Novel Unsupervised Features for Czech Multi-label Document Classification

Tomáš Brychcín[1,2], 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
{brychcin,pkral}@kiv.zcu.cz

Abstract. This paper deals with automatic multi-label document classification in the context of a real application for the Czech News Agency. The main goal of this work consists in proposing novel fully unsupervised features based on an unsupervised stemmer, Latent Dirichlet Allocation and semantic spaces (HAL and COALS). The proposed features are integrated into the document classification task. Another interesting contribution is that these two semantic spaces have never been used in the context of document classification before. The proposed approaches are evaluated on a Czech newspaper corpus. We experimentally show that almost all proposed features significantly improve the document classification score. The corpus is freely available for research purposes.

Keywords: Multi-label Document Classification, LDA, Semantic spaces, HAL, COALS, HPS, Stemming, Czech, Czech News Agency, Maximum Entropy

1 Introduction

Nowadays, the amount of electronic text documents and the size of the World Wide Web increase extremely rapidly. Therefore, automatic document classification (or categorization) becomes very important for information retrieval.

In this work, we focus on the *multi-label* document classification[1] in the context of a real application for the Czech News Agency (ČTK)[2]. ČTK produces daily about one thousand of text documents. These documents belong to different categories such as politics, sport, culture, business, etc. In the current application, documents are manually annotated. Unfortunately, the manual labeling represents a very time consuming and

[1] *Multi-label* document classification: one document is usually labeled with more than one label from a predefined set of labels vs. *Single-label* document classification: one document is assigned exactly to one label.

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

expensive task. It is thus beneficial to propose and implement an automatic document classification system.

One important issue in the document classification field is the high dimensionality and insufficient precision of the feature vector. Several feature selection methods and sophisticated language specific features have been proposed. The main drawback of these methods is that they need a significant amount of the annotated data. Furthermore, a complete re-annotation is necessary when the target language is modified.

In this work, we address these issues by proposing novel fully unsupervised features based on an unsupervised stemmer, Latent Dirichlet Allocation (LDA) and semantic spaces (HAL and COALS). We further integrate these features into the document classification task.

The next scientific contribution is evaluating a new simple LDA model, called S-LDA, which integrates stem features into the topic modeling. Another interesting contribution is the use of semantic space models (i.e. HAL and COALS), because they have not been used for the document classification yet. The last contribution consists in the evaluation of the proposed approaches on Czech, as a representative of morphologically rich language.

The paper structure is as follows. Section 2 introduces the document classification approaches with a particular focus on the document representation. Section 3 describes our proposed features and their integration into the document classification task. Section 4 deals with the experiments on the ČTK corpus. In the last section, we discuss the research results and we propose some future research directions.

2 Related Work

The today's document classification relies usually on supervised machine learning methods that exploit a manually annotated training corpus to train a classifier, which in turn identifies the class of new unlabeled documents. Most approaches are based on the Vector Space Models (VSMs), which mostly represent each document as a vector of all occurring words usually weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

Several classification algorithms have been successfully applied [3, 7], e.g. Bayesian classifiers, decision trees, k-Nearest Neighbor (kNN), rule learning algorithms, neural networks, fuzzy logic based algorithms, Maximum Entropy (ME) and Support Vector Machines (SVMs). However, one important issue of this task is that the feature space in VSM has a high dimension which negatively affects the performance of the classifiers.

Numerous feature selection/reduction approaches have been proposed in order to solve this problem. The successfully used feature selection methods include Document Frequency (DF), Mutual Information (MI), Information Gain (IG), Chi-square test or Gallavotti, Sebastiani & Simi metric [8, 9].

In the last years, multi-label document classification becomes a popular research field, because it corresponds usually better to the needs of the real applications than the single-label document classification. One popular approach presented in [27] uses n binary *class/no class* classifiers. A final classification is then given by an union of these

partial results. Another approach presented by the authors of [27] simplifies the multi-label document classification task by replacing *each different* set of labels by a new *single label*. Then, a single-label document classifier is created on such data. Note that this approach suffers by the data sparsity problem. Zhu et al. propose in [30] another multi-label document classification approach. The same classifier as in the single-label document classification task is created. The document is associated with a set of labels based on an acceptance *threshold*. The other methods are presented for instance in survey [26].

