Data Collection
for the Similar Segments in Social Speech Task

Nigel G. Ward and Steven D. Werner

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518
email: nigelward@acm.org, stevenwerner@acm.org

August 3, 2013

Information retrieval systems rely heavily on models of similarity, but for spoken
dialog such models currently use mostly standard textual-content similarity. As part
of the MediaEval Benchmarking Initiative, we have created a new corpus to support
development of similarity models for spoken dialog. This corpus includes 26 casual
dialogs among members of two semi-cohesive groups, totaling about 5 hours, with
1889 labeled regions associated into 227 sets which annotators judged to be similar
enough to share a tag. This technical report brings together information about this
corpus and its intended uses, previously only available on the project website.

Index Terms: information retrieval, multimedia, corpus, video, audio, speech, dialog, search,
recommendation systems, MediaEval

1 Goals

Information retrieval systems, recommendation systems, and other language processing systems
rely heavily on models of similarity, both document-document and document-query similarity.
The development and evaluation of these similarity models requires suitable training corpora.
For spoken dialog, such resources exist only for a few genres [Ward and Werner, 2012]. Moreover,
the design of these corpora has been based on certain assumptions, for example, that the input
is monolog, that it is deliberate speech, that it is clearly divided into topics, and that it can
Therefore we decided to create a new corpus, with the aim of support the development of systems
that overcome these assumptions.

Out of the many possible genres, we chose to collect informal dialogs between members of
the same loose social group. We call this “social speech” as it is the spoken-dialog analog of the
sorts of things found in social media. This genre was chosen not only because of the potential value [Ward and Werner, 2013, Ward et al., 2013] and the timeliness of the topic, but also to be maximally non-redundant to existing corpora. Social speech also serves as a useful proxy for many other dialog genres [Ward and Werner, 2012].

While the corpus is interesting, the real value and novelty here is in the annotations. These provide indications of which sets of dialog regions are similar.

Wanting from the start to share this data widely, we found a vehicle in MediaEval, the Benchmarking Initiative for Multimedia Information Retrieval. Proposing search in this corpus as a task, MediaEval accepted it as one of the 2013 challenge tasks. For the task we defined a novel evaluation method, incorporating a simulation of user behavior in a query-by-example scenario. We also created transcripts and prosodic features sets, to lower the barrier to entry for teams considering participating in the challenge.

The potential uses of this corpus extend beyond MediaEval. It can also support other evaluation methods and other use scenarios. In particular, it was also designed to support user studies. As the topics in this corpus include many of interest to students, it is easy for us, and for a lesser extent to other researchers with access to populations of college students, to measure, with real, motivated subjects, the actual value to users of complete systems for search or for content recommendation.

This technical report brings together information about this corpus and its intended uses. It is intended for those who have already read the task overview paper [Ward et al., 2013] and want more details.

2 Scenario and Task

While we later came to focus on support for building similarity models, originally our thoughts were focused on the construction of complete systems able to actually support users in a realistic scenario. This scenario and the task are as described in Sections 1 and 2 of [Ward et al., 2013].

3 Data Collection

The data collection is summarized in Section 3 of [Ward et al., 2013]. This section provides more detail.

3.1 Recording Conditions and Equipment

Dialogs were recorded with the conversants in different rooms seeing each other through a glass wall.

Wanting stereo with good acoustic separation between the rooms, hence the need for a wall, but also wanting to gather video, we recorded in a pair of rooms with a connecting glass wall. The resulting acoustic separation was not perfect, but the small amount of across-track bleeding is not noticeable at most listening volumes. The video was also not perfect, with some reflections and awkward angles, but acceptable in quality, at least to human eyes.
Each participant wore a Sony DR-200 headset. The inputs were fed into an Olympus DS-2 digital voice recorder. Audio from monitoring jack of the recorder was split, to 1) sent it back to the participants’ headsets so they could hear each other 2) feed the audio input of our main recording device, an iMac. The recordings were created with Quicktime on the iMac, with the video coming from the built-in camera and the audio from the voice recorder. Subsequently the recordings were edited with iMovie to remove the times where subjects were being set up and when the recording was being ended.

The first recordings were done a Macbook Pro, which had a better camera. However it lacked an audio input jack, and attempts to synchronize in the separately recorded audio using iMovie were unsatisfactory because of the inability to get exact alignments and because of clock skew. Next we used Photobooth, but discovered that it would abort the recording if a system alert popped up. We therefore settled on Quicktime, despite the fact that its narrow aspect ratio wasn’t ideal for our through-the-window setup.

