Information Retrieval Test Collection for Searching Spontaneous Czech Speech *

Pavel Ircing[1], Pavel Pecina[2], Douglas W. Oard[3], Jianqiang Wang[4], Ryen W. White[5], and Jan Hoidekr[1]

[1] University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{ircing, hoidekr}@kky.zcu.cz
[2] Charles University, Institute of Formal and Applied Linguistic
Malostranské náměstí 25, 118 00 Praha, Czech Republic
pecina@ufal.mff.cuni.cz
[3] University of Maryland, College of Information Studies/UMIACS
College Park, MD 20742, USA
oard@umd.edu
[4] State University of New York at Buffalo, Department of Library and Information Studies
Buffalo, NY 14260, USA
jw254@buffalo.edu
[5] Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
ryenw@microsoft.com

Abstract. This paper describes the design of the first large-scale IR test collection built for the Czech language. The creation of this collection also happens to be very challenging, as it is based on a continuous text stream from automatic transcription of spontaneous speech and thus lacks clearly defined document boundaries. All aspects of the collection building are presented, together with some general findings of initial experiments.

Key words: Spoken Document Retrieval; Evaluation

1 Introduction

The very essence of an information retrieval (IR) system is to satisfy user's information needs, expressed by a query submitted to the system. The degree of user satisfaction is of course inherently subjective and therefore there is a need for some form of (automatic) quantitative evaluation of the system effectiveness. Such evaluation is usually performed on a defined test collection which includes a representative set of documents, a representative set of topics (formalized information needs) and, most importantly, judgments of the relevance of

* This work was supported by projects MSMT LC536, GACR 1ET101470416, MSM0021620838 and NSF IIS-0122466.

each document to each topic. The process of relevance assessment is extremely labour-intensive, and perhaps explains why no large-scale Czech IR collection has so far been available (at least according to the authors' knowledge).

During the course of the MALACH project, which aims to improve access to the large multilingual spoken archives using advanced ASR and IR techniques [1], the need for an IR test collection arose quite naturally. The archives in question consist of the digitized videotaped testimonies given by the survivors and witnesses of the Holocaust and the ultimate goal of the project is to allow potential users to watch the passages relevant to their queries. In order to facilitate the IR itself, the soundtrack from the testimonies must be transformed into text using an Automatic Speech Recognition (ASR) decoder.

The test collection for English was created first [2]. However, a significant subset of the English interviews (approx. 4,000 of them) was manually subdivided into topically coherent segments, equipped with a three-sentence summary and indexed with keywords selected from a pre-defined thesaurus. The test collection was then built using these manually annotated data. No such manual segmentation was available in the case of the Czech interviews and this fact not only made the Czech IR more difficult but shifted the very nature of the task - the goal for Czech IR experiments is to identify appropriate replay start points rather than to select among pre-defined segments. Nevertheless, some form of manual indexing was performed even on the Czech data, as described in the following section. Both English and Czech collection were used as the reference corpora in the Cross-Language Speech Retrieval (CL-SR) track at the CLEF-2006 evaluation campaign (`http://www.clef-campaign.org/`).

2 Collection

2.1 "Documents"

The collection consists not of documents but rather of (354) interviews; a continuous text stream coming from an ASR decoder. To be precise, there are actually four text streams, generated by two different ASR systems (for details about the first one see [3], the second one is described in [4]), each of them providing the transcription of both the left and right stereo channels. Each of those channels was recorded from a separate microphone, one placed on the interviewer and one on the interviewee, thus yielding acoustic signals with non-negligible differences.

These ASR transcripts are further accompanied with:

- English thesaurus terms that were manually assigned with one-minute granularity by subject-matter experts. Two broad types of thesaurus terms are present, with some describing concepts and others describing locations. The location terms are most often pre-combined with a time clause, which reflects the fact that political boundaries and place names sometimes changed over the time frame described by interviewees in this collection. Unlike the English collection, where the keywords are associated with entire indexer-defined segments, the thesaurus terms in the Czech collection are used as

onset marks—they appear only once at the point where the indexer recognized that a discussion of a topic or a location-time pair had started; continuation and completion of discussion are not marked.
- Automatically produced Czech translations of the English thesaurus terms. These were created using the following resources: (1) professional translation of about 3,000 thesaurus terms that were selected to provide broad coverage of the constituent words, (2) volunteer translation of about 700 additional thesaurus terms, and (3) a custom-built machine translation system that reused words and phrases from manually translated thesaurus terms to produce additional translations [5]. Some words (e.g., foreign place names) remained untranslated when none of the three sources yielded a usable translation.

