The Similar Segments in Social Speech Task

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.

1. INTRODUCTION

With users' growing willingness to share personal activity information, the eventual expansion of social media to include social *multi*media, 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. This requires us to develop good models of the similarity between dialog-region pairs.

Our motivating scenario is the following: A new member has joined an organization or social group that has a small archive of conversations among 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 this region of interest and requests "more like this". 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. In this scenario users may lack specific intentions, and their behavior may resemble undirected search or even recommendation requests more than directed search.

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 such a scenario, or otherwise search in social multimedia. There is a need for evaluation support, both for examinations of the suitability for this task of existing techniques 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. In directly addressing pure similarity this task is novel; it avoids the need to use any of the typical simplifications — such as framing the problem as being topic match, search-term match, or dialog-act match — which ultimately, we believe, are distorting and limiting.

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 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, with less talk about classes and more about research, but otherwise fairly similar.

The annotations are tagsets which indicate regions similar in some way. These were done by students, mostly members of the same community, including 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 some set of somehow-related regions that some future searcher may potentially be interested in. They then did a second pass over all the dialogs, and for every region found that was relevant to some 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, 31 in the test set. There were 198 tagsets over the training set, with a total of 1697 tagged regions, and 29 and 189 for the test set.

While most tags were related to traditional-style top-

[*] nigelward@acm.org, stevenwerner@acm.org, novick@utep.edu, ees@icsi.berkeley.edu, catha@kth.se, morency@ict.usc.edu, kawahara@i.kyoto-u.ac.jp

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

ics, such as #food, #travel, #cars-and-driving, #planning-class-schedules, #TV-shows, #lack-of-money, and #family, others related instead, or in addition, to dialog activity, for example #anecdotes, #problems, #short-term-future-plans, #advice, #gossip, and #positive-things-about-classes. 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.

Participants were also given human-generated transcripts and automatic speech recognition output. The latter was far more errorful, typically having more incorrect words than correct ones, but was more faithful with the *um*s and *uh*s, as our human transcribers were told not to bother with those.

4. EVALUATION OF RESULTS

For evaluation purposes, an input to the system is a region from one of the similarity sets of an annotator, and the ideal result is a set of jump-in points that closely index all the other regions in that set. As the testset speakers and topics differ from those in the training set, systems that performed well will have demonstrated that their methods generalize, at least to some extent.

Our specific performance measures are based on the scenario, in which the user watches/listens and browses around the points suggested, rather than passively consuming some precisely delimited segments. (Despite the title of the task, the dialogs were not segmented in any way). 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 same-tagset region (a false-positive error), then the cost is 8 seconds, an estimate of the time a searcher needs to recognize 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 measure of recall, to counter for the possibility of systems doing well by generating only a handful of jump-in points, just the easiest ones. 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 two-minute maximum.

The raw Searcher Utility Ratio and raw Recall are valid for comparing systems' ability to find similar regions, but they 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. That is, because 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. Accordingly we adjusted the scores by dividing by an estimate of the best-obtainable performance values. This estimate was obtained using an algorithm that consults closely-overlapping other-annotator tags to propose jump-in points (although later we found that this underestimated the possible performance, enabling performance results to exceed 1.0). Thus for the testset the Normalized Searcher Utility Ratio (NSUR) is obtained by dividing the raw value by 0.159 and the Normalized Recall by dividing the raw value by 0.211.

The overall measure is the F-measure, with NSRU weighted higher than Normalized Recall (specifically by a factor of 9, based on consideration of how appreciative users might be of results having various NSRU and NR values).

$$\frac{10 * U * R}{U + 9R} \tag{1}$$

5. OUTLOOK

While very challenging, this task will enable researchers to explore search in dialog archives, the more-like-this task, pure-similarity models, and the social-speech domain. While our scenario is for search in social recordings, technologies developed for this task are likely to be useful also 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, Steve Renals and Khiet Truong, and the National Science Foundation for support via a REU supplement to Award IIS-0914868.

7. REFERENCES

[1] M. Larson and G. J. F. Jones. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Info. Retrieval*, 5(4-5):235–422, 2012.

[2] B. Liu and D. W. Oard. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *29th SIGIR*, pages 673–674, 2006.

[3] N. G. Ward and S. D. Werner. Thirty-two sample audio search tasks. Technical report, 2012. University of Texas at El Paso, Tech. Report, UTEP-CS-12-39.

[4] N. G. Ward and S. D. Werner. Data collection for the Similar Segments in Social Speech task. Technical report, 2013. University of Texas at El Paso, UTEP-CS-13-58.

[5] N. G. Ward and S. D. Werner. Using dialog-activity similarity for spoken information retrieval. In *Interspeech*, 2013.