Comparison of Score Normalization Methods Applied to Multi-label Classification

Lucie Skorkovská, Zbyněk Zajíc, Luděk Müller
University of West Bohemia, Faculty of Applied Sciences,
New Technologies for the Information Society, Univerzitní 22, 306 14 Pilsen, Czech Republic
Email: lskorkov@ntis.zcu.cz, zzajic@ntis.zcu.cz, muller@ntis.zcu.cz

***Abstract*—Our paper deals with the multi-label text classification of the newspaper articles, where the classifier must decide if a document does or does not belong to each topic from the predefined topic set. A generative classifier is used to tackle this task and the problem with finding a threshold for the positive classification is mainly addressed. This threshold can vary for each document depending on the content of the document (words used, length of the document, etc.). An extensive comparison of the score normalization methods, primary proposed in the speaker identification/verification task, for robustly finding the threshold defining the boundary between the "correct" and the "incorrect" topics of a document is presented. Score normalization methods (based on World Model and Unconstrained Cohort Normalization) applied to the topic identification task has shown an improvement of results in our former experiments, therefore in this paper an in-depth experiments with more score normalization techniques applied to the multi-label classification were performed. Thorough analysis of the effects of the various parameters setting is presented.**

Keywords-multi-label text classification, topic identification, Naive Bayes classification, score normalization

I. Introduction

A significant attention has been payed to the multi-label classification over the past few years. In many modern areas including newspaper topic identification, social network comments classification, web content topical organization, email routing but also images and video annotation or gene functional classification recently emerged the need for not only multi-class but also multi-label classification, where the classifier must decide if a document does or does not belong to each topic from the predefined topic set.

This paper deals with the multi-label newspaper article classification used in a real-life application for acquisition and storing huge amounts of data [1] designed to gather the training data for the estimation of the parameters of statistical language models for natural language processing. Since it has been shown that not only the size of the training data is important, but also the right scope of the language models training texts is needed [2], the topic identification algorithm is used for large scale language modeling data filtering [3]. The use of story topics for language model adaptation was shown to lower the language model perplexity and the word error rate of the ASR system [4].

Usually, the multi-label classification is handled through the set of binary classifiers, one for each label [5]. For each classifier a threshold for the positive classification must be set. This may not be a problem for a classification task with a small set of topics (ten topics for example), where for each one of them a sufficient amount of training data is available, but in a real application the set of topics is usually quite large (450 topics in our case) and for some of them very little training data can be obtained. A possible alternative is to use a single generative classifier like Naive Bayes classifier [6][7], which outputs a distribution of likelihood scores of the document belonging to the topics from the topic set. In this approach only a single threshold defining the boundary between the "correct" and the "incorrect" topics of a document has to be set. Still the selection of this one threshold is difficult, since it may vary depending on the content of each document, it can not be fixed for the whole document collection, but a dynamically set threshold is needed.

The application of score normalization methods (World Model and Unconstrained Cohort Normalization) from the open-set text-independent speaker identification domain to the topic identification task has shown promising results in robustly finding the threshold in our former experiments [8][9], therefore in this paper an in-depth experiments with more score normalization techniques applied to the multi-label classification were performed.

The remainder of this paper is organized as follows. Section II presents a short summary of the main approaches to the multi-label classification and Section II-A focuses on the related work on the problem of the threshold definition for generative classifiers. An application of various score normalization methods on the multi-label topic identification problem is described in Section III. Section IV describes the experimental setup, the evaluation measures and presents and discusses the experimental results. Finally, the conclusions are given in Section V.

II. Multi-label Text Classification

The multi-label classification methods can be divided into two main categories - *data transformation methods* and *algorithm adaptation methods* according to [10], where a detailed overview of the existing methods was given. The methods of the first group transform the problem into the single-label classification problem and the methods in the second group extend the existing algorithms to handle the multi-label data directly. The existing *data transformation methods* can be more divided into the three main approaches:

The easiest way is to transform the multi-label data set into single-label by either selecting only one label for each data instance or by discarding every multi-label data instance

from the set. Another option is to consider each set of labels as one label together [7][11].

