Building an Information Retrieval Test Collection for Spontaneous Conversational Speech

Douglas W. Oard
Dagobert Soergel
David Doermann
Xiaoli Huang
G. Craig Murray
Jianqiang Wang
CLIS/UMIACS, U. Maryland
College Park, MD 20742
oard@umd.edu

Bhuvana Ramabhadran
Martin Franz
IBM
T.J. Watson Research Center
Yorktown Heights, NY 10598
bhuvana@us.ibm.com

Samuel Gustman
Survivors of the Shoah
Visual History Foundation
Los Angeles, CA 90078
sam@vhf.org

ABSTRACT

Test collections model use cases in ways that facilitate evaluation of information retrieval systems. This paper describes the use of search-guided relevance assessment to create a test collection for retrieval of spontaneous conversational speech. Approximately 10,000 thematically coherent segments were manually identified in 625 hours of oral history interviews with 246 individuals. Automatic speech recognition results, manually prepared summaries, controlled vocabulary indexing, and name authority control are available for every segment. Those features were leveraged by a team of four relevance assessors to identify topically relevant segments for 28 topics developed from actual user requests. Search-guided assessment yielded sufficient inter-annotator agreement to support formative evaluation during system development. Baseline results for ranked retrieval are presented to illustrate use of the collection.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Automatic Speech Recognition, Search-Guided Relevance Assessment, Oral History

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

1. INTRODUCTION

Far more is spoken each day than is written, and affordable technology to acquire and store vast amounts of conversational speech currently exists. If someone were to speak every waking moment for a year, the compressed audio could be stored on a single 20 GB hard drive. Of course, few people would consider it worthwhile to record that much speech, since finding the useful nuggets in such a vast collection of speech would seem impossible. Remarkable advances in the accuracy and speed of automatic recognition of spontaneous conversational speech, however, are beginning to change this. Our challenge now is to leverage that technology to build effective search systems. To realize that goal, we will need test collections that set a gold standard against which systems can be compared. In this paper, we describe the development of what we believe is the first realistic information retrieval (IR) test collection for spontaneous conversational speech.

Spoken word collections encompass a diverse genre, including personal dictation, news broadcasts, political speeches, recorded meetings, help desk telephone calls, and tales told as folklore [1]. We chose to build our test collection using oral history interviews because many of the interviews are already digitized, some of the digitized interviews have been manually indexed, Automatic Speech Recognition (ASR) results with a useful degree of accuracy could be produced for a portion of the collection, and descriptions of the information needed from those interviews by a diverse set of real users already exists. The main challenge was therefore to leverage the existing index to create reliable relevance judgments for a representative set of topics derived from existing information need statements. The resulting test collection includes 28 topics and about 625 hours of recognized speech. Although small from the perspective of IR (17 MB), it is similar to the size of the broadcast news collections used in the spoken document retrieval tracks in the Text Retrieval Conference (TREC) and the Cross-Language Evaluation Forum (CLEF).

The remainder of this paper is organized as follows. Prior work on speech retrieval and test collection development is briefly reviewed in Sections 2 and 3. Section 4 describes the creation of the test collection for spontaneous conversational speech. A unique feature of the test collection is that multiple facets of relevance are annotated, each on a multi-valued scale, thus providing the potential to perform some types of analysis of retrieval results that would not be possible with the more common topical binary relevance judgments. Section 5 illustrates our initial use of this collection for formative evaluation with three ranked retrieval systems. The paper then concludes with a discussion of other potential uses for the test collection and our future plans.

2. SEARCHING THE SPOKEN WORD

Over the years, two basic approaches to searching spoken word collections have emerged. Word spotting was the first to be developed; the basic idea being to represent what might have been spoken (e.g., as a lattice of phonemes) and then search that representation for likely occurrences of query terms. The time needed to search a lattice increases linearly with the size of the collection, so word spotting in large collections is practical only when the query is known in advance.

