The Relevance of Timing, Pauses and Overlaps in Dialogues: Detecting Topic Changes in Scenario Based Meetings

Saturnino Luz, Jing Su

Department of Computer Science, Trinity College Dublin, Ireland
{luzs,sujing}@cs.tcd.ie

Abstract

We present an investigation of the relevance of simple conversational features as indicators of topic shifts in small-group meetings. Three proposals for representation of dialogue data are described, and their effectiveness assessed at detecting topic boundaries on a large section of the Augmented Multi-Party Interaction (AMI) corpus. These proposals consist in representing a speech event through combinations of features such as the lengths of vocalisations, pauses and speech overlaps in the immediate temporal context of the event. Results show that timing of vocalisations alone, within a 7-vocalisation window (3 on each side of the vocalisation under consideration), can be an effective predictor of topic boundaries, outperforming topic segmentation methods based on lexical features. Pause and overlap information on their own also yield comparably good segmentation accuracy, suggesting that simple methods could complement or even serve as alternatives to methods which require more demanding speech processing for meeting browsing. 

Index Terms: multi-party interaction, meeting browsing, topic segmentation, speech, dialogue, conversational features.

1. Introduction

Segmentation of dialogue into topics has been recognised as an important step toward providing support for the task of browsing, reviewing and finding information in audio (and multimedia) recordings of meetings [1]. A meeting topic can be loosely defined as a cohesive sequence of dialogue turns focusing on a clearly identifiable subject. Evidence from data-driven efforts aimed at studying multi-party interaction in meetings suggests that topics and topic boundaries can be reliably identified by independent human annotators of meeting corpora [2]. However, attempts at automatically identifying topic boundaries have so far failed to produce the levels of performance required to provide users with a reliable structure for content browsing [3].

Topic segmentation of speech data is a challenging problem which often involves solving a number of sub-problems, some of which are also challenging in their own right. It may involve, for instance, separating the speech signal from noise, identifying the different speech sources (diarisation), transcribing the spoken content (automatic speech recognition, ASR), extracting paralinguistic features from the audio, detecting non-speech cues, segmenting the speech into turns (or talk spurts) etc.

For data such as broadcast news speech [4], segmentation is facilitated by the fact that a certain amount of structure (predictable sequencing, salient audio cues, etc) is built into the recording, and that the speech signal is mostly of good quality, allowing for fairly accurate ASR. Meeting data, on the other hand, are considerably harder to deal with. Natural dialogue is marked by disfluency, interruptions, overlaps, and noise, which make processing tasks such as diarisation and ASR a lot more difficult. Furthermore, while meetings are often structured along a set of agenda items, they lack the pre-defined topic structure of a produced broadcast.

While most approaches to meeting topic segmentation to date have used transcriptions (manual or ASR-generated [5, 6, 3, 7, 8]) as the basis for boundary identification, in this study we focus solely on vocalisation events (talk spurts, pauses and speech overlaps) and investigate the extent to which sequences of such events, on their own, can predict topic shifts. Although moderately erroneous ASR (30% WER) does not seem to affect segmentation accuracy [6], methods that combine transcription and conversational features of the sort we call vocalisation events are constrained by the need to treat a potential segment as a sequence of windows over which lexical chains or lexical cohesion characteristics [5] are computed. Consequently, the conversational features can only be described as means (speaker activity change, overlap and silence rates, etc) over the duration of such windows [5, 6]. We hypothesise that this representational strategy is too coarse-grained to capture the structure of vocalisation sequences as indicators of topic boundaries, and propose an alternative strategy in a Bayesian framework.

In what follows we describe an approach which uses a simpler but finer-grained representation of the dialogue which allows for windows of variable length to be represented. This approach regards each vocalisation event as a data instance represented in terms of its duration, the role its speaker play in the meeting, and the vocalisation events in its immediate context. We tested this representation on a standard benchmark, the Augmented Multi-Party Interaction (AMI) corpus, and found that our results improve upon other results reported in the meeting topic segmentation literature.

