Combining Multiple Knowledge Sources for Dialogue Segmentation in Multimedia Archives

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

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

Abstract

Automatic segmentation is important for making multimedia archives comprehensible, and for developing downstream information retrieval and extraction modules. In this study, we explore approaches that can segment multiparty conversational speech by integrating various knowledge sources (e.g., words, audio and video recordings, speaker intention and context). In particular, we evaluate the performance of a Maximum Entropy approach, and examine the effectiveness of multimodal features on the task of dialogue segmentation. We also provide a quantitative account of the effect of using ASR transcription as opposed to human transcripts.

1 Introduction

Recent advances in multimedia technologies have led to huge archives of audio-video recordings of multiparty conversations in a wide range of areas including clinical use, online video sharing services, and meeting capture and analysis. While it is straightforward to replay such recordings, finding information from the often lengthy archives is a more challenging task. Annotating implicit semantics to enhance browsing and searching of recorded conversational speech has therefore posed new challenges to the field of multimedia information retrieval.

One critical problem is how to divide unstructured conversational speech into a number of locally coherent segments. The problem is important for two reasons: First, empirical analysis has shown that annotating transcripts with semantic information (e.g., topics) enables users to browse and find information from multimedia archives more efficiently (Banerjee et al., 2005). Second, because the automatically generated segments make up for the lack of explicit orthographic cues (e.g., story and paragraph breaks) in conversational speech, dialogue segmentation is useful in many spoken language understanding tasks, including anaphora resolution (Grosz and Sidner, 1986), information retrieval (e.g., as input for the TREC Spoken Document Retrieval (SDR) task), and summarization (Zechner and Waibel, 2000).

This study therefore aims to explore whether a Maximum Entropy (MaxEnt) classifier can integrate multiple knowledge sources for segmenting recorded speech. In this paper, we first evaluate the effectiveness of features that have been proposed in previous work, with a focus on features that can be extracted automatically. Second, we examine other knowledge sources that have not been studied systematically in previous work, but which we expect to be good predictors of dialogue segments. In addition, as our ultimate goal is to develop an information retrieval module that can be operated in a fully automatic fashion, we also investigate the impact of automatic speech recognition (ASR) errors on the task of dialogue segmentation.

2 Previous Work

In previous work, the problem of automatic dialogue segmentation is often considered as similar to the problem of topic segmentation. Therefore, research has adopted techniques previously developed

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 1016–1023,
Prague, Czech Republic, June 2007.

to segment topics in text (Kozima, 1993; Hearst, 1997; Reynar, 1998) and in read speech (e.g., broadcast news) (Ponte and Croft, 1997; Allan et al., 1998). For example, lexical cohesion-based algorithms, such as LCSEG (Galley et al., 2003), or its word frequency-based predecessor TextTile (Hearst, 1997) capture topic shifts by modeling the similarity of word repetition in adjacent windows.

However, recent work has shown that LCSEG is less successful in identifying "agenda-based conversation segments" (e.g., *presentation*, *group discussion*) that are typically signalled by differences in group activity (Hsueh and Moore, 2006). This is not surprising since LCSEG considers only lexical cohesion. Previous work has shown that training a segmentation model with features that are extracted from knowledge sources other than words, such as speaker interaction (e.g., overlap rate, pause, and speaker change) (Galley et al., 2003), or participant behaviors, e.g., note taking cues (Banerjee and Rudnicky, 2006), can outperform LCSEG on similar tasks.

In many other fields of research, a variety of features have been identified as indicative of segment boundaries in different types of recorded speech. For example, Brown et al. (1980) have shown that a discourse segment often starts with relatively high pitched sounds and ends with sounds of pitch within a more compressed range. Passonneau and Litman (1993) identified that topic shifts often occur after a pause of relatively long duration. Other prosodic cues (e.g., pitch contour, energy) have been studied for their correlation with story segments in read speech (Tur et al., 2001; Levow, 2004; Christensen et al., 2005) and with theory-based discourse segments in spontaneous speech (e.g., direction-given monologue) (Hirschberg and Nakatani, 1996). In addition, head and hand/forearm movements are used to detect group-action based segments (McCowan et al., 2005; Al-Hames et al., 2005).

However, many other features that we expect to signal segment boundaries have not been studied systematically. For instance, speaker intention (i.e., dialogue act types) and conversational context (e.g., speaker role). In addition, although these features are expected to be complementary to one another, few of the previous studies have looked at the question how to use conditional approaches to model the correlation among features.

