Automatic Segmentation of Multiparty Dialogue

Pei-Yun Hsueh
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, GB
p.hsueh@ed.ac.uk

Johanna D. Moore
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, GB
J.Moore@ed.ac.uk

Steve Renals
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, GB
s.renals@ed.ac.uk

Abstract

In this paper, we investigate the problem of automatically predicting segment boundaries in spoken multiparty dialogue. We extend prior work in two ways. We first apply approaches that have been proposed for predicting top-level topic shifts to the problem of identifying subtopic boundaries. We then explore the impact on performance of using ASR output as opposed to human transcription. Examination of the effect of features shows that predicting top-level and predicting subtopic boundaries are two distinct tasks: (1) for predicting subtopic boundaries, the lexical cohesion-based approach alone can achieve competitive results, (2) for predicting top-level boundaries, the machine learning approach that combines lexical-cohesion and conversational features performs best, and (3) conversational cues, such as cue phrases and overlapping speech, are better indicators for the top-level prediction task. We also find that the transcription errors inevitable in ASR output have a negative impact on models that combine lexical-cohesion and conversational features, but do not change the general preference of approach for the two tasks.

1 Introduction

Text segmentation, i.e., determining the points at which the topic changes in a stream of text, plays an important role in applications such as topic detection and tracking, summarization, automatic genre detection and information retrieval and extraction (Pevzner and Hearst, 2002). In recent work, researchers have applied these techniques to corpora such as newswire feeds, transcripts of radio broadcasts, and spoken dialogues, in order to facilitate browsing, information retrieval, and topic detection (Allan et al., 1998; van Mulbregt et al., 1999; Shriberg et al., 2000; Dharanipragada et al., 2000; Blei and Moreno, 2001; Christensen et al., 2005). In this paper, we focus on segmentation of multiparty dialogues, in particular recordings of small group meetings. We compare models based solely on lexical information, which are common in approaches to automatic segmentation of text, with models that combine lexical and conversational features. Because tasks as diverse as browsing, on the one hand, and summarization, on the other, require different levels of granularity of segmentation, we explore the performance of our models for two tasks: hypothesizing where major topic changes occur and hypothesizing where more subtle nested topic shifts occur.

In addition, because we do not wish to make the assumption that high quality transcripts of meeting records, such as those produced by human transcribers, will be commonly available, we require algorithms that operate directly on automatic speech recognition (ASR) output.

2 Previous Work

Prior research on segmentation of spoken "documents" uses approaches that were developed for text segmentation, and that are based solely on textual cues. These include algorithms based on lexical cohesion (Galley et al., 2003; Stokes et al., 2004), as well as models using annotated features (e.g., cue phrases, part-of-speech tags, coreference relations) that have been determined to correlate with segment boundaries (Gavalda et al., 1997; Beeferman et al., 1999). Blei et al. (2001)

and van Mulbregt et al. (1999) use topic language models and variants of the hidden Markov model (HMM) to identify topic segments. Recent systems achieve good results for predicting topic boundaries when trained and tested on human transcriptions. For example, Stokes et al. (2004) report an error rate (Pk) of 0.25 on segmenting broadcast news stories using unsupervised lexical cohesion-based approaches. However, topic segmentation of multiparty dialogue seems to be a considerably harder task. Galley et al. (2003) report an error rate (Pk) of 0.319 for the task of predicting major topic segments in meetings.[1]

Although recordings of multiparty dialogue lack the distinct segmentation cues commonly found in text (e.g., headings, paragraph breaks, and other typographic cues) or news story segmentation (e.g., the distinction between anchor and interview segments), they contain conversation-based features that may be of use for automatic segmentation. These include silence, overlap rate, speaker activity change (Galley et al., 2003), and cross-speaker linking information, such as adjacency pairs (Zechner and Waibel, 2000). Many of these features can be expected to be complimentary. For segmenting spontaneous multiparty dialogue into major topic segments, Galley et al. (2003) have shown that a model integrating lexical and conversation-based features outperforms one based on solely lexical cohesion information.

