SAPKOS: Experimental Czech Multi-label Document Classification and Analysis System

Ladislav Lenc, Pavel Král

SAPKOS: Experimental Czech Multi-label Document Classification and Analysis System

Ladislav Lenc[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
{llenc,pkral}@kiv.zcu.cz

Abstract. This paper presents an experimental multi-label document classification and analysis system called SAPKOS. The system which integrates the state-of-the-art machine learning and natural language processing approaches is intended to be used by the Czech news Agency (ČTK). Its main purpose is to save human resources in the task of annotation of newspaper articles with topics. Another important functionality is automatic comparison of the ČTK production with popular Czech media. The results of this analysis will be used to adapt the ČTK production to better correspond to the today's market requirements. An interesting contribution is that, to the best of our knowledge, no other automatic Czech document classification system exists. It is also worth mentioning that the system accuracy is very high. This score is obtained due to the unique system architecture which integrates a maximum entropy based classification engine with the novel confidence measure method.

Keywords: Document Classification, Czech News Agency, Maximum Entropy Classifier, Confidence Measure, System

1 Introduction

The amount of textual data available online has been growing rapidly since the origin of the World wide web. In the last decade, the online press substitutes to a considerable extent the traditional printed media. This trend affects in particular press agencies that are forced to organize much larger amounts of data. Unfortunately, it brings considerable demands on the people to process the data and is highly desirable to automatize this task as much as possible.

Document classification (DC) is one of the tasks that are very important for the press agencies and that can be delegated to a computer. A broad list of publications in this domain shows that a number of machine learning algorithms can be employed. In many applications, a single-label classifier is applied. It

means that each document is assigned exactly to one class (category). However, it is not sufficient in many real systems where one document usually belongs to multiple categories. The multi-label document classification became thus the today's center of interest [9].

The main goal of this paper is to present a novel experimental multi-label document classification system which integrates the state-of-the art machine learning and natural language processing methods. This system is used by the Czech News Agency (ČTK[3]), the major press agency in the Czech Republic, to automatically assign topics to newspaper documents. Another functionality of this system consists in automatic analysis of the ČTK media production and a subsequent comparison with popular Czech media (Lidové noviny, Mladá Fronta Dnes, etc.). The results of this analysis will enable ČTK to obtain an overview about what information is interesting and which not. Moreover, based on these results it is possible to adapt the direction of its production to better respond to the current market needs. It is worth mentioning that ČTK produces daily several thousands of documents that must be annotated and analysed. The processed documents are mainly in Czech language, however a significant amount is in English.

The following section summarizes important approaches in the document classification and confidence measure domains which are necessary to create the proposed system. Section 3 describes the methods and algorithms that are used in order to create the system and the following section presents its architecture. Section 5 shows experimental evaluation of the system on the real Czech ČTK document corpus. In the last section, we discuss the results and we propose some future system improvement.

2 Related Work

This section describes first the main approaches in the document classification and confidence measure fields which are beneficial for our document classification system. Then, we present some other existing document classification systems.

The methods used for document classification usually employ supervised machine learning. An annotated corpus must be available for training of a classifier which is subsequently used to annotate unlabelled documents. Vector Space Models (VSMs) are often used as document representation. It usually represents each document as a vector of word occurrences weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

There are several machine learning approaches successfully used for the document classification [2, 5]: Bayesian classifiers, Maximum Entropy (ME), Support Vector Machines (SVMs), etc. However, the main issue of this task is that the feature space in the VSM is highly dimensional which decreases the accuracy of the classifier. Numerous feature selection/reduction approaches have thus been introduced [6, 16] to solve this problem.

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

Furthermore, a better document representation should help to decrease the feature vector dimension, e.g. using lexical and syntactic features [18]. It has been demonstrated in [4] that it is beneficial to use POS-tag filtration in order to represent a document more accurately.

More recently, some interesting methods based on Labelled Latent Dirichlet Allocation (L-LDA) [25] have been proposed. Another recent work exploits partial labels to discover latent topics [26]. Principal Component Analysis (PCA) [8] incorporating semantic concepts [34] has been also introduced for DC. Nigam et al. proposed in [21] an interesting semi-supervised approach, which progressively augments labelled corpus with unlabelled documents. Another approach using enhanced grid based clustering algorithm is presented in [27].

