

AN EXTRACTIVE-SUMMARIZATION BASELINE FOR THE AUTOMATIC DETECTION OF NOTEWORTHY UTTERANCES IN MULTI-PARTY HUMAN-HUMAN DIALOG

Satanjeev Banerjee and Alexander I. Rudnicky

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

Our goal is to reduce meeting participants' note-taking effort by automatically identifying utterances whose contents meeting participants are likely to include in their notes. Though note-taking is different from meeting summarization, these two problems are related. In this paper we apply techniques developed in extractive meeting summarization research to the problem of identifying noteworthy utterances. We show that these algorithms achieve an f-measure of 0.14 over a 5-meeting sequence of related meetings. The precision – 0.15 – is triple that of the trivial baseline of simply labeling every utterance as noteworthy. We also introduce the concept of “show-worthy” utterances – utterances that contain information that could conceivably result in a note. We show that such utterances can be recognized with an 81% accuracy (compared to 53% accuracy of a majority classifier). Further, if non-show-worthy utterances are filtered out, the precision of noteworthy detection improves by 33% relative.

Index Terms— Natural language interfaces

1. INTRODUCTION

Human beings receive a large amount of spoken information from other human beings on a daily basis. Some of the received information is important to the listener, and needs to be remembered for future use. Our broad goal is to help humans capture important pieces of information from spoken dialog to facilitate future retrieval. Such a capability is particularly useful in the context of task- or goal-oriented meetings where information is presented, decisions are made, action items are created, etc. To access these pieces of information afterwards, humans must typically record them in notes during the meeting [1]. Note-taking at meetings is a particularly difficult task, however, because it must be done concurrently with participating in the discussions. Our goal is to make note-taking at meetings easier by automatically

extracting important phrases or sentences from the speech, and suggesting them as notes to the meeting participants.

Towards this end, we define an utterance spoken at a meeting as noteworthy if one or more meeting participants are willing to include the information contained in that utterance in their notes. Our goal in this paper is to identify such noteworthy utterances. Since the noteworthiness of an utterance is heavily domain- and participant-dependent, we aim to learn to perform this task for sequences of related meetings (e.g. weekly project meetings). For such meetings, our goal is to detect noteworthy utterances by learning from the notes that the participants themselves have taken in earlier meetings in the sequence.

Note-taking is both similar to and distinct from the problem of meeting summarization. Both aim to identify utterances that are in some ways more important than others. At the same time, whereas summaries aim to list the main points of discussion, notes are taken to serve as an aid to memory. As such, even though some utterances may be important in some sense, they will not be included in the notes unless participants feel (a) they need to be remembered, and (b) they are likely to forget them. Nevertheless, the techniques developed in meeting summarization research are a good starting point for noteworthy utterance detection. In this paper we evaluate the effectiveness of these techniques, when applied to the problem of noteworthy utterance detection.

2. RELATED WORK

In extractive summarization, human annotators manually identify individual utterances which, when extracted, would form a summary of the meeting. Such an approach is taken by Zhu and Penn [5] who extract a large number of features – lexical, structural, prosodic – and then train a binary classifier from the data. They experiment with various classifiers, but report best results with support vector machines and logistic regression. Murray, et al [4] also present a feature based approach, using Gaussian Mixture Models for each of the two classes. They compare this supervised approach to two unsupervised approaches – Maximum Marginal Relevance and an approach based on

Latent Semantic Analysis, and find only small differences in accuracy. Finally Maskey and Hirschberg [3] present a similar feature based extractive approach to broadcast news summarization. They compare Bayesian Network classifiers, Decision Trees and Support Vector Machines, and find Bayesian Network classifiers to be the most accurate.

4. APPROACH, DATA AND FEATURES

4.1. Overall Approach

Our goal is to learn to identify noteworthy utterances – utterances whose content one or more meeting participants are willing to include in the notes – over a sequence of related meetings. Intuitively we hope to learn the idiosyncratic words and other features in a meeting sequence that are correlated with noteworthiness. Towards this end, we have recorded sequences of weekly project meetings where participants take notes. We then manually identify those utterances in the meeting that are most closely related to these notes, and label them as noteworthy. Thus every utterance in the meeting sequence is labeled as either noteworthy or not. We then extract lexical and structural features and learn a binary noteworthy/not-noteworthy classifier similar to a typical meeting summarization approach. We evaluate this classifier on held out meetings in the same sequence.

