Evaluation of the Document Classification Approaches

Michal Hrala and Pavel Král[1]

Abstract This paper deals with one class automatic document classification. Five feature selection methods and three classifiers are evaluated on a Czech corpus in order to build an efficient Czech document classification system. Lemmatization and POS tagging are used for a precise representation of the Czech documents. We demonstrated, that POS tag filtering is very important, while the lemmatization plays a marginal role for classification. We also showed that Maximum Entropy and Support Vector Machines are very robust to the feature vector size and outperform significantly the Naive Bayes classifier from the view point of the classification accuracy. The best classification accuracy is about 90% which is enough for an application for the Czech News Agency, our commercial partner.

1 Introduction

Due to the increasing amount of electronic text documents and the rapid growth of the World Wide Web, automatic classification becomes very important for information organization and storage. The document classification task can be divided into the one class and the more class classification. In the one class classification, the document is assigned exactly to one label from a predefined set of labels, while in the more class classification (sometimes also multi-label classification), the document can be labeled with more than one label.

In this work, we focus on the one class document classification in the context of the further application for the Czech News Agency (CTK). CTK pro-

Department of Computer Science and Engineering,
Faculty of Applied Sciences,
University of West Bohemia Plzeň, Czech Republic,
e-mail: {hrala36,pkral}@kiv.zcu.cz

duces daily about one thousand of text documents. These documents belong to different categories such as weather, politics, sport, etc. Today, documents are manually annotated but this annotation is often not enough accurate. Moreover, the manual labeling represents a very time consuming and expensive job. Automatic classification is thus very beneficial.

There are three main steps in the document classification: document representation, feature selection and document modeling. Document representation consists in choosing a feature set that represents the document as accurately as possible. The full-text is transformed into the document feature vector. Feature selection is then used in order to reduce the size of this vector. The last step consists in building a document model using feature vectors. This model is used for document classification.

To the best of our knowledge, there is no complex comparative study of the document classification approaches that consider the specifics of the Czech language. The main goal of this work is thus: 1) to propose a precise Czech document representation. Morphological analysis that includes lemmatization and POS tagging is taken into account; 2) to evaluate the most promising feature selection methods and classification models on a Czech corpus in order to build an efficient Czech document classification system.

Section 2 presents a short review about the document classification approaches. Section 3 describes the presented document classification approach. Section 4 deals with the realized experiments. 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 [1, 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. The successfully used classical feature selection approaches include document frequency, mutual information, information gain, Chi-square test or Gallavotti, Sebastiani & Simi metric [2, 4, 5, 6, 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 Principal Component Analysis (PCA) [10] incorporating semantic concepts [11] have been introduced.

Unfortunately, relatively little attention has been paid to language-specific methods, such as classification methods designed exclusively for documents written in Czech. In such a case, the issues of large feature vectors become more significant due to the complexity of this language when compared to English.

3 Proposed Method

One issue of the document classification is very high data dimensionality so the number of potential features usually exceeds significantly the number of the available documents. A suitable document representation can decrease the size of the feature vector.

We would like to respect the characteristics of the Czech language in order to choose a representative feature-set that reflects the document as accurate as possible. Therefore, a morphological analysis including *lemmatization* and *Part-Of-Speech (POS) tagging* is realized.

3.1 Lemmatization

We assume that a particular word form do not contribute for the document classification. A lemmatization thus will decrease the number of features by replacing a particular word form by its *lemma* (base form) without any negative impact to the classification accuracy.

Following the definition from the Prague Dependency Treebank (PDT) 2.0[1] [12] project, we used only the first part of the *lemma*. This is a unique identifier of the lexical item (e.g. infinitive for a verb), possibly followed by a digit to disambiguate different lemmas with the same base forms. For instance, the Czech word “třeba”, having the identical lemma, can signify *necessary* or *for example* depending on the context. This is in the PDT notation differentiated by two lemmas: “třeba-1” and “třeba-2”.

The second part containing additional information about the lemma, such as semantic or derivational information, is not taken into account in this work.

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

3.2 Part-Of-Speech (POS) tagging

The *part-of-speech* is a word linguistic category, which can be defined by the syntactic or morphological behaviour of the lexical item in question [12].

The next step that will contribute to the feature vector reduction is a word filtration according to the POS tags. The words with the uniform distributions among all document classes will be removed from the feature vector. This task is usually done by using the previously defined list of words, so called *stop-list*.

We consider ten POS categories defined in the PDT 2.0 for the Czech language: nouns, adjectives, pronouns, numerals, verbs, adverbs, prepositions, conjunctions, particles and interjections.

3.3 Feature Selection

A feature selection method is then used for the further reduction of the size of the feature vector. Based on the literature (see Section 2), five most promising feature selection approaches, namely Document Frequency (DF), Mutual Information (MI), Information Gain (IG), Chi squared (χ^2) test and Gallavotti, Sebastiani & Simi (GSS) coefficient will be compared and evaluated.

