Assessing the effectiveness of conversational features for dialogue segmentation in medical team meetings and in the AMI corpus

Saturnino Luz
Department of Computer Science
Trinity College Dublinm Ireland
luzs@cs.tcd.ie

Jing Su
School of Computer Science and Statistics
Trinity College Dublin, Ireland
sujing@scss.tcd.ie

Abstract

This paper presents a comparison of two similar dialogue analysis tasks: segmenting real-life medical team meetings into patient case discussions, and segmenting scenario-based meetings into topics. In contrast to other methods which use transcribed content and prosodic features (such as pitch, loudness etc), the method used in this comparison employs only the duration of the prosodic units themselves as the basis for dialogue representation. A concept of Vocalisation Horizon (VH) allows us to treat segmentation as a classification task where each instance to be classified is represented by the duration of a talk spurt, pause or speech overlap event in the dialogue. We report on the results this method yielded in segmentation of medical meetings, and on the implications of the results of further experiments on a larger corpus, the Augmented Multiparty Meeting corpus, to our ongoing efforts to support data collection and information retrieval in medical team meetings.

1 Introduction

As computer mediated communication becomes more widespread, and data gathering devices start to make their way into the meeting rooms and the workplace in general, the need arises for modelling and analysis of dialogue and human communicative behaviour in general (Banerjee et al., 2005). The focus of our interest in this area is the study of multi-party interaction at Multidisciplinary Medical Team Meeting (MDTMs), and the eventual recording of such meetings followed by indexing and structuring for integration into electronic health records. MDTMs share a number of characteristics with more conventional business meetings, and with the meeting scenarios targeted in recent research projects (Renals et al., 2007). However, MDTMs are better structured than these meetings, and more strongly influenced by the time pressures placed upon the medical professionals who take part in them (Kane and Luz, 2006).

In order for meeting support and review systems to be truly effective, they must allow users to efficiently browse and retrieve information of interest from the recorded data. Browsing in these media can be tedious and time consuming because continuous media such as audio and video are difficult to access since they lack natural reference points. A good deal of research has been conducted on indexing recorded meetings. From a user's point of view, an important aspect of indexing continuous media, and audio in particular, is the task of structuring the recorded content. Banerjee et al. (2005), for instance, showed that users took significantly less time to retrieve answers when they had access to discourse structure annotation than in a control condition in which they had access only to unannotated recordings.

The most salient discourse structure in a meeting is the topic of conversation. The content within a given topic is cohesive and should therefore be viewed as a whole. In MDTMs, the meeting consists basically of successive patient case discussions (PCDs) in which the patient's condition is discussed among different medical specialists with the objective of agreeing diagnoses, making patient management decisions etc. Thus, the individual PCDs can be regarded as the different "topics" which make up an MDTM (Luz, 2009).

In this paper we explore the use of a content-free approach to the representation of vocalisation events for segmentation of MDTM dialogues. We start by extending the work of Luz (2009) on a small corpus of MDTM recordings, and then test our approach on a larger dataset, the AMI (Aug-

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 332–339,
The University of Tokyo, September 24-25, 2010.

mented Multi-Party Interaction) corpus (Carletta, 2007). Our ultimate goal is to analyse and apply the insights gained on the AMI corpus to our work on data gathering and representation in MDTMs.

2 Related work

Topic segmentation and detection, as an aid to meeting information retrieval and meeting indexing, has attracted the interest of many researchers in recent years. The objective of topic segmentation is to locate the beginning and end time of a cohesive segment of dialogue which can be singled out as a "topic". Meeting topic segmentation has been strongly influenced by techniques developed for topic segmentation in text (Hearst, 1997), and more recently in broadcast news audio, even though it is generally acknowledged that dialogue segmentation differs from text and scripted speech in important respects (Gruenstein et al., 2005).