4.2. Data Collection with SmartNotes

To collect such meeting data we have developed the SmartNotes [2] meeting note-taking system. Each participant in the meeting is equipped with a laptop running this system, and a head-mounted close-talking microphone connected to the laptop. The participant’s speech and all his notes are recorded, associated with his identity and time-stamped against a common NTP server. To reduce the effort of note-taking, notes are shared live between all the participants. We have deployed this system in regular weekly project meetings. These are natural meetings that would have taken place even if data collection wasn’t being performed. Further, participants took notes because they needed to remember some pieces of information and not because of the data collection effort. In this paper we use a single sequence of 5 consecutive longitudinal meetings. They were held approximately once a week in April and May 2006, were each half an hour long on average, and had 4 or 5 participants, three of whom attended each meeting. These meetings were all related to a single project, and contained discussions on related topics from meeting to meeting. As mentioned earlier, we use a single sequence because our goal is to adapt to the specifics of a single sequence of related meetings.

4.3. Data Annotation: Noteworthy Utterances

All the speech in the chosen meeting sequence was manually transcribed by humans. Next, each line of note across all the participants in the meetings was manually aligned with the fewest possible utterances such that the text of those utterances together contain all the information in the note. More than 98% of all the lines of notes could be aligned to between 1 and 5 utterances. For example the note “End pointer giving grief due to move to 16 khz models” is aligned to the utterances “Was the end pointer giving grief?” and “Yeah, that was because the model changed from 11,025 to 16khz”. The remaining 2% of lines of notes could not be aligned because they were high level summaries of discussions or were not actually spoken at the meetings. Note that because notes are shared live in SmartNotes, there were few lines of notes from different participants that referred to the same utterances. On average there were 22 lines of notes per meeting. Overall only 5% of all utterances were aligned with notes. This is a very low number compared to meeting summarization; in [4] the length of the summaries was 10% of the meeting. The alignment annotation was performed by one annotator only because the task is well defined, with little room for judgment calls. However, in the future we plan to evaluate whether this fact is true by having other humans do the alignment on a subset of the data.

4.4. Data Annotation: “Show-worthy” Utterances

Meeting participants often take scarce notes, resulting in a small number of noteworthy utterances in each meeting. A system that relies entirely on notes taken by participants to learn the concept would likely take too many meetings before it can provide real notes-assistance to the users. One option is to learn to identify utterances that *could* have resulted in a note, even if they were not aligned to any note in our data. We call such utterances show-worthy utterances because if the system can identify such utterances, they can be suggested as notes even before the participants have taken too many lines of notes. The participants’ acceptance / rejection feedback can then be used as additional data to learn to detect noteworthy utterances. Towards this end, we labeled each utterance as being show-worthy if it contained any information that could potentially be written as a note. This judgment was made very liberally, and only off-topic utterances (e.g. jokes) and utterances that are only related to the mechanics of the dialog (“What’s the agenda today?”, “Is this mic working?”) were labeled as not show-worthy.

4.5. Features Extracted from Utterances

As mentioned previously, we take a supervised binary classification approach to both problems of identifying noteworthy utterances and show-worthy utterances. We

experimented with both the Naïve Bayes classifier and with Decision Trees, but found superior performance with the latter; in this paper we only report on results using the Decision Tree classifier. We extract the following features from utterances.

Our main sets of features are word n-grams. In past approaches to meeting/speech summarization, the actual words are typically abstracted away to improve generalizability. For example, [5] and [3] use the words to identify named entities, and then use the number of named entities as features. Since our goal is to adapt to the specifics of a particular meeting sequence, and learn correlations between the words and noteworthiness of utterances, we use the words themselves. We use all n-grams with $n = 1$ to 6 that occur at least 5 times or more across all the meetings in the sequence. We designate all other n-grams that occur less than 5 times as Out of Vocabulary (OOV) words, and use the number of OOV words in each utterance as a feature.

