Overview of the CLEF-2006 Cross-Language Speech Retrieval Track

Douglas W. Oard[1], Jianqiang Wang[2], Gareth J.F. Jones[3], Ryen W. White[4], Pavel Pecina[4], Dagobert Soergel[5], Xiaoli Huang[5], and Izhak Shafran[6]

[1] College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
oard@umd.edu
[2] Department of Library and Information Studies
State University of New York at Buffalo, Buffalo, NY 14260 USA
jw254@buffalo.edu
[3] School of Computing
Dublin City University, Dublin 9, Ireland
Gareth.Jones@computing.dcu.ie
[4] Microsoft Research
One Microsoft Way, Redmond, WA 98052 USA
ryenw@microsoft.com
[5] MFF UK, Malostranske namesti 25, Room 422
Charles University, 118 00 Praha 1, Czech Republic
pecina@ufal.mff.cuni.cz
[6] College of Information Studies
University of Maryland, College Park, MD 20742 USA
dsoergel@umd.edu, xiaoli@umd.edu
[7] OGI School of Science & Engineering
Oregon Health and Sciences University
20000 NW Walker Rd, Portland, OR 97006, USA
zak@cslu.ogi.edu

Abstract. The CLEF-2006 Cross-Language Speech Retrieval (CL-SR) track included two tasks: to identify topically coherent segments of English interviews in a known-boundary condition, and to identify time stamps marking the beginning of topically relevant passages in Czech interviews in an unknown-boundary condition. Five teams participated in the English evaluation, performing both monolingual and cross-language searches of speech recognition transcripts, automatically generated metadata, and manually generated metadata. Results indicate that the 2006 English evaluation topics are more challenging than those used in 2005, but that cross-language searching continued to pose no unusual challenges when compared with monolingual searches of the same collection. Three teams participated in the monolingual Czech evaluation using a new evaluation measure based on differences between system-suggested and ground truth replay start times, with results that were broadly comparable to those observed for English.

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 744–758, 2007.

1 Introduction

The 2006 Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track continued the focus started in 2005 on ranked retrieval from spontaneous conversational speech. Automatically transcribing spontaneous speech has proven to be considerably more challenging than transcribing the speech of news anchors for the Automatic Speech Recognition (ASR) techniques on which fully-automatic content-based search systems are based.

The CLEF 2005 CL-SR task had focused solely on searching English interviews. For CLEF 2006, 30 new topics were developed for the same collection of English interviews, and an improved ASR transcript with better accuracy for the same set of interviews was added. This made it possible to validate retrieval techniques that had been shown to be effective with last year's topics, and to further explore the influence of ASR accuracy on retrieval effectiveness. The CLEF 2006 CL-SR track also added a new task of searching Czech interviews.

As with CLEF 2005, the English task was again based on a known-boundary condition for topically coherent segments. The Czech search task was based on an unknown-boundary condition in which participating systems were required to specify a replay start time for the beginning of each distinct topically relevant passage.

The first part of this paper describes the English language CL-SR task and summarizes the participants' submitted results. That is followed by a description of the Czech language task with corresponding details of submitted runs.

2 English Task

The structure of the CLEF 2006 CL-SR English task was identical to that used in 2005. Two English collections were released this year. The first release (on March 14, 2006) contained all material that was now available for training (i.e., both the training and the test topics from the CLEF 2005 CL-SR evaluation). There was one small difference from the CLEF 2005 CL-SR track data release: each person's last name that appears in the INTERVIEWDATA field (or in the associated XML data files) was reduced into its initial followed by three dots (e.g., "Smith" became "S..."). This first release contained a total of 63 topics, 8,104 topically coherent segments (the equivalent of "documents" in a classic IR evaluation), and 30,497 relevance judgments.

The second release (on June 5, 2006) included a re-release of all the training materials (unchanged), an additional 42 candidate evaluation topics (30 new topics, plus 12 other topics for which relevance judgments had not previously been released), and two new fields based on an improved ASR transcript from the IBM T. J. Watson Research Center.

