

Charles University at CLEF 2007 Ad-Hoc Track

Pavel Češka and Pavel Pecina
Institute of Formal and Applied Linguistics
Charles University, 118 00 Praha 1, Czech Republic
{ceska,pecina}@ufal.mff.cuni.cz

Abstract

In this paper we describe retrieval experiments performed at Charles University in Prague for participation in the CLEF 2007 Ad-Hoc track. We focused on the Czech monolingual task and used the LEMUR toolkit as the retrieval system. Our results demonstrate that for Czech as a highly inflectional language, lemmatization significantly improves retrieval results and manually created queries are only slightly better than queries automatically generated from topic specifications.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Ad-Hoc Retrieval

1 Introduction

This work represents the first participation of Charles University in the CLEF evaluation campaign. Our research is focused on Czech monolingual tasks and application of advanced language processing tools developed at our university - namely a morphological analyser and tagger. We also attempt to compare systems with manually and automatically created queries. For the Ad-Hoc track we submitted four experiments (runs): *Prague01*, *Prague02*, *Prague03*, and *Prague04*. Our main goal were to study influence of lemmatization and whether manual query construction can bring additional performance improvement. Similar experiments were performed also for the CLEF 2007 Cross-Language Speech Retrieval track.

2 System Description

2.1 Retrieval model

Being novices in the field of information retrieval we decided to use a freely available retrieval toolkit instead of developing our own. The final choice was the LEMUR toolkit [5] and its Indri retrieval model [3]. It is based on a combination of language modeling and inference network retrieval. It has been popular among CLEF participant in recent years and was found effective for a wide range of retrieval tasks.

An inference network (also known as a Bayesian network) consists of a document node, smoothing parameters nodes, model nodes, representation nodes, belief nodes, and information need nodes connected by edges representing independence assumptions over random variables. The document node represents documents as binary vectors where each position represents presence or absence of a certain feature of the text. The model nodes correspond to different representations of the same document (e. g. pseudo-documents made up from all titles, bodies, etc.). The representation concept nodes are related to the features extracted from the document representation. The belief nodes are used to combine probabilities of different representations, other beliefs, etc. A detailed description can be found in [6].

To improve retrieval results, we used Indri's pseudo-relevance feedback which is an adaption of Lawrenko's relevance models [4]. The basic idea behind these models is to combine the original query with a query constructed from top ranked documents of the original query.

2.2 Morphological tagging and lemmatization

State-of-the-art retrieval systems usually include at least some basic linguistically-motivated pre-processing of the documents and queries such as stemming and stopword removal. Czech is a morphologically complex language and there is no easy way how to determine stems and their endings as it can be done in English and other languages. Stemming in Czech is not sufficient and should be replaced by a proper lemmatization (substituting each word by its base form – the lemma) which involves determining the part of speech of all words. In our experiments, we employed the Czech morphological analyzer and tagger developed at Charles University [1], [2] which assigns a disambiguated lemma and a morphological tag to each word. Its accuracy is around 95%. An example of its output for one word (“serious” in English) is following:

```
<f>závažných<MDl src="a">závažný<MDt src="a">AAIP2----1A----
```

The tag <f> is followed by the original word form, tag <MDl> is followed by the lemma, and the tag <MDt> separates a 15-position morphological category (the first position represents the part-of-speech; A stands for an adjective). Lemmatization was employed in all our experiments except *Prague03*. In *Prague01*, both original word forms and lemmas were used for indexing (in two separate model representations).

2.3 Stopword list construction

We used two approaches to construct the stopword lists for our experiments. The first was based on frequency of word occurrences in the collection, the latter on part-of-speech of words. In the first three experiments (*Prague01-03*), we removed 40 most frequented words (separately from the original and lemmatized text) from the documents and the queries. In the fourth experiment (*Prague04*), we removed all words tagged as pronouns, prepositions, conjunctions, particles, interjections, and unknown words (mostly typos) and kept only open-class words.

2.4 Automatic query construction

Automatically created queries were constructed from the <title> and <description> fields of the topic specifications only. The text was simply concatenated and processed by the analyzer and tagger. A combination of the original and lemmatized query was used in the first experiment (*Prague01*). Lemmatized queries containing only nouns, adjectives, numerals, adverbs and verbs were created for the fourth experiment (*Prague04*).

Example

Step 1. The original title and description (topic 10.2452/413-AH: Reducing Diabetes Risk):

```
<title>Snižování rizika onemocnění cukrovkou</title>
<desc>Najděte dokumenty zmiňující faktory, které snižují riziko onemocnění cukrovkou.
</desc>
```

Step 2. Concatenation:

Snižování rizika onemocnění cukrovkou. Najděte dokumenty zmiňující faktory, které snižují riziko onemocnění cukrovkou.

