Using Zero-Resource Spoken Term Discovery for Ranked Retrieval

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

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

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
John Hopkins HLTCOE
Baltimore, MD USA
aren@jhu.edu

Abstract

Research on ranked retrieval of spoken content has assumed the existence of some automated (word or phonetic) transcription. Recently, however, methods have been demonstrated for matching spoken terms to spoken content without the need for language-tuned transcription. This paper describes the first application of such techniques to ranked retrieval, evaluated using a newly created test collection. Both the queries and the collection to be searched are based on Gujarati produced naturally by native speakers; relevance assessment was performed by other native speakers of Gujarati. Ranked retrieval is based on fast acoustic matching that identifies a deeply nested set of matching speech regions, coupled with ways of combining evidence from those matching regions. Results indicate that the resulting ranked lists may be useful for some practical similarity-based ranking tasks.

1 Introduction

Despite new methods of interaction, speech continues to be a dominant modality for information exchange, particularly among the half of the world's almost five billion mobile phone users who currently lack text-based Internet access. Recording speech poses no particular problems, but retrieval of spoken content using spoken queries is presently available only for the approximately two dozen languages in which there is an established path to market; English, German, or Chinese, for example. However, many of the mobile-only users who could benefit most from such systems speak only one of the several hundred other languages that each have at least a million speakers;[1] Balochi, Mossi or Quechua, for example. Addressing this challenge in a scalable manner requires an integration of speech processing and information retrieval techniques that can be effectively and affordably extended to a large number of languages.

To this end, the experiments in this paper were conducted in a conventional ranked retrieval framework consisting of spoken queries, spoken "documents" (*responses*, hereafter), graded relevance judgments, and standard evaluation measures. As with other information retrieval tasks, there is an element of uncertainty in our best representations of what was said. Our focus on speech processing techniques that are language-agnostic creates the potential for explosive growth in the uncertainty that our search techniques must accommodate. The design and evaluation of such techniques is therefore the central focus of the work explored in this paper.

Our results are both heartening and disconcerting. On the positive side, useful responses can often be found. As one measure of success, we show that a Mean Reciprocal Rank near 0.5 can be achieved when more than one relevant response exists; this corresponds to a relevant response appearing in the second position of a ranked list, on average (by the harmonic mean). On the negative side, the zero-resource speech processing technique that we rely on to generate indexing terms has quadratic time complexity, making even the hundred-hour scale of

[1] There are 393 languages with at least one million speakers according to Ethnologue.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 588–597,
Denver, Colorado, May 31 – June 5, 2015.

the collection on which we have run our experiments computationally strenuous. We believe, however, that by demonstrating the utility of the techniques introduced in this paper we can help to motivate further work on even more affordable scalable language-agnostic techniques for generating indexable terms from speech.

2 Motivation and Related Work

Extending spoken language processing to low-resource languages has been a longstanding goal of the Spoken Web Search task of MediaEval. In this task, research teams are challenged to identify instances of specific spoken terms that are provided as queries in a few hours of speech. Between 2011 and 2013, the task was run three times on a total of 16 different languages (Rajput and Metze, 2011; Metze et al., 2012; Anguera et al., 2013).[2] Two broad classes of techniques over this span proved to be practical: one based on phonetic recognition followed by phonetic matching; the other based on direct matching of acoustic features. Of the two approaches, phonetic recognition was, at the time, slightly more accurate. Directly matching acoustic features, the focus of this paper, potentially offers easier extensibility to additional languages.

From the perspective of information retrieval, the principal limitation of the "spoken term detection" design of the MediaEval task was the restriction to single-term queries. While single-term queries are common in Web search (Spink et al., 2001), the best reported Actual Term Weighted Value (ATWV) from any MediaEval Spoken Web Search participant was 0.4846 (Abad and Astudillo, 2012). This corresponds to a system that correctly detects 48 per cent of all instances of the spoken query terms, while producing at most ten false alarms for every missed detection (Fiscus et al., 2007). Thus, if users are willing to tolerate low precision, moderate levels of recall are possible. Speech search arguably demands higher precision than does Web search, however, since browsing multiple alternatives is easier in text than in speech. One way of potentially improving retrieval performance is to encourage a searcher to speak at length about what they are looking for (Oard, 2012). Such an approach, however, introduces the new challenge of properly leveraging the additional matching potential of verbose multi-term queries (White et al., 2013).

