Named Entity Recognition in the Miskolc Legal Corpus

Üveges István

Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
uvegesistvan898@gmail.com

Abstract. In this paper, a brief study will be presented with regard to the issue of Named Entity Recognition (NER) in legal texts. To get an overall picture, we examined closely the output of two existing analysers: the "*magyarlanc"* linguistic processing toolkit [1] and a Named Entity Recognition system developed by the Natural Language Processing Group at the University of Szeged [2]. Firstly, short references are made to named entity recognition projects in the literature considered important in the current framework. Secondly, quantitative analyses of the data will be presented. At the end of the study, some problematic cases and potential solutions will be discussed which will be followed by the discussion of the future research.

Keywords: Named Entity Recognition, Hungarian legal texts, magyarlanc, Szeged NER

1 Introduction

The process of finding named entities in a text and classifying them to a semantic type is called *Named Entity Recognition (NER).* The task itself was firstly introduced in the early 1990s in computational linguistics and NER is a cornerstone for tools based on Information Extraction (IE) and key issue in many fields of science nowadays.

Here we focus on NER in the legal domain, where the (semi)automatic anonymization of legal documents and the development of more informative and efficient searching tools get more and more attention. Named Entities (NEs) are not just mentions of persons and organizations in the legal domain, but we also have to take into consideration other categories like names of laws and even concepts. In the international literature, we can see many ongoing projects aimed to develop such systems and applications for Anglo-American legal documents ([3], [4], [5] etc.) and in the Hungarian literature as well ([6], [7]).

With the automatic detection and classification of such elements, legal information extraction can be enhanced for lawyers, courts, governmental organizations, or even non-professionals.

2 Data

The examination was carried out on the Miskolc Legal Corpus [8], which was created by a cooperation of lawyers, linguists and IT specialists in order to make the language of law more easily available for NLP studies.

During the creation of the corpus the main goal was to cover the largest segment possible of the Hungarian legal language. It contains six different sources (cf. [8]) of legal texts[1];

- the full text of 5 Hungarian laws (henceforth: Laws)
- randomly selected parts of other legal regulations
- texts of judgements and legal sentences
- explanatory texts (from ministerial arguments and university textbooks)
- legal forums (Forums)
- transcripts (Transcripts).

For our current analysis, the first ~6000 tokens have been chosen from the Laws, Forums and Transcripts sub-corpora.

The Forum part is, as its name may suggest, made of posts, topics and comments of online discussion sites. The transcript part consists of transcripts of courtroom discussions[2] so this section represents the spoken legal language in the corpus. The Laws part has been compiled from full texts of Hungarian laws.

In Table 1, some main properties of the selected texts from the sub-corpora are summarized.

Sub-corpus	Token number	Word count
Forums	6041	4718
Transcripts	6010	4594
Laws	6014	4660

Table 1: Basic information

When selecting texts from the Miskolc Legal Corpus, the main criterion was that they should represent (intuitively) different aspects of legal language use and text type.

3 Methods

Our main goal is to find an explicit evidence that these distinct domains of the legal language may (or may not) require a different treatment from the automatic NE recognizer tools.

To achieve this, a quantitative analysis was carried out on three levels:

[1] From each source, the corpus contains approximately 25.000 sentences, 150.000 sentences in total. The coprus was originally developed in the framework of an OTKA-project: https://sites.google.com/site/otkamiskolc2015/

[2] Recordings and transcripts were made with the consent of all participants of the discussions.

- on the one hand, after a manual annotation (see 3.1) the output of automatic POS-tagging of the *magyarlanc* toolkit, was compared with the output of the Named Entity Recogniser System on the same text,
- on the other hand, in the case of multiword NEs, where *magyarlanc* should tag the affected tokens on the level of dependency grammar with an "NE" label [1], the presence or absence of this specific tag was examined closely.

In the qualitative section of the analysis, the most frequent sources of errors will be examined closely to reveal the domain-specificity of these peculiarities and to provide useful data for increasing the efficiency of future NER-tools in the legal domain.

3.1 Manual annotation

To get comparable results, and data, at the first step, all the examined text was checked by a linguist expert, who annotated all the NEs manually. The annotation followed the *tag-for tagging* principle, but apart from this, it was match with the rules defined during the annotation of the HunNER corpus [9].

The used definition for NE categories was based on the ACE 2006 annotation guideline [10]. However, just the name mentions ("Joe Smith")[3], locations and organizations were kept as an annotated category.

The three basic category searched during the annotation process was *person*, *location* and *organization names*. Besides that, the names of legal *regulations* (e.g. Ptk. – *Civil Code*) proved to be important during the annotation process in this specific domain. Table 2 shows the manually annotated NEs in the examined texts.

