Automatically generated NE tagged corpora for English and Hungarian

Dávid Márk Nemeskey
Research Institute for
Computer Science and Automation
Hungarian Academy of Sciences
H-1111 Kende utca 13-17, Budapest
nemeskey.david@sztaki.mta.hu

Eszter Simon
Research Institute for Linguistics
Hungarian Academy of Sciences
H-1068 Benczúr utca 33, Budapest
simon.eszter@nytud.mta.hu

Abstract

Supervised Named Entity Recognizers require large amounts of annotated text. Since manual annotation is a highly costly procedure, reducing the annotation cost is essential. We present a fully automatic method to build NE annotated corpora from Wikipedia. In contrast to recent work, we apply a new method, which maps the DBpedia classes into CoNLL NE types. Since our method is mainly language-independent, we used it to generate corpora for English and Hungarian. The corpora are freely available.

1 Introduction

Named Entity Recognition (NER), the task of identifying Named Entities (NEs) in unstructured texts and classifying them into pre-selected classes, is one of the most important subtasks in many NLP tasks, such as information retrieval, information extraction or machine translation. The NER task was introduced with the 6th Message Understanding Conference (MUC) in 1995 (Grishman and Sundheim, 1996). In MUC shared tasks the NER consists of three subtasks: entity names, temporal and number expressions. Although there is a general agreement in the NER community about the inclusion of temporal expressions and some numerical expressions, the most studied types are names of persons, locations and organizations. The fourth type, called "miscellaneous", was introduced in the CoNLL NER tasks in 2002 (Tjong Kim Sang, 2002) and 2003 (Tjong Kim Sang and De Meulder, 2003), and includes proper names falling outside the three classic types. Since then, MUC and CoNLL datasets and annotation schemes have been the major standards applied in the field of NER.

The standard datasets are highly domain-specific (mostly newswire) and are restricted in size. Researchers attempting to merge these datasets to get a bigger training corpus are faced with the problem of combining different tagsets and annotation schemes. Manually annotating large amounts of text with linguistic information is a time-consuming, highly skilled and delicate job, but large, accurately annotated corpora are essential for building robust supervised machine learning NER systems. Therefore, reducing the annotation cost is a key challenge.

One approach is to generate the resources automatically, another one is to use collaborative annotation and/or collaboratively constructed resources, such as Wikipedia, Wiktionary, Linked Open Data, or DBpedia. In this paper we combine these approaches by automatically generating freely available NE tagged corpora from Wikipedia.

The paper is structured as follows. In Section 2 we give an overview of related work. Section 3 contains a description of our method, and Section 4 shows how it is applied to Hungarian. The corpus format is described in Section 5. In Section 6 we present experiments and results on the newly generated datasets. Section 7 concludes the paper with a summary.

2 Wikipedia and NER

Wikipedia (WP, see http://wikipedia.org), a free multilingual Internet encyclopedia, written collaboratively by volunteers, is a goldmine of infor-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 38–46,
Jeju, Republic of Korea, 8-14 July 2012.

mation: at the time of writing, WP contains about 21 million interlinked articles. Of these, 3,903,467 are English, and 212,120 are Hungarian. WP has been applied to several NLP tasks such as word sense disambiguation, ontology and thesaurus building, and question answering (see Medelyan et al. (2009) for a survey). It is recognized as one of the largest available collections of entities, and also as a resource that can improve the accuracy of NER. The most obvious utilization of WP for NER is extracting gazetteers containing person names, locations or organizations (e.g. Toral and Muñoz (2006)). Creating dictionaries of entities is also a common step of NE disambiguation (Bunescu and Pasca, 2006; Cucerzan, 2007). Both supervised and unsupervised NER systems use such lists, see e.g. Nadeau et al. (2006) The knowledge embodied in WP may also be incorporated in NER learning as features, e.g. Kazama and Torisawa (2007) showed that automatic extraction of category labels from WP improves the accuracy of a supervised NE tagger.

