Playing with distances: Document Similarity Amid Automatically Detected Terms @FIRE 2014

Harsh Thakkar[1], Ganesh Iyer[1], Honey Patel[2], Kesha Shah[1]

{DA-IICT[1]- Gandhinagar, Gujarat University[2]- Ahmedabad},
Gujarat, India
{harsh9t@gmail.com, me@ganeshiyer.net, honeypatel.39@gmail.com, kesha.shah1106@gmail.com}

Abstract. Spoken information retrieval is a promising domain of research. In this paper we describe our participation in the pilot Document Similarity Amid Automatically Detected Terms task of FIRE 2014. We present the findings on our experiments with variants of distance and timestamp based approaches. The de-normalized distance based variant outperformed other two delivering best results of the submitted runs. However, there is scope for further improvement in the results.

1 Introduction

In the recent days, the term ”Speech Retrieval” has gained attention of researchers from information retrieval and speech processing communities. Previous work was mostly based on retrieving useful information from semi-structured data using text-based queries. This shift of interests can prove to be extremely useful since majority of the Internet users over the world prefer voice based communication. Thus, the volume of speech based data generated every day is huge. Applying proper information retrieval and speech processing techniques can open new doors for innovative inter-disciplinary research. Processing such data also proposes challenges [1]. The query and response both should be in spoken format and there is lack of high precision speech recognition and conversion systems to tackle this task.

Since the data is audio format, a variety of systems based on non-traditional modalities are been developed [2]. Information retrieval research communities such as Forum for Information Retrieval Evaluation (FIRE)[1] have started tasks on Spoken information retrieval using spoken queries to search a set of audio files is a field that attempts to address this challenge. Advances in automatic speech recognition and speech-based information retrieval systems have driven progress within the field; however, that progress has been biased toward a relatively small number of languages. There are a large number of languages particularly localized languages within developing regions that have been left out of these discoveries.

[1] FIRE, http://www.isical.ac.in/ fire/

Addressing this problem requires either significant improvements to ASR, or viable alternatives. A promising step toward the latter is zero-resource term detection, a method that identifies matching regions amongst a collection of audio without prior knowledge of the underlying language model. Thus, given a set of audio files, a set of matching segments within and across those audio files can be created. The problem statement as described by the organizers is as "Given a set of queries and a set of responses, both represented as sets of such segments, the purpose of this task is to identify response documents that are related to each query"[2].

In this paper we propose variants of Euclidean-based distance to address the challenge of spoken information retrieval domain. The 2014 Forum for Information Retrieval Evaluation (FIRE) focuses on Indian language audio retrieval. This year it has ultimately evolved into a pilot task only at FIRE 2014, with a focus only on speech information retrieval.

2 Corpus and Task description

The test dataset collection is created from the original audio recordings from a phone-based bulletin board system for farmers in Gujarat. Farmers would call into the system to ask questions; other farmers would call in to answer those questions. Periodically system administrators would leave announcements for all system participants to hear. The entire system was automated: callers were not presented with a live operator, instead interacting with the system and making their recordings by following computer generated prompts.

These audio recordings were "transcribed" using a zero-resource term detection system. The result is a series of documents one for each recording consisting of identifier-segment pairs. That is, for each non-silent segment of the audio file, there is a demarcation of that segment, along with an identifier for the segment. Identifiers, across documents, are not necessarily unique matching identifiers denote matching regions of audio. These identifiers are known as pseudo-terms.

In the dataset provided by the task organizers, there are a total **3,148 documents**, consisting of **149 queries** and **2,999 responses**. The entire collection of responses was made available to participants. Queries are divided into test and training sets. The training set consists of **16 queries**, along with relevance judgments for those queries. The testing phase consists of some subset of the remaining **133 queries**.

The documents are CSV files with three columns: *pseudo-term, start time*, and *end time*. Pseudo-terms are regions of speech that appear throughout the corpus. Other documents may have similar pseudo-terms if regions were deemed to be similar in the audio space. The start and end regions mark where a given term appeared within an audio file. While pseudo-term itself is not necessarily unique, the (pseudo-term, start, end) generally is.

