The FIRE 2013 Question Answering for the Spoken Web Task

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

Jerome White
IBM India Research Lab
Bangalore, India
jerome.white@in.ibm.com

Jiaul Paik
University of Maryland
College Park, MD USA
jiaul@umd.edu

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

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

ABSTRACT

Question Answering for the Spoken Web (QASW) is an information retrieval evaluation in which the goal was to match spoken Gujarati questions to spoken Gujarati answers. This paper describes the design of the task, the development of the test collection, the system used by the participating teams, the runs that were submitted, and the corresponding results. This paper thus combines the track overview and the participant results.

1. INTRODUCTION

Question Answering for the Spoken Web (QASW) is an information retrieval evaluation in which the goal was to match questions spoken in Gujarati to answers spoken in Gujarati. The design of the task was motivated by a speech retrieval interaction paradigm first proposed by Oard in which a searcher, using speech for both queries and responses, speaks extensively about what they seek to find until interrupted by the system with a single potential answer [5]. For the 2013 FIRE QASW task, results are reported only for full-length questions, however, so interruption is not yet modeled. QASW was originally conceived as a joint task between the 2013 MediaEval evaluation, which focused on speech retrieval, and the 2013 Forum for Information Retrieval Evaluation (FIRE), which focused on Indian language text retrieval. As explained below, it ultimately evolved into a pilot task only at FIRE, with a focus only on speech retrieval.

2. QUESTIONS AND RESPONSES

The source of the questions and the collection of possible answers (which we call "responses") was the IBM Spoken Web Gujarati collection [7]. This collection was based on a spoken bulletin board system for Gujarati farmers. A farmer could call the system and record their question by going through a set of prompts. Other farmers would call the system to record answers to those questions. 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 or call center was involved. This collection of recorded speech, consisting of questions and responses (answers and announcements) was provided the basis for the test collection. A total of 3,557 answers were provided for specific questions. In some cases, these answers may have applied to more than one question, as the same topics might be asked about more than once. There were a total of 854 announcements.

The reuse of information retrieval test collections that are built using relevance judgments only on top-ranked responses returned by participating systems (i.e., "pooling") requires that future systems return response sets that were not too unlike the response sets produced by participating systems. Retrieval directly from speech requires speech processing components that we would expect to evolve over time, thus posing a threat to test collection reusability. Transcription can avoid this problem by allowing text-only runs to contribute to the judgment pools, so we arranged to have the questions and responses transcribed from spoken Gujarati to written Gujarati by a commercial transcription agency. The transcription agency was unable to deliver transcripts in time for use in MediaEval, which necessitated cancellation of the MediaEval 2013 QASW task.

Transcripts were later received from the transcription agency, but it turned out that only the questions had been transcribed correctly. By the time this was discovered, there was not sufficient time to arrange for correct transcription of the responses. We therefore redesigned the FIRE 2013 QASW task as a pilot study focusing only on speech retrieval, without the intent to create a reusable test collection.

The 151 longest questions were divided into a training set of 50 questions and an evaluation set of 101 questions. Training questions were those for which the largest number of answers were known beforehand (mappings between questions and known answers was available to the organizers from data collected by the operational system). Once the transcripts became available, two evaluation questions were removed for which the resulting transcripts were far shorter than would be expected based on the file length.

FIRE 2013, Delhi, India

This resulted in a total of 50 training questions and 99 evaluation questions. Of these, only the 50 training questions were subsequently used in the 2013 QASW pilot task; the 99 evaluations questions are available for use in future experiments.

We removed very short response files—those four seconds or less. We also learned from listening to some of the responses that some responses were silent and others contained only music. We therefore removed response files for which the transcript was far shorter than would be expected based on the duration of speech activity (for which we performed automatic speech activity detection). The final test collection contains 2,999 responses. We only later learned that the length matching had been based on incorrect transcripts, so it should be possible to expand the response set somewhat if new relevance judgments are made for this collection in the future.

3. SPEECH PROCESSING

In traditional speech retrieval applications, document-level features are derived from the outputs of supervised phonetic or word recognizers. Recent term discovery systems [6, 2] automatically identify repeating words and phrases in large collections of audio, providing an alternative means of extracting lexical features for retrieval tasks. Critically, this discovery is performed without the assistance of any supervised speech tools by instead resorting to a search for repeated trajectories in a suitable acoustic feature space (e.g. MFCCs, PLP) followed by a graph clustering procedure. Due to their sometimes ambiguous content, the discovered units are referred to as *pseudoterms*, and we can represent each question and response as a set of pseudoterm offsets and durations. We used an unsupervised term discovery system that consists of three steps [1]: i) search for repeated trajectories in an acoustic feature space using image processing techniques applied to sparse distance matrices, ii) cluster repetitions into pseudoterm categories; and iii) construct a bag-of-pseudoterms representation for each question and response in the collection. We summarize each step in the subsections below. Complete specifications can be found in the literature [1, 3].

