FEATURES FOR NAMED ENTITY RECOGNITION IN CZECH LANGUAGE

Pavel Král
Department of Computer Science and Engineering, University of West Bohemia, Plzeň, Czech Republic

Keywords: Conditional Random Fields (CRFs), Czech News Agency, Named entity recognition.

Abstract: This paper deals with Named Entity Recognition (NER). Our work focuses on the application for the Czech News Agency (ČTK). We propose and implement a Czech NER system that facilitates the data searching from the ČTK text news databases. The choice of the feature set is crucial for the NER task. The main contribution of this work is thus to propose and evaluate some different features for the named entity recognition and to create an "optimal" set of features. We use Conditional Random Fields (CRFs) as a classifier. Our system is tested on a Czech NER corpus with nine main named entity classes. We reached 58% of the F-measure with the best feature set which is sufficient for our target application.

1 INTRODUCTION

Named Entity Recognition has been identified as a main research topic for automatic information retrieval around 1996 (Grishman and Sundheim, 1996) and can be seen as the first step of information extraction. The objective is the identification of expressions with special meaning such as person names, organizations, times, monetary values, etc.

The named entity recognition task is composed of the three main steps:

1. Feature set creation;
2. NE models building;
3. NE classification.

The feature set is beneficial for a successful NE recognition (Ekbal et al., 2010). We thus apply some features that are evaluated positively in other languages to the Czech language. Our final feature set will be composed of the most promising ones. Conditional Random Fields (CRFs) are then used for named entity models creation and NE recognition.

The outcomes of this work will be used to improve the ČTK search engine. The current system allows only the key-words search in the text news databases. Named entity recognition will be integrated to the system such facilitates to analyse well some previously ambiguous requests and the system answers will be more accurate.

This paper is organized as follows. The next section presents a short review of the NER approaches. Section 3 describes the CRF NER approach with the particular focus on the choice of features. Section 4 evaluates the proposed approach on a Czech NE corpus. In the last section, we discuss the results and propose some future research directions.

2 RELATED WORK

Named Entity Recognition has been progressively formalized during the Message Understanding Conference series (MUC[1]) until 1998, and later on in the IREX campaign in Japan (Satoshi and Hitoshi, 2000), HAREM campaign in Portuguese (Santos et al., 2006), ESTER campaign in France (Gravier, 2005), but more importantly in the ACE[2] and CoNLL (Sang and Erik, 2002) campaigns for English.

There are two broad approaches. 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 as Hidden Markov Models (Zhou and Su, 2002), Maximum Entropy Models (Curran and Clark, 2003), Conditional Random Fields (McCallum and Li, 2003), Support Vector Machines (Isozaki and Kazawa, 2002) and other. There are also hybrid systems combining more of mentioned methods (Kozareva et al., 2007).

[1] http://www-nlpir.nist.gov/related_projects/muc/
[2] http://www.itl.nist.gov/iad/mig/tests/ace

Král P..
FEATURES FOR NAMED ENTITY RECOGNITION IN CZECH LANGUAGE.
DOI: 10.5220/0003660104370441
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 437-441
ISBN: 978-989-8425-80-5

Despite these efforts, the Named Entity recognition task still faces several fundamental issues, in particular concerning the definition of the task, of named entities and of their segmentation.

3 NAMED ENTITY RECOGNITION APPROACH

3.1 Features

Our feature choice was inspired from the works for the other inflecting languages as for Bulgarian in (Georgiev et al., 2009), for Hindi in (Ekbal and Bandyopadhyay, 2010) and for Arabic in (Abdul Hamid and Darwish, 2010). Because few works exist for the Czech (Kravalová and Žabokrtský, 2009), only few of our features are similar to those from the known Czech studies.

We distinguish two types of features: language independent and language dependent ones. The main difference is that it is possible to use language independent features without any modification in any language, however language dependant features usually perform only in one language. Note, that feature quality for one language, do not guarantees the same properties in the other language.

Language dependant features are based on the properties of a given language such as syntax, words belonging to some dictionary, etc.