The most of the proposed approaches is focused on English and only few works deal with Czech language. Hrala et al. use in [10] lemmatization and Part-Of-Speech (POS) filtering for a precise representation of Czech documents. In [9], three different multi-label classification approaches are compared and evaluated. Other recent works propose novel features based on named entities [14] or on unsupervised machine learning [3].

Confidence measure is used as a post-processing of the recognition/classification to determine whether a result is correct or not. The incorrectly recognized samples should be removed from the resulting set or another processing (e.g. manual correction) can be further realized.

This technique is mainly used in the automatic speech processing field [30, 12] and is mostly based on the *posterior* class probability. However, it can be successfully used in another research areas as shown in [31] for genome maps construction, in [11] for stereo vision, in [19] for handwriting sentence recognition or in [17] for automatic face recognition.

The confidence measures are mostly used in the single-label classification. But the nature of many real-world classification problems is multi-label. One approach using confidence measures in the multi-label setting is proposed in [29]. The authors use semi-supervised learning algorithms and include a confidence parameter when assigning the labels. Two methods for the confidence value computation are proposed.

Another possibility how to deal with the confidence measures is to use a so called Conformal Predictor (CP) [22]. CP assigns a reliable measure of confidence and is used as a complement of machine learning algorithms. Author of [23] proposes to use a modification called Cross-Conformal Predictor (CCP) to handle the multi-label classification task. He states that this modification is more suitable for this task because of its lower computational costs.

Unfortunately, to the best of our knowledge, no other automatic system for classification of the Czech text documents exists. However, there are some systems dedicated to classify the documents in other languages, particularly in English. The most important ones (often protected by the US patent) are shortly described next.

The first system [33] uses Latent Semantic Indexing (LSI) model to compute the word based feature vectors. Then these vectors are clustered to create

a centroid vector representing a category. An inductive learning from examples is further used to associate the document labels to not annotated English text documents.

Glas et al. proposed in [7] another document similarity detection and classification system to process e-mail messages. It can be thus used for example for spam detection and removal. The system is based on comparison of message fingerprints stored as their hash representations. The message is classified only if a significant resemblance level exceeds a predetermined threshold.

In [20], another document classification system is proposed. The authors have done a morphological analysis in order to keep the significant words from the document. The feature vector creation is based on the classical vector space models. The classification utilizes three different techniques, namely a Chi-square test, discriminant analysis and cluster analysis.

Another interesting system for document classification is introduced by Rocha et al. in [28]. It uses semantic information contained in the linked data to improve the document classification score. The classification itself uses both word and linked data based features. The system is used by a telecommunication operator for English and Italian document classification.

3 Technical Background

The following sections details our feature set, multi-label document classification approach and confidence measure methods. These methods are integrated into the proposed system.

3.1 Feature Set & Classification

The feature set is created according to Brychcín et al. [3]. They are used because the authors experimentally proved that the additional unsupervised features significantly improve the classification.

- **Words** – 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** – Latent Dirichlet Allocation topic probabilities for a document.
- **S-LDA** – Stem-based Latent Dirichlet Allocation topic probabilities for a document.
- **HAL** – Occurrence of a Hyperspace Analogue to Language cluster in a document. Tf-idf weighting is used.
- **COALS** – Occurrence of a Correlated Occurrence Analogue to Lexical Semantic cluster in a document. Tf-idf weighting is used.

For multi-label classification, we use an efficient approach presented by Tsoumakas et al. in [32]. This method employs 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 if 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] model is used for classification. For implementation of the multi-label classifier we used Brainy [13] implementation of Maximum entropy classifier. It has been chosen mainly because of our experience with this tool.

3.2 Confidence Measures

As already stated, the confidence measure is used as a post-processing of the classification to identify whether the result is correct or not. In this work, we combine two measures based on the *posterior* class probability by a multi-layer perceptron classifier as already presented in [15].

Posterior class probability approaches The output of an individual binary classifier C_i is the posterior probability $P(L|F)$, where $L \in \{l, \neg l\}$ represents a binary class and F represents the feature vector created from the text document d.

