First Experiments with Relevant Documents Selection for Blind Relevance Feedback in Spoken Document Retrieval

Lucie Skorkovská

University of West Bohemia, Faculty of Applied Sciences
New Technologies for the Information Society
Univerzitní 22, 306 14 Plzeň, Czech Republic
{lskorkov}@ntis.zcu.cz

Abstract. This paper presents our first experiments aimed at the automatic selection of the relevant documents for the blind relevance feedback method in speech information retrieval. Usually the relevant documents are selected only by simply determining the first *N* documents to be relevant. We consider this approach to be insufficient and we would try in this paper to outline the possibilities of the dynamical selection of the relevant documents for each query depending on the content of the retrieved documents instead of just blindly defining the number of the relevant documents to be used for the blind relevance feedback in advance. We have performed initial experiments with the application of the score normalization techniques used in the speaker identification task, which was successfully used in the multi-label classification task for finding the “correct” topics of a newspaper article in the output of a generative classifier. The experiments have shown promising results, therefore they will be used to define the possibilities of the subsequent research in this area.

Keywords: query expansion, blind relevance feedback, spoken document retrieval, score normalization

1 Introduction

The field of information retrieval (IR) has received a significant attention in the past years, mainly because of the development of World Wide Web and the rapidly increasing number of documents available in electronic form. Recently the Internet has been looked upon as an universal information media, more than the text information source it becomes the multimedia information source. Especially since large audio-visual databases are available on-line, the research in the field of information retrieval extends to the retrieval of speech content.

Experiments performed on the speech retrieval collections containing conversational speech [11][2][7] suggest that classic information retrieval methods alone are not sufficient enough for successful speech retrieval, especially when the collections contain speech data in other languages than English. The biggest issue here is that the query words are often not found in the documents from the collection. One cause of this problem is the high word error rate of the automatic speech recognition (ASR) causing the query words to be misrecognized. This problem can be dealt with through the use of ASR lattices for IR. The second cause is that the query words was actually not spoken in the recordings and thus are not contained in the documents. To deal with this issue the query expansion techniques are often used.

One of the possible query expansion methods often used in the IR field is the relevance feedback method. The idea is to take the information from the relevant documents retrieved in the first run of the search and use it to enhance the query with some new terms for the second run of the retrieval. The selection of the relevant documents can be done either by the user of the system or automatically without the human interaction - the method is then usually called the blind relevance feedback. The automatic selection is usually handled only by selecting the first *N* retrieved documents, which are considered to be relevant.

In this paper we will present the first experiments aimed at the better automatic selection of the relevant documents for the blind relevance feedback method. Our idea is to apply the score normalization techniques used in the speaker identification/verification task [12][14], which was successfully used in the multi-label classification task for finding the threshold between the "correct" and "incorrect" topics of a newspaper article in the output of a generative classifier [13], to dynamically select the relevant documents for each query depending on the content of the retrieved documents instead of just experimentally defining the number of the relevant documents to be used for the blind relevance feedback in advance.

2 Query Likelihood Model

Language modeling approach [8] was used as the information retrieval method for the experiments, specifically the query likelihood model with an linear interpolation of the unigram language model of the document with an unigram language model of the whole collection. The idea of this method is to create a language model M_d from each document d and then for each query q to find the model which most likely generated the query, that means to rank the documents according to the probability $P(d|q)$. The Bayes rule is used:

$$P(d|q) = P(q|d)P(d)/P(q), \tag{1}$$

where $P(q)$ is the same for all documents and the prior document probability $P(d)$ is uniform across all documents, so we can ignore both. We have left the probability of the query been generated by a document model $P(q|M_d)$, which can be estimated using the maximum likelihood estimate (MLE):

$$\hat{P}(q|M_d) = \prod_{t \in q} \frac{tf_{t,d}}{L_d}, \tag{2}$$

where $tf_{t,d}$ is the frequency of the term t in d and L_d is the total number of tokens in d. To deal with the sparse data for the generation of the M_d we have used the mixture model between the document-specific multinomial distribution and the multinomial distribution of the whole collection M_c with interpolation parameter λ. So the final equation for ranking the documents according to the query is:

$$P(d|q) \propto \prod_{t \in q} (\lambda P(t|M_d) + (1-\lambda) P(t|M_c)). \tag{3}$$

The retrieval performance of this IR model can differ for various levels of interpolation, therefore the λ parameter was set according to the experiments presented in [5] to the best results yielding value - $\lambda = 0.1$.

