Unsupervised Topic Modelling for Multi-Party Spoken Discourse

Matthew Purver
CSLI
Stanford University
Stanford, CA 94305, USA
mpurver@stanford.edu

Konrad P. Körding
Dept. of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
kording@mit.edu

Thomas L. Griffiths
Dept. of Cognitive & Linguistic Sciences
Brown University
Providence, RI 02912, USA
tom_griffiths@brown.edu

Joshua B. Tenenbaum
Dept. of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
jbt@mit.edu

Abstract

We present a method for unsupervised topic modelling which adapts methods used in document classification (Blei et al., 2003; Griffiths and Steyvers, 2004) to unsegmented multi-party discourse transcripts. We show how Bayesian inference in this generative model can be used to simultaneously address the problems of topic *segmentation* and topic *identification*: automatically segmenting multi-party meetings into topically coherent segments with performance which compares well with previous unsupervised segmentation-only methods (Galley et al., 2003) while simultaneously extracting topics which rate highly when assessed for coherence by human judges. We also show that this method appears robust in the face of off-topic dialogue and speech recognition errors.

1 Introduction

Topic *segmentation* – division of a text or discourse into topically coherent segments – and topic *identification* – classification of those segments by subject matter – are joint problems. Both are necessary steps in automatic indexing, retrieval and summarization from large datasets, whether spoken or written. Both have received significant attention in the past (see Section 2), but most approaches have been targeted at either text or monologue, and most address only one of the two issues (usually for the very good reason that the dataset itself provides the other, for example by the explicit separation of individual documents or news stories in a collection). Spoken multi-party meetings pose a difficult problem: firstly, neither the segmentation nor the discussed topics can be taken as given; secondly, the discourse is by nature less tidily structured and less restricted in domain; and thirdly, speech recognition results have unavoidably high levels of error due to the noisy multi-speaker environment.

In this paper we present a method for unsupervised topic modelling which allows us to approach both problems simultaneously, inferring a set of topics while providing a segmentation into topically coherent segments. We show that this model can address these problems over multi-party discourse transcripts, providing good segmentation performance on a corpus of meetings (comparable to the best previous unsupervised method that we are aware of (Galley et al., 2003)), while also inferring a set of topics rated as semantically coherent by human judges. We then show that its segmentation performance appears relatively robust to speech recognition errors, giving us confidence that it can be successfully applied in a real speech-processing system.

The plan of the paper is as follows. Section 2 below briefly discusses previous approaches to the identification and segmentation problems. Section 3 then describes the model we use here. Section 4 then details our experiments and results, and conclusions are drawn in Section 5.

2 Background and Related Work

In this paper we are interested in spoken discourse, and in particular multi-party human-human meetings. Our overall aim is to produce information which can be used to summarize, browse and/or retrieve the information contained in meetings. User studies (Lisowska et al., 2004; Banerjee et al., 2005) have shown that topic information is important here: people are likely to want to know

which topics were discussed in a particular meeting, as well as have access to the discussion on particular topics in which they are interested. Of course, this requires both identification of the topics discussed, and segmentation into the periods of topically related discussion.

Work on automatic topic segmentation of *text* and *monologue* has been prolific, with a variety of approaches used. (Hearst, 1994) uses a measure of lexical cohesion between adjoining paragraphs in text; (Reynar, 1999) and (Beeferman et al., 1999) combine a variety of features such as statistical language modelling, cue phrases, discourse information and the presence of pronouns or named entities to segment broadcast news; (Maskey and Hirschberg, 2003) use entirely non-lexical features. Recent advances have used generative models, allowing lexical models of the topics themselves to be built while segmenting (Imai et al., 1997; Barzilay and Lee, 2004), and we take a similar approach here, although with some important differences detailed below.

