Locating Case Discussion Segments in Recorded Medical Team Meetings

Saturnino Luz
Department of Computer Science
School of Computer Science and Statistics
O’Reilly Institute
Trinity College Dublin
Ireland
+353(1)869-3686
luzs@cs.tcd.ie

ABSTRACT
Although there has been great interest in the issue of indexing and providing access to multimedia records of meetings, with substantial efforts directed towards collection and analysis of meeting corpora, most research in this area is based on data collected at research labs, under somewhat artificial conditions. In contrast, this paper focuses on data recorded in a real-world setting where a number of health professionals participate in weekly meetings held as part of the work routines in a major hospital. These meetings have been observed to be highly structured, a fact that is due undoubtedly to the time pressures, as well as communication and dependability constraints characteristic of the context in which the meetings happen. The hypothesis investigated in this paper is that the conversational structure of these meetings enable their segmentation into meaningful sub-units, namely individual patient case discussions, based only on data on the roles of the participants and the duration and sequence of vocalisations. We describe the task of segmenting audio-visual records of multidisciplinary medical team meetings as a topic segmentation task, present a method for automatic segmentation based on a “content-free” representation of conversational structure, and report the results of a series of patient case segmentation experiments. The approach presented here achieves levels of segmentation accuracy (measured in terms of the standard $P_k$ and $WD$ metrics) comparable to those attained by state of the art topic segmentation algorithms based on richer and combined knowledge sources.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.5.1 [Information Interfaces]: Classifier design and evaluation

General Terms
Human Factors

Keywords
Meeting Analysis, Meeting topic segmentation, Multidisciplinary medical team meetings, Patient case discussions, Audio analysis

1. INTRODUCTION
Recording of face-to-face and remote meetings is becoming increasingly common in the workplace. So much so that human-computer interaction as well as language technology researchers have argued for the development of systems to support indexing, searching and browsing of recorded multimedia meeting data. In recent years, the topic of “meeting browsing” has received considerable attention from researchers from a variety of backgrounds, and large research projects have helped lay the foundations for automatic analysis of meeting contents by collecting large, extensively annotated corpora of meeting data and developing a number of techniques for search and visualisation of this type of data. However, in spite of its eminently practical motivations, research on meeting browsers has been based mostly on scenario-driven meetings recorded (in both cases) under laboratory recording conditions. While the relative homogeneity of the data obtained in such controlled environments facilitates intrinsic evaluation of the different machine learning and natural language processing techniques commonly employed in the analysis of meetings, questions remain as to how effective these techniques can be in more realistic application scenarios. These questions can only be answered, of course, on a case by case basis, in the context of system development. Observation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM MM–SSCS ’09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-762-2/09/10 ...$10.00.
and data collection in real workplace environments can nevertheless help assess the applicability of these techniques to different kinds of meeting data. This paper reports on one such assessment applied to audio and video data collected as part of a three-year ethnographic study of a multidisciplinary medical team and their regular meetings [23, 24]. The technique in question is segmentation of conversations into topics, in a broadly defined sense [13, 18, 12].

Multidisciplinary medical team meetings (MDTMs) are meetings in which several specialists gather in order to discuss patient cases, agree on a diagnosis, and make treatment and patient management decisions. A typical MDTM lasts over one hour and consists of a sequence of patient case discussions (PCDs). We identify each of these PCDs with “topics”, in the sense that they have well defined conceptual boundaries and can be categorised into different types, such as medical and surgical discussions, local patient and referral patients, co-located PCDs and remote PCDs, etc [24].

Current approaches to meeting topic segmentation employ combinations of feature sources, including lexical features (or bags of words) obtained from the output of a speech recogniser), conversational features (lexical cohesion statistics as well as dialogue structure, vocalisation and silence statistics) [13], prosodic features [41], video features [12], and other contextual features such as dialogue type and speaker role [13]. Only a few of these information sources can be reliably extracted from recordings obtained at a real MDTM, where the fast pace of the dialogue, the large number of participants, the diverse composition of the medical teams, and other factors make clean recording of individual speakers a practical impossibility. Very high word error rates for automatic speech recognition, for instance, would preclude the use of dialogue acts, lexical features and lexical cohesion statistics for our MDTM data (even though some topic segmentation systems have been shown to be resilient to moderate word error rates [13, 19] in other domains). This leaves us with only “content-free” features (apart from video, which we do not address in this paper) to work with.

