Overview of the CLEF-2007 Cross-Language Speech Retrieval Track

Pavel Pecina[1], Petra Hoffmannová[1], Gareth J.F. Jones[2], Ying Zhang[2], and Douglas W. Oard[3]

[1] MFF UK, Malostranske namesti 25, Room 422
Charles University, 118 00 Praha 1, Czech Republic
pecina@ufal.mff.cuni.cz, hoffmannova@knih.mff.cuni.cz
[2] School of Computing
Dublin City University, Dublin 9, Ireland
gjones@computing.dcu.ie, yzhang@computing.dcu.ie
[3] College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
oard@umd.edu

Abstract. The CLEF-2007 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. Six teams participated in the English evaluation, performing both monolingual and cross-language searches of ASR transcripts, automatically generated metadata, and manually generated metadata. Four teams participated in the Czech evaluation, performing monolingual searches of automatic speech recognition transcripts.

1 Introduction

The 2007 Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track was the third and final year for evaluation of ranked retrieval from spontaneous conversational speech from an oral history collection at CLEF. As in the CLEF 2006 CL-SR task [1], automatically transcribed interviews conducted in English could be searched using queries in one of six languages, and automatically transcribed interviews conducted in Czech could be searched using queries in one of two languages. New relevance judgments for additional topics were created to expand the Czech collection in 2007. The English collection used in 2007 was the same as that used in 2006. As in CLEF 2005 and CLEF 2006, the English task was based on a known-boundary condition for topically coherent segments. The Czech task was based on a unknown-boundary condition in which participants were required to identify a time stamp for the beginning of each distinct topically relevant passage.

The remainder of this paper is organized as follows. Section 2 describes the English task and summarizes the results for the submitted runs. Section 3 does the same for the Czech task. The paper concludes in Section 4 with a brief recap of what has been learned across all three years of the CLEF CL-SR track.

2 English Task

The structure of the CLEF 2007 CL-SR English task was identical to that used in 2006, which we review here briefly (see [1] for more details).

2.1 Segments

The "documents" searched in the English task are 8,104 segments that were designated by professional indexers as topically coherent. A detailed description of the structure and fields of the English segment collection is given in the 2005 track overview paper [2]. Automatically generated transcripts from two Automatic Speech Recognition (ASR) systems are available. The ASRTEXT2006B 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. Only 7,378 segments have text in this field. For the remaining 726 segments, no ASR output was available from that system, so in those cases the ASRTEXT2006B field includes content identical to the ASRTEXT2004A field which was generated using an earlier less accurate transcription system (with a 35% mean word error rate). An extensive set of manually and automatically generated metadata is also available for each segment.

2.2 Topics

The same 63 training topics and 33 evaluation topics were used for the English task this year as had been used in 2006. Participating teams were asked not to use the evaluation topics for system tuning. Translations into Czech, Dutch, French, German, and Spanish had been created by native speakers of those languages. Participating teams were asked to submit runs for 105 topics (the 63 training topics, the 33 evaluation topics, and 9 other topics), but results are reported here only for the 33 evaluation topics.

2.3 Evaluation Measure

As in the CLEF-2006 CL-SR track, we report uninterpolated Mean Average Precision (MAP) as the principal measure of retrieval effectiveness. Version 8.0 of the trec_eval program was used to compute this measure.[4] The Wilcoxon signed-rank signed test was employed for evaluation of significance.

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

2.4 Relevance Judgments

We reused the relevance judgments from the English task of CLEF-2005, which had been created from multi-scale and multi-level relevance assessments performed by subject matter experts [2]. These judgments were 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. The resulting binary relevance judgments were filtered to remove segments which had been judged but had not been included in the test collection. This resulted in a total of 20,560 binary judgments across the 33 topics, among which 2,449 (12%) are relevant.[5]

2.5 Techniques

This section gives a brief description of the methods used by each team participating in the English task. Additional details are available in each team's paper.

Brown University (BLLIP) The Brown Laboratory for Linguistic Information Processing (BLLIP) team extended the basic Dirichlet-smoothed unigram IR model to incorporate bigram mixing and collection smoothing. In their enhanced language model, the bigram and unigram models were mixed using a tunable mixture weight over all documents. They attempted linearly mixing the test collection with two larger text corpora, 40,000 sentences from the Wall Street Journal and 450,000 sentences from the North American News Corpus, in order to alleviate the sparse data problems in the case of small collections. They observed that bigram statistics appeared to have greater impact with pseudo-relevance feedback than without. The collection smoothing approach clearly provided a substantial improvement.

