Overview of the CLEF-2005 Cross-Language Speech Retrieval Track

Ryen W. White[1], Douglas W. Oard[1,2], Gareth J. F. Jones[3], Dagobert Soergel[2] and Xiaoli Huang[2]
[1]Institute for Advanced Computer Studies
[2]College of Information Studies
University of Maryland, College Park MD 20742, USA
[3]School of Computing, Dublin City University, Dublin 9, Ireland
{ryen,oard,dsoergel,xiaoli}@umd.edu
Gareth.Jones@computing.dcu.ie

Abstract

The task for the CLEF-2005 cross-language speech retrieval track was to identify topically coherent segments of English interviews in a known-boundary condition. Seven teams participated, performing both monolingual and cross-language searches of ASR transcripts, automatically generated metadata, and manually generated metadata. Results indicate that monolingual search technology is sufficiently accurate to be useful for some purposes (the best mean average precision was 0.18) and cross-language searching yielded results typical of those seen in other applications (with the best systems approximating monolingual mean average precision).

1. Introduction

The 2005 Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track follows two years of experimentation with cross-language retrieval of broadcast news in the CLEF-2003 and CLEF-2004 Spoken Document Retrieval (SDR) tracks [2]. CL-SR is distinguished from CL-SDR by the lack of clear topic boundaries in conversational speech. Moreover, spontaneous speech is considerably more challenging for the Large-Vocabulary Continuous Speech Recognition (referred to here generically as Automatic Speech Recognition, or ASR) techniques on which fully-automatic content-based search systems are based. Recent advances in ASR have made it possible to contemplate the design of systems that would provide a useful degree of support for searching large collections of spontaneous conversational speech, but no representative test collection that could be used to support the development of such systems was widely available for research use. The principal goal of the CLEF-2005 CL-SR track was to create such a test collection. Additional goals included benchmarking the present state of the art for ranked retrieval of spontaneous conversational speech and fostering interaction among a community of researchers with interest in that challenge.

Three factors came together to make the CLEF 2005 CL-SR track possible. First, substantial investments in research on ASR for spontaneous conversational speech have yielded systems that are able to transcribe near-field speech (e.g., telephone calls) with Word Error Rates (WER) below 20% and far-field speech (e.g., meetings) with WER near 30%. This is roughly the same WER range that was found to adequately support ranked retrieval in the original Text Retrieval Conference (TREC) SDR track evaluations [3]. Second, the Survivors of the Shoah Visual History Foundation (VHF) collected, digitized, and annotated a very large collection (116,000 hours) of interviews with Holocaust survivors, witnesses and rescuers. In particular, one 10,000-hour subset of that collection was extensively annotated in a way that allowed us to affordably decouple relevance judgment from the limitations of current speech technology. Third, a project funded by the U.S. National Science Foundation focused on Multilingual Access to Large Spoken Archives (MALACH) is producing LVSCR systems for this collection to foster research on access to spontaneous conversational speech, and automatic transcriptions from two such systems are now available [1].

Designing a CLEF track requires that we balance the effort required to participate with the potential benefits to the participants. For this first year of the track, we sought to minimize the effort required to participate, and within that constraint to maximize the potential benefit. The principal consequence of that decision was adoption of a known-boundary condition in which systems performed ranked retrieval on topically coherent segments. This yielded a test collection with the same structure that is used for CLEF ad hoc tasks, thus facilitating application of existing ranked retrieval technology to this new task. Participants in new tracks often face a chicken-and-egg dilemma, with good retrieval results needed from all participants before an a test collection can be affordably created using pooled relevance assessment techniques, but the exploration of the design space that is needed to produce good results requires that a test collection already exist. For the CLEF-2005 CL-SR track we were able to address this challenge by distributing training topics with relevance judgments that had been developed using a search-guided relevance assessment process [5]. We leveraged the availability of those training topics by distributing an extensive set of manually and automatically created metadata that participants could use as a basis for constructing contrastive conditions. In order to promote cross-site comparisons, we asked each participating team to submit one "required run" in which the same topic language and topic fields and only automatically generated transcriptions and/or metadata were used.

The remainder of this overview paper is structured as follows. In Section 2 we describe the CL-SR test collection. Section 3 identifies the sites that participated and briefly describes the techniques that they tried. Section 4 looks across the runs that were submitted to identify conclusions that can be drawn from those results. Section 5 concludes the paper with a brief description of future plans for the CLEF CL-SR track.

2. Collection

The CLEF-2005 CL-SR test collection was released in two stages. In Release One (February 15 2005), the "documents," training topics and associated relevance judgments, and scripts were made available to participants to support system development. Release Two (April 15 2005) included the 25 evaluation topics on which sites' runs would be evaluated, one additional script that could be used to perform thesaurus expansion, and some metadata fields that had been absent from Release One. This section describes the genesis of the test collection.

2.1 Documents

The fundamental goal of a ranked retrieval system is to sort a set of "documents" in decreasing order of expected utility. Commonly used evaluation frameworks rely on an implicit assumption that ground-truth document boundaries exist.[1] The nature of oral history interviews challenges this assumption, however. The average VHF interview extends for more than 2 hours, and spoken content that extensive can not presently be easily skimmed. Many users, therefore, will need systems that retrieve passages rather than entire interviews.[2] Remarkably, the VHF collection contains a 10,000 hour subset for which manual segmentation into topically coherent segments was carefully performed by subject matter experts. We therefore chose to use those segments as the "documents" for the CLEF-2005 CL-SR evaluation.

