Bizonytalanságot jelölő kifejezések azonosítása magyar nyelvű szövegekben

Vincze Veronika[1,2]

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

Kivonat A bizonytalanságot jelölő kifejezések automatikus azonosítása napjaink egyik intenzíven vizsgált területe a számítógépes nyelvészeti kutatásokban. Ebben a cikkben bemutatjuk magyar nyelvű annotált korpuszunkat, melyben kézzel bejelöltük a nyelvi bizonytalanság különféle fajtáit jelző nyelvi elemeket. A korpusz arra is lehetőséget kínál, hogy beszámoljunk az első, magyar nyelvű bizonytalanságazonosító gépi tanuló rendszer eredményeiről.
Kulcsszavak: információkinyerés, szemantika, korpusz

1. Bevezetés

A bizonytalanságot jelölő kifejezések automatikus azonosítása napjaink számítógépes nyelvészeti kutatásának egyik fontos problémaköre [1]. A feladat fontossága abban rejlik, hogy a különféle számítógépes nyelvészeti alkalmazásokban lényegi szerep jut a tényszerű és a bizonytalan, illetve tagadott információ megkülönböztetésének, hiszen például információkinyerés és szemantikus keresés esetében a felhasználónak többnyire tényszerű információra van szüksége, így alkalmazástól függően a rendszer vagy kiszűri a bizonytalan / tagadott szövegrészeket, vagy pedig a tényektől elkülönítve adja őket vissza a felhasználónak. A problémára eddig elsődlegesen angol nyelvű szövegeken nyújtottak megoldásokat [1, 2]. Ebben a cikkben bemutatjuk kézzel annotált, magyar nyelvű bizonytalansági korpuszunkat, és beszámolunk az első eredményekről a nyelvi bizonytalanságot jelölő elemek automatikus felismeréséről magyar nyelvű szövegekben.

2. A bizonytalanság típusai

A nyelvi bizonytalanságot hagyományosan a mondat szemantikájához szokták kötni, azonban vannak olyan bizonytalanságot jelző nyelvi elemek is, melyek ezzel szemben a mondat (közlés) kontextusában – diskurzusbeli tényezőknek

köszönhetően – válnak többértelművé. Például a *Lehet, hogy esik az eső* mondat alapján nem tudjuk eldönteni, hogy esik-e az eső (szemantikai bizonytalanság), viszont a *Számos kutató szerint az MSZNY a legjobb magyar konferencia* mondatból az nem derül ki, hogy pontosan kinek (illetve hány kutatónak) a véleményéről esik szó, így a közlés forrása marad bizonytalan (diskurzusszintű bizonytalanság). Ebben a cikkben követjük a [2], illetve [3] cikkekben felvázolt osztályozást a bizonytalanság különböző fajtáira nézve, illetve a magyar nyelvre alkalmazzuk azt, annotációs elveinket a fentiek alapján kialakítva.

A szemantikai bizonytalanságnak több osztálya is létezik. Egy propozíció episztemikusan bizonytalannak számít, ha a világtudásunk alapján nem tudjuk eldönteni ebben a pillanatban, hogy igaz-e vagy hamis. Ugyanez igaz a hipotetikus bizonytalanságra is, ide sorolhatók a feltételes mondatok, illetve a vizsgálati bizonytalanság – utóbbi különösen tudományos cikkekben gyakori, hiszen a kutatási kérdést gyakran a vizsgálati bizonytalanság nyelvi eszközeivel fogalmazzák meg a szerzők. A modalitás nem episztemikus típusai (például doxasztikus bizonytalanság, mely a hiedelmekkel függ össze, illetve a dinamikus modalitás különböző fajtái, melyek többek között a szükségszerűséghez kapcsolódnak) szintén ebbe a nagyob csoportba sorolhatók.