More recently, large-vocabulary ASR has offered a basis for more efficient query processing. The key idea is to search the lattice in advance for word sequences that match a language model trained using enormous amounts of representative text. This approach has proven to be useful for personal dictation (where speakers enunciate clearly and the acoustic model can be optimized for a single speaker) and recorded news broadcasts (where representative text is easily obtained and speech with good articulation is common). One weakness of present ASR techniques, however, is that they can be quite sensitive to differences between the conditions for which they are trained and the conditions to which they are applied. Differences in speaking rate, accents, background noise, emotional state, and many other factors can severely affect recognition accuracy.

Test collections have been produced using recorded news broadcasts at TREC (in the spoken document retrieval track) [10] and in the Topic Detection and Tracking Evaluations [6]. News broadcasts can be fairly naturally divided into stories, so relevance judgments are typically created at the granularity of individual stories. Some experiments have also been done with recorded telephone voice mail; in those cases, individual messages are the natural unit of retrieval [21]. Several collections of recorded telephone calls have been used for development of ASR systems for spontaneous conversational speech. One series of telephone speech collections, "Call Home," was produced by providing free calls in exchange for permission to record the call. Although this yields a collection that is suitable for ASR development, there is little recognizable topical content in the recorded calls for which meaningful queries could be constructed. The other major series, "Switchboard," was created by matching two speakers randomly and asking them to choose a topic to discuss from a list [11]. The result is an exceptional degree of focus on the selected topics, hardly a representative condition for IR.

Our goal was to build a test collection containing substantial amounts of spontaneous conversational speech, with topic descriptions based on real user needs. In the next section, we briefly survey what is known about affordable methods for building test collections that can be used to reliably identify effective systems.

3. IR TEST COLLECTIONS

Automated evaluation of ranked retrieval systems requires access to a representative set of documents, a representative set of topics, some way of automatically forming queries for those topics, and judgments of the relevance of each document to each topic.

Relevance is inherently subjective and judgments are known to vary between individuals and over time [18]. This variation seems tolerable, however, since widely used retrieval effectiveness measures such as mean uninterpolated average precision (MAP) usually rank alternative systems consistently when the judgments of equally qualified relevance assessors are substituted [20]. The most widely reported retrieval effectiveness measures are based on binary relevance judgments, but assessors generally report greater confidence in their judgments when they can express the degree relevance of an item on a multi-valued scale [14]. Rong et al suggest five-point and seven-point scales [17], and collapsing multi-point scales to binary values has been shown to give stable rankings of systems [2]. That is the approach that we adopted for the work reported here.

Early attempts at building IR test collections exhaustively judged the relevance of every document to every topic [7]. Large collections and large numbers of topics are needed to achieve stable measures, however, so this did not prove to be a scalable solution. One widely used alternative is pooled assessment, in which top-ranked documents from many systems are judged, and unjudged documents are treated as if they were not relevant [19, 20]. For moderately large collections (hundreds of thousands of documents), this typically requires judging thousands of documents per topic. Test collections built in this way can later be reused to evaluate fully automatic IR systems that did not contribute to the pools. The introduction of any human interaction or a radically different fully automated technique might invalidate that claim. The most serious limitation of pooled assessment for our purposes, however, is that it depends on contributions from a moderately large number of different systems. At present, we are working with a single ASR system and three IR systems; this does not offer sufficient diversity to build a reusable test collection using pooled assessment.

An alternative that has many of the advantages of pooled assessment is search-guided assessment [6]. In this approach, assessors first conduct detailed topic research and then interactively search the collection for relevant documents. By iterating between topic research, relevance assessment, and interactive query reformula-

tion, assessors seek to identify as many relevant documents as the design of their search system and the available time will allow. Cormack et al compared a variant of search-guided assessment with pooled assessment, finding that the resulting judgments ranked alternative systems similarly [8]. Cieri et al described a quality control process for search guided relevance assessment that further enhanced agreement with pooled assessment; subsequent pooled assessment of unjudged documents resulted in discovery of few that were relevant [6]. This approach matched our needs well, so we chose search-guided assessment with a quality control process based on independent review and team adjudication for the work reported in this paper.

