Confidence Measure for Czech Document Classification

Pavel Král[1,2] and Ladislav Lenc[1,2]

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

Abstract. This paper deals with automatic document classification in the context of a real application for the Czech News Agency (ČTK). The accuracy our classifier is high, however it is still important to improve the classification results. The main goal of this paper is thus to propose novel confidence measure approaches in order to detect and remove incorrectly classified samples. Two proposed methods are based on the *posterior* class probability and the third one is a supervised approach which uses another classifier to determine if the result is correct. The methods are evaluated on a Czech newspaper corpus. We experimentally show that it is beneficial to integrate the novel approaches into the document classification task because they significantly improve the classification accuracy.

1 Introduction

Automatic document classification is extremely important for information organization, storage and retrieval because the amount of electronic text documents is growing extremely rapidly. Multi-label document classification becomes currently significantly more important than the single-label classification because it usually corresponds better to the requirements of real applications.

Previously, we have developed an experimental multi-label document classification system for the Czech News Agency (ČTK)[1] based on the Maximum entropy classifier. The main goal of this system is to replace the manual annotation of the newspaper documents which is very expensive and time consuming. The resulting F-measure value of this system is higher than 80%, however this value is still far from perfect.

Therefore, in this paper, we propose a way how to detect incorrectly classified examples in order to improve the final classification score. Three novel Confidence Measure (CM) approaches are proposed, compared and evaluated for this task. The first two confidence measures are based on the *posterior* class probability. Then, we propose a supervised CM approach that combines these two methods by a classifier.

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

It is worthy of attention, that the confidence measure was never previously integrated to the Czech document classification. Moreover, to the best of our knowledge, no similar confidence measure approach in multi-label document classification field exists.

Section 2 is a short overview of the document classification and confidence measure approaches. Section 3 describes our document classification and confidence measure methods. Section 4 deals with the realized experiments on the ČTK corpus. We also discuss here the obtained results. In the last section, we conclude the research results and propose some future research directions.

2 Related Work

This section is composed of two parts. The document classification is described in the first one, while the second one is focused on the confidence measure task itself.

2.1 Document Classification

Document classification is usually based on supervised machine learning methods that exploit an annotated corpus to train a classifier which then assigns the classes of unlabelled documents. The most of works use Vector Space Model (VSM), which usually represents each document with a vector of all word occurrences weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

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 been introduced [1–3] to solve this problem.

Furthermore, a better document representation should help to decrease the feature vector dimension, e.g. using lexical and syntactic features as shown in [4]. Chandrasekar et al. further show in [5] that it is beneficial to use POS-tag filtration in order to represent a document more accurately. The authors of [6] and [7] use a set of linguistic features, however they do not improve the document classification accuracy.

More recently, some interesting approaches based on Latent Dirichlet Allocation (L-LDA) [8, 9] have been introduced. Another method exploits partial labels to discover latent topics [10]. Principal Component Analysis (PCA) [11] incorporating semantic concepts [12] has been also used for the document classification. Semi-supervised approaches, which progressively augment labelled corpus with unlabelled documents [13], have also been proposed.

The most of the proposed approaches is focused on English and only few works deal with Czech language. Hrala et al. use in [14] lemmatization and Part-Of-Speech (POS) filtering for a precise representation of Czech documents. In [15], three different multi-label classification approaches are compared and evaluated. The other recent works propose novel features based on the named entities [16] or on the unsupervised machine learning [9].

2.2 Confidence Measure

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 [17–20] and is mostly based on the *posterior* class probability. However, it can be successfully used in another research areas as shown in [21] for genome maps construction, in [22] for stereo vision, in [23] for handwriting sentence recognition or in [24] for automatic face recognition.

Another approach related to the confidence measure is proposed by Proedrou et al. in the pattern recognition task [25]. The authors use a classifier based on the nearest neighbours algorithm. Their confidence measure is based on the algorithmic theory of randomness and on transductive learning.

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 [26]. 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) [27]. CP assigns a reliable measure of confidence and is used as a complement of machine learning algorithms. Author of [28] 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.