3 Query Expansion - Blind Relevance Feedback

Query expansion techniques based on the blind relevance feedback (BRF) has been shown to improve the results of the information retrieval. The idea behind the blind relevance feedback is that amongst the top retrieved documents most of them are relevant to the query and the information contained in them can be used to enhance the query for acquiring better retrieval results.

First, the initial retrieval run is performed, documents are ranked according to the query likelihood computed by (3). Then the top *N* documents are selected as relevant and the top *k* terms (according to some importance weight L_t, for example *tf-idf*) from them is extracted and used to enhance the query. The second retrieval run is then performed with the expanded query.

Since we are using the language modeling approach to the information retrieval, for the terms selection we have used the importance weight defined in [8]:

$$L_t = \sum_{d \in R} \log \frac{P(t|M_d)}{P(t|M_c)}, \tag{4}$$

where *R* is the set of relevant documents.

In the standard approach to the blind relevance feedback the number of documents and terms is defined experimentally in advance the same for all queries. In our experiments we would like to find the number of relevant documents for each query automatically by selecting the "true" relevant documents for each query to dynamically define the number of top retrieved documents to be used in BRF.

4 Score Normalization for Relevant Documents Selection

The score normalization methods from the open-set text-independent speaker identification (OSTI-SI) problem were successfully used in the task of the multi-label classification to select the relevant topics for each newspaper article [13] in the output of a generative classifier. This is the same problem as in the information retrieval task, where as the result only the ranked list of documents with their likelihoods is returned. Usually the idea is, that the user of the retrieval system will look though the top *N* documents and therefore the specific selection of which document is relevant and which not is not needed. On the contrary when the blind relevance feedback is used, the selection of the true relevant documents can be very useful.

This problem is quite similar to the OSTI-SI problem. Similarly as in the speaker identification, the relevant documents selection in the retrieval results can be 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 was the system trained on.

The relevant documents selection can be described in the same way: First, we need to retrieve the documents which have the best likelihood scores for the query and second, we have to choose only the relevant documents which really generated the query. The only difference is that we try to find more than one relevant document. The normalization methods from OSTI-SI can be used in the same way, but have to be applied to all documents likelihoods.

4.1 Score Normalization Methods

After the initial retrieval run, we have the ranked list of documents with their likelihoods computed by (3). We have to find the threshold for the selection of the relevant documents. A 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 [12]. In the IR task, the likelihood score of a document is dependent on the content of the query, therefore the beforehand set number of relevant documents is not suitable.

A frequently used form to represent the normalization process [12] can be modified for the IR task:

$$L(A) = \log P(d_R|q) - \log P(d_I|q), \tag{5}$$

where $P(d_R|q)$ is the score given by the relevant document and $P(d_I|q)$ is the score given by the irrelevant document. Since the normalization score $\log P(d_I|q)$ of an irrelevant document is not known, it has to be approximated.

World Model Normalization The unknown model d_I can be approximated by the collection model M_c which was created as a language model from all documents in the retrieval collection. This technique was inspired by the World Model normalization [10]. The normalization score of a model d_I is defined as:

$$\log P(d_I|q) = \log P(M_c|q). \tag{6}$$

Even when we have the likelihood scores normalized, we still have to set the threshold for verifying the relevance of each document in the list. Based on the experiments presented in [13] we have selected only the documents which are better scoring than the collection model and we have defined the threshold as 60% of the normalized score of the best scoring document. The documents which achieved better normalized score are selected as relevant. The threshold selected in this way was experimentally proven to be robust, the change in the range of percents does not influence the result.

