Automatic Segmentation and Summarization of Meeting Speech

Gabriel Murray, Pei-Yun Hsueh, Simon Tucker
Jonathan Kilgour, Jean Carletta, Johanna Moore, Steve Renals
University of Edinburgh
Edinburgh, Scotland
{gabriel.murray,p.hsueh}@ed.ac.uk

1 Introduction

AMI Meeting Facilitator is a system that performs topic segmentation and extractive summarisation. It consists of three components: (1) a segmenter that divides a meeting into a number of locally coherent segments, (2) a summarizer that selects the most important utterances from the meeting transcripts. and (3) a compression component that removes the less important words from each utterance based on the degree of compression the user specified. The goal of the AMI Meeting Facilitator is two-fold: first, we want to provide sufficient visual aids for users to interpret what is going on in a recorded meeting; second, we want to support the development of downstream information retrieval and information extraction modules with the information about the topics and summaries in meeting segments.

2 Component Description

2.1 Segmentation

The AMI Meeting Segmenter is trained using a set of 50 meetings that are seperate from the input meeting. We first extract features from the audio and video recording of the input meeting in order to train the Maximum Entropy (MaxEnt) models for classifying topic boundaries and non-topic boundaries. Then we test each utterance in the input meeting on the Segmenter to see if it is a topic boundary or not. The features we use include the following five categories: (1) **Conversational Feature**: These include a set of seven conversational features, including the amount of overlapping speech, the amount of silence between speaker segments, the level of similarity of speaker activity, the number of cue words, and the predictions of LCSEG (i.e., the lexical cohesion statistics, the estimated posterior probability, the predicted class). (2) **Lexical Feature**: Each spurt is represented as a vector space of uni-grams, wherein a vector is 1 or 0 depending on whether the cue word appears in the spurt. (3) **Prosodic Feature**: These include dialogue-act (DA) rate-of-speech, maximum F0 of the DA, mean energy of the DA, amount of silence in the DA, precedent and subsequent pauses, and duration of the DA. (4) **Motion Feature**: These include the average magnitude of speaker movements, which is measured by the number of pixels changed, over the frames of 40 ms within the spurt. (5) **Contextual Feature**: These include the dialogue act types and the speaker role (e.g., project manager, marketing expert). In the dialogue act annotations, each dialogue act is classified as one of the 15 types.

2.2 Summarization

The AMI summarizer is trained using a set of 98 scenario meetings. We train a support vector machine (SVM) on these meetings, using 26 features relating to the following categories: (1) **Prosodic Features**: These include dialogue-act (DA) rate-of-speech, maximum F0 of the DA, mean energy of the DA, amount of silence in the DA, precedent and subsequent pauses,

NAACL HLT Demonstration Program, pages 9–10,
Rochester, New York, USA, April 2007.

and duration of the DA. (2) **Speaker Features**: These features relate to how dominant the speaker is in the meeting as a whole, and they include percentage of the total dialogue acts which each speaker utters, percentage of total words which speaker utters, and amount of time in meeting that each person is speaking. (3) **Structural Features**: These features include the DA position in the meeting, and the DA position in the speaker's turn. (4) **Term Weighting Features**: We use two types of term weighting: *tf.idf*, which is based on words that are frequent in the meeting but rare across a set of other meetings or documents, and a second weighting feature which relates to how word usage varies between the four meeting participants.

After training the SVM, we test on each meeting of the 20 meeting test set in turn, ranking the dialogue acts from most probable to least probable in terms of being extract-worthy. Such a ranking allows the user to create a summary of whatever length she desires.

2.3 Compression

Each dialogue act has its constituent words scored using *tf.idf*, and as the user compresses the meeting to a greater degree the browser gradually removes the less important words from each dialogue act, leaving only the most informative material of the meeting.

3 Related Work

Previous work has explored the effect of lexical cohesion and conversational features on characterizing topic boundaries, following Galley et al.(2003). In previous work, we have also studied the problem of predicting topic boundaries at different levels of granularity and showed that a supervised classification approach performs better on predicting a coarser level of topic segmentation (Hsueh et al., 2006).

The amount of work being done on speech summarization has accelerated in recent years. Maskey and Hirschberg(September 2005) have explored speech summarization in the domain of Broadcast News data, finding that combining prosodic, lexical and structural features yield the best results. On the ICSI meeting corpus, Murray et al.(September 2005) compared applying text summarization approaches to feature-based approaches including prosodic features, while Galley(2006) used skip-chain Conditional Random Fields to model pragmatic dependencies between meeting utterances, and ranked meeting dialogue acts using a combination or prosodic, lexical, discourse and structural features.

4 acknowledgement

This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi- party Interaction, FP6-506811)

References

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.*

M. Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of EMNLP-06, Sydney, Australia.*

P. Hsueh, J. Moore, and S. Renals. 2006. Automatic segmentation of multiparty dialogue. In *the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.*

S. Maskey and J. Hirschberg. September 2005. Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization. In *Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal.*

G. Murray, S. Renals, and J. Carletta. September 2005. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal.*