Attempts to Search Czech Spontaneous Spoken Interviews - the University of West Bohemia at CLEF 2007 CL-SR track

Pavel Ircing and Luděk Müller
University of West Bohemia
{ircing, muller}@kky.zcu.cz

Abstract

The paper presents an overview of the system build and experiments performed for the CLEF 2007 CL-SR track by the University of West Bohemia. We have concentrated on the monolingual experiments using the Czech collection only. The approach that was successfully employed by our team in the last year's campaign (simple tf.idf model with blind relevance feedback, accompanied with solid linguistic preprocessing) was used again but the set of performed experiments was broadened.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Experimentation

Keywords

Speech Retrieval

1 Introduction

The Czech subtask of the CL-SR track, which was first introduced at CLEF 2006 campaign, is enormously challenging — let us repeat once again that the goal is to identify appropriate replay points (that is, the moments where the discussion about the queried topics starts) in a continuous stream of text generated by automatic transcription of spontaneous speech. Therefore, it is neither the standard document retrieval task (as there are no true documents defined) nor the fully-fledged speech retrieval (since the participants do not have the speech data nor the lattices, so they can't explore alternative hypotheses and must rely on one-best transcription). However, in order to lower the barrier of entry for teams proficient at classic document retrieval (or, for that matter, even total IR beginners), the last year's organisers prepared a so called Quickstart collection with artificially defined "documents" that were created by sliding 3-minute window over the stream of transcriptions with a 2-minute step (i.e., the consecutive documents have a one minute overlap).[1] The last year's Quickstart collection was further equipped with both manually

[1]It turned out later that the actual timing was different due to some faulty assumptions during the Quickstart collection design, but since the principle of the document creation remains the same, we will still use the "intended" time figures instead of the actual ones, just for the sake of readability.

and automatically generated keywords (see [5] for details) but they have shown itself to be of no benefit for IR performance [3](the former for the timing problems, the latter for the problems with their assignment that yet remain to be identified) and thus have been dropped from this year's data. The scripts for generating such Quickstart collection with variable window and overlap times were also included in the data release.

2 System description

Our current system largely builds upon the one that was successful in the last year's campaign [3], with only minor modifications and larger set of tested settings.

2.1 Linguistic preprocessing

Stemming (or lemmatization) is considered to be vital for good IR performance even in the case of weakly inflected languages such as English; thus it is probably even more crucial for Czech as the representative of the richly inflectional language family. This assumption was experimentally proven by our group in the last year's CLEF CL-SR track [3]. Thus we have used the same method of linguistic preprocessing, that is, the serial combination of Czech morphological analyser and tagger [2], which provides both the lemma and stem for each input word form, together with a detailed morphological tag. This tag (namely it's first position) is used for stop-word removal — we removed from indexing all the words that were tagged as prepositions, conjunctions, particles and interjections.

2.2 Retrieval

All our retrieval experiments were performed using the Lemur toolkit [1], which offers a variety of retrieval models. We have decided to stick to the *tf.idf* model where both documents and queries are represented as weighted term vectors $\vec{d_i} = (w_{i,1}, w_{i,2}, \cdots, w_{i,n})$ and $\vec{q_k} = (w_{k,1}, w_{k,2}, \cdots, w_{k,n})$, respectively ($n$ denotes the total number of distinct terms in the collection). The inner-product of such weighted term vectors then determines the similarity between individual documents and queries. There are many different formulas for computation of the weights $w_{i,j}$, we have tested two of them, varying in the tf component:

Raw term frequency

$$w_{i,j} = tf_{i,j} \cdot \log \frac{d}{df_j} \tag{1}$$

where $tf_{i,j}$ denotes the number of occurrences of the term t_j in the document d_i (term frequency), d is the total number of documents in the collection and finally df_j denotes the number of documents that contain t_j.

BM25 term frequency

$$w_{i,j} = \frac{k_1 \cdot tf_{i,j}}{tf_{i,j} + k_1(1 - b + b\frac{l_d}{l_C})} \cdot \log \frac{d}{df_j} \tag{2}$$

where $tf_{i,j}$, d and df_j have the same meaning as in (1), l_d denotes the length of the document, l_C the average length of a document in the collection and finally k_1 and b are the parameters to be set.

The tf components for queries are defined analogously, except for the average length of a query, which obviously cannot be determined as the system is not aware of the full query set and processes one query at a time. The Lemur documentation is however not clear about the exact way of handling the l_C value for queries.

The values of k_1 and b were set according to the suggestions made by [7] and [6], that is $k_1 = 1.2$ and $b = 0.75$ for computing document weights and $k_1 = 1$ and $b = 0$[2] for query weights.

We have also tested the influence of the blind relevance feedback. The simplified version of the Rocchio's relevance feedback implemented in Lemur [7] was used for this purposes. The original Rocchio's algorithm is defined by the formula

$$\vec{q}_{new} = \vec{q}_{old} + \alpha \cdot \vec{d}_R - \beta \cdot \vec{d}_{\bar{R}}$$

where R and $\bar{R}$ denote the set of relevant and non-relevant documents, respectively, and $\vec{d}_R$ and $\vec{d}_{\bar{R}}$ denote the corresponding centroid vectors of those sets. In other words, the basic idea behind this algorithm is to move the query vector closer to the relevant documents and away from the non-relevant ones. In the case of blind feedback, the top M documents from the first-pass run are simply considered to be relevant. The Lemur modification of this algorithm sets the $\beta = 0$ and keeps only the K top-weighted terms in $\vec{d}_R$.