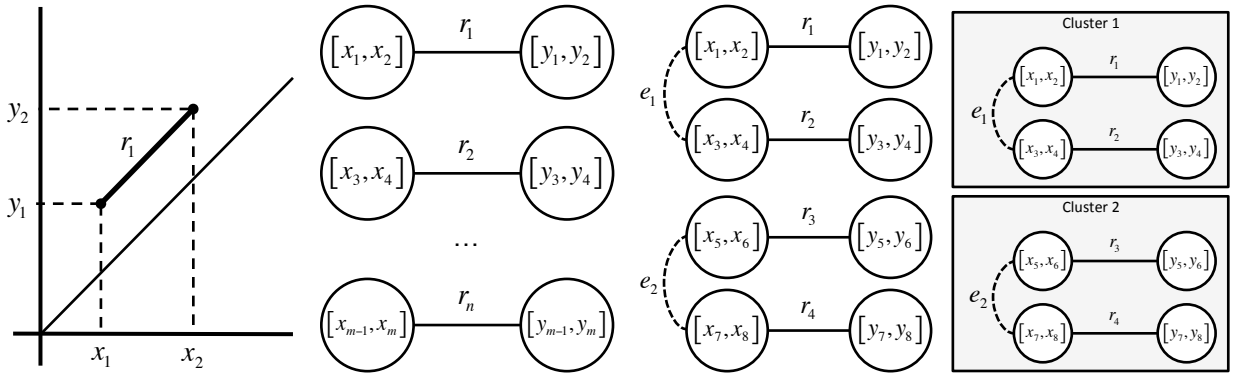
To this end, our work builds on two components: a term matching system, and a test collection. As a term matching system, we used our zero-knowledge speech matching system. In MediaEval 2012, this system achieved an ATWV of 0.321 in the Spoken Web Search task (Jansen et al., 2012). A version of this system has previously been evaluated in an example-based topic classification task using English speech, achieving a classification accuracy of 0.8683 (Drezde et al., 2010). Ranked retrieval using naturally occurring queries is more challenging, however, both because topics in information retrieval are often not easily separable, and because the form of a query may be unlike the form of the responses that are sought. Our goal now, therefore, is to use an information retrieval evaluation framework to drive the development of robust techniques for accommodating representational uncertainty.

Traditional spoken term detection (STD) tries to address uncertainty by learning speech-signal to language-model mappings; using neural networks (Cui et al., 2013; Gales et al., 2014) or Markov models (Chan et al., 2013), for example. From a broad perspective, the method utilized in our work does not use an acoustic model for its analysis. More fundamentally, however, speech signals in our collection map to dozens of smaller terms that are not necessarily the same across utterances of the same word. Thus, it is more accurate to think of the work herein as matching signal features rather than linguistic features.

For this reason, widely used techniques such as stemming, spelling correction, and stopword removal that rely to some extent on linguistic features do not apply in our setting. We therefore rely on term and corpus statistics. Even here there are limitations, since our lexical items are not easily aligned with those found in other collections. For this reason, we can not leverage external corpus statistics from, for example, Google or Wikipedia (Bendersky et al., 2011; Bendersky et al., 2010; Bendersky and Croft, 2008; Lease, 2009), or phrases from search logs (Svore et al., 2010).

Evaluation of ranked retrieval for spoken content

[2] For example, Gujarati, isiNdebele, isiXhosa, Sepedi, Setswana, Telugu, Tshivenda, and Xitsonga.



(a) Term discovery. (b) Term extraction. (c) Term overlap. (d) Term clustering.

Figure 1: Overview of the pseudo-term creation process. The term discovery system is run over the audio. A threshold, δ, dictates the acceptable length, r, and thus the number of regions extracted. Extracted regions are then made into a graph structure, where vertices are regions of speech, and edges denote a connection between those regions. A second edge set is added based on region overlap. Resulting connected components are then clustered; these clusters are known as *pseudo-terms*.

in low-resource languages has to date been hampered by a lack of suitable test collections. We have therefore made our new test collection freely available for research use in recent shared-task information retrieval evaluations (Oard et al., 2013; Joshi and White, 2014).

3 Zero-Resource Term Discovery

In traditional speech retrieval applications, document-level features are derived from the outputs of supervised phonetic or word recognizers. Recent term discovery systems automatically identify repeating words and phrases in large collections of audio (Park and Glass, 2008; Jansen et al., 2010), 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 (for example, Mel Frequency Cepstrum Coefficients (MFCC) and Perceptual Linear Prediction (PLP)) followed by a graph clustering procedure. We refer to the discovered units as *pseudo-terms* (by analogy to the terms built from character sequences that are commonly used in text retrieval), and we can represent each query and response as a set of pseudo-term offsets and durations. We summarize each step in the subsections below. Complete specifications can be found in the literature (Drezde et al., 2010; Jansen and Van Durme, 2011).