Turning to *multi-party* discourse and *meetings*, however, most previous work on automatic segmentation (Reiter and Rigoll, 2004; Dielmann and Renals, 2004; Banerjee and Rudnicky, 2004), treats segments as representing meeting phases or events which characterize the *type* or *style* of discourse taking place (presentation, briefing, discussion etc.), rather than the topic or subject matter. While we expect some correlation between these two types of segmentation, they are clearly different problems. However, one comparable study is described in (Galley et al., 2003). Here, a lexical cohesion approach was used to develop an essentially unsupervised segmentation tool (*LCSeg*) which was applied to both text and meeting transcripts, giving performance better than that achieved by applying text/monologue-based techniques (see Section 4 below), and we take this as our benchmark for the segmentation problem. Note that they improved their accuracy by combining the unsupervised output with discourse features in a supervised classifier – while we do not attempt a similar comparison here, we expect a similar technique would yield similar segmentation improvements.

In contrast, we take a generative approach, modelling the text as being generated by a sequence of mixtures of underlying topics. The approach is unsupervised, allowing both segmentation and topic extraction from unlabelled data.

3 Learning topics and segments

We specify our model to address the problem of topic segmentation: attempting to break the discourse into discrete segments in which a particular set of topics are discussed. Assume we have a corpus of U utterances, ordered in sequence. The uth utterance consists of N_u words, chosen from a vocabulary of size W. The set of words associated with the uth utterance are denoted $\mathbf{w}_u$, and indexed as $w_{u,i}$. The entire corpus is represented by $\mathbf{w}$.

Following previous work on probabilistic topic models (Hofmann, 1999; Blei et al., 2003; Griffiths and Steyvers, 2004), we model each utterance as being generated from a particular distribution over topics, where each topic is a probability distribution over words. The utterances are ordered sequentially, and we assume a Markov structure on the distribution over topics: with high probability, the distribution for utterance u is the same as for utterance $u-1$; otherwise, we sample a new distribution over topics. This pattern of dependency is produced by associating a binary switching variable with each utterance, indicating whether its topic is the same as that of the previous utterance. The joint states of all the switching variables define segments that should be semantically coherent, because their words are generated by the same topic vector. We will first describe this generative model in more detail, and then discuss inference in this model.

3.1 A hierarchical Bayesian model

We are interested in where changes occur in the set of topics discussed in these utterances. To this end, let c_u indicate whether a change in the distribution over topics occurs at the uth utterance and let $P(c_u = 1) = \pi$ (where π thus defines the expected number of segments). The distribution over topics associated with the uth utterance will be denoted $\theta^{(u)}$, and is a multinomial distribution over T topics, with the probability of topic t being $\theta_t^{(u)}$. If $c_u = 0$, then $\theta^{(u)} = \theta^{(u-1)}$. Otherwise, $\theta^{(u)}$ is drawn from a symmetric Dirichlet distribution with parameter α. The distribution is thus:

$$P(\theta^{(u)}|c_u, \theta^{(u-1)}) = \begin{cases} \delta(\theta^{(u)}, \theta^{(u-1)}) & c_u = 0 \\ \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{t=1}^{T} (\theta_t^{(u)})^{\alpha-1} & c_u = 1 \end{cases}$$