Social psychologists have argued that such content-free features as patterns of turn-taking (vocalisation) and silence can tell an analyst much about the nature and structure of a meeting or a dialogue [13, 20]. In the case of MDTMs (and patient case discussions, in particular), despite being fast-paced and apparently chaotic to an outside observer, the conversations are highly structured events where the participants have very well defined roles, according with their medical specialities, which define to a great extend their patterns of participation in the meeting.

Current approaches to topic segmentation are applied to medical team discussions, and data collection in real workplace environments can nevertheless help assess the applicability of these techniques to different kinds of meeting data. This paper reports on one such assessment applied to audio and video data collected as part of a three-year ethnographic study of a multidisciplinary medical team and their regular meetings [23, 24].

The MDTM is structured as a sequence of PCDs, where the patient’s medical record is reviewed, evidence from pathology and radiology is presented, the possibility of surgery is discussed, and a patient management plan is agreed. The meetings are generally attended by medical students and junior staff, who do not play an active role in the discussions. The presentations and discussions make intensive use of visual aids (e.g. display of pathology slides on a large screen, radiology images on high-resolution displays, etc), and are often attended by remote participants connected through teleconferencing. A detailed analysis of the MDTM, issues related to technological support for the organisational processes surrounding the meeting, its different functions in the hospital environment, and mechanisms that dependability to its decision making can be found in [23].

The physical environment in which the MDTMs recorded for this corpus take place is a dedicated teleconferencing room equipped with projection equipment, a high resolution screen for radiological images, a large plasma screen, as well as microscopes and document readers which can be connected to the large display. The recordings were taken from two separate sources: (a) the existing teleconferencing equipment fitted into the meeting room, which recorded the audio through a pressure-zone microphone and alternated recording of the video channel between a view of the participants and views of the different medical images under discussion, and (b) a high-end camcorder mounted on a tripod which recorded the audio through a highly sensitive directional microphone. These two sources were aligned (synchronised) using a multimedia annotation tool. While small and lacking in annotation detail if compared to the major meeting corpora mentioned above (Section 2), the MDTM corpus is unique in that it was collected in situ (as part of Bridget Kane’s PhD research [24]) under naturalistic conditions, with the meeting participants engaged in a complex professional task.

Over 28 hours of meeting data were collected, in total. The original purpose of data collection was to investigate the diagnosis and decision making processes of multidisciplinary medical teams [23] within an interaction analysis framework [24]. For the study reported in this paper, a dataset of 54 PCDs were segmented and annotated using the ELAN Linguistic Annotator [24]. A total of 21 different speakers actively participated in these PCDs. These speakers can be grouped into 10 different specialist roles. Unlike the speakers in the broadcast news data analysed in [13], who can play different roles at different times dur-
ing the broadcast, each speaker in the MDTM corpus has a unique (medical specialist) role throughout the PCDs. The mapping from speaker identities to roles is not one-to-one, however, as more than one speaker can perform the same role in the same PCD. A breakdown of the relative contributions of each of these roles to the discussions (in terms of amount of speech) is presented in Table 1. The remaining time is distributed between pauses (silences, 3.4%) and “group vocalisation” (overlapping speech, 1.2%).