4. BUILDING THE TEST COLLECTION

We built a test collection using a large collection of interviews with witnesses to the Holocaust. A subset of the collection has been digitized and manually indexed, and hundreds of serious users have posed questions that they hoped could be answered using the collection. In this section, we describe how we have leveraged those resources to produce an IR test collection.

4.1 Collecting the Interviews

We started with a large number of interviews that had been conducted by the Survivors of the Shoah Visual History Foundation (VHF) to record the recollections of Holocaust survivors, rescuers and witnesses. Interviewees were asked to complete a detailed questionnaire a week or so before their interview, both to help the interviewer prepare and for later use in search systems. A trained interviewer and a professional videographer then conducted the interview, typically in the interviewee's home. Interviews were recorded on Sony Beta SP videotapes. The videotaped interviews are now being digitized to create a 3 MB/sec MPEG-1 stream with 128 kb/sec (44 kHz) stereo audio. Some data from the handwritten questionnaires (e.g., biographical data, person names, family relationships, and locations) is also being typed by hand. The cost for interviewing, digitizing and data entry was approximately $2,000 per interview.

Approximately 4,000 of the original interviews were manually indexed. They were first manually subdivided into topically coherent segments by indexers with professional training appropriate to the subject matter. The indexers prepared a three-sentence summary for each segment, and assigned appropriate person names and an average of five controlled vocabulary terms from a domain-specific thesaurus to each segment. The names were obtained from the questionnaire where possible, or entered manually when necessary. This resulted in name authority control within an interview, but not across interviews. Indexers often took notes online to aid them in their work; at least some text is available from this source for about 75% of the interviews. After indexing all segments from an interview, the indexer created a half-page summary of the entire interview. The average cost of this indexing process was another $2,000 per interview. The index is now in regular use as part of a controlled vocabulary search system.

It proved to be unaffordable to scale this process up to the entire collection, thus motivating research on techniques that would minimize the required human indexing effort while still providing adequate support for subsequent access to these materials. This initial investment was not lost, however; it resulted in an exceptional set of training data for supervised machine learning techniques, and it led directly to the creation of the IR test collection described in this paper. Of the 4,000 English interviews manually indexed, 1,514 had been digitized by the time we started to build this test collection.

4.2 Topic Construction

Our oral history collection has attracted significant interest from scholars, educators, documentary film makers, and others, resulting in over 250 topic-oriented written requests for materials from the collection [12]. From that set, we selected 70 requests that we felt were representative of the types of requests and the types of subject contained in the topical requests. The requests were typically made in the form of business letters, often accompanied by questionnaire responses describing the requester's project and purpose. Additional materials (e.g., a thesis proposal) were also sometimes available. TREC-like topic descriptions consisting of title, a short description and a narrative description were then created for the 70 topics, as shown by the following example:

```
<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>
```

Our initial collection is relatively small from an information retrieval perspective, so we took two steps to ensure the presence of an adequate number of relevant segments to distinguish retrieval effectiveness among alternative systems. In some cases, we broadened 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.

4.3 Creating Relevance Judgments

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.[1] We therefore defined five relevance categories, both to guide the thinking of our assessors and to obtain judgments differentiated by category to serve as a basis for more detailed analysis than would be possible using single-facet judgments.

The relevance categories are based on the notion of evidence (rather than, for example, potential emotional impact or appropriateness to an audience). Five categories were derived from our understanding of historical methods and information seeking processes. These categories were then refined during a two-week pilot study through group discussions with our assessors [13]. The resulting categories were:

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

Each type of relevance was judged on a five-point scale (0 to 4). 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 was: "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 3 minutes, so the brevity of a mention is another factor that could prove useful when analyzing differences in retrieval effectiveness among alternative systems. We therefore asked assessors to indicate the fraction of the segment that was associated with each of the five categories. Assessors were instructed to treat brevity and degree separately (a very brief mention could be highly relevant).