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

3.4 Document Model

The last step consists in building a robust document model. Three classifiers that are successfully used for document classification in the literature (see Section 2)) are used and evaluated for this task. The evaluated classifiers are: Naive Bayes (NB), Maximal Entropy (ME) and Support Vector Machines (SVMs).

4 Experiments

4.1 Tools and Corpora

For lemmatization and POS tagging, we used the mate-tools[1]. The lemmatizer and POS tagger were trained on 5853 sentences (94.141 words) randomly taken from the PDT 2.0 corpus, which is a collection of Czech newspaper texts annotated on the morphological, syntactic and semantic layer. 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[2] [13] 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 statistic information about the corpus[3]. 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 a *Error Rate (ER)* metrics that is defined by the following equation:

$$ER = \frac{E}{A} \tag{1}$$

where E represents the number of incorrectly classified documents and A is the number of all classified documents. The resulting error rate has a confidence interval of $< 1\%$.

4.2 Impact of the Size of the Feature Vector

The first experiment studies the classification accuracy depending on the size of the feature vector. The objective is to determine the feature vector size when the decrease of the classification accuracy is still negligible. Based on the studies that deal with English document classification [2], we define an initial document representation. Lemmas are used instead of the words and the POS tag filtering is realized. Only lemmas that correspond to the nouns,

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

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

[3] This Czech document corpus is available only for research purposes for free at http://home.zcu.cz/∼pkral/sw/ or upon request to the authors.

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

Table 1 Corpus statistic information

adjectives, verbs and adverbs are used for the creation of the feature vector. As mentioned previously, five most promising feature selection methods and three classifiers are compared.

Left column of the Figure 1 shows the results of this experiment. We can conclude that Maximum Entropy and Support Vector Machines classifiers give very close results and outperform significantly the Naive Bayes classifier from the viewpoint of the classification accuracy. The feature selection method plays an important role for classification when the number of features is small. However, the differences are not significant with a great number of features. Based on this figure, we chose the size of the feature vector for the next experiments 3000 for ME and SVM classifiers and 4000 for the NB classifier.

4.3 Impact of the POS Tag Filtering

The second experiment deals with the importance of the POS tag filtering for the accuracy of the document classification. Based on the results presented in [14] for classification of English documents, we evaluate all combinations of the following POS tags: Nouns (N), Adjectives (A), Verbs (V) and Adverbs (D). The remaining POS tags are not considered because of the small or negative impact for the document classification. We assume that the nouns will be the most important for classification. Therefore, they are used for all configurations of the feature set.

The results of this experiment are presented in the second column of Figure 1. The best feature set is the same for all classifiers and is composed of the

