Topic Summarization for Multiparty Meetings

Sonjia Waxmonsky
Department of Computer Science
University of Chicago
Chicago, IL 60637
wax@cs.uchicago.edu

Gina-Anne Levow
Department of Computer Science
University of Chicago
Chicago, IL 60637
levow@cs.uchicago.edu

Abstract

We consider methods of extracting summary utterances from multiparty meetings using features not based on lexical content. We first present an SVM-based classifier and then expand this into a two-pass system using an HMM to leverage utterance context.

1 Introduction

Extractive summarization selects important segments from a document to represent its main topics and ideas. We apply this task to multiparty dialogue. In particular, our goal is to select utterances that collectively present the major topics discussed during a meeting. We view summarization as a natural second step to topic segmentation where a single meeting is divided by topic into multiple segments.

Speech summarization presents challenges over text summarization. Methods developed for text rely heavily on lexical data, but for speech this requires expensive manual transcription or imperfect automatic speech recognition. Meeting data in particular is less structured than other speech genres, and so information such as topic shifts is not as easy to identify and use. Spontaneous dialogue also contains a high number of utterances that offer little content for a summary, such as asides, fragments, and backchannels. One advantage of speech data, however, is that we can leverage prosodic information to identify areas marked as important by the speaker.

In this work, we first build a SVM-based classifier using features that are not based on lexical content. We then feed SVM output into an HMM classifier where hidden states correspond to whether or not an utterance is part of the meeting summary. We use a sequential model because we expect that local utterance context will be helpful for the summarization task.

Feature Set	Description
Length	Utterance length (in words and sec), Speaking rate (words/sec)
Pos	Position of utterance from start of segment, expressed as a ratio
Acoustic	Pitch min, max, avg, (max-min); Intensity min, max, avg, (max-min)

Table 1: SVM Feature subsets

	Non-Summary	Summary
Non-Summary	98.01	1.97
Summary	55.34	44.61

Table 2: Transition probabilities for utterances in labeled data set by topic label.

2 Data Set

Our data set consists of 15 meetings from the ICSI Meeting Recorder corpus, a collection of naturally occurring meetings of speech researchers. Meetings used in this project have between 3 and 8 participants and are between 17 and 68 minutes in duration. Each meeting has already been hand-labeled with topic boundaries for use in topic segmentation research (Galley et al., 2003). We use these same topic boundaries in this report and expand on this

Feature Set	Strict Recall	ROUGE 1 Recall	ROUGE 1 F-meas.	ROUGE 1 Prec.	Coverage
Baseline					
Initial	48.2	63.9	56.5	54.5	8.9
Longest	8.5	34.6	39.2	53.2	10.1
SVM only					
Length+Pos	37.1	64.3	59.0	57.5	9.2
Length+Acoustic	10.4	47.9	45.0	44.3	9.3
Length+Acoustic+Pos	33.5	63.2	58.2	57.0	9.3
SVM + HMM					
Length+Pos	50.0	69.0	62.3	59.6	9.6
Length+Acoustic	21.8	38.1	44.6	65.4	6.2
Length+Acoustic+Pos	55.1	73.1	61.2	55.6	11.2

Table 3: Results for Baseline, SVM-only, and two-pass SVM+HMM classifier. Coverage is the duration of the meeting summary in seconds relative to the entire meeting. Results are averaged over all meetings.

data set by manually selecting utterances from each segment to be included in a meeting summary. As a guideline, the target length of a segment summary is approximately 10% of the segment duration, although this is not strictly followed. The main goal of annotation is to select utterances that represent the most important conversation points within a meeting segment, and otherwise our criteria is open-ended.

3 Utterance Classification

We automate the summarization process by building two binary classifiers that label each utterance as summary or non-summary. First, we use the LibSVM toolkit (Chang and Lin, 2001) to train multiple SVM classifiers using combinations of feature sets that are described in Table 1.