3. RELATED WORK

For the purposes of meeting segmentation PCDs can be regarded as sequences of vocalisations that share a common topic (i.e. the discussion of a particular patient’s case). In this sense, the task of segmenting MDTMs into PCDs is similar to the topic segmentation task as defined by Galley et al [14] and tackled in recent work on meeting analysis [12, 17, 19, 4]. Meeting segmentation has been influenced by early work on text segmentation [22] and shares with it evaluation metrics and methods. The task, as approached in this paper, does not involve identification and clustering individual group actions [17] or labelling of topics (PCDs) [30]. We investigate labelling issues (PCD content categorisation) elsewhere [30]. Here we simply treat the meeting as a sequence of vocalisations and pauses, and attempt to mark out those vocalisations which signal the beginning of a PCD. These boundary vocalisations are similarly distributed for PCDs in our MDTM corpus, where only about 3.6% of all vocalisations indicate the start of a PCD, and in the AMI corpus, where about 3.3% of talk spurts indicate a topic change [19].

Despite these similarities MDTM segmentation differs from meeting topic segmentation in that the latter seeks to identify segments that are different as they appear in the vocalisation sequence, whereas the former aims to segment the stream into essentially similar sub-sequences. Topics in the AMI corpus, for instance, can be categorised as “top-level” and “functional topics” [17] denoting segments that could also be described as “meeting states” [12], such as “presentation”, “discussion”, “opening”, “closing”, “agenda” etc, which can then be subdivided into sub-topics, forming a shallow hierarchy which is usually flattened for the purposes of segmentation. For some applications, annotation focuses simply on topic changes that produce high inter-annotator agreement scores, with no further specification of topic label or discourse structure [12]. In MDTM segmentation, due to the self-contained nature of PCDs, annotators have little difficulty in identifying case discussion boundaries. It should be remarked that MDTM segmentation can also be hierarchical, since PCDs exhibit an identifiable set of internal discussion states. However, this level of topic structure is not addressed in this paper.

Different strategies have been employed for conversational topic segmentation. As mentioned above, Galley et al. [14] model meeting topic segmentation after a text segmentation approach (Texttiling [19]), relying on transcribed speech to compute lexical cohesion probabilities for adjacent analysis windows. Renals & Ellis [24], on the other hand, consider “non-lexical methods” for segmentation which bear some similarity with our approach in that their data representation is based on patterns of talk spurts encoded as transition matrices. However, their segmentation algorithm, which is analogous to acoustic speaker segmentation using the Bayesian Information Criterion, does not produce satisfactory results, leading the authors to speculate that “turn pattern boundaries are not directly related to discussion topics” [26]. The results presented in this paper contradict that conjecture.

More recent approaches have tended to work with richer data representation schemes. Sutaneev & Rudnicky [31] define their model’s data instances as short time windows over meeting segments whose features are described by low-level conversational statistics (number of speakers, number of speaker changes and speech overlap). They train a decision tree classifier to distinguish between windows that contain topic changes, obtaining an 18% accuracy gain over a baseline (random) classifier. In more recent work, implicit supervision in the form of participant notes is employed in order to segment meetings into speech intervals which correspond to agenda items [11]. Dielmann & Renals [12] segment meetings from the M4 corpus [32] into a pre-defined set of five basic “group meeting actions”. They use dynamic Bayesian networks to integrate different feature streams (prosody, turn-taking, lexical and video) into a two-level model comprising individual and group actions. Hseuh et al. [10] [33] use “talk spurts” as data instances, assessing the effectiveness of different combinations of features for topic boundary classification, including the above mentioned features as well as prosody and motion data extracted from the video source. They test feature integration using a C4.5 (decision tree) classifier [35] and maximum entropy [36] models. Although most approaches employ supervised learning, unsupervised learning has also been attempted [14] using features derived from phonotactic models [39] with some degree of success.

