CRF-based Czech Named Entity Recognizer and Consolidation of Czech NER Research

Michal Konkol and Miloslav Konopík

Department of Computer Science and Engineering
Faculty of Applied Sciences
University of West Bohemia
Univerzitní 8, 306 14 Plzeň, Czech Republic
nlp.kiv.zcu.cz
{konkol,konopik}@kiv.zcu.cz

Abstract. In this paper, we present our effort to consolidate and push further the named entity recognition (NER) research for the Czech language. The research in Czech is based upon a non-standard basis. Some systems are constructed to provide hierarchical outputs whereas the rests give flat entities. Direct comparison among these system is therefore impossible. Our first goal is to tackle this issue. We build our own NER system based upon conditional random fields (CRF) model. It is constructed to output either flat or hierarchical named entities thus enabling an evaluation with all the known systems for Czech language. We show a 3.5 – 11% absolute performance increase when compared to previously published results. As a last step we put our system in the context of the research for other languages. We show results for English, Spanish and Dutch corpora. We can conclude that our system provides solid results when compared to the foreign state of the art.

Keywords: named entity recognition, conditional random fields, Czech Named Entity Corpus

1 Introduction

Named entity recognition (NER) has proven to be an important preprocessing step for many natural language processing tasks including question answering [1], machine translation [2] or summarization [3]. The purpose of NER is to identify text phrases which carry a particular predefined meaning. The phrases are classified to a given set of classes, e.g. person, organization or location.

In this paper, we will mainly focus on NER used for texts in the Czech language. Czech is highly inflectional Slavic language with free word order. All published results [4–7] were evaluated on the Czech Named Entity Corpus [4], which is the only publicly available corpus for Czech.

Four systems were presented for Czech so far. The first two systems are similar in the architecture, but they use different methods (decision trees [4] and support vector machines [5]) and slightly different feature set. Both systems use non-standard architecture compared to conventional NER systems. They use different classifiers for one-word

named entities (NEs), two-word NEs and NEs with more words. The results of these classifiers are combined to a final result that contains structured tags. The third system [6] can be viewed as a traditional approach to NER using a maximum entropy classifier. The fourth system [7] is based on conditional random fields. The output of the latter two is not structured. This difference in the output is the reason for incompatible evaluation metrics making the direct comparison of these systems impossible.

We will present our new NER system based on conditional random fields. Our system was directly compared with all previously published systems for Czech using the original evaluation scripts and outperformed them.

We have also transformed the Czech Named Entity Corpus into CoNLL format and evaluated our system using standard CoNLL evaluation. Then we have evaluated our system on the English, Spanish and Dutch CoNLL corpora and compared the results between languages.

The paper is structured as follows. The second section describes basic properties of the Czech Named Entity Corpus. Section 3 is focused on the evaluation metrics. Section 4 presents our CRF-based NER system. Section 5 provides results of our experiments and compares them with other systems. The last section summarizes our results and provides some ideas for future research.

2 Czech Named Entity Corpus

The Czech Named Entity Corpus (CNEC) [4] is created from the Czech National Corpus. Sentences were chosen by heuristic to contain more entities than completely random sentences. The NE annotation scheme uses hierarchical types with two levels. The first level is denoted as *supertypes* and the second level as *types*. The entities in the corpus can be embedded and form structured tags. We will use the term *top level entities* for entities that are not embedded. The corpus was created in three rounds. In the first round, 7 supertypes and 42 types were used. The subsequent rounds used an extended set with 10 supertypes and 62 types. Unfortunately, this extension was not used retrospectively for the first part. The evaluation in [4, 5] is done only on the not-extended set, which is consistent across the whole corpus, but no information about this problem was given in the papers. The following example demonstrates the corpus. The NE classes are encoded using two letters, first for supertypes, second for types.

```
Generální ředitel <if Škody <gu Plzeň>> <P<pf Lubomír>
      <ps Soudek>> je stále největším akcionářem
         plzeňského strojírenského gigantu .
```

The CNEC consists of approximately 6000 sentences with 150.000 tokens. The corpus can be considered small when compared to CoNLL corpora [8, 9] which have over 300.000 tokens. While the CoNLL corpora use 4 NE types, CNEC uses 7 (10) resp. 42 (62) types, which makes the data much sparser and therefore the NER task harder.

