# Maximum Entropy Named Entity Recognition for Czech language

Michal Konkol and Miloslav Konopík

University of West Bohemia
Laboratory of Intelligent Communication Systems
Univerzitni 8, 30614 Pilsen, Czech Republic
{konkol,konopik}@kiv.zcu.cz

**Abstract.** Named Entity Recognition (NER) is an important preprocessing tool for many Natural Language Processing tasks like Information Retrieval, Question Answering or Machine Translation. This paper is focused on NER for Czech language. The proposed NER is based on knowledge and experiences acquired on other languages and adapted for Czech. Our recognizer outperforms the previously introduced recognizers for Czech. The article is also focused on the use of semantic spaces for NER. Although no significant improvement was yet achieved in this way, we believe that the research is worth of sharing.

**Keywords:** Named Entity Recognition, Maximum Entropy, Semantic Spaces, Czech

## 1 Introduction

Named Entity recognition (NER) is a very important preprocessing tool for Question Answering, Information Retrieval or Machine Translation. NER was firstly introduced at Message Understanding Conference (MUC) 6 [4] in 1995. The definition of the NER task at MUC-6 was to find 7 categories of Named Entities (NE) like persons, organizations or dates. These expressions are very important, because they are very often the key points in a text. Properly identified NEs can improve results for other Natural Language Processing tasks.

Semantic spaces are one of recent fields of research that focuses on methods to automatically find relations between words. It started with the well known LSA method [2] and continued to the very advanced Beagle method [6]. The idea is to use relations between words to help dealing with an unknown context using a similar known context. The similar known context is found using semantic spaces (see section 5). Many application such as Information Retrieval have already proven the usefulness of semantic spaces. Our intention in this article is to explore the possibility to use semantic spaces for NER.

In this paper a new recognizer based on Maximum Entropy and semantic spaces is presented. The classifier is described in section 3. Details about used features can be found in section 4. Basic information about semantic spaces is given in section 5. Experiments and results are presented in section 6. The last section contains the summary.

## 2 State of the Art

Since the time of MUC-6 a big effort has been put into NER. Many systems have been presented for different languages. Although the majority of systems was done for English, good systems were introduced for some other languages such as Chinese, Japanese or German [14]. The state of the art F-measure for English is around 90 and for German around 70. This difference around 20 percent is caused by the differences of these languages.

There are two basic approaches to NER. The first one is based on hand made rules and dictionaries and typically involves techniques like Context Free Grammars, Finite State Automata or Regular Expressions. The second one is based on statistical methods. The most used methods for statistical NER are Maximum Entropy Models [1], Conditional Random Fields [11], Hidden Markov Models [16], Support Vector Machines [5] among others. There are also hybrid systems combining more of mentioned methods [7].

There are also different types of training. The training methods can be divided into supervised, semi-supervised and unsupervised methods. Supervised methods need labelled training data to find the best parameterization of the classifier. Semi-supervised methods need some small data which are used as a seed for the training. Unsupervised methods do not need any data.

The Czech language is quite different in comparison with all mentioned languages. Czech is highly inflectional and has a more flexible word order. In addition the custom for writing proper nouns is not very helpful and in some cases not even stable. In comparison to English, the Czech NER is still unexplored. Only two systems for NER were presented for Czech language. The first system used Decision Trees and its F-measure was 68 [15]. The second one was based on SVM and achieved F-measure 71 on the same corpus [8] and was introduced in 2009.

## 3 Classifier

Our classifier is based on the maximum entropy principle. The principle says that we are looking for a model which will satisfy all our constraints and does not make any other assumptions. To define a constraint we firstly need to define a feature. A feature is a function in the following form. Typically binary features are used, but in general any non-negative function is possible. For example, consider the following feature designed to express the relation between NE class PERSON and the capitalization of the word $x$.

$$f(x,y) = \begin{cases} 1 & \text{if } y \text{ is PERSON and } x \text{ starts with capital letter} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The constraint is then defined as equality of mean values for a given feature.

$$E_p(f_i(x,y)) = E_{\tilde{p}}(f_i(x,y)) \tag{2}$$

where $E_{\tilde{p}}(f_i(x,y))$ is mean value of a feature computed over the training data and $E_p(f_i(x,y))$ is mean value of the model. It is guaranteed that such a model exists. In addition it is unique, follows the maximum-likelihood distribution and is in the following form[3].