3.1 Acoustic Repetition Search

The QASW collection consists of nearly 100 hours of speech audio. Term discovery is inherently an $O(n^2)$ search problem, and application to a corpus of this size is unprecedented in the literature. We applied the scalable system described by Jansen and Van Durme [3], which employs a coarse-to-fine strategy to achieve a very substantial (orders-of-magnitude) speedup over its predecessor state-of-the-art system [6]. The system functions by constructing a sparse (thresholded) distance matrix across the frames of the entire corpus and then searching for approximately diagonal line structures in that matrix, as such structures are indicative that a word or phrase has been repeated.

Two randomized algorithms are used to efficiently construct the sparse cosine distance matrix. The first is Locality Sensitive Hashing (LSH), which involves mapping each feature vector to a bit string such that the ability to approximate some distance metric in the original space is preserved. We use an LSH variant that preserves cosine distance, which is accomplished by applying a collection of random projections that each encodes membership in a randomly oriented halfspace. It follows that Hamming distances between bit strings can be used to approximate cosine distance, with the approximation approaching equality as the signature length approaches infinity. In our experiments, we used 64-bit signature representations of 39-dimensional short-time frequency domain linear prediction features (see Jansen and Van Durme [3] for details). The second randomized algorithm is Point Location in Equal Balls (PLEB), which is used in conjunction with LSH to find nearest neighbor sets for each frame in logarithmic time. Here, the bit signatures are lexicographically sorted such that nearby frames have some prefix of bits in common. This common prefix implies a bound on the Hamming distance and, in turn, the cosine distance, which means frames nearby in the list are strong candidates for neighbor status.

Once we have constructed the sparse distance matrix, it remains to search for runs of nearby frames. Here, we employ the two pass strategy of Jansen et al. [2] that i) thresholds the distance matrix into a binary image with only nearby frame pairs active, ii) searches for diagonal line segments using sparse image processing techniques, and iii) uses the center point of any recovered line segments to begin a local segmental Dynamic Time Warping (DTW) search. Each repetition is scored by the DTW distance between the matching segments. We also apply a minimum duration threshold of 0.6 seconds, which generally limits the pseudoterms to words or phrases of at least two syllables.

3.2 Clustering Repetitions into Pseudoterms

To cluster the individual acoustic repetitions into pseudoterm categories we apply a simple graph-based procedure. First, we construct an unweighted acoustic similarity graph, where each segment of speech involved in a discovered repetition becomes a vertex, and each match provides an edge. This produces a graph consisting of a set of disconnected dumbbells. Next, we augment the original edge list with the set of overlap edges that indicate whether two nodes correspond to the identical segment of speech in a given question or response. For two segments to be considered the same, we require a minimal fractional overlap of 0.97, which is set less than unity to allow some noise in the unit end points. These additional edges act to effectively merge vertices across the dumbbells and enable transitive matches between acoustic segments that did not match directly. The pseudoterms are defined to be the resulting connected components of the graph, each consisting of a set of acoustic segments occurring throughout the collection.

Since we construct an unweighted graph and employ a simple connected-components clustering, it is essential we apply some DTW distance threshold δ before a repetition is passed along to the clustering procedure. In the experiments described in this paper, we considered three pseudoterm feature variants arising from three settings of the DTW score threshold. Lower thresholds imply higher fidelity matches that yield purer pseudoterm clusters with (on average) lower collection frequencies. We refer to these as Weak clustering ($\delta = 0.06$, yielding 406,366 unique pseudoterms), Medium clustering ($\delta = 0.07$, yielding 1,213,223 unique pseudoterms) and Strong clustering ($\delta = 0.075$, yielding 1,503,169 unique pseudoterms). Weak clustering is precision-biased; strong clustering is recall-biased.

