Multi-label Document Classification in Czech

Michal Hrala[1] and 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
{hrala36,pkral}@kiv.zcu.cz

Abstract. This paper deals with multi-label automatic document classification in the context of a real application for the Czech news agency. The main goal of this work is to compare and evaluate three most promising multi-label document classification approaches on a Czech language. We show that the simple method based on a meta-classifier proposes by Zhu at al. outperforms significantly the other approaches. The classification error rate improvement is about 13%. The Czech document corpus is available for research purposes for free which is another contribution of this work.

Keywords: Czech, Czech News Agency, Maximal Entropy, Multi-label Document Classification, Naive Bayes, Maximal Entropy, Support Vector Machines.

1 Introduction

Automatic document classification becomes very important for information organization and storage because of the fast increasing amount of electronic text documents and the rapid growth of the World Wide Web. In this work, we focus on the *multi-label* document classification[1] in the context of the application for the Czech News Agency (CTK).[2] CTK produces daily about one thousand of text documents. These documents belong to different categories such as weather, politics, sport, etc. Nowadays, documents are manually annotated but this annotation is often not accurate enough. Moreover, the manual labeling represents a very time consuming and expensive task. Therefore, automatic document classification is very important.

In our previous work [1], we proposed a precise Czech document representation (lemmatization and POS tagging included) and evaluated five feature selection methods, namely document frequency, mutual information, information gain, Chi-square test and Gallavotti, Sebastiani & Simi metric on three classifiers (Naive Bayes (NB), Maximal

[1] One document is usually labeled with more than one label from a predefined set of labels.
[2] http://www.ctk.eu

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 343–351, 2013.

Entropy (ME) and Support Vector Machines (SVMs)) in order to build en efficient one class (sometimes also called single-label) Czech document classification[3] system.

The main goal of this work is to adapt our previously developed system to multi-label classification task. The main scientific contribution is to compare and evaluate three most promising multi-label document classification approaches on a Czech language in order to build an efficient Czech multi-label document classification system. Note that to the best of our knowledge, there is no comparative study on the multi-label document classification approaches evaluated on Czech documents. Another contribution of this work is the public availability of the Czech document corpus for the research purposes.

Section 2 presents a short review about the document classification approaches. Section 3 describes three document classification approaches that are compared. Section 4 deals with the realized experiments on the CTK corpus. In the last section, we discuss the research results and we propose some future research directions.

2 Related Work

The document classification task is basically treated as a supervised machine-learning problem, where the documents are projected into the so-called Vector Space Model (VSM), basically using the words as features. Various classification methods have been successfully applied [2,3], e.g. Bayesian classifiers, decision trees, k-Nearest Neighbour (kNN), rule learning algorithms, neural networks, fuzzy logic based algorithms, maximum entropy and support vector machines. However, the task suffers from the issue that the feature space in VSM is highly dimensional which negatively affects the performance of the classifiers.

To deal with this issue, techniques for feature selection or reduction have been proposed [4]. The successfully used classical feature selection approaches include document frequency, mutual information, information gain, Chi-square test or Gallavotti, Sebastiani & Simi metric [5,6].Furthermore, a better document representation may lead to decreasing the feature vector dimension, e.g. using lemmatization or stemming [7]. More recently, advanced techniques based on Principal Component Analysis (PCA) [8] incorporating semantic concepts [9] have been introduced.

Recently, multi-label document classification [10,11,12] becomes a popular research field, because it corresponds usually better to the needs of the real applications, than one class document classification. Unfortunately, it is much more complicated. *The choice of* 1 *class from the predefined set of* N *classes* becomes *the choice of* M *classes from* N *ones* (M value is unknown). Several approaches have been proposed as summarized for instance in a survey [13].

The most of the proposed methods (see above) deals with English and are usually evaluated on the Reuters,[4] TREC[5] or OHSUMED[6] data sets.

[3] One document is assigned exactly to one label from a predefined set of labels.

[4] `http://www.daviddlewis.com/resources/testcollections/reuters21578`

[5] `http://trec.nist.gov/data.html`

[6] `http://davis.wpi.edu/xmdv/datasets/ohsumed.html`

Only few work focuses on the document classification in other languages. Yaoyong et al. investigate in [14] learning algorithms for cross-language document classification and evaluate them on the Japanese-English NTCIR-3 patent retrieval test collection.[7] Olsson presents in [15] a Czech-English cross-language classification on the MALACH[8] data set. Wu et al. deals in [16] with a bilingual topic aspect classification of English and Chinese news articles from the Topic Detection and Tracking (TDT)[9] collection.