$$p(y|x) = \frac{1}{Z(x)} \exp \sum_i \lambda_i f_i(x,y) \tag{3}$$

$$Z(x) = \exp \sum_i \lambda_i f_i(x,y) \tag{4}$$

$Z(x)$ is just a normalizing factor and ensures that $p(y|x)$ is a probability distribution. The $\lambda_1 \ldots \lambda_n$ are parameters of the model. Various training algorithms can be used for finding appropriate parameters. The most common algorithms are Generalized Iterative Scaling (GIS) and Improved Iterative Scaling (IIS). These algorithms are not the most effective though [10]. We have implemented GIS and a limited memory BFGS (L-BFGS) method [12]. We have to approve that L-BFGS is much more effective than GIS. The classifier for one of the test mentioned later was trained using both methods. The training with L-BFGS method takes 271.053 seconds, while with GIS it was 486.258 seconds. Different stopping conditions were used and L-BFGS should be more precise. The difference in time would be probably higher with the same stopping conditions.

## 4 Features

The recognizer is built from two important parts. The first part is a classifier which was described in the previous section. The second part is a feature set. Both part are equally important. Our feature set is composed of the following features.

- *Morphological features* are used separately or in combinations. One of possible combinations can be if this word has the same case, number and gender as the previous word. We use part of speech, case, person, number and gender as features.
- *Lemma* is almost a must for highly inflectional language like Czech. Many words can have more than 10 different word forms and it causes a sparseness problem without lemmatization. Lemmas are also useful for other features like gazetteers, because they are usually using only lemmas. The F-measure decreases rapidly without usage of lemmas.
- *Regular expressions* are used for some special types of words. Most of the regular expressions are used for identification of numbers, dates and times. Some are used for special words containing mixed capitalization with numbers or symbols or for identification of short cuts.
- *Capitalization* of words has different importance among languages. Capitalization is very important for English but for German is much less important. The Czech language is somewhere between English and German.

Three classes of capitalization are used as a feature. Mixed capitalization, all capitalized and first letter capitalized.
- *Gazetteers* are used for some classes. All of our gazetteers are from publicly accessible sources and are not manually edited. We use gazetteers for forenames, surnames, cities, rivers, mountains, organizations etc.
- *Learned lists* are similar to gazetteers but are learned directly from the corpus. This feature is real valued and returns the probability that given word belongs to a particular class.

All the mentioned features are computed for some window around the classified word. A couple of different window sizes were tested. The final system uses a window $-3, \ldots, 3$ with an exception of lemmas which have window $-2, \ldots, 2$. The difference between $-3, \ldots, 3$ and $-2, \ldots, 2$ is very small in the final results. Larger windows did not improve results and smaller windows did not achieve similar results.

## 5 Semantic Spaces

Semantic space is a space of words. Every word is a single point in this space and is represented by its vector, which is often quiet large. The position of the word is somehow related to its meaning. After the creation of semantic space a distance between words can be measured. This distance can be used as a relation or similarity of the words.

There are various methods which can build a semantic space from unlabelled data. The methods use different ways to create a semantic space and therefore their results are slightly different. Some of these methods are LSA [2], HAL [9],COALS [13] or BEAGLE [6].

In our experiments we have used the COALS method. This method constructs a word matrix where each position is count of occurrences of particular words. The matrix is then normalized and a Singular Value Decomposition (SVD) is calculated. SVD reduces the dimension of vector space. We have used the dimension reduced to 1000.

Our basic idea is that if some word is similar to a known NE, there is a higher probability that this word is a NE of the same class. This idea can be extended to the context. If we know that a word is almost always person NE if the previous word is "Mr.", then it would be probably person NE if there is a word with high similarity to "Mr.".

We have tested the following usages of semantic spaces.

- *Bigger groups* – we have created lists of about 20 most frequent words for each NE category. The feature is then the average similarity of the classified word to the words on the list.

$$f(y|x) = \frac{\sum_{w \in G_y} similarity(x, w)}{|G_y|} \qquad \forall y \tag{5}$$

– *Smaller groups* – we have created about 20 smaller groups with 2-4 words. One group was for example Prague, Pilsen and Brno, which are largest Czech cities. The feature is again average similarity to these words.

$$f(y|x) = \frac{\sum_{w \in G_i} similarity(x, w)}{|G_i|} \qquad \forall i \tag{6}$$

– *Single words* – the last feature uses similarity to individual words. Each feature is a similarity to a particular word. This feature is very computationally demanding.

$$f(y|x) = similarity(x, w) \qquad \forall w \in V \tag{7}$$

## 6 Experiments

All experiments were done on the Czech Named Entity corpus [15]. The corpus was divided into two parts. The first part contains 90% of sentences and is used for training. The second part was used as test data. Each experiment was done 10 times with different parts of corpus used as test data. The presented results are averaged.