3.3 Nested Pseudoterms

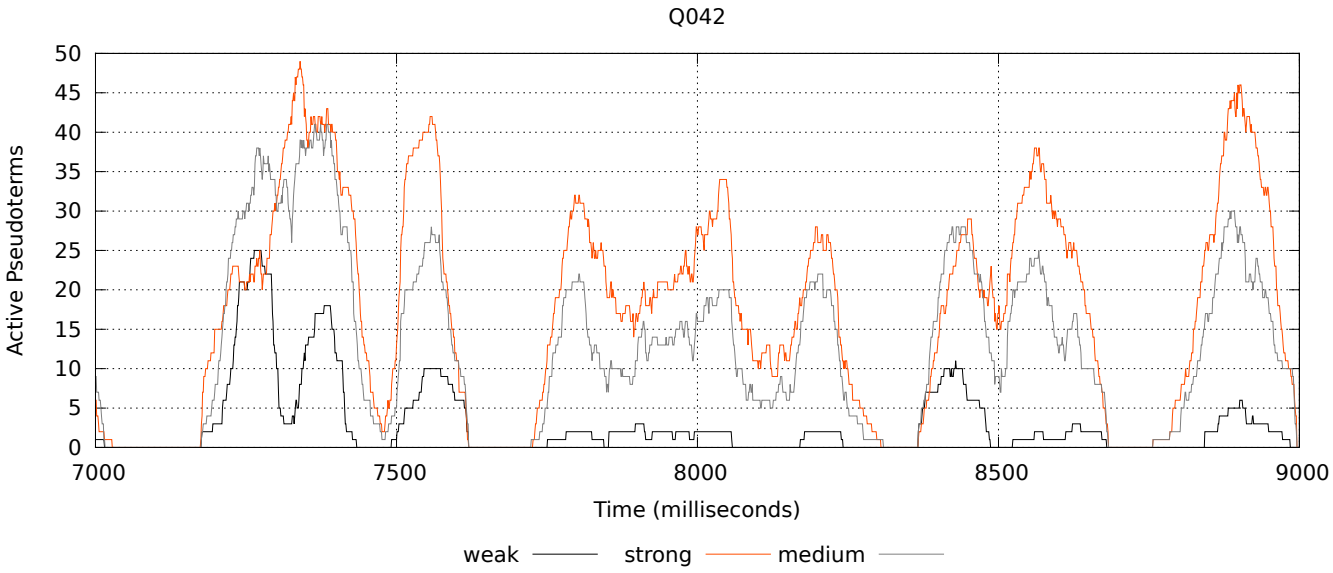

Figure 1: Different pseudo-term nesting structures for various settings of the speech-to-term extraction model. The y-axis represents the number of terms extracted at a given period in time. This figure represents a two-second interval of Question 42.

Each pseudoterm cluster consists of a list of occurrences (start and end offsets, in units of 10 ms, from the beginning of the file). It is thus a simple matter of bookkeeping to construct a bag-of-pseudoterms representation for each question and response. Moreover, because we have start and end offsets for each pseudoterm, we can also construct more sophisticated representations that are based on filtering or grouping the pseudoterms based on the ways in which they overlap temporally.

One interesting effect of pseudoterm creation is that the pseudoterms are often nested, and moreover they are often nested to a depth that has never before been seen in information retrieval experiments. Nesting has previously been explored for phrase indexing, where a longer term (e.g., White House spokesman) contains a shorter term (e.g., White House) that might also be used independently (i.e., without *spokesman*) elsewhere in the collection. Because pseudoterm detection can find any pair of matching regions, we can (by analogy) get not only pseudoterms for *White House Spokesman* and *White House*, but also for parts of those words (e.g., *Whit* and *Whi*). Indeed, it is not at all unusual to see nesting to depth 50, as Figure 1 shows. This is a fairly typical pseudoterm timeline, in which Weak clustering yields deeper nesting than Medium clustering, and much deeper nesting than Strong clustering.

4. EVALUATION DESIGN

The principal task of a participating research team in the FIRE 2013 QASW task was to rank all responses to each full question such that, to the extent possible, all correct answers were ranked ahead of all incorrect answers. Each participating system was asked to rank all responses for all training questions. Systems were evaluated on their ability to satisfy that goal using Normalized Discounted Cumulative Gain (NDCG), based on graded relevance judgments. Other measures (Mean Reciprocal Rank (MRR) of the first relevant response, mean Precision at 5 responses (P@5), mean precision at 10 responses (P@10), and Mean uninterpolated Average Precision (MAP) were also reported. All measures (except P@5) are reported only for response sets truncated at rank 10 because the top 10 responses were judged for all participating systems, and we did not have confidence that responses below rank 10 would be sufficiently densely judged to be useful as a basis for relative comparisons between system results.

5. RESEARCH TEAMS

Interest in participating had been expressed by Gujarat University (India) in participating in the QASW evaluation, and indeed they had obtained the collection before problem with the transcripts was discovered. The compressed timeline for the speech-only pilot task that resulted prevented the Gujarat University team from participating in that pilot task.

All speech processing was performed at Johns Hopkins University (USA), and runs using those speech processing results were submitted by the University of Maryland, College Park (USA) and by the IBM India Research Lab (India). Both of the research teams submitting results included QASW task organizers, so the results reported below should be considered "unofficial," useful as an indication of what can be achieved, but not the result of a formal arms-length evaluation process.

6. FIRST EXPERIMENTS

Our first set of experiments had three goals: (1) to serve as a dry run for system development, since we had no prior experience with indexing or ranked retrieval based on pseudoterms, (2) to gain experience with performing relevance judgments using only the audio responses, and (3) to begin

Measure	Criterion	$\|Q\|$	Clust.	$\Sigma\|RD\|$	BW
P@5	$\|RD\| \geq 1$	14	Weak	56	0.22
		14	Med	56	0.12
		14	Strong	56	0.08
P@10	$\|RD\| \geq 1$	14	Weak	56	0.16
		14	Med	56	0.10
		14	Strong	56	0.11

Table 1: First experiment results. $|RD|$: number of relevant responses per question, $|Q|$: number of questions meeting criterion.

to get a sense for whether pseudoterm-based speech retrieval was feasible.

6.1 Retrieval Methods

For these initial experiments, we simply treated every pseudoterm as a "word" in a Bag of Words (BW) representation using the Indri ranked retrieval system [10].