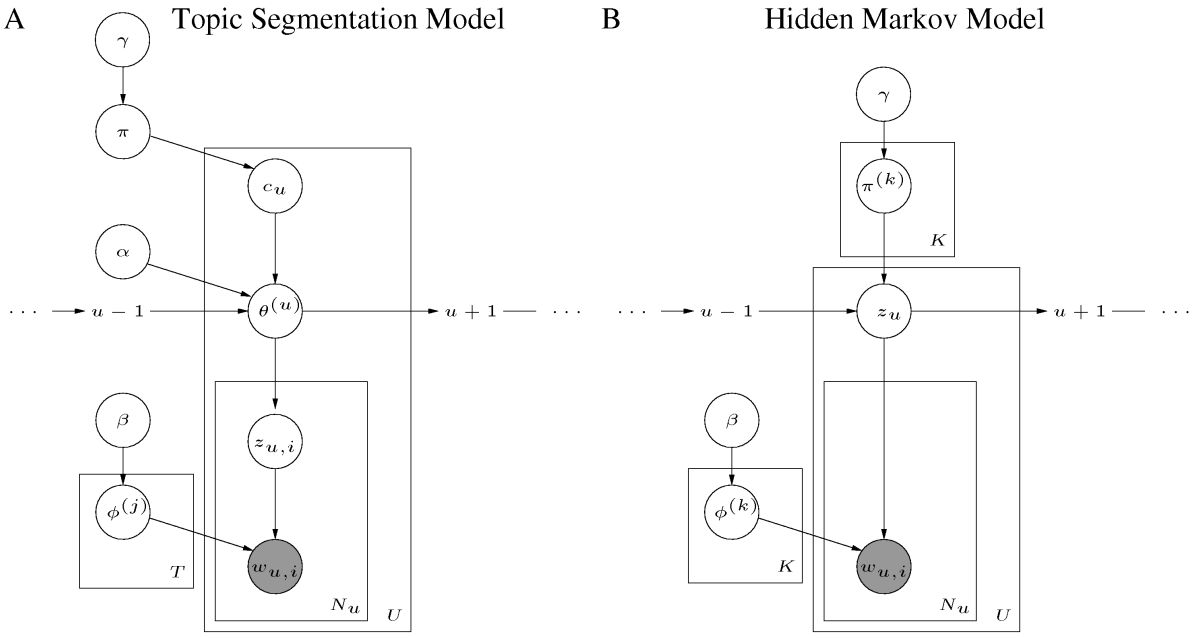

Figure 1: Graphical models indicating the dependencies among variables in (a) the topic segmentation model and (b) the hidden Markov model used as a comparison.

where $\delta(\cdot,\cdot)$ is the Dirac delta function, and $\Gamma(\cdot)$ is the generalized factorial function. This distribution is not well-defined when $u = 1$, so we set $c_1 = 1$ and draw $\theta^{(1)}$ from a symmetric Dirichlet(α) distribution accordingly.

As in (Hofmann, 1999; Blei et al., 2003; Griffiths and Steyvers, 2004), each topic T_j is a multinomial distribution $\phi^{(j)}$ over words, and the probability of the word w under that topic is $\phi_w^{(j)}$. The uth utterance is generated by sampling a topic assignment $z_{u,i}$ for each word i in that utterance with $P(z_{u,i} = t|\theta^{(u)}) = \theta_t^{(u)}$, and then sampling a word $w_{u,i}$ from $\phi^{(j)}$, with $P(w_{u,i} = w|z_{u,i} = j, \phi^{(j)}) = \phi_w^{(j)}$. If we assume that π is generated from a symmetric Beta(γ) distribution, and each $\phi^{(j)}$ is generated from a symmetric Dirichlet(β) distribution, we obtain a joint distribution over all of these variables with the dependency structure shown in Figure 1A.

3.2 Inference

Assessing the posterior probability distribution over topic changes $\mathbf{c}$ given a corpus $\mathbf{w}$ can be simplified by integrating out the parameters θ, ϕ, and π. According to Bayes rule we have:

$$P(\mathbf{z},\mathbf{c}|\mathbf{w}) = \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z}|\mathbf{c})P(\mathbf{c})}{\sum_{\mathbf{z},\mathbf{c}} P(\mathbf{w}|\mathbf{z})P(\mathbf{z}|\mathbf{c})P(\mathbf{c})} \quad (1)$$