Development of Automatic Speech Recognition (ASR) systems is an iterative process in which evaluation results from initial system designs are used to guide the development of refined systems. In order to limit the computational overhead of this process, we chose to work initially with roughly 10% of the interviews for which manual topic segmentation is available. We chose 403 interviews (totaling roughly 1,000 hours of English speech) for this purpose. Of those 403, portions of 272 interviews had been digitized and processed by two ASR systems at the time that the CLEF-2005 CL-SR test collection was released. A total of 183 of those are complete interviews; for the other 89 interviews ASR results are available for at least one, but not all, of the 30-minute tapes on which the interviews were originally recorded. In some segments, near the end of an interview, physical objects (e.g., photographs) are shown and described. Those segments are not well suited for ASR-based search because the few words are typically spoken by the interviewee (usually less then 15) and because we chose not to distribute the visual referent as a part of the test collection. Such segments were unambiguously marked by human indexers, and we automatically removed them from the test collection. The resulting test collection contains 8,104 segments from 272 interviews totaling 589 hours of speech. That works out to an average of about 4 minutes (503 words) of recognized speech per segment. A collection of this size is very small from the perspective of modern IR experiments using written sources (e.g., newswire or Web pages), but it is comparable in size to the 550-hour collection of broadcast news used in the CLEF-2004 SDR evaluation.

As Figure 1 shows, each segment was uniquely identified by a `DOCNO` field in which the `IntCode` uniquely identifies an interview within the collection, `SegId` uniquely identifies a segment within the collection, and `SequenceNum` is the sequential order of a segment within an interview. For example, VHF00009-056149.001 is the first segment in interview number 9.

[1] Note that we do not require that document boundaries be known to the system under test, only that they exist. The TREC HARD track passage retrieval task and the TREC SDR unknown boundaries condition are examples of cases in which the ground truth boundaries are not known to the system under test. Even in those cases ground-truth boundaries must be known to the evaluation software.

[2] Initial studies with 17 teachers and 6 scholars indicated that all teachers and about half the scholars needed segment-based access for the tasks in which they were engaged.

The following fields were created by VHF subject matter experts while viewing the interview. They are included in the test collection to support contrastive studies in which results from manual and automated indexing are compared:

- The INTERVIEWDATA field contains all names by which the interviewee was known (e.g., present name, maiden name, and nicknames) and the date of birth of the interviewee. The contents of this field are identical for every segment from the same interview (i.e., for every DOCNO that contains the same IntCode). This data was obtained from handwritten questionnaires that were completed before the interview (known as the Pre-Interview Questionnaire or PIQ).
- The NAME field contains the names of other persons that were mentioned in the segment. The written form of a name was standardized within an interview (a process known as "name authority control"), but not across interviews.
- The MANUALKEYWORDS field contains thesaurus descriptors that were manually assigned from a large thesaurus that was constructed by VHF. Two types of keywords are present, but not distinguished: (1) keywords that express a subject or concept; and (2) keywords that express a location, often combined with time in one pre-coordinated keyword. On average about 5 manually thesaurus descriptors were manually assigned to each segment, at least one of which was typically a pre-coordinated location-time pair (usually with one-year granularity)
- The SUMMARY field contains a three-sentence summary in which a subject matter expert used free text in a structured style to address the following questions: who? what? when? where?

The following fields were generated fully automatically by systems that did not have access to the ground truth data for any interview in the test collection. These fields could therefore be used to explore the potential of different techniques for automated processing:

- Two ASRTEXT fields contain words produced by an ASR system. The speech was automatically transcribed by ASR systems developed at the IBM T. J. Watson Research Center. The manual segmentation process at VHF was conducted using time-coded videotape without display of the acoustic envelope. The resulting segment boundaries therefore sometimes occur in the middle of a word in the one-best ASR transcript. We therefore automatically adjusted the segment boundaries to the nearest significant silence (a silence with a duration of 2 seconds or longer) if such a silence began within XX seconds of the assigned boundary time; otherwise we adjusted the segment boundary to the nearest word boundary. The words from the one-best ASR transcript were then used to create an ASR field for the resulting segments. This process was repeated for two ASR systems. The ASRTEXT2004A field of the document representation shown in Figure 1 contains an automatically created transcript using the best available ASR system, for which an overall mean WER of 38% and a mean named entity error rate of 32% was computed over portions of 15 held-out interviews. The recognizer vocabulary for this system was primed on an interview-specific basis with person names, locations, organization names and country names mentioned in an extensive pre-interview questionnaire. The ASRTEXT2003A field contains an automatically created transcript using an earlier system for which a mean WER of 40% and a mean named entity error rate of 66% was computed using the same held-out data.
- Two AUTOKEYWORD fields contain thesaurus descriptors that were automatically assigned by using text classification techniques. The AUTOKEYWORD2004A1 field contains a set of thesaurus keywords that were assigned automatically using a k-Nearest Neighbor (kNN) classifier using only words from the ASRTEXT2004A field of the segment; the top 20 keywords are included. The classifier was trained using data (manually assigned thesaurus keywords and manually written segment summaries) from segments that are not contained in the CL-SR test collection. The AUTOKEYWORD2004A2 field contains a set of thesaurus keywords that were assigned in a manner similar to those in the AUTOKEYWORD2004A1, but using a different kNN classifier that was trained (fairly) on different data; the top 16 concept keywords and the top 4 location-time pairs were included for each segment.

```
<DOC>
<DOCNO>VHF[IntCode]-[SegId].[SequenceNum]</DOCNO>
<INTERVIEWDATA>Interviewee name(s) and birthdate</INTERVIEWDATA>
<NAME>Full name of every person mentioned</NAME>
<MANUALKEYWORD>Thesaurus keywords assigned to the segment</MANUALKEYWORD>
<SUMMARY>3-sentence segment summary</SUMMARY>
<ASRTEXT2003A>ASR transcript produced in 2003</ASRTEXT2003A>
<ASRTEXT2004A>ASR transcript produced in 2004</ASRTEXT2004A>
<AUTOKEYWORD2004A1>Thesaurus keywords from a kNN classifier</AUTOKEYWORD2004A1>
<AUTOKEYWORD2004A2>Thesaurus keywords from a second kNN classifier</AUTOKEYWORD2004A2>
</DOC>
```

Figure 1. Document structure in CL-SR test collection.

The three KEYWORD fields in the test collection included only the VHF-assigned "preferred term" for each thesaurus descriptor. A script was provided with the final release of the test collection that could be used to expand the descriptors for each segment using synonymy, part-whole, and is-a thesaurus relationships. That capability could be used with automatically assigned descriptors or (for contrastive runs) with the manually assigned descriptors.

2.2 Topics

The VHF collection has attracted significant interest from scholars, educators, documentary film makers, and others, resulting in 280 topic-oriented written requests for materials from the collection. From that set, we selected 75 requests that we felt were representative of the types of requests and the types of subjects contained in the topic-oriented requests. The requests were typically made in the form of business letters, often accompanied by a filled-in request form describing the requester's project and purpose. Additional materials (e.g., a thesis proposal) were also sometimes available. TREC-style topic descriptions consisting of title, a short description and a narrative were created for the 75 topics, as shown by the example in Figure 2.

```
<top>
<num> 1148
<title> Jewish resistance in Europe
<desc> Provide testimonies or describe actions of Jewish resistance in Europe before and
during the war.
<narr> The relevant material should describe actions of only- or mostly Jewish resistance in
Europe. Both individual and group-based actions are relevant. Type of actions may include
survival (fleeing, hiding, saving children), testifying (alerting the outside world, writing,
hiding testimonies), fighting (partisans, uprising, political security) Information about
undifferentiated resistance groups is not relevant.
</top>
```

Figure 2. Example topic.

Only topics for which relevant segments exist can be used as a basis for comparing the effectiveness of ranked retrieval systems, so we sought to ensure the presence of an adequate number of relevant segments for each test topic. For the first 50 topics, we iterated between topic selection and interview selection in order to arrive at a set of topics and interviews for which the number of relevant segments was likely to be sufficient to yield reasonably stable estimates of mean average precision (we chose 30 relevant segments as our target, but allowed considerable variation). At that point we could have selected any 10% of the available fully indexed interviews for the test collection, so the process was more constrained by topic selection than by interview selection. In some cases, this required that we broaden specific requests to reflect our understanding of a more general class of information need for which the request we examined would be a specific case. This process excluded most queries that included personal names or very specific and infrequently used geographical areas. The remaining 25 topics were selected after the interview set was frozen, so in that case topic selection and broadening were the only free variables. All of the training topics are drawn from the first 50; most of the evaluation topics are from the last 25. A total of 12 topics were excluded, 6 because the number of relevant documents turned out to be too small to permit stable estimates of mean average precision (fewer than 5) or so large (over 50% of the total number of judgments) that the exhaustiveness of the search-guided assessment process was open to question. The remaining 6 topics were excluded because relevance judgments were not ready in time for release as training topics and they were not needed to complete the set of 25 evaluation topics. The resulting test collection therefore contains 63 topics, with an additional 6 topics for which embargoed relevance judgments are already available for use in the CLEF-2006 evaluation collection. Participants are asked not to perform any analysis involving topics outside the released set of 63 in order to preserve the integrity of the CLEF-2006 test collection.

All topics were originally authored in English and then re-expressed in Czech, French, German and Spanish by native speakers of those languages to support cross-language retrieval experiments. In each case, the translations were checked by a second native speaker before being released. For the French translations, resource constraints precluded translation of the narrative fields. All three fields are available for the other query languages.