Our experiments are evaluated by the standard measures for NER which are precision, recall and F-measure (or F-score). These quantities have the following meaning.

– *Precision* is percent of correctly marked entities from all marked entities.
– *Recall* is percent of correctly marked entities from all entities in the data.
– *F-measure* is a harmonic mean of precision and recall.

### 6.1 Results

Previous results for Czech NER are shown in table 1. Our primary goal was to create an effective recognizer and to improve these results.

**Table 1.** A comparison of previous results with our experiments.

| | | precision | recall | F-measure |
|---|---|---|---|---|
| Related work | DT [15] | 81 | 59 | 68 |
| | SVM [8] | 75 | 67 | 71 |
| Our experiments | Basic features | 76,78 | 69,58 | 72,94 |
| | Big groups | 76,89 | 69,71 | 73,08 |
| | Small groups | 76,84 | 69,18 | 72,76 |
| | Single words | 76,61 | 69,30 | 72,72 |

The first experiment was made with *basic features* listed in section 4 except semantic spaces. Its purpose was to create a good starting point for other

experiments and a preprocessing tool for other projects. This experiment was successful and improved the state of the art for Czech NER. The detailed results are shown in table 2.

The following experiments were focused on semantic spaces. We assumed that entity should be found in similar contexts like other entities in the same class. So we have made a list of 20 typical words for each class (denoted as *bigger groups* in table 2) and made an average similarity of these words with the classified word. The results in table 3 showed that this feature does not work. The problem was identified as too many words in the list of typical words.

We have tried to fix the problem of the second experiment with *smaller groups*. New groups had about 3 words. For example one group was made for Czech cities and contains Prague, Brno, Pilsen and Ostrava, which are four largest cities in Czech republic. There were about 20 groups focused specifically on some subclass. There was no significant change in the results (table 4).

The last experiment leaved the idea of groups and used *single words*. Similarity of each word from the corpus is taken as a feature. This feature is very computational demanding, because the similarity have to be computed $|N| \cdot |V|$ times, where $|N|$ is number of words in corpus and $|V|$ is size of vocabulary. It was very surprising for us, that even results of this experiment (table 5) did not show any significant change.

**Table 2.** Results for basic features.

| | precision | recall | F-measure |
|---|---|---|---|
| NUM | 60,54 | 62,02 | 55,98 |
| LOC | 81,08 | 74,97 | 77,82 |
| ORG | 63,33 | 51,62 | 56,76 |
| OTH | 66,29 | 42,40 | 51,48 |
| PER | 81,23 | 86,54 | 83,76 |
| DAT | 88,51 | 87,23 | 87,77 |
| Overall | 76,78 | 69,58 | 72,94 |

**Table 3.** Results for groups of 20 words.

| | precision | recall | F-measure |
|---|---|---|---|
| NUM | 60,75 | 62,18 | 56,28 |
| LOC | 80,54 | 75,23 | 77,71 |
| ORG | 63,70 | 52,07 | 57,08 |
| OTH | 67,21 | 41,91 | 51,36 |
| PER | 81,29 | 86,80 | 83,90 |
| DAT | 87,70 | 87,33 | 87,41 |
| Overall | 76,89 | 69,71 | 73,08 |

**Table 4.** Results for groups of 3 words.

| | precision | recall | F-measure |
|---|---|---|---|
| NUM | 60,25 | 61,77 | 55,55 |
| LOC | 80,89 | 74,60 | 77,55 |
| ORG | 62,53 | 50,81 | 56,03 |
| OTH | 67,98 | 41,58 | 51,24 |
| PER | 81,42 | 86,48 | 83,83 |
| DAT | 87,69 | 87,50 | 87,46 |
| Overall | 76,84 | 69,18 | 72,76 |