6.2 Relevance Assessment

Three native speakers of Gujarati at the University of Maryland performed relevance assessment; none of the three had any role in system development. The 10 highest ranked responses for each question from each participating system were pooled for relevance assessment. Relevance assessment was performed by manually selecting audio files from a list of response files that were to be assessed for each question, listening to the audio, and recording the relevance judgment. All responses in the assessment pool for each question were judged by all three assessor. Assessors could assign one of the following judgments for each response: 1) unable to assess, 2) not relevant, 3) relevant, and 4) highly relevant.

All relevance judgments were subsequently binarized by collapsing unassessable and not relevant responses into a single not relevant category and by collapsing highly relevant and relevant responses to a single relevant category. Majority voting was then used to select a consensus binary judgment for each question. Table 2 shows the inter-assessor agreement for each pair of assessors.

After completing their relevance judgments, the assessors met to discuss the assessment process. To facilitate this discussion, specific cases of disagreement were randomly sampled for each assessor pair and those cases of disagreement were used to seed the discussion. Two questions, numbered 5 and 13, were removed permanently from the collection because all three assessors reported that they did not address with any identifiable topic and thus relevance assessment could not be performed. Results for 18 questions were therefore available for analysis. Of those, 14 had one or more known relevant responses.

6.3 Results

Table 1 reports results over the 14 of the 18 questions in the first experiment for which one or more relevant responses were found by some system (comparisons on the four topics for which no system found a relevant response would not be informative). We can see that when used with BW, Weak clustering seems to be doing noticeably better than Medium or Strong clustering, although these apparent differences in averages over just 14 questions did not prove to be statistically significant. Nonetheless, the apparent preference for Weak clustering in these results was surprising because in previous work in other experimental settings stronger clustering has proven to be advantageous.

	Experiment 1			Experiment 2		
	A	B	C	A	B	C
A	-	0.407	0.387	-	0.279	0.317
B		-	0.617		-	0.867
C			-			-

Table 2: Inter-assessor agreement (kappa).

7. SECOND EXPERIMENTS

It is well known from earlier work on phrase indexing that indexing nested terms using bag of words models requires careful attention to the relative weights of shorter and longer terms [4]. The design of our initial experiments did not account for effects from pseudoterm nesting, and we therefore suspected that the tendency of Weak clustering to produce markedly shallower nesting levels might help to explain the observed results. We therefore adopted BW as a baseline for our second set of experiments and we focused our design of retrieval models on techniques that account for this effect. For this second set of experiments we increased the number of questions to 48 (numbered 1–50, excluding questions 5 and 13). Because we had previously examined results for questions 1–20, we report results separately for that question set and for questions 21–50 in the Appendix. In this section, however, we report results over the full question set.

7.1 Retrieval Methods

In order to address the nested pseudoterm issue we find the nested groups of terms in the questions and in the spoken answers (in the pseudoterm space) using the associated start and finish time tags for each pseudoterm that have been passed by the pseudoterm discovery algorithm of the previous section. We then treat each of these groups as an atomic unit.

7.1.1 Structured Query Models

Our first family of methods treats terms in each overlapping group as synonymous, an approach that has come to be called "structured queries." There are two methods we explore in dealing with such cases. The first is to treat all overlapping pseudoterms as synonyms of a single term. This is accomplished in Indri by placing each pseudoterm within an overlapping region within the `syn` operator. We refer to this as our Synonym Operator (SO) retrieval model.

One risk with our SO model is that including shorter terms may add more noise than signal. An alternative method of dealing with overlapping pseudoterms is to only use only one term term from the set. For our experiments with this technique, we chose to keep only the longest pseudoterm from each set of overlapping; all nested pseudoterms are simply deleted from the Indri query. We refer to this as our Longest Structured (LS) retrieval model.

7.1.2 Weighted Structured Query Models

Our SO and LS models represent opposite extremes on a

Measure	Criterion	Clust.	$\|Q\|$	$\Sigma\|RD\|$	BW	TW	TWD	SWD	LWD	SO	LS	Sig-Test
NDCG	$\|RD\| > 0$	Weak	38	178	0.099	0.114	0.109	0.147	0.147	0.138	**0.151**	$LS \not> BW$
		Med	38	178	0.054	0.053	0.050	0.063	0.060	0.102	0.109	
		Strong	38	178	0.064	0.069	0.067	0.076	0.075	0.114	0.132	
MRR	$\|RD\| > 0$	Weak	38	178	0.176	0.236	0.188	0.258	**0.282**	0.211	0.274	$LWD > BW$
		Med	38	178	0.088	0.100	0.083	0.116	0.133	0.235	0.167	
		Strong	38	178	0.096	0.112	0.117	0.122	0.125	0.191	0.206	
MAP	$\|RD\| \geq 3$	Weak	26	159	0.077	0.091	0.088	0.099	**0.112**	0.064	0.086	$LWD \not> BW$
		Med	26	159	0.024	0.027	0.024	0.033	0.040	0.079	0.042	
		Strong	26	159	0.021	0.024	0.035	0.027	0.033	0.071	0.093	
P@5	$\|RD\| \geq 5$	Weak	18	133	0.133	0.156	0.156	0.167	**0.189**	0.078	0.156	$LWD \not> BW$
		Med	18	133	0.056	0.067	0.056	0.078	0.111	0.167	0.089	
		Strong	18	133	0.044	0.056	0.078	0.067	0.067	0.122	0.122	
P@10	$\|RD\| \geq 5$	Weak	18	133	0.117	0.106	0.128	0.133	**0.167**	0.117	0.128	$LWD > BW$
		Med	18	133	0.078	0.078	0.083	0.078	0.072	0.128	0.089	
		Strong	18	133	0.061	0.056	0.072	0.067	0.072	0.078	0.128	