3.1 Repetition and Clustering

Our test 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 (2011), which employs a pure-to-noisy strategy to achieve a very substantial (orders-of-magnitude) speedup over its predecessor state-of-the-art system (Park and Glass, 2008). 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 (Figure 1a).

To cluster the individual acoustic repetitions into pseudo-term 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 (Figure 1b). Since we construct an unweighted graph and employ a simple connected-components clustering, it is es-

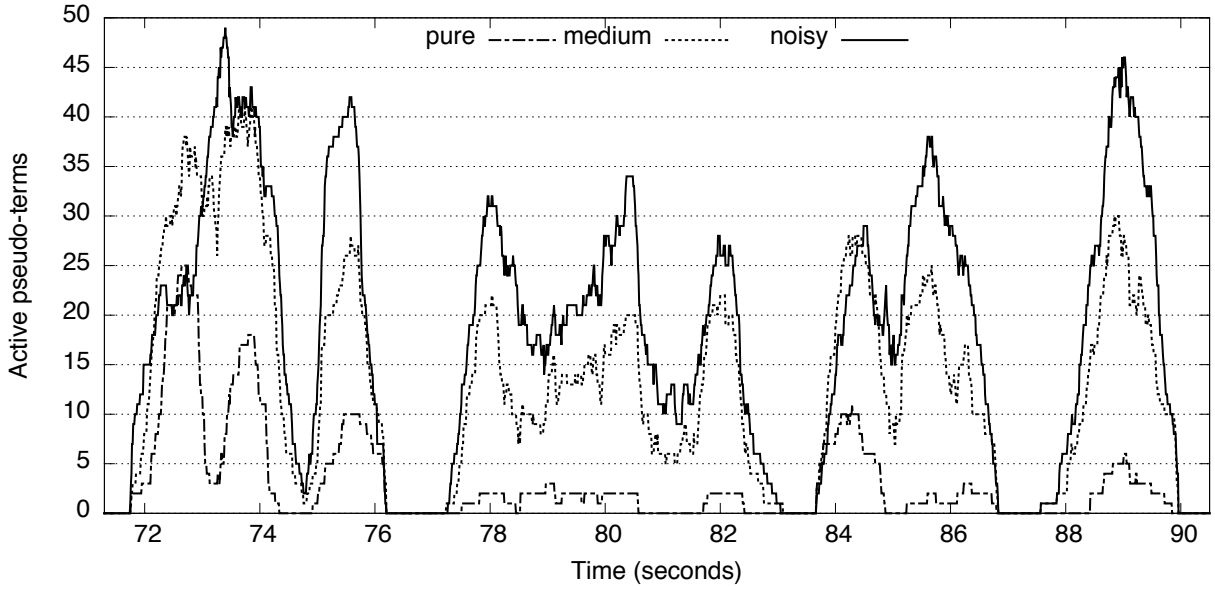


Figure 2: Different pseudo-term nesting structures for various settings of the speech-to-term extraction model. The y-axis represents the number of terms extracted at a given period in time. This figure represents an approximately twenty second interval of Query 42.

sential some DTW distance threshold δ is applied before a repetition is passed along to the clustering procedure. This produces a graph consisting of a set of disconnected "dumbbells."

Finally, the original edge list is augmented with a set of "overlap" edges between corresponding nodes in different dumbbells (Figure 1c); these overlap edges indicate that two nodes correspond to essentially the same segment of speech. For two nodes (two segments of speech) to be considered essentially the same, we require a minimal fractional overlap of 0.97, which is set less than unity to allow some noise in the segment end points. These overlap edges act to effectively merge vertexes across the dumbbells, enabling transitive matches between acoustic segments that did not match directly. The pseudo-terms are defined to be the resulting connected components of the graph, each consisting of a set of corresponding acoustic segments that can occur anywhere in the collection (Figure 1d).

In the experiments described in this paper, three pseudo-term feature variants arising from three settings of the DTW distance threshold are considered. Lower thresholds imply higher fidelity matches that yield fewer and purer pseudo-term clusters. These are referred to as *pure clustering* ($\delta = 0.06$, producing 406,366 unique pseudo-terms), *medium clustering* ($\delta = 0.07$, producing 1,213,223 unique pseudo-terms) and *noisy clustering* ($\delta = 0.075$, producing 1,503,169 unique pseudo-terms).