Furthermore, a better document representation may lead to decreasing the feature vector dimension, e.g. using lexical and syntactic features as shown in [18]. Chandrasekar et al. further show in [6] that it is beneficial to use POS-tag filtration in order to represent a document more accurate. The authors of [21] and [28] use a set of linguistic features. Unfortunately, they do not show any impact to the document classification task. However, they conclude that more complex linguistic features may improve the classification score.

More recently, an advanced technique based on Labeled Latent Dirichlet Allocation (L-LDA) [24] has been introduced. Unlike our approach, L-LDA incorporates supervision by constraining the topic model to use only those topics that correspond to document labels. Principal Component Analysis (PCA) [10] incorporating semantic concepts [29] has been also successfully proposed for the document classification. Semi-supervised approaches, which augment labeled training corpus with unlabeled data [22] were also used.

The most of the proposed approaches is focused on English. Unfortunately, only little work about the document classification in other non-mainstream languages, particularly in Czech, exits. Hrala et al. [14] use lemmatization and POS-tag filtering for a precise representation of the Czech documents. The authors further show the performance of three multi-label classification approaches [13].

3 Document Classification

In the following sections we describe the proposed unsupervised features and classification approaches.

3.1 Unsupervised Stemming

Stemming is a task to replace a particular (inflected) word form by its "stem" (an unique label for all morphological forms of a word). It is used in many Natural Language Processing (NLP) fields (e.g. information retrieval) to reduce the number of parameters with a positive impact to the classification accuracy. Therefore, we assume that stems should improve the results of the document classification.

We propose two approaches to integrate the stem features into the document classification. In the first approach, the stem occurrences are used directly as the features, while in the second one, we use stems as a preprocessing step for LDA. We use an unsupervised stemming algorithm called HPS [5] This stemmer have been already proved to be very efficient in the NLP, see for example [12].

Note that this task is very similar to lemmatization. However, the main advantage of our stemming approach is that it is fully unsupervised and thus it does not need any annotated data (only plain text).

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is a popular topic model that assigns a topic to each word in the document collection. In our first approach, we use a standard LDA model as follows. We calculate the topic probabilities for each document. The probability of each topic t is given by the number of times the topic t occurs in a document divided by the document size. These probabilities are used directly as new features for a classifier.

In our second approach, we use stems instead of words. This concept is motivated by the following assumptions. LDA is a bag-of-word model, thus the word role in a sentence is inhibited. We assume that the morphosyntactic information in a document is useless for inferring topics. Moreover, the word normalization (i.e. stemming in our case) can reduce the data sparsity problem, which is particularly significant in the processing of morphologically rich languages (e.g. Czech). The parameters of such model should be better estimated than the parameters of the standard LDA. The features for the classifier are calculated in the same way as for the word-based LDA. This model will be hereafter called the *S-LDA* (Stem-based LDA).

3.3 Semantic Spaces

Semantic spaces represent words as high dimensional vectors. Semantically close words should be represented by similar vectors and the vector space gives an opportunity to use a clustering method to create word clusters.

The authors of [4] have proved that word clusters created by the semantic spaces improve significantly language modeling. We assume that these models can play an important role for document classification. We use two semantic space models, namely: HAL (Hyperspace Analogue to Language) [19] and COALS (Correlated Occurrence Analogue to Lexical Semantic) [25]. The word clusters are created using Repeated bisection algorithm. The document is then represented as a bag of clusters and we use a tf-idf weighting scheme for each cluster to create the features.

We assume that these models should reduce (analogically as in the previous case) the data sparsity problem. It is worth of mentioning that these two semantic space models have never been used in the context of document classification before.

