Named Entities as new Features for Czech Document Classification

Pavel Král[1,2]

[1] Dept. of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
[2] NTIS - New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
pkral@kiv.zcu.cz

Abstract. This paper is focused on automatic document classification. The results will be used to develop a real application for the Czech News Agency. The main goal of this work is to propose new features based on the Named Entities (NEs) for this task. Five different approaches to employ NEs are suggested and evaluated on a Czech newspaper corpus. We show that these features do not improve significantly the score over the baseline word-based features. The classification error rate improvement is only about 0.42% when the best approach is used.

1 Introduction

Nowadays, the amount of electronic text documents and the size of the World Wide Web are extremely rapidly growing. Therefore, automatic document classification is particularly important for information organization, storage and retrieval.

This work is focused on a real application of the document classification for the Czech News Agency (CTK).[1] CTK produces daily about one thousand text documents, which belong to different classes such as sport, culture, business, etc. In the current application, documents are manually annotated. Unfortunately, the manual annotation represents a very time consuming and expensive task. Moreover, this annotation is often not sufficiently accurate. It is thus beneficial to propose and implement an automatic document classification system.

Named Entity (NE) Recognition was identified as a main research topic for automatic information retrieval around 1996 [1]. The objective is identification of expressions with special meaning such as person names, organizations, times, monetary values, etc. The named entities can be successfully used in many fields and applications, e.g. question answering, information filtering, etc.

In this paper, we propose new features for document classification of the Czech newspaper documents based on the named entities. We believe that NEs bring

[1] http://www.ctk.eu

some additional information, which can improve the performance of our document classification system. Our assumptions are supported by the following observations.

- NEs can be used to differentiate some similar words according to the context. For example, "Bush" can be American president or popular British band. Using information about the NE, the documents can be classified correctly to one of the two different categories: politics or culture.
- It is possible to use named entities to discover synonyms, e.g. "USA" and "United States" are two different words. However, the word-sense is similar and they represent the same NE, the *country*. This additional information should help to classify two different documents containing "USA" and "United States" words, respectively, into the same category.
- Named entities shall be also used to identify and connect individual words in the multiple-words entities to one token. For example, the words in the expression "Mladá fronta dnes" (*name of the Czech newspaper*) do not have any sense. They can, used separately, produce a mismatch in document classification because the word "dnes" (*today*) is mostly used in the class weather. Using one token can avoid this issue.

Five different approaches to employ this information are proposed and evaluated next.

1. add directly the named entities to the feature vector (which is composed of words (or lemmas)) as new tokens
2. concatenate words related to multiple-word entities to one individual token
3. combine (1) and (2)
4. concatenate words and named entities to one individual token
5. replace words related to the named entities by their NEs

Note that, to the best of our knowledge, named entities were never used previously as features for the document classification task of the Czech documents. Moreover, we have not found another work which uses NEs similarly for document classification.

Section 2 presents a short review about the document classification approaches with the particular focus on the use of the NE recognition in this field. Section 3 describes our approaches of the integration of named entities to the feature vector. Section 4 deals with the realized experiments on the CTK corpus. We also discuss the obtained results. In the last section, we conclude the research results and propose some future research directions.

2 Related Work

Document clustering is an unsupervised approach that aims at automatically grouping raw documents into clusters based on their words similarity, while document classification relies on supervised methods that exploit a manually annoted training corpus to train a classifier, which in turn identifies the class of new unlabeled documents. Mixed approaches have also been proposed, such as semi-supervised approaches, which augment labeled training corpus with unlabeled data [2], or methods that exploit partial labels to discover latent topics [3]. This work focuses on document classification based on the Vector Space Model (VSM), which basically represents each document with a vector of all occurring words weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

Several classification algorithms have been successfully applied [4, 5], e.g. Bayesian classifiers, decision trees, k-Nearest Neighbour (kNN), rule learning algorithms, neural networks, fuzzy logic based algorithms, Maximum Entropy (ME) and Support Vector Machines (SVMs). However, the main issue of this task is that the feature space in VSM is highly dimensional which negatively affects the performance of the classifiers.