3.2 Automatic NER methods

The selected texts were parsed with *magyarlanc* and the NER-tool, after that we checked whether a label was correctly assigned to a token, or not.

The expected label was PROPN from the *magyarlanc* and an I-TYPE tag from the NER-tool (where "TYPE" stands for one of the above mentioned 4 categories). The NER-tool's classification of tokens into PER, LOC etc. sub-categories is not investigated at this point; here the aim is just to see whether the two systems could find the expected tokens and selected them as a NE, or not.

It is important to mention that *magyarlanc* was originally trained on the Szeged Treebank, which is built up from texts from six different genres, because "the main criteria were that they should be thematically representative of different text types." [11] It contains legal texts from the field of legislation, but only one specific type of it: full texts of laws.

On the other hand, the NER-tool was developed by using the same corpora, but just with another subset of it which contains short business news articles, so the training set of the NER-tool had not contained legal texts at all. The original F-measure calculated from the metrics of different NE type's results (PER, ORG, LOC, MISC) was an overall 94.77% on the Hungarian data [2].

[3] Examples are quoted from the original guideline.

In the next section, the results connected with the actual corpora's NEs will be briefly overviewed.

Corpora	NEs count (number of annotated tokens)	Number of NEs	Multi-token NEs	Type
Transcripts	56	29	23	Person
	41	22	11	Location
	69	33	22	Organization
	25	11	7	Regulation
	191	95	63	All in the section
Forums	95	51	19	Person
	4	4	0	Location
	6	5	1	Organization
	6	6	0	Regulation
	111	66	20	All in the section
Laws	0	0	0	Person
	5	3	2	Location
	14	8	4	Organization
	6	1	1	Regulation
	25	12	7	All in the section
Sum:	327	173	90	

Table 2: Manually annotated tokens

4 Results

In Table 3 the related token-level metrics are represented. The data was calculated from the tokens, which get a PROPN label from the *magyarlanc* and/or which an I-PER, I-ORG, I-LOC or I-MISC label from the NER-tool. The criterion of getting a label[4] from both tool was not expected (so the results of the two systems was handled independently from this aspect).

It can be seen that the NER-tool consequently gets higher scores in all terms of metrics, while there is a remarkable difference in the accuracy between the text types respectively. The Law texts proved to be the less precisely predicted ones, while the best scores were achieved for Transcripts.

In the next sections, the three different genres will be analysed in a more detailed way.

[4] PROPN label form the *magyarlanc* and an I-TYPE from the NER-tool

	NER-tool	*magyarlanc*	Sub-corpus
Precision	83.10	69.51	Forums
Recall	51.75	50.00	
F-score	63.78	58.16	
Precision	**94.48**	63.22	Transcripts
Recall	70.26	56.41	
F-score	80.59	59.62	
Precision	63.33	**26.67**	Laws
Recall	73.08	61.54	
F-score	67.86	37.21	

Table 3: Precision, Recall and F-Score

5 Discussion

In this section, the detailed results of the analysis will be described from the aspect of the three sub-corpora.

5.1 Forums

In internet forums, nicknames may have almost unpredictable forms, capitalization, extent etc. The following examples represent some typical occurrences in the examined text:

(1)

Token	POS assumed by *magyarlanc*	TYPE labeled by NER- tool
55teki55	PROPN	O
heidi1115	NUM	O
ObudaFan	PROPN	I-ORG

Some “multi-token” nicknames are listed here:

(2)

Token	POS assumed by *magyarlanc*	TYPE labeled by NER- tool
Dr.	NOUN	I-PER
Attika	NOUN	I-PER
Kovács	PROPN	I-PER
_	X	I-PER
Béla	PROPN	I-PER
_	X	I-PER
Sándor	PROPN	I-PER

It can be seen that these instances are not always properly identified but we should emphasize that the original training corpora of both tools did not contain instances of NEs like these specific ones.

Handling nicknames as NEs is a more interesting issue from a linguistic point of view. One of the arguments which can support considering nicknames as proper nouns is that they meet with the most fundamental properties of proper nouns mentioned in the literature.

Although we can see that there is no unified definition of proper nouns in the literature, but there are some common points between the definitions.

One of them is usually called as identifying function [12] of proper nouns. Nicknames which are used in websites admittedly fulfil this criterion, because this is the reason why people on websites even use it; to identify themselves with a unique linguistic unit, which only refers to one user. Furthermore, another point worth mentioning is the criterion that a linguistic unit can be called proper noun, if it does not change its referent within a given argumentation (as Kripke says) [13]. Nicknames fit into this expectation as well, since we can say that they usually define more accurately an individual, then a simple first name or last name (or even the two together)[5].

Moreover, from all of the NEs in the Forum sub-corpus, 69.29% (79 out of 114) was a mention of a nickname. All these justify that web nicknames should be seen as NEs.