A diskurzusszintű bizonytalanságnak három osztályát különböztethetjük meg [3]. Először, a *weasel* kifejezésekhez nem tudunk egyértelműen forrást rendelni (azaz nem tudjuk, kihez köthető az adott információ), más esetben pedig hiányzik a közlésből egy fontos és releváns információrészlet, amely azonban az adott helyzetben szükséges lenne. Másodszor, a *hedge* szavak homályossá teszik bizonyos mennyiségek vagy minőségek pontos jelentését. Harmadszor, a *peacock* kifejezések bizonyítatlan (vagy bizonyíthatatlan) értékeléseket, minősítéseket vagy túlzásokat fejeznek ki.

A bizonytalanságot jelző kulcsszavakra itt mutatunk néhány példát:

Episztemikus: **Lehet**, hogy esik.

Dinamikus: Mennem **kell**.

Doxasztikus: Azt **hiszi**, hogy a Föld lapos.

Vizsgálat: A felvétel manipuláltságáról **vizsgálatot folytattak**.

Feltételes: **Ha** esik, itthon maradunk.

Weasel: **Egyesek** szerint inkább megszállást kellene mondani.

Hedge: A belga lakosság **kb.** 10%-a él Brüsszelben.

Peacock: Apafi négy évet **keserves** tatár fogságban töltött.

Az angolra alkalmazott osztályozást változtatások nélkül vettük át a magyarra, azonban a magyar nyelv sajátságainak megfelelően az annotációs elveket némileg átalakítottuk. Például az episztemikus bizonytalanságot a magyarban igen gyakran a *-hat/-het* képző fejezi ki, míg az angolban ez segédigék (pl. *can*, *may*) használatával történik. Ezekben az esetekben az angol korpuszban a segédigét jelöltük meg mint bizonytalanságot jelző elemet, a magyarban azonban a teljes szóalakot, mivel a képző külön címkézésére nem volt lehetőségünk morfémákra bontott nyelvi adatbázisok híján.

A [2] és [3] munkákhoz hasonlóan e cikkben is a diskurzusszintű bizonytalanság mindhárom fajtájával, illetve a szemantikus bizonytalanság négy fajtájával (episztemikus, vizsgálati, feltételes és doxasztikus) foglalkozunk.

3. Kapcsolódó irodalom

A bizonytalanságot jelző nyelvi elemek vizsgálata napjaink számítógépes nyelvészeti kutatásainak egyik népszerű témája. Ezt jelzi többek között a CoNLL-2010 verseny megrendezése, melynek témája a nyelvi bizonytalanság azonosítása volt biológiai cikkekben és Wikipedia-szócikkekben, angol nyelven [1], illetve a Computational Linguistics folyóirat tematikus különszáma (Vol. 38, No. 2), melyet a bizonytalanság és tagadás automatikus azonosításának szenteltek. Az eddigi vizsgálatok túlnyomórészt az angol nyelv köré csoportosulnak, és elsődlegesen újsághíreket, biológiai publikációkat vagy orvosi dokumentumokat, illetve Wikipedia-szócikkeket elemeznek (vö. [2, 4, 5]).

A felügyelt gépi tanulási eljárások megkövetelik egy annotált korpusz létét. Noha számos, bizonytalanságra épített korpusz elérhető a világban (a teljesség igénye nélkül megemlítve néhányat: BioScope [6], Genia [4], FactBank [5], a CoNLL-2010 verseny korpuszai [1]), ezek azonban angol nyelvűek. A magyar nyelvű kutatások egyik fontos előkészületi lépésének bizonyult tehát egy kézzel annotált, magyar nyelvű adatbázis elkészítése, melyben nyelvész szakértők bejelölték a bizonytalanságot jelző nyelvi elemeket.

