A New State-Of-The-Art Czech Named Entity Recognizer

Jana Straková, Milan Straka, and Jan Hajič

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25,
118 00 Prague, Czech Republic
{strakova,straka,hajic}@ufal.mff.cuni.cz

Abstract. We present a new named entity recognizer for the Czech language. It reaches 82.82 F-measure on the Czech Named Entity Corpus 1.0 and significantly outperforms previously published Czech named entity recognizers. On the English CoNLL-2003 shared task, we achieved 89.16 F-measure, reaching comparable results to the English state of the art. The recognizer is based on Maximum Entropy Markov Model and a Viterbi algorithm decodes an optimal sequence labeling using probabilities estimated by a maximum entropy classifier. The classification features utilize morphological analysis, two-stage prediction, word clustering and gazetteers.

Keywords: named entities, named entity recognition, Czech

1 Introduction

Named entity recognition is one of the most important tasks in natural language processing. Not only is named entity identification an important component of large applications, such as machine translation, it also belongs to one of the most useful natural language processing applications itself. Therefore it has received a great deal of attention from computational linguists. Multiple shared tasks have been organized (CoNLL-2003 [19], MUC7 [4]) for the English language and the existing state of the art systems reach remarkable results, with almost human annotator performance. Other languages, such as Czech, are with some delay also receiving attention. In this paper, we present a new state of the art named entity recognition system for Czech and English. We significantly outperform the three known Czech named entity recognizers ([20], [11], [10]) and achieve results comparable to English state of the art ([16]). The organization of this work is as follows: We describe the datasets and their evaluation methodology in Chapter 2 and present the related work in Chapter 3. Our methodology is described in Chapter 4 and results in Chapter 5. Chapter 6 concludes the paper.

2 Datasets and Task Description

2.1 English CoNLL-2003 shared task

For English, many datasets and shared tasks exist (e.g. CoNLL-2003 [19], MUC7 [4]). In this paper, we used one of the most widely recognized shared task dataset, the

CoNLL-2003 ([19]). In this task, four classes are predicted: PER (person), LOC (location), ORG (organization) and MISC (miscellaneous). It is assumed that the entities are non-embedded, non-overlapping and annotated with exactly one label. The publicly available evaluation script `conlleval`[1] evaluates the standard measures – precision, recall and F-measure.

2.2 Czech Named Entity Corpus 1.0

In 2007, the Czech Named Entity Corpus 1.0 was annotated ([17], [20]). In this corpus, Czech NEs are classified into a set of 42 classes with very detailed characterization of the predicted entities. For example, instead of English LOC, the Czech local entities are further divided into `gc` (states), `gl` (nature areas / objects), `gq` (urban parts), `gs` (streets), `gu` (cities / towns), `gh` (hydronyms), `gp` (planets / cosmic objects), `gr` (territorial names), `gt` continents and `g_` (unspecified) in Czech.

The 42 fine-grained classes are merged into 7 super-classes (called "supertypes" in [11]), which are `a` (numbers in addresses), `g` (geographic items), `i` (institutions), `m` (media names), `o` (artifact names), `p` (personal names) and `t` (time expressions).

Furthermore, one entity may be labelled with one or more classes, e.g. `<oa<gu Santa Barbara>>`, where the location (`gu`) "Santa Barbara" appears in a TV document (`oa`) "Santa Barbara". Embedded entities are allowed, and frequently appearing patterns of named entities are also embedded in so called "containers", e.g. `<P<pf Jan> <ps Stráský>>`, where the first name "Jan" and last name "Stráský" are embedded in a name container `P`.

The fine-grained and possibly embedded classification makes the Czech named entity recognition task more complicated. Our system recognizes named entities described in [17], p. 32, Table 4.1, as well as most related work ([20], [11]).[2] An evaluation script which evaluates precision, recall and F-measure for all entities, one-word entities and two-word entities is available.[3]

3 Related Work

3.1 Systems for English

For the CoNLL-2003 shared task, the winning two systems were [7] and [3].[4] For English, Stanford Named Entity Recognizer ([6]) is available online.[5] The systems which published high scores on the CoNLL-2003 task include [18], [1], and to our knowledge, the best currently known results on this dataset were published in 2009 by [16] and reached 90.80 F-measure on the test portion of the data.

[1] `http://www.cnts.ua.ac.be/conll2000/chunking/output.html`

[2] We do not recognize number usages annotated in the second annotation round.