The above mentioned approaches apply the confidence measures on other types of the data. Moreover, to the best of our knowledge, no similar confidence measure approach in multi-label document classification field exists.

3 Document Classification with Confidence Measure

The following sections are focused on our feature set, multi-label document classification approach and particularly on the proposed confidence measure methods.

3.1 Feature Set & Classification

The feature set is created according to Brychcín et al. [9]. They are used because the authors experimentally proved that the additional unsupervised features significantly improve classification results.

- **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** – 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.

For multi-label classification, we use an efficient approach presented by Tsoumakas et al. in [29]. 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 iff the label is included in the document, $\neg l$ otherwise). The classification result is given by the following equation:

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

The Maximum Entropy (ME) [30] model is used for classification.

3.2 Confidence Measure

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:

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

Only the classification results with $\varDelta P > T2$ 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. A 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.

In order to identify the best performing topology, several MLP configurations are built and evaluated. The MLP topologies will be described in detail in the experimental section.

4 Experiments

4.1 Tools and Corpus

For implementation of the multi-label classifier we used Brainy [31] implementation of Maximum entropy classifier. It has been chosen mainly because of our experience with this tool.

As already stated, the results of this work shall be used by the ČTK. Therefore, for the following experiments we used the Czech text documents provided by the ČTK. 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/`.

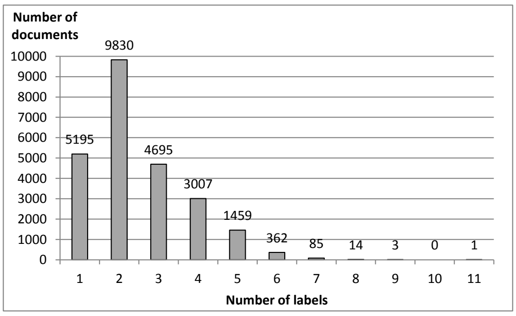

Fig. 1. Distribution of the documents depending on the number of labels

We use the five-folds cross validation procedure for all following experiments, where 20% of the corpus is reserved for testing and the remaining part for training of our models. For evaluation of the document classification accuracy, we use the standard Precision, Recall and F-measure (*F-mes*), also called F1-score, metrics [32]. The confidence interval of the experimental results is 0.6% at a confidence level of 0.95.

4.2 Experimental Results

ROC Curves of the Proposed Approaches As in many other articles in the confidence measure field, we will use the Receiver Operating Characteristic (ROC) curve [33] for evaluation of our CM methods. This curve clearly shows the relationship between the true positive and the false positive rate for different values of the *acceptance* threshold.

Figure 2 depicts the performance of the *absolute confidence value* method, while the results of the *relative confidence value* approach are given in Figure 3. These figures demonstrate that both approaches are suitable for our task in order to identify incorrectly classified documents. These figures further show, that the *relative confidence value* method slightly outperforms the *absolute confidence value* approach.

Better accuracy of this approach can be explained by the fact that the significantly higher difference in the *posterior* probabilities (between l and $\neg l$ classes) is a better metrics than the simple absolute value of this probability.

Note that this evaluation can be done only for the first two proposed methods which depend on the acceptance threshold. The third approach will be evaluated directly by the F-measure metrics.

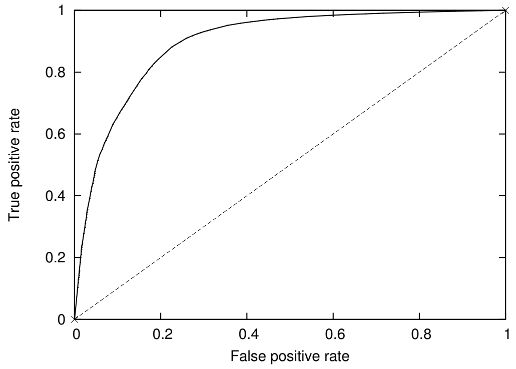

Fig. 2. ROC curve for the *absolute confidence value* method.