2. Methods

Topic segmentation in written text is generally treated as the categorisation problem of marking out the text units which signal changes of topics. Lexical cohesion statistics have been shown to be effective indicators of topic boundaries in text. This approach has been adapted for meeting transcript segmentation [5], but methods based on supervised machine learning [7, 6], and combinations of supervised and unsupervised methods in frameworks such as the Hidden Markov model (HMM) [8] have dominated most recent research.

We also define the problem as a categorisation task and treat it from a supervised machine learning perspective, employing a Naive Bayes classifier for boundary inference. However, our method differs from the above reviewed methods in that it employs a characterisation of vocalisation events which includes elements of its surrounding context in the dialogue. In this framework, segmenting the speech signal into a sequence of
topic segments involves the following steps: separating the input speech into vocalisation streams according to speaker, converting these streams into one of the representations described in Section 2.1, estimating the parameters for the model on annotated data, classifying the vocalisations according to the learnt model, and assembling the sequence of topic boundary markers for output. For naturally recorded speech, the first step would involve speaker diarisation. For our experiments with the AMI corpus, however, the first two steps were performed directly on the annotated data. This process is detailed below.

2.1. Turn representation

Unlike lexical approaches, ours assumes the basic segmentation units (i.e. the instances to be marked as boundaries) to be vocalisations. A meeting is defined as a sequence of individual vocalisations \( V = (v_1, \ldots, v_n) \) on a time line, where there might be gaps (silent pauses) or overlaps (simultaneous speech by two or more meeting participants) between vocalisations. Together these comprise the meeting’s vocalisation events.

A vocalisation can be regarded as a talk spurt of a certain duration \( d \), produced by a speaker \( s \) at time \( t \). One could also try to capture the dynamics of the conversation in which a vocalisation was produced by regarding the vocalisation events that preceded and followed it as part of its defining features. We investigate three ways of representing vocalisations which incorporate different combinations of these elements.

The first, defines a vocalisation exclusively in terms of the durations of other vocalisation events. The instances to be categorised are represented as shown in (1), where \( s \) is the speaker’s role, \( t \) is the start time of the vocalisation, \( d \) its duration, and \( d_i \) the duration of the \( i^{th} \) vocalisation preceding (for \( i < 0 \)) or following (for \( i > 0 \)) the current vocalisation. We use \( V_s \) to denote, alternatively, the set of vocalisations represented this way and the representation itself.

\[
v_s = (s, t, d, d_{i-1}, \ldots, d_{n}, d_1, \ldots, d_n)
\]  

(1)

The second alternative representation, denoted \( V_v \), is given by equation (2). In this case, the vocalisation is defined exclusively in terms of the silent pauses and the speech overlaps that surround it, in addition to its own duration, speaker label and time. The symbol \( g_i \) stands for the duration of the \( i^{th} \) pause or overlap preceding \( (i < 0) \) or following \( (i > 0) \) the current vocalisation. If \( g_i > 0 \) it represents the duration of a pause (‘gap’) between two vocalisations. If \( g_i < 0 \) it represents the duration of a speech overlap (i.e. the actual duration is \( |g_i| \), the absolute value of \( g_i \)).

\[
v_v = (s, t, d, g_{i-1}, \ldots, g_{n}, g_1, \ldots, g_n)
\]  

(2)

Finally, representation \( V_o \) combines vocalisations, pauses and overlaps in its data representation, as defined in (3). The variables \( s, d, d, \) and \( g_i \) are interpreted as in (1) and (2).

\[
v_o = (s, t, d, d_{i-1}, \ldots, d_{n}, g_{i-1}, \ldots, g_n)
\]  

(3)

2.2. Topic boundary detection

Boundary detection is defined as a mapping from vocalisations to topic boundary markers, \( f : V \rightarrow \{0, 1\} \). We employ a Naïve Bayes strategy to learn an approximation \( h \) of \( f \) from topic-annotated training data. The Naïve Bayes classifier was chosen because it can easily accommodate the combination of nominal and continuous variables described in Section 2.1 quite naturally, and can, in principle, be augmented to encode ordering dependencies \([9]\). The categorisation procedure consists in estimating the probability that a vocalisation is a begin-topic boundary given its vector representation \( P(B = 1|V = v) \), and then making a hard categorisation decision by thresholding, i.e. assigning \( h(v) = 1 \) if \( P(B = 1|V = v) \geq \tau \), where \( \tau \) could be set to 0.5 (Maximum A Posteriori, MAP, hypothesis) or empirically through cross-validation. Applying Bayes’ rule, the probabilities to be estimated are:

\[
P(b|V = v) \propto P(F_1 = x_1, \ldots, F_m = x_m|b)
\]

