosnak gondoljuk erőforrásként önmagában is közzétenni a további felhasználás érdekében. A fentiek is mutatják a kis plusz hozzáadott információt tartalmazó (például a fenti sekély elemzéssel ellátott), de nagy méretű korpuszok hasznosságát, összevetve akár a még sokkal nagyobb POS-taggelt, akár a kisebb méretű gazdag annotációval bíró korpuszokkal.

4. Példák

Alább néhány példával világítjuk meg, hogy mi mindent tartalmaz az igeiszerkezet-lista, és mire lehet alkalmas. Mint említettük, az igei szerkezetek kinyerése gyakorisági alapon történik. Ennek következtében az idiomatikus kollokációk (komplex igék) mellett megjelennek a listán az igével kompozicionális szerkezetet alkotó gyakori szavak is, a vonzatok mellett pedig az egyéb bővítmények is (eset/névutó által képviselve). Jól látszik ez, ha egy gazdag vonzatszerkezettel bíró igét vizsgálunk meg. Nézzük a *száll* legjellegzetesebb szerkezeteit a 3. ábrán.

1. száll -rA 610	11. száll =mellett sík-rA 94
2. száll 463	12. száll vonat-rA 80
3. száll vita-bA -vAl 359	13. száll maga-A-bA 72
4. száll -bA 292	14. száll -n 71
5. száll -ért sík-rA 150	15. száll sík-rA 69
6. száll -ért harc-bA 142	16. száll -bA -vAl 67
7. száll -bAn 141	17. száll -ért ring-bA 65
8. száll -vAl 134	18. száll part-rA 64
9. száll ring-bA 103	19. száll harc-bA 63
10. száll fej-A-bA 101	20. száll -rÓl -rA 61

3. ábra. A *száll* első húsz szerkezete

A 18. szerkezet (száll part-rA) tipikus komplex ige, a 12. (száll vonat-rA) talán kevésbé idiomatikus, mindenesetre itt a bővítményi helyen egyéb szavak is megjelenhetnek (villamos, busz, hajó), ahogy ez a lista további részéből kiderül. Látjuk, hogy ezek a szavak egy szemantikailag koherens osztályt alkotnak, jelen esetben a (tömeg)közlekedési eszközökét. Ilyen szóosztályokkal általában akkor találkozunk, ha egy igének egy vonzati helyén jelennek meg az odaillő, literális jelentésű szavak (vö: az eszik tárgyaként megjelenő különféle ételek). Az is gyakori megfigyelés, hogy az ilyen szemantikailag koherens osztályokból kakukktojásként ugranak ki a komplex igék, idiómák, szólások, mint például az eltörik alanyaiként szereplő testrészek közül a mécses. Vonzatra példát itt a komplex igék mellett látunk: száll vita-bA -vAl, száll sík-rA/harc-bA/ring-bA -ért, illetve

sík-rA =mellett. Az ige mellett megjelenő -bAn, -n stb. esetek különféle szabad határozók jelenlétére utalnak. Az effajta gyakori esetragok a szabad határozók miatt lényegében minden ige mellett megjelennek, vonzati funkciójukra a sokkal prominensebb megjelenés utal, például a szerepel esetében a kiemelkedően magas gyakorisági mérőszámmal bíró szerepel -bAn. A száll fej-A-bA alanyaként a teljes listában a dicsőség, vér és ital szavakat találjuk. E három szó nagyjából meg is adja azt a három fogalmi kört, ami itt előfordulhat, ez a Mazsola adatbázison ellenőrizhető. A legelöl álló száll -rA szerkezet nagyon heterogén, több különböző jelentésű szerkezetet foglal magába. A lejjebb lévő specifikusabb szerkezetek utalnak rá, hogy miféléket, de ahogy a gyakorisági mérőszám meghatározásánál erről volt szó, az itt lévő 610-es érték csakis olyan tagmondatokból állt elő, melyek mondatváza a listán szereplő egyéb szerkezetekre nem illeszkedik.

Az erőforrás hasznos lehet a vonzatok kötelezőségével foglalkozó vizsgálatokban. A listán sok olyan szerkezetpárral találkozunk, hogy az egyiket a másikból egy bővítmény/vonzat elhagyásával kaphatjuk meg. Ez arra utalhat, hogy az adott vonzat nem kötelező, elhagyható, vagy – és ez a két eset pusztán a lista alapján nem különíthető el – hogy a szerkezet sok esetben elliptikusan manifesztálódik. A felszólít, felkér és tanít esetében a sima tárgyas keret gyakoribb, mint a -t -rA keret, ez a nem kötelező -rA ragos vonzat vagy bővítmény gyanúját veti fel; a bíz, kényszerít és alapoz esetében fordított a helyzet, ekkor kötelező -rA ragos vonzatot sejthetünk.

Adott bővítményi szavakat vizsgálva megkapjuk a szót tartalmazó jellegzetes igei szerkezeteket. A *vagyon* esetében például a *rendelkezik*, *szert tesz*, *felél*, *megfoszt*, *gyarapít*, *elkoboz*, *kiforgat*, *felhalmoz* igékkel együttállva; a *tej* esetében többek között a *kifut (a) tej* vagy az *aprít (a) tejbe vmit*; a *kenyér* esetében pedig *eszik/süt/szel kenyeret*-től a *vmivel keresi (a) kenyerét*-en át a *visszadob kenyérrel*-ig.

5. A közzététel módja

A bemutatott két erőforrást oktatási, kutatási és magáncélra – az üzleti felhasználás külön megállapodás tárgyát képezheti – szabadon letölthetővé tesszük a `http://corpus.nytud.hu/isz` címen. A pontos felhasználási feltételek a honlapon olvashatók. A Mazsola adatbázist alkotó tagmondatokat ábécérend szerint, az igei szerkezeteket pedig gyakoriság szerint rendezve közöljük. Terveink szerint az erőforrások később a META-SHARE repozitóriumba is be fognak kerülni.

Néhány megjegyzés a közzététel és a szabad hozzáférés kapcsán. Van olyan álláspont [8, 4. rész], miszerint a weben szabadon elérhető anyagok korpuszépítési célú felhasználása lényegében korlátozás nélkül megengedett, főleg, ha feldolgozott, származtatott erőforrásról van szó. Ennél óvatosabb az a megközelítés, mikor az eredeti szöveg visszaállítását lényegében lehetetlenné téve ábécérendbe teszik a korpusz mondatait [9, 1. rész, „Literary texts”]. Az által, hogy esetünkben az alapegység a tagmondat, még egy lépéssel továbbmegyünk a visszaállíthatóság csökkentésében, így eljárásunk semmilyen értelemben nem tekinthető az MNSZ-ben lévő művek újraközlésének.

Azon túl, hogy a Mazsola korpuszlekérdező, illetve a Magyar igei szerkezetek szótár létrehozása során közvetlenül a bemutatott erőforrásokra építettünk, más kutatások is használták már azokat [10,11] most pedig megnyílik a lehetőség a széleskörű felhasználás előtt.

Hivatkozások

1. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389
2. Sass, B.: ,,Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi Tamás (szerk.): Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, MTA Nyelvtudományi Intézet, Budapest (2009) 117–129
3. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest (2010)
4. Sass, B.: Igei szerkezetek gyakorisági szótára - egy automatikus lexikai kinyerő eljárás és alkalmazása. PhD thesis, PPKE ITK (2011)
5. Vincze, V., Csirik, J.: Hungarian corpus of light verb constructions. In: Proceedings of COLING 2010, Beijing, China (2010) 1110–1118
6. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In Matoušek, V., ed.: Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005). Springer (2005) 123–131 Springer LNAI 3658.
7. Kornai, A.: Probabilistic grammars and languages. Journal of Logic, Language, and Information (20) (2011) 317–328
8. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**(3) (2009) 209–226
9. Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V.: Parallel corpora for medium density languages. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R., eds.: Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05. John Benjamins (2007) 247–258
10. Miháltz, M., Sass, B., Indig, B.: What do we drink? Automatically extending Hungarian WordNet with selectional preference relations. In: Proceedings of Joint Symposium on Semantic Processing, Trento (2013) 105–109
11. Pléh, Cs., Németh, K., Varga, D.: The possible role of entropy in processing argument dependencies in Hungarian. In: 16th International Morphology Meeting, Information Theory in Morphology workshop. (2014)

Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere

Sass Bálint

MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport, PPKE ITK
sass.balint@itk.ppke.hu

Korábban általánosságban bemutattuk pszicholingvisztikai indíttatású, performanciaalapú, szigorúan balról jobbra haladó magyar nyelvi elemzőnk elvi megalapozását [1]. Az igei vonzatkeretek és a tematikus szerepek kezeléséről jelen kötetben olvashatunk [2]. Most részletesebben tárgyaljuk a szűken vett központi elemző (és működtető) komponenst, illetve ennek azt a részét, ami konkrétan a balról jobbra haladást/elemzést végzi: egyesével végiglépkedve a szavakon egyfajta függőségi szintaktikai elemzésnek veti alá a szöveget.

A bemenet tokenek sorozata, ezeket veszi sorra az elemző. A balról jobbra PoS-taggernek és egyéb párhuzamosan futó komponenseknek (erőforrásszálaknak [1]) köszönhetően a szóalakon és a szótőn túl számos egyéb információ is rendelkezésre áll minden tokenhez. A tokeneket (és a belőlük képzett nagyobb egységeket, frázisokat) egy *stack*en tároljuk. Az elemző jelenleg kézzel írt *szabályok* alapján dolgozik. Egy szabály egy feltételt és egy eljárást határoz meg, ha az adott tokenre teljesül (egy vagy több) feltétel, akkor lefut a hozzá rendelt eljárás. Az eljárások tipikusan az alábbi három lépésből állnak: (1) az adott token bekerül a stackre; (2) a stackelem lezárul, ha frázis végén vagyunk; (3) az adott tokennek megfelelően valamilyen strukturális *szál* indul/lezárul, vagy valamilyen (jelenleg egyfajta: alárendeltséget kifejező függőségi) *él* keletkezik két stackelem között. Ezen élek segítségével tudunk fákat (gráfokat) építeni a stacken lévő elemek között. A strukturális szálak (röviden: szálak) két típusát különböztetjük meg: a *felkínálás*, illetve az *igény* jellegű szálat [1]. Az elváló igekötő például igét igénylő szálat indít, az igekötőmentes ige egy felkínáló szálat (az esetleges igekötő részére), a szálak típusa természetesen független az ige és igekötő sorrendjétől. Az épp futó szálakat egy halmazban tartjuk nyilván, ehhez a halmazhoz fordul az elemző minden egyes token feldolgozásakor. A (tag)mondat végén az elemző értékelést készít, összesíti a (tag)mondatban lévő felsőszintű elemeket. E reprezentáció segítségével fogjuk tudni megvalósítani a kitűzött célt, hogy a szöveg alapján válaszolni tudjunk az olyan kérdésekre, hogy ki, mit csinált, hol és mikor.