In early work (Galley et al., 2003), meeting annotation focused on changes that produce high inter-annotator agreement, with no further specification of topic label or discourse structure. Current work has paid greater attention to discourse structure, as reflected in two major meeting corpus gathering and analysis projects: the AMI project (Renals et al., 2007) and the ICSI meeting project (Morgan et al., 2001). The AMI corpus distinguishes top-level and functional topics such as "presentation", "discussion", "opening", "closing", "agenda" which are further specified into sub-topics (Hsueh et al., 2006). Gruenstein et al. (2005) sought to annotated the ICSI corpus hierarchically according to topic, identifying, in addition, action items and decision points. In contrast to these more general types of meetings, MDTMs are segmented into better defined units (i.e. PCDs) so that inter-annotator agreement on topic (patient case discussion) boundaries is less of an issue, since PCDs are collectively agreed parts of the formal structure of the meetings.

Meeting transcripts (either done manually or automatically) have formed the basis for a number of approaches to topic segmentation (Galley et al., 2003; Hsueh et al., 2006; Sherman and Liu, 2008). The transcript-based meeting segmentation described in (Galley et al., 2003) adapted the unsupervised lexical cohesion method developed for written text segmentation (Hearst, 1997). Other approaches have employed supervised machine learning methods with textual features (Hsueh et al., 2006). Prosodic and conversational features have also been integrated into text-based representations, often improving segmentation accuracy (Galley et al., 2003; Hsueh and Moore, 2007).

However, approaches that rely on transcription, and sometimes higher-level annotation on transcripts, as is the case of (Sherman and Liu, 2008), have two shortcomings which limit their applicability to MDTM indexing. First, manual transcription is unfeasible in a busy hospital setting, and automatic speech recognition of unconstrained, noisy dialogues falls short of the levels of accuracy required for good segmentation. Secondly, the contents of MDTMs are subject to stringent privacy and confidentiality constraints which limit access to training data. Regardless of such application constraints, some authors (Malioutov et al., 2007; Shriberg et al., 2000) argue for the use of prosodic features and other acoustic patterns directly from the audio signal for segmentation. The approach investigated in this paper goes a step further by representing the data solely through what is, arguably, the simplest form of content-free representation, namely: duration of talk spurts, silences and speech overlaps, optionally complemented with speaker role information (e.g. medical speciality).

3 Content-free representations

There is more to the structure (and even the semantics) of a dialogue than the textual content of the words exchanged by its participants. The role of prosody in shaping the illocutionary force of vocalisations, for instance, is well documented (Holmes, 1984), and prosodic features have been used for automatic segmentation of broadcast news data into sentences and topics (Shriberg et al., 2000). Similarly, recurring audio patterns have been employed in segmentation of recorded lectures (Malioutov et al., 2007). Works in the area of social psychology have used the simple conversational features of duration of vocalisations, pauses and overlaps to study the dynamics of group interaction. Jaffe and Feldstein (1970) characterise dialogues as Markov processes, and Dabbs and Ruback (1987) suggest that a "content-free" method based on the amount and structure of vocal interactions could complement group interaction frameworks such as the one proposed by Bales (1950). Pauses and overlap statistics alone can be used, for instance, to characterise

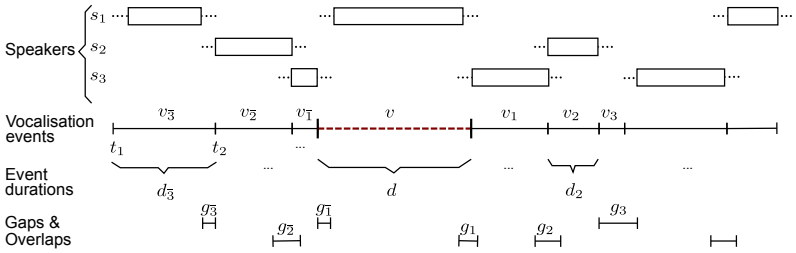

Figure 1: Vocalisation Horizon for event $\boldsymbol{v}$.

the differences between face-to-face and telephone dialogue (ten Bosch et al., 2005), and a correlation between the duration of pauses and topic boundaries has been demonstrated for recordings of spontaneous narratives (Oliveira, 2002).