In this paper, we explore the use of paralinguistic features of group interaction, following a method based exclusively on amount and structure of speech. This framework was initially developed for analysis of two-party dialogues in psychopathology research [20] and later extended to the study of interaction in small groups [11]. It builds on a technique of content-free analysis that uses vocalisation matrices as a way of summarising conversational history. An attractive aspect of this approach is that it does not rely on transcribed speech, being therefore unaffected by speech recognition errors. We employ a Naïve Bayes classifier on a combination of continuous and discrete variables [22], obtaining promising segmentation results. The method is described in detail and evaluated in the following section.
4. MDTM SEGMENTATION

Content-free analysis summarises dialogues as “vocalisation matrices” which basically encode the amount of speech produced by a speaker in a continuous talk spurt, the duration of speech pauses, and the probabilities that a particular speaker’s vocalisation will be followed by another speaker’s vocalisation. In general, a conversation is modelled as a Markov process with respect to such transition probabilities [26]. This assumption has been shown to be effective for classification of (pre-segmented) PCDs according to the nature of the discussion (medical, surgical, referral, etc) in [30], where a graph-based representation of the PCD is adopted. The approach adopted here relaxes this assumption by allowing a number of preceding vocalisations to be encoded as part of the feature set.

The data set consists of interval of silences and vocalisations to be classified as either boundary or non-boundary instances. A boundary instance indicates the beginning of a PCD. The features used to describe an instance are encoded as a vector encompassing duration of silences or vocalisations and the roles of the speakers who uttered the vocalisations, as shown in [4].

\[ s = (V_0, L_0, V_{-1}, L_{-1}, \ldots, V_n, L_n, V_1, L_1, \ldots, V_n, L_n) \] (1)

\( V_i \) is a nominal variable denoting the speaker role (or a pause type or group speech, in the cases of silences and vocalisations by more than one speaker, respectively). The speaker roles which can instantiate \( V_0, \ldots, V_n \) range over the values shown in Table 1. Although these roles are specific to MDTMs, other meetings exhibit distinct speaker roles which influence conversational structure [27]. We speculate that more general roles, such as defined in the AMI corpus, for instance, can be employed for topic segmentation in a similar way as described in this paper.

\( L_i \) is a continuous variable for the duration of the speech (or silence) interval, and the pairs \( V_{-1}, L_{-1} \) and \( V_i, L_i \) refer to the \( i^{th} \) roles and durations of vocalisation intervals preceding and following the vocalisation described by the instance, respectively.

4.1 Data preparation

As mentioned above, a dataset consisting of 54 PCDs has been segmented and manually annotated for speaker identities and roles. In addition to PCDs, the segmentation involved marking the set of dialogue states specified in Definition [1].

**Definition 1.** We distinguish the following types of dialogue states:

- **(Individual) Vocalisation:** the length of time that a speaker has the floor. A speaker takes the floor when they begin speaking to the exclusion of everyone else and speak uninterrupted without pause for at least 1 second. The vocalisation ends when a silence, another individual vocalisation or a group vocalisation begins. Talk sports shorter than 1 second (backchannels) are not annotated and are simply incorporated into the main speaker’s vocalisation.

- **Group vocalisation** occurs when an individual has fallen silent and two or more individuals are speaking together. The group vocalisation ends when any individual is again speaking alone, or a period of silence begins. Individual speaker identities are lost when a group vocalisation state is entered.

Silence represents quiet periods of over 0.9 seconds between vocalisations (including group vocalisations). A Silence ends when an individual or group vocalisation begins. A Silence can be further classified as:

- a pause: a silence between two vocalisations by the same participant,
- a switching pause: a silence between two vocalisations by different participants,
- a group pause: a silence between two group vocalisations, or
- a group switching pause: a silence between a group vocalisation and an individual vocalisation.

Annotation followed the methodology described in the social psychology and computer-supported cooperative work literature [11, 40] and therefore focused mainly on amount and structure of speech activity. The metadata created for this set of 54 PCDs are, however, much more detailed, containing information about artifacts employed during the meeting, use of informal language, roles etc. For the purposes of this paper, however, only speech activity and speaker roles are considered. The dialogue states specified in Definition [1] are similar to the ones used in [10], with an adjustment to the minimal duration of a vocalisation. Our definition of silence is similar to the concept of switching pauses described in [11]. The threshold of 0.9s in the definition of pause was determined empirically. Sellen [10], for instance, uses a threshold of 1.5s. However, her data are recorded in 2-participant remote communication scenarios in which pauses tend to be longer due to technology mediation. One could also define simplified notions of turns as sequences of vocalisations and pauses, and analogously group turns as sequences of group vocalisations and group pauses. However, we chose to avoid the term “turn” altogether, as it is used in conversation analysis [35] in a different and more complex sense.