2.3 Relevance Assessment

Relevance judgments were made for the full set of 404 interviews, including those segments that were removed from the released collection because they contained only brief descriptions of physical objects. Judging every document for every topic would have required about 750,000 relevance judgments. Even had that been affordable (e.g., by judging each segment for several topics simultaneously), such a process could not be affordably scaled up to larger collections. The usual way this challenge is addressed in CLEF, pooled relevance assessment, involves substantial risk when applied to spoken word collections. With pooled assessment, documents that are not assessed are treated as if they are not relevant when computing effectiveness measures such as mean average precision. When all systems operate on similar feature set (e.g., words), it has been shown that comparable results can be obtained even for systems that did not contribute to the assessment pools. This is enormously consequential, since it allows the cost of creating a test collection to be amortized over anticipated future uses of that collection. Systems based on automatic speech recognition with a relatively high WER violate the condition for reuse, however, since the feature set on which future systems might be based (recognized words) could well be quite different. We therefore chose an alternative technique, search-guided relevance judgment, which has been used to construct reusable test collections for spoken word collections in the Topic Detection and Tracking (TDT) evaluations [8].

Our implementation of search-guided evaluation differs from that used in TDT in that we search manually assigned metadata rather than ASR transcripts. Relevance assessors are able to search all of the metadata distributed with the test collection, plus notes made by the VHF indexers for their own use, summaries of the full interview prepared by the VHF indexer, and a fuller set of PIQ responses. For interviews that had been digitized by the time assessment was done, relevance assessors could also listen to the audio; in other cases, they could indicate whether they felt that listening to the audio might change their judgment so that re-assessment could be done once the audio became available. The relevance assessment system was based on Lucene, which supports fielded searching using both ranked and Boolean retrieval. The set of thesaurus terms assigned to each segment was expanded by adding broader terms from the thesaurus up to the root of the hierarchy. A threshold was applied to the ranked list, and retrieved segments were then re-arranged by interview and within each interview in decreasing score order. The display order was structured to place interviews with many highly ranked segments ahead of those with fewer. Relevance assessors could easily reach preceding or following segments of the same interview; those segments often provide information needed to assess the relevance of the segment under consideration, and they may also be relevant in their own right.

Our relevance assessors were 6 graduate students studying history. The assessors were experienced searchers; they made extensive use of complex structured queries and interactive query reformulation. They conducted extensive research on assigned topics using external resources before and during assessment, and kept extensive notes on their interpretation of the topics, topic-specific guidelines for deciding on the level of relevance for each relevance type, and other issues (e.g., rationale for judging specific segments). Relevance assessors did thorough searches to find as many relevant segments as possible and assessed the segments they found for each topic. We employed two processes to minimize the chance of unintentional errors during relevance assessment:

- Dual-assessment: For some training topics, segments were judged independently by two assessors with subsequent adjudication; this process resulted in two sets of independent relevance judgments that can be used to compute inter-annotator agreement plus the one set of adjudicated judgments that were released.
- Review: For the remaining training topics and all evaluation topics, an initial judgment was done by one assessor and then their results were reviewed, and if necessary revised, by a second assessor. This process resulted in one set of adjudicated relevance judgments that were released.

As a result of the above processes, for every topic-segment pair, we have two sets of relevance assessments derived from two assessors, either independent or not. This allowed us to later measure the inter-assessor agreement and thus to gain insight into the reliability of relevance assessments on selected topics.

The search-guided assessments are complemented by pooled assessments using the top 100 segments from 14 runs. Participants were requested to prioritize their runs in such a way that selecting the runs assigned the highest priority would result in the most diverse judgment pools. We selected the top two prioritized runs from each site to create the pools. Assessors judged all segments in these pools that had not already been judged as part of the search-guided assessment process. For this process, most topics had just one assessor and no review. A total of 58,152 relevance judgments were created over 3 summers for the 403 interviews and 75 topics, of which 48,881 are specific to the topics and segments in the CLEF-2005 CL-SR test collection.

Relevance is a multifaceted concept; interview segments may be relevant (in the sense that they help the searcher perform the task from which the query arose) for different reasons. We therefore defined five types of topical relevance, both to guide the thinking of our assessors and to obtain differentiated judgments that could serve as a basis for more detailed analysis than would be possible using binary single-facet judgments. The relevance types that we chose were based on the notion of evidence (rather than, for example, potential emotional impact or appropriateness to an audience). The initial inventory of five relevance types was based on our understanding of historical methods and information seeking processes. The types were then refined during a two-week pilot study through group discussions with our assessors. The resulting types are:

- Provides **direct** evidence
- Provides **indirect**/circumstantial evidence
- Provides **context**
- Useful as a basis for **comparison**
- Provides **pointer** to a source of information

The first two of these match the traditional definition of topical relevance in CLEF; the last three would normally be treated as not relevant in the sense that term is used at CLEF. Each type of relevance was judged on a five-point scale (0=none to 4=high). Assessors were also asked to assess **overall** relevance, defined as the degree of to which they felt that a segment would prove to be useful to the search that had originally posed the topic. Assessors were instructed to consider two factors in all assessments: (1) the nature of the information (i.e., level of detail and uniqueness), and (2) the nature of the report (i.e., first-hand vs. second-hand accounts vs. rumor). For example, the definition of direct relevance is: "Directly on topic ... describes the events or circumstances asked for or otherwise speaks directly to what the user is looking for. First-hand accounts are preferred ... second-hand accounts (hearsay) are acceptable." For indirect relevance, the assessors also considered the strength of the inferential connection between the segment and the phenomenon of interest. The average length of a segment is about 4 minutes, so the brevity of a mention is an additional factor that could affect the performance of search systems. We therefore asked assessors to estimate the fraction of the segment that was associated with each of the five categories.[3] Assessors were instructed to treat brevity and degree separately (a very brief mention could be highly relevant). For more detail on the types of relevance see [4].

