The Similar Segments in Social Speech Task

*** DRAFT — updated July 15, 2013 ***

Nigel G. Ward∗
Steven D. Werner
David G. Novick
University of Texas at El Paso

Elizabeth E. Shriberg
SRI and ICSI

Catharine Oertel
KTH

Louis-Philippe Morency
ICT, University of Southern California

Tatsuya Kawahara
Kyoto University

ABSTRACT

Similar Segments in Social Speech was one of the Brave New Tasks at MediaEval 2013. The task involves finding segments similar to a query segment, in a multimedia collection of informal, unstructured dialogs among members of a small community.

Categories and Subject Descriptors

H.2 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords

audio information retrieval, spoken content retrieval, more like this, evaluation, multimedia, MediaEval

1. INTRODUCTION

With users’ growing willingness to share personal activity information, the eventual expansion of social media to include social multimedia, such as video and audio recordings of casual interactions, seems inevitable. To unlock the potential value, we need to develop methods for searching such records.

Our motivating scenario is the following: A new member has joined an organization or social group that has a small archive of conversations among its members. He starts to listen, looking for any information that can help him better understand, participate in, enjoy, find friends in, and succeed in this group. As he listens to the archive (perhaps at random, perhaps based on some social tags, perhaps based on an initial keyword search), he finds something of interest. He marks what he found as a region of interest and requests more like it. The system returns a set of “jump-in” points, places in the archive to which he could jump and start listening/watching with the expectation of finding something similar. To the extent that users in this scenario may lack specific intentions, their behavior will resemble undirected searching or even recommendation requests more than directed searching.

Despite the large volume of research in technologies for audio and multimedia search, as surveyed for example by [1, 5], there has been no research addressing this scenario, or similar types of search in social multimedia. There is a need both for the evaluation of the suitability of existing techniques for this task, and for the exploration of new techniques. To support both, we provide a task, a dataset, and an evaluation method.

2. TASK DESCRIPTION

The task is, given a short audio/video region (segment) of interest, to return an ordered list of jump-in points for regions similar to it, where similarity is based on the perceptions of human searchers.

3. DATASET

We audio- and video-recorded two-person dialogs among members of the computer science community at our university. They talked about whatever they wanted, for about 10 minutes each [4]. They were told that their dialogs were going to be annotated for later searching, and many of the conversations turned out to be rich in information likely to be of interest to fellow CS students, rather than just personal talk.

The training set is 20 dialogs, 241 minutes in total, mostly involving undergraduates, with the most common topics relating to classes and class assignments, interesting new technologies, career ambitions, games, and movies. The test set is 6 dialogs, 68 minutes total, involving only research-active students, and the topics were less about classes and more about research, but otherwise fairly similar.

This data was annotated by students, mostly members of the same community, and included some who had contributed dialogs to the collection. The annotators worked mostly independently. In the first pass each listened to and viewed a few dialogs and developed a set of tags to use, each tag associated with a topic that some future searcher may potentially be interested in. They then did a second pass over all the dialogs and every time they found a region...
relevant to a tag, assigned that tag to the data. Regions could span any fragment of the dialog, regardless of any notion of topic or utterance boundary. The average durations were 50 seconds in the training set and, after clarifying the instructions to annotators slightly, 17 in the test set.

While the tags themselves are not relevant for our purposes, each tag serves to define a “similarity set” of regions in which every pair is a positive example of similarity. Task participants can use these examples to hone their similarity metrics. Those similarity metrics can then be used in a system to support the search scenario: given any new query, to return a set of similar regions.

4. EVALUATION OF RESULTS

In the scenario, a user indicates a region to the search system, and the system returns, ideally, a listing of all similar regions.

For evaluation purposes, each region input to the system is taken from one of the similarity sets of an annotator, and the target result is a listing of all the other regions in that set. As the test set speakers and topics differ, systems that perform well will have demonstrated that their methods do generalize, at least to some extent.

As each similarity set is generated by a specific annotator, with his or her own perspective and interests, no system could be expected to return the target results exactly. Nevertheless the degree of match between the results and the target is an indicator of similar-region-finding performance.

Our specific performance measures are based on the use scenario, described above, in which the user watches/listens and browses around the points suggested, rather than passively consuming some precisely delimited segments. For this reason standard metrics based on accuracy and precision are not appropriate.

Instead, we use a rough model of how searchers are likely to use the suggested jump-in points. Extending Liu and Oard’s (2006) model, we define a “Searcher Utility Ratio”, where the numerator is the estimated value to the searcher and the denominator the estimated cost, both measured in seconds.

Specifically, the value to the searcher is modeled as the number of seconds of relevant audio/video she can likely find by using the suggested jump-in points. We assume that she will find a region if a jump-in point is no earlier than 5 seconds before the region start and no later than 3 seconds before the region end.

The estimated cost to a searcher is the number of seconds needed to peruse the suggested jump-in points. There are three cases. 1) If the suggested jump-in point does not correspond to any ground-truth region (a false-positive error), then the cost is 8 seconds, which is the estimated time a user needs to determine that it is a false alarm. 2) If the suggested jump-in point is no more than 5 seconds before the actual region start point, the cost is the time from that jump-in point to the end of the actual region, reflecting the time spent to scan forward to the start of the relevant content and the time to listen to it. 3) If the suggested jump-in point is within the region, then the benefit is the remaining duration of the region, and the cost is the same.

We further assume the searcher devotes two minutes to each search. The total value is accordingly estimated as the amount of relevant audio she can find and consume in that time, according to the model above.

In addition there is a recall-evaluating component, to counter for the possibility of systems doing well by generating jump-in points for only the very easiest queries. Thus Recall is the fraction of obtainable content actually found, where the obtainable content is the total content in the other regions in the tagset, up to a maximum of two minutes.

The raw Searcher Utility Ratio and raw Recall are valid for comparing systems, but significantly understate performance. This is because regions other than those in the specific similarity set for a query may in fact be similar to that query in other respects, but will be counted as false alarms. So raw scores even for systems producing excellent jump-in points will be low. Accordingly we adjust the scores by dividing by the estimated best-obtainable raw performance values, estimated using an algorithm that consults closely-overlapping other-annotator tags to propose jump-in points. Thus for the testset the Searcher Utility Ratio is obtained by dividing the raw value by 0.159 and the Recall by dividing the raw value by 0.211.

The overall measure is the F-measure, with the utility weighted higher than the recall:

\[
\frac{10 * R * U}{U + 9R}
\]  

5. OUTLOOK

This is clearly a very challenging task, unlikely to be completely solved in this MediaEval round, or indeed anytime soon. Nevertheless it will enable researchers to demonstrate the value both of novel approaches and novel applications of existing methods. While our scenario is for search in social recordings, technologies developed for this task are also likely to be useful for other needs [3], such as search of workplace recordings, of surveillance recordings, of personal recordings, and so on.

6. ACKNOWLEDGMENTS

We thank Martha Larson and Mary Harper for suggestions, and the National Science Foundation for support via a REU supplement to Award IIS-0914868.

7. REFERENCES