In keeping the basic units of analysis simple, we expect to be able to automate their extraction from recorded audio through existing signal processing techniques [13]. It should be noted, however, that the processing steps necessary to turn the audio signal into sequences of dialogue states labelled by speaker (or silence) are not altogether straightforward. This process is called diarization and is usually performed through clustering of audio feature vectors using Gaussian mixture models as emission probabilities for continuous density hidden Markov models [13]. This process can be quite error prone, specially when the input consists of a single audio stream containing all speaker sources. The method described above can therefore be regarded as operating under idealised conditions, in this respect. This simplification seems less radical, however, than those assumed by most segmentation approaches reviewed in Section 3, which often rely on manual annotation or speaker identification through individual audio sources. Nevertheless, it would be interesting to test the method on data containing simulated levels of diarization error in order to assess its performance under more realistic recording conditions. Asking MDTM participants to wear wireless microphones may also offer a solution which could be tested with the user group in real work contexts.
The segmentation method consisted in training a Naïve Bayes classifier to identify instances marked as boundary dialogue states (i.e. vocalisation that start a new PCD). The conditional probabilities for the nominal variables (speaker roles) are estimated on the training set by maximum likelihood and combined into multinomial models [5], while the continuous variables are modelled through Gaussian kernels [22], as shown in [3], where $\mu_k$ and $\sigma_k^2$ are the mean and variance of the values taken by the features $L_i$ in the dataset.

$$P(L_i = x|b) = \frac{g(x; \mu_b, \sigma_b)}{\sigma_b \sqrt{2\pi}} e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}} \tag{3}$$

Since the feature sets used in our experiments contained relatively few features, no further pre-processing or feature selection steps were taken during training or classification. The number of PCDs in the test MDTM segments was assumed to be unknown. A fixed thresholding strategy [35] was adopted for boundary assignment, whereby dialogue states assigned probabilities greater than 0.5 were marked as PCD boundaries. Other strategies such as SCut and proportional thresholding could also be explored in future work but were not employed in the work reported in this paper. However, see Section 5 for further remarks on thresholding strategies.

### 4.3 Evaluation

Although IR metrics such as precision, recall, F scores, and accuracy, have been used to evaluate applications that combine topic segmentation and detection [4], the usual way to evaluate meeting segmentation is to employ metrics originally developed for text segmentation. For a segmentation task defined in terms of classification, as in this paper, accuracy scores are misleadingly high due to the fact that the dataset is highly unbalanced. Since only about 3% of instances are positive, a trivial classifier assigning non-boundary labels to all instances would predict accurately about 97% of the time. Precision, recall and F scores are also difficult to interpret, even if restricted to the positive class, since they penalise near misses (hypothesised boundaries that fall near true boundaries) and predictions that are wide off the mark equally. We therefore employ two slightly different error metrics from text segmentation now widely used in speech topic segmentation: $P_k$ [3], WindowDiff (WD) [35].

The $P_k$ metric gives the probability that two vocalisations occurring $k$ vocalisations apart and picked otherwise randomly from the dataset are incorrectly identified by the algorithm as belonging to the same or different PCDs. This is formally stated in equation (4), where $r$ and $h$ denote the reference and hypothesis segmentation, respectively.

$$P_k(r, h) = \sum_{1 \leq i, j \leq N} D_k(i, j)|1 - \delta(a(r_i, r_j), a(h_i, h_j))| \tag{4}$$

$D_k$ stands for a distribution with probability fixed at a distance $k$ (chosen to be half the average segment size, in number of vocalisations), $\delta$ returns 1 if $i$ and $j$ belong to the same PCD and 0 otherwise, and $\delta$ returns 1 if its two arguments are equal and 0 otherwise (Kronecker delta). This results in an increment if boundaries are assigned inconsistently within a segment.