To create binary relevance judgments, we elected to treat the union of the direct and indirect judgments with scores of 2, 3, or 4 as topically relevant, regardless of the duration of the mention within the segment.[4] A script was provided with the test collection that allowed sites to generate alternative sets of binary relevance scores as an aid to analysis of results (e.g., some systems may do well when scored with direct topical relevance but poorly when scored with indirect topical relevance).

The resulting test collection contained 63 topics (38 training, 25 evaluation topics), 8,104 segments, and 48,881 6-aspect sets of complex relevance judgments, distributed as shown in Table 1. Although the training and evaluation topic sets were disjoint, the set of segments being searched was the same.

[3] Assessments of the fraction of the segments that were judged as relevant are available, but that were not released with the CLEF-2005 CL-SR test collection because the binarization script has not yet been extended to use that information.

[4] We elected not to use the overall relevance judgments in this computation because our definition of overall relevance allowed consideration of context, comparison and pointer evidence in arriving at a judgment of overall relevance.

Table 1. Distribution of judgments across training topics and evaluation topics.

Topic set	Training	Evaluation
Total number of topics	38	25
Total judgment sets	30,743	18,138
Median judgment sets per topic	787	683
Total segments w/binary relevance true	3,105	1,846
Median relevant judgments per topic	51.5	53

Figure 3 shows the distribution of relevant and non-relevant segments for the training and evaluation topics. Topics are arranged in descending order of proportion relevant (i.e., binary relevance true) vs. judged for that topic.

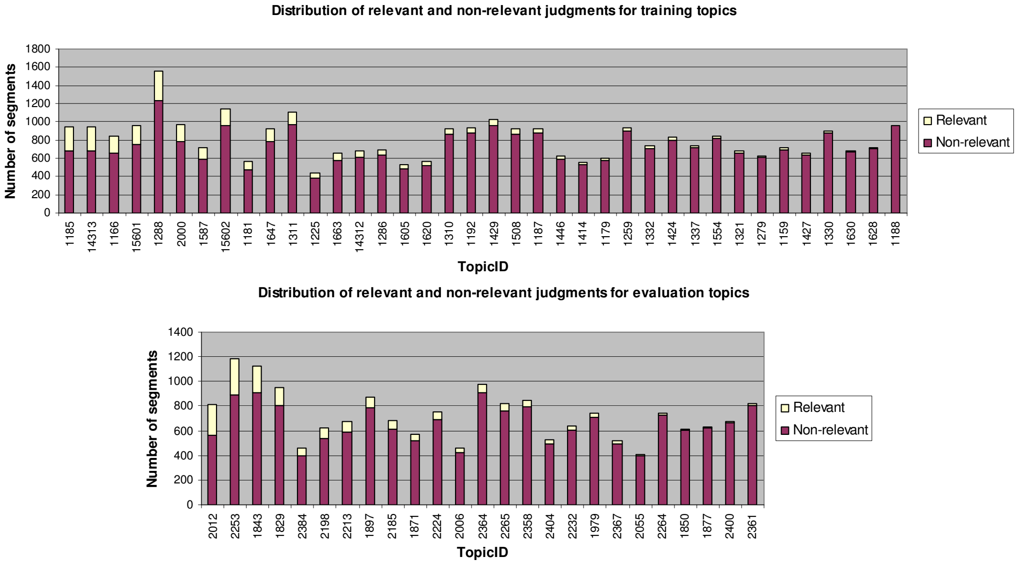

Figure 3. Distribution of relevant (binary relevance true) and non-relevant segments.

To determine the extent of individual differences, we evaluated inter-assessor agreement using two sets of independent judgments for the 28 training topics that were dual assessed. Cohen's Kappa was computed on search-guided binary relevance judgments. The average Kappa score is 0.487, with a standard deviation of 0.188, indicating moderate agreement. The distribution of Kappa scores across different levels of agreement is shown in Table 2.

Table 2. Distribution of agreement over 28 training topics.

Kappa range	**Slight** **(0.01 – 0.20)**	**Fair** **(0.21 – 0.40)**	**Moderate** **(0.41 – 0.60)**	**Substantial** **(0.61 – 0.80)**	**Almost perfect** **(0.81 – 1.00)**
Topics	4	3	12	8	1

3. Experiments

In this section, we describe the run submission procedure and the sites that participated. We accepted a maximum of 5 runs from each site for "official" (i.e., blind) scoring; sites could also score additional runs locally to further explore contrastive conditions. To facilitate comparisons across sites, we asked each site to submit one "required" run using automatically constructed queries from the English title and description fields of the topics (i.e., an automatic monolingual "TD" run) and an index that was constructed without use of human-created metadata (i.e., indexing derived from some combination of ASRTEXT2003A, ASRTEXT2004A, AUTOKEYWORD2004A1, and AUTOKEYWORD2004A2, including the optional use of synonyms and/or broader terms for one or both of the AUTOKEYWORD fields). The other submitted runs could be created in whatever way best allowed the sites to explore the research questions in which they are interested (e.g., comparing monolingual and cross-language, comparing automatic recognition with metadata, or comparing alternative techniques for exploiting ASR results). In keeping with the goals of CLEF, cross-language searching was encouraged; 40% of submitted runs used queries in a language other than English.