2.1 Segments

Other than the changes described above, the segments used for the CLEF 2006 CL-SR task were identical to those used for CLEF 2005. Two new fields contain ASR

transcripts of higher accuracy than were available in 2005 (ASRTEXT2006A and ASRTEXT2006B). The ASRTEXT2006A field contains a transcript generated using the best presently available ASR system, which has a mean word error rate of 25% on held-out data. Because of time constraints, however, only 7,378 segments have text in this field. For the remaining 726 segments, no ASR output was available from the 2006A system at the time the collection was distributed. The ASRTEXT2006B field seeks to avoid this no-content condition by including content identical to the ASRTEXT2006A field when available, and content identical to the ASRTEXT2004A field otherwise. Since ASRTEXT2003A, ASRTEXT2004A, and ASRTEXT2006B contain ASR text that was automatically generated for all 8,104 segments, any (or all) of them can be used for the required run based on automatic data. A detailed description of the structure and fields of the English segment collection is given in last year's track overview paper [1].

2.2 Topics

A total of 30 new topics were created for this year's evaluation from actual requests received by the USC Shoah Foundation Institute for Visual History and Education.[1] These were combined with 12 topics that had been developed in previous years, but for which relevance judgments had not been released. This resulted in a set of 42 topics that were candidates for use in the evaluation.

All topics were initially prepared in English. Translations into Czech, Dutch, French, German, and Spanish were created by native speakers of those languages. With the exception of Dutch, all translations were checked for reasonableness by a second native speaker of the language.[2]

A total of 33 of the 42 candidate topics were used as a basis for the official 2006 CL-SR evaluation; the remaining 9 topics were excluded because they had either too few known relevant segments (fewer than 5) or too high a density of known relevant segments among the available judgments (over 48%, suggesting that many relevant segments may not have been found). Participating teams were asked to submit results for all 105 available topics (the 63 topics in the 2006 training set and the 42 topics in the 2006 evaluation candidate set) so that new pools could be formed to perform additional judgments on the development set if additional assessment resources become available.

2.3 Evaluation Measure

As in the CLEF-2005 CL-SR track, we report Mean uninterpolated Average Precision (MAP) as the principal measure of retrieval effectiveness. Version 8.0 of the trec_eval program was used to compute this measure.[3]

[1] On January 1, 2006 the University of Southern California (USC) Shoah Foundation Institute for Visual History and Education was established as the successor to the Survivors of the Shoah Visual History Foundation, which had originally assembled and manually indexed the collection used in the CLEF CL-SR track.

[2] A subsequent quality assurance check for Dutch revealed only a few minor problems. Both the as-run and the final corrected topics will therefore be released for Dutch.

[3] The trec_eval program is available from http://trec.nist.gov/trec_eval/.

2.4 Relevance Judgments

Subject matter experts created multi-scale and multi-level relevance assessments in the same manner as was done for the CLEF-2005 CL-SR track [1]. These were then conflated into binary judgments using the same procedure as was used for CLEF-2005: the union of direct and indirect relevance judgments with scores of 2, 3, or 4 (on a 0–4 scale) were treated as topically relevant, and any other case as non-relevant. This resulted in a total of 28,223 binary judgments across the 33 topics, among which 2,450 (8.6%) are relevant.

2.5 Techniques

The following gives a brief description of the methods used by the participants in the English task. Additional details are available in each team's paper.

University of Alicante (UA). The University of Alicante used the MINIPAR parser to produce an analysis of syntactic dependencies in the topic descriptions and in the automatically generated portion of the collection. They then used these results in combination with their locally developed IR-n system to produce overlapping passages. Their experiments focused on combining these sources of evidence and on optimizing search effectiveness using pruning techniques.

Dublin City University (DCU). Dublin City University used two systems based on the Okapi retrieval model. One version used Okapi with their summary-based pseudo-relevance feedback method. The other system explored combination of multiple segment fields using the BM25F variant of Okapi weights. That second system was also used to explore the use of a field-based method for term selection in query expansion with pseudo-relevance feedback. The officially submitted DCU runs were adversely affected by a formatting error; the results reported for DCU in this paper are based on a subsequent re-submission that corrected that error.

University of Maryland (UMD). The University of Maryland team created two indexes, using the InQuery system in both cases. One index used only automatically generated fields, the other used only manually generated fields. Retrieval based on manually generated fields yielded MAP comparable to that reported in 2005, but retrieval based on automatically generated fields yielded MAP considerably below the 2005 values. Three CLIR experiments were conducted using French topics, but an apparent domain mismatch between the source of the translation probabilities and the CLEF CL-SR test collection resulted in relatively poor MAP for all cross-language runs.