These works provided the initial motivation for our content-free representation scheme and the topic segmentation method proposed in this paper. It is easy to verify by inspection of both the corpus of medical team meetings described in Section 4 and the AMI corpus that pauses and vocalisations vary significantly in duration and position on and around topic boundaries. Table 1 shows the mean durations of vocalisations that initiate new topics or PCDs in MDTMs and the scenario-based AMI meetings, as well as the durations of pauses and overlaps that surround it (within one vocalisation event to the left and right). In all cases the differences were statistically significant. These results agree with those obtained by Oliveira (2002) for discourse topics, and suggest that an approach based on representing the duration of vocalisations, pauses and overlaps in the immediate context of a vocalisation might be effective for automatic segmentation of meeting dialogues into topics or PCDs.

Table 1: Mean durations in seconds (and standard deviations) of vocalisation and pauses on and near topic boundaries in MDTM and AMI meetings.

	Boundary	Non-boundary	t-test
AMI vocal.	5.3 (8.2)	1.6 (3.5)	$p < .01$
AMI pauses	2.6 (4.9)	1.2 (2.8)	$p < .01$
AMI overlaps	0.4 (0.4)	0.3 (0.6)	$p < .01$
MDTM vocal.	12.0 (15.5)	8.1 (14.7)	$p < .05$
MDTM pauses	9.7 (12.7)	8.2 (14.8)	$p < .05$

We thus conceptualise meeting topic segmentation as a classification task approachable through supervised machine learning. A meeting can be pre-segmented into separate *vocalisations* (i.e. talk spurts uttered by meeting participants) and silences, and such basic units (henceforth referred to as *vocalisation events*) can then be classified as to whether they signal a topic transition. The basic defining features of a vocalisation event are the identity of the speaker who uttered the vocalisation (or speakers, for events containing speech overlap) and its duration, or the duration of a pause, for silence events. However, identity labels and interval durations by themselves are not enough to enable segmentation. As we have seen above, some approaches to meeting segmentation complement these basic data with text (keywords or full transcription) uttered during vocalisation events, and with prosodic features. Our proposal is to retain the content-free character of the basic representation by complementing the speaker and duration information for an event with data describing its preceding and succeeding events. We thus aim to capture an aspect of the dynamics of the dialogue by representing snapshots of vocalisation sequences. We call this general representation strategy *Vocalisation Horizon* (VH).

Figure 1 illustrates the basic idea. Vocalisation events are placed on a time line and combine utterances produced by the speakers who took part in the meeting. These events can be labelled with nominal attributes $(s_1, s_2, \ldots)$ denoting the speaker (or some other symbolic attribute, such as the speaker's role in the meeting). Silences (gaps) and group talk (overlap) can either be assigned reserved descriptors (such as "Floor" and "Group") or regarded as separate annotation layers. The general data representation scheme for, say, segment v would involve a data from its left context $(v_{\bar{1}}, v_{\bar{2}}, v_{\bar{3}}, \ldots)$ and its right context $(v_1, v_2, v_3, \ldots)$ in addition to the data for v itself. These can be a combination of symbolic labels (in Figure 1, for instance, s_1 for the current speaker, $s_3, s_2, s_1, \ldots$ for the preceding events and $s_3, s_2, s_3, \ldots$ for the following events), durations (d, $d_{\bar{1}}, d_{\bar{2}}, d_{\bar{3}}, \ldots$ etc)

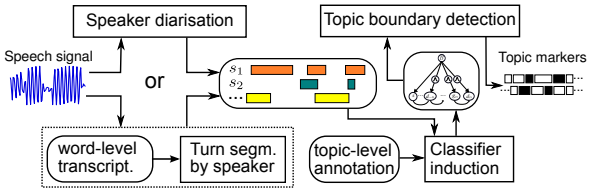

Figure 2: Meeting segmentation processing architecture.

and gaps or overlaps $g_{\bar{1}}, g_{\bar{2}}, g_{\bar{3}}, \ldots, g_1, g_2, g_3, \ldots$ etc). Specific representations depend on the type of annotation available on the speech data and on the nature of the meeting. Sections 4 and 5 present and assess various representation schemes.

