Charles University at CLEF 2007 CL-SR 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

This paper describes a system built at Charles University in Prague for participation in the CLEF 2007 Cross-Language Speech Retrieval track. We focused only on monolingual searching the Czech collection and used the LEMUR toolkit as the retrieval system. We employed own morphological tagger and lemmatized the collection before indexing to deal with the rich morphology in Czech which significantly improved our results.

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

Cross-Language Speech Retrieval

1 Introduction

Charles University in Prague has been participating in coordination of the CLEF Cross-Language Speech Retrieval track since 2006. However, this work represents its first participation in the evaluation itself. Due to the lack of experience with information retrieval, we decided to take advantage of freely available IR system and focused our effort only on the Czech monolingual task. Having a lot of experience with processing the Czech language we used own morphological analyzer and tagger to lemmatize the collection and support indexing and searching.

For the CL-SR track we submitted four quite different 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 Ad-Hoc track.

2 System Description

2.1 Retrieval model

The LEMUR toolkit [5] and its Indri retrieval model [3] is based on a combination of language modeling and inference network retrieval. It has been popular among CLEF participant in recent years and 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]. 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 preprocessing 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 ("concentration" in English) is following:

```
<f>koncentračních<MDl src="a">koncentrační<MDt src="a">AAIP6----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 1173: Children's art in Terezin):

```
<title>Dětské umění v Terezíně</title>
<desc>Hledáme popis uměleckých aktivit dětí v Terezíně, jako např.  hudby, divadla,
malování, poezie a jiných psaných děl.</desc>
```

Step 2. Concatenation:

```
Dětské umění v Terezíně.  Hledáme popis uměleckých aktivit dětí v Terezíně, jako např.
hudby, divadla, malování, poezie a jiných psaných děl.
```

Step 3. Lemmatization:

```
dětský umění v Terezín hledat popis umělecký aktivita děti v Terezín jako například hudba
divadlo malování poezie a jiný psaný dílo
```

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

```
#combine(dětské.(orig) umění.(orig) v.(orig) Terezíně.(orig) hledáme.(orig) popis.(orig)
uměleckých.(orig) aktivit.(orig) dětí.(orig) v.(orig) Terezíně.(orig) jako.(orig)
např.(orig) hudby.(orig) divadla.(orig) malování.(orig) poezie.(orig) a.(orig)
jiných.(orig) psaných.(orig) děl.(orig) dětský.(lemma) umění.(lemma) v.(lemma)
Terezín.(lemma) hledat.(lemma) popis.(lemma) umělecký.(lemma) aktivita.(lemma)
dítě.(lemma) v.(lemma) Terezín.(lemma) jako.(lemma) například.(lemma) hudba.(lemma)
divadlo.(lemma) malování.(lemma) poezie.(lemma) a.(lemma) jiný.(lemma) psaný.(lemma)
dělo.(lemma))
```

Step 5. *Prague04* query:

```
#combine(dětský umění Terezín hledat popis umělecký aktivita dítě Terezín například hudba
divadlo malování poezie jiný psaný dělo)
```

2.5 Manual query construction

The queries in two 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 1173: Children's art in Terezin):

```
<title>Dětské umění v Terezíně</title>
<desc>Hledáme popis uměleckých aktivit dětí v Terezíně, jako např.  hudby, divadla,
malování, poezie a jiných psaných děl.</desc>
```

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

```
#combine(#syn(dítě dětský) umění divadlo hudba #syn(malování kreslení)
#syn(malovat kreslit) poezie básnička)
```

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

```
#combine(dět* umění divad* hud* malov* kresl* poez* básn*)
```

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

All our experiments were performed on the Quickstart collection provided by the track coordinators. 356 holocaust survivors testimonies in Czech were automatically transcribed by an ASR system and the output segmented into 11,373 overlapping passages used as ”documents“. Word error rate of the ASR system is approximately 35%. 29 topics and their relevance assessment were available for training and other 42 topics used for the evaluation. The following table summarizes the results (mGAP scores) for the experiments described above separately for training and evaluation topics.

	topics	*Prague01*	*Prague02*	*Prague03*	*Prague04*
Mean GAP	42 evaluation	0.0187	0.0181	0.0102	**0.0190**
Mean GAP	29 training	0.0266	0.0322	**0.0328**	0.0277

Interpretation of these results is quite difficult mainly because of the difference between performance on the training and evaluation topics. One possible explanation of this discrepancy is that the relevance judgments for the training topics were obtained only by search-guided assessment and not by highly ranked assessment. Thus we consider the results on the evaluation data more credible.

The best score on evaluation topics was achieved in experiment *Prague04* but it is almost indistinguishable from scores of other experiments that employed lemmatization (*Prague01* and *Prague02*). In all these experiments we achieved significantly better results than in experiment *Prague03* where we indexed the original word forms (no lemmatization). This observation is also consistent with our results in Ad-Hoc tracks.

The best score on the training topics was achieved in experiment *Prague03* but it is almost identical to the result in experiment *Prague02*. In both these experiments we used manually created queries and they significantly outperformed experiments with queries automatically generated from the topic specifications. However, we can not conclude that manually constructed queries are better because these results were not confirmed on the evaluation set of topics.

The results we achieved are quite promising and we will continue exprimenting with this valuable collection.

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.