Dublin City University (DCU) Dublin City University concentrated on the issues of topic translation, combining this with search field combination and pseudo-relevance feedback methods used for their CLEF 2006 submissions. Non-English topics were translated into English using the Yahoo! BabelFish free online translation service and with domain-specific translation lexicons gathered automatically from Wikipedia. Combination of multiple fields using the BM25F variant of Okapi weights was explored. Additionally, the DCU team integrated their information retrieval methods based on the Okapi model with summary-based pseudo-relevance feedback.

[5] For CLEF-2006, a less careful filtering resulted in 28,223 binary judgments, of which 2,450 were relevant. The only difference in the relevant subset is that the 2007 judgments contain 33 rather than 34 relevant for topic 3032. Since the computation of uninterpolated MAP by trec_eval is affected only by the relevant subset, uninterpolated MAP values from 2006 and 2007 can reasonably be directly compared without adjustment for differences in the relevance judgments.

University of Amsterdam (UVA) The University of Amsterdam explored the use of character n-gram tokenization to improve the retrieval of documents using automatically generated text, as well as the combination of manually generated with automatically generated text. They reported that $n = 4$ provided the best retrieval effectiveness when a cross-word overlapping n-gram tokenization strategy was used. The field combination was done using the Indri query language, in which varying weights were assigned to different fields. Cross-language experiments were conducted using Dutch topics that were automatically translated into English using two different online tools, SYSTRAN and FreeTranslation. The translations generated from each MT system were then combined as a 'bag-of-words' English query.

University of Chicago (UC) The University of Chicago team focused on the contribution of automatically assigned thesaurus terms to retrieval effectiveness and the utility of different query translation strategies. For French–English cross-language retrieval, they adopted two query translation strategies: MT-based translation using the publicly available translation tool provided by Google, and dictionary-based translation. Their dictionary-based translation procedure applied a backoff stemming strategy in order to support matching with highest precision between the query terms and the bilingual word list. They noted that 27% of the French query terms remained untranslated and were thus retained.

University of Jaén (SINAI) The SINAI group at the University of Jaén investigated the effect of selection of different fields on retrieval effectiveness. An information gain measure was employed to select the best XML tags in the document collection. The tags with higher information gain values were selected to compose the final collection. Their experiments were conducted with the Lemur retrieval information system using applying KL divergence. French, German, Spanish and Dutch topics were translated to English using a translation module, SINTRAM, which works with different online machine translation systems and combines the different translations based on heuristics.

University of Ottawa (UO) The University of Ottawa used weighted summation of normalized similarity measures to combine 15 different weighting schemes from two IR systems (Terrier and SMART). Two query expansion techniques, one based on the thesaurus and the other one on blind relevance feedback, were examined. In their cross-language experiments, the queries were automatically translated from French and Spanish into English by combining the results of multiple online machine translation tools. Results for an extensive set of locally scored runs were also reported.

2.6 Results

Table 1 summarizes the evaluation results for all 29 official runs averaged over the 33 evaluation topics, listed in descending order of MAP. These 29 runs were

further categorized into four groups based on the query language used (English or non-English) and the document fields (automatic-only or at least one manual assigned) indexed: 9 automatic-only monolingual runs, 6 automatic-only cross-language runs, 9 monolingual runs with manually assigned metadata, and 5 cross-language runs with manually assigned metadata.