(4)

\[
= \prod_{i=1}^{m} P(F = x_i|b)
\]

(5)

where \( x_1, \ldots, x_m \) are the elements of the alternative feature vectors defined above (i.e. \( s, t, d, g_1, d_1 \) etc) and (5) is warranted by the conditional independence assumption.

The probabilities for the nominal variables can be learnt from the training set by maximum likelihood estimation, and the continuous variables can be modelled as Gaussian kernels:

\[
P(F_i = x|b) = \frac{1}{\sigma_b \sqrt{2\pi}} e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}}
\]

(6)

The segmentation algorithm consists of running each vocalisation event in sequence through the boundary classifier, and then assembling the results as a sequence of boundary markers (1’s) and within-topic events (0’s).

3. Experiments

We conducted a series of experiments in a section of the English language scenario-based meetings recorded in the AMI corpus \([1]\) in order to assess the performance of topic segmentation algorithms based on the representations defined in (1)-(3). The original AMI XML files were first converted into the \( V_s \), \( V_v \) and \( V_o \) formats, and a 20-fold cross validation procedure was applied to the resulting data sets. The outcomes were then compared between representations, as well as to results of other segmentation approaches reported in the literature.

3.1. Corpus preparation

We used the scenario-based portion of the AMI corpus for evaluation. In these meetings, four participants are asked to play the roles of a design team engaged in the activity of designing a new product. The roles are well defined (user interface designer, industrial designer, project manager and marketing expert) and consistently annotated in the corpus. In our data, we used role labels to identify the speakers across all meetings.

The AMI data we used are annotated on several layers of XML files containing information on participants, roles, actions, and dialogue acts as well as topics. Not all meetings contain all annotation levels, but the meetings in the subset we used all contain topic annotation. Timing information is marked at the word level. Therefore, converting the data from the AMI format to the format described in Section 2.1 required synchronising the word timings for each speaker channel into vocalisation streams. Our conversion algorithm did this by gathering all words for each speaker and sorting them by start time. Intervals of talk by the same speaker with no pause longer than 0.5s were marked as vocalisations. Pauses and overlaps were identified by
intersecting the four separate vocalisation channels. Pilot experiments were conducted to determine the optimal length n of left and right contexts. These experiments indicated that 3 ≤ n ≤ 5 produced the best results. We set n = 3 for all representations used in the experiments reported in this paper.

The topic structure of the AMI meetings is given by a shallow (3-level) hierarchy which includes top-level topics, subtopics and functional descriptions. Since the meetings were scenario-based, certain standard topics often reoccurred. We are not aware of any studies on the reliability of topic boundary placement for the AMI corpus in terms of segmentation metrics (as done in [2] for the ICSI data), but [6] calculated kappa values of .79 and .73 for topic and sub-topic boundary inter-rater agreement, respectively. Although [6] distinguish top-level from subtopic segmentation, we followed the more usual practice of flattening the hierarchy in order to assign topic boundaries. Approximately 26 hours of meetings were prepared according with the procedure described above. The mean durations of vocalisation events were: 0.3s for overlaps, 1.3s for pauses and 1.7s for vocalisations. The distributions of these vocalisation events were: 0.3s for overlaps, 1.3s for pauses and 1.7s for vocalisations. The distributions of these events were highly skewed (Zipfian). Pauses and speech overlaps accounted for a substantial portion of recording time (17% and 14%, respectively). Vocalisations marked as topic bounded pauses and 1.7s for vocalisations. The distributions of these vocalisation events were: 0.3s for overlaps, 1.3s for pauses and 1.7s for vocalisations. The distributions of these events were highly skewed (Zipfian). Pauses and speech overlaps accounted for a substantial portion of recording time (17% and 14%, respectively). Vocalisations marked as topic boundaries comprised 1.3% of all vocalisations.