3 Methodology

3.1 Meeting Corpus

This study aims to explore approaches that can integrate multimodal information to discover implicit semantics from conversation archives. As our goal is to identify multimodal cues of segmentation in face-to-face conversation, we use the AMI meeting corpus (Carletta et al., 2006), which includes audio-video recordings, to test our approach. In particular, we are using 50 scenario-based meetings from the AMI corpus, in which participants are assigned to different roles and given specific tasks related to designing a remote control. On average, AMI meetings last 26 minutes, with over 4,700 words transpired. This corpus includes annotation for dialogue segmentation and topic labels. In the annotation process, annotators were given the freedom to subdivide a segment into subsegments to indicate when the group was discussing a subtopic. Annotators were also given a set of segment descriptions to be used as labels. Annotators were instructed to add a new label only if they could not find a match in the standard set. The set of segment descriptions can be divided to three categories: activity-based (e.g., presentation, discussion), issue-based (e.g., budget, usability), and functional segments (e.g., chitchat, opening, closing).

3.2 Preprocessing

The first step is to break a recorded meeting into minimal units, which can vary from sentence chunks to blocks of sentences. In this study, we use *spurts*, that is, consecutive speech with no pause longer than 0.5 seconds, as minimal units.

Then, to examine the difference between the set of features that are characteristic of segmentation at both coarse and fine levels of granularity, this study characterizes a dialogue as a sequence of segments that may be further divided into sub-segments. We take the theory-free dialogue segmentation annotations in the corpus and flatten the sub-segment structure and consider only two levels of segmentation: top-level segments and all sub-level segments.[1] We

[1] We take the spurts which the annotators choose as the beginning of a segment as the topic boundaries. On average,

observed that annotators tended to annotate activity-based segments only at the top level, whereas they often included sub-topics when segmenting issue-based segments. For example, a top-level *interface specialist presentation* segment can be divided into *agenda/equipment issues*, *user requirements*, *existing products*, and *look and usability* sub-level segments.

3.3 Intercoder Agreement

To measure intercoder agreement, we employ three different metrics: the kappa coefficient, PK, and WD. Kappa values measure how well a pair of annotators agree on where the segments break. PK is the probability that two spurts drawn randomly from a document are incorrectly identified as belonging to the same segment. WindowDiff (WD) calculates the error rate by moving a sliding window across the transcript counting the number of times the hypothesized and reference segment boundaries are different. While not uncontroversial, the use of these metrics is widespread. Table 1 shows the intercoder agreement of the top-level and sub-level segmentation respectively.

It is unclear whether the kappa values shown here indicate reliable intercoder agreement.[2] But given the low disagreement rate among codings in terms of the PK and WD scores, we will argue for the reliability of the annotation procedure used in this study. Also, to our knowledge the reported degree of agreement is the best in the field of meeting dialogue segmentation.[3]

Intercoder	Kappa	PK	WD
TOP	0.66	0.11	0.17
SUB	0.59	0.23	0.28

Table 1: Intercoder agreement of annotations at the top-level (TOP) and sub-level (SUB) segments.

the annotators marked 8.7 top-level segments and 14.6 sub-segments per meeting.

[2] In computational linguistics, kappa values over 0.67 point to reliable intercoder agreement. But Di Eugenio and Glass (2004) have found that this interpretation does not hold true for all tasks.

[3] For example, Gruenstein et al.(2005) report kappa (PK/WD) of 0.41(0.28/0.34) for determining the top-level and 0.45(0.27/0.35) for the sub-level segments in the ICSI meeting corpus.

3.4 Feature Extraction

As reported in Section 2, there is a wide range of features that are potentially characteristic of segment boundaries, and we expect to find some of them useful for automatic recognition of segment boundaries. The features we explore can be divided into the following five classes:

Conversational Features: We follow Galley et al. (2003) and extracted a set of conversational features, including the amount of overlapping speech, the amount of silence between speaker segments, speaker activity change, the number of cue words, and the predictions of LCSEG (i.e., the lexical cohesion statistics, the estimated posterior probability, the predicted class).

Lexical Features: We compile the list of words that occur more than once in the spurts that have been marked as a top-level or sub-segment boundary in the training set. Each spurt is then represented as a vector space of unigrams from this list.

