Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006*

Pavel Ircing and Luděk Müller

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{ircing, muller}@kky.zcu.cz

Abstract. The paper describes the system built by the team from the University of West Bohemia for participation in the CLEF 2006 CL-SR track. We have decided to concentrate only on the monolingual searching in the Czech test collection and investigate the effect of proper language processing on the retrieval performance. We have employed the Czech morphological analyser and tagger for that purposes. For the actual search system, we have used the classical *tf.idf* approach with blind relevance feedback as implemented in the Lemur toolkit. The results indicate that a suitable linguistic preprocessing is indeed crucial for the Czech IR performance.

1 Introduction

This paper presents the first participation of the University of West Bohemia group in CLEF (and, for that matter, first participation of the group in an IR evaluation campaign whatsoever). Thus, being novices in the IR field, we have decided to concentrate only on the monolingual searching in the Czech test collection where we have tried to make use of the two advantages that our team might have over the others — the knowledge of the language in question (Czech — our mother tongue) and the experience with automatic NLP of that language, together with the employment of the necessary tools (morphological analyser, tagger).

As for the actual search side of the task, it has been shown by various teams experimenting with last year's English test collection that good results can be achieved simply by using some freely available IR system (see for example [1]). We have decided to use the same strategy.

Although both the English and the Czech CL-SR collections consists of the (automatic) transcriptions of the interviews with the Holocaust survivors (plus some additional metadata — see the description of the collections in the track

* 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.

overview [2]), the Czech collection lacks the manually created topical segmentation that is available for the English data. This obviously makes the retrieval more complicated. Thus, in order to facilitate the initial experiments with the Czech collection, the track organizers provided also a so-called Quickstart collection with artificially defined "documents" that were created by sliding 3-minute window over the continuous stream of transcriptions with the 1-minute step. Given the lack of time for experimentation, the presence of many other system parameters and the absence of training topics, we did not explore any other segmentation possibilities beyond this Quickstart collection in our experiments. It turned out at the workshop that all other teams applied the same approach which makes all the results well comparable.

2 System Description

2.1 Linguistic Preprocessing

At least rudimentary linguistic processing of the document collection and topics (stemming, stop-word removal) is considered to be indispensable in state-of-the-art IR systems. We have decided to use quite sophisticated NLP tools for that purpose — the morphological analyser and tagger developed by the team around Jan Hajič [3],[4]. The serial combination of these two tools assigns disambiguated lemma (basic word form) and morphological tag to the input word form and also provides the information about the stem-ending partitioning.

This is an example of the typical system output:

```
<f>holokaustem<MDl>holokaust<MDt>NNIS7-----A----<R>holokaust<E>em
```

where `<f>` introduces the actual word form, the `<MDl>` the corresponding lemma and the `<MDt>` the corresponding morphological tag (in this case the tag correctly describes the word `holokaustem` as the noun (`N` in the first position) having the masculine inanimate gender (`I`) and being in singular (`S`) instrumental (`7`) form). Finally, the `<R>` introduces the stem and `<E>` the ending of the word form in question. Note that although in this example the stem is identical to the lemma it is not the general rule. Thus one of the issues to be resolved is whether to use lemmatization or stemming for the data processing. For English, the results typically favor neither lemmatization nor stemming over each other [5] and since the stemming is simpler, it is most commonly used. As our tools perform Czech lemmatization and stemming in a single step, we investigated both strategies in our experiments.

The information provided by the NLP tools was also exploited for stop-word removal. As we were not able to find any decent stoplist of Czech words we have decided to remove words on the basis of their part-of-speech (POS). As can be seen from the example above, the POS information is present at the first position of the morphological tag. We removed from indexing all the words that were tagged as prepositions, conjunctions, particles and interjections (note that they are no articles in Czech).

Here is an example of one of the topics before and after the linguistic preprocessing
The original topic:

```
<top>
<num>1286</num>
<title>Hudba v holokaustu</title>
<desc>Svědectví o tom, zda hudba pomáhala (duševně nebo i jinak)
nebo překážela vězňům internovaným v koncentračních táborech.
</desc>
<narr>Popis toho, jakou roli hrála hudba v životě vězňů.</narr>
</top>
```

gets processed into:

```
<top>
<num>1286</num>
<title>hudba holokaust</title>
<desc>svědectví ten hudba pomáhat duševně jinak překážet vězeň
internovaný koncentrační tábor</desc>
<narr>popis ten jaký role hrát hudba život vězeň</narr> </top>
```

when using lemmatization or into:

```
<top>
<num>1286</num>
<title>hud holokaust</title>
<desc>svědectví tom hud pomáh duševně jinak překáž vězňům
internovan koncentračn tábor</desc>
<narr>popis toho jakou rol hrál hud život vězňů</narr> </top>
```

when the stemming is employed.