Table 3: Second set of experiment results. $|RD|$: number of relevant responses per question. Sig-Test: two-sided paired t-test results at 95% confidence level; $A > B$ (or $A \not> B$) means that A is significantly better than (or comparable to) B.

continuum between treating nested pseudoterms equally or ignoring them completely. Indri's `wsyn` operator offers some scope for a middle ground, however. Our second family of methods employs a term weight discounting model. Two intuitions motivate the design of this model. First, in prior work, it has been found that applying a somewhat longer minimum threshold on pseudoterm length can be helpful. As a soft way of doing this, we can treat the length of a pseudoterm as an evidence of importance. Second, it in prior work it has been shown that when both multiword phrases and the constituent terms are indexed, their weights should be adjusted in some way to reflect for overlap. We can implement this insight by adjusting the contribution of each pseudoterm based on the extent of its overlap with other pseudoterms. We could do this in a way that would give the greatest weight to either the shortest or the longest nested pseudoterm. Let $T = \{t_1, t_2 \ldots t_n\}$ be the nested term class in ascending order of term length. The weight of the term t_i is calculated as follows:

$$W(t_i, l) = \frac{a \times l}{1 + (a \times l)} \quad (1)$$

$$WD(t_i, l) = W(t_i, l) \times \prod_{j=0}^{i-1} (1 - W(t_j, l)) \quad (2)$$

where a is the free parameter and l is the length of the term (in seconds). For our experiments we simply set a to 0.5. The factor $(1 - W(t_i, l))$ is used to discount the weight of t_i due to the contribution made by the previous terms. We call this model, which gives the greatest weight to the shortest pseudoterm, Shortest Weight Discounted (SWD). Likewise, by reversing the order of T (i.e., in descending order) we get the Longest Weight Discounted (LWD) model.

7.1.3 *Unstructured Query Models*

Structured query models have proven to be effective in several applications (e.g., cross-language retrieval and document image retrieval), but at some additional cost in implementation complexity. In the simpler unstructured query models, every term is treated independently. To see if the added complexity results in improved results, we also applied our discounting model in an unstructured query model by simply using the `weight` operator in place of the synonym operator, with the weights again computed using equation 2. For the one variant of this idea that we implemented we give the longest pseudoterms the greatest weight by computing this in same order as LWD. We call this model Total Weight Discounted (TWD).

An even simpler alternative would be to simply use equation 2 to adjust the weight of each term without reference to nesting. We refer to this retrieval model as Time Weighted (TW) and implement it using Indri's `weight` operator.

Finally, our baseline BW model, used in both sets of experiments, omits all adjustments, simply counting the occurrences of each pseudoterm without taking account of its length or nesting.

7.2 Relevance Assessment

The 10 highest ranked responses for each question from each participating system were pooled for relevance assessment. Three participating systems (BW, SO and LS; see below) additionally submitted intermediate results after each 5 seconds of question duration (e.g., at 5 seconds, 10 seconds, ...) and the top-ranked response from each of those intermediate results was included in the judgment pools. For questions 43-50, 10 randomly selected responses were also added to the assessment pools.

Assessor A judged the pools for questions 1–15, Assessor B judged the pools for questions 16–30, Assessor C judged the pools for questions 21–45, and all assessors independently judged the pools for questions 46–50. As Table 2 shows, assessors B and C exhibited remarkably strong inter-assessor agreement on the five multiply-judged questions. The mean kappa value for the second experiment of 0.488 is well in line with commonly reported values of inter-assessor agreement that result in little adverse effect on relative comparisons in

ranked retrieval experiments [9].

Final relevance assessments were computed by taking the single assessor's judgment for questions 1–45, and by voting as described above for questions 46–50. There were no cases where the three assessors gave three different assessments, so voting always produced a majority, even with graded relevance judgments.

7.3 Results

In this section we summarize the results of our second set of experiments. We used the same Medium and Strong clustering conditions as in the first set of experiments, but we generated new Weak clustering results because time offsets had not been recorded for the original set of Weak clustering results. This resulted in an increase from 406,366 to 407,514 unique pseudoterms for Weak clustering.