We have transformed the corpus into the standard CoNLL format enriched with lemmas and POS tags.[1] Only supertypes and top level entities are used. This transformed

[1] http://nlp.kiv.zcu.cz

corpus is used for our CoNLL evaluation and can be easily used by multilingual NER systems for their evaluation on Czech, because these systems usually work with the CoNLL corpus format.

3 Evaluation metrics

NER is a typical example of multiclass classification problem, where the NE classes are highly skewed compared to the non-entity class. For such tasks precision, recall and f-measure are the metrics of choice. Unfortunately, it is possible to define correct answer in various ways, especially in corpora with hierarchical types and structured annotations.

3.1 Structured metric

The structured metric is used in [4, 5]. All entities (including the embedded) from the gold data set are used for the evaluation. Conventional NER systems, however, do not allow embedding. Under such conditions, their performance may seem worse then it actually is, because the embedded entities (which form an indispensable portion) are unreachable for them.

In the structured metric, an entity marked by a system is considered correct if the span and type are the same as in the gold data. All the levels in the structured annotations have the same weights. In [4, 5], the results are given for one-word NEs, two-word NEs and all NEs. We believe, that it would be better to provide results for the top level NEs and then gradually for lower levels, because it is hard to interpret usefulness of entities based on number of words and the results can be hardly compared with other systems and other languages.

3.2 Word by word metric

The word by word metric is used in [6]. It does not allow embedding, so only the top types from the corpus are used for the evaluation. In this metric, each word is a self-standing unit with only one type. For the NE `<if Škody <gu Plzeň>>` spanning over two words we would have two independent objects with one type `if`. The NER system output is transformed in individual objects as well. For each word, the gold data type is compared with the type of the system and is considered correct if both are the same. The problem of this metric is that it ignores the presence of multi-word entities.

3.3 Standard CoNLL metric

The standard CoNLL metric is used on the CoNLL-2002 and CoNLL-2003 conferences [8, 9] and it is used in the majority of papers as well. It does not allow embedding, so the embedded entities must be ignored. The output NE is considered correct, only if its span and type is exactly the same as the span and type in the gold data. It is equivalent to the structured metric if only the top level of structured annotations is used.

It was also used in [7], but they removed structure of the tags by adding higher priority to embedded entities, e.g. `<if Škody <gu Plzeň>>` became `<if Škody> <gu Plzeň>`. This choice is questionable, because the top level is significantly more useful. It can also produce weird entity sequences, e.g. `<ic <ps Smith> & <ps Delvin> Ltd.>` produces `<ps Smith> <ic &> <ps Delvin> <ic Ltd.>`.

3.4 Lenient metric extension

The lenient metric extension is based on the GATE evaluator.[2] It is used as a supplement to the CoNLL evaluation metric. It adds the information about cases, where the system correctly guessed the type, but not the span.

4 NER system

4.1 Conditional random fields

Conditional random fields (CRF) are undirected graphical models [10]. Simple chain CRF are currently the state-of-the-art method for NER [11]. They model probability distribution $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{x}$ is sequence of words and $\mathbf{y}$ is sequence of labels. The modeled probability distribution $p(\mathbf{y}|\mathbf{x})$ has the following form.

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{j=1}^{N} \exp \sum_{i=0}^{M} \lambda_i f_i(y_{j-1}, y_j, x_j) \tag{1}$$

N is number of tokens and M number of features. The parameters λ_i are estimated using L-BFGS method [12]. Gaussian prior is used for regularization [13]. Typically, the feature functions f_i are binary and are similar to the following example.

$$f_i(y_{j-1}, y_j, x_j) = \begin{cases} 1 & \text{if } y_{j-1} \text{ is CITY, } y_j \text{ is CITY and } x_j \text{ is 'York'} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

4.2 Feature set

Our CRF system works with the feature set described in this section. All the features use a window of two previous and two next words.

The feature set does not include morphological tags, which are usually used in language dependent systems. We have tried some variations of their usage, but none of them improved the results.

Lemma – The PDT [14] lemmatizer is used to get the lemmas. The lemma has to appear at least twice to be used as a feature.