Numerous feature selection/reduction approaches have been proposed [6] in order to solve this problem. The successfully used feature selection approaches include Document Frequency (DF), Mutual Information (MI), Information Gain (IG), Chi-square test or Gallavotti, Sebastiani & Simi metric [7, 8]. Furthermore, a better document representation may lead to decreasing the feature vector dimension, e.g. using lemmatization or stemming [9]. More recently, advanced techniques based on Labeled Latent Dirichlet Allocation (LDA) [10] or Principal Component Analysis (PCA) [11] incorporating semantic concepts [12] have been introduced.

Multi-label document classification [13, 14][2] becomes a popular research field, because it corresponds usually better to the needs of the real applications than one class document classification. Several methods have been proposed as presented for instance in surveys [15, 16].

The most of the proposed approaches is focused on English and is usually evaluated on the Reuters,[3] TREC[4] or OHSUMED[5] databases.

Only little work is focused on the document classification in other languages. Yaoyong et al. investigate in [17] learning algorithms for cross-language document classification and evaluate them on the Japanese-English NTCIR-3 patent retrieval test collection.[6] Olsson presents in [18] a Czech-English cross-language classification on the MALACH[7] data set. Wu et al. deals in [19] with a bilingual topic aspect classification of English and Chinese news articles from the Topic Detection and Tracking (TDT)[8] collection.

Unfortunatelly, only few work about the classification of the Czech documents exits. Hrala et al. proposes in [20] a precise representation of Czech documents (lemmatization and Part-Of-Speech (POS) tagging included) and shown that mutual information is the most accurate feature selection method which gives with the maximum entropy or support vector machines classifiers the best results in the single-label Czech document classification task[9]. It was further shown [21] that the approach proposed by Zhu et al. in [22] is the most effective one for multi-label classification of the Czech documents.

To the best of our knowledge, only little work on the use of the NEs for document classification has been done. Therefore, we will focus on the use of the named entities in the closely related tasks. Joint learning of named entities and document topics has mainly been addressed so far in different tasks than document clustering. For instance, the authors of [23] exploit both topics and named entity models for language model adaptation in speech recognition, or [24] for

[2] One document is usually labeled with more than one label from a predefined set of labels.
[3] http://www.daviddlewis.com/resources/testcollections/reuters21578
[4] http://trec.nist.gov/data.html
[5] http://davis.wpi.edu/xmdv/datasets/ohsumed.html
[6] http://research.nii.ac.jp/ntcir/permission/perm-en.html
[7] http://www.clsp.jhu.edu/research/malach/
[8] http://www.itl.nist.gov/iad/mig//tests/tdt/
[9] One document is assigned exactly to one label from a predefined set of labels.

new event detection. Topic models are also used to improve named entity recognition systems in a number of works, including [25–27], which is the inverse task to our proposed work. Joint entity-topic models have also been proposed in the context of unsupervised learning, such as in [28] and [29].
The lack of related works that exploit named entity recognition to help document classification is mainly explained in [30], which has precisely studied the impact of several NLP-derived features, including named entity recognition, for text classification, and concluded negatively. Despite this very important study, we somehow temper this conclusion and show that our intuition that suggests us that named entity features cannot be irrelevant in the context of document classification, might not be completely wrong. Indeed, nowadays NLP tools have improved and may provide richer linguistic features, and the authors of [30] only use a restricted definition of named entities, which are limited to proper nouns, while we are exploiting more complex types of named entities.

3 Document Classification with Named Entities

3.1 Preprocessing, Feature Selection and Classification

The authors of [20] have shown that morphological analysis including lemmatization and POS tagging with combination of the MI feature selection method significantly improve the document classification accuracy. Therefore, we have used the same preprocessing in our work.
Lemmatization is used in order to decrease the feature number by replacing a particular word form by its *lemma* (base form) without any negative impact to the classification score. The words that should not contribute to classification are further filter out from the feature vector according to their POS tags. The words with approximately uniform distribution among all document classes are removed from the feature vector. Therefore, only the words having the POS tags noun, adjective or adverb remain in the feature vector.
Note that the above described steps are very important, because irrelevant and redundant features can degrade the classification accuracy and the algorithm speed. In this work, we would like to evaluate the importance of new features. Absolute value of the recognition accuracy thus does not play a crucial role. Therefore, we have chosen the simple Naive Bayes classifier which has usually an inferior classification score. However, it will be sufficient for our experiments to show whether new features bring any supplementary information.