3 Experimental Evaluation

We have created 3 different indices from the collection — using original data and their lemmatized and stemmed version. There were 29 training topics and 42 evaluation topics defined by the organisers. We have first run the set of experiments for the training topics (see Table 1), comparing:

- Results obtained for the queries constructed by concatenating the tokens (either words, lemmas or stems) from the `<title>` and `<desc>` fields of the topics (TD - upper section of the table) with results for queries made from all three topic fields, i.e. `<title>`, `<desc>` and `<narr>` (TDN - lower section).
- Results achieved on the "original" Quickstart collection (i.e. 3-minute window with 1-minute overlap - Segments 3-1) with results computed using the collection created by using 2-minute window with 1-minute overlap (Segments 2.1).

In all cases the performance of raw term frequency (Raw TF) and BM25 term frequency (BM25 TF) is tested, both with (BRF) and without (no_FB) application of the blind relevance feedback. The mean Generalized Average Precision (mGAP) is used as the evaluation metric — the details about this measure can be found in [4].

		Segments 3-1				Segments 2-1			
		Raw TF		BM25 TF		Raw TF		BM25 TF	
		no_FB	BRF	no_FB	BRF	no_FB	BRF	no_FB	BRF
TD	words	0.0184	0.0183	0.0152	0.0183	0.0212	0.0246	0.0147	0.0174
	lemmas	0.0277	0.0303	0.0279	0.0324	0.0293	0.0383	0.0276	0.0346
	stems	0.0281	0.0315	0.0258	0.0322	0.0323	0.0389	0.0281	0.0335
TDN	words	0.0194	0.0209	0.0132	0.0169	0.0211	0.0234	0.0161	0.0202
	lemmas	0.0330	0.0374	0.0231	0.0325	0.0389	0.0453	0.0286	0.0376
	stems	0.0332	0.0356	0.0235	0.0341	0.0390	0.0443	0.0288	0.0374

Table 1: Mean GAP of the individual runs - training topics.

Then we identified the 5 most promising/illustrative runs from the Table 1, repeated them for the evaluation topics and send to the organisers for judgment. After receiving the relevance judgments for evaluation topics, we have replicated all the runs for those topics too (Table 2).

It turns out that the structure of the results for different experimental settings is similar for both the training and evaluation topics - thus we could observe the following trends:

[2]This is actually not a choice, as the value of b is hard-set to 0 for queries in Lemur.

		Segments 3-1				Segments 2-1			
		Raw TF		BM25 TF		Raw TF		BM25 TF	
		no_FB	BRF	no_FB	BRF	no_FB	BRF	no_FB	BRF
TD	words	0.0105	0.0121	0.0088	0.0121	0.0123	**0.0126**	0.0097	0.0108
	lemmas	0.0168	0.0189	0.0126	**0.0126**	0.0183	0.0206	0.0144	0.0133
	stems	0.0188	0.0205	0.0132	0.0161	0.0196	**0.0217**	0.0157	0.0187
TDN	words	0.0113	0.0142	0.0089	0.0108	0.0141	0.0162	0.0099	0.0125
	lemmas	0.0205	**0.0226**	0.0114	0.0150	0.0206	**0.0254**	0.0164	0.0150
	stems	0.0215	0.0215	0.0092	0.0107	0.0218	0.0246	0.0120	0.0125

Table 2: Mean GAP of the individual runs - evaluation topics. Bold runs were submitted for official scoring.

- Two minute "documents" seem to perform better than the three minute ones — probably the three minute segmentation is too coarse.
- The simplest raw term frequency weighting scheme generally outperforms the more sophisticated BM25 — one possible explanation is that in a standard document retrieval setup the BM25 scheme profits mostly from its length normalization component that is completely unnecessary in our case (remember that our documents all have approximately identical length by design).

The fact that both stemming and lemmatization boost the performance by about the same margin was already observed in the last year's experiments.

4 Conclusion

In the CLEF 2007 CL-SR task, we have made just a little step further towards successful searching of Czech spontaneous speech. In order to make a bigger progress, we would need to really take the speech part of the task into account — that is, to use the speech recognizer lattices when searching for the desired information, or even to modify the ASR components so that it will be more likely to produce output useful for IR (for example, enrich the language model with rare named entities that are currently often being misrecognized).

Acknowledgments

This work was supported by the Grant Agency of the Czech Academy of Sciences project No. 1ET101470416 and the Ministry of Education of the Czech Republic project No. LC536.

References

[1] Carnegie Mellon University and the University of Massachusetts. The Lemur Toolkit for Language Modeling and Information Retrieval. (`http://www.lemurproject.org/`), 2006.

[2] Jan Hajič. *Disambiguation of Rich Inflection. (Computational Morphology of Czech).* Karolinum, Prague, 2004.

[3] Pavel Ircing and Luděk Müller. Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Lecture Notes in Computer Science, Alicante, Spain, 2007.

[4] Baolong Liu and Douglas Oard. One-Sided Measures for Evaluating Ranked Retrieval Effectiveness with Spontaneous Conversational Speech. In *Proceedings of SIGIR 2006*, pages 673–674, Seattle, Washington, USA, 2006.

[5] Douglas Oard, Jianqiang Wang, Gareth Jones, Ryen White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Lecture Notes in Computer Science, Alicante, Spain, 2007.

[6] Stephen Robertson and Steve Walker. Okapi/Keenbow at TREC-8. In *The Eight Text REtrieval Conference (TREC-8)*, 1999.

[7] Chengxiang Zhai. Notes on the Lemur TFIDF model. Note with Lemur 1.9 documentation, School of CS, CMU, 2001.