However, the automatic segmentation models in prior work were developed for predicting top-level topic segments. In addition, compared to read speech and two-party dialogue, multi-party dialogues typically exhibit a considerably higher word error rate (WER) (Morgan et al., 2003). We expect that incorrectly recognized words will impair the robustness of lexical cohesion-based approaches and extraction of conversation-based discourse cues and other features. Past research on broadcast news story segmentation using ASR transcription has shown performance degradation from 5% to 38% using different evaluation metrics (van Mulbregt et al., 1999; Shriberg et al., 2000; Blei and Moreno, 2001). However, no prior study has reported directly on the extent of this degradation on the performance of a more subtle topic segmentation task and in spontaneous multiparty dialogue. In this paper, we extend prior work by investigating the effect of using ASR output on the models that have previously been proposed. In addition, we aim to find useful features and models for the subtopic prediction task.

[1]For the definition of Pk and Wd, please refer to section 3.4.1

3 Method

3.1 Data

In this study, we used the ICSI meeting corpus (LDC2004S02). Seventy-five natural meetings of ICSI research groups were recorded using close-talking far field head-mounted microphones and four desktop PZM microphones. The corpus includes human transcriptions of all meetings. We added ASR transcriptions of all 75 meetings which were produced by Hain (2005), with an average WER of roughly 30%.

The ASR system used a vocabulary of 50,000 words, together with a trigram language model trained on a combination of in-domain meeting data, related texts found by web search, conversational telephone speech (CTS) transcripts and broadcast news transcripts (about 10^9 words in total), resulting in a test-set perplexity of about 80. The acoustic models comprised a set of context-dependent hidden Markov models, using gaussian mixture model output distributions. These were initially trained on CTS acoustic training data, and were adapted to the ICSI meetings domain using maximum a posteriori (MAP) adaptation. Further adaptation to individual speakers was achieved using vocal tract length normalization and maximum likelihood linear regression. A four-fold cross-validation technique was employed: four recognizers were trained, with each employing 75% of the ICSI meetings as acoustic and language model training data, and then used to recognize the remaining 25% of the meetings.

3.2 Fine-grained and coarse-grained topics

We characterize a dialogue as a sequence of topical segments that may be further divided into subtopic segments. For example, the 60 minute meeting Bed003, whose theme is the planning of a research project on automatic speech recognition can be described by 4 major topics, from "opening" to "general discourse features for higher layers" to "how to proceed" to "closing". Depending on the complexity, each topic can be further divided into a number of subtopics. For example, "how to proceed" can be subdivided to 4 subtopic segments, "segmenting off regions of features",

"ad-hoc probabilities", "data collection" and "experimental setup".

Three human annotators at our site used a tailored tool to perform topic segmentation in which they could choose to decompose a topic into subtopics, with at most three levels in the resulting hierarchy. Topics are described to the annotators as what people in a meeting were talking about. Annotators were asked to provide a free text label for each topic segment; they were encouraged to use keywords drawn from the transcription in these labels, and we provided some standard labels for non-content topics, such as "opening" and "chitchat", to impose consistency. For our initial experiments with automatic segmentation at different levels of granularity, we flattened the subtopic structure and consider only two levels of segmentation–top-level topics and all subtopics.

To establish reliability of our annotation procedure, we calculated kappa statistics between the annotations of each pair of coders. Our analysis indicates human annotators achieve $\kappa = 0.79$ agreement on top-level segment boundaries and $\kappa = 0.73$ agreement on subtopic boundaries. The level of agreement confirms good replicability of the annotation procedure.

3.3 Probabilistic models

Our goal is to investigate the impact of ASR errors on the selection of features and the choice of models for segmenting topics at different levels of granularity. We compare two segmentation models: (1) an unsupervised lexical cohesion-based model (LM) using solely lexical cohesion information, and (2) feature-based combined models (CM) that are trained on a combination of lexical cohesion and conversational features.