One simple way of automatically identifying points within an interview at which replay should be started is to divide the interview into passages, and then to return the start time of the passages that best match a query. Very short passages would not contain enough words to reliably match the query, so we arbitrarily chose roughly 400 words as a passage length.[1] Non-overlapped passages of that length would yield a larger temporal granularity than would be desirable, so a 67% overlap was used (i.e., passage start times were spaced about 133 words apart). In practice, this makes the minimum temporal granularity for start times about 75 seconds, roughly five times larger than the temporal granularity of the relevance assessments for this test collection (see Section 2.3).

The 11,377 resulting passages can then be treated as "documents" for which the following fields are provided:

- DOCNO. The specification for the start time of a segment, in the "VHF[interview]-[seconds]" format required for scoring.
- INTERVIEWDATA. The first name and last initial of the interviewee; additional names (e.g., maiden names and aliases) may also be present. This field is the same for every passage that is drawn from the same interview.
- ENGLISHMANUKEYWORD. This field was intended to contain thesaurus terms that has been manually assigned to a time that fell within the segment, but a script bug resulted in inclusion of thesaurus terms from earlier in the interview. Terms found in this field are therefore not useful in this version of the collection.
- CZECHMANUKEYWORD. The automatically produced English translations of the thesaurus terms in the ENGLISHMANUALKEYWORD field. Because of the errors in that field, the CZECHMANUALKEYWORD field is similarly unusable.
- ASRSYSTEM. Usually 2006, which was the more recent (and hence the more accurate) of the two ASR systems. In the rare instances when no words were produced by the 2006 system, this value is 2004. The 2004 system had

[1] Specifically, the sum of the word durations in each passage, excluding all silences, is exactly 3 minutes.

been designed to transcribe colloquial Czech. In the 2006 system, lexical substitution was used to generate formal Czech.
- ASRTEXT. The words produced by the ASR channel that produced the largest number of words for that passage (usually this is the channel assigned to the interviewee).
- CHANNEL. The stereo ASR channel that was automatically chosen (left or right).
- ENGLISHAUTOKEYWORD. English terms from the same thesaurus that were automatically assigned based on words found in the ASR stream. A k-NN classifier was trained for this purpose using English data (manually assigned thesaurus terms and manually written segment summaries) and run on the automatically generated English translations of the Czech ASRTEXT (produced using a probabilistic dictionary - see [6] for details). Because the classifier was trained on data in which thesaurus terms were associated with segments rather than start points, the natural interpretation of an automatically assigned thesaurus term is that the classifier believes that the indicated topic is associated with the words spoken in the given passage.
- CZECHAUTOKEYWORD. Automatically produced Czech translations of the thesaurus terms in the ENGLISHAUTOKEYWORD field.

For example:

```
<DOC>
<DOCNO>VHF10325-1080.34</DOCNO>
<INTERVIEWDATA>Alexej H...</INTERVIEWDATA>
<ENGLISHMANUKEYWORD>social relations in prisons</ENGLISHMANUKEYWORD>
<CZECHMANUKEYWORD>společenská vztahy v vězení</CZECHMANUKEYWORD>
<ASRSYSTEM>2006</ASRSYSTEM>
<CHANNEL>left</CHANNEL>
<ASRTEXT> PĚKNĚ TAKŽE NĚKDY I TY I TY HLÍDAČI NE </s> <s> TO MYSLÍ
POSLOUCHALI POSLOUCHALI TO CO ZPÍVÁM HO NECHAL CELKEM ASI TO ZPÍVAL
ONI NÁS JAKO NE </s> <s> KDE MÁTE NA MYSLI </s> <s> NO TAM JSME
BYLI ASI </s> <s> DO JARA ROKU ČTYŘICET </s> <s> NO TO SEM NÁS
VOZILI NA NA NA SILNICI A NÁM SE PODAŘILO NĚKOLIKA ...</ASRTEXT>
<ENGLISHAUTOKEYWORD>fate of loved ones | living conditions in the
camps | Poland 1941 (June 21) - 1944 (July 21) | Germany 1945
(January 1 - May 7) ...</ENGLISHAUTOKEYWORD>
<CZECHAUTOKEYWORD>osudy blízkých | životní podmínky v táborech |
Polsko 1941 (21. červen) - 1944 (21. červen) | Německo 1945
(1. leden - 7. květen) ...</CZECHAUTOKEYWORD>
</DOC>
```

2.2 Topics

Currently there are 115 topics specified for the collection. All of them were originally constructed in English and then translated into Czech and in some cases

adapted in order to increase the number of relevant passages in the collection (this was usually done by removing geographic restrictions from the topic). As for the original English topics, they were mostly compiled from the real requests made by scholars, educators and documentary film makers to the administrators of the archive.