Universidad Nacional de Educacin a Distancia (UNED). The UNED team compared the utility of the 2006 ASR with manually generated summaries

and manually assigned keywords. A CLIR experiment was performed using Spanish queries with the 2006 ASR. The system used was the same as that for their participation in the CLEF 2005 CL-SR tasks [2].

University of Ottawa (UO). The University of Ottawa used two information retrieval systems in their experiments, SMART and Terrier, each with one way of computing term weights. Two query expansion techniques were tried, including a new method based on log-likelihood scores for collocations. ASR transcripts with different estimated word error rates from 2004 and 2006 were indexed individually, together, and in combination with automatic classification results. A separate index using manually generated metadata was also built. Cross-language experiments were run using French and Spanish topics by automatically translating the topics into English using the combined results of several online machine translation tools. Results for an extensive set of locally scored runs were also reported.

University of Twente (UT). The University of Twente employed a locally developed XML retrieval system to flexibly index the collection in a way that permitted experiments to be run with different combinations of fields without reindexing. Cross-language experiments were conducted using Dutch topics, both with ASR and with manually created metadata.

2.6 English Evaluation Results

Table 2.6 summarizes the results for all 30 official runs averaged over the 33 evaluation topics, listed in descending order of MAP. Teams were asked to run at least one monolingual condition using the title and description fields of the topics and indexing only automatically generated fields; those "required runs" are shown in bold as a basis for cross-system comparisons. For the required run, DCU yielded a MAP statistically significantly better (by a Wilcoxon signed rank test for paired samples, with $p < 0.05$) than the next three teams (UO, UMD,. and UT, which were statistically indistinguishable from each other). Further, the best results for this year (0.0747 from DCU) are considerably below (i.e., just 58% of) last year's best MAP. For manually generated metadata, this year's best result (0.2902) is 93% of last year's best result, however, the difference is not statistically significant. From this we conclude that this year's topic set seems somewhat less well matched with the ASR results, but that the topics are not otherwise generally much harder for information retrieval techniques based on term matching. CLIR also seemed to pose no unusual challenges with this year's topic set, with the best CLIR on automatically generated indexing data (a French run from the University of Ottawa) achieving 83% of the MAP achieved by a comparable monolingual run. Similar effects were observed with manually generated metadata (at 80% of the corresponding monolingual MAP for Dutch queries, from the University of Twente).

Table 1. English official runs. Bold runs are the required condition. N = Name (Manual metadata), MK = Manual Keywords (Manual metadata), SUM = Summary (Manual metadata), ASR4 = ASRTEXT2004A (Automatic), ASR6 = ASRTEXT2006B (Automatic), AK1 = AUTOKEYWORD2004A1 (Automatic), AK2 = AUTOKEYWORD2004A2. See [1] for descriptions of these fields. (Automatic).

Run name	MAP	Lang	Query	Doc field	Site
uoEnTDNtMan	0.2902	EN	TDN	MK,SUM	UO
3d20t40f6sta5flds	0.2765	EN	TDN	ASR6,AK1,AK2,N,SUM,MK	DCU
umd.manu	0.2350	EN	TD	N,MK,SUM	UMD
UTsummkENor	0.2058	EN	T	MK,SUM	UT
dcuEgTDall	0.2015	EN	TD	ASR6,AK1,AK2,N,SUM,MK	DCU
uneden-manualkw	0.1766	EN	TD	MK	UNED
UTsummkNl2or	0.1654	NL	T	MK,SUM	UT
dcuFchTDall	0.1598	FR	TD	ASR6,AK1,AK2,N,SUM.MK	DCU
umd.manu.fr.0.9	0.1026	FR	TD	N,MK,SUM	UMD
umd.manu.fr.0	0.0956	FR	TD	N,MK,SUM	UMD
unedes-manualkw	0.0904	ES	TD	MK	UNED
unedes-summary	0.0871	ES	TD	SUM	UNED
uoEnTDNsQEx04A	0.0768	EN	TDN	ASR4,AK1,AK2	UO
dcuEgTDauto	**0.0733**	EN	TD	ASR6,AK1,AK2	DCU
uoFrTDNs	0.0637	FR	TDN	ASR4,AK1,AK2	UO
uoSpTDNs	0.0619	ES	TDN	ASR4,AK1,AK2	UO
uoEnTDt04A06A	**0.0565**	EN	TD	ASR4,ASR6,AK1,AK2	UO
umd.auto	**0.0543**	EN	TD	ASR4,ASR6,AK1,AK2	UMD
UTasr04aEN	0.0495	EN	T	ASR4	UT
dcuFchTDauto	0.0462	FR	TD	ASR6,AK1,AK2	DCU
UA_TDN_FL_ASR06BA1A2	0.0411	EN	TDN	ASR6,AK1,AK2	UA
UA_TDN_ASR06BA1A2	0.0406	EN	TDN	ASR6,AK1,AK2	UA
UA_TDN_ASR06BA2	0.0381	EN	TDN	ASR6,AK2	UA
UTasr04aNl2	0.0381	NL	T	ASR4	UT
UTasr04aEN-TD	**0.0381**	EN	TD	ASR4	UT
uneden	**0.0376**	EN	TD	ASR6	UNED
UA_TD_ASR06B	**0.0375**	EN	TD	ASR6	UA
UA_TD_ASR06BA2	0.0365	EN	TD	ASR6,AK2	UA
unedes	0.0257	ES	TD	ASR6	UNED
umd.auto.fr.0.9	0.0209	FR	TD	ASR4,ASR6,AK1,AK2	UMD