---

**Table 2: Descriptive statistics for patient case discussions.**

<table>
<thead>
<tr>
<th>Description</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCD duration</td>
<td>243.451 sec.</td>
</tr>
<tr>
<td>Mean vocalisation length (per PCD)</td>
<td>8.554 sec.</td>
</tr>
<tr>
<td>Mean vocalisation length (overall)</td>
<td>8.226 sec.</td>
</tr>
<tr>
<td>No. speakers per PCD</td>
<td>8.149</td>
</tr>
<tr>
<td>Vocalisations per PCD</td>
<td>29.596</td>
</tr>
<tr>
<td>No. vocal. per speaker (per PCD)</td>
<td>3.507</td>
</tr>
<tr>
<td>Vocalisations per minute</td>
<td>8.758</td>
</tr>
<tr>
<td>Group Vocalisation length</td>
<td>1.744 sec.</td>
</tr>
<tr>
<td>Silence per PCD</td>
<td>4.120%</td>
</tr>
<tr>
<td>Entropy (speaker transitions)</td>
<td>0.78</td>
</tr>
<tr>
<td>Entropy (vocalisations lengths)</td>
<td>2.23</td>
</tr>
<tr>
<td>Participation ratio</td>
<td>0.388</td>
</tr>
</tbody>
</table>

General statistics on the PCDs in our corpus are presented in Table 2. These figures include the average duration and frequency of vocalisations overall (in an MDTM), the average number of vocalisations per participant during a PCD, the average number of speakers participating in each case discussion and the average duration of intervals of silence. The last row contains the mean values for a metric we call participation ratio. The participation ratio of a meeting participant is defined as the ratio between the number of case discussions in which they took active part, and the total number of cases discussed. The figures for mean participation ratio (over $n$ speakers) in Table 2 were calculated according to equation (2), where $C_i$ represent the set of PCDs in which speaker $s_i$ produced at least one vocalisation and $C$ is the entire set of PCDs.

$$r = \frac{\sum_{i=1}^{n} |C_i|}{n |C|} \tag{2}$$

Participation ratio figures summarise variability in the composition of the groups across case discussions. Table 2 indicates a high degree of variation in both medical and surgical meetings, showing that a speaker will on average take part in only around 39% of all PCDs. A different measure of variability is given by the Entropy ($H$) for the probability distribution of vocalisations by $n$ speakers, calculated by averaging over the probabilities $p_i$ that speaker $s_i$ is speaking at a given time during PCD, in the usual way: $H = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$. The $H$ score for speaker transitions ($H = 0.78$, $sd = 0.35$) reveals a predictable pattern of speaker transitions while the entropy of the vocalisation length is somewhat less predictable ($H = 2.23$, $sd = 0.46$), though still very stable, considering that the we have on average more than eight participants in a PCD.

### 4.2 Segmentation method

The annotation streams were converted from the ELAN format into an R data frame representing a collection of instances of the form specified in [6]. We created alternative datasets by varying the size of the window over previous and next dialogue state (a horizon of size $n$ role-length pairs on each side of the target dialogue state), and by distinguishing or not between different pause types (see Definitions [4], in order to assess the effect that these contextual parameters might have on segmentation accuracy.
Table 3: Results of segmentation experiments for 1 \leq n \leq 7 vocalisation horizons, with and without pause type discrimination.