Evaluating $P(\mathbf{c})$ requires integrating over π. Specifically, we have:

$$\begin{aligned} P(\mathbf{c}) &= \int_0^1 P(\mathbf{c}|\pi)P(\pi)\, d\pi \\ &= \frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2}\frac{\Gamma(n_1+\gamma)\Gamma(n_0+\gamma)}{\Gamma(N+2\gamma)} \end{aligned} \quad (2)$$

where n_1 is the number of utterances for which $c_u = 1$, and n_0 is the number of utterances for which $c_u = 0$. Computing $P(\mathbf{w}|\mathbf{z})$ proceeds along similar lines:

$$\begin{aligned} P(\mathbf{w}|\mathbf{z}) &= \int_{\Delta_W^T} P(\mathbf{w}|\mathbf{z},\phi)P(\phi)\, d\phi \\ &= \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{t=1}^T \frac{\prod_{w=1}^W \Gamma(n_w^{(t)}+\beta)}{\Gamma(n_\cdot^{(t)}+W\beta)} \end{aligned} \quad (3)$$

where Δ_W^T is the T-dimensional cross-product of the multinomial simplex on W points, $n_w^{(t)}$ is the number of times word w is assigned to topic t in $\mathbf{z}$, and $n_\cdot^{(t)}$ is the total number of words assigned to topic t in $\mathbf{z}$. To evaluate $P(\mathbf{z}|\mathbf{c})$ we have:

$$P(\mathbf{z}|\mathbf{c}) = \int_{\Delta_T^U} P(\mathbf{z}|\theta)P(\theta|\mathbf{c})\, d\theta \quad (4)$$

The fact that the c_u variables effectively divide the sequence of utterances into segments that use the same distribution over topics simplifies solving the integral and we obtain:

$$P(\mathbf{z}|\mathbf{c}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^{n_1} \prod_{u\in\mathcal{U}_1} \frac{\prod_{t=1}^T \Gamma(n_t^{(\mathcal{S}_u)}+\alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_u)}+T\alpha)}. \quad (5)$$

$$P(c_u|\mathbf{c}_{-u},\mathbf{z},\mathbf{w}) \propto \begin{cases} \frac{\prod_{t=1}^{T}\Gamma(n_t^{(\mathcal{S}_u^0)}+\alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_u^0)}+T\alpha)} & \frac{n_0+\gamma}{N+2\gamma} \quad c_u=0 \\ \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\frac{\prod_{t=1}^{T}\Gamma(n_t^{(\mathcal{S}_{u-1}^1)}+\alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_{u-1}^1)}+T\alpha)}\frac{\prod_{t=1}^{T}\Gamma(n_t^{(\mathcal{S}_u^1)}+\alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_u^1)}+T\alpha)} & \frac{n_1+\gamma}{N+2\gamma} \quad c_u=1 \end{cases} \tag{7}$$

where $\mathcal{U}_1 = \{u|c_u = 1\}$, $\mathcal{U}_0 = \{u|c_u = 0\}$, $\mathcal{S}_u$ denotes the set of utterances that share the same topic distribution (i.e. belong to the same segment) as u, and $n_t^{(\mathcal{S}_u)}$ is the number of times topic t appears in the segment $\mathcal{S}_u$ (i.e. in the values of $z_{u'}$ corresponding for $u' \in \mathcal{S}_u$).

Equations 2, 3, and 5 allow us to evaluate the numerator of the expression in Equation 1. However, computing the denominator is intractable. Consequently, we sample from the posterior distribution $P(\mathbf{z}, \mathbf{c}|\mathbf{w})$ using Markov chain Monte Carlo (MCMC) (Gilks et al., 1996). We use Gibbs sampling, drawing the topic assignment for each word, $z_{u,i}$, conditioned on all other topic assignments, $\mathbf{z}_{-(u,i)}$, all topic change indicators, $\mathbf{c}$, and all words, $\mathbf{w}$; and then drawing the topic change indicator for each utterance, c_u, conditioned on all other topic change indicators, $\mathbf{c}_{-u}$, all topic assignments $\mathbf{z}$, and all words $\mathbf{w}$.