At the same time, mentions of organisations can rise up questions about what is considered to be a proper noun. There are numerous instances where the same expression (which obviously refers to the same entity or object) occurs twice in the data; one with a capitalized first letter and one in lowercase:

(3) "...ez volt a legfőbb érve a **törvényszéknek**, hogy szabálytalanul lett kézbesítve az idézés."
"... the main argument of the court of law was that the summon was delivered irregularly."

But:

(4) "…...a végzés ellen fellebbezést nyújtsak be a várossal egy megyében található **Törvényszéknek** címezve 3 példányban."
"...against the order, I should submit an appeal in 3 copies to the Court of Law, which is in the same county as the town."

[5] Let's suppose that there is a class full of students. Although it is not likely, but possible, that there are more than one child in the room whose name is Tamás. It is less likely, but again, it is statistically possible, that there are more than one Kovács Tamás in the room. On the other hand, the list of First Names and Last Names in every language is a well-defined set of linguistic expressions (a definitely finite list). However, the potential combinations of characters (alphanumeric and special ones) are a more extensive set, therefore, the chance of having a unique nickname in a given site can be higher than having a unique name in a class (but indeed, it is not proved statistically yet). Moreover, a unique nickname is necessary in many websites.

In such cases, the two forms of mentioning these organizations are assumed to be distinct in the sense of what they refer to; the capitalized one is assumed to refer to a specific organization (e.g. in (4): Szegedi Törvényszék – Court of Law, Szeged) while the lowercase one is assumed to refer to the "type", or "role" of the organization (e.g. in (3): a type of court which can help you in this problem).

In the statistical data, only the capitalized mentions were included.

5.2 Transcripts

In the case of transcripts and in the output of *magyarlanc*, the most typical sources of errors may be related to the beginning of sentence. Within this, two typical problems occur most frequently.

In transcripts the main tool of discourse segmentation is the explicit marking of the speaker in the beginning of every utterance. These marks are abbreviations of the roles which the given person plays in that specific procedure, e.g. "V." stands for "vádlott" (*suspect*), "B." for "Bíró" (*judge*), "Ü." or "Ü / Ügyv." for "ügyvéd" (Lawyer) and so on. (5) is a typical case, where both the abbreviations are parsed incorrectly.

(5)	1	Ü	Ü	PROPN	Case=Nom|Number=Sing	0	ROOT
	2	/	/	PUNCT	_	1	PUNCT
	3	Ügyv	Ügyv	PROPN	Case=Nom|Number=Sing	1	COORD
	4	:	:	PUNCT	_	1	PUNCT
	5	Nem	nem	ADV	PronType=Neg	1	NEG
	6	.	.	PUNCT	_	0	PUNCT

"*Lawyer: No.*"

The remarkable majority (60.93%, 39 out of 64 instances) of falsely predicted PROPN labels was due to this phenomenon.

The other incorrectly predicted labels have miscellaneous reasons. For instance, it was frequent that the word "Bíró" (judge) at the beginning of the sentence was predicted to be a PROPN (because of the similar capitalization with the Hungarian surname: "Bíró").

Examining the false positive labels of the NER-tool, here we can see some examples for the falsely predicted tags:

(6)	a)	.	I-ORG	
	b)	Urat	I-PER	(*Sir, ACC*)
	c)	Interneten	I-ORG	(*on the Internet*)

(6) a) is a clear case, while b) and c) are more interesting ones. The word *internet* originally had a capitalized and a lowercase version depending on the referent of the word (Internet as an "organization" or internet as a notion), while the title, "úr" (*sir*) can be attributed as a part of the former proper noun. For instance, if we mention a bare last name, like Kovács, the referent of it can be vague in some cases. If we have two names; Kovács úr (*Sir Kovács*) and a Kovács néni (*Mrs. Kovács*) then, without the title, we cannot decide clearly who the name Kovács actually refers to. In this case, the title can be considered as a part of the NE.

5.3 Laws

Both within the laws and transcripts, there are numerous mentions of paragraphs of laws, such as:

- Btk. 236 § (1), *(236§ (1) from the Penal Code)*
- Ptk 6: 494§ (2), *(494§ (2) from the Civil Code)*
- Tht 1§ (2), *(1§ (2) from the Act on Condominium buildings)*

As a convenience, only the name of the acts are considered to be a NE here (for instance; Btk., Ptk., Tht. from the aforementioned ones).

Within the current part of texts, the main reason behind the relatively low scores of *magyarlanc* may be traced back to two distinct sources. Firstly, many of the typographical elements devoted to determine items of lists are predicted to be proper nouns:

(7) "(3a) A (3) bekezdésben foglalt szankciókat..."
"(3a) Sanctions mentioned in the (3) paragraph…"

Example (7) illustrates one of the sentences where this happened: the token "3a" was predicted to be a PROPN. On the other hand, there were a negligible number of cases when real NEs were not predicted as a proper name.