3 Czech Task

The goal of the Czech task was to automatically identify the start points of topically-relevant passages in unsegmented interviews. Ranked lists for each topic were submitted by each system in the same form as the English task, with the single exception that a system-generated starting point was specified rather than a document identifier. The format for this was "VHF[IntCode].[starting-time]," where "IntCode" is the five-digit interview code (with leading zeroes added) and "starting-time" is the system-suggested replay starting point (in seconds) with

reference to the beginning of the interview.[4] Lists were to be ranked by systems in the order that the system would suggest for listening to passages beginning at the indicated points.

3.1 Interviews

The Czech task was broadly similar to the English task in that the goal was to design systems that could help searchers identify sections of an interview that they might wish to listen to. The processing of the Czech interviews was, however, different from that used for English in three important ways:

- No manual segmentation was performed. This alters the format of the interviews (which for Czech is time-oriented rather than segment-oriented), it alters the nature of the task (which for Czech is to identify replay start points rather than to select among predefined segments), and it alters the nature of the manually assigned metadata (there are no manually written summaries for Czech and the assignment of a manual thesaurus term implies that the discussion of a topic started at that time).
- The two available Czech ASR transcripts (2004 and 2006) were generated using different ASR systems. In both cases, the acoustic models were trained using 15-minutes snippets from 336 speakers, all of whom are present in the test set as well. However, the language model for the 2006 system was created by interpolating two models—an in-domain model from transcripts, and an out-of-domain model from selected portions of the Czech National Corpus. For details, see the baseline systems described in [3,4]. Apart from the improvement in transcription accuracy, the 2006 system differs from the 2004 system in that the transcripts are produced in formal Czech, rather than the colloquial Czech that was produced in 2004. Since the topics were written in formal Czech, the 2006 ASR transcripts would be expected to yield better matching. Interview-specific vocabulary priming (adding proper names to the recognizer vocabulary based on names present in a pre-interview questionnaire) was not done for either Czech system. Thus, a somewhat higher error rate on named entities might be expected for the Czech systems than for the two English systems (2004 and 2006) in which vocabulary priming was included.
- ASR is available for both the left and right stereo channels (which were usually recorded from microphones with different positions and orientations).

Because the task design for Czech is not directly compatible with the design of document-oriented IR systems, we provided a "quickstart" package containing the following:

[4] The interviews were initially recorded on separate tapes, and we adopted the convention that the start time of a tape was defined to be the end time of the last word recognized by the 2006 system for the channel that yielded the largest number of recognized words (across all tapes) for that system. Note that these tape start times are the same for all four transcripts (2004 and 2006, left and right channels).

- A quickstart script for generating overlapping passages directly from the ASR transcripts. The passage duration (in seconds), the spacing between passage start times (also in seconds), and the desired ASR system (2004 or 2006) could be specified. The default settings (180, 60, and 2006) were intended to result in 3-minute passages with start times spaced one minute apart. A design error in the quickstart scripts resulted in use of the sum of the word durations rather than the word start times; hence the default passage length turned out to be about 4 minutes (and, somewhat more precisely, an average of 403 words).
- A quickstart collection created by running the quickstart script with the default settings. This collection contains 11,377 overlapping passages.