3.2 Named Entity Integration

For better understanding, the features obtained by the proposed approaches will be demonstrated on the Czech simple sentence "Český prezident Miloš Zeman dnes navštívil Spojené státy." (*The Czech president Miloš Zeman visited today the United States*) (see Table 1). The baseline features after lemmatization and POS-tag filtration are shown in the first line of this table. The second line corresponds to the English translation and the third line illustrates the recognized named entities.

Table 1. Examples of the NE-based features obtained by the five proposed approaches

	Český	Prezident	Miloš	Zeman	dnes	Spojené	státy.			
	Czech	president	Miloš	Zeman	today	United	States			
	O	O	Figure-B	Figure-I	Datetime-B	Country-B	Country-I			
1.	Český	Prezident	Miloš	Zeman	Figure	dnes	Datetime	Spojené	státy	Country
2.	Český	Prezident	Miloš-Zeman	dnes	Spojené-státy					
3.	Český	Prezident	Miloš-Zeman	Figure	dnes	Datetime	Spojené-státy	Country		
4.	Český	Prezident	Miloš-Zeman-Figure	dnes-Datetime	Spojené-státy-Country					
5.	Český	Prezident	Figure	Datetime	Country					

Named entities as new tokens in the feature vector - (1) The baseline feature vector is composed of words (lemmas in our case) and their values are calculated by the TF-IDF approach. In this approach, we insert directly the named entity labels to the feature vector as new tokens. The values of the NE features are calculated similarly as the values of the word features using the TF-IDF method. One example of the resulting features of this approach is shown in the first line of the second section of Table 1.

Note that the feature values in all following approaches will be also computed by the TF-IDF weighting.

Concatenation of words (lemmas) related to multiple-word entities to one individual token - (2) As mentioned previously, the individual words of the multiple-word entities have usually the different meaning than connected to one single token. In this approach, all words which create a multiple-word NE are connected together and the NE labels are further discarded. The second line of the second section of Table 1 shows the features created by this approach.

Combination of the approach (1) and (2) - (3) We assume that the NE labels can bring other information than the connected words of the multiple-word NEs. Therefore, in this approach we combine both previously proposed methods as illustrated in the third line of the second section of Table 1.

Concatenation of words (lemmas) and named entities into one individual token - (4) The concatenated words of the multiple-word NEs and their NE labels are used in the previous approach as two separated tokens. In this approach, they are linked together to create one token. This approach should play an important role for word sense disambiguation (e.g. “Bush-Figure” vs. “Bush-Organization”). One example of the features obtained by this approach is shown in the fourth line of the second section of Table 1.

Named entities instead of the corresponding words - (5) The previously proposed approaches increase the size of the feature vector. In this last approach, the size of the vector is reduced replacing the words corresponding to the named entities by their NE labels. The last line of Table 1 shows the features created by this approach.

Weighting of the named entities We assume that named entities represent the most important words in the documents. Therefore, we further slightly modify the TF-IDF weighting in order to increase the importance of the named entities. The original weight is multiplied by K when named entity identified.

Note that we often use the term "words" in the text while in the experiment we use rather their "lemmas" instead.

4 Experiments

4.1 Tools and Corpora

We used the mate-tools[10] for lemmatization and POS tagging The lemmatizer and POS tagger were trained on 5853 sentences (94.141 words) randomly taken from the PDT 2.0[11] [31] corpus. The performance of the lemmatizer and POS tagger are evaluated on a different set of 5181 sentences (94.845 words) extracted from the same corpus. The accuracy of the lemmatizer is 81.09%, while the accuracy of our POS tagger is 99.99%. Our tag set contains 10 POS tags as shown in Table 2.