3.2 Nested Pseudo-Terms

Each pseudo-term cluster consists of a list of occurrences. A term is denoted using start and end offsets, in units of 10 milliseconds, from the beginning of the file. It is thus a simple matter of bookkeeping to construct a bag-of-pseudo-terms representation for each query and response. Moreover, because we have start and end offsets for each pseudo-term, we can also construct more sophisticated representations that are based on filtering or grouping the pseudo-terms based on the ways in which they overlap temporally.

One interesting effect of pseudo-term creation is that the pseudo-terms are often "nested," and they are often nested quite deeply. This sort of nesting has previously been explored for phrase indexing, where a longer term contains a shorter term that might also be used independently elsewhere in the collection. As an English text analogy, if we index "White House spokesman" we might well also want to index "White House" and "spokesman"

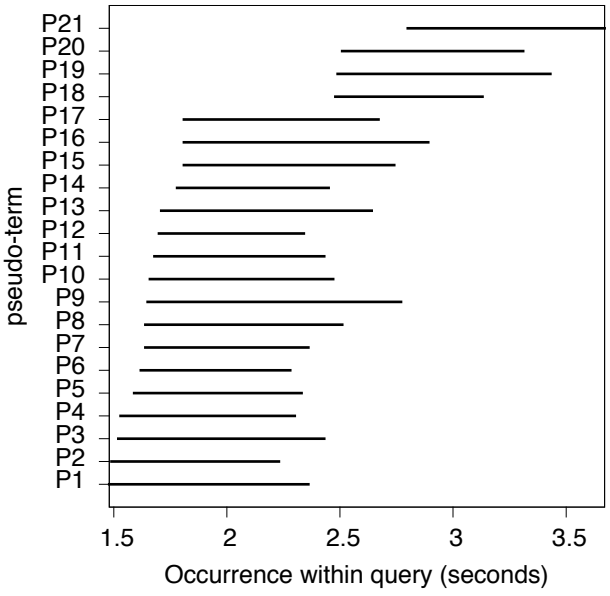


Figure 3: Example of overlapping pseudo-terms within Query 42 under medium clustering. Terms are presented as horizontal bars denoting their start and end time.

separately to support partial matching. Because pseudo-term detection can find any pair of matching regions, we could, continuing the analogy, not only get pseudo-terms for "White House Spokesman" and "White House," but also for parts of those words such as "Whit" and "Whi". Indeed, nesting to depth 50 has been observed in practice for noisy clustering, as displayed in Figure 2. This is a fairly typical pseudo-term nesting graph, in which noisy clustering yields deeper nesting than medium clustering, and much deeper nesting than pure clustering.

Figure 3 shows a collection of pseudo-terms within an overlapping region; in this case a medium clustering representation of the 1.48 second to 3.67 second region of Query 42.[3] As can be seen, calling this "nesting" is somewhat of an oversimplification, the region is actually a set of pseudo-terms that generally overlap to some degree, although not all pseudo-term pairs in one of these "nested" regions actually overlap—pseudo-terms P1 and P21, for example. What gives a nested region its depth is the overlap between pseudo-terms that have adjacent start times. Although in this case, as is typical, there is no one dominating pseudo-term for the entire nested region, there are some cases in which one pseudo-terms is entirely subsumed by another; pseudo-terms P5 and P6, for example. This trait can be leveraged during term matching.

[3]Figure 2 shows the same query between 70 and 90 seconds.

4 Retrieval Models

The development of ranking functions, referred to as "retrieval models," proceeded in three stages. To establish a baseline, we first implemented a standard bag-of-words approach. We then looked to techniques from Cross-Language Information Retrieval (CLIR) for inspiration, since CLIR techniques must accommodate some degree of translation ambiguity and for which robust techniques have been established. Our zero-resource pseudo-term discovery techniques result in representations that differ from the CLIR case in two key ways, however: 1) in CLIR the translation relationship is normally represented such that one side (query or document) exhibits no ambiguity, whereas we have ambiguity on both sides; and 2) in CLIR the typical scope of all translation alternatives are aligned, whereas we have complex nested units that contain terms with differing temporal extents. We therefore developed a new class of techniques that leverage the temporal extent of a pseudo-term as a measure of specificity (Figure 2) and the fraction of a nested unit covered by a pseudo-term as a measure of descriptiveness (Figure 3). This section describes each of these three types of retrieval models in turn.

Indri (Strohman et al., 2004) indexes were built using pseudo-terms from pure, medium or noisy clustering; in each case, stemming and stopword removal were disabled. Indri's query language provides operators that make it possible to implement all of our retrieval models using query-time processing from a single index.