Another approach to improve NER with WP is the automatic creation of training data. Richman and Schone (2008) built corpora for less commonly taught languages annotated with NE tags. They used the inherent category structure of WP to determine the NE type of a proposed entity. Nothman et al. (2008) used a similar method to create a NE annotated text in English. They transformed the WP links into NE annotations by classifying the target articles into standard entity classes. Their approach to classification is based primarily on category head nouns and the opening sentences of articles where definitions are often given.

Our approach to recognize and classify NEs in corpora generated from WP was to map the DBpedia ontology classes to standard NE tags and assign these to WP entities (see more details in Section 3.1). Except for the Semantically Annotated Snapshot of the English WP (SASWP) (Zaragoza et al., 2007), no such automatically built corpora are freely available. SASWP provides a wide range of linguistic information: POS tags, dependency labels, WordNet super senses and NE annotation according to WSJ and CoNLL tagsets. Even though the SASWP NEs were tagged by the best available open source taggers, the tags provided here, being based on the manual judgement of thousands of WP volunteers, are more reliable. Given the huge number of WP articles we can build sufficiently large corpora for less resourced languages as well, as our method is largely language-independent. We demonstrate this on Hungarian, a highly agglutinative language, with free word order and other typological characteristics detailed later in Section 4. There are smaller, manually annotated CoNLL-style datasets, but the one presented here is the first automatically NE annotated corpus for Hungarian.

3 Creating the English Corpus

Our goal is to create a large NE annotated corpus, automatically generated from WP articles. We followed a similar path to Nothman et al. (2008) and broke down the process into four steps:

1. Classify WP articles into entity classes.
2. Parse WP and split articles into sentences.
3. Label named entities in the text.
4. Select the sentences for inclusion in the corpus.

In this section, we describe how these steps were implemented. This section explains the general approach and its execution for English; Section 4 describes how the idea is adapted to Hungarian.

3.1 Articles as Entities

Many authors, such as Kazama and Torisawa (2007) and Nothman et al. (2008) used semi-supervised methods based on WP categories and text to classify articles into NE types. To avoid the inevitable classification errors, we obtain entity type information from the DBpedia knowledge base (Bizer et al., 2009), which presents type, properties, home pages, etc. information about pages in WP in structured form. With DBpedia we have high precision information about entity types at the expense of recall: of the 3,903,467 English WP pages, 1,470,293 are covered by DBpedia (as of 18 March, 2012).

The types in DBpedia are organized into a class hierarchy, available as an OWL[1] ontology containing 320 frequent entity categories, arranged into a taxonomy under the base class `owl:Thing`.

[1] http://www.w3.org/TR/owl-ref/

Most of the classes belong to the 6 largest sub-hierarchies: `Person`, `Organisation`, `Event`, `Place`, `Species` and `Work`. The taxonomy is rather flat: the top level contains 44 classes and there are several nodes with a branching factor of 20.

The type of entities is extracted automatically from WP categories. However, the mapping between WP categories and classes in the DBpedia ontology is manually defined. This, together with the fact that the existence of the reference ontology prevents the proliferation of categories observable in WP (Bizer et al., 2009), ensures that type information in DBpedia can be considered gold quality.

From the available NER annotation standards we elected to use the CoNLL (Tjong Kim Sang and De Meulder, 2003) NE types. It is not difficult to see the parallels between the DBpedia sub-hierarchies `Person`, `Organisation` and `Place` and the CoNLL NE types `PER`, `ORG` and `LOC`. The fourth category, `MISC` is more elusive; according to the CoNLL NER annotation guide[2], the sub-hierarchies `Event` and `Work` belong to this category, as well as various other classes outside the main hierarchies.