5 Experiments

In this section the experiments with the score normalization method are presented. All experiments were performed on the spoken document retrieval collection.

5.1 Information Retrieval Collection

Our experiments were performed on the spoken document retrieval collection that was used in the Czech task of the Cross-Language Speech Retrieval track organized within the CLEF 2007 evaluation campaign [2]. This collection contains automatically transcribed spontaneous interviews (segmented by sliding a fixed-size window over the transcribed text into 22 581 "documents") and two sets of TREC-like topics - 29 training and 42 evaluation topics. Each topic consists of 3 fields - `<title>` (T), `<desc>` (D) and `<narr>` (N) (an example of a topic can be seen on Figure 1). Both topic sets were used for our first experiments and the queries were created from all terms from the fields T, D and N since is has been shown to achieve better results than when only T and D fields are used [3]. Stop words were omitted from all sets of query terms. All the terms were also lemmatized, since lemmatization was shown to improve the effectiveness of information retrieval in highly inflected languages (as is the Czech language) [1][9][3]. For the lemmatization an automatically trained lemmatizer described in [4][5] was used.

```
<top>
<num>1166
<title>Chasidismus
<desc>Chasidové a jejich nezlomná víra
<narr>Relevantní materiál by měl vypovídat o Chasidismu
v období před holokaustem, v průběhu holokaustu a po
něm. Informace o chasidských dynastiích a založených a
zničených geografických lokalitách.
</top>
```

Fig. 1. Example of a topic (query) from Czech task of the CLEF 2007 evaluation campaign

5.2 Evaluation Metrics

The mean Generalized Average Precision (mGAP) measure that was used in the CLEF 2007 Czech task was used as an evaluation measure. The measure (described in detail in [6]) is designed for the evaluation of the retrieval performance on the conversational speech data, where the topic shifts in the conversation are not separated as documents. The mGAP measure is based on the evaluation of the precision of finding the correct beginning of the relevant part of the data.

5.3 Results

This section shows the results for our first experiment with score normalization methods. For the standard blind relevance feedback we have chosen the settings used for BRF in the paper dealing with the experiments on this collection [3] - take first 20 documents as relevant and extract 5 terms with the best score for the query enhancement. On the training topic set the experiments with the selection of another number of top retrieved documents to be chosen as relevant for standard BRF was also performed.

As can be seen from the Table 1, for the training topic set the results for the BRF with the score normalization method used are better than with the standard BRF with the predefined number of documents. For the evaluation topic set the results with the standard BRF are even slightly worse than without BRF, but with the score normalization used, the results are better than without BRF. It can be also seen that the results for the standard BRF are almost the same for different number of documents. This is caused by the fact that for each query different number of documents is relevant. For one query the result for BRF with 10 documents was better than with 30 documents, for another one the other way around. The dynamic number of relevant documents chosen by the score normalization method deals with this problem.

Table 1. IR results (mGAP score) for no blind relevance feedback, with standard BRF and BRF with score normalization (SN). 5 terms were used to enhance each query in all cases.

query set / method	**no BRF**	**standard BRF**	**standard BRF**	**standard BRF**	**BRF with SN**
# of documents	**-**	**10**	**20**	**30**	**SN**
train TDN	0.0392	0.0436	0.0432	0.0438	**0.0442**
eval TDN	0.0255	-	0.0245	-	**0.0272**

6 Future Work

Since this were only our first experiments on this subject, there is a lot of future work which can be done. We plan to try different methods for the score normalization from the area of speaker identification task. The score normalization methods can also be tested with another IR method, for example the Vector Space method, where the Rocchio's relevance feedback can be used. We would like to use the score normalization method to dynamically find also the number of irrelevant documents used in Rocchio's relevance feedback formula. We would also like to try the query expansion with the different collection and use the terms extracted from the relevant documents from that collection (selected with the use of score normalization) to enhance the query.

The number of terms to be selected for query expansion was chosen the same as used in [3]. We have performed experiments on the training query set for the BRF with score normalization with different number k of terms to be selected, the results can be seen in Table 2. It can be seen that the number of terms significantly affects the retrieval results. The experiments on how to select this number automatically will also be the subject of our future research.