Unfortunately, to the best of our knowledge, there are no language-specific multi-label document classification method for documents written in Czech language. In such a case, the issues of large feature vectors become more significant due to the complexity of Czech language when compared to English.

3 Multi-label Document Classification

3.1 Preprocessing, Feature Selection and Classification

The same preprocessing as in our previous work [1] is used, i. e. a morphological analysis including *lemmatization* and *Part-Of-Speech (POS) tagging*. The lemmatization decreases the number of features by replacing a particular word form by its *lemma* (base form) without any negative impact to the classification accuracy.

The knowledge of the POS tags is used for the further feature vector reduction. We filter out the words that should not contribute to classification according to theirs POS tags. The words with the uniform distribution among all document classes are removed from the feature vector. After this filtration, only words with the POS tags noun, adjective or adverb remain in the feature vector.

As a feature selection, the mutual information method is used because it achieves the best results in our previous work.

Note, that the above described steps are very important, because irrelevant and redundant features can degrade the classification accuracy and the algorithm speed.

Three classifiers that are successfully used for document classification in the literature (see previous section) and in our previous work are used: Naive Bayes (NB), Maximal Entropy (ME) and Support Vector Machines (SVMs).

3.2 Multi-label Document Representation

The existing approaches can be divided into two groups: 1) problem transformation methods; and 2) algorithm adaptation methods. We focus here only on the first group. According to the authors of the survey [13], we have implemented two approaches that give the best classification scores. These approaches are described next. Then, we present a simple approach proposed by Zhu et al. in [17].

[7] http://research.nii.ac.jp/ntcir/permission/perm-en.html
[8] http://www.clsp.jhu.edu/research/malach/
[9] http://www.itl.nist.gov/iad/mig//tests/tdt/

Class and Complement. Let N be the number of the classes. The first approach uses N binary classifiers $C_{i=1}^{N} : x \rightarrow l, \neg l$, i. e. each binary classifier assigns the document x to the label l iff the label is included in the document, $\neg l$ otherwise.

The final classification result is given by:

$$C(x) = \cup_{i=1}^{N} : C_i(x) = l \tag{1}$$

The main drawback of this method is a very long training and classification time. This approach is hereafter called *Class & complement*.

Merged Categories. Let K be the number of the different sets of labels existing in the corpus. The second approach uses each different set of labels as a new single label:

$$L = \cup_{k=1}^{K} l_k \tag{2}$$

One class document classifier $C : x \rightarrow L$ is then used for the document classification. Authors of [13] state, that this approach brings the best classification results. The principal weakness of this method is the data sparsity, i.e. some new classes with few document occurrences are created. This approach is further called *Merged categories*.

Threshold Classification. In this approach, the corpus is transformed as follows: the document with K labels is considered as K one class documents for training. The same classifier C as in the one label document classification task is created. This classifier produces a sorted list of the N labels l_i according to their classification scores s_i.

The core of the method consists in building a meta-classifier C_M in order to separate K classes belonging to the document and the rest $\neg K$. In this work, we distinguish these two classes by a *threshold* T. The document x is associated with a label l_i iff:

$$s_i(x) > T \tag{3}$$

The resulting set of labels L is given by:

$$L = \cup l_i \leftrightarrow C_M : x \rightarrow l_i \tag{4}$$

The threshold value is determined experimentally on the development corpus.

Note, that this approach is very simple. Nevertheless, there are two main advantages of this method: 1) minimal adaptation of our previously developed system is necessary; 2) algorithm speed. This approach is hereafter called *Threshold classification*.

4 Experiments

4.1 Tools and Corpora

For lemmatization and POS tagging, we used the mate-tools.[10] The lemmatizer and POS tagger were trained on 5853 sentences (94.141 words) randomly taken from the PDT

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

2.0[11] [18] 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 1.

We used an adapted version of the MinorThird[12] tool for implementation of the document classification methods. This tool has been chosen mainly because the three evaluated classification algorithms were already implemented.