We use two different approaches. The first approach, called ***absolute confidence value***, assumes that higher recognition score confirms the classification result. For the correct classification $\hat{L}$ the following two equations must be satisfied:

$$\hat{L} = \arg\max_{L}(P(L|F)) \tag{2}$$

$$P(\hat{L}|F) > T1 \tag{3}$$

The second approach, called ***relative confidence value***, computes the difference between the l score and the $\neg l$ score by the following equation:

$$\Delta P = abs(P(l|F) - P(\neg l|F)) \tag{4}$$

Only the classification results with $\Delta P > T2$ are accepted. This approach assumes that the significant difference between l and $\neg l$ classification scores confirms the classification result.

$T1$ and $T2$ are the acceptance thresholds and their optimal values are set experimentally.

Composed supervised approach Let R_{abs} and R_{rel} be the scores obtained by the *absolute confidence value* and *relative confidence value* methods, respectively. Let variable H determine whether the document is classified correctly or not. The Multi-Layer Perceptron (MLP) classifier which models the *posterior* probability $P(H|R_{abs}, R_{rel})$ is used to combine the two partial measures in a supervised way.

The following MLP topology was set experimentally using a small development corpus (containing 100 randomly chosen examples) as the best one: two input nodes (R_{abs} and R_{rel}), ten nodes in the hidden layer and two output nodes (classes *correct / not correct*).

4 System Architecture

The system is designed as a web application and Java Vaadin 7[4] framework is used to facilitate the implementation. The open source search engine SOLR[5] is used for text documents storage, indexing and retrieval. This platform is used because it provides many advanced features in text processing field and is freely available even for commercial usage. The PostgreSQL[6] database management system is used for storing the classification models, results and statistics. The machine learning algorithms are implemented using the Brainy [13] library.

The system has a modular architecture and is composed from three main modules. Its architecture is depicted in Figure 1.

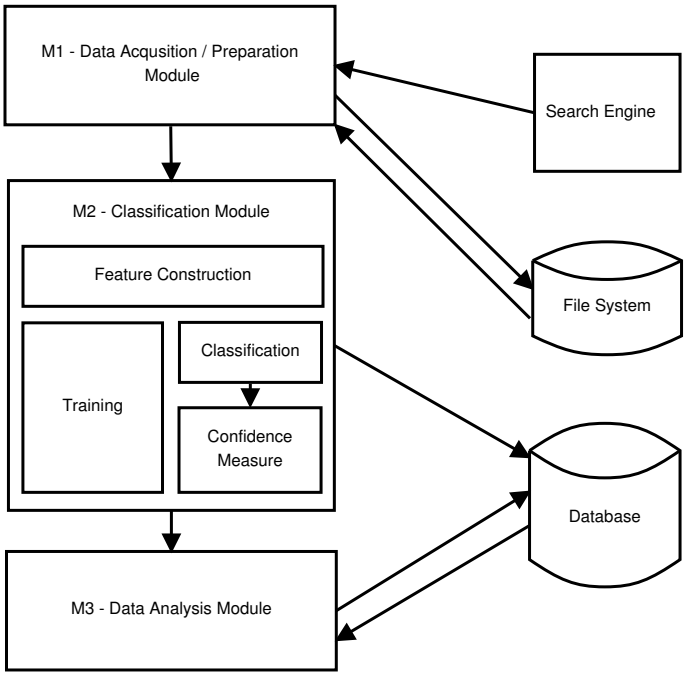

Fig. 1. SAPKOS system architecture

M1 - Data Acquisition/Preparation Module

This module is used to manage all main data flows in the SAPKOS system. It is connected with the SOLR engine which serves as the source of text documents.

[4] https://vaadin.com/home
[5] http://lucene.apache.org/solr/
[6] http://www.postgresql.org/

Its main goal is to load/store text (annotated/unannotated) documents to/from the system. The annotated documents are then used for training of the model in the classification module while the unannotated documents are used for classification. The documents from the SOLR engine are loaded to the system as *xml* files. The exact form of the input documents is *a priori* not known and therefore, the highly configurable interface is available.