While the correspondence described above holds for most classes in the sub-hierarchies, there are some exceptions. For instance, the class `SportsLeague` is part of the `Organisation` sub-hierarchy, but according to the CoNLL annotation scheme, they should be tagged as `MISC`. To avoid misclassification, we created a file of DBpedia class–NE category mappings. Whenever an entity is evaluated, we look up its class and the ancestors of its class, and assign to it the category of the class that matches it most closely. If no match is found, the entity is tagged with `O`.

As of version 3.7, the DBpedia ontology allows multiple superclasses, making a directed acyclic graph[3]. Since selecting the right superclass, and hence, CoNLL tag, for classes with more than one parent cannot be reliably done automatically, the class-to-category mapping had to be determined manually. The only such class in version 3.7, `Library`, can be traced back to both `Place` and `Organisation`; its CoNLL tag is `LOC`. Using the mapping thus created, we compile a list that contains all entities in DBpedia tagged with the appropriate CoNLL category.

We note here that our method can be trivially modified to work with any tagset compatible with the DBpedia ontology (indeed, the DBpedia classes define a NE tagset themselves), but we leave the exploration of these possibilities for future work.

3.2 Parsing Wikipedia

WP is a rich source of information; in addition to the article text, a huge amount of data is embedded in infoboxes, templates, and the category structure. Our task requires only the links between the articles and the article text. In addition to in-article links, our method takes advantage of the redirect and interlanguage links, available as SQL dumps. The English corpus is based on the WP snapshot as of January 15, 2011. The XML files were parsed by the mwlib parser[4], the raw text was tokenized by a modified version of the Punkt sentence and word tokenizers (Kiss and Strunk, 2002). For lemmatization we used the Wordnet Lemmatizer in NLTK (Bird et al., 2009), and for part-of-speech tagging the HunPOS tagger (Halácsy et al., 2007).

3.3 Named Entity Labeling

In order to automatically prepare sentences where NEs are accurately tagged, two tasks need to be performed: identifying entities in the sentence and tagging them with the correct tag. Sentences for which accurate tagging could not be accomplished must be removed from the corpus. Our approach is based on the work of Nothman et al. (2008). The WP cross-references found in the article text are used to identify entities. We assume that individual WP articles describe NEs. A link to an article can then be perceived as a mapping that identifies its anchor text with a particular NE.

The discovered entities are tagged with the CoNLL label assigned to them in the entity list extracted from DBpedia. If the link target is not in the entity list, or the link points to a disambiguation page, we cannot determine the type of the entity, and tag it as UNK for subsequent removal from the corpus. Links to redirect pages are resolved to point instead to the redirect target, after which they are han-

[2] http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt

[3] http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions

[4] http://code.pediapress.com

dled as regular cross-references. Finally, sentences with UNK links in them are removed from the corpus.

The following sub-sections describe how the method explained above can be improved to increase precision, sentence coverage and to account for peculiarities in the English orthography and the CoNLL guidelines.

3.3.1 Non-entity Links

Strictly speaking, our original assumption of equating WP articles with NEs is not valid: many pages describe common nouns (Book, Aircraft), calendar-related concepts (March 15, 2007), or other concepts that fall outside the scope of NER. To increase sentence coverage, we modified the algorithm to prevent it from misclassifying links to these pages as unknown entities and discarding the sentence.

Common noun links are filtered by POS tags; if a link contains no `NNP`s, it is ignored.

Time expression links require special attention, because dates and months are often linked to the respective WP pages. We circumvented this problem by compiling a list of calendar-related pages and adding them to the main entity list tagged with the CoNLL category `O`.

Lowercase links for entities referred to by common nouns, such as *republic* to *Roman Republic* are not considered NEs and are ignored.

3.3.2 Unmarked Entities

In a WP article, typically only the first occurrence of a particular entity is linked to the corresponding page. Subsequent mentions are unmarked and often incomplete – e.g. family names are used instead of full names. To account for such mentions, we apply Nothman's (2008) solution. For each page, we maintain a list of entities discovered in the page so far and try to associate capitalized words in the article text with these entities. We augment the list with the aliases of every entity, such as titles of redirect pages that target it, the first and last names in case of a `PER` entity and any numbers in the name. If the current page is a NE, the title and its aliases are added to the list as well; moreover, as WP usually includes the original name of foreign entities in the article text, localized versions of the title are also added to the list as aliases. Nothman's solution used a trie to store the entity list, while we use a set, with more alias types than what he used. We expect more precise tagging from our slightly more rigorous solution.