3.1.1 Language Independent Features

The evaluated language independent features are described below:

- **Word and Surrounding Words** (w_i): current word w_i and words that are placed in the neighbourhood of this current word.
- **First Word** (fw): *true*, when the current word is the first one in the sentence, *false* otherwise.
- **Word Length** ($wLen_i$): number of characters in the word w_i.
- **Is Word Longer than?** ($wLenB$): *true* , when the word is longer than a given threshold *B*, *false* otherwise.
- **Few Occurrence Words** ($infreq$): *true*, when the word occurs fewer than a given threshold, *false* otherwise.
- **Previous Word Tag** (tag): NE tag of the previous word obtained during the recognition.
- **Word Prefix** (p_i): i characters from the beginning of the word.
- **Word Suffix** (s_i): i characters from the end of the word.
- **Surrounding Word Prefix** ($c_j p_i$): word character prefix p_i of the surrounding words to the word at position j.
- **Surrounding Word Suffix** ($c_j s_i$): word character suffix s_i of the surrounding words to the word at position j.
- **Is Number?** (dig): word (or its part) is composed of digits.
- **Word n-gram** (N_i): n-grams created from the words.
- **Character n-gram** (chN_i): n-grams created from the characters.

3.1.2 Language Dependent Features

The first feature subset is based on the manual defined dictionaries. The most important ones are summarized further:

- **World Capitals** ($capD$): available at http://www.mestasveta.cz/hlavni-mesta.
- **World Countries** ($statesD$): available at http://www.mifin.cz/WU-seznam_statu.doc.
- **Currencies** ($currencyD$): available at http://www.finance.cz/bankovnictvi/meny/vse/.
- **Days of the Week** ($daysD$).
- **Months of the Year** ($monthsD$).
- **Companies** ($compD$): a list provided by the ČTK.
- **Czech Cities and Villages** ($CZcityD$): available at http:// www2.czso.cz/ csu/ 2009edicniplan.nsf/ p/ 1302-09.
- **Numerals** ($numsD$).
- **Organizations** ($orgD$).
- **Czech First and Last Names** ($namesD$): available at http://www.mvcr.cz/clanek/cetnost-jmena-prijmeni-722752.aspx?q=Y2hudW09Mw%3d%3d and at http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx?q=Y2hudW09Mg%3d%3d.
- **Physical Units** ($unitD$): available at http://www.converter.cz.
- **Word Prefixes** ($c_j pD_i$) : list of word prefixes (features are created from prefixes p_i of the surrounding words to the word at position j).
- **Word Suffixes** ($c_j sD_i$) : list of word suffixes (features are created from prefixes p_i of the surrounding words to the word at position j).

The second subset is based on the Czech syntactic properties as follows:

- **Word Lemma** (*lemma*): base form of the word (e.g. the lemma of the English word "walking" is "walk").
- **Part of Speech Tag** (POS_i): main part of the part of speech (POS) tag of the word w_i (Jan Hajic, 2005).
- **Sub Category of the POS-tag** (*sPOS*): this sub category specifies the main POS-tag (Jan Hajic, 2005).
- **Verbal Voice** ($voice_i$).
- **Case** (mc_i): to express one or more particular syntactic relationships of the word w_i to other words in a sentence.
- **Singular vs. Plural** (mn_i): *true*, when the word w_i is singular, *false* otherwise.

3.2 CRF Classifier

We use conditional random fields as a classifier due to their performances superior to the other classification models on several NLP tasks. Laferty et al. shown in (Lafferty et al., 2001) that CRFs avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models on a part-of-speech tagging task. Favre et al. compared in (Favre et al., 2009) three sequence models: Hidden-Event Language Model (HELM), factored-HELM and Conditional Random Fields (CRF) for comma prediction. They have shown that the best results have been obtained with CRF, although CRFs may not scale easily to large databases.

We use thus CRFs to calculate the conditional probability $P(NE|f_1, f_2, .., f_n)$ where $F =< f_1, f_2, .., f_n >$ represents the features sequence and NE is the sequence of the named entity classes.