The most common option is to train a binary classifier for each class. The labels for which the binary classifier yields a positive result are then assigned to the tested data item. The disadvantage of this method is that you have to transform the data set into $|L|$ data sets, where L is the set of possible labels, containing only the positive and negative examples. The second disadvantage is that you have to find the threshold for each binary classifier. This method was used for example in [12][13].

Another possibility is to decompose each training data with n labels into n data items each with only one label. One generative classifier with the distribution of likelihoods for all labels is learned from the transformed data set. The distribution is then processed to find the correct labels of the data item. This approach is used in the works [7][14] and also in our experiments.

A. Threshold Definition for Generative Classifiers

A related work on the problem how to select the set of correct topics from the output distribution of a generative classifier is presented in this section. A straightforward approach is to select the labels for which the likelihood is greater than a specific threshold or select a predefined number of topics. In the work [6] only the one best label is assigned to each news article. In our later work [3], we selected 3 topics for each article. In the work [7] this problem is bypassed by creating a mixture topic model from all possible topic subsets and then choosing the subset for which the corresponding mixture model has achieved the maximum likelihood.

To our knowledge, the only work concerning the finding of a threshold for choosing the correct topics in the distribution output of a classifier is described in [14]. A dynamic threshold is set as the mean plus one standard deviation of the topic likelihoods. The assumption is that topics that have a likelihood greater than this threshold are the best choices for the article.

III. Score Normalization Applied to Multi-label Topic Identification

The topic identification problem is quite similar to the open-set text-independent speaker identification (OSTI-SI) problem. The speaker identification is described as a twofold problem: First, the speaker model best matching the utterance has to be found and secondly, it has to be decided, if the utterance has really been produced by this best-matching model or by some other speaker outside the set. The difficulty in this task is that the speakers are not obliged to provide the same utterance that the system was trained on.

The document classification problem can be described in the same way: First, we need to find the topic models which have the best likelihood score for the tested document and second, we have to choose only the correct topic models which really generated the document. The only difference in topic identification is that we try to find more than one correct topic model. The normalization methods from OSTI-SI can be used in the same way, but they have to be applied to all topic models likelihoods.

A. Naive Bayes Classification

For the first phase of the topic identification the multinomial Naive Bayes (NB) classifier is used, which is formally equal to the language modeling based approach in the information retrieval [15]. The reasons for the selection of NB classifier are more addressed in Section IV-A. Each topic is defined by its unigram language model and a probability of a document A being generated by a topic model T is expressed by a conditional model $P(T|A)$. Using the Bayes' theorem, leaving out the prior probability of an article $P(A)$ and under the "naive" conditional independence assumption, the following equation can be written:

$$P(T|A) \propto \frac{P(T)p(A|T)}{P(A)} \propto p(A|T) = \prod_{t \in A} p(t|T), \quad (1)$$

where $P(T)$ is the prior probability of the topic T, which can be estimated as a relative frequency of the articles belonging to a topic in the training data, or considered uniform and be left out as in our case [8]. The distribution of topic likelihoods $p(A|T)$ is then used to find the most likely topics of an article. The probability $p(t|T)$ is estimated as the relative frequency of the term t in the training data of the topic T. The uniform prior smoothing was used in the estimation of $p(t|T)$.