Seven groups submitted runs, and each has provided the following brief description of their experiments; additional details can be found in the working notes paper submitted by each group.

3.1 University of Alicante (ualicante)

The University of Alicante used a passage retrieval system for their experiments in the track this year. Passages in such systems are usually composed of a fixed number of sentences, but the lack of sentence boundaries in the ASR that composed the collection of this track does not allow this feature. To address this issue they used fixed word length overlapping passages and distinct similarity measures (e.g., Okapi) to calculate the weights of the words of the topic according to the document collection. Their experimental system applied heuristics to the representation of the topics in the way of logic forms. The University of Alicante's runs all used English queries and automatic metadata.

3.2 Dublin City University (dcu)

As in Dublin City University's previous participations in CLEF, the basis of their experimental retrieval system was the City University research distribution version of the Okapi probabilistic model. Queries were expanded using pseudo relevance feedback (PRF). Expansion terms were selected from sentence-based summaries of the top 5 most assumed relevant documents. All terms within the chosen sentences were then ranked and the top 20 ranking terms selected as expansion terms. Non-English topics were translated to English using SYSTRAN version 3.0. Runs explored various combinations of the ASR transcription, autokeyword and summary fields.

3.3 University of Maryland (umaryland)

The University of Maryland tried automatic retrieval techniques (including blind relevance feedback) with two types of data: manually created metadata and automatically generated data. Three runs used automatic metadata. Submission of the two runs with manual metadata has two main purposes: to set up the best monolingual upper-bound and to compare CLIR with monolingual IR. All runs used the InQuery search engine (version 3.1p1) from the University of Massachusetts.

3.4 Universidad Nacional de Educación a Distancia (uned)

UNED tested different ways to clean documents in the collection. They erased all duplicate words and joined the characters that forms spelled words like "l i e b b a c h a r d" into the whole word (i.e., "liebbachard"). Using this cleaned collection they tried a monolingual trigrams approach. They also tried to clean the documents, erasing the less informative words using two different approaches: morphological analysis and part of speech tagging. Their runs were monolingual and cross-lingual.

3.5 University of Pittsburgh (upittsburgh)

The University of Pittsburgh explored two ideas: (1) to study the evidence combination techniques for merging retrieval results based on ASR outputs with human generated metadata at the post-retrieval stage, (2) to explore the usage of Self-Organizing Map (SOM) as a retrieval method by first obtaining the most similar cell on the map to a given search query, then using the cell to generate a ranked list of documents. Their submitted runs used English queries and a mixture of manual and automatically generated document fields.

3.6 University of Ottawa (uottawa)

The University of Ottawa employed an experimental system built using off-the-shelf components. To translate topics from French, Spanish, and German into English, six free online machine translation tools were used. Their output was merged in order to allow for variety in lexical choices. The SMART IR system was tested with many different weighting schemes for indexing the collection and the topics. The University of Ottawa used a variety of query languages and only automatically generated document fields for their submitted runs.

3.7 University of Waterloo (uwaterloo)

The University of Waterloo submitted three English automatic runs, a Czech automatic run and a French automatic run. The basic retrieval method for all runs was Okapi BM25. All submitted runs used a combination of several query formulation and expansion techniques, including the use of phonetic n-grams and feedback query expansion over a topic-specific external corpus crawled from the Web. The French and Czech runs used translated queries supplied by the University of Ottawa group.

4. Results

Table 3 summarizes the results for all 35 official runs averaged over the 25 evaluation topics, listed in descending order of mean uninterpolated average precision (MAP). Table 3 also reports precision at the rank where the number of retrieved documents equals the number of known relevant documents (Rprec), the fraction of the cases in which judged non-relevant documents are retrieved before judged relevant documents (Bpref) and the precision at 10 documents (P10). Required runs are shown in bold.

Table 3. Official runs.