4 EXPERIMENTS

We assume that tokenization task is done. Text is thus already split into the basic text units, *tokens*, that correspond mostly to the words.

We use a CRF classification algorithm of the LingPipe[3] tool as a background of our system. The input of this tool have been modified in order to accept the previously described features. Numerous experiments have been realized. However, only these that have a positive impact for the NER feature set creation are reported in this paper.

[3]http://alias-i.com/lingpipe/

The feature selection algorithm was as follows:

1. Start (to compose the basic feature set, i.e. words only);
2. Add one feature into the feature set;
3. Train CRF models and perform NE recognition;
4. If F-measure value is increasing, keep this feature; remove it otherwise;
5. If F-measure is smaller than the value given, go to step 2; finish the algorithm otherwise.

4.1 Corpus

Czech Named Entity Corpus 1.0 (Kravalová et al., 2009), which is composed of some text samples from the Czech National Corpus, is used to validate the proposed methods. It was created at the Institute of Formal and Applied Linguistics of Faculty of Mathematics and Physics at the Charles University in Prague. It has been labelled manually with 12 main named entity classes and with 61 sub-classes. The whole corpus containing 5866 sentences (such represents 150451 words) is randomly divided into two disjoint sets, 500 sentences for testing and the rest for training.

The corpus structure is detailed in Table 1.

Table 1: Distribution of the 12 main named entity classes in the corpus.

Tag description	Tag	Sent. no.
Number as a part of the address	a	195
Part of bibliographic information	c	153
Geographic names	g	2353
Institution names	i	1380
Media names	m	174
Number with a specific meaning	n	537
Object names	o	1196
Person names	p	2710
Ordinal number, count number	q	1047
Time values	t	1427
Abbreviations	s	747
Foreign language words	f	373
Not specified	?	231

We used F-measure as an evaluation metric of our experiments. 9 main named entity classes related to our application is recognized.

4.2 Language Independent Features

Table 2 relates the language independent feature sets with the experiment numbers. The F-measure of these experiments is reported in Table 3.

New features included into the experiments no. 2, 3 and 4 don't improve F-measure. Therefore, these

Table 2: Composition of the evaluated language independent feature sets with the corresponding experiment numbers.

Exp. no.	Features
1	w_i $(i=-1,..,1)$, fw
2	w_i $(i=-1,..,1)$, fw, $wLenB$ $(B=4)$
3	w_i $(i=-2,..,2)$, fw, $wLenB$ $(B=3)$
4	w_i $(i=-3,..,3)$, fw, $wLenB$ $(B=3)$
5	w_i $(i=-1,..,1)$, fw, $wLen_i$ $(i=-2,..,2)$, $infreq$
6	w_i $(i=-1,..,1)$, fw, $wLen_i$ $(i=-2,..,2)$, $infreq$, tag
7	w_i $(i=-1,..,1)$, fw, $wLen_i$ $(i=-2,..,2)$, $infreq$, tag, p_i $(i=1,..,4)$, s_i $(i=1,..,4)$
8	w_i $(i=-1,..,1)$, fw, $wLen_i$ $(i=-2,..,2)$, $infreq$, tag, p_i $(i=1,..,3)$, s_i $(i=1,..,3)$
9	w_i $(i=-1,..,1)$, fw, $wLen_i$ $(i=-2,..,2)$, $infreq$, tag, c_jp_i $(j=-1,..,1, i=0,..,3)$, c_js_i $(j=-1,..,1, i=0,..,3)$
10	w_i $(i=-1,..,1)$, fw, $wLen_i$ $(i=-2,..,2)$, $infreq$, tag, p_i $(i=1,..,3)$, s_i $(i=1,..,3)$, dig

Table 3: F-measure of the NER experiments with the language independent features in %.

