Czech Monolingual Information Retrieval Using Off-The-Shelf Components - the University of West Bohemia at CLEF 2007 Ad-Hoc track

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

Abstract

The paper provides a brief description of the system assembled for the CLEF 2007 Ad-Hoc track by the University of West Bohemia. We have performed only monolingual experiments (Czech documents - Czech queries) using two incarnations of the tf.idf model — one with raw term frequency and the other with the BM25 term frequency weighting — as implemented in the Lemur toolkit. The effect of the blind relevance feedback was also explored. Czech morphological analyser and tagger were used for lemmatization and stop word removal. The results achieved seem to be quite reasonable, with MAP ranging from 0.11. to 0.30.

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

Monolingual Ad-Hoc Information Retrieval

1 Introduction

Although our group is mainly interested in the CL-SR track in the CLEF campaign, we could not resist participating in Ad-Hoc once our native language was introduced to the track. Our runs were generated essentially just by putting together off-the-shelf components available either for Czech NLP or general IR. Such seemingly unambitious approach has, however, proven to be quite successful in the past CLEF campaigns. We have performed monolingual Czech experiments only.

2 System description

2.1 Linguistic preprocessing

Stemming (or lemmatization) is considered to be vital for good IR performance. This assumption was experimentally proven by our group also for the Czech language IR 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 its 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 [5] and [4], that is $k_1 = 1.2$ and $b = 0.75$ for computing document weights and $k_1 = 1$ and $b = 0$[1] 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 [5] 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 $\vec{d}_R$.

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

3 Experimental Evaluation

There were 50 topics defined for Ad-Hoc track, in a variety of languages. As we have already mentioned, we have used only the Czech topics for searching Czech documents. The document set consists of electronic versions of articles from two nationwide newspapers (Mladá Fronta Dnes, Lidové Noviny); following the track organisers' instructions, we have indexed only the `<TITLE>` and `<TEXT>` fields, in both the original (non-lemmatized) and the lemmatized version.

The results are summarized in Table 1. The upper section shows the MAP for queries constructed by concatenating the tokens (either words or lemmas) from the `<title>` and `<desc>` fields of the topics (TD), the lower section then the results for queries made from all three topic fields, i.e. `<title>`, `<desc>` and `<narr>` (TDN). Both results with (BRF) and without (no_FB) application of the blind relevance feedback are shown.

		Raw TF		BM25 TF	
		no_FB	BRF	no_FB	BRF
TD	words	0.1405	**0.1101**	0.2053	0.2500
	lemmas	0.1765	0.1247	**0.2569**	0.3025
TDN	words	0.1491	0.1162	**0.2219**	0.2480
	lemmas	0.1869	0.1415	**0.2277**	0.2405

Table 1: MAP of the individual runs - bold runs were submitted for official scoring.

The table reveals several findings. First of them is that the length normalization contained in the BM25 formula seems to have an immense effect on the performance — this is probably something not very surprising to an experienced IR researcher, it did however surprise us as we were dealing with documents of approximately uniform length in last year's CL-SR track (again see [3] for details). What is truly puzzling is the negligible effect of lemmatization for the runs using BM25 term frequency component and TDN queries; especially when you compare those runs with the other "quadrants" of the table.

4 Conclusion

Our participation in the Ad-Hoc track was motivated mainly by two factors — we wanted to enrich the diversity of the pool of results and we wanted to know how our quite strong experience of dealing with Czech language processing and a rather poor experience of designing IR systems will hold up in competition. While we have hopefully succeeded in the former, we still have no idea how we have done in the latter as the organisers did not publish any cross-site comparison. Thus we look forward to seeing such ranking in the track overview paper.

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

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