4.1 Types of Retrieval Models

To explore the balance between specificity and descriptiveness, retrieval models were developed that primarily differed along three dimensions: structured versus unstructured, selective versus inclusive, and weighted versus unweighted. Structured mod-

els (*S*) treat nested pseudo-terms with varying levels of synonymy. Unstructured models (*U*) treat nested pseudo-terms as independent. Selective models retain only a subset (1 or n) of the pseudo-terms from each nested region; inclusive models retain them all (*a*). Finally, weighted models (*W*) include a heuristic adjustment to give some pseudo-terms (in our experiments, longer ones) greater influence; unweighted models treat each pseudo-term in the same manner. Table 1 illustrates the weights given to each term by each of the retrieval models defined below. Unweighted models implicitly take a binary approach to term weighting—with unweighted selective models omitting many pseudo-terms—while structured and weighted models yield real values between zero and one. Note that both weighted and unweighted models reward term repetition (term frequency) and term specificity (inverse collection frequency).

4.2 Bag-of-Words Baseline (Ua)

Our first set of experiments had three goals: 1) to serve as a dry run for system development, as we had no prior experience with indexing or ranked retrieval based on pseudo-terms; 2) to gain experience with performing relevance judgments using only the audio responses; and 3) to understand the feasibility of speech retrieval based on pseudo-terms. For these initial experiments, each pseudo-term was treated as a "word" in a bag-of-words representation (coded Ua). No consideration was given to term length or nesting. Although this set of runs was largely exploratory, it provided a good baseline for comparison to other methods considered.

4.3 Terms as Synonyms (Sa, U1)

Moving beyond the bag of words method of term selection involves various forms of term analysis within an overlapping region. The first family of methods treats terms in each overlapping group as synonymous. Aside from being straightforward, treating terms as unweighted synonyms has been a successful technique in cross-language IR. There are generally two methods that can be used in such cases. The first is to treat all overlapping pseudo-terms as synonyms of a single term. This is accomplished in Indri by placing each pseudo-term in an overlapping region within the `syn` operator. This model is coded Sa.

	Retrieval Model					
P. Term	Ua	Sa	U1	Un	UaW	SaW
P21	1.00	0.05		1.00	0.45	0.45
P20	1.00	0.05			0.43	0.22
P19	1.00	0.05			0.48	0.48
P18	1.00	0.05			0.36	0.36
P17	1.00	0.05			0.45	0.06
P16	1.00	0.05		1.00	0.53	0.53
P15	1.00	0.05			0.48	0.11
P14	1.00	0.05			0.37	0.12
P13	1.00	0.05			0.48	0.22
P12	1.00	0.05			0.36	0.02
P11	1.00	0.05			0.41	0.22
P10	1.00	0.05			0.43	0.24
P9	1.00	0.05	1.00		0.54	0.54
P8	1.00	0.05			0.45	0.45
P7	1.00	0.05			0.39	0.04
P6	1.00	0.05			0.37	0.03
P5	1.00	0.05			0.40	0.13
P4	1.00	0.05			0.41	0.08
P3	1.00	0.05			0.47	0.47
P2	1.00	0.05			0.40	0.22
P1	1.00	0.05		1.00	0.46	0.46

Table 1: Weights assigned to pseudo-terms in Figure 3 by each retrieval model (zero values shown as blank).

One risk with the Sa model is that including shorter terms may add more noise than signal. Another method of dealing with alternatives in the cross-language IR literature is to somehow select a single term from the set. For our experiments with this technique, only the longest pseudo-term from an overlapping set is retained; all other ("nested") pseudo-terms are simply deleted from the query. The thinking behind this is that the longest term should contain the greatest amount of information. This method is coded U1.

4.4 Length Measure of Specificity (UaW, SaW)

The U1 and Sa models are two extremes on a spectrum of possibilities; thus, models in which some pseudo-terms receive less weight, rather than being ignored entirely, were also explored. Care must be

taken, however, to do so in a way that emphasizes coverage rather than nesting depth: more weight should not be given to some region in a query or a response just because it is deeply nested (indicating extreme uncertainty). Both the U1 and Sa models do this, but in a rather unnuanced manner. For a more nuanced approach, inspiration can be found in techniques from cross-language IR that give more weight to some term choices than to others.

Our basic approach is to downweight terms that are dominated temporally by several other terms, where the amount of downweighting is proportional to the number of terms that cover it. This is implemented by adjusting the contribution of each pseudo-term based on the extent of its overlap with other pseudo-terms. This could be done in a way that would give the greatest weight to either the shortest or the longest nested pseudo-term.

