A Test Collection for Spoken Gujarati Queries

Douglas W. Oard
University of Maryland
College Park, MD USA
oard@umd.edu

Rashmi Sankepally
University of Maryland
College Park, MD USA
rashmi@umd.edu

Jerome White
New York University
Abu Dhabi, UAE
jerome.white@nyu.edu

Aren Jansen
Johns Hopkins HLTCOE
Baltimore, MD USA
aren@jhu.edu

Craig Harman
Johns Hopkins HLTCOE
Baltimore, MD USA
craig@craigharman.net

ABSTRACT

The development of a new test collection is described in which the task is to search naturally occurring spoken content using naturally occurring spoken queries. To support research on speech retrieval for low-resource settings, the collection includes terms learned by zero-resource term discovery techniques. Use of a new tool designed for exploration of spoken collections provides some additional insight into characteristics of the collection.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

speech retrieval; relevance judgment; test collection

1. INTRODUCTION

Like many sciences, information retrieval must accommodate some differences between what can be studied in the lab and what naturally occurs in practice. Information retrieval in general, and speech retrieval in particular, relies on test collections in which researchers can test and compare their ideas in a setting that is reproducible. For results obtained using such test collections to be transferable, we need collections that reflect the characteristics of specific settings. For research in speech retrieval, striking this balance can be particularly challenging. The difficulty lies in *what* speech is being used for testing, and *how* that speech is indexed. In an ideal situation, studies would cover speech that occurs naturally in the intended application setting, indexed in a manner that is practical for that setting. There are now several test collections that model high-resource settings fairly well (e.g., retrieval of English-language news broadcasts [5, 3]), but we are not aware of other test collections that model search by low-literacy users, our focus in this paper. In such cases both spoken content and spoken queries are required; we only know of one present test collection (for retrieval of Japanese technical talks) with such a structure [1].

It has been repeatedly shown that Large-Vocabulary Continuous Speech Recognition (LVCSR) can produce useful features on which ranked retrieval can be based, but creating robust LVCSR systems with current techniques is an expensive undertaking, costing on the order of $100,000 U.S. dollars or more. For that reason, sufficiently accurate LVCSR systems are presently available for just a few dozen languages. An alternative is to bypass pronunciation and language models altogether and focus on terms identified in some acoustic or phonetic feature space. This approach to "zero-resource term discovery" has been shown to be effective in text clustering and classification experiments when applied to English [4], and in detection of single terms in several languages for which information retrieval test collections are not yet available [2]. Extending the application of zero-resource term discovery to more complex queries requires the development of new test collections. This paper describes the construction of one such test collection.

2. TEST COLLECTION FRAMEWORK

We often think of test collections as containing documents, topics, and relevance judgments. This, however, may be too narrow a view when evaluation of the test collection is the main goal. Thinking somewhat more comprehensively, test collections can include:

Content: The content to be searched, in its original form. In a test collection, this is often a combination of words and structure; in a multimedia collection, this is typically digital content or some digitization of the original artifact.

Representation(s): Additional representations of the content that can be used as a basis for search. Such representations are typically included to foster specific ways of using the collection, or to facilitate certain comparisons. For the test collection in this paper, we use terms that are discovered automatically from acoustic features based on detectable repetition of relatively long acoustic units.

Targets: What is sought. In a text collection, this is often documents, which are an implicit packaging of the content into units that serve as the target for retrieval. For rela-

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767791 .

tively long spoken content, this could be the point where replay should begin. For the test collection described in this paper, the target is a relatively short recording that we call a "response."

Queries: The basis for search. For fully automated evaluations, it is common to specify one or more standard query forms (e.g., TREC "title queries") so as to enhance comparability across systems. In our collection, we use a complete spoken voice forum post as the query.

Topics: The basis for relevance judgment. Topics and queries are often confused: topics are a mental state of the assessor; topics can be summarized in writing, but the assessor's opinion, not the summary, is the basis for each judgment. We provide written summaries of each topic.

Relevance Judgments: Human-generated encoding of the degree to which each target satisfies a topic. It is typically not possible to judge every target for relevance to every topic, so some sampling is required. For our collection, we sample by pooling highly ranked targets from several systems for each query, a common design.

Evaluation Measures: A characterization of experiment results. Although rarely made explicit, test collections are generally designed with some class of evaluation measures in mind. Our test collection is intended for use with ranked retrieval measures such as Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR), and with measures such as bPref and xinfAP which are designed to be relatively robust to incomplete relevance judgments [10].