3.3.1 Lexical cohesion-based model

In this study, we use Galley et al.'s (2003) LCSeg algorithm, a variant of TextTiling (Hearst, 1997). LCSeg hypothesizes that a major topic shift is likely to occur where strong term repetitions start and end. The algorithm works with two adjacent analysis windows, each of a fixed size which is empirically determined. For each utterance boundary, LCSeg calculates a lexical cohesion score by computing the cosine similarity at the transition between the two windows. Low similarity indicates low lexical cohesion, and a sharp change in lexical cohesion score indicates a high probability of an actual topic boundary. The principal difference between LCSeg and TextTiling is that LCSeg measures similarity in terms of lexical chains (i.e., term repetitions), whereas TextTiling computes similarity using word counts.

3.3.2 Integrating lexical and conversation-based features

We also used machine learning approaches that integrate features into a combined model, casting topic segmentation as a binary classification task. Under this supervised learning scheme, a training set in which each potential topic boundary[2] is labelled as either positive (POS) or negative (NEG) is used to train a classifier to predict whether each unseen example in the test set belongs to the class POS or NEG. Our objective here is to determine whether the advantage of integrating lexical and conversational features also improves automatic topic segmentation at the finer granularity of subtopic levels, as well as when ASR transcriptions are used.

For this study, we trained decision trees (c4.5) to learn the best indicators of topic boundaries. We first used features extracted with the optimal window size reported to perform best in Galley et al. (2003) for segmenting meeting transcripts into major topical units. In particular, this study uses the following features: (1) lexical cohesion features: the raw lexical cohesion score and probability of topic shift indicated by the sharpness of change in lexical cohesion score, and (2) conversational features: the number of cue phrases in an analysis window of 5 seconds preceding and following the potential boundary, and other interactional features, including similarity of speaker activity (measured as a change in probability distribution of number of words spoken by each speaker) within 5 seconds preceding and following each potential boundary, the amount of overlapping speech within 30 seconds following each potential boundary, and the amount of silence between speaker turns within 30 seconds preceding each potential boundary.

3.4 Evaluation

To compare to prior work, we perform a 25-fold leave-one-out cross validation on the set of 25 ICSI meetings that were used in Galley et

[2] In this study, the end of each speaker turn is a potential segment boundary. If there is a pause of more than 1 second within a single speaker turn, the turn is divided at the beginning of the pause creating a potential segment boundary.

al. (2003). We repeated the procedure to evaluate the accuracy using the lexical cohesion and combined models on both human and ASR transcriptions. In each evaluation, we trained the automatic segmentation models for two tasks: predicting subtopic boundaries (SUB) and predicting only top-level boundaries (TOP).

3.4.1 Evaluation metrics

In order to be able to compare our results directly with previous work, we first report our results using the standard error rate metrics of Pk and Wd. Pk (Beeferman et al., 1999) is the probability that two utterances drawn randomly from a document (in our case, a meeting transcript) are incorrectly identified as belonging to the same topic segment. WindowDiff (Wd) (Pevzner and Hearst, 2002) calculates the error rate by moving a sliding window across the meeting transcript counting the number of times the hypothesized and reference segment boundaries are different.

3.4.2 Baseline

To compute a baseline, we follow Kan (2003) and Hearst (1997) in using Monte Carlo simulated segments. For the corpus used as training data in the experiments, the probability of a potential topic boundary being an actual one is approximately 2.2% for all subtopic segments, and 0.69% for top-level topic segments. Therefore, the Monte Carlo simulation algorithm predicts that a speaker turn is a segment boundary with these probabilities for the two different segmentation tasks. We executed the algorithm 10,000 times on each meeting and averaged the scores to form the baseline for our experiments.

3.4.3 Topline

For the 24 meetings that were used in training, we have top-level topic boundaries annotated by coders at Columbia University (Col) and in our lab at Edinburgh (Edi). We take the majority opinion on each segment boundary from the Col annotators as reference segments. For the Edi annotations of top-level topic segments, where multiple annotations exist, we choose one randomly. The topline is then computed as the Pk score comparing the Col majority annotation to the Edi annotation.