The NER-tool's most conspicuous missed NE was "1952. évi III. Törvény a Polgári perrendtartásról" (*1952. 3rd Act on the Rules of the Court*), because it is fully missed. It is important to mention here that although the structure of the NE is actually very typical in the nomenclature of Laws (for example YYYY, Roman Numeral, Act on *something*), but if the tool did not have access to annotated instances like that, they are very hard to predict

5.4 Dependency relations

To get a full picture about the recognition of NEs, the last approach is the analysis of the multi-token NEs in the syntactic level of *magyarlanc*.

Table 4 represents the calculated metrics of the syntax-based NE labelling in each legal domain:

		Sub-corpus
Precision	80.75	Forums
Recall	70.00	
F-score	74.99	
Precision	76.92	Transcripts
Recall	66.67	
F-score	71.43	
Precision	100.00	Laws
Recall	57.14	
F-score	72.73	

Table 4: Token-level metrics of syntactic parsing

However, it is important to note that the approach could identify much fewer multiword NEs than expected. In Table 5, the actual count of the NEs represented.

Corpora	NEs labeled by the *magyarlanc* on the syntactic level	multi-token NEs manually annotated[6]
Transcripts	23	63
Forums	15	20
Laws	7	7

Table5: Syntactic-level data

In the second column, the number of multi-token NEs can be seen, which get a *NE* tag from the *magyarlanc* at the syntactic level, while on the third column, the manually annotated multi-token NEs indicated.

The number of multi-token NEs (both in terms of manually annotated ones and which labelled by the *magyarlanc*) are much lower than expected originally, and not yet suitable for an exhausting statistical analysis.

Therefore, no general conclusions can be determined at this point and further investigations are needed.

6. Conclusions

In this paper, we focused on the identification of NEs in legal texts. We analyzed the output of the *magyarlanc* and the Szeged NER system and compared it to manually annotated NEs. The most typical sources of errors of the two negotiated NLP tools were also presented that cause most of the problems in the POS-tagging and in the NE-tagging approaches.

With regard to domain specificity, our investigations supported that all three sub-corpora had some unique peculiarities that need to be handled in order to get a higher rate of correctly recognized and classified NEs.

After a thorough examination, it turned out the multi-token NE category has a much lower presence in all investigated sub-corpora than expected, so a more massive amount of text should be analysed to be able to conclude more precise statements connected to the syntactic labelling efficiency.

References

1. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In Proceedings of RANLP (2013) 763-771

[6] cf. Table2

2. Szarvas Gy., Farkas R., Kocsor A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: The Ninth International Conference on Discovery Science LNAI 4265 (2006)
3. Quaresma, P., Gonçalves, T.: Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.): Number 6036 in Lecture Notes in AI. Springer-Verlag (2010) 44–59
4. Surdeanu, M., Nallapati, R., Manning, C.-D.: Legal claim identification: Information extraction with hierarchically labeled data. In: Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT) (2010)
5. Lenci, A., Montemagni, S., Pirrelli, V., Venturi, G.: Ontology learning from italian legal texts. In: Proceeding of the 2009 Conference on Law, ontologies and the Semantic Web: Channelling the Legal information Flood (2009)
6. Vincze, V., Farkas, R.: Tulajdonnevek a számítógépes nyelvészetben. In: Általános Nyelvészeti Tanulmányok XXIV (2012) 97-119
7. Móra, Gy., Vincze, V., Zsibrita, J.: Szófaji kódok és névelemek együttes osztályozása. In:Tanács, A., Vincze, V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem (2011) 131-142
8. Vincze V.: A Miskolc Jogi Korpusz nyelvi jellemzői. In: Szabó, M. (szerk.): A törvény szavai. Miskolc (2018)
9. Simon E., Farkas R., Halácsy P., Sass B., Szarvas Gy., Varga D.: A HunNER korpusz. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (2006) 373-376
10. Linguistic Data Consortium. ACE (automatic content extraction) English annotation guidelines for entities. https://www.ldc.upenn.edu/ collaborations/past-projects/ace, Version 5.6.6 2006.08.01. (2006)
11. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC 2004) at The 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland (2004)
12. Farkas, T.: A tulajdonnevek fordíthatóságáról és napjaink fordítási hibáiról, közszók és tulajdonnevek példáján. In: Névtani Értesítő 29 (2007) 167–188.
13. Kripke, S.: Naming and Necessity. Cambridge, Massachusetts: Harvard University Press (1980)
14. Várnai, J.-Sz.: A tulajdonnév a nyelvben és a nyelvészetben – A tulajdonnevek lehetséges megközelítéseiről, PhD értekezés, Debreceni Egyetem (2009)