A párjukat kereső felkínálás-igény szálakat tartalmazó architektúránk sokban hasonlít a Link Grammar [3] felépítéséhez, de sokban el is tér tőle (performanciaközpontúság vs. kompetencia-központúság; láncolás vs. fejhez kötés; kategoriális vs. lexikalizált stb.). Már ez a klasszikus cikk nehézségként említi a koordinációkat: bizonyos koordinációk a Link Grammar eszközeivel – a linkek kereszteződése miatt – közvetlenül nem kezelhetők. Egy friss magyar tanulmány [4], mely az adatvezérelt függőségi elemzés hibaelemzésével foglalkozik, számos más probléma mellett – talán legnehezebbként – szintén a koordináció kérdését emeli ki. Más problémákkal ellentétben a koordináció kezelésére nem is ad javaslatot a

cikk, hanem a szemantika területére utalja, azzal érvelve, hogy a koordinációk gyakran csak kontextuális vagy szemantikai háttértudás segítségével értelmezhetők helyesen.

A továbbiakban arról lesz szó, hogy hogyan kezeljük a koordinációt a vázolt architektúrában. Alapelv, hogy a felsorolás valamilyen értelemben azonos típusú elemekből áll. Az elemzőben a koordinációt (felsorolást) egy speciális fajtájú felkínáló szál segítségével kezeljük. Ez a szálfajta tartalmaz egy külön adattagot (*pattern*), mely a felsorolás kezdete óta feldolgozott felső szintű egységekről szóló információt (PoS-tag stb.) tárolja annak érdekében, hogy ellenőrizni tudjuk az említett alapelv teljesülését. A konjunktív elemeknél (*és*, *vagy*, vessző) elindítunk egy ilyen felsorolás-szálat, úgy, hogy természetesen a konjunktív elemet megelőző egységről szóló információt is hozzávesszük. A szál külső (pl. névutó, *és* utáni vessző megjelenése, mondat vége) vagy belső esemény (pl. *A és A* alakú *pattern*) hatására zárul le. Lezáráskor kiértékelődik a *pattern* adattag: nem egyszerű azonosságot vizsgálunk, például egyes szám harmadik személyű ige és alanyesetű melléknév megengedett ugyanazon felsoroláson belül. A beazonosított felsorolásokból egy új komplex elemet képezünk a stacken, a komplex elem feje a koordinált elemek fejeinek összessége lesz. Az új komplex elem egy egységként funkcionál aztán tovább, akár egy újabb koordinációs szerkezet egyik tagjaként.

Alább néhány példmondat elemzésének eredményével illusztráljuk a fentieket.

- *„A kutya megkergette és megharapta Marit ."* ([3] problematikusként említett példájának fordítása.) A felsorolás lezárulását az *A és A* mintázat váltja ki, a mondat alanyt, tárgyat és állítmányt (*megkergette+megharapta*) tartalmaz.
- *„Józsi és Pisti mellett nem fut el ."* A felsorolás lezárását itt a névutó váltja ki, a felsorolás feje/lemmája a *Józsi+Pisti* lesz, és utána a felsorolás egységhez kapcsolódik hozzá a névutó.
- *„Aláírják a finanszírozási szerződést a Budapesti Közlekedési Központ igazgatósága és a Fővárosi Közgyűlés jóváhagyásával ."* Az egy egységként azonosított terjedelmes koordinációhoz járul hozzá a birtok: a mondatban állítmány, tárgy és *-vAl* ragos bővítmény (valamint kikövetkeztetett alany) van.
- *„Romulus és Remus, Róma későbbi két városalapítója egy fügefa árnyékában szopta a farkasanya tejét ."* A *Romulus és Remus* koordináció lezárul a követő vessző hatására, majd első eleme lesz ugyanezen vessző által indított másik koordinációnak. Végül ez utóbbi értelmezőként (*Róma későbbi két városalapítója*) elemződik.

Hivatkozások

1. Prószéky, G., Indig, B., Miháltz, M., Sass, B.: Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2014), SZTE, Szeged (2014) 79–87
2. Miháltz, M., Indig, B., Prószéky, G.: Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű elemzőben. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2015). (2015) 298–302

3. Sleator, D., Temperley, D.: Parsing English with a link grammar. In: Proceedings of the Third International Workshop on Parsing Technologies. (1993)
4. Farkas, R., Vincze, V., Schmid, H.: Dependency parsing of Hungarian: Baseline results and challenges. In: Proceedings of the 13th Conference of the EACL, Avignon, France (2012) 55–65

SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz

Vincze Veronika[1,2], Hegedűs Klára[3], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged Árpád tér 2., e-mail: {vinczev,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[3]Szegedi Tudományegyetem, Pszichológiai Intézet
e-mail: klarahegedus92@gmail.com

Kivonat Ebben a munkában bemutatjuk a SzegedKoref nevű, magyar nyelvű, teljes egészében kézzel annotált koreferenciakorpuszunkat, amely nagy méretének köszönhetően a későbbiekben alkalmas lehet különféle koreferenciafeloldó algoritmusok tanítására és kiértékelésére is. Ismertetjük a korpusz felépítését, az annotációs elveket, majd statisztikai adatokat közlünk az annotált nyelvi jelenségekről.
Kulcsszavak: koreferencia, anafora, korpusz

1. Bevezetés

A természetes nyelven írt szövegekben általában megfigyelhető, hogy a szerzők szóhasználatukban változatosságra törekednek, kerülik az ismétléseket, többféle kifejezést használnak ugyanarra az entitásra. A nyelvek mind lexikai, mind grammatikai síkon is lehetőséget biztosítanak erre a változatosságra. A grammatikai eszközök közé sorolhatjuk a koreferenciajelenségeket, azaz az anaforát és a kataforát. Anaforának nevezzük azt a jelenséget, amikor a szövegen belül egy elem visszautal egy másik elemre (antecedensre) oly módon, hogy a két elem azonos entitást jelöl, azaz koreferensek. Ennek ellentéte a – jóval ritkábban előforduló – katafora, amikor egy elem a szövegen belül előreutal egy később előforduló elemre. Koreferenciaviszonyokat leggyakrabban névmások, határozószók és bizonyos főnevek (például személyek esetén általában nemet vagy rangot jelölő főnevek) fejeznek ki. Lexikai szinten pedig többnyire szinonimák alkalmazása segíti a szóhasználatbeli változatosság elérését.

A szövegek jelentésének megértéséhez szükséges annak ismerete, hogy a szövegbeli egyedek a világ mely egyedeire referálnak, illetve melyek azok a szövegbeli egyedek, amelyek azonos egyedre utalnak a világban. A számítógépes nyelvészetben egyrészt a normalizálás feladata a szövegbeli egyedek egységes formára hozása (például az *OTP*, *Országos Takarékpénztár* és *OTP Bank* kifejezések egymáshoz rendelése), másrészt pedig a koreferenciafeloldás segítségével lehetséges meghatározni az azonos egyedre utaló szövegrészeket. Míg a normalizálás általában névelemekre alkalmazott eljárás, így azzal a sajátsággal rendelkezik, hogy a szövegben előforduló minden egyes *OTP Bank* kifejezés az *Országos*

Takarékpénztárra utal, addig a koreferenciafeloldás nem csak tulajdonnevekre alkalmazható, mivel a szövegbeli összes antecedens azonosítása fontos részfeladat nyelvtechnológiai célalkalmazások (főleg az információkinyerés) számára, továbbá ugyanannak az anaforikus elemnek akár mondatról mondatra is változhat az antecedense (például személyes névmások esetében).

Ebben a munkában bemutatjuk a SzegedKoref nevű, magyar nyelvű, teljes egészében kézzel annotált koreferenciakorpuszunkat, amely nagy méretének köszönhetően a későbbiekben alkalmas lehet különféle koreferenciafeloldó algoritmusok tanítására és kiértékelésére is. Cikkünkben ismertetjük a korpusz felépítését, az annotációs elveket, majd statisztikai adatokat közlünk az annotált nyelvi jelenségekről.

2. Kapcsolódó irodalom

A világ számos nyelvére létezik koreferenciára annotált korpusz, például az OntoNotes adatbázis [1, 2] angol, kínai és arab nyelvre tartalmaz koreferenciaannotációt. Ez a adatbázis szolgált a CoNLL-2011 [3] és CoNLL-2012 [4] versenyek alapjául, ahol a feladat automatikus koreferenciafeloldás volt.

Francia és német nyelvre beszélt nyelvi korpuszokban, a DIRNDL és ANCOR_Centre korpuszokban találhatunk koreferenciaannotációt [5,6]. Japán nyelven a NAIST Text korpusz tartalmaz koreferenciajelölést, a predikátum-argumentum viszonyok jelölése mellett [7]. Lengyel nyelvre is készült nagyméretű, koreferenciaannotált korpusz [8, 9], emellett holland [10] és cseh [11] nyelvekre is elérhetők korpuszok.

Magyar nyelvre is készült már egy kisméretű, kézzel annotált koreferenciakorpusz [12]. Jelen cikkben egy nagyméretű, teljes egészében kézzel annotált magyar nyelvű koreferenciakorpusz elkészítését ismertetjük, mely a későbbiekben méreténél fogva alkalmas lehet gépi tanuláson alapuló koreferenciafeloldó rendszerek tanítására és kiértékelésére is, ami a manapság legelterjedtebb eljárás koreferenciaviszonyok azonosítására (vö. [4]).

Morfológiailag gazdag nyelvek esetében – mint amilyen a magyar is – a koreferenciaviszonyok jelölése nehézségekbe ütközhet bizonyos nyelvi jelenségek kapcsán. Többek között a fonológiailag meg nem jelenő személyes névmások kezelése igényel különös figyelmet, vö. [13] a lengyel nyelvben tapasztalt nehézségekről. Emellett az utalószavak és mellékmondatok kapcsolatának jelölésére is külön figyelmet kell fordítani. Cikkünkben erről a két jelenségről is szót ejtünk.