Affixes – Prefixes and suffixes of length from 2 to 4 were used. Both affixes are taken from lemmas and have to appear at least 5 times.

[2] `http://gate.ac.uk/sale/tao/splitch10.html#x14-26900010.2`

Bag of words – Bag of words is similar to lemma feature, but ignores order of the lemmas in the window.

Bi-grams – Bi-grams of lemmas have to appear at least twice to be used. Higher level n-grams did not improve the results, probably due to the size of the corpora.

Orthographic features – Standard orthographic features. Specifically: *firstLetterUpper; allUpper; mixedCaps; contains ., ', –, &; upperWithDot; acronym*

Orthographic patterns – Orthographic pattern [15] is a lemma, where every lower case letter is rewritten to `a`, upper case letter to `A`, number to `1` and symbol to `-`. We are using these patterns directly and also compressed, where every sequence of the same type is represented by only one character. The uncompressed pattern has to appear at least 5 times, compressed 20 times.

Orthographic word pattern – Combines compressed orthographic patterns (by joining them) of all the words in the window. This pattern has to appear at least 5 times.

Gazetteers – We use various gazetteers. The majority of them are from publicly available sources (e.g. list of cities provided by the Czech Ministry of Regional Development).

Table 1. Comparison of Czech NER systems.

	structured	word by word	CoNLL(inner)	CoNLL	Lenient
DecTree [4]	68%	—	—	—	—
SVM [5]	71%	—	—	—	—
MaxEnt [6]	—	72.94%	—	—	—
CRF [7]	—	—	58.00%	—	—
Our system	79%	75.61%	62.07%	74.08%	79.50%

5 Experiments

5.1 Czech systems comparison

Our priority for the Czech language is to compare our system with all the other Czech systems using their evaluation scripts. All the systems use Czech lemmatiaztion and morphological analysis on a similar level, but use different gazetteers.

For comparison with [4, 5], we have altered the system slightly to get structured output. We have added one more CRF model with the same features, but trained only on the embedded entities. For all the multiword entities found by the original CRF model, the second model tries to find entities inside.

We have also evaluated our system using the CoNLL evaluation. Comparison between the systems is given in table 1. Detailed results for individual classes are given in table 2.

Table 2. Detailed CoNLL evaluation on the Czech Named Entity Corpus of our system.

	Lenient			Strict		
	Precission	Recall	F-measure	Precission	Recall	F-measure
Time	92.11%	88.62%	90.33%	88.17%	85.24%	86.68%
Geography	80.13%	80.65%	80.39%	77.10%	77.61%	77.37%
Person	83.62%	88.41%	85.95%	79.82%	84.91%	82.29%
Address	50.00%	70.00%	58.33%	50.00%	70.00%	58.33%
Media	76.92%	30.30%	43.48%	69.23%	27.27%	39.13%
Other	67.70%	66.55%	67.13%	56.25%	55.67%	55.96%
Institution	79.95%	70.06%	74.68%	71.52%	63.05%	67.02%
Total	80.46%	78.56%	79.50%	74.81%	73.38%	74.08%

Our system outperforms the other systems by a large margin. An interesting fact is, that marking the embedded entities (using the original structural evaluation) is actually easier task than finding only the not-embedded entities.

5.2 Comparison among languages

We have evaluated our system on the CoNLL corpora [8, 9] for other languages. The CoNLL corpora use four types of entities – organizations (ORG), persons (PER), locations (LOC) and miscellaneous (MISC). The results for English, Spanish and Dutch are shown in table 3. Our system is not tweaked for these languages. Lemmas are replaced with words and gazetteers are removed, because they are unavailable for some languages. For Spanish and Dutch we compare the results with the best system [16] from the original CoNLL conference. For English we report the best results from the conference [17] along with the overall best result so far [11]. For the latter we present two results – a language independent baseline (LIB) and a complete system.

The results of our system are obviously worse then from systems tweaked for given languages, because we have not tweaked our feature set (features and their parameters) for given languages at all. We still believe that this comparison shows the state-of-the-art quality of our system. It is also possible to roughly compare Czech with the other languages. We can see, that Czech has the lowest f-measure, even though our system uses some Czech language specific features for the Czech corpus. This can be partly caused by the smaller size of the Czech corpus and more entity types, but we believe that the major reason is the difficulty of Czech NER task.