The quickstart collection contains the following automatically generated fields:

DOCNO. The DOCNO field contains a unique document number in the same format as the start times that systems were required to produce in a ranked list. This design allowed the output of a typical IR system to be used directly as a list of correctly formatted (although perhaps not very accurate) start times for scoring purposes.[5]

ASRSYSTEM. specifying the type of the ASR transcript, where "2004" and "2006" denote colloquial and formal Czech transcipts respectively. In addition, the "2006" transcript benefited from improvements in the ASR system, which were developed at Johns Hopkins University with input from University of West Bohemia. By default, the quickstart collection would use 2006 transcripts when they were available; otherwise 2004 transcripts were used.[6]

CHANNEL. The CHANNEL field specifies which recorded channel (left or right) was used to produce the transcript. The channel that produced the greatest number of total words over the entire transcript (which is usually the channel that produced the best ASR accuracy for words spoken by the interviewee) was automatically selected by default. This automatic selection process was hardcoded in the script, although the script could be modified to generate either or both channels.

ASRTEXT. The ASRTEXT field contains words in order from the transcript selected by ASRSYSTEM and CHANNEL for a passage beginning at the start time indicated in DOCNO. When the selected transcript contains no words at all for a tape, words are drawn from one alternate source that is chosen in the following priority order: (1) the same ASRSYSTEM from the other CHANNEL, (2) the same CHANNEL from the other ASRSYSTEM, or (3) the other CHANNEL from the other ASRSYSTEM.

[5] The design error in the script that generated the quickstart collection resulted in incorrect times appearing in the DOCNO field in the distributed collection. Results reported in this volume are based on automatic correction of the times in each DOCNO.

[6] Transcripts for some tapes were not available from the 2006 Czech ASR system at the time the collection was built.

ENGLISHAUTOKEYWORD. The ENGLISHAUTOKEYWORD field contains a set of thesaurus terms that were assigned automatically using a k-Nearest Neighbor (kNN) classifier based solely on words from the ASRTEXT field of the passage; the top 20 thesaurus terms are included in best-first order. Thesaurus terms (which may be phrases) are separated with a vertical bar character. The classifier was trained using English data (manually assigned thesaurus terms and manually written segment summaries) and run on automatically produced English translations of the 2006 Czech ASRTEXT [5]. Two types of thesaurus terms are present, but not distinguished: (1) terms that express a subject or concept; (2) terms that express a location, often combined with time in one precombined term [6]. Because the classifier was trained on the English collection, in which thesaurus terms were assigned to segments, the natural interpretation of an automatically assigned thesaurus term is that the classifier believes the indicated topic is associated with the words spoken in this passage. Note that this differs from the way in which the presence of a manually assigned thesaurus term (described below) should be interpreted.

CZECHAUTOKEYWORD. The CZECHAUTOKEYWORD field contains Czech translations of the ENGLISHAUTOKEYWORD field. These translations were obtained from three sources: (1) professional translation of about 3,000 thesaurus terms, (2) volunteer translation of about 700 thesaurus terms, and (3) a custom-built machine translation system that reused words and phrases from manually translated thesaurus terms to produce additional translations [6]. Some words (e.g., foreign place names) remained untranslated when none of the three sources yielded a usable translation.

Three additional fields containing data produced by human indexers at the Survivors of the Shoah Visual History Foundation were also available for use in contrastive conditions:

INTERVIEWDATA. The INTERVIEWDATA field contains the first name and last initial for the person being interviewed. This field is identical for every passage that was generated from the same interview.

ENGLISHMANUKEYWORD. The ENGLISHMANUALKEYWORD field was intended to contain thesaurus terms that were manually assigned with one-minute granularity from a custom-built thesaurus by subject matter experts at the Survivors of the Shoah Visual History Foundation while viewing the interview. The format is the same as that described for the ENGLISHAUTOKEYWORD field, but the meaning of a keyword assignment is different. In the Czech collection, manually assigned thesaurus terms are used as onset marks—they appear only once at the point where the indexer recognized that a discussion of a topic or location-time pair had started; continuation and completion of discussion are not marked. Unfortunately, the design error in the quickstart script resulted in misplacement of the manually assigned thesaurus terms in later segments of the quickstart collection than was appropriate.

CZECHMANUKEYWORD. The CZECHMANUALKEYWORD field contains Czech translations of the English thesaurus terms that were produced from the ENGLISHMANUALKEYWORD field using the process described above.[7]

All three teams used the quickstart collection; no other approaches to segmentation and no other settings for passage length or passage start time spacing were tried.