Run ID	MAP	Lng	Query	Document Fields	Site
dcuEnTDNmanual	0.2847	EN	TDN	MK,SUM	DCU
uoEnTDtManF1	0.2761	EN	TD	MK,SUM	UO
brown.TDN.man	0.2577	EN	TDN	MK,SUM	BLLIP
dcuEnTDmanualauto	0.2459	EN	TD	MK,SUM,ASR06B	DCU
brown.TD.man	0.2366	EN	TD	MK,SUM	BLLIP
brown.T.man	0.2348	EN	T	MK,SUM	BLLIP
UvA_4_enopt	0.2088	EN	TD	MK,SUM,ASR06B	UVA
dcuFrTDmanualauto	0.1980	FR	TD	MK,SUM,ASR06B	DCU
UvA_5_nlopt	0.1408	NL	TD	MK,SUM,AK2,ASR06B	UVA
uoEnTDtQExF1	0.0855	EN	TD	AK1,AK2,ASR04	UO
uoEnTDtQExF2	0.0841	EN	TD	AK1,AK2,ASR04	UO
brown.TDN.auto	0.0831	EN	TDN	AK1,AK2,ASR06B	BLLIP
dcuEnTDauto	0.0787	EN	TD	AK1,AK2,ASR06B	DCU
brown.TD.auto	0.0785	EN	TD	AK1,AK2,ASR06B	BLLIP
SinaiSp100	0.0737	ES	TD	ALL	SINAI
dcuFrTDauto	0.0636	FR	TD	AK1,AK2,ASR06B	DCU
uoEsTDtF2	0.0619	ES	TD	AK1,AK2,ASR04	UO
uoFrTDtF2	0.0603	FR	TD	AK1,AK2,ASR04	UO
SinaiFr100	0.0597	FR	TD	ALL	SINAI
SinaiEn100	0.0597	EN	TD	ALL	SINAI
SinaiSp050	0.0579	ES	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B	SINAI
UCkwENTD	0.0571	EN	TD	AK1,AK2,ASR06B	UC
SinaiEn050	0.0515	EN	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B	SINAI
UCbaseENTD1	0.0512	EN	TD	ASR06B	UC
UvA_2_en4g	0.0444	EN	TD	AK2,ASR06B	UVA
UvA_1_base	0.0430	EN	TD	ASR06B	UVA
UCkwFRTD1	0.0406	FR	TD	AK1,AK2,ASR06B	UC
UvA_3_nl4g	0.0400	NL	TD	AK2,ASR06B	UVA
UCbaseFRTD1	0.0322	FR	TD	ASR06B	UC

Table 1. Evaluation results for all English official runs. MK = MANUALKEYWORD (Manual metadata), SUM = SUMMARY (Manual metadata), AK1 = AUTOKEYWORD2004A1 (Automatic), AK2 = AUTOKEYWORD2004A2, ASR03 = ASRTEXT2003A (Automatic), ASR04 = ASRTEXT2004A (Automatic), ASR06A = ASRTEXT2006A (Automatic), ASR06B = ASRTEXT2006B (Automatic), and ALL = all fields.

Run ID	MAP	Lng	Query	Document Fields	Site
uoEnTDtQExF1	**0.0855**	EN	TD	AK1,AK2,ASR04	UO
uoEnTDtQExF2	0.0841	EN	TD	AK1,AK2,ASR04	UO
brown.TDN.auto	0.0831	EN	TDN	AK1,AK2,ASR06B	BLLIP
dcuEnTDauto	**0.0787**	EN	TD	AK1,AK2,ASR06B	DCU
brown.TD.auto	**0.0785**	EN	TD	AK1,AK2,ASR06B	BLLIP
UCkwENTD	**0.0571**	EN	TD	AK1,AK2,ASR06B	UC
UCbaseENTD1	**0.0512**	EN	TD	ASR06B	UC
UvA_2_en4g	**0.0444**	EN	TD	AK2,ASR06B	UVA
UvA_1_base	0.0430	EN	TD	ASR06B	UVA

Table 2. Evaluation results for automatic English monolingual runs. Bold runs are the required condition. AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR03 = ASRTEXT2003A, ASR04 = ASRTEXT2004A, ASR06A = ASRTEXT2006A, and ASR06B = ASRTEXT2006B.

Run ID	MAP	Lng	Query	Document Fields	Site
dcuFrTDauto	0.0636	FR	TD	AK1,AK2,ASR06B	DCU
uoEsTDtF2	0.0619	ES	TD	AK1,AK2,ASR04	UO
uoFrTDtF2	0.0603	FR	TD	AK1,AK2,ASR04	UO
UCkwFRTD1	0.0406	FR	TD	AK1,AK2,ASR06B	UC
UvA_3_nl4g	0.0400	NL	TD	AK2,ASR06B	UVA
UCbaseFRTD1	0.0322	FR	TD	ASR06B	UC

Table 3. Evaluation results for automatic cross-language runs. AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR04 = ASRTEXT2004A, and ASR06B = ASRTEXT2006B.