The relevance judgments were created using search-guided assessment. This was done before ASR results were available, so all searches were based on manual indexing. Using Lucene, we indexed the segments using thesaurus terms, person names, segment summaries, and (when available) the indexers' online notes. The set of thesaurus terms assigned to each segment was expanded by adding broader terms from the thesaurus. Interview-level metadata (questionnaire responses and the interview summary) were also added to the index for each summary. The system supported fielded searching, using both unstructured queries for ranked retrieval and structured Boolean queries. Retrieved segments were arranged by interview and within each interview by the order in which they appear. The display order was structured to place interviews with many highly ranked segments ahead of those with fewer.

The assessor interface was designed to make examination and entering assessments as efficient as possible. The screen included regions for query entry, display of a result list, and display of detailed information for a single segment, and assessment. The result list shows the summary for each segment found. Segments that are clearly not relevant can be checked off quickly there. Clicking on a segment in the result list shows the segment summary, the indexer's online notes, the (unexpanded) thesaurus terms, person names assigned to the segment, and (if available) the indexer's personal notes. The interview summary and questionnaire responses are also available by selecting tabs within the detailed result display. There are arrows to look at the 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. There are also affordances to play the audio for digitized segments.

The assessment region included a drop-down menu with which to designate the degree of relevance for each category (with 0 selected by default) and a slider to indicate roughly the portion of the segment that pertains to that category. There were also a number of check boxes to enter data about the relevance assessment (e.g., "difficult judgment," or "judgment based on indexer's notes"). The assessor could also highlight passages of text (e.g., from the segment summary) and designate that text as evidence for a relevance category.

Four graduate students studying history or library science worked about 700 hours over 3 months to create 15,343 relevance judgments in 404 full interviews for 31 of the 70 available topics. Interviews that had not yet been digitized were then removed from the collection, and 28 topics with at least 5 relevant segments among the remaining interviews were then selected for inclusion in the test collection. The resulting test collection contains 28 topics, 199 full interviews, and 47 partial interviews.[2]

The relevance assessors were experienced searchers; they made extensive use of Boolean 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 query, query-specific guidelines for deciding on the level of relevance for each relevance category, and other issues (e.g., the rationale for judging specific segments). These notes supported adjudication, and they can be used in the future to support additional judgments for segments that are highly ranked by IR systems.

Of the 28 topics, 14 were independently assessed by two assessors. The assessors then met to adjudicate cases in which at least one had assigned a high score (3 or 4) to some facet. Assessors referred to their notes during adjudication and could run new queries, possibly discovering additional relevant segments. Other judg-

[1]This broader concept of "relevance" is sometimes referred to as "utility."

[2]Partial interviews contain at least one 30-minute tape.

ments on which there were differences were automatically averaged (rounding up). The remaining 14 topics were reviewed by a second assessor. Reviewers were instructed to review the entire process (based on the first assessor's queries and notes), and to run new queries if appropriate. Reviewers checked any high-scoring judgments (3 or 4) for any facet and a selective sampling of the other judged segments. The decision of a reviewer was final.

We produced two sets of binary relevance judgments from this data for the preliminary experiments reported in this paper. The first, representing the opinion of a single individual, was based on independent assessment by whichever assessor recorded more judgments. The second set of judgments captured the full effect of our quality assurance process, including independent review or adjudicated assessment by multiple assessors. 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.

4.4 Automatic Speech Recognition

Our interviews contain natural speech filled with disfluencies, heavy accents, age-related coarticulations, uncued speaker and language switching, and emotional speech. The speaking rate is highly variable across the collection, averaging 146 words per minute (with a dynamic range of 100 to 200).[3] The recordings were made under a wide variety of conditions (e.g., quiet room, periodic airplane overflights, wind or highway noise, and background conversations). Microphone positions varied, but typically one channel was intended to record the interviewer and the other to record the interviewee. In practice, both channels picked up both speakers. The audio channel with the greatest average energy (generally, from the interviewee) was downsampled to 16 kHz and parameterized using 24-dimensional mel frequency cepstral coefficients; the other channel was not processed. Acoustic features were derived using linear discriminant and maximum-likelihood based linear transformations. Speaker-specific transformations (SAT and MLLR) were then applied. Details are described in [3, 15].