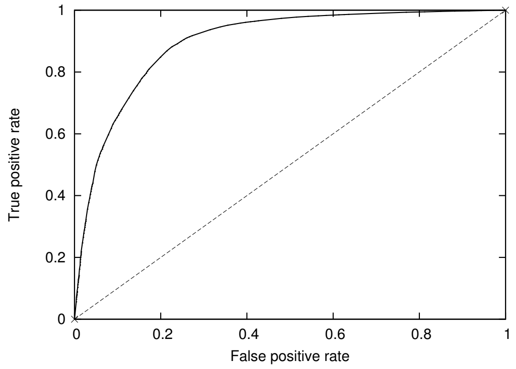

Fig. 3. ROC curve for the *relative confidence value* method.

Dependency of the F-measure on the Acceptance Threshold We deal with the multi-label classification task. The proposed confidence measure approaches thus significantly influence the resulting F-measure score. In this experiment, we would like to identify optimal acceptance thresholds for both CM methods.

Figure 4 shows the dependency of the F-measure value on the acceptance threshold for the *absolute confidence value* method, while the Figure 5 depicts the same dependency for the *relative confidence value* approach. These curves show that both optimal threshold values are close to 1. We can conclude that the correct classification must be associated with the significantly high level of the *posterior* probability (or significantly high difference between $P(l|F)$ and $P(\neg l|F)$ probability values).

Similarly as in the previous experiment, this evaluation is realized only for two first CM methods.

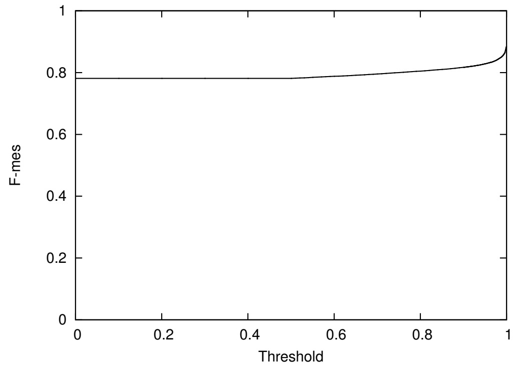

Fig. 4. Dependency of the F-measure on the acceptancce threshold for the *absolute confidence value* method.

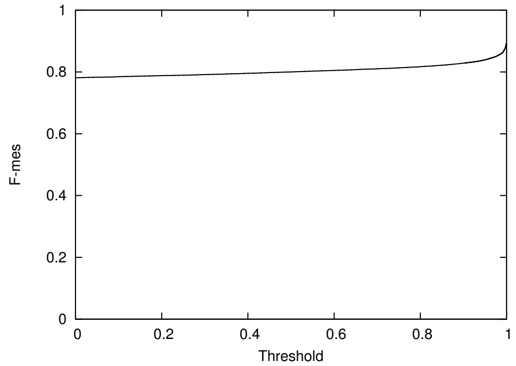

Fig. 5. Dependency of the F-measure on the acceptancce threshold for the *relative confidence value* method.

Classification Results without and with the Proposed Confidence Measure Approaches In the last experiment, we show the document classification scores in two cases: without and with the confidence measure. We also evaluate and compare the performance of the proposed confidence measure methods. As already stated, we use the standard Precision (*Prec*), Recall (*Rec*) and F-measure (*F-mes*) metrics for evaluation.

The results of this experiments are given by Table 1. The first line shows the classification scores without any confidence measure (the *baseline*). The two following lines depict the results of the *absolute* and *relative confidence value* methods. The optimal values of the thresholds $T1$ and $T2$ are based on the results of the previous experiment and are set in both cases to 0.99. The last line shows the results of the composed supervised approach which uses an MLP classifier. We set experimentally the following MLP topology 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*).

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.