The general processing architecture for meeting segmentation assumed in this paper is shown in Figure 2. The system will received the speech signal, possibly on a single channel, and pre-segment it into separate channels (one per speaker) with intervals of speech activity and silence labelled for each stream. Depending on the quality of the recording and the characteristics of the environment, this initial processing stage can be accomplished automatically through existing speaker diarisation methods — e.g. (Ajmera and Wooters, 2003). In the experiments reported below manual annotation was employed. In the AMI corpus, speaker and speech activity annotation is done on the word level and include transcription (Carletta, 2007). We parsed these word-level labels, ignoring the transcriptions, in order to build the content-free representation described above. Once the data representation has been created it is then used, along with topic boundary annotations, to train a probabilistic classifier. Finally, the topic detection module uses the models generated in the training phase to hypothesise boundaries in unannotated vocalisation event sequences and, optionally, performs post-processing of these sequences before returning the final hypothesis. These modules are described in more detail below.

4 MDTM Segmentation

The MDTM corpus was collected over a period of three years as part of a detailed ethnographic study of medical teams (Kane and Luz, 2006). The corpus consists in 28 hours or meetings recorded in a dedicated teleconferencing room at a major primary care hospital. The audio sources included a pressure-zone microphone attached to the teleconferencing system and a highly sensitive directional microphone. Video was gathered through two separate sources: the teleconferencing system, which showed the participants and, at times, the medical images (pathology slides, radiology) relevant to the case under discussion, and a high-end camcorder mounted on a tripod. All data were imported into a multimedia annotation tool and synchronised. Of these, two meetings encompassing 54 PCDs were chosen an annotated for vocalisations (including speaker identity and duration) and PCD boundaries.

Vocalisation events were encoded as vectors $v = (s, d, s_{\bar{1}}, d_{\bar{1}}, \ldots, s_{\bar{n}}, d_{\bar{n}}, s_1, d_1, \ldots, s_n, d_n)$, where the variables are as explained in Section 3. The speaker labels s, $s_{\bar{i}}$ and s_i are replaced, for the sake of generality, by "role" labels denoting medical specialties, such as "radiologist", "surgeon", "clinical oncologist", "pathologist" etc. In addition to these roles, we reserved the special labels "Pause" (a period of silence between two vocalisations by the same speaker), "SwitchingPause" (pause between vocalisations by different speakers), and "Group" (vocalisations containing overlaps, i.e. speech by more than one speaker). We set a minimum duration of 1s for a talk spurt to count as a speech vocalisation event and a 0.9s minimum duration for silence period to be a pause. Shorter intervals (depicted in Figure 1 as the fuzzy ends of the speech lines on the top of the chart) are incorporated into an adjacent vocalisation event.

The segmentation process can be defined as the process of mapping the set of vocalisation events V to $\{0, 1\}$ where 1 represents a topic boundary and 0 represents a non-boundary vocalisation event. In order to implement this mapping we employ a Naive Bayes classifier. The conditional probabilities for the nominal variables (speaker roles) are estimated on the training set by maximum likelihood and combined into multinomial models, while the continuous variables are log transformed and modelled through Gaussian kernels (John and Langley, 1995).