4 Results

4.1 Experiment 1: Predicting top-level and subtopic segment boundaries

The meetings in the ICSI corpus last approximately 1 hour and have an average of 8-10 top-level topic segments. In order to facilitate meeting browsing and question-answering, we believe it is useful to include subtopic boundaries in order to narrow in more accurately on the portion of the meeting that contains the information the user needs. Therefore, we performed experiments aimed at analysing how the LM and CM segmentation models behave in predicting segment boundaries at the two different levels of granularity.

All of the results are reported on the test set. Table 1 shows the performance of the lexical cohesion model (LM) and the combined model (CM) integrating the lexical cohesion and conversational features discussed in Section 3.3.2.[3] For the task of predicting top-level topic boundaries from human transcripts, CM outperforms LM. LM tends to over-predict on the top-level, resulting in a higher false alarm rate. However, for the task of predicting subtopic shifts, LM alone is considerably better than CM.

Error Rate		Transcript		ASR	
Models		Pk	Wd	Pk	Wd
LM	SUB	32.31%	38.18%	32.91%	37.13%
(LCSeg)	TOP	36.50%	46.57%	38.02%	48.18%
CM	SUB	36.90%	38.68%	38.19%	n/a
(C4.5)	TOP	28.35%	29.52%	28.38%	n/a

Table 1: *Performance comparison of probabilistic segmentation models.*

In order to support browsing during the meeting or shortly thereafter, automatic topic segmentation will have to operate on the transcriptions produced by ASR. First note from Table 1 that the preference of models for segmentation at the two different levels of granularity is the same for ASR and human transcriptions. CM is better for predicting top-level boundaries and LM is better for predicting subtopic boundaries. This suggests that these

[3] We do not report Wd scores for the combined model (CM) on ASR output because this model predicted 0 segment boundaries when operating on ASR output. In our experience, CM routinely underpredicted the number of segment boundaries, and due to the nature of the Wd metric, it should not be used when there are 0 hypothesized topic boundaries.

are two distinct tasks, regardless of whether the system operates on human produced transcription or ASR output. Subtopics are better characterized by lexical cohesion, whereas top-level topic shifts are signalled by conversational features as well as lexical-cohesion based features.

4.1.1 Effect of feature combinations: predicting from human transcripts

Next, we wish to determine which features in the combined model are most effective for predicting topic segments at the two levels of granularity. Table 2 gives the average Pk for all 25 meetings in the test set, using the features described in Section 3.3.2. We group the features into four classes: (1) lexical cohesion-based features (LF): including lexical cohesion value (LCV) and estimated posterior probability (LCP); (2) interaction features (IF): the amount of overlapping speech (OVR), the amount of silence between speaker segments (GAP), similarity of speaker activity (ACT); (3) cue phrase feature (CUE); and (4) all available features (ALL). For comparison we also report the baseline (see Section 3.4.2) generated by Monte Carlo algorithm (MC-B). All of the models using one or more features from these classes outperform the baseline model. A one-way ANOVA revealed this reliable effect on the top-level segmentation ($F(7, 192) = 17.46, p < 0.01$) as well as on the subtopic segmentation task ($F(7, 192) = 5.862, p < 0.01$).

TRANSCRIPT Feature set	Error Rate(Pk) SUB	TOP
MC-B	46.61%	48.43%
LF(LCV+LCP)	38.13%	29.92%
IF(ACT+OVR+GAP)	38.87%	30.11%
IF+CUE	38.87%	30.11%
LF+ACT	38.70%	30.10%
LF+OVR	38.56%	29.48%
LF+GAP	38.50%	29.87%
LF+IF	38.11%	29.61%
LF+CUE	37.46%	29.18%
ALL(LF+IF+CUE)	36.90%	28.35%

Table 2: *Effect of different feature combinations for predicting topic boundaries from human transcripts. MC-B is the randomly generated baseline.*

As shown in Table 2, the best performing model for predicting top-level segments is the one using all of the features (ALL). This is not surprising, because these were the features that Galley et al. (2003) found to be most effective for predicting top-level segment boundaries in their combined model. Looking at the results in more detail, we see that when we begin with LF features alone and add other features one by one, the only model (other than ALL) that achieves significant[4] improvement ($p < 0.05$) over LF is LF+CUE, the model that combines lexical cohesion features with cue phrases.