3.2. Results

The experiment aimed primarily at evaluating the performance of the segmentation algorithm on the different data representation schemes, Vg, Vp and Vd. Cross validation was run on each data set by splitting it into 20 folds, and iterating through them: training on 19 folds and testing it on the remaining one.

We also tested two thresholding strategies: MAP hypothesis and proportional thresholding (PT). This was done in order to assess whether topic generality information could be useful for segmentation. Based on our experience with segmentation of medical meetings [10] we hypothesised it would.

Finally, we tested a slightly different version of the algorithm, modified to take advantage of some basic prior knowledge. This version explored the fact that two or more topic boundaries cannot occur immediately next to each other. A simple post-processing procedure was applied which searched the classifier output sequence for such clusters of consecutive 1’s, assigning 0 to all vocalisations in them except the one with the highest probability Ph(h). We denote the original version of the algorithm An and the modified version A1.

Segmentation is usually evaluated by two closely related error metrics which, unlike the categorisation metrics of precision and recall, allow for near misses to be taken into account [11]. These metrics are known as Pk and WindowDiff, or WD.

$$P_k(r, h) = \sum_{1 \leq i \leq j \leq N} D_0(i, j) [1 - \delta(a(r_i, r_j), a(h_i, h_j))]$$ (7)

$$WD(r, h) = \sum_{i=1}^{E} E^{-1}[1 - \delta(b(r_i, r_{i+k}), b(h_i, h_{i+k}))]$$ (8)

Since the data set is highly unbalanced, accuracy, precision and recall scores would be misleadingly high for negative instances and, as mentioned above, would fail to distinguish near-misses for positive instances. They are therefore not reported, in accordance with the literature on meeting segmentation.

A summary of the evaluation is presented in Table 1. Baseline scores are displayed in the last row. They were obtained by repeatedly (100 iterations) taking an unmarked sequence and marking boundaries on it at random but in the same proportion as in the training reference sequences, and averaging the evaluation scores. All methods significantly outperformed the baseline (paired t-tests, p < 0.01, for all conditions).

It is apparent that the combination of Vd, MAP hypothesis thresholding, and the algorithm for filtering of adjacent boundary hypothesis (A1) produced the best scores in general. However, analysis of variance shows that this apparent advantage is not statistically significant (F[2, 57] = 0.05, p = 0.95). In fact, the Pk scores were remarkably close for Vg and Vp, which is somewhat surprising since it suggests that pauses and overlaps may be as good predictors of topic shifts as vocalisations, even though vocalisation features might work better combined (Vd) than separately. However, differences in WD scores between Vd and Vp (paired t-test; t[19] = 3.2, p = 0.005) and between Vg and Vp (t[19] = 3.2, p = 0.004) are significant. Since WD penalises multiple spurious boundaries more than Pk, this indicates that Vp tends to lead to overprediction of boundaries.

The comparison between the MAP hypothesis thresholding approach and proportional thresholding proved inconclusive when considered with respect to representations. A three-way analysis of variance showed no interactions for Pk (F[11, 228] = 0.49, p = 0.91). There were interactions between representation type and post processing (F[2, 228] = 4.91, p < 0.01) and between thresholding strategy and post processing (F[1, 228] = 11.40, p < 0.001) for WD scores. For aggregated scores (including An and A1) the slight advantage of MAP is significant with respect to Pk (p = 0.04, Tukey
Table 2: Comparison with other meeting segmentation methods. Only the best reported results are presented.