These models are used to estimate the probability, given by equation (1), of a vocalisation being marked as a topic boundary given the above described data representation, and the usual conditional independence assumptions applies.

$$P(B = b|V = v) = P(B = b|S_{\bar{n}} = s_{\bar{n}}, D_{\bar{n}} = d_{\bar{n}}, \ldots, S = s, \ldots, D_n = d_n) \quad (1)$$

The model can therefore be represented as a simple Bayesian network where the only depen-

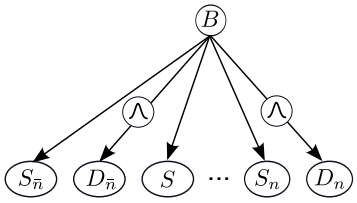

Figure 3: Bayesian model employed for dialogue segmentation.

dencies are between the boundary variable and each feature of the vocalisation event, as shown in Figure 3.

Luz (2009) reports that, for a similar data representation, horizons of length $2 < n < 6$ produced the best segmentation results. Following this finding, we adopt $n = 3$ for all our experiments. We tested two variants of the representation: V_{pd} that discriminated between pause types (pauses, switching pauses, group pauses, and group switching pauses), as in (Dabbs and Ruback, 1987), and V_{sp} which labelled all pauses equally. The evaluation metrics employed include the standard classification metrics of *accuracy* (A), the proportion of correctly classified segments, boundary precision (P), the proportion of correctly assigned boundaries among all events marked as topic boundaries, boundary recall (R), the proportion of target boundaries correctly assigned, and the F_1 score, the harmonic mean of P and R.

Although these standard metrics provide an initial approximation to segmentation effectiveness, they have been criticised as tools for evaluating segmentation because they are hard to interpret and are not sensitive to near misses (Pevzner and Hearst, 2002). Furthermore, due to the highly unbalanced nature of the classification task (boundary vocalisation events are only 3.3% of all instances), accuracy scores tend to produce overoptimistic results. Therefore, to give a fairer picture of the effectiveness of our method, we also report values for two error metrics proposed specifically for segmentation: P_k (Beeferman et al., 1999) and WindowDiff, or *WD*, (Pevzner and Hearst, 2002).

The P_k metric gives the probability that two vocalisation events occurring k vocalisations apart and picked otherwise randomly from the dataset are incorrectly identified by the algorithm as belonging to the same or to different topics. P_k is computed by sliding two pairs of pointers over the reference and the hypothesis sequences and observing whether each pair of pointers rests in the same or in different segments. An error is counted if the pairs disagree (i.e. if they point to the same segment in one sequence and to different segments in the other).

WD is as an estimate of inconsistencies between reference and hypothesis, obtained by sliding a window of length equal k segments over the time line and counting disagreements between true and hypothesised boundaries. Like the standard IR metrics, P_k and *WD* range over the $[0, 1]$ interval. Since they are error metrics, the greater the value, the worse the segmentation.

Table 2: PCD segmentation results for 5-fold cross validation, horizon $\boldsymbol{n = 3}$ (mean values).

Threshold	Filter	Data	A	P	R	F_1	P_k	*WD*
MAP	no	V_{sp}	0.94	0.20	0.21	0.18	0.33	0.44
		V_{pd}	0.95	0.17	0.20	0.16	0.30	0.38
	yes	V_{sp}	0.95	0.20	0.16	0.16	0.32	0.38
		V_{pd}	**0.95**	0.16	0.12	0.13	0.29	0.34
Proport.	no	V_{sp}	0.95	0.28	**0.28**	**0.28**	0.26	0.36
		V_{pd}	0.95	0.26	0.27	0.26	0.27	0.42
	yes	V_{sp}	0.95	**0.30**	0.22	0.25	0.25	**0.31**
		V_{pd}	0.95	0.22	0.14	0.17	0.27	0.33

Table 2 shows the results for segmentation of MDTMs into PCDs under the representational variants mentioned above and two different thresholding strategies: *maximum a posteriori* hypothesis (MAP) and proportional threshold. The latter is a strategy that varies the threshold probability above which an event is marked as a boundary according to the generality of boundaries found in the training set. The motivation for testing proportional thresholds is illustrated by Figure 4, which shows a step plot of MAP hypothesis (h) superimposed on the true segmentation (peaks represent boundaries) and the corresponding values for $p(b|v)$. It is clear that a number of false positives would be removed if the threshold were set above the MAP level[1] with no effect on the number of false negatives.