[3] `http://ufal.mff.cuni.cz/tectomt/releases/czech_named_entity_corpus_10/`

[4] The difference between the two systems was statistically insignificant.

[5] `http://nlp.stanford.edu/software/CRF-NER.shtml`

3.2 Systems for Czech

Together with the Czech Named Entity Corpus 1.0 ([17], [20]), a decision tree classifier was published. It achieved 62 F-measure on the embedded fine-grained classification and 68 F-measure on the embedded supertypes classification.

Another Czech named entity recognizer is [11] from 2009. On the Czech task, the authors achieved 68 F-measure for the embedded fine-grained classification and 71 F-measure on the embedded supertypes. The system used a combination of simple n-gram SVM-based recognizers.

In 2011, Konkol and Konopík ([10]) published a maximum-entropy based recognizer. They achieved 72.94 F-measure on the supertypes. The results for the fine-grained classification were not published.

4 Methods

4.1 System Overview

A simple overview of our named entity recognizer is described in Figure 4.1. The system is based on Maximum Entropy Markov Model (MEMM) and a Viterbi decoder decodes probabilities estimated by a maximum entropy classifier.

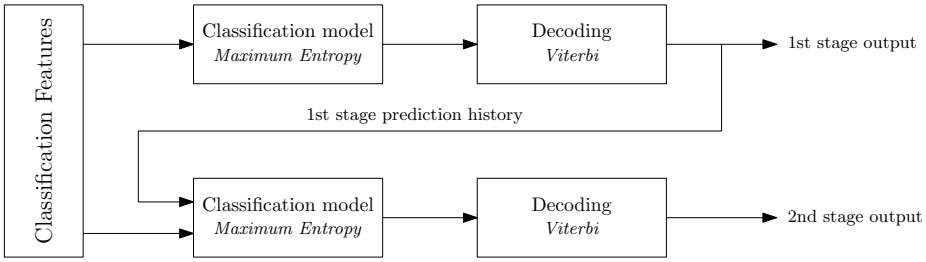

Fig. 1. System overview

First, maximum entropy model predicts for each word in a sentence the full probability distribution of its classes and positions with respect to an entity. Consequently, a global optimization via dynamic programming determines the optimal combination of classes and named entities chunks (lengths). This procedure deals with the most inner embedded entities and the system outputs one label per entity. Finally, the Czech system output is post-edited with four rules to add containers.[6]

The whole pipeline runs two times, utilizing the output from the first stage as additional classification features in the second stage.

[6] We automatically selected a subset of embedding patterns appearing in the training data by sequential adding the rule that increased F-measure the most. There are no such rules for English because the dataset does not contain embedded entities.

4.2 Maximum Entropy Classifier

In the first step, the maximum entropy classifier task is to predict for each word the named entity type and position within the entity. The positions are described with a BILOU scheme ([16]): B for multiword entity Beginning, I for Inside multiword entity, L for Last word of multiword entity, U for unit word entity and O for outside any entity. This scheme results in a large combination of predicted classes ($4 \times |C| + 1$, where $|C|$ is the number of classes, 4 is for B-X, I-X, L-X, U-X and $+1$ for O).

For the classifier training, we implemented our own gradient ascent parameter estimation.

4.3 Decoding

For decoding, global sequence decoders are often used, such as HMM ([15]) or CRF ([12]).

We decode the probabilities estimated by the maximum entropy model via dynamic programming. In our implementation of the Viterbi algorithm, we prune the impossible trellis transitions (e.g., once B-X starts, it can be followed either by I-X or L-X). Using this observation we can decode a whole sentence using dynamic programming with $\mathcal{O}(N \cdot C)$ complexity, where N is the number of words in the sentence.

Also, we were concerned with large growth of classes predicted by the maximum entropy classifier in the first step. With the full BILOU scheme, there are 17 classes for English and 169 classes for Czech. With the previous observation, we simplified the BILOU scheme from full B-X, I-X, L-X, U-X and O, to B-X, I, L, U-X and O. With this simplified scheme, the number of predicted classes is halved.

4.4 Classification Features

For maximum entropy classifier, we use a standard set of classification features: form, lemma, tag, chunk (only English) of current word and surrounding words in window ± 2, orthographic features (capitalization, punctuation, lowercase and uppercase form of the word), suffixes and prefixes of length 4 and regular expressions identifying possible year, date and time (in Czech). Feature selection was done by sequentially (manually) adding new classification features to the feature set; we retrained those features that have improved the classification based on development data. In English, we used forms in most of the classification features, while in Czech, we had to use lemmas because of data sparsity due to the Czech being a morphologically rich language.