3. A korpusz

Az annotálás alapjául a Szeged Korpuszt [14] választottuk, újabb kézi annotációs réteggel bővítve a szövegeket. Mivel koreferenciaviszonyokat hosszabb, összefüggő szövegekben érdemes vizsgálni, a Szeged Korpuszon belül is ki kellett választani, mely alkorpuszokban hasznos bejelölni a koreferenciakapcsolatokat. A gazdasági rövidhíreket tartalmazó alkorpuszban a hírek pusztán 1-2 mondatból állnak, így úgy döntöttünk, ezen az alkorpuszon nem végezzük el az annotálást.

Az annotálási munkálatok jelenleg is folyamatban vannak. 2014 novemberéig 181 dokumentum (újsághír, illetve iskolai fogalmazás) annotációja készült el, azonban ez a szám folyamatosan nő. A teljes koreferenciakorpusz tehát – az eddig elkészült anyagokon túl – további újságcikkeket, regényeket, jogi és számítástechnikai szövegeket, valamint iskolai fogalmazásokat fog tartalmazni.

4. Annotációs elvek

Az annotáció során összekötjük az antecedenseket és a velük koreferens elemeket. Jelöljük a névmási, főnévi, határozószói és igei anaforákat is, ahogy a következő példák is mutatják:

- Névmási anafora
  - Személyes névmás: ***Mari*** *észrevette Józsit, de a fiú nem látta* ***őt****.*
  - Mutató névmás: *Megvettem* ***a labdát****, de* ***az*** *hamarosan kidurrant.*
  - Kölcsönös névmás: ***Józsi és Mari*** *látta* ***egymást****.*
  - Visszaható névmás: ***Józsi*** *látta* ***magát*** *a tükörben.*
  - Vonatkozó névmás: *Ismertem* ***a lányt****,* ***aki*** *épp átjött az úton.*
  - Birtokos névmás: ***Józsi*** *nem tudta eldönteni, melyik labda az* ***övé****.*
  - Zéró névmás: ***A tanárok*** *látták előre* ***a konfliktust****, de (****ők****) nem tudták megakadályozni (****azt****).*
    Józsi *bejött a szobába. A(z* ***ő****) kutyája követte.*
- Főnévi anafora (NP)
  - Ismétlés: *Józsi este találkozott* ***a lánnyal****.* ***A lány*** *piros ruhát viselt.*
  - Variáns: ***Pálffy János gróf*** *személyében magyar főparancsnokot neveztek ki a császári sereg élére.* ***Pálffy*** *tárgyalásokat kezdett Károlyi Sándor báróval.*
  - Szinonima: *Józsi kapott* ***egy biciklit****. Másnap* ***az új kerékpárral*** *jött munkába.*
  - Hipernima: *Az udvaron volt* ***egy kutya****.* ***Az állat*** *keservesen ugatott.*
  - Hiponima: *Az udvaron volt* ***egy kutya****.* ***Szegény uszkár*** *meg volt kötve.*
  - Meronima: *Jól játszott* ***a csapat****,* ***a kapus*** *különösen kiemelkedett a mezőnyből.*
  - Holonima: *Defektes lett* ***a jobb első kerék****, így* ***az autónak*** *ki kellett állnia a versenyből.*
  - Epitheton: ***Józsi*** *nem tudott bejutni, mert* ***a szerencsétlen*** *otthon hagyta a kulcsot.*
  - Appozíció: *Pálffy tárgyalásokat kezdett* ***Rákóczi megbízottjával****,* ***Károlyi Sándor báróval****.*
- Határozói anafora:
  - Mutató határozószó: *Elindultunk* ***a hotelba****, a többiekkel* ***ott*** *találkozunk.*
  - Vonatkozó határozószó: *Hol jársz* ***itt****,* ***ahol*** *a madár se jár?*
- Igei anafora: *Juli* ***elénekelt*** *tegnap* ***egy dalt****, ma pedig Józsi is* ***így tett****.*
- Anafora képzett alakokkal: *Józsi mindig* ***énekel*** *a fürdőben.* ***Az éneklés*** *nagyon zavarja a többi lakót.*
  Józsi mindig ***énekel*** *a fürdőben. Az* ***éneklő*** *férfi nagyon zavarja a többi lakót.*

Az annotáció során a fenti fő kategóriákat jelöltük a szövegben. Főnévi anafora esetében jelöltük az altípust is, a névmási és határozói anaforák esetében azonban a szavak morfológiai elemzéséből kiderül, hogy melyik altípusról van szó, ezek külön jelölését tehát mellőztük.

Magyar nyelvű szövegekben az anaforák bejelölését nehezítik az ún. zéró névmások. Az alanyi és tárgyas igeragozás különbségének megléte folytán nem szükséges kitenni a tárgyi névmásokat, illetve az alanyt jelző személyes névmás kitétele sem kötelező, sőt birtokos szerkezetben is elmarad(hat) a személyes névmási birtokos. A koreferenciaviszonyok szempontjából ez annyit tesz, hogy az anaforikus elem látható formában nincs jelen a mondatban, csak zéró névmás (pro) formájában, így azokat az annotáció megkezdése előtt be kellett illeszteni a szövegbe. Egy példa:

Látta a kertjében. → ***proSUBJ*** *látta* ***proOBJ*** *a* ***proPOSS*** *kertjében.*

A zéró anaforikus névmások beszúrása a szövegbe automatikusan, morfológiai és szintaktikai megkötésekre épülő nyelvészeti szabályok alapján történt.

Külön figyelmet fordítottunk arra is, hogy a mellékmondatokra vonatkozó utalószavak is össze legyenek kötve az adott mellékmondattal, akár teljes alakban, akár zéró névmás formájában jelennek meg. Így tehát az alábbi példák mindegyikében jelöltük a névmás és a mellékmondat kapcsolatát:

Mondtam ***proOBJ****, hogy mindjárt itt a karácsony.*
Azt *mondtam, hogy mindjárt itt a karácsony.*

Az alábbiakban közlünk egy példát az annotált szövegre, indexekkel jelölve az összetartozó elemeket, illetve külön szerepeltetve az anaforikus láncokat.

Az úton **[sok ismerőssel]**$_i$ találkoztunk, **[akik]**$_i$ újságolták **[proOBJ]**$_j$ nekünk, hogy **[milyen jó a hangulat a majálison]**$_j$. Amikor leérkeztünk, már nagy volt a nyüzsgés, finom illatok szálltak a levegőben, és folytak **[a koncert]**$_k$ előkészületei, ugyanis – ha még nem írtam **[proOBJ]**$_l$ volna – **[a [Bestiák]**$_m$ **énekeltek aznap nekünk]**$_l$. Én ugyan nem nagyon szeretem **[ezt az együttest]**$_m$, de **[miattuk]**$_m$ nem hagyhattam ki **[ezt az eseményt]**$_k$. Amíg **[a koncert]**$_k$ nem kezdődött el, addig édességet ettünk a haverjaimmal, és hülyéskedtünk. Aztán egyszer csak **[sipító hangot]**$_n$ hallottunk. **[Azt]**$_o$ hittük, hogy **[a Bestiák**$_m$ **egyik énekese [az]**$_n$**]**$_o$, de **[proSUBJ]**$_p$ kiderült, hogy **[csak egy mikrofon hibásodott meg]**$_p$. Rövid várakozás után végül elkezdődött **[a koncert]**$_k$. **[A hangulat]**$_q$ a **[proPOSS]**$_q$ tetőfokára hágott, **[mindenki]**$_r$ tombolt, és együtt **[proSUBJ]**$_r$ énekelt **[a lányokkal]**$_m$. Több visszatapsolás és ráadás-dal után véget ért **[a koncert]**$_k$.

Anaforikus láncok:

sok ismerőssel – akik
proOBJ – milyen jó a hangulat a majálison

a koncert – ezt az eseményt – a koncert –a koncert – a koncert
proOBJ – a Bestiák énekeltek aznap nekünk
Bestiák – ezt az együttest – miattuk – Bestiák – a lányokkal
sipító hangot – az
azt – a Bestiák egyik énekese az
proSUBJ – csak egy mikrofon hibásodott meg
a hangulat – proPOSS
mindenki – proSUBJ

5. Statisztikai adatok

A korpusz jelenleg 309 mondatot és 9782 tokent tartalmaz az újsághírekből, illetve 3712 mondatot és 45981 tokent az iskolai fogalmazásokból, összesen 4021 mondat és 55763 token szerepel tehát a korpusz 2014 novemberi változatában. Ezekben összesen 2456 anaforikus lánc található (2191 az iskolai fogalmazásokban, 265 pedig az újsághírekben). Az anafora típusa szerinti (százalékos) eloszlást az 1. táblázat mutatja.

1. táblázat. Az anaforatípusok eloszlása.

Anafora	**Fogalmazás**	**%**	**Újsághír**	**%**	**Összesen**	**%**
névmási	1531	33,51	320	39,22	1851	34,37
ismétlés	1176	25,74	86	10,54	1262	23,44
szinonima	329	7,20	252	30,88	581	10,79
hipernímia	445	9,74	0	0,00	445	8,26
holonímia	350	7,66	34	4,17	384	7,13
epitheton	17	0,37	23	2,82	40	0,74
appozíció	117	2,56	70	8,58	187	3,47
határozói	339	7,42	1	0,12	340	6,31
igei	5	0,11	0	0,00	5	0,09
képzés	76	1,66	30	3,68	106	1,97
egyéb	184	4,03	0	0,00	184	3,42
Összesen	4569	100	816	100	5385	100

A táblázatból látszik, hogy a névmási anafora és az ismétlés a leggyakoribb anaforatípusok, e két kategória együttesen lefedi az adatok mintegy felét. Így tehát az automatikus koreferenciafeloldó rendszereknek e kategóriákra fokozott figyelmet kell fordítaniuk.

A 2. táblázat azt is elárulja, hogy a szövegekben számos zéró névmás szerepel anaforikus lánc részeként, sőt a névmási anaforák jelentős részében (mintegy kétharmadában) zéró névmás szerepel. Így a magyar nyelvű koreferenciafeloldó algoritmusoknak ezeknek a kezelésére is célszerű felkészülniük.