Step 3. Lemmatization:

snižování riziko onemocnění cukrovka najít dokument zmiňovat faktor který snížit riziko onemocnění cukrovka

Step 4. *Prague01* query (original word forms plus lemmas; the suffixes `.(orig)` and `.(lemma)` refer to the corresponding model representations):

```
#combine(snižování.(orig) rizika.(orig) onemocnění.(orig) cukrovkou.(orig) najdete.(orig)
dokumenty.(orig) zmiňující.(orig) faktory.(orig) které.(orig) snižují.(orig)
riziko.(orig) onemocnění.(orig) cukrovkou.(orig) snižování.(lemma) riziko.(lemma)
onemocnění.(lemma) cukrovka.(lemma) najít.(lemma) dokument.(lemma) zmiňující.(lemma)
faktor.(lemma) kter.(lemma) snižovat.(lemma) riziko.(lemma) onemocnění.(lemma)
cukrovka.(lemma))
```

Step 5. *Prague04* query:

```
#combine(snižování riziko onemocnění cukrovka zmiňující faktor snižovat riziko onemocnění
cukrovka)
```

2.5 Manual query construction

The queries in two of our experiments were created manually. In *Prague02* they were constructed from lemmas (to match the lemmatized documents) and their synonyms and in *Prague03* with the use of “stems“ and wildcard operators to cover all possible word forms (documents indexed in the original forms).

Example

Step 1. The original title and description (topic 10.2452/413-AH: Reducing Diabetes Risk):

```
<title>Snižování rizika onemocnění cukrovkou</title>
<desc>Najděte dokumenty zmiňující faktory, které snižují riziko onemocnění cukrovkou.
</desc>
```

Step 2. The *Prague02* query based on lemmas (the operator `#combine()` combines beliefs of the nested operators, operator `#syn()` represents synonymic line of equal expressions and operator `#2()` represents ordered window with width 2 words):

```
#combine(#syn(diabetes cukrovka úplavice) #2(snížení riziko) prevence)
```

Step 3. The *Prague03* query with wildcard operators (which can be used as a suffix only).

```
#combine(diabet* cukrovk* úplavic* sníž* rizik* preven*)
```

3 Experiment Specification

Prague01

Topic fields: <title>, <desc>
Query construction: *automatic*
Document fields: <title>, <heading>, <text>
Word forms: *original + lemmas*
Stop words: *40 most frequent original forms + 40 most frequent lemmas*

Prague02

Topic fields: <title>, <desc>
Query construction: *manual*
Document fields: <title>, <heading>, <text>
Word forms: *lemmas*
Stop words: *40 most frequent lemmas*

Prague03

Topic fields: <title>, <desc>
Query construction: *manual (with wildcard operators)*
Document fields: <title>, <heading>, <text>
Word forms: *original*
Stop words: *40 most frequent word forms*

Prague04

Topic fields: <title>, <desc>
Query construction: *automatic*
Document fields: <title>, <heading>, <text>
Word forms: *lemmas*
Stop words: *pronouns, prepositions, conjunctions, particles, interjections, and unknown words*

4 Results and Conclusion

The Czech Ad-Hoc collection consists of 81,735 documents and 50 topics. The following table summarizes the results for the experiments described above.

	<i>Prague01</i>	<i>Prague02</i>	<i>Prague03</i>	<i>Prague04</i>
Mean Average Precision	0.3419	0.3336	0.3202	0.2969
Mean R Precision	0.3201	0.3349	0.3147	0.2886
Mean Binary Preference	0.2977	0.3022	0.2801	0.2601
Precision at 10 interpolated recall level	0.5733	0.6314	0.5299	0.5367

In terms of Mean Average Precision, the best score was achieved in experiment *Prague01*. Indexing both original word forms and lemmas in combination with automatically generated queries seems to be a reasonable way how to build a retrieval system. In terms of other performance measures, the scores of *Prague02* are slightly better but this is probably due to the use of synonyms in the manually created queries – not in the manual approach itself.

By comparing scores of *Prague03* with results of *Prague01* and *Prague02* we can confirm that lemmatization is quite useful for searching in highly flectional languages such a Czech and can not be fully substituted by stemming.

The last lesson we learned is that using extensive stopword lists based on part-of-speech can seriously harm the performance of a retrieval system as can be seen on the results of experiment *Prague04*.

We found these results quite encouraging and motivating for our future work.

Acknowledgments

This work has been supported by the Ministry of Education of the Czech Republic, projects MSM 0021620838 and #1P05ME786.

References

- [1] Jan Hajič and Barbora Vidová-Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of the Conference COLING - ACL '98*. Montreal, Canada, 1998.
- [2] Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague, 2004.
- [3] <http://www.lemurproject.org/indri/>.
- [4] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2001. ACM Press.
- [5] <http://www.lemurproject.org/>.
- [6] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language-model based search engine for complex queries (extended version). Technical Report IR-407, CIIR, UMass, 2005.