The topics are represented in the well-known TREC-style format as shown in the following example:

```
<top>
<num>1225</num>
<title>Osvobození Buchenwaldu a Dachau </title>
<desc>Výpovědi svědků osvobození koncentračních táborů Buchenwald
a Dachau.</desc>
<narr>Relevantní materiál by měl zahrnovat příběhy přeživších nebo
osvoboditelů popisující tyto události. Osvobození jiných táborů
není relevantní.</narr>
</top>
```

where the `<title>`, `<desc>` (description) and `<narr>` (narrative) fields gradually provide more detailed specification about the user's request. Both Czech and English versions of the topics are available for searching the Czech collection to the CLEF participants.

2.3 Relevance judgments

Relevance judgments are prepared within the CLEF evaluation campaign at Charles University in Prague. In 2006 the judgments were completed for a total of 29 topics by five domain experts. The assessors were Czech native speakers with a good knowledge of English. They were working 20 hours a week in average for a period of five months. The two-sided relevance assessment process was performed in two phases supported by an advanced search system designed especially for this task at the University of Maryland in College Park. Full Czech ASR transcript of the best audio channel and the manually assigned keywords from English thesaurus were indexed as overlapping passages as described in Section 2.1.

Search-guided assessment. Each assessor processed one topic at a time. The assessor started with an individual topic research using external resources (such as books, encyclopedias, and web pages) followed by a presentation of the topic to the other assessors and discussion aimed at detailed specification of the topic relevance that all assessors agreed on. Then the assessor iterated between formulating the topic-related queries (using either ASR or thesaurus terms, or both) and searching the collection for interviews containing potentially relevant passages.

Each promising interview was displayed in a detailed view providing an interactive search capability, showing the Czech ASR transcript, English thesaurus

terms, and the possibly relevant passages identified by graphical depiction of the retrieval status value. The assessor could scroll through the interview, search using either type of thesaurus terms, and also replay the audio from any point in order to identify the start time and end time of the relevant periods by indicating points on the transcript. The mGAP measure (see Section 2.4) employs only the start times (converted to 15-second granularity), however both start and end times are available for future research. A screenshot from the interface used by the assessors is shown in Fig. 1.

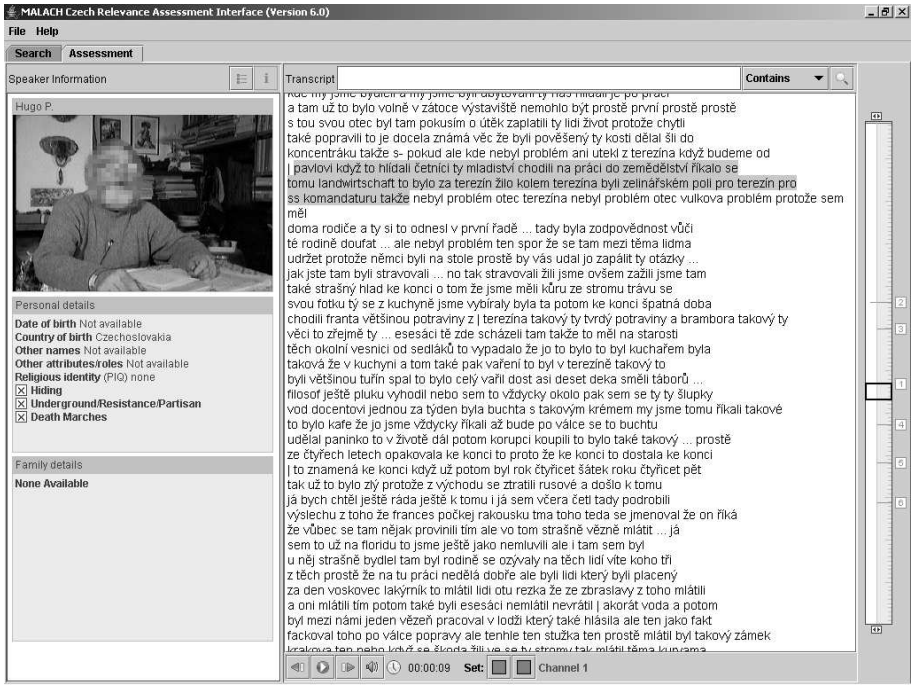

Fig. 1. Screenshot of the assessment interface

At least five relevant passages were required to minimize the effect of quantization noise on the computation of mGAP. Twenty nine such topics were distributed to the CLEF 2006 participants as evaluation topics. Each participating team employed its existing information retrieval systems on these topics and submitted maximum of five official runs.