Table 2. IR results (mGAP score) for BRF with score normalization for different number *k* of terms selected.

query set / # of terms k	5	10	20
train TDN	0.0442	0.0480	0.0501

7 Conclusions

This article has shown the first experiments with the use of score normalization method for selection of the relevant documents for the blind relevance feedback in speech information retrieval. The result are showing that with the score normalization better retrieval results can be achieved than with the standard blind relevance feedback with the number of relevant documents set beforehand. We have also confirmed that the blind relevance feedback in any form is very useful in the speech information retrieval.

The retrieval results are for each query the best with different number of documents used (because the number of truly relevant documents is different for each query). In the standard BRF the number of relevant documents is set the same for all the queries, therefore the mean results for the set of queries can not be the best which can be achieved. The use of score normalization methods for the automatic dynamic selection of relevant documents for each query independently solves this problem.

Acknowledgments. The work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LM2010013 and by the grant of the University of West Bohemia, project No. SGS-2013-032.

References

1. Ircing, P., Müller, L.: 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. pp. 759–765. Lecture Notes in Computer Science, Alicante, Spain (2007)
2. Ircing, P., Pecina, P., Oard, D.W., Wang, J., White, R.W., Hoidekr, J.: Information Retrieval Test Collection for Searching Spontaneous Czech Speech. In: Proceedings of TSD 2007. pp. 439–446. Lecture Notes in Artificial Intelligence, Plzeň, Czech Republic (2007)
3. Ircing, P., Psutka, J., Vavruška, J.: What Can and Cannot Be Found in Czech Spontaneous Speech Using Document-Oriented IR Methods – UWB at CLEF 2007 CL-SR Track, pp. 712–718. Springer-Verlag, Berlin, Heidelberg (2008), `http://portal.acm.org/citation.cfm?id=1428850.1428952`
4. Kanis, J., Müller, L.: Automatic lemmatizer construction with focus on oov words lemmatization. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 3658, pp. 742–742. Springer Berlin / Heidelberg (2005)
5. Kanis, J., Skorkovská, L.: Comparison of different lemmatization approaches through the means of information retrieval performance. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010, LNCS, vol. 6231, pp. 93–100. Springer, Heidelberg (2010)
6. Liu, B., Oard, D.W.: One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In: Proceedings of the 29th annual international ACM

SIGIR conference on Research and development in information retrieval. pp. 673–674. SIGIR '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1148170.1148311

7. Mamou, J., Carmel, D., Hoory, R.: Spoken document retrieval from call-center conversations. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 51–58. SIGIR '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1148170.1148183
8. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 275–281. ACM, New York, NY, USA (1998)
9. Psutka, J., Švec, J., Psutka, J.V., Vaněk, J., Pražák, A., Šmídl, L., Ircing, P.: System for fast lexical and phonetic spoken term detection in a czech cultural heritage archive. EURASIP J. Audio, Speech and Music Processing 2011 (2011)
10. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. In: Digital Signal Processing. p. 2000 (2000)
11. Saraclar, M., Sproat, R.: Lattice-based search for spoken utterance retrieval. In: Proceedings of HLT-NAACL 2004. pp. 129–136 (2004)
12. Sivakumaran, P., Fortuna, J., Ariyaeeinia, M., A.: Score normalisation applied to open-set, text-independent speaker identification. In: Proceedings of Eurospeech 2003. pp. 2669–2672. Geneva (2003)
13. Skorkovská, L.: Dynamic threshold selection method for multi-label newspaper topic identification. In: Habernal, I., Matoušek, V. (eds.) Text, Speech, and Dialogue, Lecture Notes in Computer Science, vol. 8082, pp. 209–216. Springer Berlin Heidelberg (2013)
14. Zajíc, Z., Machlica, L., Padrta, A., Vaněk, J., Radová, V.: An expert system in speaker verification task. In: Proceedings of Interspeech. vol. 9, pp. 355–358. International Speech Communication Association, Brisbane, AU (2008)