2. táblázat. Az anaforikus zéró névmások eloszlása.

Zéró névmás	**Fogalmazás**	**Újsághír**	**Összesen**
proSUBJ	594	119	713
proOBJ	181	9	190
proPOSS	212	128	340
Összesen	987	256	1243

6. Alkalmazási lehetőségek

A koreferenciaviszonyokra annotált korpusz, illetve a rá épülő automatikus koreferenciafeloldó rendszer felhasználási lehetőségei számos területre terjednek ki. A koreferenciaviszonyok információkinyerő rendszerek számára is hasznosak, hiszen például egy adott cégről szóló információkat nemcsak a cég nevére keresve lehet így megtalálni, hanem a cégre anaforikusan utaló elemek kikeresésével is többletinformációkra lehet szert tenni.

Fordítóprogramok is hasznosíthatják a bejelölt koreferenciakapcsolatokat, hiszen például míg a magyarban nincsenek nyelvtani nemek, addig számos nyelvben léteznek. Ha egy magyar névmás össze van kapcsolva antecedensével, ennek segítségével meg lehet határozni, hogy az idegen nyelven hímnemű, nőnemű vagy semlegesnemű névmás felel-e meg neki.

7. Összegzés

Ebben a munkában bemutattuk a SzegedKoref korpuszt, melyben kézzel megjelöltük a koreferenciaviszonyokat. Példákon keresztül ismertettük az annotálás alapelveit, illetve statisztikai adatokat közöltünk az elkészült anyagról. A jövőben szeretnénk a korpuszt bővíteni, illetőleg az annotált anyagra építve egy automatikus koreferenciafeloldó rendszert létrehozni.

Az annotált korpuszt kutatási és oktatási célokra ingyenesen elérhetővé tesszük.

Köszönetnyilvánítás

Szeretnénk megköszönni Miháltz Mártonnak, Anders Björkelundnak és Arndt Riesnernek az annotációs elvek kialakításában nyújtott önzetlen segítségüket.

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradan, S., Ramshaw, L., Xue, N.: OntoNotes: A Large Training Corpus for Enhanced Processing. In: Handbook of Natural Language Processing and Machine Translation. (2011)

2. Pradhan, S.S., Ramshaw, L., Weischedel, R.M., MacBride, J., Micciulla, L.: Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In: ICSC, IEEE Computer Society (2007) 446–453
3. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. CONLL Shared Task '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1–27
4. Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y.: CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012), Jeju, Korea (2012)
5. Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J.: ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 843–847 ACL Anthology Identifier: L14-1169.
6. Björkelund, A., Eckart, K., Riester, A., Schauffler, N., Schweitzer, K.: The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 3222–3228 ACL Anthology Identifier: L14-1683.
7. Iida, R., Komachi, M., Inui, K., Matsumoto, Y.: Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In: Proceedings of the Linguistic Annotation Workshop, Prague, Czech Republic, Association for Computational Linguistics (2007) 132–139
8. Ogrodniczuk, M., Kopeć, M., Savary, A.: Polish Coreference Corpus in Numbers. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 3234–3238 ACL Anthology Identifier: L14-1066.
9. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Polish coreference corpus. In: 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Volume 3., Wydawnictwo Poznańskie (2013) 494–498
10. Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Vloet, J.V.D., Verschelde, J.L.: A Coreference Corpus and Resolution System for Dutch. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) http://www.lrec-conf.org/proceedings/lrec2008/.
11. Nedoluzhko, A., Mírovskỳ, J., Ocelák, R., Pergler, J.: Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In: Proceedings

of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India. (2009) 1–16
12. Miháltz, M.: Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. Általános Nyelvészeti Tanulmányok **XXIV** (2012) 151–166
13. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In Sun, M., Zhang, M., Lin, D., Wang, H., eds.: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Volume 8202 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 97–108
14. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged TreeBank. In Matousek, V., Mautner, P., Pavelka, T., eds.: Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005. Lecture Notes in Computer Science, Berlin / Heidelberg, Springer (2005) 123–132

VIII. Laptopos bemutatók

Yako: egy intelligens üzenetváltó alkalmazás nyelvtechnológiai kihívásai

Farkas Richárd, Kojedzinszky Tamás, Zsibrita János, Wieszner Vilmos

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{rfarkas,zsibrita,wieszner}@inf.u-szeged.hu

1 Intelligens üzenetváltó alkalmazások

Az infokommunikációs technológiák fejlődésével egyidejűleg megfigyelhető az a tendencia, hogy az ember-ember kommunikációra egyre nagyobb arányban használunk szöveges formát (gondoljunk csak a Facebookra, chatre, e-mailre, SMS-re stb.). Számos szöveges kommunikációt támogató, üzenetváltó alkalmazás érhető el különféle platformokon, azonban ezek többsége nem rendelkezik „intelligens" funkciókkal [1]. Intelligens funkció lehet az üzenetek kategorizálása/csoportosítása, rangsorolása, összefoglalása, megjelölése vagy a szövegbevitelt támogató funkciók. Ezek a funkciók különösen hasznosak az okostelefonokon futó üzenetváltó alkalmazások számára, mert ott a kis megjelenítő- és beviteli eszközök miatt a hagyományos alkalmazások kényelmetlenek.

A Szegedi Tudományegyetemen, a FuturICT.hu projekt keretében kifejlesztettünk egy Android-alkalmazást (Yako), amely egységes felületen fogad és küld SMS-eket, e-maileket és alkalmas Facebook-beszélgetésre. Az alkalmazás felhasználóinak egy csoportjának folyamatosan mentjük, hogy hogyan váltanak üzeneteket ismerőseikkel (természetesen az adatvédelmi irányelvek tiszteletben tartása mellett[1]). Az így gyűjtött adatbázis jó alapul szolgál az üzenetváltásokhoz kapcsolódó intelligens funkciók fejlesztéséhez. A jelenleg fejlesztés alatt álló intelligens funkciók nagy része a nyelvtechnológia területéhez kapcsolódik: üzenetek fontosság szerinti rangsorolása, automatikus válaszgenerálás, témaszálak azonosítása, összefoglalás/kulcsszavazás, információk kiemelése. Az összegyűjtött adatbázis alapján megállapíthatjuk, hogy számos speciális nyelvtechnológiai kihívásokkal kell szembenéznie azoknak, akik magyar nyelvű szöveges üzenetváltást támogató intelligens funkciókat terveznek megvalósítani.

2 Ékezetesítés

Az első ilyen kihívás az automatikus ékezetesítés. Míg a desktopokon írt magyar üzenetek 95%-a ékezethelyesnek mondható, a mobil eszközökről írt üzeneteknél ez az arány mindössze 63%. A nyelvi előfeldolgozás egyik fontos lépése az ékezetek hely-

[1] A projekt keretében egy adatvédelmi hatástanulmány is elkészült.

reállítása [3]. A [2]-ben alkalmazott módszerekhez hasonlóan kidolgoztunk egy egyszerű eljárást, ami nagy ékezethelyes szótár segítségével, illetve többértelműség esetén a környező szavak figyelembe vételével valószínűségi döntést hoz.

3 Kérdések elemzése

Míg az üzenetváltások a szóbeli dialógusokhoz hasonlítanak, addig az elérhető annotált korpuszok leíró dokumentációkat tartalmaznak, melyekben jellemzően kevés kérdés fordul elő. Az üzenetváltások esetében nagyon gyakori a kérdés-válasz jellegű párbeszéd. A sztenderd korpuszokon tanított elemzők azonban a kérdéseken kifejezetten rosszul teljesítenek (mert kevés tanítópélda állt rendelkezésükre). Első funkcióként implementáltunk egy eldöntendő kérdéseket, illetve a döntés tárgyát azonosító egyszerű módszert. A kérdések elemzését egy gyorsválasz funkcióban használjuk fel. Egyszerű eldöntendő kérdés esetén *igen*/*nem* a lehetséges gyorsválasz, míg alternatívák közti választás esetén maguk az alternatívák. Például a *A TIK parkban vagy a Pivo Várban találkozzunk?* kérdésre a *TIK parkban* és a *Pivo Várban* lesz a két gyorsválasz. A felismerő szabályok a magyarlanc [4] szófaji elemzésén alapulnak, így a *Hogy vagy barátom?* kérdésre nem lesz gyorsválaszi lehetőség a *Hogy* és a *barátom*.

4 Szövegen kívüli kontextus

Az intelligens funkciók megvalósításához nem elégséges a szöveges üzenetek elszigetelt elemzése, a tágabb értelemben vett környezet (pl. ki a partner) elemzése engedhetetlenül fontos. Az alkalmazásba bevezettünk ún. zónákat, amelyek közt a mobil eszköz a geolokáció alapján vált. Célunk az, hogy a rendszer különböző zónákban, különböző élethelyzetekhez igazodva másképp működjön. Különböző üzenetek fontosak például akkor ha otthon vagyunk, mint ha a munkahelyünkön tartózkodunk. A yako alkalmazásba építettünk egy szabályalapú fontosüzenet-detektáló módszert, ami a különböző zónákban különböző üzeneteket tekint fontosnak (a fontos üzenetek a beérkezett lista elejére kerülnek és az értesítésük is figyelemfelkeltőbb).

A bejövő üzenetek fontosságának eldöntésekor a szöveges tartalom mellett természetesen nagyon fontos a feladó személye és a fogadó és a feladó viszonya, mint szövegen kívüli kontextus. A rendszerünk egyszerű statisztikák felhasználásával egy felhasználó kapcsolatait öt különböző kategóriába sorolja be és ezeket a kategóriacímkéket felhasználjuk a fontosság eldöntéséhez.