Table 3 summarizes the performance of our different retrieval models with different clustering methods, as characterized by different evaluation measures. MRR is an easily interpreted measure (the reciprocal of MRR is the harmonic mean of the rank of the first relevant response), but MRR is known to exhibit significant quantization noise (e.g., because the difference between the first relevant response at rank 1 and rank 2 for a single question changes that questions contribution to the average score by 0.5) [8]. NDCG provides a more nuanced view of the differences, both because that measure uses graded relevance judgments (our other measures are based on relevance judgments that are binarized as described above) and because the discounting rate is lower for NDCG than for MRR.[1] Both NDCG and MRR are averaged over all questions for which at least one relevant response is known to exist. MAP, by contrast, suffers from the same sharp discounting rate of MRR, so to minimize quantization noise effects in that case we report MAP results averaged across questions for which at least 3 relevant responses are known to exist. P@5 and P@10 are easily interpretable measures, but when fewer relevant responses than the cutoff are known to exist they do not distinguish well between systems. We therefore average those measures across only questions for which at least five relevant responses are known to exist (there are too few questions with 10 or more relevant responses to make averaging over that few questions an insightful option).

Several outcomes are evident from Table 3. We see that weak clustering consistently yields better numerical results than the other two clustering options that we tried, irrespective of the retrieval model or the evaluation measure. With weak clustering the LWD model gives the best numerical results by most measures (and very close to the best by all measures). With Medium clustering, SO numerically outperforms LWD (and all other retrieval models) by all measures. With Strong clustering, the LS model is consistently preferred (or, in one case, the equal of SO).

Statistical significance testing using a two-sided paired t-test at a 95% confidence level is generally inconclusive, with tests on only two measures detecting an improvement of LWD over BW and no statistical significant test detecting an improvement from Weak clustering over Strong clustering (when the numerically best approach is used for each).

From these observations we can draw the following conclusions. First, we do not yet seem to have relevance judgments for enough questions to reliably see statistically significant differences. This is consistent with prior results that show that 40 or more queries are typically needed to see statistically significant results with moderate effect sizes. Second, the consistent dominance of Weak clustering (which yields far less pseudoterm nesting) and the relatively strong (numerical) performance of LWD, LS and SO (all of which attempt to compensate for the effects of pseudoterm nesting), suggests that pseudoterm nesting is indeed an important factor. Third, the relatively strong (numerical) performance of LWD and LS, both of which exhibit a strong bias in favor of longer pseudoterms, tends to confirm earlier results that indicate that very short pseudoterms may introduce more noise than signal. We note, however, that the relatively good results from SO with Medium clustering is not consistent with that interpretation, which suggests that caution is called for when interpreting differences that have not been shown to be statistically significant. Finally, we note that 10 of the 48 questions had no relevant responses found. Since every question in the collection has at least one known response (according to the operational systems from which the questions and responses were extracted), this suggests that the values of the measures that we report for comparison purposes are somewhat optimistic estimates of what can be accomplished on average over the full set or questions that might be asked. Moreover, our best run by MRR (LWD with Weak clustering) only placed a relevant response in first rank 7 of the 48 full-length questions, which suggests that substantial improvement will be needed before operational systems will be able to use pseudoterms as a basis for responding effectively to incomplete questions with a single answer, as envisioned in the "query by babbling" interaction design.

8. RANDOM BASELINE COMPARISON

Pooled relevance assessment can't tell us how many relevant responses actually exist for each question, but we can use random sampling to determine that. More specifically, we can compute a low baseline for any evaluation measure by randomly sampling the test collection. We did this for questions 43–50. As Table 4 shows, our BW baseline (and indeed every retrieval model that we tried) very substantially outperforms the resulting random baseline. These results are averaged over 6 questions (because two questions in this set have no known relevant responses) and are shown only for Weak clustering (which yielded the best overall results). From this we can conclude that speech retrieval based on automatically detected pseudoterms actually works.

9. CONCLUSION AND FUTURE WORK

Despite the challenges that we encountered with development of the test collection, we have made substantial progress on a number of important problems. Most notably, we have demonstrated that pseudoterms can be generated in a language-independent manner for a hundred-hour speech collection, and that the resulting pseudoterms can be used as a basis for effective ranked retrieval. Additionally, our experiment results suggest that nested pseudoterms pose considerable challenges, but that some combination of constrained

[1]Our NDCG computation was subtly different from our computation of other measures. For NDCG, unassessable responses were removed from all ranked lists before scoring. For all other measures, unassessable responses were treated as not relevant, as described above.

Measure	$\|Q\|$	RAND	BW	TW	TWD	SWD	LWD	SO	LS
NDCG	5	0.038	0.354	0.396	0.423	0.441	**0.456**	0.236	0.336
MRR	5	0.090	0.527	0.708	0.687	0.700	0.740	0.403	**0.767**
MAP	5	0.011	0.192	0.232	0.262	0.286	**0.309**	0.149	0.202
P@5	5	0.080	0.240	0.280	0.320	0.320	**0.360**	0.160	0.240
P@10	5	0.040	0.240	0.200	0.260	0.260	**0.280**	0.140	0.180

Table 4: Comparison with random baseline, Weak clustering. RAND: random baseline.

generation of pseudoterm clusters and filtering or weighting the resulting pseudoterms based at least in part on length offers some promise for improving retrieval effectiveness. A third clear result is a test collection with 48 questions, 2,999 responses, and several thousand relevance judgments that is available to interested parties on a license that makes it freely usable for research purposes.