	F-measure of the different NEs in [%]									
No.	**a**	**c**	**g**	**i**	**n**	**o**	**p**	**q**	**t**	**all**
1	13.3	45.7	44.9	34.7	17.8	36.4	48.2	48.3	62.9	44.0
2	13.2	47.9	44.8	34.8	22.4	37.6	48.4	44.1	62.8	44.0
3	13.2	45.7	44.0	33.3	20.6	37.7	48.5	46.7	62.2	43.7
4	17.4	46.4	42.5	33.9	20.4	39.9	46.2	47.1	60.8	43.0
5	21.7	47.9	47.2	35.6	21.4	42.2	51.4	49.3	65.1	46.0
6	31.3	58.0	43.6	32.8	23.3	44.6	49.4	47.7	64.2	46.1
7	36.2	62.2	**59.2**	**42.4**	**30.1**	**53.2**	**65.8**	54.7	**73.4**	**58.0**
8	36.2	**69.2**	58.7	40.6	28.6	52.9	64.5	55.8	72.5	57.4
9	29.4	71.8	44.9	34.0	18.6	47.3	51.1	47.7	66.3	47.5
10	**36.7**	61.4	58.8	40.1	29.9	53.9	64.4	**54.9**	72.1	57.3

features are not further used. The best F-measure is 73.4% and have been obtained for the class *t* (time) in the 7^{th} experiment. This experiment contains also most of the best results (for six classes). Its global F-measure is 58% which represents a maximal value. Additional features (see experiments no. 8, 9 and 10) don't improve the recognition F-measure. The final feature set is thus chosen from experiment no. 7.

This table shows further that F-measure values for the different NEs differ significantly. This fact may be due to two reasons:

1. There are not enough occurrences of some NEs in the corpus for the correct estimation of the models (classes with enough NE occurrences are generally recognized better that few occurrence classes);
2. The proposed feature set in not discriminating enough for all NEs.

4.3 Language Dependent Features

The previously created language independent feature set (see 7^{th} experiment at Table 3) is progressively completed by the language dependent features. Table 4 relates the features with the experiment numbers. Note that language independent features are similar for all experiments and are thus not reported in this table. The F-measure of these experiments is reported in Table 5.

Table 4: Composition of the evaluated language dependent feature sets with the corresponding experiment numbers.

Exp. no.	Features
1	c_jpD_i $(j=-2,..,2, i=1,..,3)$, c_jsD_i $(j=-2,..,2, i=1,..,3)$, $numsD$, POS_i $(i=-1,..,1)$
2	$numsD$, POS_i $(i=-1,..,1)$, $lemma$
3	$voice_i$ $(i=-2,..,2)$
4	$sPOS_i$ $(i=-1,..,1)$
5	mc_i $(i=-2,..,2)$, mn_i $(i=-2,..,2)$
6	mc_i $(i=-3,..,3)$, mn_i $(i=-3,..,3)$, $orgD$
7	mc_i $(i=-3,..,3)$, mn_i $(i=-3,..,3)$, $daysD$, $monthsD$
8	mc_i $(i=-3,..,3)$, mn_i $(i=-3,..,3)$, $namesD$
9	mc_i $(i=-3,..,3)$, mn_i $(i=-3,..,3)$, $namesD$, $compD$, $CZcityD$, $statesD$
10	mc_i $(i=-3,..,3)$, mn_i $(i=-3,..,3)$, $numsD$, $currencyD$, $unitD$

Table 5: F-measure of the NER experiments with the added language dependent features in %.

	F-measure of the different NEs in [%]									
No.	**a**	**c**	**g**	**i**	**n**	**o**	**p**	**q**	**t**	**all**
1	29.3	64.1	57.8	40.6	27.8	53.7	65.5	55.1	72.6	57.4
2	36.4	**69.1**	58.3	**41.4**	31.4	53.1	65.7	54.8	72.6	57.7
3	34.8	61.4	58.6	39.5	30.3	54.2	64.6	54.0	72.9	57.2
4	36.2	65.1	**58.7**	40.5	31.8	53.1	64.3	54.0	72.7	**58.4**
5	31.7	65.1	58.0	40.9	33.3	53.8	**66.3**	55.2	72.2	57.9
6	26.6	68.3	57.7	39.3	33.3	53.5	64.6	55.3	72.5	57.2
7	26.8	68.3	58.0	39.6	32.8	54.0	65.0	**55.5**	72.6	57.5
8	32.5	67.5	57.5	39.9	**34.4**	53.2	65.4	54.4	**72.9**	57.5
9	**36.7**	67.5	57.8	40.6	**34.4**	53.3	65.6	55.1	72.6	57.8
10	35.3	68.3	57.9	39.4	33.3	**55.0**	64.6	**55.5**	72.9	57.7