Köszönetnyilvánítás

A kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

Hivatkozások

1. Dredze, M.: Intelligent Email: Aiding Users with AI. PhD thesis, Computer and Information Science, University of Pennsylvania (2009)
2. Zainkó, Cs., Németh, G.: Ékezetek gépi helyreállítása. In: A magyar beszéd: Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek. Budapest: Akadémiai Kiadó (2010) 485–488
3. Kornai, A., Tóth, G.: Gépi ékezés. Magyar Tudomány No.4 (1997) 400–410
4. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelmûsítés. In: Tanács, A., Vincze, V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem (2013) 368–374

HumInA projektcsoport a ℜeALIS1.1 bázisán

Nőthig László, Alberti Gábor

PTE BTK Nyelvtudományi Tanszék
ℜeALIS Emléleti, Számítógépes és Kognitív Nyelvészeti Kutatócsoport
nothig.laszlo@gmail.com, alberti.gabor@pte.hu

Kivonat: A laptopos bemutatóra benyújtott szoftver egy mobiltelefonra szánt alkalmazás személyi számítógépes verziója. Létrehozásának fő célja a mindennapi kommunikációt meghatározó kapcsolati információk gyors áttekintése, menedzselése. A szoftver a ℜeALIS főbb felismerésein alapuló, annak szellemiségét hordozó, könnyen kezelhető alkalmazás, amelyhez hamarosan kapcsolódhat egy, a korábbi elméleti munkákon és számítógépes programjainkon alapuló nyelvi elemző, ahol a rendszer megkeresi az ismerhető szereplőket, entitásokat, eseményeket. A ℜeALIS rendszer HumInA – Humán Intelligenciájú Adatbázis – modulja egy potenciálisan széles felhasználói kör részére elkerülhetővé teszi azt a munkát, amit egy világmodell részletes kidolgozása jelent. Az elnevezés onnan ered, hogy az információt nem önmagában, hanem az emberi kommunikációs folyamattal együtt próbáljuk megragadni, hogy ezáltal további értékes adatokhoz jussunk. A bemutatásra kerülő szoftver legegyszerűbb funkciója különböző mondatok tárolása, egy használható és áttekinthető világocska-rendszer létrehozása, valamint a más-más kontextusban előforduló entitások és események közötti kapcsolatok rögzítése. A mindennapi életben általában másoktól szerzett információra támaszkodunk, ahol forrásaink is csak részleges tudással rendelkeznek. Amikor egy kijelentés igazságáról szeretnénk dönteni, valamilyen súlyozás szerint vesszük figyelembe a környezetünkből érkező információkat. A program különböző stratégiák alapján megbízhatósági kereséseket tud végrehajtani, ahol eltérő heurisztikák alkalmazásával próbálja figyelembe venni az egyes források hitelességét.

1. Az információtartalom új dimenziói

1.1. Már nem az arany, a vas, a szén vagy a kőolaj megszerzése, birtoklása, továbbítása és elosztása mozgatja a világot, hanem az információ vált a jelenkor legfontosabb „ásványkincsévé”. Olyan mennyiségű adat vált elérhetővé az interneten, ami csodálatos lehetőségeket kínál egyfelől, másfelől viszont ott a veszély, hogy belefulladunk, elveszünk benne.

A területtel nyelvészeti és informatikai kongresszusok külön szekciói foglalkoznak, mi most egy új megközelítésre hívjuk fel a figyelmet.

Nyilvánvaló, hogy adatbázisainkat a lehető legintelligensebb módon kell megszervezni… Mi azonban ez a „legintelligensebb mód”? Az emberi intelligencia?

Aligha választhatunk más kiindulópontot. A legóvatosabb megfogalmazás mellett is kijelenthetjük, hogy érdemes a humán intelligenciából a lehető legtöbbet ellesni, és

az eltanult elemeket kamatoztatni az adatbázis-szervezés területén. Mintha az eddigi adatbázisokat „kommunikációs tudatra” ébresztenénk.

1.2. A nyelvtudományi háttérben a döntő felismerés az, hogy egy-egy mondat információtartalma nem pusztán a világban értékelhető logikai tény (*igaz vagy hamis?*). Hanem az emberi kommunikációban megsokszorozódik, bizonyos értelemben újabb dimenziókat nyer [1], és az ezekben a dimenziókban rejlő információ gyakran fontosabb, mint a „csupasz tények” [7]. Ha nem ragadjuk ki az információt a kommunikációs folyamatból, hanem azzal együtt ragadjuk meg, akkor e kommunikációs folyamat úgy működik, mint egy prizma, amelynek sok-sok lapján megsokszorozva és némileg mutálódva észleljük az elemi információs egységeket.

1.3. Vegyük górcső alá az (1a) példában megadott egyszerű mondatot:

1. példa. Információ, túl az igazságon…

a. Anna tegnap felhívta Bélát.
b. Anna tegnap felhívta Bélát?
c. Dóra a kérdés elhangzását követően tisztában van azzal, hogy
   Eszter megtudta:
   Csaba úgy gondolja, hogy
   ő, Dóra, tudja, hogy
   felhívta-e tegnap Anna Bélát.
d. Anna egy adott napon felhívta Bélát.
e. Mire vágyik Csaba, hogy
   Eszternek milyen kérdésekben legyen tévhiedelme?

Ha kérdő mondatként hangzott el (1b), akkor még csak az sem derül ki, hogy igaz-e az említett hívás ténye, vagy sem. Vegyünk tehát ilyen jellegű esetet: tegyük fel, hogy Csaba éppen ezt kérdezte Dórától, Eszter jelenlétében! Mi ebben a nyelvileg megragadható információ?

1. Az, hogy Csaba a kérdésével kinyilvánítja, hogy ő nem tudja, hogy Anna tegnap felhívta-e Bélát. 2. Dóráról viszont azt feltételezi, hogy ő tudja az igazat. 3. Túl a tudáson és feltételezésen, vágyak és szándékok is megfogalmazódnak. Csaba például a szóban forgó kérdéssel kinyilvánítja azt a vágyát, hogy szeretné tudni, felhívta-e Anna Bélát. 4. Csaba szándéka: felébreszteni Dórában a szándékot kíváncsisága kielégítésére. 5. Mindezt (1–4.) érdemes lehet regisztrálni egy adatbázisban, például egy nyomozás vagy egy bírósági tárgyalás bármelyik szereplője számára, megragadva ily módon Csaba (adott pillanatbeli) „információállapotát”. Arról sem megfeledkezve az adatbázis-építés során, hogy Dóra és Eszter is levonhatta az 1–4. következtetéseket, hosszabb távon akár olyan összetett gondolatokra jutva, amelyet a fenti (1c) pontban említünk.

A fenti példával azt kívántuk bemutatni, hogy érdemes olyan adatbázist építeni, amelyben a fenti (1d) pontban megadott információ mellett nem pusztán egy IGAZ vagy HAMIS értékelés áll, hanem egy rugalmasan bővíthető hipotézisegyüttes egymással kommunikációs kapcsolatban álló szereplők tudásáról, vágyairól, szándékairól. Az is értékes információ, hogy valaki nem IGAZ vagy HAMIS választ adna egy kérdésre, hanem valamiféle „0” igazságértéket („nem tudom”).

Szögezzük le gyorsan: a regisztrált adattömeget nem kell készpénznek venni. Csaba akár „hazudhatott" is a kérdésével, amennyiben például valójában biztos tudása van arról, hogy Anna felhívta Bélát. Ez a fajta mímelés érvénytelenné teszi a fenti 1–4. pontokban regisztrált információ egy részét. Ez azonban egyáltalán nem jelenti azt, hogy az információ rögzítése fölösleges vagy hibás lépés volt. Csupán azt jelenti, hogy az adatbázis felhasználójának fel kell készülnie arra, hogy egyes adatcsoportokra átértékelő műveleteket kell majd alkalmaznia, bizonyos eséllyel. Amennyiben például biztosan tudja, hogy Csaba tisztában van a szóban forgó telefonhívás létével, rendkívül fontos információhoz jut: ahhoz, hogy Csaba meg akarja téveszteni a megkérdezett Dórát és/vagy a beszélgetésüket figyelő Esztert. Egy intelligensen strukturált adatbázison alternatív értékelési mechanizmusokat lehet lefuttatni, a felhasználói igények függvényében. Vagy szűrni lehet bizonyos adattípusokra, például a fenti (1e) pontban megadottra (pl.).

2. Felhasználói adatbázisok építése

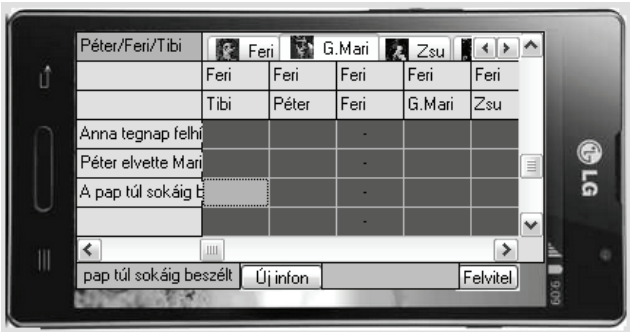
2.1. A címben említett HumInA projekt keretében ilyen jellegű *humán intelligencia szerint szervezett adatbázisokat* építünk ki a ℜeALIS1.1 keretrendszerre alapozva [2] [6].

Az elmúlt években a ℜeALIS [1] számítógépes implementációja során két eltérő, de egymást kiegészítő megközelítés merült fel a nyelvi elemzés megvalósításakor. Az egyik esetben a vonzatokra és az egyéb szintaktikai összetevőkre fókuszáltunk, hogy az adott (magyar) nyelv szabályait vezérfonalnak választva építsük fel a mondatot. A másik – „nyelvfüggetlen" – esetben elképzelésünk szerint a mondat elemeinek összerendelése eleinte a felhasználó feladata kell, hogy legyen – bár a program természetesen ezt több eszközzel is támogatja. Ha az elsődleges cél a mondatok igazságértékelése, hosszabb szövegek pragmatikai-szemantikai elemzése – ami egyébként a valódi kihívást is jelenti –, akkor a második megközelítés is vállalható, hiszen a szintaktikai elemzés „csak" ahhoz szükséges, hogy a mondatban megtaláljuk a – megfelelő – kapcsolatokat az egyes összetevők között.

A ℜeALIS rendszer HumInA modulja a saját területén feloldja ezt a dilemmát, és egy potenciálisan széles felhasználói kör részére elkerülhetővé teszi azt az egyébként nem megspórolható munkát, amit egy világmodell aprólékos kidolgozása jelent.