There are several scenarios to load the data into the system. The basic one uses directly the annotated documents loaded from the search engine. Another possibility is loading documents from the engine and annotate them manually in the system. This module also allows exporting the loaded documents to *xml* format, setting/adjusting the assigned categories and importing the documents back into the system.

Figure 2 shows the configurable interface for loading the documents using the SOLR server. This figure shows that a user can set several filters to chose only the appropriate documents (e.g. date, priority, source media, keywords, text, etc.).

M2 - Classification Module

This module is the main part of the system. It provides three separate tasks: 1) feature construction; 2) model training and 3) classification itself. These tasks are described in details in Section 3.

The feature construction is a shared task which is used both for training and classification. During training, a classification model is built using annotated text documents provided by the module M1. This model is stored to the PostgreSQL database. During classification (testing) the learned model is used to assign topics to unannotated documents. The classification results are then stored into the database.

Note that the confidence measure method is also integrated to this module after the classification task.

M3 - Data Analysis Module

This module is used to prepare statistical data from the processed documents and to analyse, compare and visualize the statistics. There is also a possibility to export the data to a spreadsheet. As already stated, this analysis is very important to automatically compare the ČTK production with popular Czech media. The results of this analysis will be used to adapt the ČTK production to better correspond to the today's market requirements.

Figure 3 shows one example of comparison of the ČTK production with one Czech media. This figure shows that the document topics frequencies are almost the same[7]. Therefore, the ČTK production in this example corresponds to the market needs. This graph is generated by JFreeChart[8].

[7] red colour=ČTK production, blue colour=Czech newspaper *Lidové noviny*
[8] http://www.jfree.org/jfreechart/