A bizonytalanságot azonosító rendszerek eleinte szakértői szabályok alapján működtek (pl. [7,8]), az utóbbi időben azonban gépi tanulásra épülnek, többnyire felügyelt tanulási módszereket hasznosítva (pl. [9,10] és a CoNLL-2010 versenyen részt vevő rendszerek [1]). A legutóbbi tendenciákkal összhangban e cikkben bemutatunk egy felügyelt tanulásra épülő modellt, mely gazdag jellemzőtérrel rendelkezik: lexikai, morfológiai, szintaktikai és szemantikai jegyekre egyaránt épít, továbbá kontextuális jellemzőket is figyelembe vesz.

4. A korpusz

A hUnCertainty korpusz magyar nyelvű Wikipédia-szócikkekből áll, összesen 1081 bekezdést, 9722 mondatot és 180 000 tokent tartalmaz. A szövegek kiválogatása során összegyűjtöttük a legtipikusabb angol nyelvű bizonytalan kulcsszavak magyar megfelelőit, majd az olyan bekezdések kerültek bele a korpuszba, amelyek legalább egyet tartalmaztak e kulcsszavak közül. Mindemellett olyan bekezdések is a korpusz részét képezik, amelyek nem tartalmazták ezen kulcsszavak egyikét sem, így törekedve a korpuszbeli adatok kiegyensúlyozottságára.

A korpuszban kézzel jelöltük meg a bizonytalanságért felelős nyelvi elemek (kulcsszavak) több fajtáját. A korpuszban előforduló kulcsszavak arányát az 1. táblázat mutatja.

Mint látható, a korpuszban a diskurzusszintű bizonytalanság kulcsszavai dominálnak. Ez összhangban van a korábban angol nyelvű Wikipedia-szócikkeken

elért eredményekkel [3], így valószínűleg a kulcsszavak ilyen eloszlása a Wikipédia-szövegek sajátja nyelvtől függetlenül.

1. táblázat. Bizonytalanságot jelző kulcsszavak.

Kulcsszó típusa	#	%	Eltérő kulcsszavak száma
Hedge	2100	35,12	439
Weasel	2150	35,95	598
Peacock	788	13,18	400
Diskurzusszintű összesen	5038	84,25	1437
Episztemikus	441	7,37	184
Doxasztikus	316	5,28	67
Feltételes	154	2,58	46
Vizsgálat	31	0,52	22
Szemantikus összesen	942	15,75	319
Összesen	5980	100	1756

Ha a mondatok szintjén vizsgáljuk a bizonytalanságot, azt találjuk, hogy a korpuszban 3710 (39,22%) bizonytalan mondat szerepel (azaz legalább egy kulcsszót tartalmaznak). Ezek közül 3344 mondat tartalmaz diskurzusszintű bizonytalanságot jelölő kulcsszót (35,35%), és 746 pedig szemantikus bizonytalanságra utaló kulcsszót (7,89%).

A 2. táblázat foglalja össze a leggyakoribb magyar episztemikus és doxasztikus kulcsszavakat. Az első tíz kulcsszó adja az összes előfordulás 42 és 79%-át ezen kulcsszavak esetében. Mivel a feltételes és a vizsgálati kulcsszavak nem mutatnak nagy változatosságot a korpuszban, csak a legalább háromszor előforduló elemeket soroljuk fel itt: a *vizsgál* és *tanulmányoz* szavak adják a vizsgálati kulcsszavak 29%-át, illetve a *ha*, *akkor* és *amennyiben* szavak a feltételes kulcsszavak 68%-át.

2. táblázat. A leggyakoribb episztemikus és doxasztikus kulcsszavak.

Episztemikus	#	%	Doxasztikus	#	%
valószínűleg	79	17,87	szerint	151	47,63
talán	28	6,33	tart	25	7,89
feltehetőleg	15	3,39	tekint	19	5,99
állítólag	14	3,17	állít	18	5,68
feltehető	11	2,49	vél	10	3,15
lehet	10	2,26	tulajdonít	7	2,21
lehetséges	10	2,26	gondol	6	1,89
feltételez	7	1,58	tesz	5	1,58
tekinthető	7	1,58	hisz	4	1,26
lehetőség	6	1,36	vall	4	1,26

A 3. táblázatban találhatók meg a leggyakoribb, diskurzusszintű bizonytalanságot jelölő kulcsszavak. A tíz leggyakoribb kulcsszó az esetek 40, 31 és 26%-át fedi le a weasel, hedge és peacock előfordulásoknak.