When we look at the results for predicting subtopic boundaries, we again see that the best performing model is the one using all features (ALL). Models using lexical-cohesion features alone (LF) and lexical cohesion features with cue phrases (LF+CUE) both yield significantly better results than using interactional features (IF) alone ($p < 0.01$), or using them with cue phrase features (IF+CUE) ($p < 0.01$). Again, none of the interactional features used in combination with LF significantly improves performance. Indeed, adding speaker activity change (LF+ACT) degrades the performance ($p < 0.05$).

Therefore, we conclude that for predicting both top-level and subtopic boundaries from human transcriptions, the most important features are the lexical cohesion based features (LF), followed by cue phrases (CUE), with interactional features contributing to improved performance only when used in combination with LF and CUE.

However, a closer look at the Pk scores in Table 2, adds further evidence to our hypothesis that predicting subtopics may be a different task from predicting top-level topics. Subtopic shifts occur more frequently, and often without clear conversational cues. This is suggested by the fact that absolute performance on subtopic prediction degrades when any of the interactional features are combined with the lexical cohesion features. In contrast, the interactional features slightly improve performance when predicting top-level segments. Moreover, the fact that the feature OVR has a positive impact on the model for predicting top-level topic boundaries, but does not improve the model for predicting subtopic boundaries reveals that having less overlapping speech is a more prominent phenomenon in major topic shifts than

[4]Because we do not wish to make assumptions about the underlying distribution of error rates, and error rates are not measured on an interval level, we use a non-parametric sign test throughout these experiments to compute statistical significance.

in subtopic shifts.

4.1.2 Effect of feature combinations: predicting from ASR output

Features extracted from ASR transcripts are distinct from those extracted from human transcripts in at least three ways: (1) incorrectly recognized words incur erroneous lexical cohesion features (LF), (2) incorrectly recognized words incur erroneous cue phrase features (CUE), and (3) the ASR system recognizes less overlapping speech (OVR).

In contrast to the finding that integrating conversational features with lexical cohesion features is useful for prediction from human transcripts, Table 3 shows that when operating on ASR output, neither adding interactional nor cue phrase features improves the performance of the model using only lexical cohesion features. In fact, the model using all features (ALL) is significantly worse than the model using only lexical cohesion based features (LF). This suggests that we must explore new features that can lessen the perplexity introduced by ASR outputs in order to train a better model.

ASR Feature set	Error Rate(Pk) SUB	TOP
MC-B	43.41%	45.22%
LF(LCV+LCP)	36.83%	25.27%
IF(ACT+OVR+GAP)	36.83%	25.27%
IF+CUE	36.83%	25.27%
LF+GAP	36.67%	24.62%
LF+IF	36.83%	28.24%
LF+CUE	37.42%	25.27%
ALL(LF+IF+CUE)	38.19%	28.38%

Table 3: *Effect of different feature combinations for predicting topic boundaries from ASR output.*

4.2 Experiment 2: Statistically learned cue phrases

In prior work, Galley et al. (2003) empirically identified cue phrases that are indicators of segment boundaries, and then eliminated all cues that had not previously been identified as cue phrases in the literature. Here, we conduct an experiment to explore how different ways of identifying cue phrases can help identify useful new features for the two boundary prediction tasks.

In each fold of the 25-fold leave-one-out cross validation, we use a modified[5] Chi-square test to calculate statistics for each word (unigram) and word pair (bi-gram) that occurred in the 24 training meetings. We then rank unigrams and bigrams according to their Chi-square scores, filtering out those with values under 6.64, the threshold for the Chi-square statistic at the 0.01 significance level. The unigrams and bigrams in this ranked list are the learned cue phrases. We then use the occurrence counts of cue phrases in an analysis window around each potential topic boundary in the test meeting as a feature.