[2] The Document Similarity Amid Automatically Detected Terms page, online at http://14.139.122.23/8000_HJJoshi/4000_Document_Similarity.html

3 Proposed methodology

We employ distance-based similarity approach as the document similarity measure. We prepared three variants as described below. From these three distances based we submitted two best performing runs for the final evaluation. We used the standard Euclidean based distance calculation method with a combination of the start and end time as required. The three variants are summarized as:

- **Normalized distance** (Time + distance):
  In this variant we applied the Euclidean distance method on the pseudo terms in conjunction with the time intervals. We considered a combination of the difference of pseudo terms and timestamps. The cumulative score obtained was then normalized for each document to provide precise statistics. Thus, we considered the (document, query) pair for every pseudo term present in the query which produced the least cumulative difference and marked it as relevant for that particular query. Thus this method gives the best global
- **De-normalized distance**(Distance only):
  In this variant we employed the same methodology as in the above method, except we did not aggregate/normalized the results for each query. It was observed that the local minimum difference proved to be a better method than the previous global minimum difference. The findings are discussed in section 4.
- **Normalized distance** (Distance only):
  This variant is based on variant 1 with a modification, in which we consider only the pseudo-term difference for every document per each query. We do not consider difference of the timestamps.

4 Experimentation and result analysis

It was discovered during the analysis of the training corpus that not all the queries have relevant documents in the QREL file. This remains a mystery for the participants. The number of total queries with relevant documents are shown in the figure 1 below. **There are in total 12 queries with 34 relevant documents in the training corpus**. The other 4 queries however have no relevant documents in the QREL or they are not mentioned retrieved in the QREL file.

Figure 2 represents the comparison of normalized and de-normalized distance variants respectively. The de-normalized variant retrieved relevant documents for 8 distinct queries of the total 12. While the normalized variant managed to retrieve only for 5 queries.

Figure 3 represents the comparison of all the three distance based variants with the total number of relevant documents in the training corpus. It can be observed from the results shown below that the de-normalized based variant outperforms the other two on an overall basis. While it fails to retrieve relevant

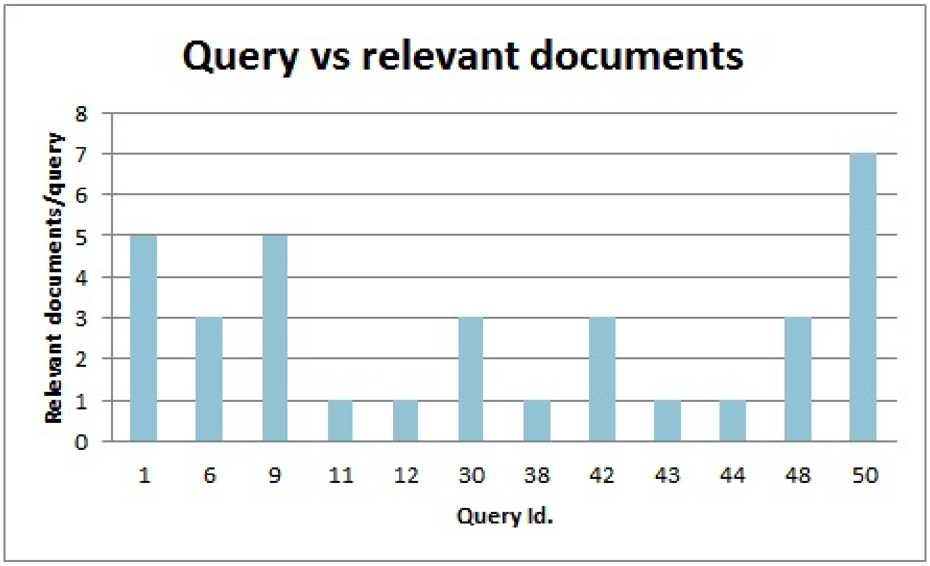

Fig. 1. The figure represents the graph of query id. vs. relevant documents per query of the training dataset as per the qrel file provided by organizers.

