The University of West Bohemia at CLEF 2006, the CL-SR track

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

Abstract

The paper describes the system build 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. We have employed the Czech morphological analyser and tagger in order to perform necessary linguistic preprocessing (lemmatization and stop-word removal). As for the actual search system, we have employed the classical tf.idf approach with blind relevance feedback as implemented in the LEMUR toolkit. Since the results are currently very close to zero and appear to behave rather randomly, it is not possible to draw any conclusion at this moment. There are several hypothesis concerning the possible causes of the system failure that are currently a subject of investigation.

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

Measurement, Performance, Experimentation

Keywords

Spoken Document Retrieval

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 disposal of the necessary tools (morphological analyzer, tagger).

As for the actual search side of the task, it has been shown by various teams experimenting with the last year English test collection that good results can be achieved simply by using some freely available IR system (see for example [4]). We have decided to use the same strategy.

Although both the English and the Czech CL-SR collections consists of the (automatic) transcriptions[1] of the interviews with the Holocaust survivors, the Czech collection lacks the manually

[1] Plus some additional metadata - see the description of the collections in the track overview.

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. The total number of those "documents" in the collection is 11,377. Given the lack of time for experimentation, the presence of many other system parameters and the absence of the training topics, we did not explore any other segmentation possibilities beyond this Quickstart collection in our experiments.

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 purposes - the morphological analyzer and tagger developed by the team around Jan Hajič [2],[3]. 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 and we believe that the lemmatization should be used instead of stemming in the IR experiments with highly inflectional languages such as Czech. Therefore all our submitted experiments use the lemmatized version of the collection and topics.

The information provided by the NLP tools was also exploited for the 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 topic 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>
```

2.2 Retrieval

For the actual IR we have used the freely available LEMUR toolkit [1] that allows 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 $\vec{d_i} = (w_{i,1}, w_{i,2}, \cdots, w_{i,n})$ 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. 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

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

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.

In order to boost the performance, we also used the simplified version of the Rocchio's blind relevance feedback implemented in LEMUR [7]. 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 Description of the runs

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. The linguistic preprocessing and the retrieval method as described in sections 2.1 and 2.2, respectively, were the same for all the runs submitted to the official evaluation. Those runs were the following:

UWB_aTD

Query fields used: `<title>` (T) and `<desc>` (D)
Collection fields used: `<ASRTEXT>` only

UWB_a_akTD

Query fields used: TD
Collection fields used: `<ASRTEXT>` and `<CZECHAUTOKEYWORD>`

UWB_mk_aTD

Query fields used: TD
Collection fields used: `<CZECHMANUKEYWORD>` and `<ASRTEXT>`

UWB_mk_a_akTD

Query fields used: TD
Collection fields used: `<CZECHMANUKEYWORD>` and `<ASRTEXT>` and `<CZECHAUTOKEYWORD>`

UWB_mk_a_akTDN

Query fields used: TD and `<narr>` (N)
Collection fields used: `<CZECHMANUKEYWORD>` and `<ASRTEXT>` and `<CZECHAUTOKEYWORD>`

Since the official results of the other teams participating in the track revealed that using only the manual keywords gives the best results, we have generated this one additional run:

UWB_mkTD

Query fields used: `<title>` (T) and `<description>` (D)
Collection fields used: `<CZECHMANUKEYWORD>` only

The total of six runs seems to be small to assess the behavior of the task. However, the reasons why we stopped additional runs are explained in detail in the following section.

4 Results and their analysis

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. Table 1 summarizes the results for the runs described above (the prefix UWB is omitted due to formatting issues). The mean Generalized Average Precision (GAP) is used as the evaluation metric - the details about this measure can be found in [5].

Run	aTD	a_akTD	mk_aTD	mk_a_akTD	mk_a_akTDN	mkTD
Mean GAP	0.0003	0.0003	0.0004	0.0004	0.0004	0.0015

Table 1: Mean GAP of the individual runs

As you can see from the table, the achieved results are very close to zero, especially for the runs employing any of the automatic fields. Moreover, when we tried the same runs with the original word forms or the stems instead of the lemmas we have discovered that the mean GAP generally remains unchanged but at the same time the GAP for individual topics varies quite wildly. This lead us to the hypothesis that the results are completely random. In order to prove or reject such hypothesis we have generated 100 different runs putting random 1,000 documents in the ranked list for each of the topics. The average mean GAP of these runs exceeds 0.0005 which is actually more than the results achieved by runs involving any of the automatic fields.

As for the run using only the manual keywords, the GAP is slightly better there but behaves no less wildly. Generally, the mean GAP result depends on the result achieved on the single topic (number 14312) as the rest of the GAPs is more or less zero. Moreover, when we accidentally run the non-lemmatized topics on the lemmatized collection, the GAP of that topic jumped from 0.0325 to 0.1000 causing the mean GAP to jump from 0.0015 to 0.0046.

Therefore we are prone to conclude that the results are indeed currently at the random level. The question is why is it so. The first possibility that comes to mind is some fundamental flaw in the design of the search system itself. This is quite improbable as all the three teams that participated in the track achieved comparable results and at least two of these teams (that means, all besides us) have a significant experience with designing the IR systems for similar evaluation campaigns.

According to our opinion, one of the reasons of the failure is the immense difficulty of the task in question. This difficulty stems mainly from the following factors:

1. The collection lacks the topical segmentation - segments in the Quickstart collection are in most cases not topically coherent.

2. The quality of the ASR transcriptions is rather poor (around 40% WER) - but that is a problem which is shared by both the Czech and the English collections.

3. The quality of the automatic keyword assignment is generally very low - this is probably caused by the fact that this assignment had to be done in a complicated cross-language manner due to the lack of annotated Czech training data (see [6]).

4. 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 the 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[2], the word "Sonderkommando" was written with double "m" in the topics and in the `<ASRTEXT>` field and with single "m" in the keyword fields.

5. Some of the salient words from the topics hardly appear in the collection at all. For example, the most important word from the topic 1166 ("Hasidism", or its modifications) appears only 3 times in the entire collection, in all cases in the `<CZECHAUTOKEYWORD>` field and in all cases it is placed there incorrectly.

There is also a remote possibility that the problem is on the evaluation side. Anyway, all those issues will be a subject of a more detailed investigation in the near future.

5 Conclusion

The Czech CL-SR track presents a first attempt to create and test the collection of the Czech spontaneous speech. As such (and given the inherent challenge of the task in question) seems to suffer some initial difficulties that have to be first precisely identified and then hopefully solved.

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] http://www.lemurproject.org/.

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

[3] Jan Hajič and Barbora Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pages 483–490, Montreal, Canada, 1998.

[4] Diana Inkpen, Muath Alzghool, and Aminul Islam. University of Ottawa's Contribution to CLEF 2005, the CL-SR Track. In *Working notes for CLEF 2005 Workshop*, Vienna, Austria, 2005.

[5] 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.

[2]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.

[6] Scott Olsson, Douglas Oard, and Jan Hajič. Cross-Language Text Classification. In *Proceedings of SIGIR 2005*, pages 645–646, Salvador, Brazil, 2005.

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