The conditional probabilities we need can be derived directly from Equations 2, 3, and 5. The conditional probability of $z_{u,i}$ indicates the probability that $w_{u,i}$ should be assigned to a particular topic, given other assignments, the current segmentation, and the words in the utterances. Cancelling constant terms, we obtain:

$$P(z_{u,i}|\mathbf{z}_{-(u,i)},\mathbf{c},\mathbf{w}) \propto \frac{n_{w_{u,i}}^{(t)}+\beta}{n_\cdot^{(t)}+W\beta}\frac{n_{z_{u,i}}^{(\mathcal{S}_u)}+\alpha}{n_\cdot^{(\mathcal{S}_u)}+T\alpha}. \tag{6}$$

where all counts (i.e. the n terms) exclude $z_{u,i}$. The conditional probability of c_u indicates the probability that a new segment should start at u. In sampling c_u from this distribution, we are splitting or merging segments. Similarly we obtain the expression in (7), where $\mathcal{S}_u^1$ is $\mathcal{S}_u$ for the segmentation when $c_u = 1$, $\mathcal{S}_u^0$ is $\mathcal{S}_u$ for the segmentation when $c_u = 0$, and all counts (e.g. n_1) exclude c_u. For this paper, we fixed α, β and γ at 0.01.

Our algorithm is related to (Barzilay and Lee, 2004)'s approach to text segmentation, which uses a hidden Markov model (HMM) to model segmentation and topic inference for text using a bigram representation in restricted domains. Due to the adaptive combination of different topics our algorithm can be expected to generalize well to larger domains. It also relates to earlier work by (Blei and Moreno, 2001) that uses a topic representation but also does not allow adaptively combining different topics. However, while HMM approaches allow a segmentation of the data by topic, they do not allow adaptively combining different topics into segments: while a new segment can be modelled as being identical to a topic that has already been observed, it can not be modelled as a combination of the previously observed topics.[1] Note that while (Imai et al., 1997)'s HMM approach allows topic mixtures, it requires supervision with hand-labelled topics.

In our experiments we therefore compared our results with those obtained by a similar but simpler 10 state HMM, using a similar Gibbs sampling algorithm. The key difference between the two models is shown in Figure 1. In the HMM, all variation in the content of utterances is modelled at a single level, with each segment having a distribution over words corresponding to a single state. The hierarchical structure of our topic segmentation model allows variation in content to be expressed at two levels, with each segment being produced from a linear combination of the distributions associated with each topic. Consequently, our model can often capture the content of a sequence of words by postulating a single segment with a novel distribution over topics, while the HMM has to frequently switch between states.

4 Experiments

4.1 Experiment 0: Simulated data

To analyze the properties of this algorithm we first applied it to a simulated dataset: a sequence of 10,000 words chosen from a vocabulary of 25. Each segment of 100 successive words had a con-

[1] Say that a particular corpus leads us to infer topics corresponding to "speech recognition" and "discourse understanding". A single discussion concerning *speech recognition for discourse understanding* could be modelled by our algorithm as a single segment with a suitable weighted mixture of the two topics; a HMM approach would tend to split it into multiple segments (or require a specific topic for this segment).

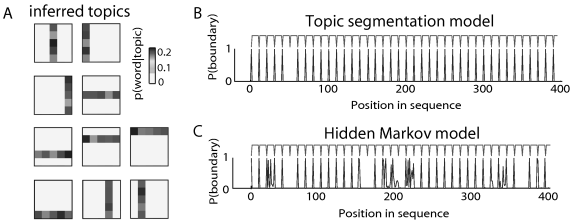

Figure 2: Simulated data: A) inferred topics; B) segmentation probabilities; C) HMM version.

stant topic distribution (with distributions for different segments drawn from a Dirichlet distribution with $\beta = 0.1$), and each subsequence of 10 words was taken to be one utterance. The topic-word assignments were chosen such that when the vocabulary is aligned in a 5×5 grid the topics were binary bars. The inference algorithm was then run for 200,000 iterations, with samples collected after every 1,000 iterations to minimize autocorrelation. Figure 2 shows the inferred topic-word distributions and segment boundaries, which correspond well with those used to generate the data.