ASR systems are trained using representative examples of transcribed speech. A 200-hour training corpus was created from contiguous 15-minute excerpts starting at randomly selected points in 800 randomly selected interviews. Male and female speakers are more or less equally represented, and many accents are present. Manual transcription proved to be challenging, typically requiring 8 to 12 hours per hour of speech, in part because of unfamiliar names and places and occasional use of words from languages other than English.

> "...so I didn't I never left New York before I didn't know anything else so some fellow I knew mentioned that uh he sa- said I have a friend that lives in Arizona in Tucson Arizona so I went to the map looked it up um I never heard of Tucson uh and any- anyhow he says well I'll write him a letter and when you go there you could uh stay with him so he did he wrote a letter and his friend he was a dentist he invited me to come over there and spend a week with him..."

[3]For comparison, the average speaking rate in the well-studied Switchboard corpus is 100 words per minute.

System	Word Error Rate
Speaker-Independent	51.3%
SAT	43.6%
MLLR + SAT	39.6%

Table 1: Recognition effectiveness on the ASR test set.

The language model used for decoding was built using the modified Kneser-Ney algorithm [5] by interpolating this relatively small (1.7 million word) training corpus with the Broadcast News and Switchboard corpora (158 million and 3.4 million words, respectively), optimizing the interpolation weights to achieve minimum perplexity on held-out manually prepared transcripts. ASR systems can only hypothesize words that appear in their lexicon. Person and place names have proven to be important to searchers of this collection [12], so person names obtained from the questionnaires, and common place names were added from a large domain-specific list. These rare terms are inevitably modeled less well than more common ones, but their presence in the lexicon makes their recognition possible. The resulting ASR lexicon contained approximately 30,000 words.

Preprocessing prior to decoding includes channel selection, acoustic segmentation of the audio into contiguous periods of speech or (possibly noisy) silence, and clustering to generate coherent-speaker sets of speech periods. Acoustic segmentation avoids some types of insertion errors, enhances robustness to background noise, and improves computational efficiency. Speaker labeling allows adaptation on speaker-coherent clusters. The entire recognition process, including preprocessing and speaker adaptation, requires about ten machine-hours per hour of speech. We created a one-hour ASR test set by manually transcribing brief segments taken from 20 randomly chosen previously unseen speakers. Table 1 presents the ASR results obtained on the ASR test set.

The resulting test collection contains approximately 625 hours of recognized speech by interviewees that do not appear in the ASR training collection.[4] The topical segment boundaries defined by the indexers were adjusted to the nearest significant silence (2 seconds or longer), and the words produced by ASR were treated as the text of that segment, resulting in 9,947 segments with an average length of 380 words. For contrastive studies, metadata for each segment (summaries, notes, thesaurus terms, and person names) were included as additional fields that can optionally be indexed.

[4]Some interviewers appear in both collections because interviewers typically conducted several interviews.

	Single Assessor		Adjudicated	
	Title	Full	Title	Full
ASR	0.0717	0.0713	0.0694	0.0720
Notes	0.1015	0.1169	0.1022	0.1231
ThesTerm	0.2834	0.3385	0.2817	0.3367
Summary	0.2783	0.3029	0.2823	0.3029
Manual	0.3799	0.4122	0.3819	0.4188
All	0.3439	0.3887	0.3460	0.3952

Table 2: Maryland experiments. MAP for alternative sources of indexing terms, title and full queries, adjudicated and unadjudicated judgments.