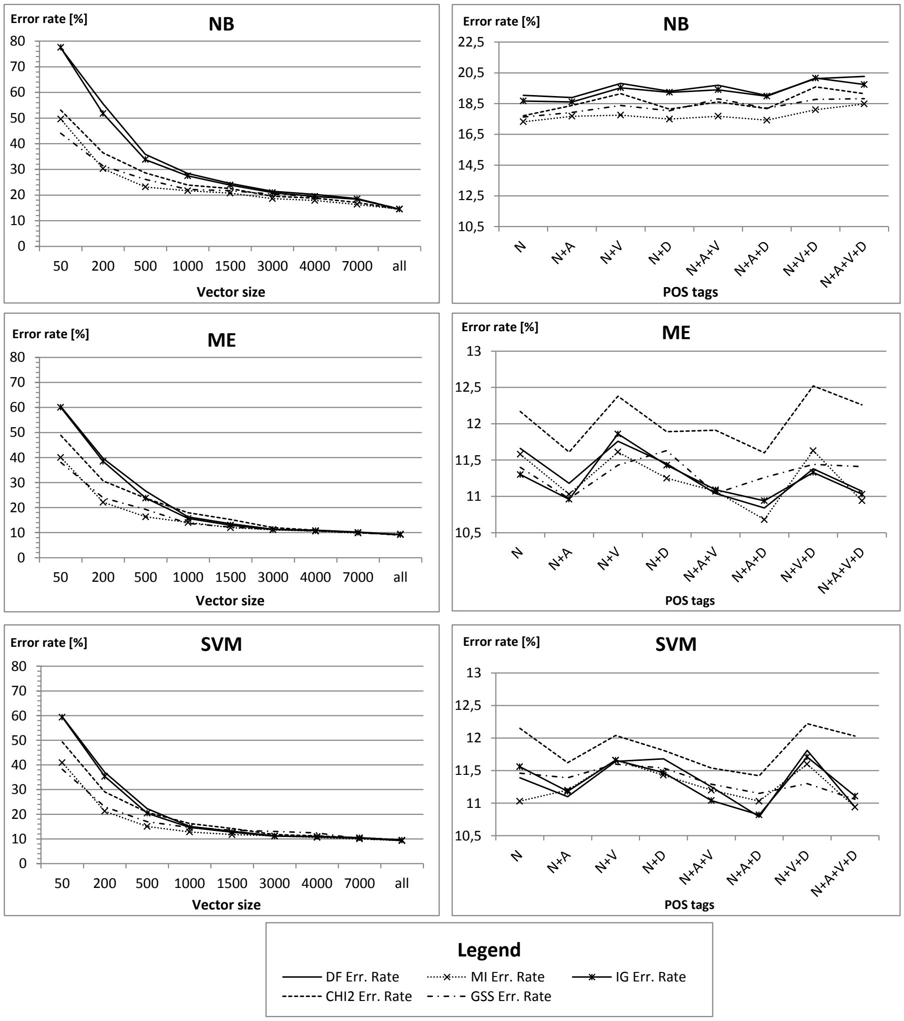

Fig. 1 Impact of the feature vector size (left column) and of the POS tags (right column) on the document classification accuracy. Five feature selection methods (document frequency, mutual information, information gain, Chi squared test and GSS coefficient) and three classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines - from the top) are compared and evaluated.

nouns, adjectives and adverbs (i.e. *N+A+D*). The use of verbs decreases the classification score in all cases. The best feature selection metrics is mutual information. Based on these results, we will consider only the words (lemmas) with the POS tags: *N*, *A* and *D* in the following experiments.

4.4 Impact of the Lemmas

This experiment deals with the impact of the lemmas on the classification accuracy. As already stated in the Section 3.1, we assume that lemmas will have a positive impact on the document classification if the size of the feature vector remains constant.

The results of this experiment are shown in Table 2. The use of lemmas instead of words has a small positive impact on classification, however the obtained increase of the classification accuracy is less than 1% and is statistically not significant. Note, that 193.399 words was replaced by 152.462 lemmas.

Classifier		Feature selection method				
	lemmas/words	DF	MI	IG	χ^2 test	GSS
NB	lemmas	19.14	17.26	18.75	18.19	17.98
	words	20.35	17.71	19.5	18.41	18.2
ME	lemmas	10.94	10.92	11.21	11.89	11.14
	words	11.66	11.09	11.31	11.96	11.45
SVM	lemmas	10.84	10.78	11.07	11.66	11.03
	words	11.33	10.93	11.38	11.7	11.02

Table 2 Document classification Error Rate (ER [in %]) depending on the use of lemmas. Five feature selection methods (document frequency, mutual information, information gain, Chi squared test and GSS coefficient - columns) and three classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines - table lines) are compared and evaluated.

4.5 Document Classification using the Best Configuration of the Classifiers

The last experiment compares the document classification accuracy when the best configuration of the classifiers is used. Two cases are evaluated: the reduced feature set and all feature set (about 152 000 features). Only the lemmas according to the nouns, adjectives and adverbs are considered and the mutual information feature selection method is used.

Table 3 shows the classification accuracy of this experiment. As expected, the maximum entropy and support vector machine classifiers outperform significantly the Naive Bayes classifier from the viewpoint of the classification accuracy. Moreover, the scores of these classifiers are almost similar in the case with reduced features set and with all features.

Vector size	4000	3000	3000	All	All	All
Classifier	NB	ME	SVM	NB	ME	SVM
Error Rate	17.26	10.92	10.78	13.94	9.1	8.79

Table 3 Document classification error rate when the best configuration of the classifiers is used [in %].

5 Conclusions and Future Work

In this work, we evaluated the five feature selection methods and the three classifiers on a Czech corpus in order to build an efficient Czech document classification system. We used lemmatization and POS tagging for a precise representation of the Czech documents. We showed the impact of the feature vector length, of the POS tag filtering and of the lemmas on the classification accuracy of Czech documents. We demonstrated, that POS tag filtering is very important, while the lemmatization plays a very small role for the classification score. We also showed that Maximum Entropy and Support Vector Machines outperform significantly the Naive Bayes classifier from the viewpoint of the classification accuracy. Moreover, these two classifiers are very robust to the size of the feature vector when the mutual information feature selection method is used. Based on the experiments, we set an optimal configuration of the classifiers. The best classification accuracy is about 90%.

In this paper, we presented the results obtained with one class document classification. The first perspective consists in the adaptation of our current system to a multi-label classification task. This extension is beneficial for our commercial partner, CTK. The next perspective is to propose 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 Systems 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. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management. (2004) 679–694

2. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. 1 edn. Cambridge University Press (2008)
3. 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
4. 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
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 '97, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1997) 412–420
6. Luo, X., Zincir-Heywood, A.N.: Incorporating temporal information for document classification. In: ICDE Workshops. (2007) 780–789
7. 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
8. Cover, T., Thomas, J.: Elements of information theory. Wiley, New York (1991)
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. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis **56** (2012) 741–751
11. 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
12. 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
13. Cohen, W.W.: Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. (2004)
14. P. Ponmuthuramalingam, T.D.: Effective term based text clustering algorithms. In: International Journal on Computer Science and Engineering, International Journal on Computer Science and Engineering (2010) 1665–1673