4.2 Experiment 1: The ICSI corpus

We applied the algorithm to the ICSI meeting corpus transcripts (Janin et al., 2003), consisting of manual transcriptions of 75 meetings. For evaluation, we use (Galley et al., 2003)'s set of human-annotated segmentations, which covers a sub-portion of 25 meetings and takes a relatively coarse-grained approach to topic with an average of 5-6 topic segments per meeting. Note that these segmentations were not used in training the model: topic inference and segmentation was unsupervised, with the human annotations used only to provide some knowledge of the overall segmentation density and to evaluate performance.

The transcripts from all 75 meetings were linearized by utterance start time and merged into a single dataset that contained 607,263 word tokens. We sampled for 200,000 iterations of MCMC, taking samples every 1,000 iterations, and then averaged the sampled c_u variables over the last 100 samples to derive an estimate for the posterior probability of a segmentation boundary at each utterance start. This probability was then thresholded to derive a final segmentation which was compared to the manual annotations. More precisely, we apply a small amount of smoothing (Gaussian kernel convolution) and take the midpoints of any areas above a set threshold to be the segment boundaries. Varying this threshold allows us to segment the discourse in a more or less fine-grained way (and we anticipate that this could be user-settable in a meeting browsing application). If the correct number of segments is known for a meeting, this can be used directly to determine the optimum threshold, increasing performance; if not, we must set it at a level which corresponds to the desired general level of granularity. For each set of annotations, we therefore performed two sets of segmentations: one in which the threshold was set for each meeting to give the known gold-standard number of segments, and one in which the threshold was set on a separate development set to give the overall corpus-wide average number of segments, and held constant for all test meetings.[2] This also allows us to compare our results with those of (Galley et al., 2003), who apply a similar threshold to their lexical cohesion function and give corresponding results produced with known/unknown numbers of segments.

Segmentation We assessed segmentation performance using the P_k and *WindowDiff* (W_D) error measures proposed by (Beeferman et al., 1999) and (Pevzner and Hearst, 2002) respectively; both intuitively provide a measure of the probability that two points drawn from the meeting will be *in*correctly separated by a hypothesized segment boundary – thus, lower P_k and W_D figures indicate better agreement with the human-annotated results.[3] For the numbers of segments we are dealing with, a baseline of segmenting the discourse into equal-length segments gives both P_k and W_D about 50%. In order to investigate the effect of the number of underlying topics T, we tested models using 2, 5, 10 and 20 topics. We then compared performance with (Galley et al., 2003)'s *LCSeg* tool, and with a 10-state HMM model as described above. Results are shown in Table 1, averaged over the 25 test meetings.

Results show that our model significantly outperforms the HMM equivalent – because the HMM cannot combine different topics, it places a lot of segmentation boundaries, resulting in inferior performance. Using stemming and a bigram

[2] The development set was formed from the other meetings in the same ICSI subject areas as the annotated test meetings.

[3] W_D takes into account the likely number of incorrectly separating hypothesized boundaries; P_k only a binary correct/incorrect classification.