### 3.2 Notes on Specific Dialogs

Those wishing to get an intuitive feel for the corpus are recommended to start by listening to dialogs 004, 006, 008, and 012, as these are rich in interesting and diverse audio regions of the sort that one might want find as results to a search.

The release includes all English-language dialogs collected, including some with flaws and some not strictly comparable to the others:

- Dialogs 000–002: recorded using a MacBook (whereas the body of the corpus was recording using an iMac).
- Dialogs 000, 002: include participants knowledgeable about our systems, processes, and information retrieval technology.
- Dialogs 014, 016–018: include one non-CS major (although he/she was taking a CS class).
- Dialog 000–002, 004, 007: audio recorded separately from the video and later imperfectly aligned.
- Dialogs 000, 004, 005, 010, 011, 013–015: video incomplete.
- Dialogs 001 and 007: topic suggestions were given strongly and/or participants took them literally.

One of the dialog participants offered to speak Spanish, a few dropped in a few Spanish words, and several were not native English speakers, as indicated in the metadata.

### 3.3 Other Information

While the dialogs were solicited, all of them were among people who might have had a conversation anyway that day. Most of the speakers were engaged in their conversations, most had to be stopped when the time was up, and several remarked that they gladly would have continued talking. Overall the dialogs were quite natural.
Appendix A, the official Data Collection Protocol, as submitted to the UTEP Institutional Review Board (human subjects experimentation committee), overviews the collection.

Appendix B is a flier used to recruit subjects.

Appendix C is the explanation the subjects read and the consent form they signed. Notable are the restrictions on the use of the data, which must be respected.

Appendix D is the verbal instructions given to the subjects.

Appendix E is the metadata for the dialogs, including for each dialog information on the participants including gender, age range, class status, and native languages, and prior relationships.

4 Annotation

The annotation is summarized in Section 3 of [Ward et al., 2013]. This section provides more detail on who did the tagging and on the different tagging styles observed.

4.1 Annotators

For the training set:

- Annotator 1 was speaker 6. Atypically he did the annotations before the Annotators Guide was available, he’s not a student, and he’s not naive about information retrieval technology.
- Annotator 2 was speaker 1. Atypically he knows a lot about one technique of possible value for this task.
- Annotator 3 was speaker 28. Atypically she’s not quite part of the CS community, being a math major, although she was taking a CS course and does IT work in her job.
- Annotator 4 was speaker 2. Atypically, she has experience doing search in audio archives, having spent about 10 hours doing so as part of the experiment reported in [Ward and Werner, 2013].
- Annotator 5 was speaker 30.
- Annotator 6 was speaker 5. Atypically, he was a member of the research lab, having joined a few months ago, and somewhat knowledgeable about our favorite analysis methods.

For the testset:

- Annotator 1 was the same as annotator 3 for the testset.
- Annotators 2 and 3 were CS undergraduates. They worked at the the same table, and may have shared some thoughts on the tags.
- Annotator 4 was a high-school student.
4.2 Tagging Styles

The most salient differences in tagging styles involved the length of the tagged regions.

Compared to the training set tags, there were many fewer long regions tagged in the test set. In the training set there was one region tagged over 4 minutes long (tv-shows), four more over 3 minutes (playing-video-games, programming-projects, movies-tv-shows, course-experiences), and 58 over two minutes. In the test set there were only 5 longer than 2 minutes, and these were all for just one category of one annotator. One reason for the difference is that, having noticed the over-long tags in the training set, we suggested to the test set annotators that it was okay to leave large sections of the data without any tags, and that could be appropriate to break up a long region on a single vague topic (for example entertainment) into more specific contributions. Another possible factor is that, with only 6 conversations to work on, the test set annotators may not have felt the same degree of desire to get it over with.

There was a small problem with short tags: most of the regions tagged by test set annotators 2 and 3 turned out to be only 1-4 seconds in length; it seemed that they had interpreted their job as being the identification of semantically related words in the corpus. While this relates to a potentially interesting task, it is different from the one in our scenario, and so their tag sets were dropped.

4.3 Other Information

The relevant appendices are:

Appendix F, the Annotator Training Overview, lists the steps used to train the first six annotators.

Appendix G, the Annotators Guide, specifies the goal, guidelines, and procedure for annotation.

Appendix H, How to Tag Social Speech using Elan, explains the software used.

5 Features

In order to help teams participating in this task, we provided both transcripts and prosodic features.

5.1 Transcriptions

Initially attempts to obtain transcripts using a free Google transcription service and from Dragon Dictate gave outputs that were correct for fewer than one word in ten.