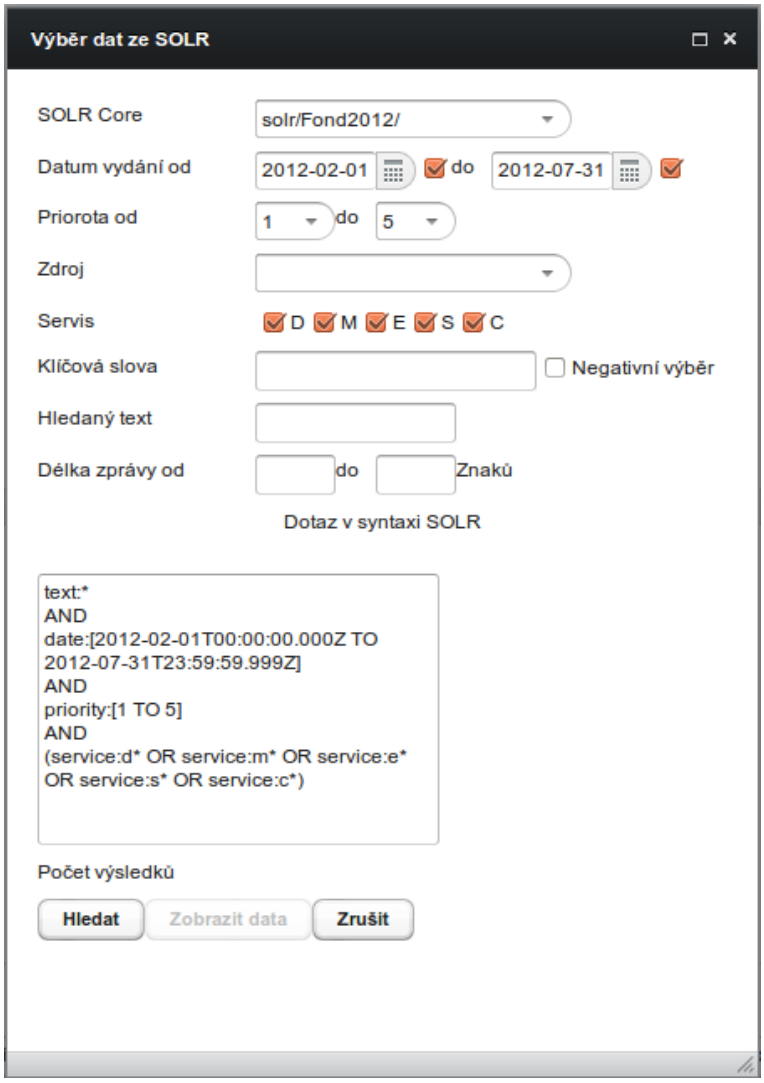

Fig. 2. Configurable dialogue window for data selection and loading from the SOLR search engine. The main field captions (from above) with English translations: "Datum vydání od/do" (*publication date from/to*), "Priorita od/do" (*priority from/to*), "Zdroj" (*source media*), "Servis" (*service*), "Klíčová slova" (*keywords*), "Hledaný text" (*searched text*) and "Délka zprávy od/do" (*length of the document from/to*); the button captions: "Hledat" (*search*), "Zobrazit data" (*show the data*) and "Zrušit (*cancel*). The large text area contains a query example in the SOLR syntax.

5 System Evaluation

We use two different real ČTK datasets to evaluate the performance of the proposed system. The first one is relatively small and is chosen because it

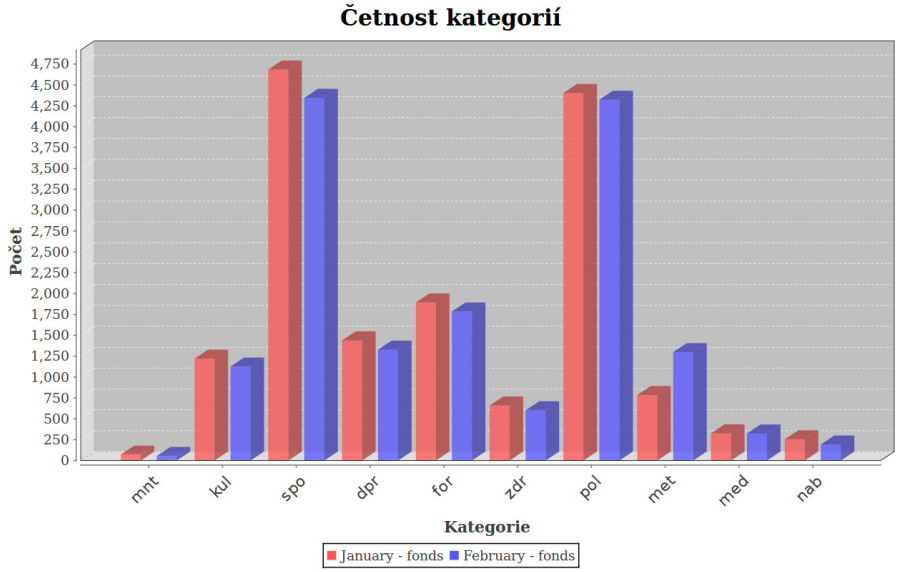

Fig. 3. Comparison example of the ČTK production with one Czech media. The main captions with English translations: "Četnost kategorií" (*category frequencies*), "Kategorie" (*categories*) and "Počet" (*number*). The horizontal axis depicts the categories while the vertical one shows the number of the documents containing the category.

is freely available for research purposes (download at `http://home.zcu.cz/~pkral/sw/`). Therefore, it is possible to compare our results with other state-of-the-art methods.

The second corpus is significantly larger. Unfortunately, this dataset is not publicly available yet, due to the commercial politics of the ČTK . However, we would like to show on this corpus that our system performance is independent to the size of the dataset.

5.1 Evaluation on the Free Czech ČTK Document Corpus

This corpus contains 11,955 documents composed of 2,974,040 individual words. The documents are manually annotated with at least one of 37 categories. Figure 4 shows the distribution of labels depending to the document numbers.

Table 1 shows the document classification scores in two cases: without and with the confidence measure. The five-folds cross validation procedure is used (20% of the corpus is reserved for testing and the remaining part for training of our system) with the standard Precision (*Prec*), Recall (*Rec*) and F-measure (*F-mes*) [24] metrics.