We use acoustic features to identify utterances where the speaker uses prosodic variation to indicate importance. In particular we include pitch, intensity, and speaking rate (in words/second). Similar features have been found to be useful in other speech processing tasks. For example, (Shriberg and Stolcke, 2004) describes how prosody can reduce error rates for tasks including topic segmentation and dialog act tagging, particularly in the presence of lexical ambiguity.

Our feature set also includes utterance length with the intuition that longer utterances have more content and thus are better summary candidates, and utterance position within the segment, since summary utterances tend to follow topic shifts in our data set.

Pitch and intensity features are extracted by aligning audio data with manual transcripts at the utterance level. Utterance length and speaking rate are taken directly from manual transcripts.

One challenge of an SVM is that our data set is very unbalanced, with less than 4% of utterances carrying the summary label. As a result our SVM classifier tends to apply the non-summary label almost exclusively. To overcome this we use the SVM to generate posterior probabilities rather than class labels and then select utterances from each segment with the highest probability of a summary label. LibSVM already includes functionality to compute posterior probabilities using methods presented in (Lin et al., 2003). As with hand annotation, the target summary length covers approximately 10% of the entire segment.

As Table 2 shows, summary utterances tend to precede other summary utterances in our data set, indicating that utterance context may be useful in this task. With this motivation, we build an HMM classifier to model the relationship between neighboring utterances. We order utterances sequentially by start time and use SVM posteriors as input to the HMM. Viterbi decoding can be used to find the most likely state sequence, but our unbalanced data set means that the most likely sequence is often one that only includes the non-summary label. We then constrain the Viterbi algorithm so that the returned sequence

must contain some summary utterances. Results from Viterbi are not constrained to cover exactly 10% of a segment, although we do enforce an upper limit of 20%. Utterances shorter than 0.25 seconds are automatically filtered out in this approach.

4 Evaluation and Results

We test our classifiers using 15-fold cross-validation, holding one meeting out for testing in each fold. Results are shown in Table 3. We use the ROUGE evaluation metric (Lin and Hovy, 2003) to account for the fact that multiple extractive summaries are possible for a single meeting. Here we apply ROUGE-1, which counts over word-unigrams. Results are also evaluated by strict recall, which counts the number of utterances in the reference summary that appear in the machine summary.

We construct two baseline summaries for each meeting. First, we select utterances from the start of each meeting segment until we have 10% coverage, and then do the same with longest utterances. We see that the initial-utterance baseline shows much better results and compares well with the feature-based approaches.

Looking at ROUGE-1 F-measure, we see that the SVM classifiers using the Position feature outperform the initial-utterance baseline, otherwise the baseline does better. The two-pass HMM system shows further improvements on the best performing SVMs (Length+Acoustic+Pos and Length+Pos). The HMM allows for longer summaries which can inflate recall scores, so we include Coverage data, or summary length as a percentage of meeting length, in Table 2. We see the upper limit of 20% is not always reached and average summary size approximates the baseline method. For individual meetings the HMMs summaries cover 3 to 19% across all feature sets. Examining HMM results shows that this approach tends tends to build summaries by taking utterances strictly from the start of each segment. However, results differs from the baseline method in where the end point of the summary is placed.

Table 4 shows the result from the HMM and Baseline for one meeting segment, and Table 5 shows the corresponding reference summary. We see that only the first utterance is matched exactly, but the concepts in the reference summary are still represented in the machine summary.

5 Related Works

(Murray et al., 2005) also considers the problem of summarization over the ICSI data set. This work compares purely lexical approaches from text retrieval (LSA and MMR) to feature-based methods that use both lexical and prosodic information. Evaluation was done over manual and ASR transcripts. Results showed that MMR and LSA were comparable to each other and outperformed the combined feature-based models. A later work on the same data set specifically examines discourse-based features such as listener feedback as discourse cues (Murray et al., 2006). Here, a system based on discourse and structural features performed better than lexical features alone as well as a combined system of both lexical and non-lexical features. This work used shorter utterances of 350 words, and interestingly anticipates that summary utterances will occur at the end of a segment rather than at the beginning as is the case in this report.