Table 4 shows the performance of models that use statistically learned cue phrases in their feature sets compared with models using no cue phrase features and Galley's model, which only uses cue phrases that correspond to those identified in the literature (Col-cue). We see that for predicting subtopics, models using the cue word features (1gram) and the combination of cue words and bigrams (1+2gram) yield a 15% and 8.24% improvement over models using no cue features (NOCUE) ($p < 0.01$) respectively, while models using only cue phrases found in the literature (Col-cue) improve performance by just 3.18%. In contrast, for predicting top-level topics, the model using cue phrases from the literature (Col-cue) achieves a 4.2% improvement, and this is the only model that produces statistically significantly better results than the model using no cue phrases (NOCUE). The superior performance of models using statistically learned cue phrases as features for predicting subtopic boundaries suggests there may exist a different set of cue phrases that serve as segmentation cues for subtopic boundaries.

5 Discussion

As observed in the corpus of meetings, the lack of macro-level segment units (e.g., story breaks, paragraph breaks) makes the task of segmenting spontaneous multiparty dialogue, such as meetings, different from segmenting text or broadcast news. Compared to the task of segmenting expository texts reported in Hearst (1997) with a 39.1% chance of each paragraph end being a target topic boundary, the chance of each speaker turn being a top-level or sub-topic boundary in our ICSI corpus is just 2.2% and 0.69%. The imbalanced class distribution has a negative effect on the per-

[5]In order to satisfy the mathematical assumptions underlying the test, we removed cases with an expected value that is under a threshold (in this study, we use 1), and we apply Yate's correction, $(|ObservedValue - ExpectedValue| - 0.5)^2/ExpectedValue$.

	NOCUE	Col-cue	1gram	2gram	1+2gram	MC-B	Topline
SUB	38.11%	36.90%	32.39%	36.86%	34.97%	46.61%	n/a
TOP	29.61%	28.35%	28.95%	29.20%	29.27%	48.43%	13.48%

Table 4: *Performance of models trained with cue phrases from the literature (Col-cue) and cue phrases learned from statistical tests, including cue words (1gram), cue word pairs (2gram), and cue phrases composed of both words and word pairs (1+2gram). NOCUE is the model using no cue phrase features. The Topline is the agreement of human annotators on top-level segments.*

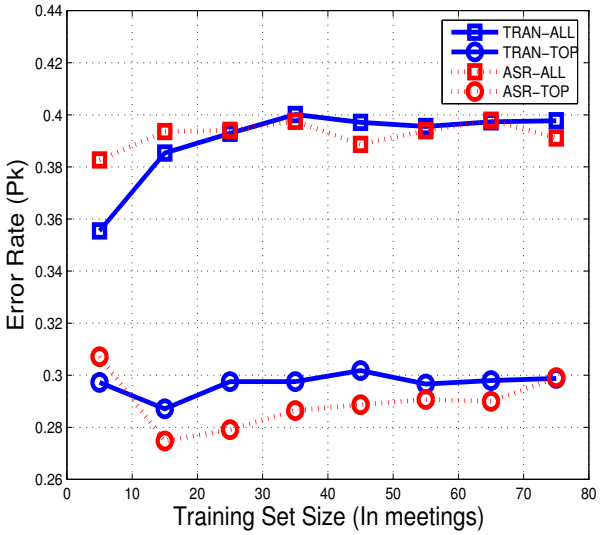

Figure 1: *Performance of the combined model over the increase of the training set size.*

formance of machine learning approaches. In a pilot study, we investigated sampling techniques that rebalance the class distribution in the training set. We found that sampling techniques previously reported in Liu et al (2004) as useful for dealing with an imbalanced class distribution in the task of disfluency detection and sentence segmentation do not work for this particular data set. The implicit assumption of some classifiers (such as pruned decision trees) that the class distribution of the test set matches that of the training set, and that the costs of false positives and false negatives are equivalent, may account for the failure of these sampling techniques to yield improvements in performance, when measured using Pk and Wd.