Another possible improvement suggested by Figure 4 is the *filtering* of adjacent boundary hypotheses. Wider peaks, such as the ones on instances 14 and 172 indicate that two or more boundaries were hypothesised in immediate succession. Since this is clearly impossible, a straightforward improvement of the segmentation

[1]I.e. $p(b|v) > 0.5$; above the horizontal line in the centre.

hypothesis can be achieved by choosing a single boundary marker among a cluster of adjacent ones. This has been done as a post-processing step by choosing a single event with maximal estimated probability within a cluster of adjacent boundary hypotheses as the new hypothesis.

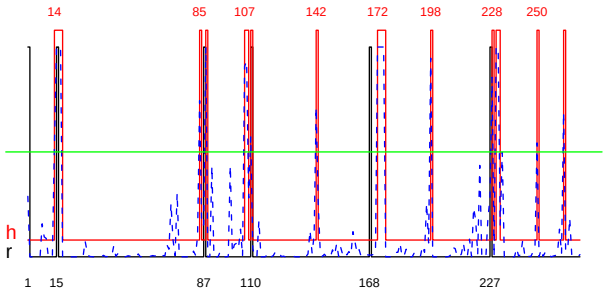

Figure 4: Segmentation profile showing true boundaries (r), boundaries hypothesised by a MAP classifier (h) and probabilities (dotted line).

The results suggest that both proportional thresholding and filtering improve segmentation. As expected, accuracy figures were generally high (an uninformative) reflecting the great imbalance in favour of negative instances and the conservative nature of the classifier. Precision, recall and F_1 (for positive instances only) were also predictably low, with V_{sp} under a proportional threshold attaining the best results. However, in meeting browsing marking the topic boundary precisely is far less important than retrieving the right text is in information retrieval or text categorisation, since the user can easily scan the neighbouring intervals with a slider (Banerjee et al., 2005). Therefore, P_k and *WD* are the most appropriate measures of success in this task. Here our results seem quite encouraging, given that they all represent great improvements over the (rather reasonable) baselines of $P_k = .46$ and $WD = .51$ estimated by Monte Carlo simulation as in (Hsueh et al., 2006) by hypothesising the same proportion of boundaries found in the training set. Our results also compare favourably with some of the best results reported in the meeting segmentation literature to date, namely $P_k = 0.32$ and $WD = 0.36$, for a lexical cohesion algorithm on the ICSI corpus (Galley et al., 2003), and $P_k = 0.34$ and $WD = 0.36$, for a maximum entropy approach combining lexical, conversational and video features on the AMI corpus (Hsueh et al., 2006).

Although these results are promising, they pose a question as regards data representation. While V_{pd} yielded the best results under MAP, V_{sp} worked best overall under a proportional threshold. What is the effect of encoding more detailed pause and overlap information? Unfortunately, the MDTM corpus has not been annotated to the level of detail required to allow in-depth investigation of this question. We therefore turn to the far larger and more detailed AMI corpus for our next experiments. In addition to helping clarify the representation issue, testing our method on this corpus will give us a better idea of how our method performs in a more standard topic segmentation task.

5 AMI Segmentation

The AMI corpus is a collection of meetings recorded under controlled conditions, many of which have a fixed scenario, goals and assigned participant roles. The corpus is manually transcribed, and annotated with word-level timings and a variety of metadata, including topics and sub-topics (Carletta, 2007). Transcriptions in the AMI corpus are extracted from redundant recording channels (lapel, headset and array microphones), and stored separately for each participant. Because timing information in AMI is so detailed, we were able to create much richer VH representations, including finer grained pause and overlap information.