Formally, let $T = \{t_1, t_2, \ldots, t_n\}$ be the nested term class, ordered by term length. Let $l(t_i)$ denote the length of term t_i, in seconds. Further, let

$$w(t_i) = \frac{\alpha \times l(t_i)}{1 + \alpha \times l(t_i)}$$

be the weight of term t_i, where α is a free parameter. For our experiments, $\alpha = 0.5$. The *discounted weight* is

$$d(t_i) = \begin{cases} w(t_i) & i = 1 \\ w(t_i) \times \prod_{j=1}^{i-1} (1 - w(t_j)) & \text{otherwise,} \end{cases}$$

where t_j refers, implicitly, to other members of T. The factor $1 - w(t_i)$ is used to discount the weight of t_i due to the contribution made by the previous term(s). We assume T to be in descending order and define two heuristics: *total weight discounted* (UaW) and *longest weight discounted* (SaW). The former uses Indri's `weight` operator to specify term weights at query time; the latter uses `wsyn`.

4.5 Coverage Measure of Descriptiveness (Un)

Recall Figure 3, a visual display of pseudo-term overlap within an arbitrary region of speech. Outside of the bounds of that figure there is either silence—no terms to describe a particular segment of time—or a region of terms that describe some other utterance within the overall speech. Of particular note, however, is that within the bounds there are a potentially large number of terms that can be used to *describe* a region of speech. Thus, the larger the number of terms present, the larger the amount of redundancy in the segment of speech each term describes. This observation motivates our final query methodology: removing redundancy within a region by extracting a seemingly descriptive subset of terms from that region. Here we begin to move beyond the ideas inspired by cross-language IR.

Specifically, we posit that an optimal subset contains the beginning and ending terms of the region, along with a series of intra-terms that connect the two. It is with this logic that the *unweighted shortest path* (coded Un) was conceived. Un attempts to find the subset that captures the most information using the smallest number of terms. Formally, consider a directed graph in which the set of vertexes is the set of pseudo-terms within an overlapping region. For an arbitrary pair of vertexes, $u, v \in V$, there is an outgoing edge from u to v if $y(u) \geq x(v)$, where $x(\cdot)$ and $y(\cdot)$ denote the start and end time, respectively, of a given pseudo-term. Further, the weight of such an edge is the difference between these times: $w(u, v) = y(u) - x(v)$. Note that an edge between u and v does not exist if they have the same start time, $x(u) = x(v)$.

Let $\hat{u}$ and $\hat{v}$ be the endpoints of the graph; that is, for all $u, v \in P$, $x(\hat{u}) \leq x(u)$, and $y(\hat{v}) \geq y(v)$. Our objective is to find the shortest path from $\hat{u}$ to $\hat{v}$ that minimizes the standard deviation of the edge weights. Minimizing standard deviation results in a set of terms with more uniform overlaps.

5 Building a Test Collection

The test collection was built using actual spoken content from the Avaj Otalo (Patel et al., 2010) "speech forum," an information service that was regularly used by a select group of farmers in Gujarat. These farmers spoke Gujarati, a language native to western parts of India and spoken by more than 65 million people worldwide. Most of the farmers knew no other language, and approximately 30 per cent were unable to read or write. The idea was to provide a resource for the local farming community to exchange ideas and have their questions an-

swered. To this end, farmers would call into an Interactive Voice Response (IVR) system and peruse answers to existing questions, or would pose their own questions for the community. Other farmers would 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 or call center was involved.

Avaj Otalo's recorded speech was divided into 50 *queries* and 2,999 *responses*. Queries were statements on a particular topic, sometimes phrased as a question, sometimes phrased as an announcement. Responses were sometimes answers to questions, sometimes they were related announcements, and sometimes they were questions on a similar topic. This represented approximately two-thirds of the total audio present in the system. Very short recordings were omitted, as were those in which little speech activity was automatically detected. The average length of a query is approximately 70 seconds ($SD = 14.40$s), or approximately 61 seconds ($SD = 15.76$s) after automated silence removal. Raw response lengths averaged 110 seconds ($SD = 88.80$s), and 96.52 seconds ($SD = 82.75$s) after silence was removed.

5.1 Relevance Judgments and Evaluation

Pools for judgment were formed by combining the results from every system reported in our results section below, along with several other systems that yielded less interesting results that we omit for space reasons. Three native speakers of Gujarati performed relevance assessment; none of the three had any role in system development. Relevance assessment was performed by listening to the audio and making a graded relevance judgment. Assessors could assign one of the following judgments for each response: 1) unable to assess, 2) not relevant, 3) relevant, and 4) highly relevant.