As mentioned previously, the results of this work will be used by the CTK. Therefore, for the following experiments we used the Czech text documents provided by the CTK. Table 1 shows the statistical information about the corpus. Figure 1 illustrates the distribution of the documents depending on the number of labels. 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. For evaluation of the classification accuracy, we used as frequently in some other studies a *Error Rate (ER)* metric. The resulting error rate has a confidence interval of $< 1\%$.

Table 1. Corpus statistical information

Unit name	Unit number	Unit name	Unit number
Document	11955	Numeral	216986
Category	60	Verb	366246
Word	2974040	Adverb	140726
Unique word	193399	Preposition	346690
Unique lemma	152462	Conjunction	144648
Noun	1243111	Particle	10983
Adjective	349932	Interjection	8
Pronoun	154232		

4.2 Class and Complement

The first section of the Table 2 shows the classification results of the *class & complement* approach. These results show clearly, that SVM and ME classifiers having comparable scores outperform significantly the NB.

Note that the ER metric is very strict, because the document is considered as classified incorrectly when only one label (from K) is not correct.

4.3 Merged Categories

As already stated, this approach suffers from the data sparsity problem. There are some classes with few document occurrences and a correct estimation of such models is very difficult. One solution is not to consider the classes with few occurrences and remove them from the classification.

[11] `http://ufal.mff.cuni.cz/pdt2.0/`
[12] `http://sourceforge.net/apps/trac/minorthird`

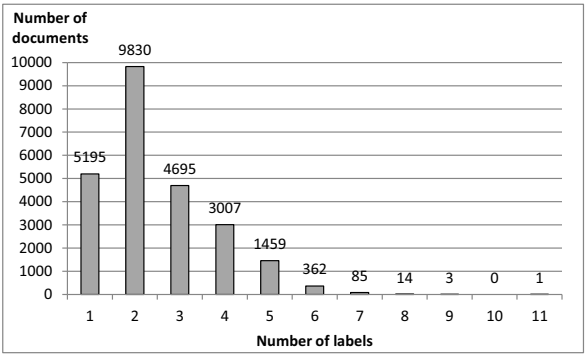

Fig. 1. Distribution of the documents depending on the number of labels

Table 2. Classification error rates [in %] of the all evaluated approaches using NB, SVM and ME classifiers

	Approach	Classifier NB	SVM	ME
1	Class & complement	63.45	**36.62**	37.60
2	Merged categories	52.88	60.87	**32.26**
3	Threshold classification	38.40	25.67	**23.12**
4	Number of classes given	19.04	**9.44**	9.96

Figure 2 illustrates the classification error rates of the *merged categories* approach using NB, SVM and ME classifiers depending on the number of classes. This number is given by the value of the minimal number of documents per class. The figure demonstrates that the ME classifier brings better results than NB and SVM. The difference is most significant in the case when all classes are considered.

The second section of the Table 2 shows the error rates when all classes are used. The best error rate value is 32.26% which outperforms the best score of the previous experiment by 4% in the absolute value.

4.4 Threshold Classification

Figure 3 illustrates the classification error rates of the *threshold classification* approach using NB, SVM and ME classifiers when the different thresholds are used. The best error rates are reported in the third section of the Table 2. These results show that this approach outperforms significantly both previous methods. The best result is given by the ME classifier as in the previous approach.

4.5 Results Analysis

In this section, we would like to analyze the most accurate method from the two aspects:

1. Evaluation of the classification result where the correct number of the classes is given (see the fourth section of the Table 2). We can conclude that it is possible to

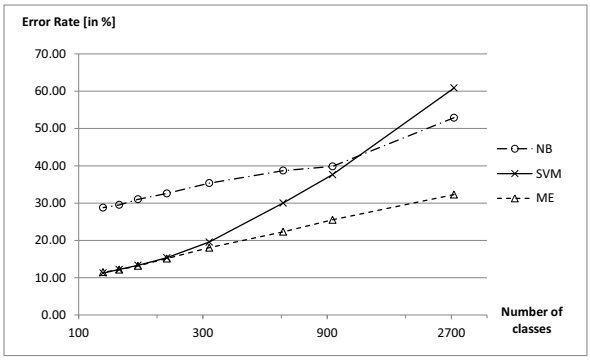

Fig. 2. Document classification error rates of the *merged categories* approach using NB, SVM and ME classifiers depending on the number of classes (x-axis in logarithmic scale)