3.3.3 Special Cases

Derived words According to the CoNLL guidelines, words derived from NEs are tagged as `MISC`. We complied with this rule by tagging each entity whose head is not a noun, as well as when the link's anchor text is not contained in the entity's name, as `MISC`. The most prominent example for such entities are nationalities, which can be linked to their home country, a `LOC`; e.g. *Turkish* to *Turkey*. Our solution assigns the correct tag to these entities.

First word in a sentence As first words are always capitalized, labeling them is difficult if they are unlinked and not contained in the entity alias set. We base the decision on the POS tag of the first word: if it is `NNP`, we tag it as `UNK`; otherwise, `O`.

Reference cleansing Page titles and anchor texts may contain more than just the entity name. Personal titles are part of the entity name in WP, but not in CoNLL, and punctuation marks around the entity may become part of the link by mistake. We tag all punctuation marks after the entity name as `O`.

To handle personal titles, we extracted a list from the WP page *List of titles*, which contains titles in many languages. We manually removed all titles that also function as given names, such as *Regina*. If a link to a `PER` or `UNK` entity, or an unlinked entity starts with, or consists solely of a title in the list, we tag the words that make up the title as `O`.

Incidental capitalization Various non-`NNP` words in English are capitalized: names of months, the pronoun *I*, and non-entity acronyms such as *RSVP*. While the latter two types are unlikely to appear in WP text, we assembled a list of these words and tag them as `O` unless they are part of the alias set.

3.4 Sentence Filtering

As mentioned above, sentences with words tagged as UNK are discarded. Furthermore, there are many incomplete sentences in the WP text: image captions, enumerations items, contents of table cells, etc. On the one hand, these sentence fragments may be of too low quality to be of any use in the traditional NER task. On the other hand, they could prove to be invaluable when training a NER tagger for User Generated Content, which is known to be noisy and fragmented. As a compromise we included these fragments in the corpus, but labelled them as "low quality", so that users of the corpus can decide whether they want to use them or not. A sentence is labelled as such if it either lacks a punctuation mark at the end, or it contains no finite verb.

4 Creating the Hungarian Corpus

The procedure described in the previous section was used to generate the Hungarian corpus as well. However, typological differences posed several problems. In this section we describe the differences between the two languages related to labeling NEs, and the changes they prompted in the method.

4.1 Parsing the Hungarian Wikipedia

Although Hungarian is reckoned to be a less resourced language, and it is not supported in NLTK, several high quality language processing tools have been developed for Hungarian in recent years. For tokenization and sentence segmentation we used an in-house statistical tool tailored for Hungarian. It has been trained on the largest manually annotated Hungarian corpus (Csendes et al., 2004), and it handles the peculiarities of Hungarian orthography, such as the periods placed after numbers in date expressions. Lemmatization was performed by HunMorph (Trón et al., 2005) and HunDisambig, an in-house disambiguator to select the right analysis based on the word context.

For the most part Hungarian expresses grammatical elements within a word form using affixes. HunMorph outputs KR-codes (Kornai et al., 2004), which, in addition to the POS category, also include inflectional information, making it much better suited to agglutinative languages than Penn Treebank POS tags. One shortcoming of the KR-code is that it does not differentiate between common and proper nouns. Since in Hungarian only proper nouns are capitalized, we can usually decide whether a noun is proper based on the initial letter. However, this rule can not be used if the noun is at the beginning of a sentence, so sentences that begin with nouns have been removed from the corpus.