Highly-ranked assessment. Following completion of the search-guided relevance assessment process, the assessors were provided with a set of additional interviews to check for relevant segments. These interviews were derived from highly ranked passages identified by the systems from the teams participating in

the evaluation. Each such interview was checked and relevant passages found in this way were added to those coming from search-guided assessment to produce the final set of relevance judgments comprising of a total of 1,322 start times for relevant passages identified with an average of 46 relevant passages per topic.

In 2007 the currently ongoing relevance assesment process follows the same rules and the collection is expected to be enriched by judgments for new 25-30 topics.

2.4 Evaluation measure

The evaluation measure called mean Generalized Average Precision (mGAP) is designed to suit the specific needs of this collection — it is sensitive to errors in the start time, but not in the end time, of passages retrieved by the system. It is computed in the same manner as well-know mean average precision, but with one important difference: partial credit is awarded in a way that rewards system-recommended start times that are close to those chosen by assessors. After a simulation study, we chose a symmetric linear penalty function that reduces the credit for a match by 0.1 (absolute) for every 15 seconds of mismatch (either early or late) (see [7] for details). Thus differences at or beyond a 150 second error are treated as a no-match condition.

Relevance judgments are drawn without replacement so that only the highest-ranked match (including partial matches) can be scored for any relevance assessment; other potential matches receive a score of zero. Such approach might represent a pitfall, especially in the case of using the overlapping passages. Specifically, the start time of the highest-ranking passage that matches (however poorly) a passage start time in the relevance judgments will "use up" that judgment. Subsequent passages in which the same matching terms were present would then receive no credit at all, even if they were closer matches than the highest-ranking one.

3 Initial experiments

The first experiments were performed on the set of 11,377 passages (see Section 2.1), using quite a simple document-oriented IR system based on the *tf.idf* model. Detailed description of the experiments can be found in [8], here we will only summarize the most important findings:

- The artificially created "documents", however not topically coherent, are usable for initial experiments with the described collection as the system is indeed able to identify a significant number of relevant starting points.
- The best result was achieved when only the ASRTEXT field was indexed. We knew that the manually assigned keywords were misaligned, but the poor performance of the indexes involving automatic keywords was surprising. Manual examination of a few CZECHAUTOKEYWORD fields indeed indicates a low density of terms that appear as if they match the content

of the passage, but additional analysis will be needed before we can ascribe blame between the transcription, classification and translation stages in the cascade that produced those keyword assignments.
– Proper linguistic preprocessing seems to be indispensable for good performance of the Czech IR system - both lemmatization and stemming boosted the performance almost by a factor of two in comparison with the runs using the original word forms.

4 Conclusion and future work

The presented test collection constitutes a valuable resource for facilitating research into IR for the Czech language. As was already mentioned, the collection is going to be further enriched for this year's CLEF campaign. Moreover, the assessors at Charles University are preparing also the document-oriented, text-based Czech collection for the CLEF Ad-Hoc track. Once these collections are completed, we will have a rather rich set of resources for experiments with Czech IR systems. The development of such systems is our current top priority — so far we have employed only the standard (document-oriented) IR approaches, not reflecting the specific nature of the collection described in this paper and taking into account the properties of the Czech language only to a limited extent.

References

1. Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.J.: Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. IEEE Transactions on Speech and Audio Processing **12**(4) (2004) 420–435
2. Oard, D., Soergel, D., Doermann, D., Huang, X., Murray, G.C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S.: Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In: Proceedings of SIGIR 2004, Sheffield, UK (2004) 41–48
3. Shafran, I., Byrne, W.: Task-Specific Minimum Bayes-risk Decoding Using Learned Edit Distance. In: Proceedings of ICSLP 2004, Jeju Island, South Korea (2004) 1945–1948
4. Shafran, I., Hall, K.: Corrective Models for Speech Recognition of Inflected Languages. In: Proceedings of EMNLP 2006, Sydney, Australia (2006) 390–398
5. Murray, C., Dorr, B.J., Lin, J., Hajič, J., Pecina, P.: Leveraging Reusability: Cost-effective Lexical Acquisition for Large-scale Ontology Translation. In: Proceedings of ACL 2006, Sydney, Australia (2006) 945–952
6. Olsson, S., Oard, D., Hajič, J.: Cross-Language Text Classification. In: Proceedings of SIGIR 2005, Salvador, Brazil (2005) 645–646
7. 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
8. Ircing, P., Oard, D., Hoidekr, J.: First Experiments Searching Spontaneous Czech Speech. In: Proceedings of SIGIR 2007, Amsterdam, The Netherlands (2007)