The first line of the table shows the classification scores without the confidence measure module. The two following lines depict the results of the *absolute*

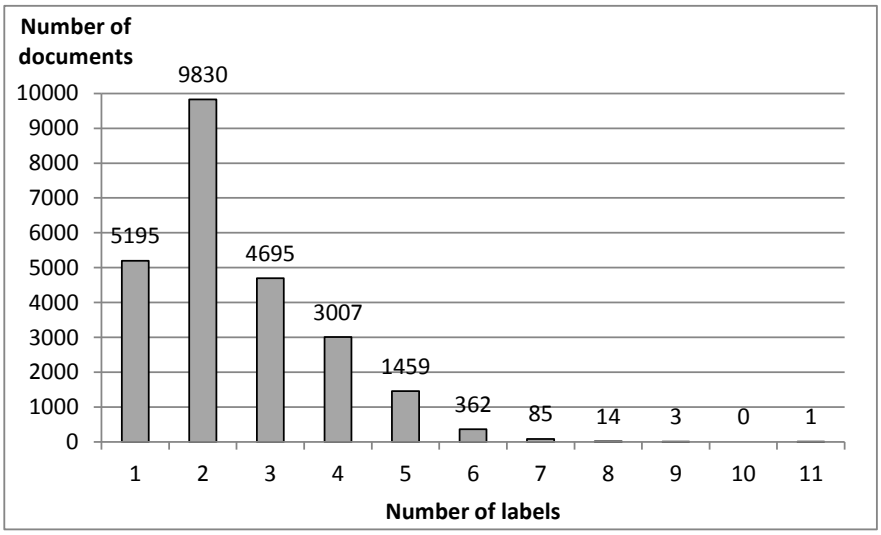

Fig. 4. Distribution of label numbers for documents in the ČTK corpus

and *relative confidence value* methods. The optimal values of the thresholds $T1$ and $T2$ are set to 0.99. These values are found experimentally using small development corpus. The last line shows the results of the composed supervised approach based on the MLP classifier.

It is clearly visible that every individual confidence measure method improves the classification results. The improvement is then further significantly increased when the MLP is used to combine the two measures. Therefore, this approach is integrated into our system.

Table 1. Classification results without / with the proposed confidence measures [in %] for the free Czech ČTK document corpus

Confidence Measure Approach	**Prec**	**Rec**	**F-mes**
-	89.0	75.6	81.7
Absolute confidence value	93.8	78.3	85.3
Relative confidence value	94.3	79.4	86.2
Composed supervised approach (MLP)	97.4	99.3	98.3

5.2 Evaluation on the Private Czech ČTK Document Corpus

This corpus contains 120,000 documents: 20,000 documents is used for training and the testing part is composed of the remaining 100,000 documents. The num-

ber of classes is similar as in the previous case, i.e. 37 ones. The distribution of the document labels is also similar. The documents are collected from one Czech newspaper for a period of 6 months.

Table 2 details the results of this experiment. This table clearly shows that the document number does not influence the recognition accuracy. When the best supervised confidence measure is used, the resulting F-measure is reduced only slightly. This experiment confirms high robustness of our system.

Table 2. Classification results without / with the proposed confidence measures [in %] for the private Czech ČTK document corpus

Confidence Measure Approach	**Prec**	**Rec**	**F-mes**
-	85.9	68.0	77.1
Absolute confidence value	92.8	74.4	82.6
Relative confidence value	93.4	75.6	83.6
Composed supervised approach (MLP)	97.1	98.5	97.8

6 Conclusions and Future Work

In this paper, we presented a novel experimental multi-label document classification and analysis system called SAPKOS. The system will be used by the Czech news Agency to save human resources in the task of annotation of newspaper articles with the topics. Another important functionality is an automatic comparison of the ČTK production with the newspaper articles from popular Czech media. This analysis will be used to adapt ČTK production to better correspond to the today's market requirements.

An interesting contribution is that, to the best of our knowledge, no other automatic Czech document classification system exists. The system accuracy is very high (F-measure 98.3% on the free and 97.8% on the private ČTK document corpus). These results are obtained due to the unique system architecture which integrates a maximum entropy based classifier with the novel confidence measure.

The current version of the system is not optimized to the speed and memory requirements. Therefore, our first perspective consists in the optimization of the system in this way. The system is adapted to the Czech language. However, the majority of the methods is language independent. Therefore another perspective consists in the adaptation of the system to the other languages.