3. táblázat. A leggyakoribb diskurzusszintű kulcsszavak.

Weasel	#	%	Hedge	#	%	Peacock	#	%
számos	150	8,60	általában	127	6,18	fontos	50	6,36
egyes	134	7,68	gyakran	119	5,79	jelentős	39	4,96
egyik	118	6,76	később	99	4,82	ismert	25	3,18
más	100	5,73	nagyon	50	2,43	híres	23	2,93
néhány	66	3,78	főleg	47	2,29	nagy	17	2,16
különböző	34	1,95	nagy	46	2,24	kiemelkedő	15	1,91
egyéb	29	1,66	igen	43	2,09	komoly	11	1,40
sok	27	1,55	néhány	40	1,95	erős	10	1,27
bizonyos	22	1,26	főként	37	1,80	kiváló	9	1,15
többek között	19	1,09	mintegy	36	1,75	egyszerű	9	1,15

Néhány kulcsszó több bizonytalansági osztályt is jelölhet, ugyanakkor a kulcsszavak nem minden előfordulása jelöl ténylegesen bizonytalanságot az adott kontextusban. Az első esetre példa a *nagy* szó, amely hedge és peacock kulcsszó is lehet attól függően, hogy fizikai vagy minőségi nagyságra utal-e. A második esetet illusztrálja az *igen* szó: határozószóként előfordulhat hedge-ként, mondatszóként azonban nem jelöl bizonytalanságot.

Minthogy a hUnCertainty korpusz annotációs elvei angol korpuszok építése során használt elveken alapulnak [2,3], az angol és magyar korpuszokból származó adatok összevethetők egymással. Például a szemantikai és diskurzusszintű bizonytalanság kulcsszavai hasonló arányban fordulnak elő mindkét nyelvű Wikipédia-szövegekben. A kulcsszavak szintjén pedig megfigyelhetjük, hogy azonos jelentésű szavak szerepelnek a leggyakoribb kulcsszavak között, például *valószínű*, *lehetséges*, *hisz*. E tények arra utalnak, hogy a [2] és [3] munkákban bemutatott osztályozás több nyelvre is alkalmazható.

5. A bizonytalanság automatikus azonosítása

Annak érdekében, hogy automatikus úton azonosítsuk a bizonytalanságot jelölő kulcsszavakat, kifejlesztettünk egy gépi tanuláson alapuló módszert, melyet a következőkben ismertetünk részletesen. Méréseinkhez a hUnCertainty korpuszt vettük alapul, melyet a magyarlanc elemzőt [11] felhasználva morfológiailag és szintaktikailag elemeztünk.

5.1. Gépi tanulási módszerek

Korábbi angol nyelvű kísérleteink alapján a szekvenciajelölés bizonyult a legeredményesebbnek a bizonytalanság automatikus azonosításában [2], így a magyar nyelvű anyagon végzett méréseinket is feltételes véletlen mezőkön (CRF)

[12] alapuló módszerrel kiviteleztük. Kísérleteink kiindulópontjaként egy magyar nyelvre implementált, MALLET alapú névelem-felismerő rendszer [13] szolgált, a felhasznált jellemzőket természetesen a bizonytalanságazonosítási feladat sajátságaira szabva, melyeket az alábbiakban ismertetünk:

- **Felszíni jellemzők:** a szó írásmódjával kapcsolatos jellemzők (tartalmaze írásjelet, számot, kis/nagybetűket, szóhossz, mássalhangzó bi- és trigramok...)
- **Lexikai jellemzők:** a hasonló elvek alapján annotált, rendelkezésre álló angol nyelvű korpuszokból [2] minden bizonytalansági típushoz kigyűjtöttük a leggyakoribb kulcsszavakat, és ezeket magyarítva listákba rendeztük őket. A listákat bináris jellemzőként használtuk fel: ha az adott szó lemmája előfordult valamelyik listában, akkor igaz értéket kapott az adott jellemzőre nézve.
- **Morfológiai jellemzők:** minden szó esetében felvettük annak fő szófaját, illetve lemmáját a jellemzők közé. Igék esetében továbbá megvizsgáltuk, hogy ható igéről van-e szó, feltételes módú-e az ige, illetve T/1. vagy T/3. alakban fordul-e elő. Főnevek esetében felvettük jellemzőként, hogy egyes vagy többes számban állnak-e. Külön jelöltük a névmások esetében azt is, ha határozatlan névmásról volt szó, illetve mellékneveknél a fokot is felvettük a jellemzők közé.
- **Szintaktikai jellemzők:** minden szóhoz felvettük annak szintaktikai címkéjét, továbbá főnevek esetében megvizsgáltuk, hogy rendelkezik-e névelővel, illetve igék esetében felvettük, hogy van-e alanya.
- **Szemantikai/pragmatikai jellemzők:** egy általunk összeállított, beszédaktusokat tartalmazó lista alapján megvizsgáltuk, hogy az adott szó beszédaktust jelölő ige-e. Mindemellett a kulcsszavakhoz hasonlóan, angol nyelvű, pozitív és negatív jelentéstartalmú szavakat tartalmazó listákat [14] is magyarítottunk, és megnéztük, hogy a szó lemmája szerepel-e az adott listában.

Az adott szó környezeti jellemzőjeként felvettük a tőle egy vagy két szó távolságra levő szavak szófaji kódját és szintaktikai címkéjét is.

A fentiekben leírt jellemzőkészlet alapján tízszeres keresztvalidációt használva hajtottuk végre méréseinket a hUnCertainty korpuszon. Mivel csak a tokenek körülbelül 3%-a funkcionál kulcsszóként a korpuszban, így szükségesnek láttuk a tanító adatbázis szűrését: a kulsszót nem tartalmazó mondatoknak csak a fele került bele a tanító halmazba. Továbbá mivel a vizsgálati bizonytalanság kulcszavai összesen 31 előfordulást mutattak, ezt az ritka osztályt nem vettük figyelembe a rendszerünk létrehozásánál, így a kiértékelésben sem szerepel.

5.2. Baseline mérések

Baseline mérésként egyszerű szótárillesztést használtunk. A lexikai jellemzők között említett listákat jelöltük rá a korpuszra: amennyiben a szó lemmája megegyezett az adott lista egyik elemével, a bizonytalanság adott típusának címkéztük fel.

6. Eredmények

A 4. táblázat mutatja a baseline, valamint a gépi tanuló kísérletek eredményeit. A kiértékelés során a pontosság, fedés és F-mérték metrikákat alkalmaztuk.

4. táblázat. Eredmények.

	Szótárillesztés			**Gépi tanuló**			**Különbség**
Típus	Pontosság	Fedés	F-mérték	Pontosság	Fedés	F-mérték	
Weasel	26,03	38,50	31,06	59,26	34,74	43,80	+12,74
Hedge	55,86	29,92	38,97	64,59	50,02	56,38	+17,41
Peacock	23,29	30,63	26,46	37,85	13,80	20,22	-6,38
Episztemikus	49,57	37,34	42,59	63,95	36,03	46,09	+3,5
Doxasztikus	25,24	65,20	36,40	54,31	33,54	41,47	+5,07
Feltételes	29,66	67,74	41,26	47,12	31,61	37,84	-3,42