Automatic-Only Monolingual Runs Teams were required to run at least one monolingual condition using the title (T) and description (D) fields of the topics and indexing only automatically generated fields; the best of these "required runs" for each team are shown in bold in Tables 1 and 2 to facilitate comparison of results between different teams. The University of Ottawa (0.0855), Dublin City University (0.0787), and the BLLIP team (0.0785) reported comparable results (no significant difference at the 95% confidence level). These results are statistically significant better than those reported by the next two teams, the University of Chicago (0.0571) and the University of Amsterdam (0.0444), which were statistically indistinguishable from each other.

Automatic-Only Cross-Language Runs As shown in Table 3, the best result (0.0636) for cross-language runs on automatically generated indexing data (a French–English run from Dublin City University) achieved 81% of the mono-

lingual retrieval effectiveness with comparable conditions (0.0787 as shown in Table 2).

Run ID	MAP	Lng	Query	Document Fields	Site
dcuEnTDNmanual	0.2847	EN	TDN	MK,SUM	DCU
uoEnTDtManF1	0.2761	EN	TD	MK,SUM	UO
brown.TDN.man	0.2577	EN	TDN	MK,SUM	BLLIP
dcuEnTDmanualauto	0.2459	EN	TD	MK,SUM,ASR06B	DCU
brown.TD.man	0.2366	EN	TD	MK,SUM	BLLIP
brown.T.man	0.2348	EN	T	MK,SUM	BLLIP
UvA_4_enopt	0.2088	EN	TD	MK,SUM,ASR06B	UVA
SinaiEn100	0.0597	EN	TD	ALL	SINAI
SinaiEn050	0.0515	EN	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B	SINAI

Table 4. Evaluation results for monolingual English runs with manual metadata. MK = MANUALKEYWORD, SUM = SUMMARY, AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR04 = ASRTEXT2004A, ASR06A = ASRTEXT2006A, ASR06B = ASRTEXT2006B, and ALL = all fields.

Run ID	MAP	Lng	Query	Document Fields	Site
dcuFrTDmanualauto	0.1980	FR	TD	MK,SUM,ASR06B	DCU
UvA_5_nlopt	0.1408	NL	TD	MK,SUM,AK2,ASR06B	UVA
SinaiSp100	0.0737	ES	TD	ALL	SINAI
SinaiFr100	0.0597	FR	TD	ALL	SINAI
SinaiSp050	0.0579	ES	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B	SINAI

Table 5. Evaluation results for cross-language runs with manual metadata. MK = MANUALKEYWORD, SUM = SUMMARY, AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR04 = ASRTEXT2004A, ASR06A = ASRTEXT2006A, ASR06B = ASRTEXT2006B, and ALL = all fields.

Monolingual Runs With Manual Metadata For monolingual TD runs on manually generated indexing data, the University of Ottawa achieved the best result (0.2761), which is statistically significantly better than all other runs under comparable conditions, as shown in Table 4. For TDN runs, the DCU result (0.2847) it not statistically significantly better than that obtained by BLLIP (0.2577).

Cross-Language Runs With Manual Metadata The evaluation results for cross-language runs on manually generated indexing data are shown in Table 5. The best cross-language result (0.1980), representing 81% of monolingual retrieval effectiveness under comparable conditions (0.2459 shown in Table 4), was achieved by DCU's French-English run.

3 Czech Task

The structure of the Czech task was quite similar to the one used in the 2006, with differences which we describe in the following subsections. Further details can be found in the 2006 track overview paper [1].

3.1 Interviews

A "quickstart" collection was generated from the same set of 357 Czech interviews as in 2006. It contained 11,377 overlapping passages with the following fields:

DOCNO containing a unique document number in the same format as the start times that systems were required to produce in a ranked list.

INTERVIEWDATA containing 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.

ASRSYSTEM specifying the type of the ASR transcript, where "2004" and "2006" denote colloquial and formal Czech transcripts respectively.

CHANNEL specifying which recorded channel (left or right) was used to produce the transcript.

ASRTEXT containing words in order from the transcript selected by ASRSYSTEM and CHANNEL for a passage beginning at the start time indicated in DOCNO.