Az alkalmazás legegyszerűbb funkciója különböző mondatok tárolása, amelyeket egymástól eltérő nézetekben láthatunk. Az alapképernyő egy munkaasztal, amely megjeleníti az adatbázis mondatait, a respektált szereplőket, ismerősöket, a mobil tulajdonosát és a szereplők viszonyát az állításokhoz, vagyis melyiküknek milyen világocskája, vagy világocskái „tartalmazza" (tartalmazzák) az adott információt. A megvalósítás egyik nehézsége egy többdimenziós halmaz két dimenzióba való ergonomikus leképezése volt. Ez egy idő után ahhoz az immár technikai kérdéshez vezetett, hogy miként lehet az egyik „összevont" dimenzió mentén az információt könnyen áttekinthetővé és gyorsan bejárhatóvá tenni. Bár már egy tucatnál alig több „ismerős" és 3 szintű világocskacímke-rendszer esetén is több ezer az egy mondathoz tartozó világocskacímkelánc-helyek száma és ezek mindegyike több világocskacímkét

„tartalmazhat”, a kész alkalmazásban – ennél sokkal több szereplő esetén is – a navigálás könnyen és gyorsan történik.



1. ábra. A program áttekintő képernyője.

2.2. A program használata során a felhasználó egy karaktersorozatot – általában egy mondatot – ír az adatbázisba. Ezt olyan mélységig elemzi, amilyen mélységig akarja. A bejegyzés egysége az infon [7], azaz valamilyen eseménybe való „belelátás” egy kiválasztott időpillanat aspektusából [1,3]. Magát az infont a rendszer nem elemzi, viszont rögzíthetők hozzá a felhasználó számára releváns szereplők, a további entitások és a kapcsolódó események. Az egyes elemek összekapcsolhatók, hierarchiába rendezhetők. A felhasználó maga dönti el, hogy egy esemény ugyanaz-e, mint ami egy korábbi bejegyzésben szerepelt, esetleg az esemény egy más aspektusában (pl. korábban mint egy jövőbeli terv jelent meg, most pedig éppen a kumulatív szakaszában tart [3]). Létre lehet hozni forgatókönyveket (definiálható címkével ellátott űrlapokat, rendezett n-esekből álló sablonokat) az egyes eseményekre, amelyekkel kapcsolat létesíthető az események és a szereplők között. Az infonok különböző világocskákba – a külvilágnak az egyes szereplők elméjében a tudás, a vágyak, a szándékok szerint megsokszorozott alternatív világába – kerül(het)nek, amelyek a „Tudom, hogy Jóska tudja, hogy én tudom…” állításhoz hasonló kijelentésekhez vezetnek. A világocskákban lévő információ áttekinthető struktúrákban jelenik meg, a bevitt adatok és kapcsolataik különféle nézetekben láthatók, a rendszerben különböző lekérdezések és listák készíthetők.

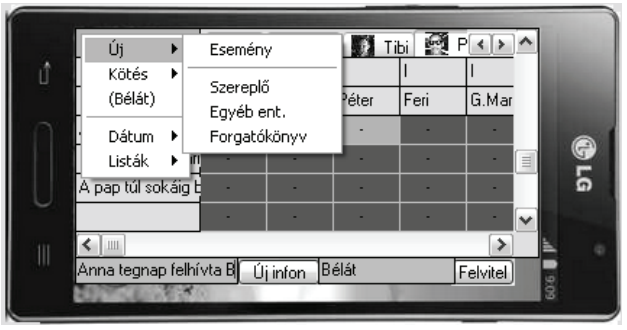
2.3. A fentieket is szeretnénk megvilágítani egy példával. Tekintsük a következő két mondatot [4]:

2. példa.

a. Mari múlt szombaton feleségül ment Péterhez.
b. … A pap másfél órán át beszélt a házasság nehézségeiről.
b'. … ??A kutya veszettül ugatott.

Ebben a klasszikus példában az első mondatból bár logikailag nem következik a pap jelenléte, kulturális enciklopédikus tudásunkban az esküvőhöz asszociálódik (2a+b), szemben egy kutya említésével (2a+b').

A felhasználó a mondatokat minden további információ megadása nélkül is beírhatja, így azonban csak jegyzet írására használja a programot. Ha azonban bármilyen okból úgy gondolja, hogy érdemes más adatokat is rögzíteni (pl. több Péter nevű ismerőse van, vagy valamelyik szereplő később várhatóan újra elő fog fordulni az adatbázisban), megteheti a következőket: Felvesz egy Péter nevű entitást, mint új személyt, vagy azonosítja egy korábban már felvett szereplővel (pl. *az egyetlen unokatestvérem*). Mari szintén bekerülhet, mint új szereplő, vagy azonosítható egy korábbival. Nemcsak az entitásokat, hanem az eseményt is "nevesíthetjük", létrehozhatunk egy új eseményt (vagy ha erre már korábban sor került, ahhoz hozzákötheţjük), aminek címkéje pl. *Mariék esküvője*. A második mondatot már ehhez az eseményhez tudjuk kötni. Ezek után a *pap* szintén felvehető, mint (valószínűleg új) személy. Az esküvő általában sok szereplő részvételével zajlik, amelyre előre megírt forgatókönyv készülhet, és később az aktuális esemény sablonjaként használható. Ezen az űrlapon lesz tehát egy menyasszony, egy vőlegény és további rovatok esetleg opcionális szereplők – például a pap – számára. Ezt a szereposztást a sablon alapján könnyen elkészíthetjük a beírt mondatokhoz az adott eseményre. A tipikus forgatókönyvek megírása nem szükségképpen a felhasználó feladata, célszerűen azok az alkalmazás letöltésekor, frissítésekor kerülhetnek a mobil eszközre.



2. ábra. Új szereplő felvitele.

2.4. A sokszor látszólag magától értetődő kapcsolatok létrehozásakor döntő a felhasználóbarát működés – különösen egy mobiltelefonos alkalmazás esetén. Ennek elérésére szolgál például a kiválasztott szó kattintással történő szerkesztőterületre hozása, az aspektus menü nélküli kiválasztása, vagy a megjelenítésben annak az elvnek a végigvitele, hogy a mondatrészek soha nem önmagukban, hanem előfordulásukkal, teljes környezetükkel együtt jelennek meg. A munkaasztal fő vizuális komponense egy csúszka segítségével a rácsméret változtatásával az olvashatóság határáig „fokozatmentesen” átméretezhető, de a kurzor mellett minden esetben olvasható az alatta lévő cella, vagy mező tartalma.

2.5. A megfelelő kötések létrehozása hasznos lehet a felhasználó számára, ugyanakkor modellezi az információ tárolását a mindennapi kommunikáció során. Bár a háttér lényegében ugyanaz, mint a ℜeALIS alaprendszernél, akadnak eltérések is. Az alkalmazással a mobilkészüléket készségszinten használókat céloztuk meg:

teljes világmodell építésre nincs szükség, nem „igazságértékelés” a cél. A beszélő és a hallgató (sőt további jelenlévők is) viszont ugyanúgy rögzíthetők és akár az alapmunkaasztalon is megjeleníthetők. Ez egy, a világocska-rendszer ábrázoláshoz képest elemibb információt tartalmazó nézet, amelynek szerkezete, kereshetősége ugyanolyan. Ebből a nézetből a felvitt mondatok, pl. (1b) megfelelő világocskákba történő „szétosztása” egy menüpont meghívásával történik, ahol a funkció végrehajtása előtt a felhasználónak be kell állítani a főbb kommunikációs szerepeket (a példánál elsősorban a kérdezőét). Egyáltalán nem mindegy ugyanis – amint láttuk – hogy a prototipikus alapkérdés hangzik el, vagy a korábban említett mímelésről, esetleg tanári vizsgakérdésről, netalán nyomozói keresztkérdésről van szó. A kialakuló adatbázis többféleképpen lekérdezhető. Egy magától értetődő használati mód az egyszerű keresés, ahol már nemcsak szótöredékes egyezés alapján fogunk találatokat kapni, hanem az egymással azonosított, ill. az asszociált entitások és események is az eredménylistára kerülnek.

2.6. A mindennapi életben az információ kis része jut közvetlenül, érzékszerveinken keresztül tudomásunkra. Általában másoktól szerzett adatokra támaszkodunk, ahol forrásaink is csak parciális tudással rendelkeznek. Amikor egy kijelentés igazságáról szeretnénk dönteni, figyelembe vesszük a környezetünkből érkező – lehetőleg mértékadó – információt. A program megbízhatósági kereséseket képes végrehajtani különböző stratégiák szerint. A figyelembe vett szereplők megfelelő világocskái alapján kiértékeli az ugyanarra az állításra vonatkozó véleményeket, és elfogadásra vagy elutasításra javasolja a felhasználó számára. Ahhoz, hogy súlyozni lehessen az „informátorokat”, különböző eljárások készültek. A legegyszerűbb, amikor ismert igazságértékű állításokon teszteljük a szavazatokat. Más esetben különböző heurisztikák alkalmazhatók, amelyek jellemzően több iterációs lépést tartalmaznak.

Köszönetnyilvánítás

A jelen tudományos közleményt a szerzők a Pécsi Tudományegyetem alapításának 650. évfordulója emlékének szentelik. A szerzőket e cikk alapjait jelentő kutatásaikban és a konferencia-részvételben a TÁMOP 4.2.2.C-11/1/KONV-2012-0005 (Jól-lét az információs társadalomban) kutatási projektum támogatta.

Hivatkozások

1. Alberti, G.: ℜeALIS: Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
2. Alberti, G., Nőthig, L.: ℜeALIS1.1: The Toolbox of Generalized Intensional Truth Evalation. In: Grzymala-Busse, J., Schwab, I., eds.: INTELLI 2014, The Third International Conference on Intelligent Systems and Applications (INTELLI) (2014) 60–66

3. Farkas, J., Ohnmacht, M.: Aspect and Eventuality Structure in a Representational Dynamic Semantics. In: Alberti, G., Kleiber, J., Farkas, J.: Vonzásban és változásban, PTE Nyelvtudományi Doktori Iskola, Pécs (2012) 353–379
4. Kálmán, L.: Deferred Information: The Semantics of Commitment. In: Kálmán, L., Pólos, L., eds.: Papers from the Second Symposium on Logic and Language, Akadémiai Kiadó, Budapest (1990) 125–157
5. Nőthig, L., Alberti, G., Dóla, M.: ℜeALIS1.1. In: Tanács, A., Varga, V., Vincze, V., eds. : X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2014) 364–372
6. Seligman, J., Moss, L.: Situation Theory. In: van Benthem, J., ter Meulen, A., eds.: Handbook of Logic and Language, Elsevier, Amsterdam / Cambridge (1997) 239–309
7. Vadász, N., Alberti, G., Kleiber, J.: The Matrix of Beliefs, Desires and Intentions. International Journal of Computational Linguistics & Applications 4/1 (2013) 95–110

Neticle – Megmutatjuk, mit gondol a web

Szekeres Péter

Neticle Technologies Kft. (http://www.neticle.hu)
Budapest, Magyarország
peter.szekeres@neticle.hu

Kivonat: A Neticle rendszer (http://www.neticle.hu/) összefoglalja a web véleményét egy adott témával, ha úgy tetszik, kulcsszóval kapcsolatban. Ehhez az egyik legfontosabb mutatónk az úgynevezett webes véleményárfolyam, mellyel egyszerűen követhető a webes jelenlét.