3. THE TEST COLLECTION

Queries and responses in our collection were formulated using recorded spoken content from the Avaj Otalo "speech forum," an information service that was regularly used by farmers in Gujarat, India [8]. The goal was to provide a resource for the local farming community to exchange ideas and have their questions answered. To this end, farmers would call into the system and peruse answers to existing questions, or would pose their own questions for the community. Other farmers could call into the system to leave answers to those questions. On occasion, there were also a small group of system administrators who would periodically call in to leave announcements that they expected would be of interest to the broader farming community. The system was completely automated—no human intervention was involved. Avaj Otalo's recorded speech was divided into 50 *queries* and 2,999 *responses*. Queries were intended to be statements on a particular topic, sometimes phrased as a question, sometimes phrased as an announcement. As an example, consider an English translated summary of Query 6: *"This question is about the crop of "jeera" (cumin). Medicine Pento Methyl was provided at some first stage to prevent growing of "nindaman" but yet it has grown. Then can the medicine be provided at the second stage too?"* The 2,999 responses varied between answers to questions, additions to announcements, and new questions on similar topics. Very short recordings were omitted, as were those in which little speech activity was automatically detected. The average length of a query was approximately 70 seconds (SD = 14.40s), or approximately 61 seconds (SD = 15.76s) after silence was automatically removed. Raw response lengths averaged 110 seconds (SD = 88.80s), or 96.52 seconds (SD = 82.75s) after silence was removed.

	Round 1			Round 2			Round 3		
	A	B	C	A	B	C	A	B	C
A	-	0.41	0.39	-	0.28	0.32	-	×	×
B		-	0.62		-	0.87		-	0.85

Table 1: Inter-assessor agreement (κ).

3.1 Representation

Audio was processed using a zero-resource term discovery system described by Dredze et al. [4]. The system detects repetitions of similar speech patterns, assigning a unique (and arbitrary) term identifier to each set of similar patterns. Because the system lacks knowledge of word, syllable, or phoneme boundaries, several terms of different temporal scope can be co-active; it is not uncommon to have a dozen or more such terms overlapping. The resulting terms can be indexed by any information retrieval system, although systems that model term length and term overlap have been shown to yield better results than systems that treat all terms similarly [7].

3.2 Relevance Judgments

Judgment pools were formed by combining top-ranked results from several ranked retrieval systems developed by White et al. [9]. Three native speakers of Gujarati performed relevance assessment; none had any role in system development. Assessment was performed by listening to the audio and making a graded relevance judgment. Assessors could assign one of the following judgments to each response: 1) unable to assess, 2) not relevant, 3) relevant, or 4) highly relevant. To support computation of NDCG, the relevant and highly relevant categories were coded as 1 and 2, respectively; non-relevant judgments were coded as 0. For evaluation measures such as Mean Average Precision (MAP) that require binary judgments, and for evaluating annotator agreement, relevance judgments were subsequently binarized by collapsing highly relevant and relevant responses to relevant. Three rounds of assessments were conducted.

The first 20 queries were used for Round 1 of relevance judgments. For each query, the top 10 responses retrieved by each of three basic systems [7] were judged. All assessors judged each pooled response. As outlined in Table 1, assessors B and C were largely in agreement ($\kappa > 0.6$), while Assessor A was an outlier. The assessors then met to discuss the assessment process. To facilitate this discussion, specific cases of disagreement were randomly sampled for each assessor pair and used to seed the discussion.

The second set of judgments were created by pooling the top 10 responses for all 50 queries from 21 more sophisticated systems [9], most of which used term length and term overlap as features. Assessor A judged queries 1–15, Assessor B judged queries 16–30, and Assessor C judged queries 31–45. All assessors independently judged queries 46–50. Agreement between assessors B and C improved from Round 1 ($\kappa > 0.8$), while Assessor A remained an outlier. Assessor A's judgments of queries 1–15 were retained, as those judgments had been discussed with assessors B and C. Assessor C's judgments for queries 46–50 were retained.

During the first two rounds of judging, assessors occasionally mentioned that some queries were difficult to assess

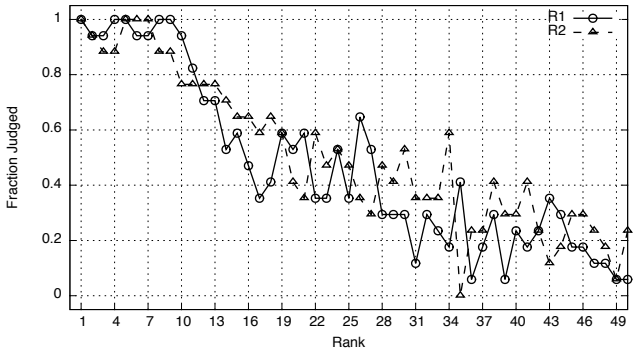

Figure 1: Judged responses by rank.