3.4 Document Classification

For multi-label classification, we use (as presented in [27]) n binary classifiers $C_{i=1}^{n}$: $d \rightarrow l, \neg l$ (i.e. each binary classifier assigns the document d to the label l iff the label is included in the document, $\neg l$ otherwise). The classification result is given by the following equation:

$$C(d) = \cup_{i=1}^{n} : C_i(d) \tag{1}$$

The Maximum Entropy (ME) [1] classifier is used. As a baseline, we use the tf-idf weighting of the word features. Then, this set is progressively extended by the novel unsupervised features. In order to facilitate the reading of the paper, all features are summarized next.

- **Words (baseline)** – Occurrence of a word in a document. Tf-idf weighting is used.
- **Stems** – Occurrence of a stem in a document. Tf-idf weighting is used.
- **LDA** – LDA topic probabilities for a document.
- **S-LDA** – S-LDA topic probabilities for a document.
- **HAL** – Occurrence of a HAL cluster in a document. Tf-idf weighting is used.
- **COALS** – Occurrence of a COALS cluster in a document. Tf-idf weighting is used.

4 Experiments

In our experiments we use LDA implementation from the MALLET [20] tool-kit. For each experiment, we train LDA with 1,000 iterations of the Gibbs sampling. The hyperparameters of the Dirichlet distributions were (as proposed in [11]) initially set to $\alpha = 50/K$, where K is the number of topics and $\beta = 0.1$.

The S-Space package [15] is used for implementation of the HAL and COALS algorithms. For each semantic space, we use a four-word context window (in both directions). HAL uses a matrix consisting of 50,000 columns. COALS uses a matrix with 14,000 columns (as suggested by the authors of the algorithm). SVD (Singular Value Decomposition) was not used in our experiments.

We created the word clusters in the similar way as described in [4], i.e. by using Repeated Bisection algorithm and cosine similarity metric. For clustering, we use an implementation from the CLUTO software package [16]. For both semantic spaces, the word vectors are clustered into four depths: 100, 500, 1,000, and 5,000 clusters.

For multi-label classification we use Brainy [17] implementation of Maximum Entropy classifier.

4.1 Corpus

As mentioned previously, the results of this work will be used by the ČTK. Therefore, we use Czech document collection provided by the ČTK for the training of our models (i.e. LDA, S-LDA, semantic spaces and multi-label classifier).

This corpus contains 2,974,040 words belonging to 11,955 documents annotated from a set of 37 categories. Figure 1 illustrates the distribution of the documents depending on the number of labels. This corpus is freely available for research purposes at `http://home.zcu.cz/~pkral/sw/`.

In all experiments, we use the five-fold cross-validation procedure, where 20% of the corpus is reserved for the test. For evaluation of the document classification accuracy, we use the standard Precision (P), Recall (R) and F-measure (F_m) metrics [23]. The confidence interval of the experimental results is 0.6% at a confidence level of 0.95.

No feature selection has been done in our experiments to clearly show the impact of the proposed features. In the following tables, the term *words* denotes the word features and *stems* denotes the stem features.

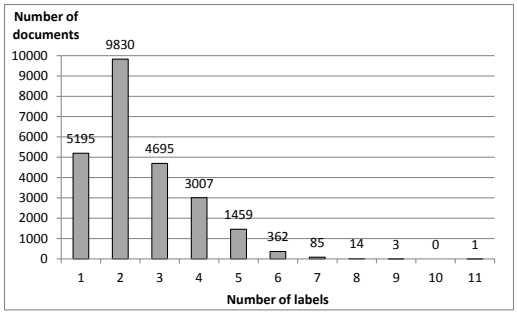

Fig. 1. Distribution of the documents depending on the number of labels

4.2 Classification Results of the LDA and S-LDA Models

In this experiment, we would like to compare the classification results of the stand-alone LDA and S-LDA model (see Table 1). This table shows that the larger number of the topics is better for document classification. Moreover, the proposed S-LDA slightly outperforms the stand-alone LDA model for all topic numbers.