The actual transcripts were kindly provided by Steve Renals of the University of Edinburgh. According to his note:

These were produced by a system which should be considered to be very much a baseline system using acoustic models trained on meetings data (primarily AMI cor-
pus) and a language model optimised for NIST RT meeting transcription. For each recording there’s a “rec” file, giving the detected speaker, start and end times (in centiseconds) and transcript for each utterance, and an “mlf” file, with detailed timings of words and phones within the utterances (but the utterances not in chronological order). We imagine it will be mainly the “rec” files that participants want to use.

In addition we provided human-generated transcripts. The bulk were done by trainingset annotator 3, with the remainder, 20, 27, and the first half of 26, done by an inexperienced volunteer. They used TranscriberAG. The instructions were:

What we want is each of the 26 dialogs transcribed (all dialogs except the Spanish ones.) We’re using Transcriber because it gives timestamps at the utterance level; so please select roughly utterance-sized regions (probably 4-10 seconds) and label each one.

Speed is more important for this task than accuracy. In particular, it’s okay to skip nonlexical items (uh-huh, um, and laughter), word fragments, and anything that’s not clear on first listening. (Although it’s also okay to include them, at times when that’s more convenient.) If there are technical terms or other words you can’t catch, just leave them out or enter xxx as a placeholder. Spelling errors should be avoided, but other minor errors are fine; you should never go back and correct mistakes.

It’s okay to be sloppy here because the aim is to be as good as powerful future automatic speech recognition, but not perfect, since that would be unrealistic for us when we test the systems.

5.2 Prosodic Features

There are three sets of prosodic features.

5.2.1 Basic $F_0$ (Pitch) Values

These were generated using the Hirose-Seto algorithm [Hirose et al., 1992]. There is a $f_0$ file for each track, left and right, of each conversation. Each file has one line every 10 milliseconds. Each line in turn has three fields: a timestamp in 100-nanoseconds, the most likely pitch value, and an boolean flag indicating whether or not pitch was in fact likely to be present at that point.

5.2.2 78 Contextual Prosodic Features

Out of the thousands of possible prosodic features that one might consider, this is a set of the sort that we have found helpful for characterizing dialog activity. They are track-normalized, computed over fixed windows (rather than being utterance-, word-, or syllable-aligned), and computed at various offsets. More about such features, including reasons for using them, how they are computed, and experiences with them are described elsewhere [Ward and Vega, 2012, Ward and Werner, 2013].
Specifically for each conversation in the corpus there are two .pc (prosodic context) files, one for the right track and one for the left. Both contain one line for every 10-milliseconds of the conversation. Each line specifies the values for 78 features. Each line in that file below indicates the base feature type, then the start time of the window over which that feature is being computed in milliseconds relative to the current frame, the word “to”, the end time, and the track. The abbreviations are:

- **vo** volume
- **ph** pitch height
- **pr** pitch range
- **sr** speaking-rate proxy
- **self** speaker in this track
- **inte** speaker in the other track (interlocutor)

Appendix I, slimcrunch.fss, lists the items in this featureset.

5.2.3 78 prosodic dimensions

These are PCA-rotated composite features, derived from the above set of 78 original features.

6 Evaluation

The evaluation method is summarized in Section 3 of [Ward et al., 2013].

One complexity is that our idea of evaluation by user simulation is not supported perfectly by our tagset-based algorithm, since region-pairs which do not share a tagset may still be similar. While the metrics are adjusted for this, this remains the biggest issue with our evaluation method.

Appendix J, Guide to Interpreting the Metrics, discusses the realism and stability of the main measures, explains the normalization factors, and describes the ancillary measures.

Appendix K, Baselines on the Testset, gives the performance of a random baseline and a clever baseline.

Appendix L, score5.py, is the python code that computes the performance metrics.

7 Availability

The recordings and transcripts are available by request to either of the authors.

The tagsets, scoring software, prosodic features, and documentation are also available at http://www.cs.utep.edu/nigel/ssss/.
All of the above are also archived in the UTEP Library Special Collections Section. **PENDING**

8 Frequently Asked Questions

This section addresses some questions asked by task participant teams.

_How does this differ from topic detection?_

Some of the similarity sets will probably be largely or entirely topic-based, but most will probably also involve on other factors, such as user goals (talk about childcare experiences in order to find a new daycare provider, versus in order to find ways to make the child feel comfortable), and such as attitudes and purposes (talk about a course for the sake of deciding whether to take it, or for figuring out how best to study for it, or for just sharing stories about the professor).

What exactly do systems have to do?