3.2 Topics

At the time the Czech evaluation topics were released, it was not yet clear which of the available topics were likely to yield a sufficient number of relevant passages in the Czech collection. Participating teams were therefore asked to run 115 topics—every available topic at that time. This included the full 105 topic set that was available in 2006 for English (including all training and all evaluation candidate topics) and adaptations of 10 topics from that set in which geographic restrictions had been removed (as insurance against the possibility that the smaller Czech collection might not have adequate coverage for exactly the same topics).

All 115 topics had originally been constructed in English and then translated into Czech by native speakers. Since translations into languages other than Czech were not available for the 10 adapted topics, only English and Czech topics were distributed with the Czech collection. No teams used the English topics this year; all official runs this year with the Czech collection were monolingual.

Two additional topics were created as part of the process of training relevance assessors, and those topics were distributed to participants along with a (possibly incomplete) set of relevance judgments. However, this distribution occurred too late to influence the design of any participating system.

3.3 Evaluation Measure

The evaluation measure that we chose for Czech is designed to be sensitive to errors in the start time, but not in the end time, of system-recommended passages. It is computed in the same manner as 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) [7]. This results in the same computation as the well-known mean Generalized Average Precision (mGAP) measure that was introduced to deal with human assessments of partial relevance [8]. In our case, the human assessments are binary; it is the degree of match to those assessments that can be partial. Relevance judgments are used without replacement, which means that only the highest

[7] Because the ENGLISHMANUALKEYWORD field is not correct in the quickstart collection, the CZECHMANUALKEYWORD field is also incorrect.

ranked match (including partial matches) can be scored for any relevance assessment; other potential matches receive a score of zero. Differences at or beyond a 150 second error are treated as a no-match condition, thus not "using up" a relevance assessment. Systems could therefore optimize their mGAP score by automatically removing near-duplicates (i.e., start points that are close in time) that occur lower in the list, although no system did so this year.

3.4 Relevance Judgments

Relevance judgments were completed at Charles University in Prague for a total of 29 Czech topics by subject matter experts who were native speakers of Czech. All relevance assessors had good English reading skills. Topic selection was performed by individual assessors, subject to the following factors:

- At least five relevant start times in the Czech collection were required in order to minimize the effect of quantization noise on the computation of mGAP, so assessors were encouraged to perform initial triage searches and then focus on topics that they expected to meet this criterion.
- The greatest practical degree of overlap with topics for which relevance judgments were available in the English collection was desirable, so choosing to assess one of the "adapted" topics (for which no English relevance judgments are available) was discouraged.

Once a topic was selected, the assessor iterated between topic research (using external resources) and searching the collection. A new search system was designed to support this interactive search process. The best channel of the Czech ASR and the manually assigned English thesaurus terms were indexed as overlapping passages, and queries could be formed using either or both. Once a promising interview was found, an interactive search within the interview could be performed using either type of term. Promising regions (according to the system) were then highlighted for the assessor using a graphical depiction of the retrieval status value. Assessors could then scroll through the interview using that graphical depiction in conjunction with the displayed English thesaurus terms and the displayed ASR transcript as a basis for recognizing regions that they believed might be topically relevant. They could then replay the audio from any point in order to confirm topical relevance. As they did this, they could indicate the onset and conclusion of the relevant period by designating points on the transcript that were then automatically converted to times with 15-second granularity.[8] Only the start times are used for computation of the mGAP measure, but both start and end times are available for future research.

[8] Several different types of time spans arise when describing evaluation of speech indexing systems. For clarity, we have tried to stick to the following terms when appropriate: manually defined segments (for English indexing), 15-minute snippets (for ASR training), 15-second increments (for the start and end time of Czech relevance judgments), relevant passages (identified by Czech relevance assessors), and automatically generated passages (for the quickstart collection).