A táblázatból jól látszik, hogy a gépi tanuló megközelítés eredményei két osztály kivételével minden esetben meghaladták a baseline szótárillesztés által elért eredményeket. Ez elsődlegesen a pontosság javulásának köszönhető, mely kivétel nélkül minden osztályra nézve jóval magasabb lett a szekvenciajelölő megközelítés esetén. Ezzel szemben a fedési értékek nagyobb változatosságot mutatnak: míg a hedge osztály esetében ez is nőtt, a weasel és episztemikus kulcsszavaknál nem változott jelentős mértékben, addig a peacock, doxasztikus és feltételes kulcsszavaknál drasztikus visszaesést figyelhetünk meg. Vélhetően a gyenge fedésre vezethető vissza az is, hogy a peacock és feltételes kulcsszavaknál a szótárjelölő megközelítés magasabb F-mértéket ért el, mint a gépi tanuló algoritmus.

7. Az eredmények megvitatása

Elért eredményeink azt igazolják, hogy a magyar nyelvben is lehetséges a bizonytalanságot jelölő kifejezések automatikus azonosítása szekvenciajelölő megközelítéssel. A szótárillesztés során a legjobb eredményeket az episztemikus, feltételes és hedge kulcsszavakon értük el, míg a szekvenciajelöléssel a hedge, episztemikus és weasel osztályokon születtek a legjobb eredmények. Mindezek alapján a hedge és episztemikus osztályok tűnnek a legkönnyebben felismerhetőknek. Az eredmények arra is utalnak, hogy azon (szemantikai) osztályok esetében, ahol kicsi volt a különbség a szótárillesztés és gépi tanulás eredményei között, az adott bizonytalanságtípus nyelvi jelölésmódja elsődlegesen lexikális (és kevésbé többértelmű) eszközökkel valósul meg. Ugyanakkor a diskurzusszintű bizonytalanság kulcsszavainak felismerésében nagyobb szerepet játszik a gépi tanulás, ami annak köszönhető, hogy esetükben igen fontos szerepe van a kontextusnak (diskurzusnak), így egy szekvenciajelölő algoritmus sikeresebben tudja megoldani a feladatot.

Amennyiben eredményeinket összevetjük a korábban angol nyelvű Wikipedia-szócikkeken elért, szemantikai bizonytalanságot azonosító rendszer által elértekkel [2], azt láthatjuk, hogy angol nyelven könnyebbnek tűnik a feladat: 0,6 és 0,8 közötti F-mértékekről számol be a cikk. Azonban nem szabad figyelmen kívül hagynunk két fontos tényezőt. Egyrészt a két nyelv közti tipológiai különbségeknek köszönhetően az angolban inkább lexikálisan meghatározott a bizonytalanság jelölése, a magyarban pedig inkább morfológiai eszközök valósítják meg ezt: például a ható igéket a magyarban a *-hat/-het* képző jelöli, az angolban pedig a *may*, *might* stb. segédigék. Így a szóalak, illetve lemma jellemzőként való szerepeltetése angolban már viszonylag jó eredményekhez vezethet, magyarban azonban ezek a jellemzők önmagukban (morfológiai jellemzők felvétele nélkül) kevésbé hatékonyak. Másrészt az adatbázis nagysága jelentősen különbözik a két esetben: míg körülbelül 20000 annotált angol mondat állt rendelkezésre, addig a magyarban ez a szám nem érte el a 10000-et. Az annotált adatok mennyiségének fontosságát igazolják az angol nyelvű mérések is: azokban az esetekben, amikor csupán néhány ezer annotált mondat állt rendelkezésre, az elért F-mértékek – doméntől és kulcsszótípustól függően – 0,1-0,8 között mozogtak.