1 Bevezetés

Az internezők értékelnek, beszámolnak, kritizálnak. Terméket, szolgáltatást, céget, piacot, eseményt. Sok hasznos információ és tudás hever szerte a világhálón.

A Neticle-lel az volt a célunk, hogy egy olyan webes szolgáltatást hozzunk létre, amelynek segítségével a lehető legegyszerűbben felhasználhatjuk ezeket az információkat az üzleti döntéseinkhez. A Neticle (http://www.neticle.hu) olyan böngészőből elérhető szolgáltatás, amely a magyar nyelvű webes szövegek automatikus elemzésével, értékelésével és vizualizálásával a közel valós idejű nyomon követést és a tény alapú döntéshozatalt támogatja.

2 Adatok gyűjtése

A rendszer alapja egy olyan crawler (kereső), mely a web magyar nyelvű tartalmait megkeresi és kategorizálja előre meghatározott csoportok szerint. Az így kialakított osztályok segítségével a crawler meghatározza, érdemes-e az oldalt újralátogatni, és ha igen, akkor milyen gyakran, hogy a frissített tartalmakat vagy bővülő hozzászólásokat minél hamarabb megtalálja a rendszer. A webes médiumtípusok automatikus meghatározása a rendszer által talált tartalmak későbbi felhasználását és feldolgozását is elősegíti. [1] A weboldalak mellett a fő közösségi oldalak (Twitter, Facebook, Google+) nyilvános posztjait is feldolgozzuk.

3 Véleményárfolyam számítása

A Neticle a megtalált szövegek elemzésével számokban foglalja össze a web véleményét, hangulatát egy adott témával, ha úgy tetszik, kulcsszóval kapcsolatban. Ehhez az egyik legfontosabb mutatónk az úgynevezett webes véleményárfolyam.

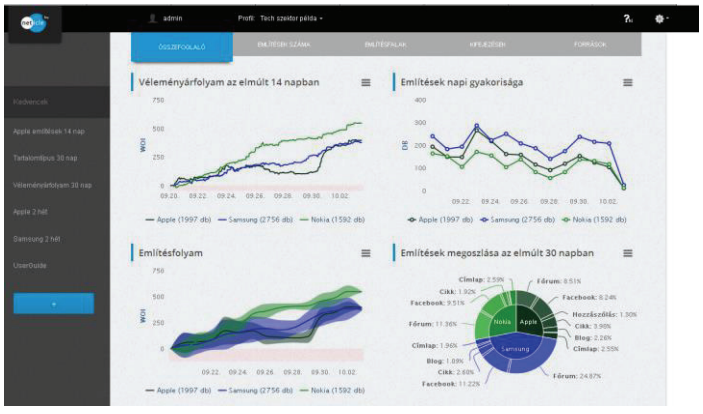
A Neticle véleményárfolyam (**WOI, Web Opinion Index**) egy univerzális mutató, amely összefoglalja a webes tartalmak véleményét egyetlen dinamikusan változó számba. (Ez a mutató tulajdonképpen a tőzsdei részvényárfolyam analógiája, a weben publikálók véleményét, hangulatát tükrözi.)

A pozitív és negatív webes tartalmak alapján számolt index egyértelműen mutatja egy cég/termék/téma webes megítélését illetve annak változását: a tartalmak polaritását számszerűsíti a rendszer, polaritásindexet rendel a szövegekhez. Így ha egy pozitív írás jelenik meg a témában a weben, akkor az árfolyam növekszik, ha pedig valaki negatívan nyilatkozik egy fórumon például a termékről, akkor a véleményárfolyam csökken. Az árfolyam összevethető múltbeli adatokkal és a versenytársak árfolyamaival, vizsgálhatóak a marketingkampányok hatásai is például.

A fejlesztés során az elképzelésünk az volt, hogy a magyar nyelvű webes mondatok véleménypolaritásának (tehát pozitív-negatív voltának) számítógépes meghatározása a megfelelő algoritmussal elérheti az emberi ítélőképesség határát. Azaz közel 82%-ban egyezhet egy ember által elvégzett manuális pozitív-semleges-negatív értékeléssel Az eddigi tesztek alapján különböző témakörökben 75-85%-os pontosságot sikerült elérni az automatikus polaritásmérő algoritmusunkkal. [2]

4 Eredmények

Az automatikus polaritásméréssel megvalósított közel valós idejű webes véleményárfolyam számítással egyszerűen követhető a Neticle-ben, hogy mit gondol a web a cégünkről, termékünkről.



1. ábra. A Neticle rendszer egy képernyője.

Hivatkozások

1. Csikós, Z., Szekeres, P.: Friss tartalmak gyors megtalálása a magyar weben. Budapesti Corvinus Egyetem Tudományos Diákköri dolgozat (2012)
2. Szekeres, P.: Polaritásmérés magyar nyelvű webes szövegekben. Budapesti Corvinus Egyetem, Budapest (2012)

Magyar nyelvű hasonló tartalmú orvosi leletek azonosítása

Wieszner Vilmos[1], Farkas Richárd[1], Csizmadia Sándor[2], Palkó András[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{wieszner,rfarkas}@inf.u-szeged.hu
[2]Szegedi Tudományegyetem, Radiológiai Klinika
6725 Szeged, Semmelweis u. 6

A radiológiai praxist támogató számítógépes nyelvészeti alkalmazás lehet egy az éppen gépelt dokumentumhoz hasonló leletek megtalálása. Demónkban egy ilyen, általunk készített rendszert mutatunk be. A kitűzött cél részszövegekre a leghasonlóbb leletek megtalálása, valamint a megtalált leletek rangsorolása.

A rangsor helyességének értékelésére rendelkezésünkre áll egy 200 elemű adatbázis, ahol az orvosok által kézzel lettek megadva a dokumentumokhoz leghasonlóbb találatok. Továbbá fontos kritérium volt, hogy a találatok között nem szerepelnek olyan leletek, melyek diagnózisa negatív.

Két lelet közötti hasonlóság kiszámítását nagyban befolyásolják az emberi tényezők, ilyenek a helyesírás, valamint a leletekből hiányzó információ vagy eltérő írásmód. A helytelen helyesírás, mint a kisbetűk, illetve gépelési hibák a figyelmetlenségből fakadnak, míg a hiányos információ és a különböző írásmód a páciens aktuális orvosán múlik. Ebből következik, hogy a rendszernek képesnek kell lennie az ilyen jellegű hibákból keletkező eltérések figyelmen kívül hagyására. A leletek tárolását a Solr rendszer [2] segítségével végeztük, ami lehetővé teszi a valós időben történő komplex kereséseket még rendkívül terjedelmes dokumentumhalmaz esetében is.

A figyelmetlenségből eredő hibák javítása könnyen megoldható [1], de a leletek közötti orvosok stílusának eltérései több kihívást rejtenek. Ha az orvos már tudja a beteg egy lehetséges diagnózisát, akkor legtöbbször nem írja le azt egy másik leletbe, valamint a leletekben eltérő lehet a rövidítések értelmezése, még ugyanazon orvos által írt leletek esetében is. A rövidítés értelmezése a kontextustól is függ, ilyen például a *CA* jelölés, ami a szövegkörnyezettől függően jelenthet szívrohamot vagy rosszindulatú daganatot a hámszöveten. A találatokat továbbá befolyásolják a tünetek, illetve a már diagnosztizált betegségek fizikai helye, valamint a mérete. A *két milliméteres csomó* az agyban, illetve a bélrendszerben teljesen más következményekkel járhat, így ezeket más méretkategóriába soroljuk, amit a hely és a nagyság határoz meg. A kór és a tünet megnevezése is változhat, leletenként, részben az orvos szóhasználatától, részben a lehetséges szinonimáktól függően. Az *infarktus* előfordulhat strokeként, vagy a bekövetkezés helyétől függően agyvérzésként is. Ennek megoldása a kontextustól függő szinonima-, illetve rövidítésfeloldás, azaz a szövegkörnyezetből kinyert részinformációkból adjuk meg a rövidítés legvalószínűbb jelentését.

Előfeldolgozó lépések után – mint például a szótövezés és írásjelek eltávolítása – a dokumentumok reprezentációját az unigramok és a kinyert numerikus

tulajdonság-érték párok alkotják. A leghasonlóbb találatokat a leleteken tf-idf normalizálással számítjuk ki, azzal a módosítással, hogy meghatározott szavak, mint a szervek megnevezése, a méretkategóriák és bizonyos előre megadott tünetek nagyobb súllyal legyenek figyelembe véve.

Köszönetnyilvánítás

Jelen kutatást a Telemedicina fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Siklósi B., Novák A., Prószéky G.: Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Magyar Számítógépes Nyelvészeti Konferencia (2013)
2. Smiley, D., Pugh, E., Parisa, K., Mitchell, M.: Apache Solr 4 Enterprise Search Server (2014)

IX. Angol nyelvű absztraktok

Natural Language Processing for Mixed Speech-Music Playlist Generation

Ivett Benyeda[1], Mátyás Jani[2], Gergely Lukács[2]

[1] Research Institute for Linguistics, Hungarian Academy of Sciences
33, Benczúr str., Budapest, HU-1068, Hungary
benyeda.ivett@nytud.mta.hu
[2] PPCU Faculty of Information Technology and Bionics
50/A, Práter str., Budapest, HU-1083, Hungary
{jani.matyas, lukacs}@itk.ppke.hu

Abstract

Music listening habits are changing with the spread of online media consumption and the usage of smartphones. Large online music collections have become available and there is a need for selecting and ordering pieces of music automatically, for a customised listening experience. This process, the playlist generation, has gained much research attention recently and got implemented recently in popular music streaming services. The mainstream focuses on the acoustics of the playlist generation. Some current studies have revealed that natural language processing can also improve the results, especially in the mood detection of the songs. These approaches focus on music only playlists.