Once that search-guided relevance assessment process was completed, the assessors were provided with a set of additional "highly ranked" points to check for topical relevance that were computed using a pooling technique similar to that used for English. The top 50 start times from every official run were pooled, duplicates (at one minute granularity) were removed, and the results were inserted into the assessment system as system recommendations. Every system recommendation was checked, although assessors exercised judgment regarding when it would be worthwhile to actually listen to the audio in order to limit the cost of this process. Relevant passages identified in this way were added to those found using search-guided assessment to produce the final set of relevance judgments (topic 4000 was generalized from a pre-existing topic).[9]

Table 2. Number of the relevant passages identified for each of the 29 topics in the Czech collection

topid	#rel	topid	#rel	topid	#rel	topid	#rel	topid	#rel
1166	8	1181	21	1185	50	1187	26	1225	9
1286	70	1288	9	1310	20	1311	27	1321	27
14312	14	1508	83	1620	35	1630	17	1663	34
1843	52	2198	18	2253	124	3004	43	3005	84
3009	77	3014	87	3015	50	3017	83	3018	26
3020	67	3025	51	3033	45	4000	65		

A total of 1,322 start times for relevant passages were identified, thus yielding an average of 46 relevant passages per topic (minimum 8, maximum 124). Table 2 shows the number of relevant start times for each of the 29 topics, 28 of which are the same as topics used in the English test collection (the exception being topic 4000, an "adapted" topic).

3.5 Techniques

The participating teams all employed existing information retrieval systems to perform monolingual searches of the quickstart collection.

University of Maryland (UMD). The University of Maryland submitted three runs in which they tried combining all the fields (the Czech ASR transcript, automatically and manually assigned Czech thesaurus terms, and the English translations of those thesaurus terms) to form a unified passage index again using Inquery. They compared the retrieval results based on that index with results based on indexing ASR alone or based on indexing only a combination of automatically assigned thesaurus terms and the ASR transcript.

[9] Because all three participating teams used the quickstart collection, which had incorrect segment start times, this highly ranked assessment process was likely less effective than it should have been. For purposes of characterizing the relevance judgments for the 2006 Czech collection, it is therefore reasonable to consider them as having been created principally using a search-guided assessment process.

University of Ottawa (UO). Four runs were submitted from the University of Ottawa using SMART, and a fifth run was submitted using Terrier. The configuration of the SMART runs explored indexed field combinations similar to those tried by the University of Maryland; the Terrier run used ASR in combination with automatically assigned thesaurus terms.

University of West Bohemia (UWB). The University of West Bohemia was the only team to apply morphological normalization and stopword removal for Czech. A TF*IDF model was implemented in Lemur, along with the Lemur implementation of blind relevance feedback. Five runs were submitted for official scoring, and additional runs were scored locally.

Results. The results reported in Table 3 were computed after correction of the times in the DOCID field of the quickstart collection. Because this correction invalidates the use of manually assigned thesaurus terms, their use (and use of their translations) should be interpreted only as a source of noise. As a comparison of runs UWB_mk_aTD and UWB_mk_a_akTD shows, this effect is small. Somewhat surprisingly, there is no evidence that automatically assigned English thesaurus terms or their Czech translations was helpful. Examination of a sample of translation pairs by a native speaker identified remarkably few translation errors, so it seems reasonable to dismiss that factor as a possible cause. The automatic assignment process was based on the Czech terms from the same segment, so time misalignment also seems unlikely to be the cause. Manual examination of a sample of the English thesaurus terms does however, indicate that the assignments appear to bear little relation to the topical content of the segment. We therefore suspect either a deficiency in the classifier design or some broader-scale alignment error in associating the classifier results with ASR segments. That issue has not yet been resolved, and for the analysis in this paper we simply treat the use of automatically assigned thesaurus terms (in either language) as an additional source of noise.

Every run except uoCzEnTDNsMan used the Czech ASR transcript, so we consider those 12 official runs as a group for the purpose of this analysis (ignoring the use of other fields). A strict ordering by participating site is evident, with UWB achieving better results with every one of their runs than any run from any other site. This suggests that handling Czech morphology in some way is particularly important.

These results can almost certainly be improved upon in future experiments. We had originally intended the quickstart collection to be used only for out-of-the-box sanity checks, with the idea that teams would either modify the quickstart scripts or create new systems outright to explore a broader range of possible system designs. Time pressure and a lack of a suitable training collection precluded that sort of experimentation, however, and the result was that this undesirable effect of passage overlap affected every system. The use of relatively long overlapping passages in the quickstart collection probably reduced

Table 3. Czech official runs with passage start times corrected to match ASR. ASR = ASRTEXT, CAK = CZECHAUTOKEYWORD (Automatic), EAK = ENGLISHAUTOKEYWORD (Automatic), CMK = CZECHMANUKEYWORD (Manual metadata), EMK = ENGLISHMANUKEYWORD (Manual metadata).