<table>
<thead>
<tr>
<th>n</th>
<th>Pause types included</th>
<th>No pause types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P_k, pause types incl.</td>
<td>P_k, no pause types</td>
</tr>
<tr>
<td>1</td>
<td>31.9%</td>
<td>37.0%</td>
</tr>
<tr>
<td>2</td>
<td>31.0%</td>
<td>43.0%</td>
</tr>
<tr>
<td>3</td>
<td>27.6%</td>
<td>38.8%</td>
</tr>
<tr>
<td>4</td>
<td>27.8%</td>
<td>35.7%</td>
</tr>
<tr>
<td>5</td>
<td>28.1%</td>
<td>34.7%</td>
</tr>
<tr>
<td>6</td>
<td>31.8%</td>
<td>40.0%</td>
</tr>
<tr>
<td>7</td>
<td>31.1%</td>
<td>43.4%</td>
</tr>
</tbody>
</table>

The WD metric is based on a similar idea. It can also be regarded as an estimate of inconsistencies between reference and hypothesis, obtained by sliding a window of length equal to k segments over the MDTM and counting disagreements. WD, however, takes into account number of boundaries predicted by the algorithm and the number actually contained in the reference for the calculation of the error score. The score is calculated as shown in (5). N is the number of sub-segments of size k, as before, and \( b(i, j) \) gives the number of PCD boundaries between segments \( i \) and \( j \).

\[
WD(r, h) = \frac{\sum_{i=1}^{N-k} [1 - \delta(b(r_i, r_{i+k}), b(h_i, h_{i+k}))]}{N - k} \tag{5}
\]

Table 3 shows the performance of the segmentation algorithm in a 5-fold cross validation experiment in which we compared different window sizes, and data representations. Two alternative representations were assessed. In one of them the algorithm distinguished between the various types of pauses specified in Definition 1. In the other, it labelled all types of pauses simply as “silence”. Results showed that discriminating between pause types (switching pauses, group switching pauses, vocalisation pauses and group pauses) and increasing the vocalisation context horizon both have a positive effect on segmentation accuracy. As the context horizon is increased past 5 vocalisations on each side of the current segment performance degrades considerably (as shown clearly in Figure 1), as a consequence of data sparsity.

A closer analysis of the predictions revealed that WD scores are considerably higher than \( P_k \) scores due to the fact that the algorithm over-predicts boundaries around the true boundary (sometimes predicting as many as 4 hypothetical boundaries adjacent to the true boundary). This is an interesting phenomenon which further supports the hypothesis that the sequential structure of speech exchanges is indicative of higher level (topic) structures. In addition, from a pragmatic perspective, since adjacent boundaries do not occur in practice, this algorithm’s behaviour offers a straightforward possibility for improvement simply by filtering the excess boundaries at a post-processing step. Figure 2 shows the segmentation profile for an interval of an MDTM. Peaks represent dialogue states (vocalisations or silences) marked as PCD boundaries. The black lines starting at the bottom of each chart represent the manually annotated data (reference) while the raised red lines represent the boundaries found by the system (hypothesis). The wider peaks seen on the top chart indicate clusters of adjacent dialogue states identified as boundaries. An overview of the probabilities assigned to each dialogue state by the classifier is shown in the chart in the middle. The bottom chart shows the boundary assignment results after a simple filtering algorithm has been applied which selects among a cluster of adjacent hypothetical boundaries the one with the highest probability (as assigned by the Naïve Bayes classifier) as the true boundary state.

5. DISCUSSION

The system presented above achieves performance levels comparable to those achieved by state of the art supervised systems for segmentation of meetings by topic, while using much simpler content-free features. The decision tree approach presented in [14], which is based on lexical cohesion features extracted from hand-transcribed speech from the ICSI corpus, has error rates of 31.9% (\( P_k \)) and 35.9% (WD). The authors report these results to be significantly better than results of other approaches originally designed for text segmentation [13, 10], whose error scores on the same corpus range from 37.4% to 58%. Hsueh et al. [19] report that a lexical cohesion segmentation approach applied to topic segmentation of the AMI corpus produces a \( P_k \) score of about 40% and a WD score of 47%. Their improved maximum entropy segmentation algorithm, which combines lexical, conversational, prosody, video and contextual features achieves 34% (\( P_k \)) and 36% (WD). These scores were obtained on the task that includes sub-topics, whose ratio of boundary segments to total number of segments is similar to the same ratio for the MDTM corpus. The authors also show that moderate levels of word error rates in speech recognition cause only slight degradation in performance, and that not all classes of features are equally important. Somewhat in agreement with the hypothesis investigated in this paper, they find that conversational features are the most essential non-lexical features for topic segmentation.