Prosodic Features: We use the direct modelling approach proposed in Shriberg and Stolcke (2001) and include maximum F0 and energy of the spurt, mean F0 and energy of the spurt, pitch contour (i.e., slope) and energy at multiple points (e.g., the first and last 100 and 200 ms, the first and last quarter, the first and second half) of a spurt. We also include rate of speech, in-spurt silence, preceding and subsequent pauses, and duration. The rate of speech is calculated as both the number of words and the number of syllables spoken per second.

Motion Features: We measure the magnitude of relevant movements in the meeting room using methods that detect movements directly from video recordings in frames of 40 ms. Of special interest are the frontal shots as recorded by the close up cameras, the hand movements as recorded by the overview cameras, and shots of the areas of the room where presentations are made. We then average the magnitude of movements over the frames within a spurt as its feature value.

Contextual Features: These include dialogue act type[4] and speaker role (e.g., project manager, mar-

[4] In the annotations, each dialogue act is classified as one of 15 types, including acts about information exchange (e.g., Inform), acts about possible actions (e.g., Suggest), acts whose primary purpose is to smooth the social functioning (e.g., Be-positive), acts that are commenting on previous discussion (e.g.,

keting expert). As each spurt may consist of multiple dialogue acts, we represent each spurt as a vector of dialogue act types, wherein a component is 1 or 0 depending on whether the type occurs in the spurt.

3.5 Multimodal Integration Using Maximum Entropy Models

Previous work has used MaxEnt models for sentence and topic segmentation and shown that conditional approaches can yield competitive results on these tasks (Christensen et al., 2005; Hsueh and Moore, 2006). In this study, we also use a MaxEnt classifier[5] for dialogue segmentation under the typical supervised learning scheme, that is, to train the classifier to maximize the conditional likelihood over the training data and then to use the trained model to predict whether an unseen spurt in the test set is a segment boundary or not. Because continuous features have to be discretized for MaxEnt, we applied a histogram binning approach, which divides the value range into N intervals that contain an equal number of counts as specified in the histogram, to discretize the data.

4 Experimental Results

4.1 Probabilistic Models

The first question we want to address is whether the different types of characteristic multimodal features can be integrated, using the conditional MaxEnt model, to automatically detect segment boundaries. In this study, we use a set of 50 meetings, which consists of 17,977 spurts. Among these spurts, only 1.7% and 3.3% are top-level and subsegment boundaries. For our experiments we use 10-fold cross validation. The baseline is the result obtained by using LCSEG, an unsupervised approach exploiting only lexical cohesion statistics.

Table 2 shows the results obtained by using the same set of conversational (CONV) features used in previous work (Galley et al., 2003; Hsueh and Moore, 2006), and results obtained by using all the available features (ALL). The evaluation metrics PK and WD are conventional measures of error rates in segmentation (see Section 3.3). In Row 2, we see that using a MaxEnt classifier trained on the conversational features (CONV) alone improves over the LCSEG baseline by 15.3% for top-level segments and 6.8% for sub-level segements. Row 3 shows that combining additional knowledge sources, including lexical features (LX1) and the non-verbal features, prosody (PROS), motion (MOT), and context (CTXT), yields a further improvement (of 8.8% for top-level segmentation and 5.4% for sub-level segmentation) over the model trained on conversational features.

	TOP		SUB	
Error Rate	PK	WD	PK	WD
BASELINE(LCSEG)	0.40	0.49	0.40	0.47
MAXENT(CONV)	0.34	0.34	0.37	0.37
MAXENT(ALL)	0.30	0.33	0.34	0.36

Table 2: Compare the result of MaxEnt models trained with only conversational features (CONV) and with all available features (ALL).

4.2 Feature Effects

The second question we want to address is which knowledge sources (and combinations) are good predictors for segment boundaries. In this round of experiments, we evaluate the performance of different feature combinations. Table 3 further illustrates the impact of each feature class on the error rate metrics (PK/WD). In addition, as the PK and WD score do not reflect the magnitude of over- or underprediction, we also report on the average number of hypothesized segment boundaries (Hyp). The number of reference segments in the annotations is 8.7 at the top-level and 14.6 at the sub-level.

Rows 2-6 in Table 3 show the results of models trained with each individual feature class. We performed a one-way ANOVA to examine the effect of different feature classes. The ANOVA suggests a reliable effect of feature class ($F(5, 54) = 36.1$; $p < .001$). We performed post-hoc tests (Tukey HSD) to test for significant differences. Analysis shows that the model that is trained with lexical features alone (LX1) performs significantly worse than the LCSEG baseline ($p < .001$). This is due to the fact that cue words, such as *okay* and *now*, learned from the training data to signal seg-

Elicit-Assessment), and acts that allow complete segmentation (e.g., Stall).