5. USING THE TEST COLLECTION

This section draws on early experiments at three sites to illustrate some of the ways that the collection can be used. For our experiments at Maryland we used InQuery to index six types of terms using the standard InQuery stemmer (kstem) and stopword list:

ASR: Terms from ASR

Notes: Terms from the indexer's notes, when available

ThesTerm: Thesaurus terms assigned to the segment

Summary: Terms from the segment summary

Manual: All terms from Notes, ThesTerm and Summary, plus person names assigned to the segment

All: All terms from ASR and Manual

As Table 2 shows, title-only and full (title, description and narrative) queries yielded similar results. Moreover, adjudication and review did not markedly alter the relative effectiveness of reported collections. We have not yet evaluated inter-annotator agreement directly, but the similarity of single-assessor and adjudicated-judgment results suggests that this will not be a serious source of error when comparing alternative systems. Use of all available manual indexing yields an average of 4.8 relevant segments in the top 10, which matches our expectation for a collection of this size and relevant segment density. ASR alone does relatively poorly (18% of Manual). Segment summaries and thesaurus terms are clearly the most useful sources of index terms.

As Figure 1 shows, the relative effectiveness of ASR and Manual indexing depends strongly on the topic. The 16 topics with an average precision above 0.2 for manual indexing divide fairly clearly into three groups. At the low end, there were 7 topics for which ASR is below 5% of Manual. This essentially amounts to complete failure. In 4 of the 7 cases, the title query contains at least one highly selective person or organization name that was not recognized by ASR in any segment (although in one case the missed term was present in the ASR lexicon). In the middle, there were 8 topics for which ASR is between 8% and 36% of Manual. For each of these topics, every term in the title query was recognized in at least one segment. The lowest four of these

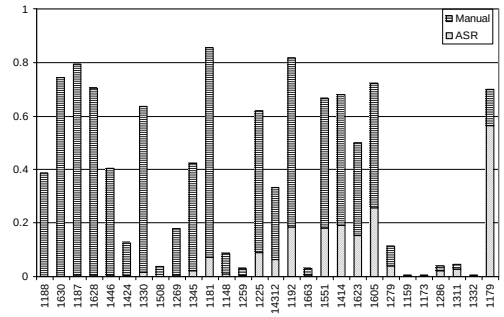

Figure 1: Average precision by topic, title queries. Full bar: Manual. Partial bar: ASR.

	Single Assessor	Adjudicated
Unexpanded	0.0698	0.0681
BRF	0.0747	0.0740
Merged	0.0951	0.0941

Table 3: IBM experiments. MAP, title queries, ASR segments. Merged scores are 30% BRF and 70% Category-based.

(below 23%) included at least one domain-specific term, while the remaining four (above 27%) used only more common terms that would be expected to be present in the larger interpolated language models. Document expansion might be expected to be effective in this range. Finally, at the high end, there was 1 topic for which ASR achieved 81% of Manual. The title query for that topic again makes no use of domain-specific terminology.

For our experiments at the IBM T.J. Watson Research Center, we use an IR system similar to the one we previously applied as a core scoring component in our TREC participation [9]. Text preprocessing consists of tokenization using a decision tree tokenizer, stemming with the Porter stemmer, stop-word removal using a list of 508 stop words, and division into overlapping 50-word passages. Relevance scores are computed using an approach based on the Okapi formula [16]. As Table 3 shows, Blind Relevance Feedback (BRF) yields an 8% relative improvement in MAP (with adjudicated judgments) that is not significant at $p < 0.05$ by a Wilcoxon signed-rank test for paired samples.

The results from Maryland suggest that manually assigned thesaurus terms are a useful basis for search, so we have also explored one way of leveraging our previous work on automatic classification in this domain [3]. We trained a k Nearest Neighbor (kNN) classifier [22] using 3,199 manually transcribed segments from the ASR training set. We then applied the resulting model to the ASR results in the IR test collection. The accuracy of the resulting classifier is relatively poor (with a microaveraged balanced F measure of 0.192), in part

	Single Assessor	Adjudicated
Title	0.0692	0.0695
Title+Desc	0.0719	0.0695
Full	0.0774	0.0715

Table 4: APL experiments. MAP, three query lengths, ASR segments, character 5-grams.

because the training set is relatively small.