because they did not clearly express an information need. To help clarify the issue, assessors wrote brief English summaries of each query after the second round. Based on these summaries, and further discussion with assessors, 33 queries that did not pose a clear information need were removed from the collection. The final round of judgment pools were then constructed by pooling the top 10 responses from 15 sophisticated systems. The 17 retained queries were assessed as follows: Assessor C judged queries 1, 2, 6, 9, 11, 12, 23, 44, 48, and 50; Assessor B judged queries 24, 26, 28, 30, 38, 42–44, 48, and 50. The multiple judgments for queries 44, 48, and 50 were used to measure inter-annotator agreement ($\kappa > 0.8$). Assessor C's judgments were used for these cases.

Because pools from Rounds 2 and 3 differed somewhat, the completeness of the assessments was enriched by combining judgments from those pools as follows: 1) if only one judgment is available, keep it; 2) if judgments from two rounds are the same, keep it; 3) if one marked "highly relevant" and the other marked "relevant", keep it "highly relevant". 4) if only one marked "unable to assess", keep the other; 5) remove cases of clear disagreement (e.g, one marked "non-relevant," the other marked "relevant").

The test collection was designed to support three principal evaluation measures: NDCG, MAP, and MRR. MRR can be computed for all 17 topics. MAP yields the same result as MRR for the seven topics for which only one relevant response is known, so MAP is best used for the other 10 topics (each having three or more relevant responses). There are a total of 61 relevant responses for those 10 topics, an average of 6.1 responses per topic ($SD = 2.13$).

4. UTILITY OF THE TEST COLLECTION

It can be useful when introducing a new test collection to review characteristics that assess its utility—aspects that make it effective and generalizable. Although not a complete list, there are at least three such characteristics that deserve attention, and in which this test collection is couched:

Insightfulness: Ultimately test collections are built in order to answer questions, so the degree of insight that a test collection offers is a key criterion. On the positive side, this test collection includes real content, both as queries and as responses. It also includes relevance judgments made by native speakers of Gujarati, whose judgments exhibit good inter-annotator agreement. On the negative side, relevance judgments were not made by actual users of the system, and removal of non-topical queries late in the process resulted in a query set that is too small to reliably support statistical significance testing. Further, the number of responses within the collection is at the low end of what is typically expected of a real-world application.

Affordability: Acquiring content, creating representations, constructing queries, generating documentation, and sharing the collection all incur some cost. Those costs, however, are often dwarfed by the costs of creating relevance judgments. The design of the relevance assessment process is thus often the central focus of affordability concerns, so affordability and insightfulness are invariably in some tension. In this case, the two were balanced by using pooling to produce incomplete but useful relevance judgments, and by making judgments directly on the audio.

Reusability: The reuse of relevance judgments can enhance affordability by amortizing assessment costs over multiple future uses. Reuse also benefits insightfulness by eliminating otherwise uncontrolled differences in assessor opinions. Reusability thus receives special attention in test collection design. Reusability and insightfulness are sometimes in tension, however, as newly designed systems may find responses that are unlike those that have been previously assessed. Speech retrieval test collections are particularly vulnerable to this effect, since changes in either the speech processing or the retrieval algorithms could result in new systems returning results from previously unassessed parts of the collection. Figure 1 shows the fraction of responses placed at a given rank, across all 17 queries. When this fraction is high, every evaluation measure can reasonably be compared. As the fraction decays, the attention should shift to measures such as bPref and xinfAP that can accommodate a moderate level of missing judgments. Notably, the robustness of such measures is typically assessed using random ablation of a rich set of judgments, and there is no assurance that the missing judgments in any specific case are well modeled by random selection. Nonetheless, in the absence of evidence to the contrary, use of measures that are designed to be robust to missing judgments would be a reasonable choice when the ranks of interest exhibit moderate judged fractions (e.g., 0.3 to 0.7). The example shown in Figure 1 is for the best system (by MAP) that contributed to the judgment pools (R1) and for an intentionally somewhat different system that did not contribute to the judgment pools (R2). In the case of R1, most of the top 10 responses have been judged for relevance, as would be expected given our assessment methodology. The observed exceptions are the few responses that were determined by the assessor to be unjudgable. Although R2 exhibits a larger drop-off by rank 10 than R1, that drop-off is moderate: 76% of the responses have been judged. Overall, the fraction of judged responses for R2 maintains a strong correlation to R1 throughout the ranked range ($r(998) = 0.91, p < 0.001$). While it is necessary to look at such plots on a case-by-case basis, such a result is evidence for collection reusability.