The average passage duration in the default 2007 quickstart collection is 3.75 minutes, and each passage has a 33% overlap with the subsequent passage (i.e., passages begin about every 2.5 minutes).

No thesaurus terms (neither manual nor automatic, neither English nor Czech) were distributed with the collection this year because it was not practical to correct the time misalignment that was present in the 2006 quickstart collection for the manually assigned thesaurus terms (and because the available automatically assigned thesaurus terms had not proven to be useful in 2006).

3.2 Topics

A total of 29 training topics and 42 evaluation topics were selected as follows. Participating teams were asked to submit results for a total of 118 topics: 105 topics from 2006 that had originally been created for the English collection,

10 topics from 2006 that were variants of 10 of the English topics that were "broadened" in a way that we expected to result in more matches in the Czech collection, and 3 new broadened topics that were constructed this year. For example, topic 1187 (Title: "IG Farben Labor Camps") was broadened to create topic 4003 (Title: "Labor Camps"). All of these topics were originally created in English and then translated into Czech by native speakers.[6] Some minor errors in the Czech translations from last year were corrected.[7] No teams used the English topics this year; all official runs with the Czech collection were monolingual.

Two of the 118 topics were used for assessor training and excluded from the evaluation, 29 topics were available for training systems (with relevance judgments from 2006), and 50 of the remaining 87 topics were initially selected as possible evaluation topics. This set of 50 includes all available topics that were not used for assessor or system training for which at least 6 relevant passages were identified during the search-guided assessment phase. This cutoff at six segments was selected to balance quantization noise in the evaluation measure with the risk of sampling error that would result from too few topics. An additional "pooled" assessment process was conducted after submission of results by participating teams to judge highly-ranked passages for which judgments had not been recorded during search-guided assessment. This pooled assessment process was completed for 42 of the 50 topics in the available time, so those 42 were chosen as the evaluation topics for the 2007 Czech task.

3.3 Evaluation Measure

The evaluation measure used for the Czech task was the same as in 2006: mean Generalized Average Precision (mGAP). This measure was originally designed to accommodate human assessments of partial relevance [3]. In our case, the human assessments are binary but the degree of match to those assessments can be partial. An exact match between the system-specified start time and the closest assessor-assigned start time yielded full credit for the match, with a linear decay to zero credit for system start time errors of plus or minus 90 seconds from the nearest assessor-assigned start time.[8] The Wilcoxon signed-rank signed test was employed for evaluation of significance.

3.4 Relevance Judgments

Relevance judgments were completed at Charles University in Prague for the 42 evaluation topics this year under the same conditions as in 2006 by the same six

[6] Dutch, French, German and Spanish versions are also available for the topics that were designed originally for the English task, but the 13 broadened topics have not been translated into those languages.

[7] The corrected topics are 1259, 1282, 1551, 14313, and 24313. Of these, only topic 14313 had been selected as an evaluation topic in the 2006 Czech task. None of these have been used as evaluation topics in any year of the English task.

[8] The window size was incorrectly reported as plus or minute 150 seconds in the 2006 CL-SR track overview paper, but a 90-second window was actually used in both 2006 and 2007.

relevance assessors. A total of 2,389 start (and end) times for relevant passages were identified, thus yielding an average of 56 relevant passages per topic (minimum 6, maximum 199). Table 6 shows the number of relevant start times for each of the 42 evaluation topics. A total of 34 of these 42 topics are also present in the CLEF CL-SR English task collection (as training, evaluation, or unused topics; the exceptions are 8 broadened topics, which are the 4000-series).

Topic	# rel	Topic	# rel	Topic	# rel	Topic	# rel
1192	18	2265	113	3019	14	4005	68
1345	12	2358	126	3021	16	4006	135
1554	46	2384	37	3022	29	4007	51
1829	6	2404	8	3023	78	4009	10
1897	31	3000	41	3024	105	4011	132
1979	17	3001	102	3026	33	4012	61
2000	114	3002	95	3027	86	14313	17
2006	63	3007	107	3028	199	15601	108
2012	90	3008	53	3032	9	15602	25
2185	25	3010	18	4001	35		
2224	63	3016	40	4004	13		

Table 6. Number of relevant passages identified for each of the evaluation topics.