The original XML-encoded AMI data were parsed and collated to produce our variants of the VH scheme. We tested four types of VH: V_v, which includes only vocalisation events; V_g, which includes only pause and speech overlap events; V_a, which includes all vocalisations, pauses and overlaps; and V_r, which is similar to V_{pd} in that it includes speaker roles in addition to vocalisations. Pauses and overlaps were encoded by the same variable g_i, where $g_i > 0$ indicates a pause $g_i < 0$ an overlap, as shown in Figure 1. Unlike MDTM, no arbitrary threshold was imposed on the identification of pause and overlap events. As before, we tested on a horizon $n = 3$, in order to allow comparison with MDTM results.

The training and boundary inference process also remained as in the MDTM experiment, except that the larger amount of meeting data available enabled us to increase the number of folds for cross validation so that the results could be tested for statistical significance.

The error scores and the number of boundaries predicted for the different representational vari-

ants, filtering an thresholding strategies are shown in Table 3. Although all methods significantly outperformed the baseline scores of $P_k = 0.473$ and *WD* $= 0.542$ (paired t-tests, $p < 0.01$, for all conditions), there were hardly any differences in P_k scores across the different representations, even when conservative boundary filtering is performed. Filtering, however, caused a significant improvement for *WD* in all cases, though the combined effects of proportional thresholding and filtering caused the classifier to err on the side of underprediction. A 3-way analysis of variance including non-filtered scores for proportional threshold resulted in $F[4, 235] = 31.82$, $p < 0.01$ for *WD* scores. These outcomes agree with the results of the smaller-scale MDTM segmentation experiment, showing that categorisation based on conversational features tend to mark clusters of segments around the true topic boundary. In addition, the trend for better performance of proportional thresholding exhibited in the MDTM data was not as clearly observed in the AMI data, where only *WD* scores were significantly better than MAP ($p < 0.01$, Tukey HSD).

Table 3: Segmentation results for 16-fold cross validation on AMI corpus, horizon $\boldsymbol{n = 3}$. Correct number of boundaries in reference is 724.

Threshold	Filter	Data	P_k	*WD*	# bound.
MAP	no	V_a	0.270	0.462	3322
		V_g	0.278	0.433	1875
		V_v	0.273	0.449	3075
		V_r	0.271	0.448	3073
	yes	V_a	0.272	0.362	574
		V_g	0.277	0.391	851
		V_v	0.275	0.358	468
		V_r	0.274	0.357	469
Proport.	no	V_a	0.289	0.398	1233
		V_g	0.290	0.382	735
		V_v	0.293	0.387	1002
		V_r	0.293	0.387	1002
	yes	V_a	0.293	0.353	241
		V_g	0.290	0.362	383
		V_v	0.297	0.350	183
		V_r	0.297	0.350	182

It is noteworthy that the finer-grained representations from which speaker roles were excluded (V_v, V_g, and V_a) yielded segmentation performance comparable to the MDTM segmentation performance under V_{sp} and V_{pd}. In fact, adding speaker role information in V_r did not result in improvement for AMI segmentation. Also interesting is the fact that representations based solely on pause and overlap information also produced good performance, thus confirming our initial intuition.

5.1 MDTM revisited

Since V_v, V_g and V_a seem to perform well without including speaker role information (except for the current vocalisation's speaker role) we would like to see how a similar representation might affect segmentation performance for MDTM. We therefore tested whether excluding preceding and following speaker role information from V_{sp} and V_{pd} had a positive impact on PCD segmentation performance. However, contrary to our expectations neither of the modified representations yielded better scores. The best results, achieved for the modified V_{pd} under proportional thresholding ($P_K = 0.27$ and *WD* $= 0.34$), failed to match the results obtained with the original representation. It seems that the various and more specialised speaker roles found in medical meetings can be good predictors of PCD boundaries. For example: a typical pattern at the start of a PCD is the recounting of the patient's initial symptoms and clinical findings by the registrar in a narrative style. In AMI, on other hand, the roles are much fewer, being only acted out by the participants as part of the given scenario, which might explain the irrelevance of these roles for segmentation.