5. EXPLORING THE COLLECTION

Unlike text retrieval, where visualizing representations can be straightforward, conventional ways of exploring untranscribed speech rely on listening. We therefore developed VaporEngine, a collection browsing and annotation tool.[1] In

[1]https://vapor.umiacs.umd.edu

Figure 2: Using VaporEngine to explore a response.

VaporEngine's term-centric view, waveforms for several occurrences of the term are displayed and the user can play those examples in quick succession. As each occurrence plays, the link to the response in which the term was found is highlighted. Following a link shows the corresponding response-centric view. Text entry fields are available in either view for a user to provide an English gloss translation for the term. In the response-centric view shown in Figure 2, a term cloud is shown, where each term initially has a generic identifier (e.g., "pt8834"), sized by the term frequency in the response. As English gloss translations are provided by the user, generic identifiers are replaced with the corresponding English term, both in this term cloud and in term clouds for every other response. Thus, as the user annotates terms, the generic clouds progressively transform into English term clouds. In this way, the response-centric view supports efficient collection browsing with a relatively modest annotation effort.

State-of-the-art zero resource term discovery systems are most easily able to detect repetitions of longer terms that are spoken by speakers with similar vocal characteristics. We found, from a bilingual annotator who used VaporEngine to browse the Avaj Otalo collection, that an unexpectedly large number of such terms arose from speaking conventions used in announcements; which consistently included long salutations, and which were recorded by a relatively small number of speakers (e.g., agricultural outreach agents). Such insights could help with the design of retrieval systems for this type of content.

We therefore generated six term importance criteria that could be used to prioritize the annotation effort for any language, each combining document frequency and median term duration in some way. The top 50 terms by each criterion were then pooled to form a single set and presented to the annotator in VaporEngine. The annotator was able to translate 200 terms in four hours, which covered the top 50 terms of each term ranking. Each translation was then manually labeled as a "content term" if the assessor felt it was indicative of the content of the response in which it was found. For each criterion, two evaluation measures were computed for each ranking: 1) term precision (the fraction of the 50 annotated terms that were labeled as content-bearing), and 2) collection coverage (a recall-like measure computed as the fraction of responses that contain at least one term annotated as content-bearing). At one end of the spectrum, emphasizing document frequency resulted in 11/50 content terms that together covered 16% of the collection. At the other end, emphasizing duration resulted in 28/50 content terms that covered only 2% of the collection. Balancing the two yielded a term precision of 25/50 terms as content-bearing that covered 8% of the collection. From this we conclude that the collection is sufficiently rich in content-bearing terms to be used in retrieval experiments, that visualization systems such as VaporEngine that leverage repetition in speech are potentially useful tools when designing ranking functions, and that additional work on optimizing the design of such systems is therefore called for.

6. CONCLUSION AND FUTURE WORK

We have built a test collection for zero-resource ranked retrieval of spoken Gujarati content based on spoken Gujarati queries. Our experience has highlighted the difficulty of topic selection when the researchers themselves do not know the language, the practicality of performing relevance judgments directly on the audio, and unusual characteristics of voice forum content that have implications for retrieval system design. Our evaluation of the collection shows a balance between insightfulness, affordability, and reusability that is suitable for formative evaluation of ranked-retrieval methods for the type of zero-resource features provided. The test collection, which is available for research use, has been used in Forum for Information Retrieval Evaluation (FIRE) shared tasks [6, 7].[2]

7. ACKNOWLEDGMENTS

We would like to thank Komal Kamdar, Dhwani Patel, and Yash Patel for performing relevance assessments. This work has been supported in part by NSF award 1218159.

References

[1] T. Akiba et al. Overview of the NTCIR-11 spoken query and doc task. In *NTCIR-11*, 2014.
[2] X. Anguera et al. The spoken web search task. In *MediaEval*, 2013.
[3] P. Comas et al. Sibyl, a factoid question-answering system for spoken documents. *ACM TOIS*, 30(3):19, 2012.
[4] M. Dredze et al. NLP on spoken documents without ASR. In *EMNLP*, 2010.
[5] J. Garofolo et al. The TREC spoken document retrieval track: A success story. In *RIAO*, 2000.
[6] H. Joshi and J. White. Document silmilarity amid automatically detected terms. In *FIRE*, 2014.
[7] D. Oard et al. The FIRE 2013 question answering for the spoken web task. In *FIRE*, 2013.
[8] N. Patel et al. Avaaj Otalo: A field study of an interactive voice forum for small farmers in rural India. In *CHI*, 2010.
[9] J. White et al. Using zero-resource spoken term discovery for ranked retrieval. In *NAACL-HLT*, 2015.
[10] E. Yilmaz et al. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR*, 2008.

[2]http://www.umiacs.umd.edu/∼oard/qasw/