3.5 Techniques

All participating teams employed existing information retrieval systems to perform monolingual retrieval and submitted total of 15 runs for official scoring. To facilitate cross-team comparisons, each participating team submitted at least one run with the quickstart collection and with queries that were automatically created from the title and description topic fields. The narrative topic field was used only by University of West Bohemia. Most teams used only automatically generated queries; manual query construction was performed only by Charles University. The University of West Bohemia also used the quickstart scripts with different parameters to generate another collection for some experiments.

Brown University (BLLIP) The Brown University system was based on a language model paradigm and implemented using Indri. A unigram language model, Czech-specific stemming, and pseudo-relevance feedback were applied in three officially submitted runs.

Charles University (CUNI) The Charles University team performed experiments with Indri using blind relevance feedback, stopword removal, and lemmatization obtained using a morphological analysis system that also performed

Run name	mGAP score	Query construction	Topic fields	Term normalization	Site name
UWB_2-1_tdn_l	0.0274	Auto	TDN	lemma	UWB
UWB_3-1_tdn_l	0.0241	Auto	TDN	lemma	UWB
UWB_2-1_td_s	0.0229	Auto	TD	stem	UWB
UCcsaTD2	0.0213	Auto	TD	aggressive stem	UC
UCcslTD1	0.0196	Auto	TD	light stem	UC
prague04	0.0195	Auto	TD	lemma	CUNI
prague01	0.0192	Auto	TD	lemma	CUNI
prague02	0.0183	Manual	TD	lemma	CUNI
UWB_3-1_td_l	0.0134	Auto	TD	lemma	UWB
UWB_2-1_td_w	0.0132	Auto	TD	none	UWB
UCunstTD3	0.0126	Auto	TD	none	UC
brown.s.f	0.0113	Auto	TD	light stem	BLLIP
brown.sA.f	0.0106	Auto	TD	aggressive stem	BLLIP
prague03	0.0098	Manual	TD	none	CUNI
brown.f	0.0049	Auto	TD	none	BLLIP

Table 7. Corrected scores for Czech official runs (Query language: CZ, Document fields: ASR2006, 90-second window).

part-of-speech tagging. The team submitted four official runs; two of which employed manual query construction.

University of Chicago (UC) The University of Chicago employed the InQuery information retrieval system with stopword removal and three different stemming approaches: no stemming, light stemming, and aggressive stemming. Three runs were submitted for official scoring.

University of West Bohemia (UWB) The University of West Bohemia employed a TF*IDF model implemented in Lemur with blind relevance feedback. Five runs were submitted for official scoring which differed in methods used for word normalization (none, lemmatization, stemming), in formulas used for term weighting (Raw TF, BM25), and in the topic fields used (TDN, TD).

Results A computation error was discovered in the mGAP scoring script that was corrected after the CLEF-2007 meeting. Corrected results for all official runs (evaluated on 42 topics) are reported in Table 7, with bold indicating the highest-scoring run by each team with standard conditions (TD queries, standard quickstart collection),[9] and the Charles University and University of West Bohemia papers in this volume report corrected mGAP scores as well. The effect of

[9] Corrected scores generally improved slightly, and the only reversal in system preference order was between two systems separated by 0.0001 in both the original and the corrected scores.

Run name	mGAP score	mGAP increase	Query construction	Term normalization	Site name
UWB_2-1_td_s	0.0229	+73%	Auto	stem	UWB
UWB_2-1_td_w	0.0132		Auto	none	UWB
UCcsaTD2	0.0213	+69%	Auto	aggressive stem	UC
UCunstTD3	0.0126		Auto	none	UC
prague02	0.0183	+87%	Manual	lemma	CUNI
prague03	0.0098		Manual	none	CUNI
brown.s.f	0.0113	+131%	Auto	light stem	BLLIP
brown.f	0.0049		Auto	none	BLLIP

Table 8. Comparison of systems with and without term normalization (Topic fields: TD, corrected results).

term normalization handling the rich Czech morphology is quite significant. The runs employing any type of term normalization (stemming or lemmatization) outperform systems indexing only original word forms with no normalization by 69–131%. The scores of directly comparable runs are given in Table 8, all the differences are statistically significant at a 95% confidence level.