Table 1. Results of stand-alone LDA and S-LDA models

topics	LDA			S-LDA		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	82.9	65.9	73.4	83.1	66.0	73.6
200	83.4	69.1	75.6	83.7	70.3	76.4
300	84.0	71.1	77.0	85.3	72.6	78.4
400	83.3	70.7	77.5	85.5	73.2	78.8
500	84.7	72.6	78.2	85.9	74.0	79.5

4.3 Classification Results of the LDA and S-LDA Models with baseline Word Features

This experiment compares the classification results of the stand-alone LDA and S-LDA models when the baseline word features are also used (see Table 2). Unlike the previous experiment, the recognition score remains almost constant for every topic number and both LDA models. The topic number and LDA type thus no longer play any role for document classification.

4.4 Classification Results of the Semantic Space Models

This experiment compares the classification results of the HAL and COALS models (see Table 3). The table shows that with rising number of clusters the classification score

Table 2. Results of LDA models with baseline word features

topics	Words+LDA			Words+S-LDA		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	89.0	74.0	80.8	88.9	74.0	80.8
200	88.9	73.8	80.7	88.9	73.6	80.5
300	88.9	73.6	80.6	89.0	73.6	80.6
400	88.8	73.7	80.5	88.8	73.7	80.5
500	88.8	73.7	80.5	88.8	73.5	80.4

increases. At the level of 5,000 clusters the score is almost the same as for the baseline. However, the number of the parameters in the classifier is significantly reduced.

In the case of COALS and 5,000 clusters the F-measure is slightly better than the baseline. However, we believe this deviation is caused by a chance. In all these experiments COALS outperforms the HAL model.

Table 3. Results of semantic space models

clusters	HAL			COALS		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	58.5	14.7	23.6	66.9	25.2	36.6
500	76.1	51.3	61.3	79.6	59.3	68.0
1000	80.2	62.0	70.0	81.6	64.8	72.2
5000	87.9	72.1	79.2	88.5	73.5	80.3

4.5 Classification Results of the Semantic Space Models with baseline Word Features

This experiment compares the classification results of the HAL and COALS models when the baseline word features are also used. The results are reported in Table 4. Unlike the previous experiment, the recognition score remains almost constant for all clusters and for both semantic space models.

We can explain this behavior by the fact that the clusters from semantic spaces do not bring any useful additional information compared to the baseline.

Table 4. Results of semantic space models with baseline word features

clusters	Words+HAL			Words+COALS		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	88.2	72.6	79.7	88.2	72.8	79.7
500	88.2	72.7	79.7	88.2	72.7	79.7
1000	88.3	72.8	79.8	88.2	72.7	79.7
5000	88.3	72.8	79.8	88.3	72.7	79.7

4.6 Classification Results of the Different Model Combinations

In this section we evaluate and compare several combinations of our models (see Table 5). The best model configurations from the previous experiments are used. These configurations are compared over the baseline "word" approach (first line in the table). This experiment clearly shows that almost all proposed features significantly improve the document classification accuracy. The F-measure improvement is 2.1% in the absolute value when all proposed features are used. Only the semantic space models do not have any significant impact to improve the classification score. Note that this behavior has been already justified in the previous section.

Table 5. Results of different model combinations. The term COALS denotes the combination of all four COALS models (i.e. 100, 500, 1000, and 5000 clusters). The term HAL denotes the combination of all HAL models. The term S-LDA means the combination of the S-LDA models with 100 and 400 topics.

model	$P[\%]$	$R[\%]$	$F_m[\%]$	impr. $F_m[\%]$
words	88.1	72.7	79.7	
stems	86.4	75.0	80.3	+0.7
words+stems	88.3	74.8	81.0	+1.3
words+HAL	88.4	72.8	79.9	+0.2
words+COALS	88.5	72.8	79.9	+0.2
words+S-LDA	89.2	74.6	81.2	+1.6
words+stems+S-LDA	88.8	75.5	81.6	+1.9
words+stems+S-LDA+COALS	89.0	75.6	81.7	+2.1

5 Conclusions and Future Work

In this work, we have proposed novel fully unsupervised features based on an unsupervised stemmer HPS, Latent Dirichlet Allocation and semantic spaces (HAL and COALS). These features were further integrated into the multi-label document classification task.