Acknowledgements

This work has been partly supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports. We would like also 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. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management. pp. 679–694 (2004)
3. Brychcín, T., Král, P.: Novel unsupervised features for Czech multi-label document classification. In: 13th Mexican International Conference on Artificial Intelligence (MICAI 2014). pp. 70–79. Springer, Tuxtla Gutierrez, Chiapas, Mexic (16-22 November 2014)
4. Chandrasekar, R., Srinivas, B.: Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging (1996)
5. 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
6. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)
7. Glass, J., Derr, E.: Document similarity detection and classification system (Mar 17 2005), http://www.google.com/patents/US20050060643, uS Patent App. 10/710,918
8. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis 56(3), 741–751 (2012)
9. 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)
10. 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)
11. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2121–2133 (2012)
12. Jiang, H.: Confidence measures for speech recognition: A survey. Speech Communication 45(4), 455–470 (2005)
13. Konkol, M.: Brainy: A machine learning library. In: Artificial Intelligence and Soft Computing, Lecture Notes in Computer Science, vol. 8468. Springer Berlin Heidelberg (2014)
14. Král, P.: Named entities as new features for Czech document classification. In: 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014). vol. 8404 LNCS, pp. 417–427. Kathmandu, Nepal (6-12 April 2014), http://link.springer.com/chapter/10.1007/978-3-642-54903-8_35
15. Kral, P., Lenc, L.: Confidence measure for Czech document classification. In: 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2015). pp. 525–534. Springer (14-20 April 2015)
16. Lamirel, J.C., Cuxac, P., Chivukula, A.S., Hajlaoui, K.: Optimizing text classification through efficient feature selection based on quality metric. Journal of Intelligent Information Systems pp. 1–18 (2014)

17. Li, F., Wechsler, H.: Open world face recognition with credibility and confidence measures. In: Audio-and Video-Based Biometric Person Authentication. pp. 462–469. Springer (2003)
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. Marukatat, S., Artières, T., Gallinari, P., Dorizzi, B.: Rejection measures for handwriting sentence recognition. In: Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on. pp. 24–29. IEEE (2002)
20. Nagatsuka, T., Miyachi, T., Shimada, A., Takeya, K., Kemmochi, E., Nakajima, A., Yamasaki, M., Fujita, K.: Document classification system and method for classifying a document according to contents of the document (Mar 20 2007), http://www.google.com/patents/US7194471, uS Patent 7,194,471
21. 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
22. Nouretdinov, I., Costafreda, S.G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., Fu, C.H.: Machine learning classification with confidence: application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression. Neuroimage 56(2), 809–813 (2011)
23. Papadopoulos, H.: A cross-conformal predictor for multi-label classification. In: Artificial Intelligence Applications and Innovations, pp. 241–250. Springer (2014)
24. Powers, D.: Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies 2(1), 37–63 (2011)
25. 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
26. 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. pp. 457–465. KDD '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2020408.2020481
27. Rashad, M.A., El-Deeb, H., Fakhr, M.W.: Document classification using enhanced grid based clustering algorithm. In: New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering, pp. 207–215. Springer (2015)
28. Rocha, O.R., Vagliano, I., Figueroa, C., Cairo, F., Futia, G., Licciardi, C.A., Marengo, M., Morando, F.: Semantic annotation and classification in practice. IT Professional 17(2), 33–39 (2015)
29. Rodrigues, F.M., de M Santos, A., Canuto, A.M.: Using confidence values in multi-label classification problems with semi-supervised learning. In: Neural Networks (IJCNN), The 2013 International Joint Conference on. pp. 1–8. IEEE (2013)
30. Senay, G., Linares, G.: Confidence measure for speech indexing based on latent dirichlet allocation. In: INTERSPEECH (2012)
31. Servin, B., de Givry, S., Faraut, T.: Statistical confidence measures for genome maps: application to the validation of genome assemblies. Bioinformatics 26(24), 3035–3042 (2010)

32. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3(3), 1–13 (2007)
33. Wnek, J.: Multi-strategy document classification system and method (Dec 28 2006), http://www.google.com/patents/US20060294101, uS Patent App. 11/473,131
34. 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)