**Table 5.** Results for single words.

| | precision | recall | F-measure |
|---|---|---|---|
| NUM | 60,45 | 62,28 | 56,09 |
| LOC | 81,08 | 74,45 | 77,53 |
| ORG | 63,14 | 51,07 | 56,30 |
| OTH | 65,76 | 41,75 | 50,83 |
| PER | 80,96 | 86,48 | 83,59 |
| DAT | 88,69 | 87,22 | 87,87 |
| Overall | 76,61 | 69,30 | 72,72 |

Tables 2–5 show the performance of the classifier for particular NE classes. The results for numbers (see the NUM row) are influenced by the fact, that 2 parts of the data used for testing do not contain any numbers. This leads to 0

precision, undefined recall and 0 F-measure. The average is then the F-measure around 56, but by leaving this two parts it is increased the value to around 75.

Apparently, for dates and times (DAT), persons (PER), numbers (NUM) and locations (LOC) the results are satisfactory in comparison to organizations (ORG) and other NEs (OTH). The results of organizations and other NEs are worse, because these classes are open and not well defined. The names of organizations are fuzzy and often can be distinguished from other classes only from the context or by using global knowledge (e.g. Johnie Walker). The NEs class "other" is even worse already from the definition since it covers a very wide range of entities. Entities from these classes are also not very often repeated in the corpus.

## 7 Conclusion and Future Work

We have created a new NE recognizer based on maximum entropy. Our recognizer achieved 76.78 recall, 69.58 precision and 72.94 F-measure and according to our best knowledge outperformed the previously published results. We have also conducted tests with semantic spaces used as a feature for our classifier. Our tests have so far shown no statistically significant improvement.

Our plan is to continue in the development of NER systems. We still believe in semantic spaces however a more proper way to use them is probably necessary. The combination of classifiers may also be a way to improve the results. At the moment a new corpus for Czech named entity recognition is being prepared. The corpus is designed to avoid some problems found in the Czech Named Entity corpus [8].

**Acknowledgments.** This research was supported by Advanced Computer and Information Systems project no. SGS-2010-028.

## References

1. Curran, J. R., and Clark, S. Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4* (Morristown, NJ, USA, 2003), Association for Computational Linguistics, pp. 164–167.
2. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41* (1990), 391–407.
3. Della Pietra, S., Della Pietra, V., and Lafferty, J. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell. 19* (April 1997), 380–393.
4. Grishman, R., and Sundheim, B. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1* (Stroudsburg, PA, USA, 1996), COLING '96, Association for Computational Linguistics, pp. 466–471.

5. Isozaki, H., and Kazawa, H. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1* (Morristown, NJ, USA, 2002), Association for Computational Linguistics, pp. 1–7.
6. Jones, M. N., and Mewhort, D. J. K. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review 114* (2007), 1–37.
7. Kozareva, Z., Ferrández, O., Montoyo, A., Muñoz, R., Suárez, A., and Gómez, J. Combining data-driven systems for improving named entity recognition. *Data Knowl. Eng. 61* (June 2007), 449–466.
8. Kravalová, J., and Žabokrtský, Z. Czech named entity corpus and svm-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration* (Stroudsburg, PA, USA, 2009), NEWS '09, Association for Computational Linguistics, pp. 194–201.
9. Lund, K., and Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments and Computers 28*, 2 (1996), 203–208.
10. Malouf, R. A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning - Volume 20* (Stroudsburg, PA, USA, 2002), COLING-02, Association for Computational Linguistics, pp. 1–7.
11. McCallum, A., and Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4* (Morristown, NJ, USA, 2003), Association for Computational Linguistics, pp. 188–191.
12. Nocedal, J. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation 35*, 151 (1980), 773–782.
13. Rohde, D. L. T., Gonnerman, L. M., and Plaut, D. C. An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Psychology 7* (2004), 573–605.
14. Tjong Kim Sang, E. F., and De Meulder, F. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4* (Stroudsburg, PA, USA, 2003), CONLL '03, Association for Computational Linguistics, pp. 142–147.
15. Ševčíková, M., Žabokrtsky, Z., and Kruza, O. Named entities in czech: annotating data and developing ne tagger. In *Proceedings of the 10th international conference on Text, speech and dialogue* (Berlin, Heidelberg, 2007), TSD'07, Springer-Verlag, pp. 188–195.
16. Zhou, G., and Su, J. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Morristown, NJ, USA, 2002), ACL '02, Association for Computational Linguistics, pp. 473–480.