We use the top scoring Czech NER system [32]. It is based on Conditional Random Fields. The overall F-measure on the CoNLL format version of Czech Named Entity Corpus 1.0 (CNEC) is 74.08%, which is the best result so far. We have used the model trained on the private CTK Named Entity Corpus (CTKNEC). The F-measure obtained on this corpus is about 65%.

For implementation of the classifier we used an adapted version of the MinorThird[12] tool. It has been chosen mainly because of our experience with this system.

As already stated, the results of this work shall be used by the CTK. Therefore, for the following experiments we used the Czech text documents provided by the CTK. Table 2 shows the statistical information about the corpus. This corpus is available only for research purposes for free at `http://home.zcu.cz/~pkral/sw/` or upon request to the authors.

In all experiments, we used the five-folds cross validation procedure, where 20% of the corpus is reserved for the test. All experiments are repeated 10 times with randomly reshuffled documents in the corpus. The final result of the experiment is then a mean of all obtained values. For evaluation of the classification accuracy, we used, as frequently in some other studies, a standard *Error Rate (ER)* metric. The resulting error rate has a confidence interval of $< 0.5\%$.

Our NE tag-set is composed of 16 named entities (see Table 3). This table further shows the numbers of the NE occurrences in the corpus. The total number of the NE occurrences is about 700,000 which represents a significant part of the corpus (approximately 13%).

Note that, this named entities have been identified fully-automatically. Some labeling errors are thus available. This fact can influence the following experiments negatively.

[10] http://code.google.com/p/mate-tools/

[11] http://ufal.mff.cuni.cz/pdt2.0/

[12] http://sourceforge.net/apps/trac/minorthird

Table 2. Corpus statistical information

Unit name	Unit number	Unit name	Unit number
Document	11,955	Numeral	216,986
Category	60	Verb	366,246
Word	5,145,788	Adverb	140,726
Unique word	193,399	Preposition	346,690
Unique lemma	152,462	Conjunction	144,648
Noun	1,243,111	Particle	10,983
Adjective	349,932	Interjection	8
Pronoun	154,232		

Table 3. NE tag-set and distribution in the CTK document corpus

NE	No.	NE	No.	NE	No.	NE	No.
City	55,370	E-subject	5,447	Number	160,633	Religion	24
Country	56,081	Figure	133,317	Organization	119,021	Sport	12,524
Currency	25,429	Geography	7,418	Problematic	17	Sport-club	38,745
Datetime	108,594	Nationality	5,836	Region	14,988	Uknown	1,750

4.2 Analysis of the Named Entity Distribution according to the Document Classes and Classification with only NEs

This experiment should support our assumption that named entities bring useful information for document classification. Therefore, we realize a statistical study of the distribution of the named entities according to the document classes in the corpus (see Figure 1). This figure shows that some NEs (e.g. E-subject, Region, Sport, etc.) are clearly discriminant across the document classes. The analysis supports our assumption that the NEs can have a positive impact to the document classification.

We further realize another experiment in order to show whether only named entities (without the word features) are useful for the document classification. The results of this experiment (see the first line of Table 4) shows that NEs bring some information for document classification. However, their impact is small.

4.3 Classification Results of the proposed Approaches

The Table 4 further shows the recognition error rates of the proposed approaches. We evaluate the NE weights $K \in \{1, 2, 3\}$. The greater weight values are not used because the classification scores is decreasing according to this value in all experiments.

This table shows that the named entities help for document classification only slightly and this improvement is unfortunately statistically not significant. The best score is obtained by the second approach when the words are concatenated across the NEs and the information about the NE labels is completely removed from the feature vector.