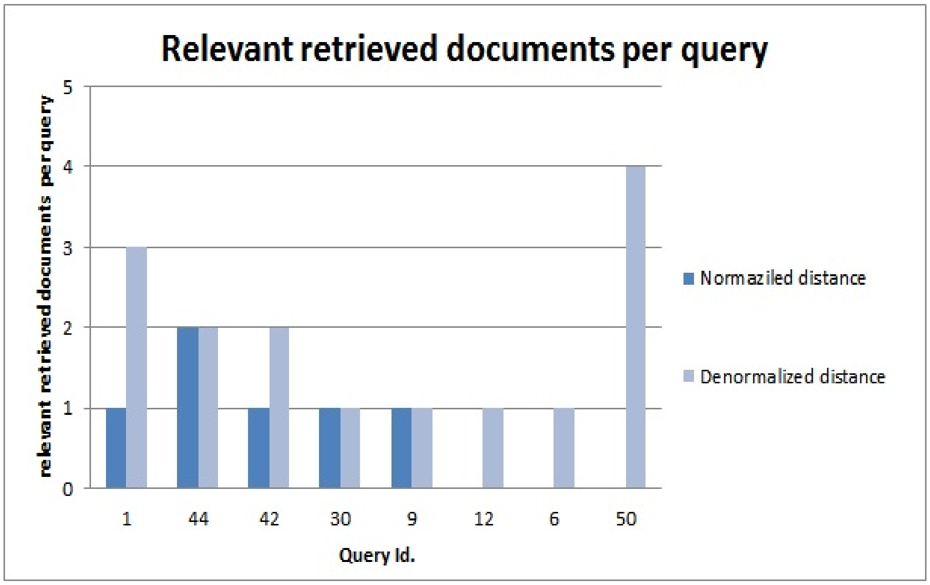

Fig. 2. The figure represents a comparison of relevant retrieved documents per query for normalized distance vs. de-normalized distance.

documents for query number 43. The reason for which is currently unknown. The normalized distance approach outperforms the de-normalized based variant for queries 44 and 6. It is observed that for query 44, all the three variants retrieve more documents than the relevant documents mentioned in the QREL file of the training data. The possible explanations can be: Since for all the other queries the variants work fine and any other relevant information is not known regarding the relevance parameter/threshold of the documents, all the three models fail drastically to differentiate between relevant and non-relevant models; OR; this can be an error in the QREL file provided by the organizers. The normalized (distance only) variant, purple column bar, manages to perform slightly better than the normalized (distance + time) based variant by retrieving relevant documents for 7 queries as compared to 5 to the later. For query 43 the normalized (distance only) variant outperforms all the other variants. For queries 11 and 38 all the three variants display a disappointing zero retrieval of relevant documents.

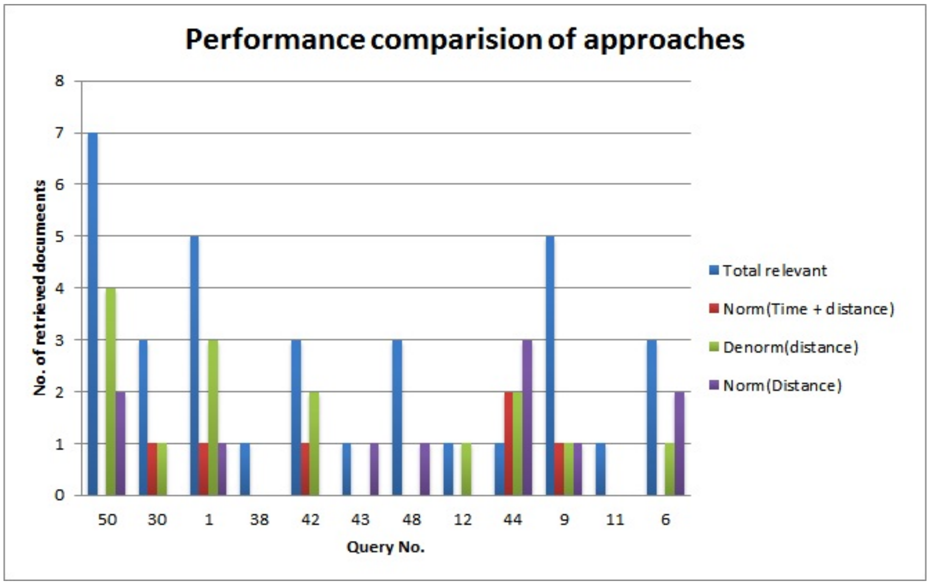

Fig. 3. The figure represents a comparison of relevant retrieved documents per query for normalized distance vs. de-normalized distance.

5 Conclusion

We conclude that the time based variants are not sufficient for the purpose of this task due to a variety of reasons. The parameters are not sufficient for the approaches to differentiate between relevant and non-relevant files. Not all queries have relevant documents, only 12 out of 16 have relevant documents in the QREL file, reason is not known. These variants do not retrieve documents for all the 12 queries present in the dataset, the **best retrieving model is the de-normalized variant** retrieving for a maximum of 8 queries of the total 12.

References

1. White, Jerome, Douglas W. Oard, Nitendra Rajput, and Marion Zalk. "Simulating Early-Termination Search for Verbose Spoken Queries." In EMNLP, pp. 1270-1280. 2013.
2. Oard, Douglas W. "Query by babbling: a research agenda." In Proceedings of the first workshop on information and knowledge management for developing region, pp. 17-22. ACM, 2012.