Run name	MAP	Rprec	Bpref	P10	Lang	Query	Document fields	Site
metadata+syn.en.qe	0.3129	0.3494	0.3423	0.4800	EN	TD	N,MK,SUM	umaryland
metadata+syn.fr2en.qe	0.2476	0.2877	0.2819	0.3680	FR	TD	N,MK,SUM	umaryland
uoEnTDN	0.2176	0.2364	0.2005	0.3200	EN	TDN	ASR04,AK1,AK2	uottawa
titdes-all	0.1878	0.2306	0.2009	0.3640	EN	TD	All	upitt
uoSpTDN	0.1863	0.2078	0.1750	0.2640	SP	TDN	ASR04,AK1,AK2	uottawa
uoFrTD	0.1685	0.1923	0.1599	0.2960	FR	TD	ASR04,AK1,AK2	uottawa
dcusumtit40ffr	0.1654	0.2117	0.1750	0.3080	FR	T	ASR03,ASR04,AK1,AK2,SUM	dcu
uoEnTD	**0.1653**	**0.2088**	**0.1705**	**0.2960**	**EN**	**TD**	**ASR04,AK1,AK2**	**uottawa**
dcusumtiteng	0.1429	0.1994	0.1561	0.2560	EN	T	ASR03,ASR04,AK1,AK2	dcu
titdes-combined	0.1415	0.1779	0.1489	0.3600	EN	TD	Mixed	upitt
autokey+asr.en.qe	**0.1288**	**0.1719**	**0.1440**	**0.2720**	**EN**	**TD**	**ASR04,AK2**	**umaryland**
uoGrTDN	0.1281	0.1493	0.1331	0.2000	DE	TDN	ASR04,AK1,AK2	uottawa
asr.de.en.qe	0.1275	0.1882	0.1461	0.2760	EN	TD	ASR04	umaryland
uw5XETDNfs	0.1138	0.1907	0.1414	0.2720	EN	TDN	ASR03,ASR04	uwaterloo
uw5XETDfs	**0.1121**	**0.1744**	**0.1388**	**0.2760**	**EN**	**TD**	**ASR03,ASR04**	**uwaterloo**
asr.en.qe	0.1102	0.1712	0.1292	0.2800	EN	TD	ASR04	umaryland
dcua1a2tit40feng	0.1101	0.1559	0.1312	0.2520	EN	T	ASR03,ASR04,AK1,AK2	dcu
dcua1a2tit40ffr	0.1064	0.1571	0.1322	0.2600	FR	T	ASR03,ASR04,AK1,AK2	dcu
uw5XETfs	0.0980	0.1559	0.1270	0.2680	EN	T	ASR03,ASR04	uwaterloo
unedMpos	**0.0934**	**0.1522**	**0.1096**	**0.2400**	**EN**	**TD**	**ASR04**	**uned**
unedMmorpho	0.0918	0.1532	0.1097	0.2360	EN	TD	ASR04	uned
uw5XFTph	0.0848	0.1421	0.1160	0.2560	FR	T	ASR03,ASR04	uwaterloo
UATDASR04AUTOA2	0.0769	0.1181	0.0980	0.2240	EN	D	ASR04,AK2	ualicante
UATDASR04LF	**0.0768**	**0.1230**	**0.0949**	**0.1920**	**EN**	**TD**	**ASR04**	**ualicante**
titdes-text04a	**0.0757**	**0.1341**	**0.1045**	**0.2120**	**EN**	**TD**	**ASR04**	**upitt**
UATDASR04AUTOS	0.0739	0.1274	0.1056	0.2400	EN	D	ASR04,AK1,AK2	ualicante
UATDASR04AUTOA1	0.0727	0.1206	0.1018	0.2200	EN	D	ASR04,AK1	ualicante
UATDASR04	0.0724	0.1246	0.0899	0.1600	EN	D	ASR04	ualicante
uned3gram	0.0706	0.1119	0.0994	0.1800	EN	TD	ASR04	uned
dcua2desc40feng	**0.0654**	**0.1196**	**0.0944**	**0.1760**	**EN**	**TD**	**ASR03,ASR04,AK2**	**dcu**
uw5XCTph	0.0471	0.0751	0.0928	0.1320	CZ	T	ASR03,ASR04	uwaterloo
unedCLpos	0.0373	0.0750	0.0535	0.1200	SP	TD	ASR04	uned
unedCLmorpho	0.0370	0.0759	0.0536	0.1200	SP	TD	ASR04	uned
som-allelb	0.0124	0.0132	0.0397	0.0120	EN	TDN	All	upitt
som-titdes-com	0.0041	0.0147	0.0408	0.0120	EN	TD	Mixed	upitt

N = Name (Manual metadata), MK = Manual Keywords (Manual metadata), SUM = Summary (Manual metadata)
ASR03 = ASRTEXT2003A (Automatic), ASR04 = ASRTEXT2004A (Automatic)
AK1 = AUTOKEYWORDS2004A1 (Automatic), AK2 = AUTOKEYWORDS2004A2 (Automatic)

Figure 4 compares the required runs across the seven participating sites. The University of Ottawa results were statistically significantly better than all others for this condition (using a two-tailed Wilcoxon Signed-Rank Test for paired samples at $p<0.05$ across the 25 evaluation topics). The ovals in that figure group runs that are statistically indistinguishable. The best official run using manual metadata yielded a statistically significant improvement over the strongest results obtained using only automatically generated data.