5.2 Conclusion

MDTM segmentation differs from topic segmentation of the AMI meetings in that PCDs are more regular in their occurrence than meeting topics proper. Speaker role information was also found to help MDTM segmentation, which was expected since there are many more very distinct active speaker roles in MDTM (10 specialties, in total). Furthermore, V_{sp} and V_{pd} represent pauses and overlaps as reserved roles, so that the information encoded in V_g and V_a as separate variables appear in the speaker role horizon of V_{sp} and V_{pd}. It is possible that the finer-grained timing annotation of the AMI corpus (including detailed overlap and pause information unavailable in the MDTM data) contributed to the relatively good segmentation performance achieved on AMI even in the absence of speaker role cues. It would be interesting to investigate whether finer pause and overlap timings can also improve MDTM segmentation. This suggests some requirements for MDTM data collection and pre-processing, such as the use of individual close-talking and the use of a speech recogniser to derive word-level timings. We plan on conducting further experiments in that regard.

Acknowledgements

This research was funded by Science Foundation Ireland under the Research Frontiers program. The presentation was funded by Grant 07/CE/1142, Centre for Next Generation Localisation (CNGL).

References

J. Ajmera and C. Wooters. 2003. A robust speaker clustering algorithm. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'03*, pages 411–416. IEEE Press.

R. F. Bales. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, Cambridge, Mass.

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 *Proceedings of the 10th International Conference on Human-Computer Interaction (INTERACT'05)*, pages 643–656.

D. Beeferman, A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34:177–210, Feb. 10.1023/A:1007506220214.

J. Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

J. M. J. Dabbs and B. Ruback. 1987. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology*, 20(123–169).

M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multiparty conversation. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 562–569.

A. Gruenstein, J. Niekrasz, and M. Purver. 2005. Meeting structure annotation: Data and tools. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 117–127, Lisbon, Portugal, September.

M. A. Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

J. Holmes. 1984. Modifying illocutionary force. *Journal of Pragmatics*, 8(3):345 – 365.

P. Hsueh and J. D. Moore. 2007. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. ACL Press.

P. Hsueh, J. D. Moore, and S. Renals. 2006. Automatic segmentation of multiparty dialogue. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*, pages 273–277. ACL Press.

J. Jaffe and S. Feldstein. 1970. *Rhythms of dialogue*. Academic Press, New York.

G. H. John and P. Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI'95)*, pages 338–345, San Francisco, CA, USA, August. Morgan Kaufmann Publishers.

B. Kane and S. Luz. 2006. Multidisciplinary medical team meetings: An analysis of collaborative working with special attention to timing and teleconferencing. *Computer Supported Cooperative Work (CSCW)*, 15(5):501–535.

S. Luz. 2009. Locating case discussion segments in recorded medical team meetings. In *Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech (SSCS'09)*, pages 21–30, Beijing, China, October. ACM Press.

I. Malioutov, A. Park, R. Barzilay, and J. Glass. 2007. Making sense of sound: Unsupervised topic segmentation over acoustic input. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 504–511, Prague, Czech Republic, June. Association for Computational Linguistics.

N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. 2001. The meeting project at ICSI. In *Procs. of Human Language Technologies Conference*, San Diego.

M. Oliveira, 2002. *The role of pause occurrence and pause duration in the signaling of narrative structure*, volume 2389 of *LNAI*, pages 43–51. Springer.

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

S. Renals, T. Hain, and H. Bourlard. 2007. Recognition and interpretation of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*.

M. Sherman and Y. Liu. 2008. Using hidden Markov models for topic segmentation of meeting transcripts. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 185–188.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1-2):127–154.

L. ten Bosch, N. Oostdijk, and L. Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47:80–86.