The thesaurus terms associated with the top 20 category assignments for each segment were then indexed, and the resulting index was used as a basis for a second search. This second search alone was less effective than a search using blind relevance feedback that is based on the ASR results alone, but as the merged entry in Table 3 shows, a linear combination of scores from the two searches can yield as much as a 30% (statistically significant) improvement in MAP.

For our experiments at Johns Hopkins Applied Physics Lab (APL), we indexed the ASR results using word-spanning character n-grams. As Table 4 shows, 5-grams yielded MAP values comparable to those obtained at Maryland and IBM using stemming. These results are slightly better than we obtained with 4-grams. From this we conclude that character n-grams exhibited little or no benefit with this test collection from their potential to conflate acoustically confusable words.

5.1 Assessing the Assessments

Search-guided relevance assessment was completed before these experiment results became available, so we have not yet had an opportunity to formally assess highly ranked unjudged segments from any of these systems. There is, however, good evidence that the present sets of relevance judgments are useful. We calculated agreement between their judgments using two measures. Overlap is calculated as the proportion of items judged relevant by either assessor which were also found to be relevant by the other assessor. Topic-averaged overlap is 44% (with a minimum of 4% and a maximum of 83%) Our computation of kappa, by contrast, characterizes agreement on all segments that were judged by at least one of the two assessors [4]. Topic-averaged kappa is 0.63 (with a minimum of 0.24 and a maximum of 1.0). These values agree well with results reported by Voorhees using pooled assessment [20].

MAP is sensitive to accurate assessment of highly ranked segments, so the existence of judgments for such segments is another indicator of the utility of the assessments when comparing systems. Adjudicated relevance judgments are available for an average 72% of the top 10 segments for our best run (Maryland, manual, full queries) and for 30% of the top 10 segments for our best ASR run (IBM, merged). To get a feel for whether unassessed segments should be a matter of concern, we randomly selected ten highly ranked unjudged segments from several runs. Two of those segments appear (to our untrained ear) to be relevant; that seems to be within the expected range of inter-annotator agreement.

Finally, changing from single-assessor to adjudicated judgments changed the preference order between alternative techniques only in two cases that were quite close to begin with. From this we can conclude that the adjudicated judgments are likely to prove useful as a basis for computing stable system rankings. We will assess additional topics this summer and we do plan to perform additional assessments of some highly ranked segments for the present topic set at that time. But we believe the preset set of judgments to be sufficiently thorough for use in formative evaluation of alternative techniques for searching spontaneous conversational speech.

6. CONCLUSIONS/FUTURE WORK

Our test collection can support much richer exploration of retrieval from spontaneous conversational speech than we have been able to exercise in the few experiments that we have reported here. Our next step will be to assess inter-annotator agreement across the full range of relevance categories, degrees of relevance, and duration of mention. With that understanding, we should be able to probe our systems to determine whether improved tuning or new approaches could yield result sets that better match the needs of our users as we understand them. We also plan to conduct a more detailed analysis for cases in which ASR did relatively poorly, manually transcribing some known relevant segments to accurately characterize the spoken content of those segments. We must, however, guard against drawing overly general conclusions from results on a single relatively small test collection; additional test collections will ultimately be needed. The Shoah Foundation is presently engaged in a massive digitization and indexing effort that will ultimately make it possible to build test collections at least an order of magnitude larger than the collection described in this paper. At the same time, we are building a community of users from whom we can continue to learn about the true information needs that motivate those who seek access to this collection [12]. As additional research teams begin to conduct experiments with the collections that we build, the balance between pooled and search-guided relevance assessment will likely shift in favor of pooled judgments.

The collections we build can be used in several other ways. The marked segment boundaries make our existing collection well suited to research on automatic segmentation of spontaneous conversational speech, and the availability of manually assigned index terms make the collection useful for exploring the application of text classification techniques to spontaneous conversational speech. Cross-language IR might be explored by translating the topic descriptions, and we ultimately expect to support an even richer range of multilingual experiments as more ASR systems become available. The full collection includes interviews in 32 languages, and our colleagues at the University of West Bohemia and Charles University have already developed an ASR system with similar accuracy on interviews in Czech.