This table shows that the increase of the F-measure is very small. The best obtained F-measure is 58.4% by the $sPOS_i$ $(i=-1,..,1)$ feature set. Moreover, the majority of the features degrades the results. This fact may be due to the character of our NE corpus which is very small and contains less NEs from the defined dictionaries.

5 CONCLUSIONS AND PERSPECTIVES

The main objective of this work is to propose and implement a Czech NER system that facilitates the data searching from the ČTK text news databases. We presented the CRF based method for automatic named entity recognition. The particular focus was on the NE feature selection. We have proved that the feature choice plays a crucial role for the NE recognition. We have also shown that the language independent features are much more important than the language dependent ones. The best obtained F-measure is 58.4% which represents the increase of F-measure 14.4% over the baseline in absolute value.

The first perspective consists of evaluation of the system over the ČTK data. This includes the data annotation, because the current text data is not annotated by named entities. We further propose to integrate in the system new information coming from the other knowledge sources, in particular from the syntactic parsing module. Detailed syntactic features are most often underexploited. However they bring clearly additional information to the NE recognition.

REFERENCES

Abdul Hamid, A. and Darwish, K. (2010). Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115. Association for Computational Linguistics.

Curran, J. R. and Clark, S. (2003). Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 164–167, Edmonton, Canada. Association for Computational Linguistics.

Ekbal, A. and Bandyopadhyay, S. (2010). Named entity recognition using support vector machine: A language independent approach.

Ekbal, A., Saha, S., and Garbe, C. S. (2010). Feature selection using multiobjective optimization for named entity recognition. In *International Conference on Pattern Recognition*, pages 1937–1940.

Favre, B., Hakkani-Tür, D., and Shriberg, E. (2009). Syntactically-informed models for comma prediction. pages 4697–4700, Taipei, Taiwan.

Georgiev, G., Nakov, P., Ganchev, K., and Osenova, P. (2009). Feature-rich named entity recognition for bulgarian using conditional random fields. *aclweborg*, pages 113–117.

Gravier, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *European Conf. on Speech Communication and Technology*.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Copenhagen, Denmark. Association for Computational Linguistics.

Isozaki, H. and Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Taipei, Taiwan. Association for Computational Linguistics.

Jan Hajic, e. a. (2005). Manual for morphological annotation, revision for the prague dependency treebank 2.0. Technical Report TR-2005-27, ÚFAL MFF UK, Praha, Czechia.

Kozareva, Z., Ferrández, O., Montoyo, A., Muñoz, R., Suárez, A., and Gómez, J. (2007). Combining data-driven systems for improving named entity recognition. *Data & Knowledge Engineering*, 61:449–466.

Kravalová, J., Ševčíková, M., and Žabokrtský, Z. (2009). Czech Named Entity Corpus 1.0.

Kravalová, J. and Žabokrtský, Z. (2009). Czech named entity corpus and svm-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, NEWS '09, pages 194–201, Suntec, Singapore. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

McCallum, A. and Li, W. (2003). 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*, CONLL '03, pages 188–191, Edmonton, Canada. Association for Computational Linguistics.

Sang, T. K. and Erik, F. (2002). Introduction to the conll-2002 shared task: language-independent named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–4, Taipei, Taiwan.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). Harem: An advanced ner evaluation contest for portuguese. In *Odjik and Daniel Tapias (eds.), Proceedings of LREC 2006 (LREC'2006) (Genoa*, pages 22–28.

Satoshi, S. and Hitoshi, I. (2000). Ir and ie evaluation project in japanese. In *LREC*.

Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 473–480, Philadelphia, Pennsylvania. Association for Computational Linguistics.