We have evaluated the proposed approaches on the ČTK corpus in Czech that is a representative of morphologically rich languages.

We have experimentally shown that almost all proposed unsupervised features significantly improve the document classification score. The F-measure improvement over the baseline is 2.1% absolute, when all proposed features are used.

We plan to extend our work by experiments with different languages and language families. Due to the unsupervised character of the proposed methods, no additional annotations are required.

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 (ČTK) for support and for providing the data.

References

1. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics 22(1), 39–71 (1996)
2. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 2003 (2003)
3. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management. pp. 679–694 (2004)
4. Brychcín, T., Konopík, M.: Semantic spaces for improving language modeling. Computer Speech & Language 28(1), 192 – 209 (2014)
5. Brychcín, T., Konopík, M.: Hps: High precision stemmer. Information Processing & Management 51(1), 68 – 91 (2015), http://www.sciencedirect.com/science/article/pii/S0306457314000843
6. Chandrasekar, R., Srinivas, B.: Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging (1996)
7. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(4), 380–393 (1997), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=588021
8. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)
9. 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. pp. 59–68. ECDL '00, Springer-Verlag, London, UK, UK (2000), http://dl.acm.org/citation.cfm?id=646633.699638
10. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis 56(3), 741–751 (2012)
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(Suppl 1), 5228–5235 (Apr 2004)
12. Habernal, I., Ptáček, T., Steinberger, J.: Sentiment analysis in czech social media using supervised machine learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 65–74. Association for Computational Linguistics, Atlanta, Georgia (June 2013)
13. Hrala, M., Kral, P.: Multi-label document classification in Czech. In: 16th International conference on Text, Speech and Dialogue (TSD 2013). pp. 343–351. Springer, Pilsen, Czech Republic (1-5 September 2013)
14. Hrala, M., Král, P.: Evaluation of the Document Classification Approaches. In: 8th International Conference on Computer Recognition Systems (CORES 2013). pp. 877–885. Springer, Milkow, Poland (27-29 May 2013)
15. Jurgens, D., Stevens, K.: The s-space package: An open source package for word space models. In System Papers of the Association of Computational Linguistics (2010)
16. Karypis, G.: Cluto - a clustering toolkit (2003), www.cs.umn.edu/~karypis/cluto
17. Konkol, M.: Brainy: A machine learning library. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) Artificial Intelligence and Soft Computing, Lecture Notes in Computer Science, vol. 8468. Springer Berlin Heidelberg (2014)

18. 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(5), 1263 – 1276 (2005), http://www.sciencedirect.com/science/article/pii/S0306457304000676
19. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods Instruments and Computers 28(2), 203–208 (1996)
20. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
21. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In: McDonald, S., Tait, J. (eds.) Advances in Information Retrieval. Lecture Notes in Computer Science, vol. 2997, pp. 181–196. Springer Berlin Heidelberg (2004), http://dx.doi.org/10.1007/978-3-540-24752-4_14
22. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Mach. Learn. 39(2-3), 103–134 (May 2000), http://dx.doi.org/10.1023/A:1007692713085
23. Powers, D.: Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies 2(1), 37–63 (2011)
24. 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. pp. 248–256. EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1699510.1699543
25. Rohde, D.L.T., Gonnerman, L.M., Plaut, D.C.: An improved method for deriving word meaning from lexical co-occurrence. Cognitive Psychology 7, 573–605 (2004)
26. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
27. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3(3), 1–13 (2007)
28. Wong, A.K., Lee, J.W., Yeung, D.S.: Using complex linguistic features in context-sensitive text classification techniques. In: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. vol. 5, pp. 3183–3188. IEEE (2005)
29. 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(2), 2035–2046 (2012)
30. 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)