Apart from the features based on the current word and its immediate vicinity, we tried to incorporate also global features. We use two-stage prediction, that is, we run our system two times in a row and in the second run, we use the predictions made in the first run. We used the information about the prediction of the previous and following five words and about the previous predictions of the candidate word in the preceding window of 500 words.

Named entity recognizers rely substantially on external knowledge. For English, we used 24 gazetteers of $1.8M$ items and for the Czech language, we used 17 gazetteers of $148K$ items. We collected both manually maintained gazetteers and automatically

retrieved gazetteers from the English and Czech Wikipedia ([9]). We did not parse the whole Wikipedia article content, we only listed the title in gazetteer when it was filed under an appropriate category (e.g. “people”, “births”, “cities”, etc.)

In the Czech morphology, lemmas are manually annotated with labels marking proper names, such as `Y` for given names, `S` for surnames and `G` for geographical names ([8], p. 121). These labels act as gazetteers built inside morphology.

Furthermore, we utilized Brown mutual information bigram clusters ([2], [13]), which we trained on Czech Wikipedia and downloaded for English.[7] We added these clusters respective to forms (English) and lemmas (Czech) and cluster prefixes of length 4, 6, 10 and 20 (see [16]) as new classification features.

4.5 Preprocessing and Other Experiments

We did not use the original morphological analysis annotation in the data and instead, we retagged both the English and the Czech data with Featurama tagger[8], based on average perceptron sequence labeling. The Czech data was lemmatized by Featurama and the English data with an algorithm by [14]. We chunked the English data with TagChunk ([5]).[9]

We also experimented with classifier combination. In English, we used the publicly available Stanford NER ([6]) and interpolated its output with our maximum entropy classifier predicted probability distribution just before the dynamic programming step. The probability distributions were interpolated using a linear combination in which the weight was discovered via grid search.[10] Our future work involves experiments with more English named entity recognizers. Unfortunately in Czech, the previously published named entity recognizers ([20], [11], [10]) are not available for such a combination approach.

5 Results and Discussion

We call “baseline” the simplest model where we used the common set of classification features in maximum entropy model, then decoded the probability distribution given by the classifier with dynamic programming and in Czech, post-edited the result with three automatically discovered rules.

Table 1 shows the effect of more sophisticated classification features or processing: (A) new tagging, lemmatization and chunking, (B) two stage prediction, (C) gazetteers, (D) Brown clusters, (E) linear combination with the Stanford NER. The experiments (A), (B), (C), (D) and (E) show the system improvement after adding the respective feature to the baseline. The last line of the table shows results after combining all features. All new features and preprocessing steps improved the system performance over the baseline and the gains were similar in both languages. In the Czech language, most of the impact of adding gazetteers (C) is formed by the manually annotated proper name

[7] `http://people.csail.mit.edu/maestro/papers/bllip-clusters.gz`
[8] `http://sourceforge.net/projects/featurama/`
[9] `http://www.umiacs.umd.edu/~hal/TagChunk/`
[10] Our maximum entropy classifier weight $= 10$, Stanford NER weight $= 3$.

labels in the morphology and the manually collected and Wikipedia extracted gazetteers did not yield substantial improvement.

Table 1. System development. The experiments (A), (B), (C), (D), (E) show F-measure gains over the baseline on the test portion of the English and Czech data.

	English	Czech
baseline	83.80	74.87
(A) new tags, lemmas and chunks	84.20	75.47
(B) two stage prediction	84.93	76.14
(C) gazetteers	86.20	76.15
(D) Brown clusters	85.88	76.67
(E) linear combination with Stanford NER	84.21	NA
all	89.16	79.23

Table 2 shows detailed results with precision, recall and F-measure for Czech one-word, two-word and all named entities for comparison with similar tables published in [20] and [11]. Table 3 compares the related work for Czech and English on the respective datasets.

Table 2. Detailed results for Czech language. The table shows results for one-word, two-word and all named entities. The three measures evaluated are precision (P), recall (R) and F-measure (F).