4.2 Named Entity Labeling in Hungarian

For well-resourced languages, DBpedia has internationalized chapters, but not for Hungarian. Instead, the Hungarian entity list comprises of the pages in the English list that have their equivalents in the Hungarian WP. Two consequences follow. First, in order to identify which pages denote entities in the Hungarian WP, an additional step is required, in which the Hungarian equivalents of the English pages are added to the entity list. The English titles are retained because (due to the medium size of the Hungarian WP) in-article links sometimes point to English articles.

Second, entities without a page in the English WP are absent from the entity list. This gives rise to two potential problems. One is that compared to English, the list is relatively shorter: the entity/page ratio is 12.12%, as opposed to the 37.66% of the English WP. The other, since mostly Hungarian people, places and organizations are missing, a NER tagger that takes the surface forms of words into account might be mislead as to the language model of entity names. To overcome these problems, the list has to be extended with Hungarian entity pages that do not have a corresponding English page. We leave this for future work.

To annotate our corpus with NE tags, we chose to follow the annotation guidelines of the largest human-annotated NER corpus for Hungarian, the Szeged NER corpus (Szarvas et al., 2006). It is similar to CoNLL standards: contains newswire texts, comprises ca. 200,000 tokens, and is annotated with NE class labels in line with the CoNLL annotation scheme. However, the convention of what constitutes a NE is slightly different for Hungarian.

4.2.1 Special cases

The Szeged NER guideline relies heavily on the rules of capitalization to decide which words should be marked as NEs. The following concepts are not

train	test	precision	recall	F-measure
Szeged NER	Szeged NER	94.50	94.35	94.43
huwiki	huwiki	90.64	88.91	89.76
huwiki	Szeged NER	63.08	70.46	66.57
Szeged NER with wikilists	Szeged NER	95.48	95.48	95.48
Szeged NER with wikitags	Szeged NER	95.38	94.92	95.15

Table 1: Hungarian results.

proper nouns in Hungarian, and thus are not considered as NEs: names of languages, nationalities, religions, political ideologies; adjectives derived from NEs; names of months, days, holidays; names of special events and wars.

There is another special case in Hungarian: unlike in English, the number of compound words is quite large, and NEs can also be subject to compounding. In this case the common noun following the NE is joined with a hyphen, so they constitute one token. However, the joint common noun can modify the original sense of NE, depending on the semantics of the common noun. For example in the compound *Nobel-díj* ['Nobel Prize'] the common noun changes the labeling from `PER` to `MISC`, while in the case of the compound *WorldCom-botrány* ['WorldCom scandal'] the NE tag changes from `ORG` to `O`. The solution to this problem is not obvious, and needs more investigation.

5 Data Description

The corpora are available under the Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA), the same license under which the text of WP is released. The data files can be freely downloaded from `http://hlt.sztaki.hu`. The corpora will also be distributed through the META-SHARE network, which is an open, distributed facility for exchanging and sharing resources, and is one of the lines of action of META-NET, a Network of Excellence funded by the European Commission.

The files are in multitag format. Content lines are tab separated; there is one column for the tokens plus one column per tagset. Sentence boundaries are marked by empty lines. The linguistic features include the lemmatized form of the word and its POS tag. Two NE tags are included with each word: the most specific DBpedia category it belongs to and the CoNLL NE tag. While the NE tags can be considered as a "silver standard", the linguistic features are provided on a "best-effort" basis.

6 Evaluation

Having the obvious advantages, an automatically generated corpus can not serve as a gold standard dataset. Then what can we do with silver standard corpora? They can be very useful for improving NER in several ways: (a) for less resourced languages, they can serve as training corpora in lieu of gold standard datasets; (b) they can serve as supplementary or independent training sets for domains differing from newswire; (c) they can be sources of huge entity lists, and (d) feature extraction.