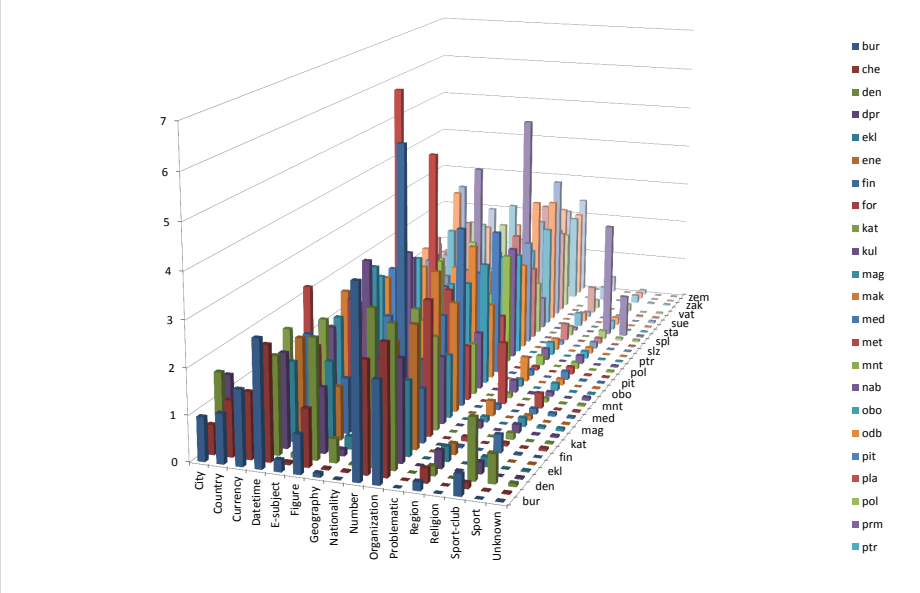

Fig. 1. Distribution of the named entities according to the document classes

Table 4. Document classification error rates [in %] of the different implementations of the named entity features (NE weights $K \in \{1, 2, 3\}$)

	Approach	NE weights 1	2	3
NEs only		84.25		
Lemmas (baseline)		17.83		
1.	Lemmas + NEs	17.60	17.75	17.79
2.	Concatenated lemmas	**17.41**	18.01	18.71
3.	1 + 2 together	17.48	18.02	18.61
4.	Concatenated Lemmas + NEs as one token	17.54	18.20	18.59
5.	NEs instead of the corresponding words	17.81	18.27	19.03

4.4 Analysis of the Confusion Matrices

In this experiment, we analyze the confusion matrices in order to compare the errors when the baseline word-bases features and the proposed features used (see Table 5). This table illustrates the number of errors, number of different errors in absolute value and in %, respectively. It is depicted that about 27% of errors (except the first proposed method) is different. Therefore, this experiment confirms that named entities bring some additional information. Unfortunately, this

information is not sufficient to improve significantly the document classification accuracy on the CTK document corpus.
Note that the error number of the baseline approach is 2,131.

Table 5. Analysis of the confusion matrices errors between the baseline and five proposed approaches

	Baseline vs. Proposed Approach	Error no.	Diff. err. no.	Diff. err. no [in %]
1.	Lemmas + NEs	2,104	357	16.97
2.	Concatenated lemmas	2,081	577	27.73
3.	1 + 2 together	2,089	581	27.81
4.	Concatenated Lemmas + NEs as one token	2,097	569	27.13
5.	NEs instead of the corresponding words	2,129	594	27.9

5 Conclusions and Future Work

In this paper, we have proposed new features for the document classification based on the named entities. We have introduced five different approaches to employ NEs in order to improve the document classification accuracy. We have evaluated these methods on the Czech CTK corpus of the newspaper text documents. The experimental results have shown that these features do not improve significantly the score over the baseline word-based features. The improvement of the classification error rate was only about 0.42% when the best approach is used. We have further analyzed and compared the confusion matrices of the baseline approach with our proposed methods. This analysis has shown that named entities bring some additional information for document classification. Unfortunately, this information is not sufficient to improve significantly the document classification accuracy.
However, we assume that this information could play more important role on smaller corpora with more unknown words in the testing part of the corpus. The first perspective thus consists in evaluation of the proposed features on the other (smaller) corpora including more European languages. Then, we would like to propose other sophisticated features which introduce the semantic similarity of word-based features. These features should be useful for example for word-sense disambiguation and can be created for instance by the semantic spaces.