Our second set of features is based on term frequency – inverse document frequency (TFIDF). Following [4], for each word w , we define document frequency $df(w)$ as the number of utterances in which w occurs, and term frequency $tf(w, i)$ as the number of times word w occurs in utterance i . $TFIDF(w, i)$ is then computed as $tf(w, i) * \log(N/df(w))$, where N is the total number of utterances in the meeting sequences. For each utterance we use the following four TFIDF based features: the maximum and minimum TFIDF scores in that utterance, and the average and standard deviation of the TFIDF scores in that utterance.

In addition to these lexical features we used two types of structural features. The first is speaker information – who spoke the current utterance, who spoke how much in the preceding set of utterances and who spoke immediately after. Note that while these features are similar to the anchor/reporter features used for summarization of broadcast news in [3], the difference is that we use the actual identity of the speakers themselves (similar to using the words themselves, without abstraction). Besides speaker information we also use the length of the current utterance as a feature, as has been used in previous work.

Two evaluation metrics have been used in speech summarization work in the past – ROUGE and f-measure (with precision and recall). ROUGE compares units of text (n-grams or words) from the system output to one or more reference summaries. Precision is computed as the number of true positives divided by the number of utterances labeled as noteworthy by the system, while recall divides the number of true positives by the total number of utterances manually marked as noteworthy. F-measure is computed as a harmonic mean of precision and recall, typically with equal weight on both values. F-measure, precision and recall are considered a more stringent evaluation mechanism than ROUGE; we use f-measure as the metric of evaluation here.

5. RESULTS AND ANALYSIS

5.1. Noteworthy Utterance Detection

The following results were obtained by doing a leave-one-out cross-validation at the level of meetings. That is, for each meeting in the sequence, we trained the noteworthy and show-worthy classifiers on the manually labeled data from all the remaining meetings in the sequence, and then evaluated this classifier on the test meeting. We average the test results over all the meetings, and present results here. Note that we do not perform any tuning of the parameters because of the small amount of available data. Also, while a deployed system would only have access to previously held meetings, here we use future meetings as well in order to maximize the yield from our data.

Using all the lexical and structural features mentioned above, the noteworthiness classifier achieves an overall classification accuracy of 91%. That is, 91% of the utterances are correctly labeled by the system as being noteworthy or not-noteworthy. Considering only the noteworthy utterances, we achieve precision of 0.15, recall of 0.12 and f-measure of 0.14. The high accuracy number of 91% is not surprising given the large skew in the data, and in fact is worse than the majority baseline that labels all utterances as non-noteworthy. Such a baseline would get an accuracy of 95%, but would result in 0 precision and recall. For the precision/recall/fmeasure metric, another trivial baseline is to simply label all the utterances as noteworthy. Such a method would achieve a recall of 1.0, but a precision of 0.05 (and an f-measure of 0.1). Thus, our algorithm triples the precision from this trivial baseline, and improves f-measure by 40% relative. (Because the data is limited, we do not report significance values).

5.2. Balancing Positive/Negative Examples

One of the problems of the data set as mentioned earlier is its skew – only 5% of the utterances are noteworthy. In order to see what effect this has on the accuracy of the classifier, we re-sampled the training data as follows. For each training set, we retained all the noteworthy utterances, and randomly picked non-noteworthy utterances (without replacement) until we had a desired ratio of noteworthy to non-noteworthy utterances. We trained the classifier on this training data, and then evaluated the classifier on the test data. We did not change the ratio of utterances in the test data. We did this entire process 10 times for each chosen ratio, each time picking non-noteworthy utterances randomly, and computed the average f-measure across the 10 trials. We performed this experiment using 5 different ratios; the results of which are presented in Table 1. The results show that while recall increases greatly, the precision remains stable. This implies that as the fraction of non-noteworthy utterances is reduced in the training data, many more utterances in the test data

previously labeled non-noteworthy by the system are labeled noteworthy. Additionally, utterances whose true labels are noteworthy and non-noteworthy are both labeled noteworthy by the system in equal proportions, resulting in the stable value of precision. Training data re-balancing is thus a useful way of changing the number of utterances that the system suggests to the user as a note, while keeping precision relatively stable.