For evaluation measures that require binary judgments, and for computing inter-annotator agreement, the relevance judgments were subsequently binarized by removing all the unassessable cases. Highly relevant and relevant responses were then collapsed into a single relevant category. To compute NDCG, relevant and highly relevant categories were assigned the scores 1 and 2, respectively, while non-relevant judgments retained a score of 0. Three rounds of relevance assessments were conducted as query models were developed and assessor agreement was characterized.

	Retrieval Model					
	U1	Un	Ua	UaW	Sa	SaW
MRR	0.447	0.281	0.169	0.204	0.235	0.432
	0.139	0.071	0.081	0.089	0.242	0.075
	0.188	0.104	0.109	0.193	0.252	0.105
MAP	0.106*	0.057	0.047	0.060*	0.058	0.111
	0.023	0.011	0.015	0.018	0.050	0.010
	0.045	0.013	0.018	0.050	0.058	0.022
NDCG	0.237	0.216	0.206	0.219	0.214	0.284*
	0.122	0.098*	0.187	0.195	0.243	0.194
	0.142	0.089*	0.219	0.191	0.285	0.230

Table 2: Results for pure (top), medium (middle) and noisy (bottom) clustering for the 10 queries for which more than one relevant response is known. Shaded cells are best-performers, per measure; starred values indicate NDCG or MAP is significantly better or worse than same-row Ua (two-sided paired t-test, $p < 0.05$).

6 Results

Each retrieval model was run for each of the three clustering results. For each method, there were three metrics of interest: normalized discounted cumulative gain (NDCG), mean reciprocal rank (MRR), and mean average precision (MAP). Results are outlined in Table 2. To limit the effect of quantization noise on the evaluation measures, results are reported for queries having three or more relevant documents. There were a total of 10 such queries, having a total of 61 relevant documents and yielding an average of 6.10 documents per query ($SD = 2.13$).

Low baselines for each evaluation were established—as there were none in prior existence—by randomly sampling 60 documents from the test collection. For each of the six randomly selected topics, 10 of the 60 randomly selected documents were add to the judgment pool without replacement.

Relevance judgments were performed in an order that obscured, from the assessor, the source of the response being judged. The 10 random selections were then evaluated for each of the six topics as if they had been a system run. None of the 60 randomly selected documents were judged by assessors to be relevant to their respective randomly selected topic; thus the random baseline for each of our measures is zero. Without multiple draws, confidence intervals on this value cannot be established. However, we are confident that random baselines even as high as 0.1 for any of our measures would be surprising.

Pure clustering produced the best results with respect to other clustering domains. SaW was, generally, the best performing retrieval model. Although SaW did not produce the highest pure cluster MRR numbers, it was within 0.015 of U1, the best performing method. This is notable given that the difference between U1 and the third best method was 0.166. Further, given the highly quantized nature of MRR, a difference of 0.015 says little about any overall difference between the rankings. In the case of NDCG, SaW was the best performer with pure clustering, significantly better than BoW with pure clustering and second best overall. Sa with noisy clustering was best numerically with NDCG, but the difference is minuscule (1/1000th).

Under pure clustering, Ua was generally the worst performer. Thus, query refinement using the temporal extent of pseudo-terms is a good idea. Further, the MRR of U1 and SaW both approach one-half. Since MRR is the inverse of the harmonic mean of the rank, we can interpret this as meaning that it is likely that a user will get a relevant document somewhere in the first three positions of the result set. Such a result is encouraging, as it means that, under the correct conditions, a retrieval system built using zero-resource term detection is a potentially useful tool in practice. We should note, however, that this result was obtained for result-rich queries in which three or more relevant responses were known to exist; MRR results on needle-in-a-haystack queries for which only a single relevance response exists would likely be lower. As with all search, precision-biased measures benefit from collection richness.

7 Conclusions and Future Work

Recent advances in zero-resource term discovery have facilitated spoken document retrieval without the need for traditional transcription or ASR. There are still open questions, however, as to best practices around building useful IR systems on top of these tools. This work has been a step in filling that void. The results show that these zero-resource methods can be used to find relevant responses, and that in some cases such relevant responses can also be highly ranked. Retrieval results vary depending on how much redundancy exists in the transcribed data, and how that redundancy is handled within the query. One common theme, at least for the techniques that we have explored, is that pure clustering seems to be the best overall choice when ranked retrieval is the goal. A promising next step is to look to techniques from speech retrieval for insights that might be applicable to the zero-resource setting. One possibility in this regard is to explore extending the zero-resource term matching techniques to generate a lattice representation from which expected pseudo-term counts could be computed.