B. Score Normalization

As a result of the NB classification we get the distribution of the topic likelihoods $p(A|T)$ and we now have to find the threshold for the selection of the correct topics of an article. Score normalization methods have been used to tackle the problem of the compensation for the distortions in the utterances in the second phase of the open-set text-independent speaker identification problem [16]. In the topic identification task, the likelihood score of a topic obtained from the classifier is dependent on the characteristics of the document (words used, length of the document, ...). Similarly as in the OSTI-SI [16] we can define the decision formula:

$$P(T_C|A) > P(T_I|A) \rightarrow A \in T_C \quad \text{else} \quad A \in T_I, \quad (2)$$

where $P(T_C|A)$ is the score given by the correct topic model T_C and $P(T_I|A)$ is the score given by the incorrect topic model T_I. By the application of the Bayes' theorem, formula (2) can be rewritten as:

$$\frac{p(A|T_C)}{p(A|T_I)} > \frac{P(T_I)}{P(T_C)} \rightarrow A \in T_C \quad \text{else} \quad A \in T_I, \quad (3)$$

where $l(A) = \frac{p(A|T_C)}{p(A|T_I)}$ is the normalized likelihood score and $\theta = \frac{P(T_I)}{P(T_C)}$ is a threshold that has to be determined. Setting this threshold θ a priori is a difficult task, since we do not know the prior probabilities $P(T_I)$ and $P(T_C)$. Similarly as in the OSTI-SI task the topic set is open - an article belonging to a topic not contained in our set can easily occur.

A frequently used form to represent the normalization process is the following [16]:

$$L(A) = \log p(A|T_C) - \log p(A|T_I). \quad (4)$$

The score $\log p(A|T_C)$ is affected by the document characteristics as well as the score $\log p(A|T_I)$. Thus, the distance between them should stand constant for various documents and finding the threshold experimentally for the whole collection of documents can be achieved.

Since the normalization score $\log p(A|T_I)$ of an incorrect topic is not known, there are several possibilities how to approximate it.

1) World Model Normalization (WMN): In the OSTI-SI task, the unknown model can be approximated by a model based on a very large number of speakers, commonly called the world model [17]. This method was adopted as the General topic model normalization (GTMN) in [8]. The model T_I can be approximated as the General topic model G, which was created as a language model from all documents in the training collection. The normalization score of a topic model T_I is defined as:

$$\log p(A|T_I) = \log p(A|G). \tag{5}$$

2) Unconstrained Cohort Normalization (UCN): This score normalization method from OSTI-SI domain [18], was applied to the topic identification in our work [9]. For every topic model a set (cohort) of N similar models $C = \{T_1, ..., T_N\}$ is chosen. These models in the set C are the most competitive models with the reference topic model, i.e. models which yield the next N highest likelihood scores. The normalization score is given by:

$$\log p(A|T_I) = \log p(A|T_{UCN}) = \frac{1}{N}\sum_{n=1}^{N} \log p(A|T_n). \tag{6}$$

3) Cohort Normalization (CN): In cohort normalization method [19] a set C of similar models is chosen in advance - before the classification, when the tested article is not known. For OSTI-SI the most competitive models are chosen depending on the closeness in the speaker space [17][20]. In this work the competitiveness of two models is defined by its position in our topic tree (see Section IV-A for the topic set description). For the reference topic model the cohort C consists of the topic models on the same level with the same upper node. The set selected in such way can have different size N for a different topic model, because the size of the set C depends on the number of the competitive topics in the topic tree. The normalization score is given by the same formula as (6), only the selection of the set C is different.

4) Standardizing a Score Distribution: Another solution called Test normalization (T-norm) stated in [16] is to transform a score distribution, resulting from a different test conditions, into a standard form (we have assumed Gaussian score distributions). The transformation of the formula (4) has the form:

$$L(A) = (\log p(A|T_C) - \mu(A))/\sigma(A), \tag{7}$$

where $\mu(A)$ and $\sigma(A)$ are the mean and standard deviation of the whole topic likelihood distribution. This approach has similarities to WMN, the main difference here is the use of the standard deviation of the distribution.