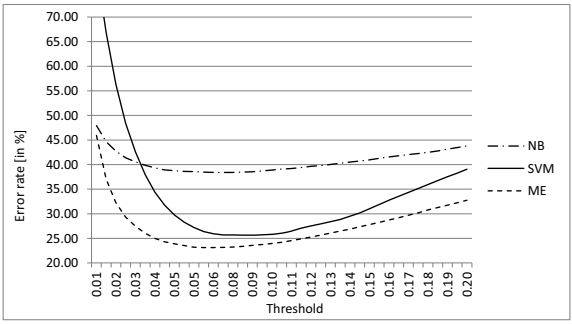

Fig. 3. Document classification error rates [in %] of the *threshold classification* approach

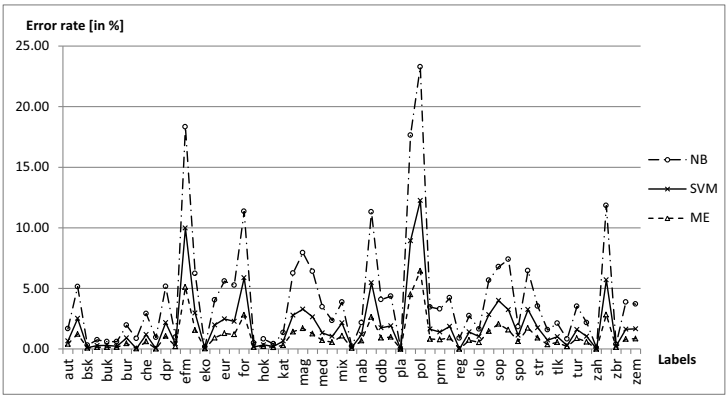

Fig. 4. Error rates of the separated classes (without any combination using a meta-classifier)

improve the classification score by 13% when the "ideal" meta-classifier is used. The ME and SVM give comparable results, while the error rate of the NB is significantly worse.
2. Evaluation of the single-label classification (see Figure 4), i. e. error rates of all classes separately without any combination. This analysis confirms that the single-label classification is much easier than the multi-label ones. Note, that the best global error rate is given by the SVM classifier and is 9.44%.

5 Conclusions and Future Work

In this work, we have implemented three promising multi-label document classification methods. Then, we have evaluated these methods on the Czech CTK corpus. We have shown that the simplest method based on the meta-classifier outperforms significantly both other approaches. The classification error rate improvement is about 13%.

The main perspective consists in proposing a more suitable document representation. For this task, we would like to study the impact of the syntactic structure of the sentence, semantic spaces, etc.

Acknowledgements. This work has been partly supported by the UWB grant SGS-2013-029 Advanced Computer and Information Sstems and 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. Hrala, M., Král, P.: Evaluation of the Document Classification Approaches. In: Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierek, A. (eds.) CORES 2013. AISC, vol. 226, pp. 875–885. Springer, Heidelberg (2013)
2. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management, pp. 679–694 (2004)
3. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 380–393 (1997)
4. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)
5. 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 1997, pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
6. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 59–68. Springer, Heidelberg (2000)
7. 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, 1263–1276 (2005)
8. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis 56, 741–751 (2012)

9. Yun, J., Jing, L., Yu, J., Huang, H.: A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications 39, 2035–2046 (2012)
10. Novovičová, J., Malík, A., Pudil, P.: Feature selection using improved mutual information for text classification. In: Fred, A., Caelli, T.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds.) SSPR&SPR 2004. LNCS, vol. 3138, pp. 1010–1017. Springer, Heidelberg (2004)
11. Novovičová, J., Somol, P., Haindl, M., Pudil, P.: Conditional mutual information based feature selection for classification task. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 417–426. Springer, Heidelberg (2007)
12. Forman, G., Guyon, I., Elisseeff, A.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
13. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3, 1–13 (2007)
14. Yaoyong, L., Shawe-Taylor, J.: Advanced learning algorithms for cross-language patent retrieval and classification. Information Processing & Management 43, 1183–1199 (2007)
15. Olsson, J.S.: Cross language text classification for malach (2004)
16. 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, pp. 203–210. ACM (2008)
17. 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, pp. 274–281. ACM (2005)
18. 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, pp. 103–127. Kluwer, Amsterdam (2000)