[5] The parameters of the MaxEnt classifier are optimized using Limited-Memory Variable Metrics.

	TOP			SUB		
	Hyp	PK	WD	Hyp	PK	WD
BASELINE (LCSEG)	17.6	0.40	0.49	17.6	0.40	0.47
LX1	61.2	0.53	0.72	65.1	0.49	0.66
CONV	3.1	0.34	0.34	2.9	0.37	0.37
PROS	2.3	0.35	0.35	2.5	0.37	0.37
MOT	96.2	0.36	0.40	96.2	0.38	0.41
CTXT	2.6	0.34	0.34	2.2	0.37	0.37
ALL	7.7	0.29	0.33	7.6	0.35	0.38

Table 3: Effects of individual feature classes and their combination on detecting segment boundaries.

ment boundaries, are often used for non-discourse purposes, such as making a semantic contribution to an utterance.[6] Thus, we hypothesize that these ambiguous cue words have led the LX1 model to overpredict. Row 7 further shows that when all available features (including LX1) are used, the combined model (ALL) yields performance that is significantly better than that obtained with individual feature classes ($F(5, 54) = 32.2$; $p < .001$).

	TOP			SUB		
	Hyp	PK	WD	Hyp	PK	WD
ALL	7.7	0.29	0.33	7.6	0.35	0.38
ALL-LX1	3.9	0.35	0.35	3.5	0.37	0.38
ALL-CONV	6.6	0.30	0.34	6.8	0.35	0.37
ALL-PROS	5.6	0.29	0.31	7.4	0.33	0.35
ALL-MOTION	7.5	0.30	0.35	7.3	0.35	0.37
ALL-CTXT	7.2	0.29	0.33	6.7	0.36	0.38

Table 4: Performance change of taking out each individual feature class from the ALL model.

Table 4 illustrates the error rate change (i.e., increased or decreased PK and WD score)[7] that is incurred by leaving out one feature class from the ALL model. Results show that CONV, PROS, MOTION and CTXT can be taken out from the ALL model individually without increasing the error rate significantly.[8] Morevoer, the combined models always perform better than the LX1 model ($p < .01$), cf. Table 3.

This suggests that the non-lexical feature classes are complementary to LX1, and thus it is essential to incorporate some, but not necessarily all, of the non-lexical classes into the model.

	TOP			SUB		
	Hyp	PK	WD	Hyp	PK	WD
LX1	61.2	0.53	0.72	65.1	0.49	0.66
MOT	96.2	0.36	0.40	96.2	0.38	0.41
LX1+CONV	5.3	0.27	0.30	6.9	0.32	0.35
LX1+PROS	6.2	0.30	0.33	7.3	0.36	0.38
LX1+MOT	20.2	0.39	0.49	24.8	0.39	0.47
LX1+CTXT	6.3	0.28	0.31	7.2	0.33	0.35
MOT+PROS	62.0	0.34	0.34	62.1	0.37	0.37
MOT+CTXT	2.7	0.33	0.33	2.3	0.37	0.37

Table 5: Effects of combining complementary features on detecting segment boundaries.

Table 5 further illustrates the performance of different feature combinations on detecting segment boundaries. By subtracting the PK or WD score in Row 1, the LX1 model, from that in Rows 3-6, we can tell how essential each of the non-lexical classes is to be combined with LX1 into one model. Results show that CONV is the most essential, followed by CTXT, PROS and MOT. The advantage of incorporating the non-lexical feature classes is also shown in the noticeably reduced number of overpredictions as compared to that of the LX1 model.

To analyze whether there is a significant interaction between feature classes, we performed another round of ANOVA tests to examine the effect of LX1 and each of the non-lexical feature classes on detecting segment boundaries. This analysis shows that there is a significant interaction effect on detecting both top-level and sub-level segment boundaries ($p < .01$), suggesting that the performance of LX1 is significantly improved when combined with any non-lexical feature class. Also, among the non-lexical feature classes, combining prosodic features significantly improves the performance of the model in which the motion features are combined to detect top-level segment boundaries ($p < .05$).

[6] Hirschberg and Litman (1987) have proposed to discriminate the different uses intonationally.

[7] Note that the increase in error rate indicates performance degradation, and vice versa.

[8] Sign tests were used to test for significant differences between means in each fold of cross validation.