To evaluate our corpora we used a maximum entropy NE tagger (Varga and Simon, 2007), which was originally developed for labeling NEs in Hungarian texts, but can be tuned for different languages as well. Corpus-specific features (e.g. NP chunks, WP links) were removed to get better comparability, so the feature set consists of gazetteer features; sentence start and end position; Boolean-valued orthographic properties of the word form; string-valued surface properties of the word form; and morphological information.

We used the CoNLL standard method for evaluation. According to this, an automatic labeling is correct if it gives the same start and end position, and the same NE class as the gold standard. Based on this, precision and recall can be calculated, and the F-measure, as usual, the harmonic mean of these two values.

6.1 Wikipedia data

Our automatic annotation process retains all of the WP sentences which remained after our two-step filtering method, so sentences without NEs are also in-

	enwiki	enwiki filtered	CoNLL	huwiki	huwiki filtered	Szeged NER
token	60,520,819	21,718,854	302,811	19,108,027	3,512,249	225,963
NE	3,169,863	3,169,863	50,758	456,281	456,281	25,896
NE density	5.23%	14.59%	16.76%	2.38%	12.99%	11.46%

Table 2: Corpus size and NE density.

train	test	precision	recall	F-measure
CoNLL	CoNLL	85.13	85.13	85.13
enwiki	enwiki	72.46	73.33	72.89
enwiki	CoNLL	56.55	49.77	52.94
CoNLL with wikilists	CoNLL	86.33	86.35	86.34
CoNLL with wikitags	CoNLL	85.88	85.94	85.91

Table 3: English results.

cluded in the corpus. The rationale behind this is that we wanted to reserve the original distribution of names in WP as much as possible. However, after further investigation of the NE density in our corpora and gold standard corpora, we decided not to include the sentences without NEs in evaluation datasets.

Table 2 summarizes the data regarding corpus size and NE density. The English (enwiki) and the Hungarian WP (huwiki) corpora originally have the NE density of 5.23% and 2.38%, respectively. In comparison to the gold standard datasets (CoNLL, Szeged NER) these counts are quite low. It can be due to the difference between domains: newswire articles usually contain more NEs, typically ORG. The other reason might be that we discarded sentences containing unidentified NEs (cf. Section 3).

6.2 Experiments and results

The English WP corpus was evaluated against itself and a manually annotated English corpus. Since the filtered English WP corpus, containing only the sentences with NEs, is still very large, our experiments were performed with a sample of 3.5 million tokens, the size of our filtered Hungarian corpus, divided into train and test sets (90%-10%).

For English cross-corpus evaluation the CoNLL-2003 corpus was chosen. As is well known, training and testing across different corpora decreases F-measure. Domain differences certainly affect NER performance, and the different annotation schemes pose several compatibility problems. Nothman et al. (2008) showed that each set of gold standard training data performs better on corresponding test sets than on test sets from other sources. The situation here is similar (see Table 3 for results): the NE tagger trained on WP does not achieve as high performance tested against CoNLL test set (enwiki-CoNLL) as one trained on its own train set (enwiki-enwiki).

WP-derived corpora can also be used for improving NER accuracy in other ways. First, we collected gazetteer lists from the corpus for each NE category, which improved the overall F-measure given to the NE tagger training and testing on CoNLL dataset (CoNLL with wikilists). A second trial was labeling the CoNLL datasets by the model trained on WP corpus, and giving these labels as extra features to the next CoNLL train (CoNLL with wikitags). Both methods result in improved F-measure on CoNLL test set.

Since in Hungarian NE tagging we followed the Szeged NER corpus annotation guidelines, we performed the experiments on this dataset. Hungarian results are similar to the English ones (see Table 1), the only difference is that F-measures for Hungarian are significantly higher. This can be due to the fact that the MISC category for Hungarian contains less types of names, thus the inconsistency of this class is smaller (cf. Section 4). In contrast to the CoNLL corpus, the Szeged NER corpus was accurately annotated with an inter-annotator agreement over 99%.

Due to the quite good F-measure of training on

our Hungarian train corpus and testing on the corresponding test set, our Hungarian corpus can serve as a training corpus to build NE taggers for non-newswire domains.