5) Threshold Selection: Even when we have the topic likelihood score normalized, we still have to set the threshold θ in 3 for verifying the correctness of each topic in the list. Selecting a threshold defining the boundary between the correct and the incorrect topics in a list of normalized likelihood is more robust, because the normalization removes the influence of the various document characteristics. Since in our former experiments [8][9] we have successfully defined the threshold as 80% of the normalized score of the best scoring topic, the threshold θ will be similarly defined as the ratio k of the best normalized score. A thorough analysis of different parameters setting is presented in Section IV-D and the dependency between the threshold ratio k setting and the size of the set C for the UCN method is examined.

IV. Performed Experiments

All experiments were performed within the System for acquisition and storing data [1] designed to gather the training data for the estimation of the parameters of statistical language models for natural language processing. For the multi-label classification experiments the text preprocessing modules of the system were used. On each article a *tokenization*, *text normalization*, *vocabulary-based substitution* and *decapitalization* algorithms are applied. Automatic *text lemmatization* [21] is also applied in our work, since it has been shown to improve the results when dealing with sparse data [22][23] in highly inflected languages.

A. Topic Identification Module

The topic identification module uses a multinomial Naive Bayes classifier described in Section III-A, since based on the nature of our application (every day more than 600 new articles are downloaded containing more than 130 new topic training articles) we needed the topic identification algorithm which will be fast and can use the easily updatable statistics stored in the database tables as the trained classifier data [3][8].

The topics are chosen from a hierarchical system - topic tree created based on our expert findings in the topic distribution in the articles on the Czech news servers, it contains about 450 topics. The advantage of the hierarchical organization of the topics is currently used only for the selection of documents to be used as the training data for the estimation of statistical language models and now also for the selection of the most competitive models for the CN method. For the classification all topics are used only as the set of topics on an equal level. This is caused by the nature of the training data since we use as training data the real articles from the different news servers and we do not want to change it in any way. The authors of these articles to our knowledge do not use any topic hierarchy, or at least not strictly. Sometimes the articles have assigned also the more general topic for some detailed topic, but mostly it does not.

B. Evaluation Metrics

The commonly used evaluation metric in the multi-label classification is somewhat similar to the evaluation used in the information retrieval (IR), each tested article is considered a query in IR and precision and recall is computed for the answer

TABLE I. RESULTS OF APPLICATION OF SCORE NORMALIZATION METHODS IN COMPARISON TO OTHER THRESHOLD FINDING METHODS ON THE TEST SET

method parameters	**1 topic** -	**3 topics** -	**MpSD** -	**WMN** k=0.8	**CN** k=0.8	**UCN** k=0.7,N=15	**UCN** k=0.8,N=80	**T-norm** k=0.9
$P(H,D)$	0.7968	0.5702	0.0581	0.5746	0.5967	**0.6771**	0.6623	0.6667
$R(H,D)$	0.3080	0.5956	0.9520	0.6693	0.5256	**0.5825**	0.5909	0.5844
$F_1(H,D)$	0.4442	0.5826	0.1096	0.6183	0.5589	**0.6263**	0.6246	0.6228

topic set [5][11][12]. For the article set D and the classifier H precision ($P(H,D)$) and recall ($R(H,D)$) is computed:

$$P(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{T_C}{T_A} \quad R(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{T_C}{T_R}, \qquad (8)$$

where T_A is the number of topics assigned to the article, T_C is the number of correctly assigned topics and T_R is the number of the relevant reference topics. The $F_1(H,D)$-measure, which is used for the straightforward comparison of methods, is then computed from the $P(H,D)$ and $R(H,D)$ measures:

$$F_1(H,D) = 2\frac{P(H,D)\cdot R(H,D)}{P(H,D)+R(H,D)}. \qquad (9)$$

These metrics express the partial match of the classification result, for each data item being classified we obtain P and R values expressed as a percentage of the full match between the correct topics set and the assigned topics set.