	All NEs			One-word NEs			Two-word NEs		
	P	R	F	P	R	F	P	R	F
Type:	84.46	74.61	79.23	87.70	79.97	83.66	81.85	77.10	79.40
Suptype:	88.27	78.00	82.82	92.07	84.00	87.85	84.12	79.24	81.60
Span:	91.56	82.56	86.83	94.00	87.90	90.85	90.28	86.09	88.13

6 Conclusions

We have presented a new named entity recognizer and evaluated it for Czech and English. We have reached 82.82 F-measure for the Czech language and significantly outperformed the existing Czech state of the art. For English, we achieved 89.16 F-measure. Our future work includes publicly releasing the recognizer and experimenting with named entity recognizer combination.

7 Acknowledgements

This work has been partially supported and has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of

Table 3. System comparison for English and Czech language (F-measure on test data).

Czech	Types	Supertypes
this work	**79.23**	**82.82**
Konkol and Konopík, 2011 ([10])	NA	72.94
Kravalová and Žabokrtský, 2009 ([11])	68.00	71.00
Ševčíková et al., 2007 ([20])	62.00	68.00

English	Test F-measure
Ratinov and Roth, 2009 ([16])	90.80
Suzuki and Isozaki, 2008 ([18])	89.92
Ando and Zhang, 2005 ([1])	89.31
this work	**89.16**
Florian et al. 2003 ([7])	88.76
Chieu and Ng, 2003 ([3])	88.31
Finkel et al. 2005 ([6], Stanford parser)	86.86

Education of the Czech Republic (project LM2010013). This work was also partially supported by SVV project number 267 314. We are grateful to the reviewers of this paper for comments which helped us to improve the paper.

References

1. Ando, R.K., Zhang, T.: A high-performance semi-supervised learning method for text chunking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 1–9. ACL '05, Association for Computational Linguistics (2005)
2. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 467–479 (Dec 1992)
3. Chieu, H.L., Ng, H.T.: Named entity recognition with a maximum entropy approach. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4. pp. 160–163. CONLL '03, Association for Computational Linguistics (2003)
4. Chinchor, N.A.: Proceedings of the Seventh Message Understanding Conference (MUC-7) Named Entity Task Definition. In: Proceedings of the Seventh Message Understanding Conference (MUC-7). p. 21 pages (April 1998)
5. Daumé III, H., Marcu, D.: Learning as search optimization: approximate large margin methods for structured prediction. In: Proceedings of the 22nd international conference on Machine learning. pp. 169–176. ICML '05, ACM (2005)
6. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. ACL '05, Association for Computational Linguistics (2005)
7. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named Entity Recognition through Classifier Combination. In: Proceedings of CoNLL-2003. pp. 168–171. Edmonton, Canada (2003)
8. Hajič, J.: Disambiguation of Rich Inflection: Computational Morphology of Czech. Karolinum Press (2004), `http://books.google.cz/books?id=sB63AAAACAAJ`
9. Kazama, J., Torisawa, K.: 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 (EMNLP-CoNLL). pp. 698–707. Association for Computational Linguistics (June 2007)
10. Konkol, M., Konopík, M.: Maximum Entropy Named Entity Recognition for Czech Language. In: Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 6836, pp. 203–210. Springer Berlin Heidelberg (2011)
11. Kravalová, J., Žabokrtský, Z.: Czech named entity corpus and SVM-based recognizer. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. pp. 194–201. NEWS '09, Association for Computational Linguistics (2009)
12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc. (2001)
13. Liang, P.: Semi-Supervised Learning for Natural Language. Master's thesis, Massachusetts Institute of Technology (2005)
14. Popel, M.: Ways to Improve the Quality of English-Czech Machine Translation. Master's thesis, ÚFAL, MFF UK, Prague, Czech Republic (2009)
15. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
16. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147–155. Association for Computational Linguistics (2009)
17. Ševčíková, M., Žabokrtský, Z., Krůza, O.: Zpracování pojmenovaných entit v českých textech. Tech. Rep. TR-2007-36 (2007)
18. Suzuki, J., Isozaki, H.: Semi-Supervised Sequential Labeling and Segmentation using Gigaword Scale Unlabeled Data. Computational Linguistics (June), 665–673 (2008)
19. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003. pp. 142–147. Edmonton, Canada (2003)
20. Ševčíková, M., Žabokrtský, Z., Krůza, O.: Named entities in Czech: annotating data and developing NE tagger. In: Proceedings of the 10th international conference on Text, speech and dialogue. pp. 188–195. TSD'07, Springer-Verlag (2007)