4.3 Degradation Using ASR

The third question we want to address here is whether using the output of ASR will cause significant degradation to the performance of the segmentation approaches. The ASR transcripts used in this experiment are obtained using standard technology including HMM based acoustic modeling and N-gram based language models (Hain et al., 2005). The average word error rates (WER) are 39.1%. We also applied a word alignment algorithm to ensure that the number of words in the ASR transcripts is the same as that in the human-produced transcripts. In this way we can compare the PK and WD metrics obtained on the ASR outputs directly with that on the human transcripts.

In this study, we again use a set of 50 meetings and 10-fold cross validation. We compare the performance of the reference models, which are trained on human transcripts and tested on human transcripts, with that of the ASR models, which are trained on ASR transcripts and tested on ASR transcripts. Table 6 shows that despite the word recognition errors, none of the LCSEG, the MaxEnt models trained with conversational features, and the MaxEnt models trained with all available features perform significantly worse on ASR transcripts than on reference transcripts. One possible explanation for this, which we have observed in our corpus, is that the ASR system is likely to mis-recognize different occurences of words in the same way, and thus the lexical cohesion statistic, which captures the similarity of word repetition between two adjacency windows, is also likely to remain unchanged. In addition, when the models are trained with other features that are not affected by the recognition errors, such as pause and overlap, the negative impacts of recognition errors are further reduced to an insignificant level.

	TOP		SUB	
Error Rate	PK	WD	PK	WD
LCSEG(REF)	0.45	0.57	0.42	0.47
LCSEG(ASR)	0.45	0.58	0.40	0.47
MAXENT-CONV(REF)	0.34	0.34	0.37	0.37
MAXENT-CONV(ASR)	0.34	0.33	0.38	0.38
MAXENT-ALL(REF)	0.30	0.33	0.34	0.36
MAXENT-ALL(ASR)	0.31	0.34	0.34	0.37

Table 6: Effects of word recognition errors on detecting segments boundaries.

5 Discussion

The results in Section 4 show the benefits of including additional knowledge sources for recognizing segment boundaries. The next question to be addressed is what features in these sources are most useful for recognition. To provide a qualitative account of the segmentation cues, we performed an analysis to determine whether each proposed feature discriminates the class of segment boundaries. Previous work has identified statistical measures (e.g., Log Likelihood ratio) that are useful for determining the statistical association strength (relevance) of the occurrence of an n-gram feature to target class (Hsueh and Moore, 2006). Here we extend that study to calculate the LogLikelihood relevance of all of the features used in the experiments, and use the statistics to rank the features.

Our analysis shows that people do speak and behave differently near segment boundaries. Some of the identified segmentation cues match previous findings. For example, a segment is likely to start with higher pitched sounds (Brown et al., 1980; Ayers, 1994) and a lower rate of speech (Lehiste, 1980). Also, interlocutors pause longer than usual to make sure that everyone is ready to move on to a new discussion (Brown et al., 1980; Passonneau and Litman, 1993) and use some conventional expressions (e.g., *now*, *okay*, *let's*, *um*, *so*).

Our analysis also identified segmentation cues that have not been mentioned in previous research. For example, interlocutors do not move around a lot when a new discussion is brought up; interlocutors mention agenda items (e.g., *presentation*, *meeting*) or content words more often when initiating a new discussion. Also, from the analysis of current dialogue act types and their immediate contexts, we also observe that at segment boundaries interlocutors do the following more often than usual: start speaking before they are ready (*Stall*), give information (*Inform*), elicit an assessment of what has been said so far (*Elicit-assessment*), or act to smooth social functioning and make the group happier (*Be-positive*).

6 Conclusions and Future Work

This study explores the use of features from multiple knowledge sources (i.e., words, prosody, motion, interaction cues, speaker intention and role) for developing an automatic segmentation component in spontaneous, multiparty conversational speech. In particular, we addressed the following questions: (1) Can a MaxEnt classifier integrate the potentially characteristic multimodal features for automatic dialogue segmentation? (2) What are the most discriminative knowledge sources for detecting segment boundaries? (3) Does the use of ASR transcription significantly degrade the performance of a segmentation model?

First of all, our results show that a well performing MaxEnt model can be trained with available knowledge sources. Our results improve on previous work, which uses only conversational features, by 8.8% for top-level segmentation and 5.4% for sub-level segmentation. Analysis of the effectiveness of the various features shows that lexical features (i.e., cue words) are the most essential feature class to be combined into the segmentation model. However, lexical features must be combined with other features, in particular, conversational features (i.e., lexical cohesion, overlap, pause, speaker change), to train well performing models.