C. Data Description

A collection separated from the whole corpus used in our previous work [8] was also used for the experiments. The collection contains 31k articles published in the year 2011(January to October) and is divided into 27k training and 4k test articles. The test articles were used in this work as development data for the experiments with the size of the cohort and the threshold selection. Another set of 5k articles from the year 2012 was separated to be used as final test data. The articles were not rearranged in any way, therefore all the test articles were published after the training articles.

D. Results

A thorough analysis of different parameters setting for all score normalization methods was done on the development 2011 collection. For the WMN, T-norm a CN method only the threshold has to be set. In Fig. 2b) the dependency of these methods on the different threshold ratio k can be seen, from which the best threshold can be selected. For the UCN method, the best combination of the threshold and the cohort size has to be found. As can be seen in Fig. 1 the dependency of the threshold is directly proportional to the cohort size, because the normalization score in (6) is bigger (an average from the higher topic likelihoods) for a smaller cohort size. Fig. 2a) shows the comparison of the dependency on the cohort size for two different settings of the threshold ratio. For the threshold of about 80% ($k = 0.8$) of the best score, the results are most stable for the different setting of the cohort size. On the other hand, better result can be found for the threshold of about

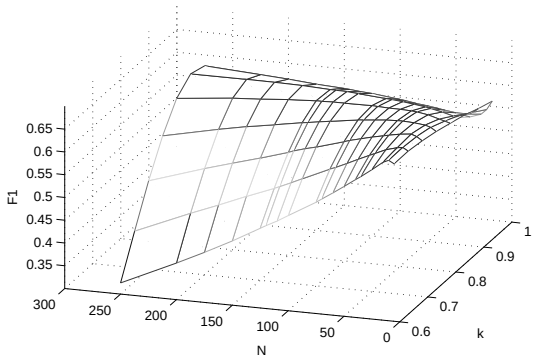

Fig. 1. Dependency of the UCN method on the size of the cohort N and threshold ratio k on the development set

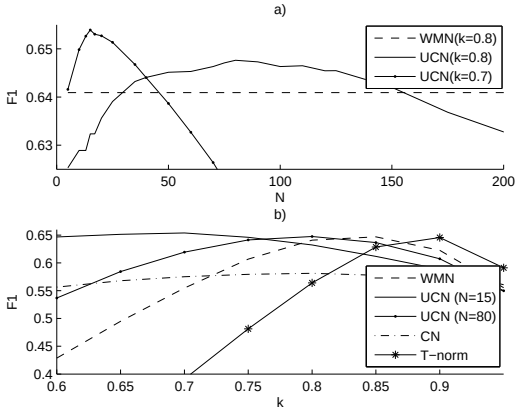

Fig. 2. Comparison of score normalization methods on the development set a) dependency on the size of the cohort N for the fixed threshold ratio $k = 0.7$ and 0.8 respectively b) dependency on the threshold ratio k (for the size of the cohort in UCN $N = 15$ and $N = 80$ respectively)

70% ($k = 0.7$) for a smaller cohort size ($N = 15$). From the different perspective for the smaller cohort size, the UCN method is more stable for different threshold selection than the UCN with bigger cohort size and also than other methods (see Fig. 2b)).

Table I shows the comparison of the results on the test 2012 collection. For the score normalization methods the size of the cohort and the threshold ratio yielding the best F_1-measure on the development set was selected. For the UCN two possibilities of the settings are presented, smaller cohort size $N = 15$ for the stability in the threshold selection and threshold ratio setting to $k = 0.8$ for the stability in the cohort size selection. The results are also compared to the previously used selection of 1 and 3 topics for each article [6], [3] resp. and setting the threshold as the mean plus one standard deviation (MpSD) of the topic likelihoods [14].

V. Conclusions

The score normalization methods from the OSTI-SI domain have shown significantly better results than other techniques used for threshold selection in multi-label document classification. The UCN method yields better results than the rest of the score normalization methods, furthermore the UCN method is more stable in the selection of the parameters setting, the selection of the threshold is most robust for the smaller cohort size ($N = 5 - 15$).