Perhaps the most important contribution of this work will not be the new answers we find, but rather the new

questions. For example, topic segmentation makes sense for broadcast news materials, but it is not clear that this is even the right way to think about supporting access to spontaneous conversational speech. We have chosen to rely on topic segmentation for the test collection described in this paper simply because by doing so we gain access to a wealth of prior art on the evaluation of IR systems. But we have already shifted our manual indexing effort from segment-based annotation to time-based annotation, and in future test collections we hope to explore evaluation designs that reflect our growing understanding of access strategies for unsegmented spontaneous conversational speech.

Taken together, this is an audacious research agenda that demands a concerted effort from a broad community. But the potential payoffs are immense. For thousands of years, writing has occupied a privileged place in our society because it possessed two unique characteristics: it could be preserved over long periods, and things that had been written could later be found. Our generation will ultimately achieve the same capabilities for the human voice. We can only begin to imagine how that will change the world in which we live.

Acknowledgments

Thanks to Anton Leuski for help building queries and Meghan Glenn for comments. This work has been supported in part by NSF IIS Award 0122466 and NSF CISE RI Award EIA0130422. 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. ADDITIONAL AUTHORS

James Mayfield (The Johns Hopkins University Applied Physics Laboratory, james.mayfield@jhuapl.edu), Liliya Kharevych (California Institute of Technology, lily@cs.caltech.edu), Stephanie Strassel (Linguistic Data Consortium, strassel@ldc.upenn.edu).

8. REFERENCES

[1] EU-US working group on spoken-word audio collections, 2003. http://www.dcs.shef.ac.uk/spandh/projects/swag/.

[2] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *SIGIR 2000*, pages 33–40, 2000.

[3] William Byrne et al. Automated recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4), 2004.

[4] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[5] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Computer Speech and Language*, 1999.

[6] C. Cieri et al. Corpora for topic detection and tracking. In *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic, Boston, 2002.

[7] C. Cleverdon. The Cranfield tests on index language devices. *ASLIB Proceedings*, 19(6):173–194, 1967.

[8] G.V. Cormack et al. Efficient construction of large test collections. In *SIGIR '98*, pages 282–289, 1998.

[9] Martin Franz et al. Ad hoc and multilingual information retrieval at IBM. In *TREC-7*, 1998.

[10] J.S. Garofolo et al. The TREC spoken document retrieval track: A success story. In *TREC-8*, 1999.

[11] J. Godfrey et al. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 517–520, 1992.

[12] Samuel Gustman et al. Supporting access to large digital oral history archives. In *The Second Joint Conference on Digital Libraries*, pages 18–27, 2002.

[13] Xiaoli Huang and Dagobert Soergel. Relevance judges' understanding of topical relevance types: An explication of an enriched concept of topical relevance. In *Annual Meeting of the American Society for Information Science and Technology*, 2004. to appear.

[14] R.V. Katter. The influence of scale form on relevance judgment. *Information Storage and Retrieval*, 4(1):1–11, 1968.

[15] B. Ramabhadran et al. Towards automatic transcription of large spoken archives - English ASR for the MALACH project. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2003.

[16] S. E. Robertson et al. Okapi at TREC-3. In *TREC-3*, pages 109–126, 1994.

[17] T. Rong et al. Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science*, 50(3):254–264, 1998.

[18] L. Schamber. Variations in relevance and information behavior. In *Annual Review of Information Science and Technology*, volume 29, pages 3–48. 2000.

[19] K. Sparck-Jones and C.J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–72, 1976.

[20] E.M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[21] S. Whittaker et al. Play it again: A study of the factors underlying search browsing behavior. In *CHI '98*, pages 247–248, 1998.

[22] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99*, 1999.