Acknowledgements

This work has been partly supported by the European Regional Development Fund (ERDF), project “NTIS - New Technologies for Information Society”, European Centre of Excellence, CZ.1.05/1.1.00/02.0090. We also would like to thank Czech New Agency (CTK) for support and for providing the data.

References

1. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th conference on Computational linguistics - Volume 1. COLING '96, Copenhagen, Denmark, Association for Computational Linguistics (1996) 466–471
2. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Mach. Learn. **39** (2000) 103–134
3. Ramage, D., Manning, C.D., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11, New York, NY, USA, ACM (2011) 457–465
4. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management. (2004) 679–694
5. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 380–393
6. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research **3** (2003) 1289–1305
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. ICML '97, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1997) 412–420
8. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries. ECDL '00, London, UK, UK, Springer-Verlag (2000) 59–68
9. Lim, C.S., Lee, K.J., Kim, G.C.: Multiple sets of features for automatic genre classification of web documents. Information Processing and Management **41** (2005) 1263 – 1276
10. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 248–256
11. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis **56** (2012) 741–751
12. Yun, J., Jing, L., J., Y., Huang, H.: A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications **39** (2012) 2035–2046
13. Novovicova, J., al. et: Conditional mutual information based feature selection for classification task. Progress in Pattern Recognition, Image Analysis and Applications (2007) 417–426
14. Forman, G., Guyon, I., Elisseeff, A.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research **3** (2003) 1289–1305

15. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34** (2002) 1–47
16. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3** (2007) 1–13
17. Yaoyong, L., Shawe-Taylor, J.: Advanced learning algorithms for cross-language patent retrieval and classification. Information processing & management **43** (2007) 1183–1199
18. Olsson, J.S.: Cross language text classification for malach. (2004)
19. Wu, Y., Oard, D.W.: Bilingual topic aspect classification with a few training examples. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2008) 203–210
20. Hrala, M., Král, P.: Evaluation of the Document Classification Approaches. In: 8th International Conference on Computer Recognition Systems (CORES 2013), Milkow, Poland, Springer (2013) 877–885
21. Hrala, M., Kral, P.: Multi-label document classification in Czech. In: 16th International conference on Text, Speech and Dialogue (TSD 2013), Pilsen, Czech Republic, Springer (2013) 343–351
22. Zhu, S., Ji, X., Xu, W., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2005) 274–281
23. Liu, Y., Liu, F.: Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In: Proceedings ASSP. (2008)
24. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2004) 297–304
25. Guo, H., Zhu, H., Guo, Z., Zhang, X., Wu, X., Su, Z.: Domain adaptation with latent semantic association for named entity recognition. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 281–289
26. Knopp, J., Frank, A., Riezler, S.: Classification of named entities in a large multilingual resource using the Wikipedia category system. PhD thesis, Masters thesis, University of Heidelberg (2010)
27. Zhang, Z., Cohn, T., Ciravegna, F.: Topic-oriented words as features for named entity recognition. In: Computational Linguistics and Intelligent Text Processing. Springer (2013) 304–316
28. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06, New York, NY, USA, ACM (2006) 680–686
29. Vosecky, J., Jiang, D., Leung, K.W.T., Ng, W.: Dynamic multi-faceted topic discovery in twitter. In: Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management. CIKM '13, New York, NY, USA, ACM (2013) 879–884
30. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In McDonald, S., Tait, J., eds.: Advances in In-

formation Retrieval. Volume 2997 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2004) 181–196
31. Hajič, J., Böhmová, A., Hajičová, E., Vidová-Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abeillé, A., ed.: Treebanks: Building and Using Parsed Corpora. Amsterdam: Kluwer (2000) 103–127
32. Konkol, M., Konopík, M.: Crf-based Czech named entity recognizer and consolidation of Czech NER research. In Habernal, I., Matoušek, V., eds.: Text, Speech and Dialogue. Volume 8082 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2013) 153–160