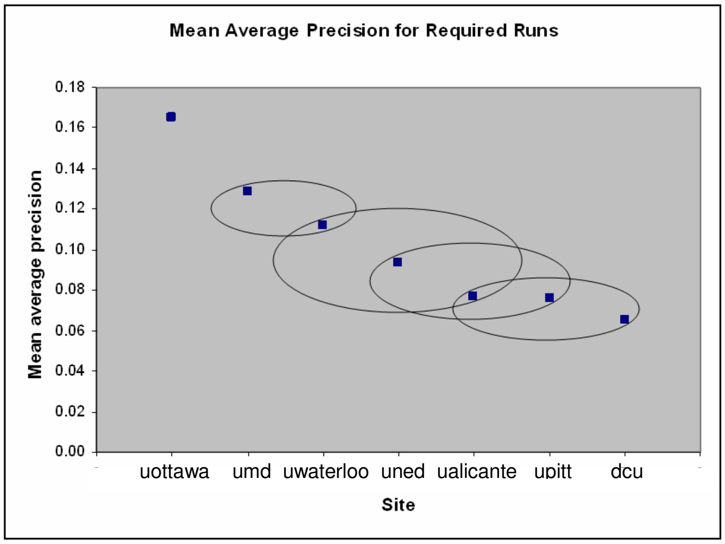

Figure 4. Plot of mean average precision for required runs

There were 8 cases in which the same site submitted both monolingual and cross-language runs under comparable experimental conditions (i.e., the same query fields and same document fields). Table 4 summarizes those results. Every query language was used. French topics proved to be the most popular for cross-language searching, being used by four of the seven participating teams. Notably, two teams achieved cross-language results for French that numerically exceeded their English monolingual mean average precision (although neither difference was statistically significant). Monolingual baselines constructed in this way are known to be deficient because cross-language retrieval introduces a natural query expansion effect. They are nonetheless useful as a reference condition.

Table 4. Percentage difference in MAP between English and non-English comparable runs.

Site (query – document)	En	Cz	De	Fr	Sp
uottawa (TD - ASR04,AK1,AK2)	0.1653	–	–	+2%	–
uottawa (TDN - ASR04,AK1,AK2)	0.2176	–	**–41%**	–	–14%
umaryland (TD - N,K,SUM)	0.3129	–	–	**–21%**	–
uwaterloo (T - ASR03,ASR04)	0.0980	**–52%**	–	–13%	–
uned (TD – ASR04)	0.0934	–	–	–	**–60%**
dcu (T – ASR03,ASR04,AK1,AK2)	0.1429	–	–	+16%	–

Two sites submitted official runs in which manual metadata and automatic metadata were used under otherwise comparable conditions (i.e., the same query length). As Table 5 shows, the use of manual metadata yielded substantial improvements that were statistically significant. This most likely reflects some combination of indexing by subject matter experts of concepts that were not lexicalized within the segment, ASR deficiencies, and a possible bias in word choices made when writing topic descriptions in favor of more formal language. We do not presently have sufficient evidence to differentiate among these three effects.

Table 5. Comparing retrieval effectiveness for Automatic and Manual metadata.

Site	MAP(Manual Metadata)	MAP(Automatic)	Automatic/Manual
umaryland – TD	0.3129	0.1288	41%
upitt – TD	0.1878	0.0757	40%

5. Conclusion and Future Plans

Overall, the CLEF-2005 CL-SR track succeeded in creating a reusable test collection, bringing together a group of researchers with similar interests, and exploring alternative techniques to facilitate access to a large collection of spontaneous conversational speech. We therefore plan to continue the track in 2006. The following options are under consideration: (1) addition of an unknown boundary condition for English using the retrieval effectiveness measures first developed for the TREC SDR evaluation, (2) release of a larger English collection (approximately 900 hours of speech) with an improved word error rate (approximately 25%), (3) release of a word lattice to permit searching alternative recognition hypotheses, and (4) creation of a second test collection containing Czech interviews. We look forward to discussing these and other when we meet in Vienna!

6. Acknowledgments

The authors would like to thank those who performed the topic translation: Christelle Ayache, Catherine Bosio, Clara Cabezas, Pavel Ircing, Fernando Lopez Ostenero, Pavel Pecina and Victor Peinado. We would also like to thank our relevance assessors: Stefan Papaioannou, Daniel Rubin, Ingo Trauschweizer, Ruth Schachter, Benzion Schwinn and Ariel Segal. Finally we are grateful to Bhuvana Ramabhadran for providing the ASR results, to Martin Franz and J. Scott Olsson for providing the text classification results, and to G. Craig Murray for the thesaurus expansion script. This work has been supported in part by NSF IIS Award 0122466. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

7. References

[1] Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B. Soergel, D., Ward, T. and Zhu, W.-J. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4), pp. 420-435.

[2] Federico, M., Bertoldi, N., Levow, G.-A. and Jones, G. J. F. (2004). CLEF 2004 Cross-Language Spoken Document Retrieval Track. In *Proceedings of the CLEF 2004: Workshop on Cross-Language Information Retrieval and Evaluation.*

[3] Garafolo, J.S., Auzanne, C.G.P. and Voorhees, E.M. (2000). The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the RIAO Conference: Content-Based Multimedia Information Access*, pp. 1-20.

[4] Huang, X. and Soergel, D. (2004). Relevance judges' understanding of topical relevance types: An explication of an enriched concept of topical relevance. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, pp. 156-167.

[5] Oard, D., Soergel, D., Doermann, D., Huang, X., Murray, G.C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S., Mayfield, J., Kharevych, L. and Strassel, S. (2004). Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of 27th Annual ACM Conference on Research and Development in Information Retrieval*, pp. 41-48.

[6] Soergel, D. and Oard, D. (2005). The MALACH English Speech Retrieval Test Collection. Technical Report, 11 pages, available with the test collection from the Evaluations and Language Resources Distribution Agency (ELDA).