Ratio of noteworthy to non-noteworthy utts	Precision	Recall	F-measure
1 : 0.5	0.09	0.70	0.16
1 : 1	0.10	0.51	0.16
1 : 2	0.10	0.39	0.16
1 : 5	0.09	0.20	0.13
1 : 10	0.10	0.14	0.11
~ 1 : 20 (entire data)	0.15	0.12	0.14

Table 1 Balancing Positive/Negative Training Examples

5.3. Other Analyses

In order to see the usefulness of each group of features, we performed training/testing using each feature group separately (and with the entire skewed training data). Interestingly, none of the TFIDF based features, nor the structural features – speaker and length information – are sufficient to learn a decision tree any better than a one-node stump that simply labels all utterances as “non-noteworthy”. This indicates that on their own these features do not have the discriminating power to overcome the skew in the training data. The n-gram features on the other hand do not result in a degenerate tree. Unigrams alone (without any other feature) achieve an f-measure score of 0.06. Unigrams and bigrams together achieve an f-measure of 0.08. Trigrams and higher n-grams do not seem to improve this f-measure. Thus individually, the features get a highest f-measure of 0.08, whereas when used together, they result in an f-measure of 0.14.

5.4. Show-worthy Utterance Detection

We used the same set of features to train a show-worthy utterance classifier. We followed the same hold-one-out protocol as for noteworthy utterances, and achieved an overall accuracy of 81% for both show-worthy and non-show-worthy utterances. This is a 28% absolute improvement over the majority baseline of 53%. Considering only show-worthy utterances, this algorithm achieves a precision of 0.81, recall of 0.78 and f-measure of 0.80. These results imply that the chosen features have good discriminative power for show-worthy utterances. In inspecting the decision trees learned, the topmost features are typically TFIDF features – tests on the maximum and average TFIDF scores of the utterance being classified.

While detecting show-worthy utterances has its own use (doing notes assistance before much note data is available), we investigated whether show-worthy information about an utterance can help noteworthy detection. There are two ways to use show-worthy information. The simple approach is to simply filter out all utterances in the test meeting labeled not show-worthy and then classify the remaining utterances as noteworthy or not. Doing so improves precision of noteworthy detection from 0.15 to 0.20, while the recall remains fixed at 0.12 (f-measure increases to 0.15). Another approach is to use the show-worthy information as a feature in learning the noteworthy classifier. This approach also resulted in an f-measure of 0.15, but with no increase in precision and a modest increase in recall – from 0.12 to 0.14. Thus the show-worthy information can result in a slight improvement in the quality of noteworthy detection. (In both these results we used the manual show-worthy labels). How useful show-worthy utterance detection is in eliciting accept/reject feedback from meeting participants is an avenue for future research.

7. CONCLUSIONS

In this paper we have investigated the effectiveness of extractive-summarization techniques when applied to the problem of noteworthy utterance detection. We have shown that these algorithms achieve an f-measure of 0.15, and that recall improves greatly when the training data is well-balanced. We have also introduced the concept of “show-worthy” utterances, and shown that such utterances can be accurately labeled. One of the limitations of the work here is the small amount of data. In future work, we plan to evaluate the ideas in this paper on multiple long meeting sequences. We also plan to experiment with sequence classification methods, as suggested by a reviewer.

8. REFERENCES

- [1] Banerjee, S., et. Al. (2005). The Necessity of a Meeting Recording and Playback System, and the Benefit of Topic-Level Annotations to Meeting Browsing. *Proc. of INTERACT*. Rome.
- [2] Banerjee, S., & Rudnicky, A. I. (2007). Segmenting Meetings into Agenda Items by Extracting Implicit Supervision from Human Note-Taking. *Proceedings of IUI 2007*. Honolulu, HI.
- [3] Maskey, S., & Hirschberg, J. (2005). Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. *Proceedings of Interspeech 2005*. Lisbon, Portugal.
- [4] Murray, G., Renals, S., & Carletta, J. (2005). Extractive Summarization of Meeting Recordings. *Proceedings of Interspeech 2005*. Lisbon, Portugal.
- [5] Zhu, X., & Penn, G. (2006). Summarization of Spontaneous Conversations. *Proceedings of Interspeech 2006*. Pittsburgh, PA.