8 Acknowledgments

The authors wish to thank Nitendra Rajput for providing the spoken queries and responses, and for early discussions about evaluation design; Komal Kamdar, Dhwani Patel, and Yash Patel for performing relevance assessments; and Nizar Habash for his insightful comments on early drafts. Thanks is also extended to the anonymous reviewers for their comments and suggestions. This work has been supported in part by NSF award 1218159.

References

Alberto Abad and Ramón Fernandez Astudillo. 2012. The L2F spoken web search system. In *MediaEval*.

Xavier Anguera, Florian Metze, Andi Buzo, Igor Szöke, and Luis Javier Rodríguez-Fuentes. 2013. The spoken web search task. In *MediaEval*.

Michael Bendersky and W. Bruce Croft. 2008. Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Rnformation Retrieval*, pages 491–498.

Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2010. Learning concept importance using a weighted dependence model. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 31–40.

Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2011. Parameterized concept weighting in verbose queries. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 605–614.

Chun-an Chan, Cheng-Tao Chung, Yu-Hsin Kuo, and Lin shan Lee. 2013. Toward unsupervised model-based spoken term detection with spoken queries without annotated data. In *International Conference on Acoustics, Speech and Signal Processing*, pages 8550–8554, May.

Jia Cui, Xiaodong Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T.N. Sainath, and A. Sethy. 2013. Developing speech recognition systems for corpus indexing under the IARPA Babel program. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6753–6757, May.

Mark Drezde, Aren Jansen, Glen Coppersmith, and Ken Church. 2010. NLP on spoken documents without ASR. In *Conference on Empirical Methods on Natural Language Processing*, pages 460–470.

Jonathan Fiscus, Jerome Ajot, John Garofolo, and George Doddington. 2007. Results of the 2006 spoken term detection evaluation. In *SIGIR Workshop on Searching Spontaneous Conversational Speech*, pages 51–57.

Mark Gales, Kate Knill, Anton Ragni, and Shakti Rath. 2014. Speech recognition and keyword spotting for low resource languages: Babel project research at CUED. In *Spoken Language Technologies for Under-Resourced Languages*.

Aren Jansen and Benjamin Van Durme. 2011. Efficient spoken term discovery using randomized algorithms. In *Automatic Speech Recognition and Understanding*.

Aren Jansen, Kenneth Church, and Hynek Hermansky. 2010. Towards spoken term discovery at scale with zero resources. In *Interspeech Conference*, pages 1676–1679.

Aren Jansen, Benjamin Van Durme, and Pascal Clark. 2012. The JHU-HLTCOE spoken web search system for MediaEval. In *MediaEval*.

Hardik Joshi and Jerome White. 2014. Document similarity amid automatically detected terms. Forum for Information Retrieval Evaluation, December.

Matthew Lease. 2009. An improved Markov random field model for supporting verbose queries. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 476–483.

Florian Metze, Nitendra Rajput, Xavier Anguera, Marelie Davel, Guillaume Gravier, Charl van Heerden, Gautam Mantena, Armando Muscariello, Kishore Prahallad, Igor Szoke, and Javier Tejedor. 2012. The spoken web search task at MediaEval 2011. In *International Conference on Acoustics, Speech and Signal Proccessing*, pages 3487–3491.

Douglas Oard, Jerome White, Jiaul Paik, Rashmi Sankepally, and Aren Jansen. 2013. The FIRE 2013 question answering for the spoken web task. Forum for Information Retrieval Evaluation, December.

Douglas W. Oard. 2012. Query by babbling: A research agenda. In *Workshop on Information and Knowledge Management for Developing Regions*, pages 17–22.

Alex Park and James R. Glass. 2008. Unsupervised pattern discovery in speech. *Transactions on Audio, Speech, and Language Processing*, 16(1):186–197.

Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. 2010. Avaaj Otalo: A field study of an interactive voice forum for small farmers in rural India. In *Human Factors in Computing Systems*, pages 733–742.

Nitendra Rajput and Florian Metze. 2011. Spoken web search. In *MediaEval*.

Amanda Spink, Dietman Wolfram, Bernard Jansen, and Tefko Saracevic. 2001. Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234.

Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2004. Indri: A language model-based search engine for complex queries. In *International Conference on Intelligence Analysis*.

Krysta Svore, Pallika Kanani, and Nazan Khan. 2010. How good is a span of terms? exploiting proximity to improve Web retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161.

Jerome White, Douglas W. Oard, Nitendra Rajput, and Marion Zalk. 2013. Simulating early-termination search for verbose spoken queries. In *Empirical Methods on Natural Language Processing*, pages 1270–1280.