2.2 Retrieval

For the actual IR we have used the freely available Lemur toolkit [6] that allows us to employ various retrieval strategies, including among others the classical vector space model and the language modeling approach.

We have decided to stick to the *tf.idf* model where both documents and queries are represented as weighted term vectors $\boldsymbol{d}_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,n})$ and $\boldsymbol{q}_k = (w_{k,1}, w_{k,2}, \cdots, w_{k,n})$, 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. As there are many ways to compute the weights $w_{i,j}$ without any of them performing consistently better than the others, we employed the very basic formula

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

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. We have not used any document length normalization as the length of the "documents" in the Quickstart collection is approximately uniform.

In order to boost the performance, we also used the simplified version of the blind relevance feedback implemented in Lemur [7]. The original Rocchio's algorithm is defined by the formula

$$\boldsymbol{q}_{new} = \boldsymbol{q}_{old} + \alpha \cdot \boldsymbol{d}_R - \beta \cdot \boldsymbol{d}_{\bar{R}} \tag{2}$$

where R and $\bar{R}$ denote the set of relevant and non-relevant documents, respectively, and $\boldsymbol{d}_R$ and $\boldsymbol{d}_{\bar{R}}$ 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 treated as if they were relevant. The Lemur modification of this algorithm sets the $\beta = 0$ and keeps only the K top-weighted terms in $\boldsymbol{d}_R$.

3 Experimental Evaluation

As we already mentioned in the Introduction, all the experiments were carried out on the Czech Quickstart collection, using only the Czech version of the queries. There are 115 queries defined for searching in the Czech test collection. However, only 29 of them were manually evaluated by the assessors and used to generate the `qrel` files.

3.1 Evaluated Runs

Table 1 summarizes the results for the runs that we consider to be important. The mean Generalized Average Precision (mGAP) is used as the evaluation metric — the details about this measure can be found in [8]. The table contains two main horizontal sections — the upper one contains the results achieved when using only `<title>` and `<desc>` fields of the topics as queries (TD), the lower one then the results with all `<title>`, `<desc>` and `<narr>` fields used (TDN). Inside each of these sections, the runs are further divided according to the type of the performed linguistic processing — both the queries and the collection were always stoplisted and either left in the original word forms (w) or stemmed (s) or lemmatized (l). Finally, the vertical division denotes the fields from the collection that were indexed for the corresponding runs — `<CZECHMANUKEYWORD>` (mk), `<ASRTEXT>` (asr), `<CZECHAUTOKEYWORD>` (ak) and their combinations. The runs typeset in boldface are the ones that were submitted for the official CLEF scoring.

Table 1. Mean GAP of the individual runs - bold runs were submitted for official scoring.

		mk	asr	ak	mk.asr	asr.ak	mk.asr.ak
TD	w	0.0026	0.0271	0.0022	0.0270	0.0240	0.0247
	s	0.0030	0.0438	0.0024	0.0441	0.0405	0.0401
	l	0.0026	**0.0435**	0.0024	**0.0416**	**0.0402**	**0.0377**
TDN	w	0.0025	0.0256	0.0018	0.0266	0.0241	0.0242
	s	0.0029	0.0494	0.0022	0.0488	0.0447	0.0456
	l	0.0024	0.0506	0.0023	0.0518	0.0467	**0.0456**

3.2 Analysis of the Results

First, we cannot resist to mention that all our submitted runs significantly outperformed the runs of the other two Czech sub-track participants. The reason of this is quite simple and readily apparent from the table of results — as far as we know, we were the only team that used Czech lemmatization and/or stemming and the table shows that either one of these operations boosts the IR performance almost by a factor of two (that is, at least in the columns where the asr field is indexed).