(Maskey and Hirschberg, 2006) applies HMMs to summarization of broadcast news data, also using only non-lexical features. This work incorporates utterance position by expanding the topology of the HMM, rather than including it as a feature, and shows a recall score on utterances of 95 in the best case. Earlier work on broadcast news (Maskey and Hirschberg, 2005) combines lexical features with prosodic, structural, and discourse features in a Bayes Network classifier. A combination of lexical and non-lexical features showed the best results, but a combination of all non-lexical features compared well with lexical features alone. As in (Murray et al., 2006) the most useful feature was found to be utterance length.

(Purver et al, 2006) considers the task of topic segmentation and identification on the ICSI data set using a purely lexical generative model; here the goal is to extract words that are good indicators of topic. (Hsueh and Moore, 2006) also looks at lexically based topic labeling over the AMI meeting corpus, where predefined topic labels are assigned.

Speaker	Utterance
8	**well y- - you didn't really talk about the non english speaker categories .**
0	**so ==**
3	**so there's another area which is non native english speakers which asks for native language region and variety of english.**
3	**and that's it .**
8	because it used to be that germans would learn british english and now i think there's you know is a certain percentage of them who - who've @reject@ to tend toward american english .
4	n- - detec- ==
0	what about proficiency in english as well ?
3	proficiency .
3	that's a good idea .
1	uh — i ==
8	that's hard though for self identification ==
0	you don't want to do it for political reasons ?
2	huh .

Table 4: Machine summary for segment 4 of meeting Bmr008 output by HMM approach with the (Length+Acoustic+Pos) feature set. The first four utterances (bold) form the shorter summary selected by the initial-utterance baseline method.

Speaker	Utterance
8	well y- - you didn't really talk about the non english speaker categories .
3	no i think proficiency is actually a good thing to have on this sort of form .
3	the - the question are - what are - what would the categories be ?
0	how long have you been in an english speaking country .
8	that's - that's non threatening and it's also an i- - a good indicator .

Table 5: Reference summary for segment 4 of meeting Bmr008

6 Conclusion

We have have presented methods for extracting summary sentences from multiparty meetings. We find that a sequential model identifies collocated summary utterances that are good summary candidates.

Our current approach assumes that a reliable segmentation is available, and data in this report has been segmented by hand. Our next step is to feed in machine-generated topic boundaries to evaluate the robustness of our system.

Going forward, we also plan to incorporate lexical information and to examine whether our sequential model can improve performance in this case.

References

Paul Boersma. PRAAT, a system for doing phonetics by computer. Glot International 5(9/10): 341-345.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL 2003*, pages 562-569.

Pei-Yun Hsueh and Johanna Moore. 2006. Automatic Topic Segmentation and Labeling in Multiparty Dialogue. In *Proceedings of the first IEEE/ACM workshop on Spoken Language Technology (SLT)*, pages 98-101.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 71-78.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. 2003. A note on Platt's probabilistic outputs for support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University.

Sameer Maskey and Julia Hirschberg. 2005. Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. In *Proceedings of Interspeech 2005*, pages 621-624.

Sameer Maskey and Julia Hirschberg. 2006. Summarizing Speech Without Text Using Hidden Markov Models. In *Proceedings of HLT-NAACL 2006*, pages 89-92.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005 Extractive summarization of meeting recordings. In *Proceedings of Interspeech 2005*, pages 593-596.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006 Incorporating speaker and discourse features into speech summarization. In *Proceedings of HLT-NAACL 2006*, pages 367-374.

Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. 2005. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of ACL 2006*, pages 17-24.

Elizabeth Shriberg and Andreas Stolcke. 2004. Prosody modeling for automatic speech recognition and understanding. In *Proceedings of the Workshop on Mathematical Foundations of Natural Language Modeling*, pages 105-114.