Table 3. Results for the (a) Spanish, (b) Dutch and (c) English CoNLL corpus. The comparison with other systems is shown at the bottom.

(a)

	Precission	Recall	F-measure
ORG	74.54%	64.40%	69.10%
PER	80.97%	82.51%	81.73%
LOC	81.54%	76.49%	78.93%
MISC	75.71%	71.19%	73.38%
Total	78.18%	73.86%	75.97%
Best [16]	77.83%	76.29%	77.05%

(b)

	Precission	Recall	F-measure
ORG	79.48%	80.79%	80.13%
PER	86.35%	89.52%	87.91%
LOC	80.21%	77.77%	78.97%
MISC	62.58%	54.12%	58.04%
Total	79.77%	79.12%	79.44%
Best [16]	81.38%	81.40%	81.39%

(c)

	Precission	Recall	F-measure
ORG	81.15%	75.68%	78.32%
PER	87.28%	86.58%	86.93%
LOC	87.81%	86.81%	87.31%
MISC	78.55%	74.07%	76.24%
Total	84.64%	81.89%	83.24%
Best CoNLL [17]	88.99%	88.54%	88.76%
Best LIB [11]	—	—	83.78%
Best [11]	—	—	90.90%

6 Conclusion and future work

We have developed new CRF-based Czech NER system which outperforms all other published systems by a large margin. Using the standard CoNLL evaluation metric we achieved f-measure 74.08%. We have also evaluated the system on English, Spanish and Dutch and discussed the differences in performance. The whole paper has set the new state of the art for Czech language.

We believe that the results can be further improved by feature extraction. Another possibility is to combine various machine learning methods into one model.

Acknowledgements

This work was supported by grant no. SGS-2010-028, by grant no. SGS-2013-029 Advanced computing and information systems, by the European Regional Development Fund (ERDF). Access to the MetaCentrum computing facilities provided under the program “Projects of Large Infrastructure for Research, Development, and Innovations” LM2010005, funded by the Ministry of Education, Youth, and Sports of the Czech Republic, is highly appreciated.

References

1. Mollá, D., Van Zaanen, M., Smith, D.: Named entity recognition for question answering. (2006)
2. Babych, B., Hartley, A.: Improving machine translation quality with automatic named entity recognition. In: Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. EAMT '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 1–8
3. Nobata, C., Sekine, S., Isahara, H., Grishman, R.: Summarization System Integrated with Named Entity Tagging and IE pattern Discovery. In: LREC. (2002)
4. Ševčíková, M., Žabokrtsky, 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. TSD'07, Berlin, Heidelberg, Springer-Verlag (2007) 188–195
5. Kravalová, J., Žabokrtský, Z.: Czech named entity corpus and SVM-based recognizer. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. NEWS '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 194–201
6. Konkol, M., Konopík, M.: Maximum entropy named entity recognition for Czech language. In: Proceedings of the 14th international conference on Text, speech and dialogue. TSD'11, Berlin, Heidelberg, Springer-Verlag (2011) 203–210
7. Král, P.: Features for Named Entity Recognition in Czech Language. In: KEOD. (2011) 437–441
8. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: proceedings of the 6th conference on Natural language learning - Volume 20. COLING-02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 1–4
9. Tjong Kim Sang, E.F., 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. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 142–147
10. 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. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
11. Lin, D., Wu, X.: Phrase clustering for discriminative learning. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. ACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 1030–1038
12. Nocedal, J.: Updating Quasi-Newton Matrices with Limited Storage. Mathematics of Computation **35**(151) (1980) 773–782
13. Chen, S.F., Rosenfeld, R.: A gaussian prior for smoothing maximum entropy models (1999)
14. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Charles University Press, Prague, Czech Republic (2004)
15. Ciaramita, M., Altun, Y.: Named-Entity Recognition in Novel Domains with External Lexical Knowledge. (2005)
16. Carreras, X., Màrques, L., Padró, L.: (named entity extraction using adaboost). In: Proceedings of CoNLL-2002, Taipei, Taiwan (2002) 167–170
17. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: (named entity recognition through classifier combination). In Daelemans, W., Osborne, M., eds.: Proceedings of CoNLL-2003, Edmonton, Canada (2003) 168–171