Three quantization factors are present in the Czech evaluation: (1) the 15-second resolution of assessor-assigned start times; (2) the 90-second window size for mGAP computation, and (3) the 150-second spacing between passage start times in the standard quickstart collection. The 150-second passage start time spacing is clearly somewhat problematic when coupled with a 90-second evaluation window size. The University of West Bohemia demonstrated the effect by reducing the passage start time spacing to 75 seconds (the UWB_2-1 runs, in which the average passage duration was also reduced to 2.5 minutes). This yielded an apparent 14% increase in mGAP (compare UWB_2-1_tdn_l: mGAP=0.0274 and UWB_3-1_tdn_l: mGAP=0.0241) that turned out not to be statistically significant (perhaps because of quantization noise).

Although we compute evaluation results only from start times, our assessors marked both start and end times. The average duration of assessor-marked relevant passages is 2.83 minutes, which seems to be somewhat better matched to the 2.5 minutes passages used in the University of West Bohemia's alternate condition (2.5 minutes for UWB_2-1_tdn_l vs. 3.75 minutes for UWB_3-1_tdn_l and all runs from other sites).

The Charles University team reported on the first experiments with interactive use of the Czech collection. Their best run based on manual query construction (prague02) turned out to be statistically indistinguishable from a run under comparable conditions from the same team with queries that were generated automatically (prague04).

4 Conclusion and Future Plans

Like all CLEF tracks, the CL-SR track had three key goals: (1) to develop evaluation methods and reusable evaluation resources for an important information access problem in which cross-language access is a natural part of the task, (2) to generate results that can provide a strong baseline against which future research results with the same evaluation resources can be compared, and (3) to foster the development of a research community with the experience and expertise to make those future advances. In the case of the CL-SR track, those goals have now been achieved. Over three years, research teams from 14 universities in 6 countries submitted 123 runs for official scoring, and many additional locally scored runs have been reported in papers published by those research teams. The resulting English and Czech collections are the first standard information retrieval test collections for spontaneous conversational speech, unique characteristics of the English collection have fostered new research comparing searches based on automatic speech recognition and manually assigned metadata, and unique characteristics of the Czech collection have inspired new research on evaluation of information retrieval from unsegmented speech.

Now that the CL-SR track has been completed, these new CLEF test collections will be made available to nonparticipants through the Evaluations and Language Resources Distribution Agency (ELDA). The training data for the automatic speech retrieval systems that were used to generate the transcripts in those collections is also expected to become available soon, most likely through the Linguistic Data Consortium (LDC). It is our hope that these resources will be used together to investigate more closely coupled techniques than have been possible to date with just the present CLEF CL-SR test collections. Looking further forward, we believe that it is now time for the information retrieval research community to look beyond oral history to other instances of spontaneous conversational speech such as recordings of meetings, historically significant telephone conversations, and broadcast conversations (e.g., radio "talk shows"). We also believe that it would be productive to begin to explore application of some of the technology developed for this track to improve access to a broad range of oral history collections and similar cultural heritage materials (e.g., interviews contained in broadcast archives). Together, these directions for future work will likely continue to extend the legacy and impact of this initial investment in exploring the retrieval of information from spontaneous conversational speech.

Acknowledgments

This year's 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 Prague without whom the Czech collection simply would not exist, to Scott Olsson for helping to prepare the English collection this year, to Ayelet Goldin and Jianqiang Wang for their timely help with critical details of the Czech relevance assessment and scoring process, to Pavel Ceske for creating the new

Czech scoring script, to Jan Hajic for his support and advice throughout, and to Carol Peters for her seemingly endless patience. This work has been supported in part by NSF IIS award 0122466 (MALACH), by the Ministry of Education of the Czech Republic, projects MSM 0021620838 and #1P05ME786, and by the European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD—project MultiMATCH contract IST-033104. The authors are solely responsible for the content of this paper.

References

1. Oard, D.W., Wang, J., Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 cross-language speech retrieval track. In: Evaluation of Multilingual and Multi-modal Information Retrieval, Revised Selected Papers, CLEF-2006, Springer-Verlag, LNCS 4730. (2006)
2. 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: Multilingual Information Repositories, Revised Selected Papers, CLEF-2005, Springer-Verlag, LNCS 4022. (2005)
3. Kekalainen, J., Jarvelin, K.: Using graded relevance assessments in IR evaluation. In: Journal of the American Society for Information Science and Technology. (2002)