Table 1. Classification results without / with the proposed confidence measures [in %]

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 Conclusions and Future Work

In this paper, we proposed three confidence measure methods and integrated them into multi-label document classification scenario. The first two measures are based on the *posterior* class probability of the output of our binary classifiers, while the third method is a supervised one and incorporates an MLP to decide whether the classification is correct or not. The methods are evaluated on the Czech ČTK corpus of the newspaper text documents. The experiments show that all these measures improve significantly the classification results. Moreover, we further show that the composed supervised CM approach gives the best classification score. The improvement over the baseline (no CM used) reaches 16.6% in the absolute value when this approach is used. Therefore, we conclude that the confidence measure approach will be integrated into our document classification system.

The first perspective is proposing a semi-supervised confidence measure. In this approach, the CM model will be progressively adapted according to the processed data. We will further integrate other suitable individual measures into our composed MLP approach (use for example the so called *predictor* features [19]). The last perspective consists in evaluation of our proposed methods on different languages and language families.

Acknowledgements

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

References

1. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research **3** (2003) 1289–1305
2. 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
3. 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 (2014) 1–18
4. 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
5. Chandrasekar, R., Srinivas, B.: Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging. (1996)
6. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In McDonald, S., Tait, J., eds.: Advances in Information Retrieval. Volume 2997 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2004) 181–196
7. 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. Volume 5., IEEE (2005) 3183–3188
8. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 248–256
9. Brychcín, T., Král, P.: Novel unsupervised features for Czech multi-label document classification. In: 13th Mexican International Conference on Artificial Intelligence (MICAI 2014), Tuxtla Gutierrez, Chiapas, Mexic, Springer (2014) 70–79
10. Ramage, D., Manning, C.D., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11, New York, NY, USA, ACM (2011) 457–465
11. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis **56** (2012) 741–751
12. Yun, J., Jing, L., J., Y., Huang, H.: A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications **39** (2012) 2035–2046
13. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Mach. Learn. **39** (2000) 103–134
14. Hrala, M., Král, P.: Evaluation of the Document Classification Approaches. In: 8th International Conference on Computer Recognition Systems (CORES 2013), Milkow, Poland, Springer (2013) 877–885
15. Hrala, M., Kral, P.: Multi-label document classification in Czech. In: 16th International conference on Text, Speech and Dialogue (TSD 2013), Pilsen, Czech Republic, Springer (2013) 343–351

16. Král, P.: Named entities as new features for Czech document classification. In: 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014). Volume 8404 LNCS., Kathmandu, Nepal (2014) 417–427
17. Senay, G., Linares, G., Lecouteux, B.: A segment-level confidence measure for spoken document retrieval. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE (2011) 5548–5551
18. Senay, G., Linares, G.: Confidence measure for speech indexing based on latent dirichlet allocation. In: INTERSPEECH. (2012)
19. Jiang, H.: Confidence measures for speech recognition: A survey. Speech Communication **45** (2005) 455–470
20. Wessel, F., Schluter, R., Macherey, K., Ney, H.: Confidence measures for large vocabulary continuous speech recognition. Speech and Audio Processing, IEEE Transactions on **9** (2001) 288–298
21. Servin, B., de Givry, S., Faraut, T.: Statistical confidence measures for genome maps: application to the validation of genome assemblies. Bioinformatics **26** (2010) 3035–3042
22. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2121–2133
23. 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, IEEE (2002) 24–29
24. Li, F., Wechsler, H.: Open world face recognition with credibility and confidence measures. In: Audio-and Video-Based Biometric Person Authentication, Springer (2003) 462–469
25. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. In: ECML'02. (2002) 381–390
26. 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, IEEE (2013) 1–8
27. 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** (2011) 809–813
28. Papadopoulos, H.: A cross-conformal predictor for multi-label classification. In: Artificial Intelligence Applications and Innovations. Springer (2014) 241–250
29. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3** (2007) 1–13
30. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics **22** (1996) 39–71
31. Konkol, M.: Brainy: A machine learning library. In: Artificial Intelligence and Soft Computing. Volume 8468 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2014)
32. Powers, D.: Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies **2** (2011) 37–63
33. Brown, C.D., Davis, H.T.: Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems **80** (2006) 24–38