A peacock és a feltételes kulcsszavak esetében a szekvenciajelölő módszer rosszabbul teljesített a szótárjelölő megközelítésnél: mindkét esetben a pontosság nőtt ugyan, de a fedés jelentős visszaesést mutatott. Ez alapján szükségesnek ígérkezik a rendszer felülvizsgálata, továbbá új, speciálisan ezekre az osztályokra kifejlesztett jellemzők definiálása.

A gépi tanuló rendszer kimenetét részletesen is megvizsgáltuk hibaelemzés céljából. Azt találtuk, hogy elsődlegesen a többértelmű kulcsszavak egyértelműsítése jelent problémát. Például a *számos* vagy *sok* szavak lehetnek szövegkörnyezettől függően weasel és hedge kulcsszavak is, vagy a *nagy* lehet peacock és hedge is. Az ehhez hasonló eseteket a rendszer időnként rossz osztályba sorolta. Gyakori hibaforrásnak számítottak azok a kulcsszavak is, amelyek gyakran használatosak nem kulcsszó jelentésben is, mint például a *tart* ige, amely lehet doxasztikus kulcsszó (*vki vmilyennek tart vkit/vmit*), azonban más jelentésben nem kulcsszó (pl. *vki vhol tart vmit*, *vki vhol tart vmiben* stb.). Egy sajátos hibának bizonyult az episztemikus osztálynál a tagadást tartalmazó kulcsszavak fel nem ismerése: a *nem zárható ki*, *nem tudni* stb. alakokat a rendszer nem jelölte meg kulcsszóként.

8. Összegzés

Ebben a cikkben bemutattuk a hUnCertainty korpuszt, amely az első kézzel annotált, magyar nyelvű bizonytalansági korpusz. A korpusz lehetőséget adott arra, hogy beszámoljunk az első eredményekről a nyelvi bizonytalanságot jelölő elemek automatikus felismeréséről magyar nyelvű szövegekben. A szekvenciajelölésen alapuló, gazdag jellemzőtérrel dolgozó megközelítésünk által elért eredményeink bizonyítják, hogy magyar nyelvre is alkalmazható a bizonytalanság nyelvi modellje, illetve a bizonytalanságot jelölő kulcsszavak automatikus azonosítása is megoldható.

A jövőben módszereinket szeretnénk továbbfejleszteni, elsősorban a jobb fedés elérésének irányába, mindemellett más jellegű szövegekben is szeretnénk annotálni, illetve automatikusan azonosítani a bizonytalanságot jelölő kifejezéseket.

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. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task, Uppsala, Sweden, Association for Computational Linguistics (2010) 1–12
2. 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
3. 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
4. Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics **9**(Suppl 10) (2008)
5. Saurí, R., Pustejovsky, J.: FactBank: a corpus annotated with event factuality. Language Resources and Evaluation **43** (2009) 227–268
6. Vincze, V., Szarvas, Gy., Farkas, R., Móra, Gy., Csirik, J.: The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. BMC Bioinformatics **9**(Suppl 11) (2008) S9
7. Light, M., Qiu, X.Y., Srinivasan, P.: The language of bioscience: Facts, speculations, and statements in between. In: Proc. of the HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases. (2004) 17–24
8. Chapman, W.W., Chu, D., Dowling, J.N.: Context: An algorithm for identifying contextual features from clinical text. In: Proceedings of the ACL Workshop on BioNLP 2007. (2007) 81–88
9. Medlock, B., Briscoe, T.: Weakly Supervised Learning for Hedge Classification in Scientific Literature. In: Proceedings of the ACL, Prague, Czech Republic (2007) 992–999
10. Özgür, A., Radev, D.R.: Detecting speculations and their scopes in scientific text. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, Association for Computational Linguistics (2009) 1398–1407
11. 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

12. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML-01, 18th Int. Conf. on Machine Learning, Morgan Kaufmann (2001) 282–289
13. Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: Proceedings of the 9th international conference on Discovery Science. DS'06, Berlin, Heidelberg, Springer-Verlag (2006) 267–278
14. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)