In addition, many of the non-lexical feature classes, including those that have been identified as indicative of segment boundaries in previous work (e.g., prosody) and those that we hypothesized as good predictors of segment boundaries (e.g., motion, context), are not beneficial for recognizing boundaries when used in isolation. However, these non-lexical features are useful when combined with lexical features, as the presence of the non-lexical features can balance the tendency of models trained with lexical cues alone to overpredict.

Experiments also show that it is possible to segment conversational speech directly on the ASR outputs. These results encouragingly show that we can segment conversational speech using features extracted from different knowledge sources, and in turn, facilitate the development of a fully automatic segmentation component for multimedia archives.

With the segmentation models developed and discriminative knowledge sources identified, a remaining question is whether it is possible to automatically select the discriminative features for recognition. This is particularly important for prosodic features, because the direct modelling approach we adopted resulted in a large number of features. We expect that by applying feature selection methods we can further improve the performance of automatic segmentation models. In the field of machine learning and pattern analysis, many methods and selection criteria have been proposed. Our next step will be to examine the effectiveness of these methods for the task of automatic segmentation. Also, we will further explore how to choose the best performing ensemble of knowledge sources so as to facilitate automatic selection of knowledge sources to be included.

Acknowledgement

This work was supported by the EU 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811). Our special thanks to Wessel Kraaij, Stephan Raaijmakers, Steve Renals, Gabriel Murray, Jean Carletta, and the anonymous reviewers for valuable comments. Thanks also to the AMI ASR group for producing the ASR transcriptions, and to our research partners in TNO for generating motion features.

References

M. Al-Hames, A. Dielmann, D. GaticaPerez, S. Reiter, S. Renals, and D. Zhang. 2005. Multimodal integration for meeting group action segmentation and recognition. In *Proc. of MLMI 2005*.

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

G. M. Ayers. 1994. Discourse functions of pitch range in spontaneous and read speech. In Jennifer J. Venditti, editor, *OSU Working Papers in Linguistics*, volume 44, pages 1–49.

S. Banerjee and A. Rudnicky. 2006. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proc. of IUI 2006*.

S. Banerjee, C. Rose, and A. I. Rudnicky. 2005. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proc. of the Tenth International Conference on Human-Computer Interaction*.

G. Brown, K. L. Currie, and J. Kenworthe. 1980. *Questions of Intonation*. University Park Press.

J. Carletta et al. 2006. The AMI meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Springer-Verlag Lecture Notes in Computer Science*, volume 3869. Springer-Verlag.

H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. 2005. Maximum entropy segmentation of broadcast news. In *Proc. of ICASP*, Philadelphia USA.

B. Di Eugenio and M. G. Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *Proc. of ACL 2003*.

B. Grosz and C. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3).

A. Gruenstein, J. Niekrasz, and M. Purver. 2005. Meeting structure annotation: Data and tools. In *Proc. of the SIGdial Workshop on Discourse and Dialogue*.

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 *Proc. of Interspeech 2005*.

M. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 25(3):527–571.

J. Hirschberg and D. Litman. 1987. Now let's talk about now: identifying cue phrases intonationally. In *Proc. of ACL 1987*.

J. Hirschberg and C. H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of ACL 1996*.

P. Hsueh and J.D. Moore. 2006. Automatic topic segmentation and lablelling in multiparty dialogue. In *the first IEEE/ACM workshop on Spoken Language Technology (SLT) 2006*.

H. Kozima. 1993. Text segmentation based on similarity between words. In *Proc. of ACL 1993*.

I. Lehiste. 1980. Phonetic characteristics of discourse. In *the Meeting of the Committee on Speech Research, Acoustical Society of Japan*.

G. Levow. 2004. Prosody-based topic segmentation for mandarin broadcast news. In *Proc. of HLT 2004*.

I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317.

R. Passonneau and D. Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of ACL 1993*.

J. Ponte and W. Croft. 1997. Text segmentation by topic. In *Proc. of the Conference on Research and Advanced Technology for Digital Libraries 1997*.

J. Reynar. 1998. *Topic Segmentation: Algorithms and Applications*. Ph.D. thesis, UPenn, PA USA.

E. Shriberg and A. Stolcke. 2001. Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proc. International Conference on Speech Prosody 2004*.

G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

K. Zechner and A. Waibel. 2000. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proc. of COLING-2000*.