<table>
<thead>
<tr>
<th>Method</th>
<th>Corpus</th>
<th>Segm. level</th>
<th>( P_k )</th>
<th>WD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCSeg [5]</td>
<td>ICSI</td>
<td>top-level</td>
<td>31.91%</td>
<td>35.88%</td>
</tr>
<tr>
<td>LCSeg [8]</td>
<td>ICSI</td>
<td>sub-topic</td>
<td>35.29%</td>
<td>42.00%</td>
</tr>
<tr>
<td>HMM [8]</td>
<td>ICSI</td>
<td>sub-topic</td>
<td>32.70%</td>
<td>39.80%</td>
</tr>
<tr>
<td>MAXENT [7]</td>
<td>AMI</td>
<td>top-level</td>
<td>30.00%</td>
<td>33.00%</td>
</tr>
<tr>
<td>LCSeg [7]</td>
<td>AMI</td>
<td>sub-topic</td>
<td>40.00%</td>
<td>47.00%</td>
</tr>
<tr>
<td>MAXENT [7]</td>
<td>AMI</td>
<td>sub-topic</td>
<td>34.00%</td>
<td>36.00%</td>
</tr>
<tr>
<td>( V_a + MAP + A_f )</td>
<td>AMI</td>
<td>sub-topic</td>
<td>27.67%</td>
<td>36.00%</td>
</tr>
</tbody>
</table>

HSD3), while PT yields significantly better scores in terms of WD (\( p < 0.01, \) Tukey HSD). These mixed results indicate that, contrary to our hypothesis, proportional thresholding was unable to generalise topic shifts from the training data even if, as the WD scores show, it succeeded in mitigating overprediction.

Output post-processing also improved segmentation with respect to WD (3-way ANOVA \( F[1, 235] = 31.82, p < 0.01 \)). The improvement observed in WD scores though significant was unsurprising since WD penalises placement of repeated boundary hypothesis in a segment and post-processing removes impossible boundaries. On the other hand, \( P_k \) scores were not statistically significantly different for \( A_a \) and \( A_f \). This suggests that there is still room for improvement in post-processing.

### 3.3. Discussion

The results reported above seem to support the idea that, taken in context, speaker role information and conversational features such as vocalisation timing, pauses and overlaps can be good predictors of topic changes in meeting dialogues. These results compared favourably to those of methods which employ much more informative representation schemes, as summarised in Table 2. It should be noted, however, that an entirely fair comparison of performance based on these reported results is impossible. Some evaluations were performed on a different corpus (ICSI1), and some systems targeted a somewhat different segmentation task (top-level topic segmentation), as indicated.

The sub-topic segmentation setting in [7] (results shown in rows 5 and 6 in Table 2) is probably the closest to our experimental setup. The maximum entropy approach [7] employed a combination of lexical, prosodic, and conversational features extracted from audio, as well as motion features extracted from video. The main difference in the use of conversational features in this method and ours (apart from the fact that they are used in combination with other data) is that we use these features to encode the context of the vocalisation to be represented. Although our classifier does not model the influences of these context variables on one another directly, encoding the features explicitly as we do seems to be better than recording them simply as rates over a time window.

The HMM framework employed by [8] also tries to account for context, but with respect to sentences, as it relies on transcribed speech input. Their results are reported in terms of modified versions of the standard metrics, which might slightly overestimate the error. It would be interesting to attempt to combine our method with the HMM approach in future.

Despite the fact that the AMI corpus is a highly unbalanced data set with respect to topic boundaries, the MAP hypothesis worked fairly well in converting the probabilities estimated according to the model into hard classification decisions. This is in part due to the well known fact that Naïve Bayes classifiers perform well under zero-one loss functions, even though their probability estimates are often inaccurate. The disappointing performance of proportional thresholding may also be due to the fact that topic shifts in the AMI corpus are far less regular than one might expect from scenario based meetings.

### 4. Conclusion

While the potential relevance of conversational features such as pauses and overlaps have been acknowledged [5, 7], to the best of our knowledge this is the first work to systematically explore the relevance of these features to topic segmentation by representing vocalisations in terms of their adjacent pauses and overlaps. These features proved effective as predictors of topic boundaries. Surprisingly, pauses and overlaps alone were found to be as effective as vocalisations at this task.

The experimental results point to some research avenues, such as: the development of more sophisticated algorithms for post-processing of output segmentation sequences, and a relaxation of the independence assumption in the manner suggested by [9] to allow the Naïve Bayes model to account for the ordering of variables in the left and right contexts of a vocalisation.

### 5. Acknowledgements

This research was supported by Science Foundation Ireland Research Frontiers grant under the National Development Plan. We also thank the reviewers for their feedback and suggestions.

### 6. References