Mixed speech-music playlists are different from those in the approach that they contain audio recordings with speech (interviews, actual news, etc.) alongside with musical pieces. Such playlists allow new, innovative applications, through which users can listen to music matching their tastes, and they are also connected with the external world and actual events. The first approaches on mixed speech-music playlists focused on the acoustics of the audio clips.

In this paper preliminary experiments are presented towards the generation of mixed speech-music playlists with the help of language technology, an earlier untouched area. In our work, first the relevant connecting points between recordings containing speech and music pieces were examined with the help of professional radio editors. This revealed that the most important connecting points are (1) the mood of the parts and in some cases, especially in the case of feasts (2) the matching of topics.

The most straightforward natural language processing approaches for both parts are to use special mood and feast lexicons. Experiments were conducted based on English language radio podcasts and on their transcripts. A major challenge is that automatic speech recognition (ASR) technologies are required to produce the transcripts. ASR can be used either to recognise the whole speech, this is the so called spoken term detection, or only to recognise some selected keywords, the so called keyword search. Our experiments on a limited dataset using an ASR system suggest that the limited quality achievable with ASR does not affect significantly the quality of the mood detection.

The Reliability of Statistics in Linguistics Notes to a Dictionary Extension

Mátyás Naszódi

MorphoLogic
1122 Budapest, Ráth György utca 36.
naszodim@morphologic.hu

Abstract: Nowadays statistical tools are often used tool in linguistics, but the reliability of these methods is rarely examined. In natural language processing, statistical methods have their boundaries, and one should pay more attention to them. I try to show, when and how can we estimate its boundaries.

1 Uncertainty in languages

Due to the natural features of the languages, there are many types of uncertainty. To decide whether a word form is correct or not is sometimes questionable. The syntax of a language depends on its creator, and even linguists are unconvinced about the correctness of certain sentences. It is impossible to find the only right translation; there are, however, bad, good and better translations of a text. Probabilistic models can describe the problems in all cases.

In spite of the uncertainty, a user would hate a spell checker that marks words with "perhaps", "sure", or percentage of correctness; or would hate a translation tool offering hundreds of possible solutions.

2 Characteristics of linguistic statistics – Zipf Law

If a linguistic phenomenon has more then thousand distinguishable classes, the distribution of the phenomena by the row of classes show similar characters. It is described by the Zipf Law. It tells that the probability of a class and the order in distribution row are in correlation:

$$f(n) \approx C/n^s \quad \textbf{(1)}$$

where n is the ordinal in the probability order of the n^{th} class, C is a constant for normalizing the equation, s is an exponent that is a little bit less than 1 (1).

For low n, that is for classes of high probability, the estimation is far from correct. For classes of low probability, the estimations are biased by measuring errors. In the middle part of the row however, the Zipf Law works well. In linguistic cases the exponent s is as nearer to 1 as larger the number of classes.

3 Mathematics of quality in linguistic works

While collecting or processing a linguistic database (creating a dictionary), the time (of the work) for the collection of N items might be in linear correspondence with the number of items, if the work is homogenous in the set of items.

$$T(N) = \sum_i e_i = N*e \quad (2)$$

If you take it in account that rare items need larger corpora to find them in it, and they need more time to code them, the equation should be changed:

$$T(N) = \sum_i e_i = C*\sum_i i^s = C* N^{1+s}/(1+s) > O(n^2) \quad (3)$$

The quality of coding of individual items gets worse by the rarity of the items. For gaining a quality level, the necessary time is fast growing with the requirement of quality:

$$T(q) > O(1/q^2) \quad (4)$$

where $1-q$ is the covering, that can be a measure of quality.

If the time for preparing an item is limited, the quality of the work gets worse by the number of items. In that case, the (dictionary) work loses required quality in a well-defined number of items. That is why a dictionary may reach its optimal size, and machine translation based on memory or statistics reaches quickly its maximal quality level.

Quality barrier may be broken by independent evaluations and reduction of the number of classes. (It may decrease the constant s of the Zipf equation). An example for the first one is when parallel coding is used for better quality. Here are some examples for the second one:

1. User dictionaries (spelling checkers, dictionaries)
2. Thematic terminologies (spelling checkers, dictionaries, translators)
3. Morpheme-based statistics instead of wordform-based one for translations

I tried to estimate the value s in translations of the same text to several languages. The results cause some surprise because of the following reasons:

1. The corpora are too small.
2. The measured numbers depend not only on the languages, but on the novel and on the translation as well.
3. Coding errors also biased the data.

Despite of that, data show that in case of languages where the number of word forms is large, the probability of word-forms is nearer to the reciprocal value than in languages with poor morphologies.

Automatic Conversion of Constituency Trees into Dependency Trees or Manual Annotation?

Katalin Ilona Simkó[1], Veronika Vincze[1,2], Zsolt Szántó[1], Richárd Farkas[1]

[1]University of Szeged, Department of Informatics
Szeged, Árpád tér 2.
kata.simko@gmail.com, szantozs@inf.u-szeged.hu, rfarkas@inf.u-szeged.hu
[2]MTA-SZTE Research Group on Artificial Intelligence
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Nowadays, two popular approaches to data-driven syntactic parsing are based on constituency grammar on the one hand and dependency grammar on the other hand. Hungarian is one of those rare examples where there exist manual annotations for both constituency and dependency syntax on the same bunch of texts, the Szeged (Dependency) Treebank, which makes it possible to evaluate the quality of a rule-based automatic conversion from constituency to dependency trees, to compare the two sets of manual annotations and also the output of constituency and dependency parsers trained on converted and gold standard dependency trees.

We investigate the effect of automatic conversions related to the two parsing paradigms as well. It is well known that for English, the automatic conversion of a constituency parser's output to dependency format can achieve competitive unlabeled attachment scores (ULA) to a dependency parser's output trained on automatically converted data. One of the possible explanations for this is that English is a configurational language, hence constituency parsers have advantages over dependency parsers here. We check whether this hypothesis holds for Hungarian too, which is the prototype of free word order languages.

In this paper, we compare three pairs of dependency analyses in order to evaluate the usefulness of converted trees. First, we examine the errors of the conversion itself by comparing the converted dependency trees with the manually annotated gold standard ones. Second, we argue for the importance of training parsers on gold standard trees by looking at the typical differences between the outputs of dependency parsers trained on converted (silver standard) trees, parsers trained on gold standard trees and the manual annotation itself. Third, we demonstrate that similar to English, training on a constituency treebank and converting the results to dependency format can achieve similar results in terms of ULA to the dependency parser trained on the automatically converted treebank, but the typical errors they make differ in both cases.

We present the details of the results achieved by different parsing methods as well as a linguistic analysis and categorization of the types of errors they made. For instance, analysing multiword names seems to be easier for the constituency parser, while the dependency parser is better at finding the arguments of verbs.

SzegedKoref: A Manually Annotated Coreference Corpus of Hungarian

Veronika Vincze[1,2], Klára Hegedűs[3], Richárd Farkas[1]

[1]University of Szeged, Department of Informatics
Szeged Árpád tér 2., e-mail: {vinczev,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Research Group on Artificial Intelligence
[3]University of Szeged, Institute of Psychology
e-mail: klarahegedus92@gmail.com

Here we introduce the SzegedKoref corpus, in which coreference relations are manually annotated. For annotation, we selected the texts of Szeged Treebank, the biggest treebank of Hungarian with manual annotation at several linguistic layers.

We present the annotated texts, we describe the annotated categories of anaphoric relations, and we offer several examples of each annotated category. Currently, the corpus contains 309 sentences and 9,782 tokens from the newspaper domain and 3,712 sentences and 45,981 tokens from the student essay subcorpus. Altogether, there are 4021 sentences and 55,763 tokens in the current version of the corpus, however, the annotation process is still going on, and the amount of annotated texts is continuously growing.

In Hungarian, zero pronouns also mean a challenge to coreference resolution systems. We automatically inserted zero pronouns into the text before the manual annotation process, so they are also annotated in the data.

There are 2191 anaphoric chains in the student essay subcorpus and 265 in the newspaper domain, adding up to 2456 anaphoric chains altogether. The most frequent types of anaphor are pronominal anaphors and repetition, indicating that automatic coreference resolution systems should pay extra attention to these categories, together with zero pronouns.

Due to its size, the corpus can be exploited in training and testing machine learning based coreference resolution systems, which we would like to implement in the near future.

Morphological and Syntactic Annotation of Hungarian Webtext

Veronika Vincze[1,2], Viktor Varga[1], Petra Anna Papp[1],
Katalin Ilona Simkó[1], János Zsibrita[1], Richárd Farkas[1]

[1]University of Szeged, Department of Informatics
Szeged, Árpád tér 2.
{vinczev,zsibrita,rfarkas}@inf.u-szeged.hu,
{varga.viktor.1991,papp.petra.anna,kata.simko}@gmail.com
[2]MTA-SZTE Research Group on Artificial Intelligence
Szeged, Tisza Lajos körút 103.

For a while now, internet communication has been used as a source of data for research. Texts on the web trying to mimic oral communication include many abbreviations and errors that make their linguistic processing more difficult. Our goal was to create a corpus of texts from the web and manually annotate it for morphology and syntax in order to make it useful for the development of future natural language processing applications for this domain.

Our corpus is made up of public Facebook comments (1208 sentences, 8615 tokens) and questions and answers from gyakorikerdesek.hu (728 sentences, 9702 tokens). Most posts are about users' hobbies, personal interests and lifestyle.

First, we manually segmented the sentences and tokenised the text, then, using one of the modules of magyarlanc, we built a corpus, structurally similar to the Szeged Korpusz, in which the annotators manually assigned the contextually correct morphological code to each word. Similar to Szeged Treebank and Szeged Dependency Treebank, we also created manual constituent and dependency syntax analysis for each sentence. We mainly followed the principles used in the development of our two previous, bigger treebanks, but some modifications were unavoidable given the special form of this text. The corpus is also annotated for semantic and discourse level uncertainty markers and we plan to annotate named entities in it as well.

This first Hungarian, manually annotated web corpus will be used as a test database in developing a morphological and syntactic parser, optimalised for the analysis of texts from the web. The corpus is currently too small to train statistical parsers, however, our goal was to create a benchmark database. We believe that as web texts are so varied both in topic and genre, the application of supervised machine learning techniques would not be a suitable solution, instead, we plan to use domain adaptation methods.

Névmutató