Of course, much remains to be done. Our most urgent task will be to actually characterize the extent to which the test collection is reusable. For this purpose, we plan to leverage the known responses to each question as a probe to see what fraction of those responses have been discovered through pooling and relevance assessment. We also plan to plot the decay in the fraction of assessed responses for the runs in our second experiment below rank 10, and for a few new runs with different filtering or weighting designs at all ranks. If, as we expect, reusability turns out to be limited, then transcription of the responses will become our highest priority. We also plan to compute a P@1 measure for the runs that were submitted at 5-second intervals on partial questions in order to compare those results with earlier traces we have seen for the temporal evolution of that measure in a simulation study. Of course, much remains to be done on the question of how best to leverage evidence from nested pseudoterms and on the question of how evidence from multiple ways of doing that might best be combined, but until we have a reusable test collection we will not be in a position to efficiently try very many alternative designs.

10. ACKNOWLEDGMENTS

We are grateful to Nitendra Rajput for providing the spoken questions and responses and for early discussions about evaluation design, to Komal Kamdar, Dhwani Patel, and Yash Patel for performing the relevance assessments, to Apruva Bhatt and Hardik Joshi for their assistance with detection of the transcription problem, and to Kundan Shrivastava and Joe Webster for their help with building and installing the relevance assessment system that would have been used with those transcripts. This work has been supported in part by DARPA contract HR0011-12-C-0015 and by NSF award 1219130.

11. REFERENCES

[1] M. Dredze, A. Jansen, G. Coppersmith, and K. Church. NLP on spoken documents without asr. In *Proc. EMNLP*, pages 460–470. Association for Computational Linguistics, 2010.

[2] A. Jansen, K. Church, and H. Hermansky. Towards spoken term discovery at scale with zero resources. In *INTERSPEECH*, pages 1676–1679, 2010.

[3] A. Jansen and B. Van Durme. Efficient spoken term discovery using randomized algorithms. In *Proc. ASRU*, 2011.

[4] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proc. RIAO*, pages 200–217, 1997.

[5] D. W. Oard. Query by babbling: A research agenda. In *Proceedings of the First Workshop on Information and Knowledge Management for Developing Regions*, IKM4DR '12, pages 17–22, 2012.

[6] A. Park and J. R. Glass. Unsupervised pattern discovery in speech. *IEEE T-ASLP*, 16(1):186–197, 2008.

[7] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj Otalo: A field study of an interactive voice forum for small farmers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 733–742. ACM, 2010.

[8] S. Tomlinson. Measuring robustness with first relevant score in the TREC 2012 microblog track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, 2012.

[9] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[10] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

APPENDIX

Table 5: Results on Weak pseudoterms, cutoff at rank 10 (except P@5).

	# rel doc > 0							# rel doc > 2						
	Quest.	BW	TW	TWD	SWD	LWD	SO	LS	BW	TW	TWD	SWD	LWD	SO	LS
MAP	1-20	0.059	0.056	0.056	0.070	0.066	0.048	0.052	0.077	0.066	0.073	0.072	0.086	0.050	0.069
	21-50	0.059	0.088	0.079	0.097	0.109	0.069	0.109	0.077	0.116	0.104	0.126	0.138	0.079	0.104
	All	0.053	0.065	0.061	0.083	0.080	0.063	0.076	0.077	0.091	0.088	0.099	0.112	0.064	0.086
P@5	1-20	0.106	0.118	0.118	0.129	0.106	0.059	0.129	0.138	0.138	0.154	0.154	0.138	0.062	0.169
	21-50	0.071	0.106	0.094	0.118	0.141	0.071	0.129	0.092	0.138	0.123	0.154	0.185	0.092	0.138
	All	0.079	0.100	0.095	0.116	0.111	0.068	0.121	0.115	0.138	0.138	0.154	0.162	0.077	0.154
P@10	1-20	0.076	0.071	0.082	0.088	0.100	0.088	0.088	0.100	0.085	0.108	0.108	0.131	0.108	0.115
	21-50	0.082	0.071	0.088	0.088	0.112	0.076	0.106	0.108	0.092	0.115	0.115	0.138	0.092	0.115
	All	0.071	0.063	0.076	0.084	0.097	0.079	0.089	0.104	0.088	0.112	0.112	0.135	0.100	0.115
MRR	1-20	0.215	0.190	0.193	0.241	0.291	0.182	0.243	0.278	0.231	0.249	0.275	0.376	0.211	0.317
	21-50	0.178	0.335	0.226	0.299	0.329	0.241	0.348	0.224	0.430	0.288	0.379	0.415	0.295	0.414
	All	0.176	0.236	0.188	0.258	0.282	0.211	0.274	0.251	0.330	0.268	0.327	0.396	0.253	0.366
NDCG	1-20	0.107	0.103	0.108	0.129	0.137	0.125	0.130	0.140	0.117	0.141	0.140	0.179	0.134	0.170
	21-50	0.115	0.152	0.136	0.158	0.180	0.072	0.196	0.150	0.198	0.177	0.206	0.227	0.069	0.193
	All	0.099	0.114	0.109	0.147	0.147	0.108	0.151	0.145	0.158	0.159	0.173	0.203	0.101	0.182

Table 6: Results on Medium pseudoterms, cutoff at rank 10 (except P@5).