This article has shown that score normalization techniques are very useful in the multi-label classification task. Although we still have to set the threshold for verifying the correctness of the topics, the selection of a threshold defining the boundary between the correct and the incorrect topics is more robust, because the normalization removes the influence of the various document characteristics.

Acknowledgment

This research was funded by the Ministry of Culture Czech Republic, project No.DF12P01OVV022.

Acknowledgment also to the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

References

[1] J. Švec, J. Hoidekr, D. Soutner, and J. Vavruška, "Web text data mining for building large scale language modelling corpus," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, I. Habernal and V. Matoušek, Eds. Springer Berlin / Heidelberg, 2011, vol. 6836, pp. 356–363.

[2] J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. Byrne, J. Hajič, J. Mírovský, and S. Gustman, "Large vocabulary ASR for spontaneous Czech in the MALACH project," in *Proceedings of Eurospeech 2003*, Geneva, 2003, pp. 1821–1824.

[3] L. Skorkovská, P. Ircing, A. Pražák, and J. Lehečka, "Automatic topic identification for large scale language modeling data filtering," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, I. Habernal and V. Matoušek, Eds. Springer Berlin / Heidelberg, 2011, vol. 6836, pp. 64–71.

[4] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," in *Proceedings of Eurospeech*, 1997.

[5] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, Sep. 2012.

[6] A. D. Asy'arie and A. W. Pribadi, "Automatic news articles classification in indonesian language by using naive bayes classifier method," in *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*. New York, USA: ACM, 2009, pp. 658–662.

[7] A. K. McCallum, "Multi-label text classification with a mixture model trained by em," in *AAAI 99 Workshop on Text Learning*, 1999.

[8] L. Skorkovská, "Dynamic threshold selection method for multi-label newspaper topic identification," in *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science, I. Habernal and V. Matoušek, Eds. Springer Berlin Heidelberg, 2013, vol. 8082, pp. 209–216.

[9] L. Skorkovská and Z. Zajíc, "Score normalization methods applied to topic identification," in *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer Berlin Heidelberg, 2014, p. in press.

[10] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int J Data Warehousing and Mining*, vol. 2007, pp. 1–13, 2007.

[11] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," 2004.

[12] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *In Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2004, pp. 22–30.

[13] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Granular Computing, 2005 IEEE International Conference on*, vol. 2, 2005, pp. 718–721.

[14] D. B. Bracewell, J. Yan, F. Ren, and S. Kuroiwa, "Category classification and topic discovery of japanese and english news articles," *Electron. Notes Theor. Comput. Sci.*, vol. 225, pp. 51–65, 2009.

[15] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

[16] P. Sivakumaran, J. Fortuna, Ariyaeeinia, and A. M., "Score normalisation applied to open-set, text-independent speaker identification," in *Proceedings of Eurospeech 2003*, Geneva, 2003, pp. 2669–2672.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 13, pp. 42 – 54, 2000.

[19] Y. Zigel and A. Cohen, "On cohort selection for speaker verification," in *Proceedings of EUROSPEECH*, Geneva, 2003, pp. 2977–2980.

[20] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Second International Conference on Spoken Language Processing*, 1992.

[21] J. Kanis and L. Müller, "Automatic lemmatizer construction with focus on oov words lemmatization," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, V. Matoušek, P. Mautner, and T. Pavelka, Eds. Springer Berlin / Heidelberg, 2005, vol. 3658, pp. 742–742.

[22] P. Ircing and L. Müller, "Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006," in *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum*, ser. Lecture Notes in Computer Science, Alicante, Spain, 2007, pp. 759–765.

[23] J. Psutka, J. Švec, J. V. Psutka, J. Vaněk, A. Pražák, L. Šmídl, and P. Ircing, "System for fast lexical and phonetic spoken term detection in a czech cultural heritage archive," *EURASIP J. Audio, Speech and Music Processing*, vol. 2011, 2011.