Although task and corpus differences do not allow a detailed comparison of our results with the ones reported for the above mentioned systems, we note that for a comparable proportion of target boundaries our approach, based solely on amount of speech, speaker transition and role description features, attains lower error rates (27.6% and 34.7%
for $P_k$ and WD respectively) than those more elaborate approaches. It is likely that PCDs are better structured and homogeneous with respect to turn-taking than topics in more general meetings (even scenario-based ones) and that this structure is captured by our model. It would be interesting to test this hypothesis by applying the technique presented in this paper to topic segmentation in other corpora.

From the practical point of view of implementing a searchable multimedia archive of MDTMs usable in a real-world application, segmentation is an initial but important step. Due to its relatively high error rates, it is unlikely that current segmentation methods could be used for storage of PCD discussion records as separate units on a database system. Rather, we envision an interaction mode in which the user, for instance, "browses" time-based media containing recordings of MDTMs in order to locate the information of interest. The method presented above, even though it clearly over-predicts, could usefully support this interaction mode. In browsing, high recall is often favoured over precision. When presented with a misidentified PCD boundary (a false positive), the user can usually identify it as such after a few seconds of listening and skip over to the next boundary. In that regard, it is worth pointing again to Figure 2 and noting that the profile is dominated by zero or very low probabilities (representing true negatives), and that for all missed boundaries (false negatives) the probabilities peak to values greater than those of true negatives. Therefore, if one were to adjust the classification threshold one could optimise the utility of the classifier (in a decision-theoretic sense, valuing recall over precision) for this particular interaction mode. Usability studies to determine and test such parameters are a promising area for future work.

Although the information generated at MDTMs constitute valuable resources for a number of processes in healthcare, from patient management to teaching, the incorporation of MDTM-generated data into existing patient-centred models is far from straightforward. Given that MDT meeting participants work under tight time constraints, automatic recording seems to be the only viable option for data gathering. Recording and storage of multimedia meeting data in digital form have become relatively commonplace in recent years. The challenge consists in finding effective ways of structuring and providing easy access to these data.

6. CONCLUSION

This paper demonstrated the use of a simple data representation technique inspired by research on dialogue in the fields of social psychology and computer supported cooperative work in an automatic topic segmentation task. The combination of nominal and continuous features derived from amount and sequence of speech and speaker roles through a Naïve Bayes classifier yielded promising results when applied to the segmentation of multidisciplinary medical team meetings into patient case discussions, achieving...
performance levels that compare favourably to state of the art meeting segmentation techniques.

The work described here forms part of an ongoing study aimed at understanding the task and process at play in MDTMs with a view to identifying ways in which computer technology might be deployed in such settings. This includes an investigation into the possibility of enriching existing electronic health records with automatic segmentation and indexing of patient case discussions. Such a system would potentially allow users to easily retrieve PCDs for teaching and healthcare management purposes. In order to achieve these goals, in addition to segmentation, we are currently tackling the issue of automatically categorising PCDs as well as carrying out further fieldwork studies in the hospital environment.

Work in progress also includes further analysis of annotations of the MDTM corpus and identification of patient case discussion stages. Future work will explore the detection of specific salient events related to such discussion stages. An example of a salient event is the TNM (Tumour, Nodes, Metastases) “staging” (categorisation) by the meeting participants. Finally, we intend to test our approach to segmentation on larger, publicly available datasets such as the ICSI and the AMI meeting corpora.

7. ACKNOWLEDGEMENTS

The author would like to thank Bridget Kane for introducing him to the world of multidisciplinary medical team meetings and for many interesting discussions on the subject. The feedback from the SSCS reviewers is kindly acknowledged. This research is funded by Science Foundation Ireland under the Research Frontiers program.

8. REFERENCES