Given a query (in the form of a region of one of the dialogs), the system should return a set of pointers to hypothesized similar regions. The ideal system will return the onset points of all such similar regions, as identified by the human annotators, and no other regions.

_Could you provide us some more details on how a baseline system would work?_

One the obvious way to build a baseline system would be to gather all the words in the query region, then find all regions elsewhere in the corpus which densely contain those words or similar words. For this any traditional IR technique would work, although probably with modifications to deal with the special properties of spoken language (noisy, interlocutor-track information available, prosodic features also available), and with the lack of a segmentation into “documents”, meaning that system can return regions from anywhere in the corpus and of any size.

_Will the final dataset contain a fixed number of tags?_

No. In fact, the tags are there as comments only. The system you build will not be able to rely on the tags being there or meaning anything. The meaning of each similarity-set is just the set of regions in that set. And in particular, the test set will include as queries regions which were not seen in the training set, and which may not relate to any of the tags seen in the training set. We think this is realistic. For example, our campus recently had a bomb threat, the first in 10 years, so there is no talk anywhere in the corpus about anything like that, but we’d still want a system running today to be able to find other regions of talk about campus security issues if a segment related to a bomb threat was submitted as a query.

_How many training segments will there be available for each tag? Right now, we are seeing tags that only have 1-2 training segments; we do not see how we can train classification models from these segments._

Training a classifier for each tag would be a poor strategy, since the tagset is not fixed. The goal of the task is to find similar segments, and the similarity-sets provided as examples of what counts as similar in this corpus, for these users. If you use these to build and refine a general similarity metric, then that similarity metric can be used for any retrieval request. For example,
if you use a vector-space model, then for any query (e.g. a couple of utterances about the bomb threat), you can find some speech regions that are close to it in the vector space, and return those. In a real system you’d probably use a nearest-neighbors algorithm to find these quickly, but for this task, due to the small corpus and lack of a real-time requirement, exhaustive search will probably be just fine. (But you’re right to note that having only 1 example of some tags is not useful for anything; in the actual data release we’ll aim to have 5-15 examples for most tags.)

Will you provide a training, development, and test set? Right now, there is only mention of a training and testing set.

While most teams will want to split the training set themselves into one part for training and one part for tuning, we aren’t imposing any specific partition.

Why did the developers themselves contribute to the recordings and annotations? Mightn’t that skew things somehow?

The test data will be pristine, so there is no risk. However for training purposes we decided to release also the pilot recordings (000–002) and annotations, thinking that participants would like to have as much data as possible. However the metadata shows which files these are, so it’s possible to exclude them from training if desired.

Why are the value weights for?

As described in section step 7 of the Annotators Guide, each similarity set over the training data was assigned a value, from 0 to 3. These numbers may be useful in the training process, since similarity sets ones with higher values may be more informative/valuable, and participants may want to tune their parameters to perform best on the higher-valued sets.

Some regions appear in multiple places in the tagsets; is this because some segments were assigned multiple tags?

Yes, and this is why the key performance metrics are adjusted, as described in the Guide to Interpreting the Metrics document.

There are some similarity sets that are semantically very much close to each other but are nevertheless seen as different similarity sets. For example, the sets #Courses, #courses, #course material, and #Course Work are all very similar but in evaluation (if you test with all possible queries over the training set), the system will be penalized if, for example, a test query from #Courses results in retrieval of a segment from #courses. One could consider merging some of these simsets in order to obtain a more accurate view of the system’s current performance.

Yes, that’s true. However we are reluctant to try merging the simsets, as that would involve a lot of subjectivity, and would lose some information.

If this is a problem in training, task participants may consider reducing the penalty for false alarms (falsePositivePenalty in score5.py), since, as noted, many of the false alarms are not really errors, at least during the early stages of training.
Acknowledgments

We thank our dialog participants, annotators, and transcribers, especially Jeanette M. Mendoza.

Martha Larson helped us formulate the problem as a challenge task.

Khiet Truong found several bugs in the initial releases, and posed most of the questions above.

Steve Renals provided the automatic-speech-recognition transcriptions.

The Similar Segments in Social Speech organizing committee, Elizabeth Shriberg, Catharine Oertel, Tatsuya Kawahara, David Novick, and Louis-Philippe Morency, provided encouragement and advice.

Shirley Moore, Olac Fuentes, and Kay Roy helped us recruit participants for the dialogs.

Alejandro Vega, Shreyas Karkhedkar, and David Novick provided infrastructure and support.

This work was supported in part by the NSF under project IIS-0914868 and REU supplements thereto.

References