	# rel doc > 0							# rel doc > 2						
	Quest.	BW	TW	TWD	SWD	LWD	SO	LS	BW	TW	TWD	SWD	LWD	SO	LS
MAP	1-20	0.032	0.026	0.030	0.035	0.048	0.041	0.037	0.029	0.024	0.032	0.029	0.050	0.054	0.034
	21-50	0.014	0.023	0.012	0.028	0.023	0.080	0.047	0.019	0.031	0.016	0.037	0.030	0.104	0.049
	All	0.021	0.022	0.019	0.029	0.032	0.054	0.053	0.024	0.027	0.024	0.033	0.040	0.079	0.042
P@5	1-20	0.047	0.059	0.035	0.071	0.106	0.082	0.082	0.046	0.062	0.031	0.062	0.123	0.108	0.092
	21-50	0.024	0.024	0.035	0.035	0.035	0.094	0.059	0.031	0.031	0.046	0.046	0.046	0.123	0.077
	All	0.032	0.037	0.032	0.047	0.063	0.079	0.068	0.038	0.046	0.038	0.054	0.085	0.115	0.085
P@10	1-20	0.071	0.059	0.076	0.065	0.071	0.071	0.076	0.085	0.069	0.092	0.069	0.085	0.092	0.085
	21-50	0.035	0.041	0.029	0.041	0.024	0.076	0.071	0.046	0.054	0.038	0.054	0.031	0.100	0.085
	All	0.047	0.045	0.047	0.047	0.042	0.066	0.071	0.065	0.062	0.065	0.062	0.058	0.096	0.085
MRR	1-20	0.109	0.102	0.096	0.129	0.184	0.189	0.140	0.109	0.106	0.100	0.130	0.205	0.238	0.152
	21-50	0.087	0.121	0.088	0.129	0.114	0.334	0.166	0.109	0.155	0.112	0.165	0.144	0.432	0.201
	All	0.088	0.100	0.083	0.116	0.133	0.235	0.167	0.109	0.131	0.106	0.148	0.175	0.335	0.177
NDCG	1-20	0.073	0.061	0.076	0.082	0.098	0.083	0.100	0.080	0.067	0.087	0.076	0.113	0.109	0.095
	21-50	0.047	0.057	0.037	0.060	0.036	0.145	0.108	0.061	0.074	0.048	0.078	0.047	0.190	0.113
	All	0.054	0.053	0.050	0.063	0.060	0.102	0.109	0.071	0.071	0.068	0.077	0.080	0.149	0.104

Table 7: Results on Strong pseudoterms, cutoff at rank 10 (except P@5).

	# rel doc > 0							# rel doc > 2						
	Quest.	BW	TW	TWD	SWD	LWD	SO	LS	BW	TW	TWD	SWD	LWD	SO	LS
MAP	1-20	0.023	0.022	0.051	0.026	0.028	0.005	0.050	0.020	0.019	0.048	0.021	0.026	0.006	0.066
	21-50	0.016	0.023	0.018	0.025	0.031	0.104	0.093	0.021	0.029	0.023	0.033	0.040	0.136	0.121
	All	0.026	0.033	0.031	0.036	0.031	0.077	0.069	0.021	0.024	0.035	0.027	0.033	0.071	0.093
P@5	1-20	0.035	0.035	0.082	0.059	0.047	0.024	0.071	0.046	0.046	0.092	0.062	0.062	0.031	0.092
	21-50	0.012	0.035	0.024	0.035	0.035	0.129	0.106	0.015	0.046	0.031	0.046	0.046	0.169	0.138
	All	0.026	0.037	0.047	0.047	0.042	0.074	0.084	0.031	0.046	0.062	0.054	0.054	0.100	0.115
P@10	1-20	0.065	0.053	0.076	0.065	0.076	0.018	0.094	0.069	0.054	0.085	0.069	0.085	0.023	0.123
	21-50	0.018	0.024	0.018	0.024	0.024	0.076	0.088	0.023	0.031	0.023	0.031	0.031	0.100	0.115
	All	0.039	0.037	0.042	0.042	0.047	0.047	0.087	0.046	0.042	0.054	0.050	0.058	0.062	0.119
MRR	1-20	0.102	0.097	0.175	0.109	0.108	0.075	0.172	0.114	0.106	0.195	0.114	0.118	0.090	0.225
	21-50	0.094	0.122	0.086	0.133	0.159	0.284	0.264	0.122	0.159	0.112	0.173	0.208	0.371	0.346
	All	0.096	0.112	0.117	0.122	0.125	0.191	0.206	0.118	0.133	0.154	0.144	0.163	0.230	0.285
NDCG	1-20	0.073	0.065	0.113	0.076	0.084	0.020	0.122	0.071	0.061	0.110	0.071	0.086	0.026	0.160
	21-50	0.042	0.053	0.037	0.056	0.061	0.164	0.152	0.055	0.069	0.049	0.073	0.080	0.214	0.199
	All	0.064	0.069	0.067	0.076	0.075	0.114	0.132	0.063	0.065	0.080	0.072	0.083	0.120	0.179