Run name	mGAP	Lang	Query	Doc field	Site
UWB_mk_a_akTDN	0.0456	CZ	TDN	ASR,CAK,CMK	UWB
UWB_aTD	0.0435	CZ	TD	ASR	UWB
UWB_mk_aTD	0.0416	CZ	TD	ASR,CMK	UWB
UWB_a_akTD	0.0402	CZ	TD	ASR,CAK	UWB
UWB_mk_a_akTD	0.0377	CZ	TD	ASR,CAK,CMK	UWB
uoCzEnTDNsMan	0.0235	CZ,EN	TDN	CAK,CMK,EAK,EMK	UO
uoCzEnTDt	0.0218	CZ,EN	TD	ASR,CAK	UO
uoCzTDs	0.0211	CZ	TD	ASR,CAK	UO
uoCzTDNsMan	0.0200	CZ	TDN	ASR,CAK,CMK	UO
uoCzTDNs	0.0182	CZ	TDN	ASR,CAK	UO
umd.all	0.0070	CZ	TD	ASR,CAK,CMK,EAK,EMK	UMD
umd.asr	0.0066	CZ	TD	ASR	UMD
umd.akey.asr	0.0060	CZ	TD	ASR,CAK,EAK	UMD

mGAP values somewhat because the (roughly) 80-second passage spacing was nearly five times the 15-second granularity of the relevance judgments.

4 Conclusion and Future Plans

The CLEF 2006 CL-SR track extended the previous year's work on the English task by adding new topics, and by introducing a new Czech task with a new unknown-boundary evaluation condition. The results of the English task suggest that the evaluation topics this year posed somewhat greater difficulty for systems doing fully automatic indexing. Studying what made these topics more difficult would be an interesting focus for future work. However, the most significant achievement of this year's track was the development of a CL-SR test collection based on a more realistic unknown-boundary condition. Now that we have both that collection and an initial set of system designs, we are in a good position to explore issues of system and evaluation design.

The CLEF CL-SR track will continue in 2007. For Czech, relevance judgments will be created for additional topics, with a goal of having at least 50 topic for Czech as a legacy from the track for use by future researchers. The present set of 96 topics for English is already adequate for many future purposes, and indeed for a collection of this size construction of additional topics that are representative of real information needs could prove to be impractical. We therefore will run the English task again in 2007 with the same evaluation topics in order to support comparative evaluation of improved systems, but no new topics are planned. Some unique characteristics of the CL-SR collection (e.g., location-oriented topics) may also be of interest to other tracks, including

domain-specific retrieval and geoCLEF. By sharing our resources widely, we hope to maximize the impact of the pioneering research teams that have contributed to the construction of these unique resources.

Acknowledgments

This track would not have been possible without the efforts of a great many people. Our heartfelt thanks go to the dedicated group of relevance assessors in Maryland and Prague, to the Dutch, French and Spanish teams that helped with topic translation, and to Bill Byrne, Martin Cetkovsky, Bonnie Dorr, Ayelet Goldin, Sam Gustman, Jan Hajic, Jimmy Lin, Baolong Liu, Craig Murray, Scott Olsson, Bhuvana Ramabhadran and Deborah Wallace for their help with creating the techniques, software, and data sets on which we have relied.

References

1. White, R.W., Oard, D.W., Jones, G.J.F., Soergel, D., Huang, X.: Overview of the clef-2005 cross-language speech retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. López-Ostenero, F., Peinado, V., Sama, V., Verdejo, F.: UNED@CL-SR CLEF 2005: Mixing different strategies to retrieve automatic speech transcriptions. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
3. Shafran, I., Byrne, W.: Task-specific minimum bayes-risk decoding using learned edit distance. In: Proceedings of INTERSPEECH2004-ICSLP (2004)
4. Shafran, I., Hall, K.: Corrective models for speech recognition of inflected languages. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2006)
5. Olsson, J.S., Oard, D.W., Hajic, J.: Cross-language text classification. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York (2005)
6. Murray, G.C., Dorr, B.J., Lin, J., Hajic, J., Pecina, P.: Leveraging reusability: Cost-effective lexical acquisition for large-scale ontology translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2006)
7. Liu, B., Oard, D.W.: One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York (2006)
8. Kekalainen, J., Jarvelin, K.: Using graded relevance assessments in ir evaluation. Journal of the American Society for Information Science and Technology (2002)