7 Conclusion

We have presented freely available NE tagged corpora for English and Hungarian, fully automatically generated from WP. In contrast to the methods used so far for automatic annotation of NEs in WP texts, we applied a new approach, namely mapping DBpedia ontology classes to standard CoNLL NE tags, and assigning them to WP entities. Following Nothman (2008), the process can be divided into four main steps: classifying WP articles into entity classes; parsing WP and splitting articles into sentences; labeling NEs in the text; and selecting sentences for inclusion in the corpus.

The huge amount of WP articles opens the possibility of building large enough corpora for otherwise less resourced languages such as Hungarian. Due to the particularities of Hungarian, some steps are slightly different, and special linguistic phenomena pose several problems related to the NER task to solve.

Automatically generated corpora can be useful for improving NER in more ways. We showed that gazetteer lists extracted from our corpora, and training with extra features given by the model trained on our corpora, improve F-measure. Moreover, our Hungarian corpus can serve as a training corpus for more general domains than the classic newswire.

Acknowledgements

This research was supported by OTKA grant no. 82333 and the CESAR project under the ICT Policy Support Programme (grant no. 271022). The authors are grateful to Attila Zséder for his work on Wikipedia parsing and to András Kornai for his insightful comments.

References

Steven Bird, Ewan Klein, Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. 2009. DBpedia – A Crystallization Point for the Web of Data. In: *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Issue 7, pages 154–165.

B. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16.

Dóra Csendes, János Csirik, Tibor Gyimóthy. 2004. The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: *Proceedings of TSD 2004*, vol. 3206, pages 41–49.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, June 2007. pages 708–716.

Ralph Grishman and B. Sundheim. 1996. Message Understanding Conference – 6. In: *Proc. International Conference on Computational Linguistics*.

P. Halácsy, A. Kornai and Cs. Oravecz. 2007. Hunpos – an open source trigram tagger. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

Tibor Kiss, Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. In: *Computational Linguistics*, 32 (4): pages 485–525.

András Kornai, Péter Rebrus, Péter Vajda, Péter Halácsy, András Rung, Viktor Trón. 2004. Általános célú morfológiai elemző kimeneti formalizmusa (The output formalism of a general-purpose morphological analyzer). In: *Proceedings of the 2nd Hungarian Computational Linguistics Conference*.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67: 716–754.

David Nadeau, Peter D. Turney and Stan Matwin. 2006. Unsupervised named entity recognition: Generating gazetteers and resolving ambiguity. In: *Proceedings of the 19th Canadian Conference on Artificial Intelligence*, volume 4013 of *LNCS*, pages 266–277.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into Named Entity Training Data. In: *Proceedings of the Australasian Language Technology Workshop*, Vol 6., pages 124–132.

Joel Nothman, Tara Murphy and James R. Curran. 2009. Analysing Wikipedia and Gold-Standard Corpora for NER Training. In: *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 612–620.

Alexander E. Richman and Patrick Schone. 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. In: *Proceedings of ACL-08: HLT*, pages 1–9.

György Szarvas, Richárd Farkas, András Kocsor. 2006. A highly accurate Named Entity corpus for Hungarian. In: *Proceedings of International Conference on Language Resources and Evaluation*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: *Proceedings of the 6th Conference on Natural Language Learning*, pages 1–4, Taipei, Taiwan.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147, Edmonton, Canada.

A. Toral and R. Muñoz. 2006. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In: *EACL 2006*.

Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, Dániel Varga. 2005. Hunmorph: open source word analysis. In: *Proceedings of the ACL 2005 Workshop on Software*.

Dániel Varga and Eszter Simon. 2007. Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 18: 293–301.

H. Zaragoza and J. Atserias and M. Ciaramita and G. Attardi. 2007. Semantically Annotated Snapshot of the English Wikipedia v.1 (SW1). `http://www.yr-bcn.es/semanticWikipedia`