What the experiments did not help to resolve is whether to use stemming or lemmatization in the preprocessing stage — both methods yielded comparable results. On the other hand, the results show that both the manual and the automatic keyword fields are virtually useless for the retrieval — they perform poorly when indexed alone and do not bring any noticeable improvement when added to the asr field. Further investigation revealed that the scripts that were used for the creation of the Quickstart collection contain systematic error that caused manual keywords to be assigned to the wrong segments. Manual examination of a sample of the automatically assigned English thesaurus terms indicates that the automatic keyword assignment is not overly accurate either. Therefore both the manual and the automatic keywords most probably bring more noise than useful information to the IR system.

We have also performed a couple of experiments in order to investigate the effect of the stop-word removal (not shown). We have found out that removing the stop words on the basis of morphological tags did not help the IR performance in comparison with working with the "non-stopped" collection and queries. On the other hand, it substantially reduced the size of the index files.

3.3 Tips for Future Improvement

First of all, let us point out again that the segments in the Quickstart collections are not well-formed "documents", especially in the sense that they are not, in most cases, topically coherent. Thus a more sophisticated way of collection segmentation might be useful. Moreover, the quality of the ASR transcriptions is rather poor — around 35% WER in general but even more for the named

entities (NEs) which are extremely important for searching. The application of the specially designed language model focused on a more accurate transcription of NEs could partially alleviate this problem [9]. Finally, there appears to be a non-negligible vocabulary mismatch between the topics and the collection or even between the different fields in the collection. For example, just looking at the first two topics that were evaluated by the assessors we have discovered that in topic 1181 the name of the infamous concentration camp "Auschwitz" was kept untranslated in the topics but it was translated into its Czech form ("Osvětim") in the `<CZECHMANUKEYWORD>` and `<CZECHAUTOKEYWORD>` fields,[1] the word "Sonderkommando" was written with double "m" in the topics and in the `<ASRTEXT>` field and with single "m" in the keyword fields. A coherent treatment of such different variants is therefore desirable.

4 Conclusion

The Czech CL-SR track presents a first attempt to create and test the collection of the Czech spontaneous speech. As such, it suffered some initial difficulties that were for the most part identified by the end of the actual CLEF workshop and just recently fixed. As a result, we are currently at the performance level comparable with the English part of the track but the lack of time prevented us from doing even more detailed result analysis (especially significance testing). Nevertheless, we are already in a good position for the next year's campaign where the current evaluation set of topics will most probably serve as a training data.

References

1. Inkpen, D., Alzghool, M., Islam, A.: Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. In Peters, C., Gey, F., Gonzalo, J., Mueller, H., Jones, G., Kluck, M., Magnini, B., de Rijke, M., eds.: Accessing Multilingual Information Repositories - 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Volume 4022 of Lecture Notes in Computer Science., Vienna, Austria (2006) 760–768
2. Oard, D., Wang, J., Jones, G., White, R., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In Peters, C., Clough, P., Gey, F., Karlgren, J., Magnini, B., Oard, D., de Rijke, M., Stempfhuber, M., eds.: 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)
3. Hajič, J.: Disambiguation of Rich Inflection. (Computational Morphology of Czech). Karolinum, Prague (2004)

[1] Note that neither one of those variants is the original name of the Polish town in question — "Oświęcim." Consequently, all three forms are routinely used by the Czech speakers and therefore appear in the ASR transcripts.

4. Hajič, J., Hladká, B.: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings of COLING-ACL Conference, Montreal, Canada (1998) 483–490
5. Hull, D.A.: Stemming Algorithms: A Case Study for Detailed Evaluation. Journal of the American Society of Information Science **47**(1) (1996) 70–84
6. Carnegie Mellon University and the University of Massachusetts: The Lemur Toolkit for Language Modeling and Information Retrieval. (`http://www.lemurproject.org/`) (2006)
7. Zhai, C.: Notes on the Lemur TFIDF model. Note with Lemur 1.9 documentation, School of CS, CMU (2001)
8. Liu, B., Oard, D.: One-Sided Measures for Evaluating Ranked Retrieval Effectiveness with Spontaneous Conversational Speech. In: Proceedings of SIGIR 2006, Seattle, Washington, USA (2006) 673–674
9. Ircing, P., Psutka, J., Radová, V.: Automatic Transcription of Audio Archives for Spoken Document Retrieval. In: Proceedings of Computational Intelligence 2006, San Francisco, USA (2006) 448–452