Another approach that copes with the imbalanced class prediction problem but does not change the natural class distribution is to increase the size of the training set. We conducted an experiment in which we incrementally increased the training set size by randomly choosing ten meetings each time until all meetings were selected. We executed the process three times and averaged the scores to obtain the results shown in Figure 1. However, increasing training set size adds to the perplexity in the training phase. We see that increasing the size of the training set only improves the accuracy of segment boundary prediction for predicting top-level topics on ASR output. The figure also indicates that training a model to predict top-level boundaries requires no more than fifteen meetings in the training set to reach a reasonable level of performance.

6 Conclusions

Discovering major topic shifts and finding nested subtopics are essential for the success of speech document browsing and retrieval. Meeting records contain rich information, in both content and conversation behavioral form, that enable automatic topic segmentation at different levels of granularity. The current study demonstrates that the two tasks – predicting top-level and subtopic boundaries – are distinct in many ways: (1) for predicting subtopic boundaries, the lexical cohesion-based approach achieves results that are competitive with the machine learning approach that combines lexical and conversational features; (2) for predicting top-level boundaries, the machine learning approach performs the best; and (3) many conversational cues, such as overlapping speech and cue phrases discussed in the literature, are better indicators for top-level topic shifts than for subtopic shifts, but new features such as cue phrases can be learned statistically for the subtopic prediction task. Even in the presence of a relatively higher word error rate, using ASR output makes no difference to the preference of model for the two tasks. The conversational features also did not help improve the performance for predicting from ASR output.

In order to further identify useful features for automatic segmentation of meetings at different levels of granularity, we will explore the use of

multimodal, i.e., acoustic and visual, cues. In addition, in the current study, we only extracted features from within the analysis windows immediately preceding and following each potential topic boundary; we will explore models that take into account features of longer range dependencies.

7 Acknowledgements

Many thanks to Jean Carletta for her invaluable help in managing the data, and for advice and comments on the work reported in this paper. Thanks also to the AMI ASR group for producing the ASR transcriptions, and to the anonymous reviewers for their helpful comments. This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811).

References

J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

D. Beeferman, A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34:177–210.

D. M. Blei and P. J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press.

H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. 2005. Maximum entropy segmentation of broadcast news. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, Philadelphia, USA.

S. Dharanipragada, M. Franz, J.S. McCarley, K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2000. Statistical methods for topic segmentation. In *Proceedings of the International Conference on Spoken Language Processing*, pages 516–519.

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

M. Gavalda, K. Zechner, and G. Aist. 1997. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Proceedings of the Fifth ANLP Conference*, pages 12–15.

T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. 2005. Transcription of conference room meetings: an investigation. In *Proceedings of Interspeech*.

M. Hearst. 1997. Texttiling: Segmenting text into multiparagraph subtopic passages. *Computational Linguistics*, 25(3):527–571.

M. Kan. 2003. *Automatic text summarization as applied to information retrieval: Using indicative and informative summaries*. Ph.D. thesis, Columbia University, New York USA.

Y. Liu, E. Shriberg, A. Stolcke, and M. Harper. 2004. Using machine learning to cope with imbalanced classes in natural sppech: Evidence from sentence boundary and disfluency detection. In *Proceedings of the Intl. Conf. Spoken Language Processing*.

N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, , and C. Wooters. 2003. Meetings about meetings: research at icsi on speech in multiparty conversations. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*.

L. Pevzner and M. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

E. Shriberg, A. Stolcke, D. Hakkani-tur, and G. Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communications*, 31(1-2):127–254.

N. Stokes, J. Carthy, and A.F. Smeaton. 2004. Select: a lexical cohesion based news story segmentation system. *AI Communications*, 17(1):3–12, January.

P. van Mulbregt, J. Carp, L. Gillick, S. Lowe, and J. Yamron. 1999. Segmentation of automatically transcribed broadcast news text. In *Proceedings of the DARPA Broadcast News Workshop*, pages 77–80. Morgan Kaufman Publishers.

Klaus Zechner and Alex Waibel. 2000. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000*, pages 968–974.