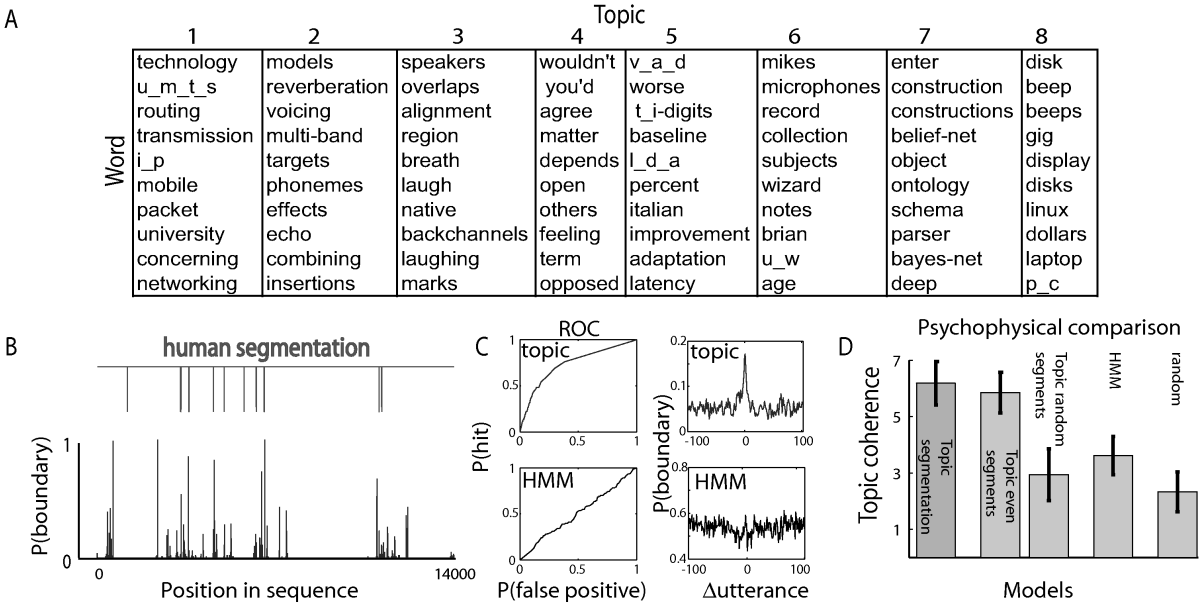

Figure 3: Results from the ICSI corpus: A) the words most indicative for each topic; B) Probability of a segment boundary, compared with human segmentation, for an arbitrary subset of the data; C) Receiver-operator characteristic (ROC) curves for predicting human segmentation, and conditional probabilities of placing a boundary at an offset from a human boundary; D) subjective topic coherence ratings.

Model	Number of topics T 2	5	10	20	HMM	*LCSeg*
P_k	.284	.297	.329	.290	.375	.319

Model	known P_k	known W_D	unknown P_k	unknown W_D
$T = 10$	.289	.329	.329	.353
LCSeg	.264	.294	.319	.359

Table 1: Results on the ICSI meeting corpus.

representation, however, might improve its performance (Barzilay and Lee, 2004), although similar benefits might equally apply to our model. It also performs comparably to (Galley et al., 2003)'s unsupervised performance (exceeding it for some settings of T). It does not perform as well as their hybrid supervised system, which combined *LCSeg* with supervised learning over discourse features ($P_k = .23$); but we expect that a similar approach would be possible here, combining our segmentation probabilities with other discourse-based features in a supervised way for improved performance. Interestingly, segmentation quality, at least at this relatively coarse-grained level, seems hardly affected by the overall number of topics T.

Figure 3B shows an example for one meeting of how the inferred topic segmentation probabilities at each utterance compare with the gold-standard segment boundaries. Figure 3C illustrates the performance difference between our model and the HMM equivalent at an example segment boundary: for this example, the HMM model gives almost no discrimination.

Identification Figure 3A shows the most indicative words for a subset of the topics inferred at the last iteration. Encouragingly, most topics seem intuitively to reflect the subjects we know were discussed in the ICSI meetings – the majority of them (67 meetings) are taken from the weekly meetings of 3 distinct research groups, where discussions centered around speech recognition techniques (topics 2, 5), meeting recording, annotation and hardware setup (topics 6, 3, 1, 8), robust language processing (topic 7). Others reflect general classes of words which are independent of subject matter (topic 4).

To compare the quality of these inferred topics we performed an experiment in which 7 human observers rated (on a scale of 1 to 9) the semantic coherence of 50 lists of 10 words each. Of these lists, 40 contained the most indicative words for each of the 10 topics from different models: the topic segmentation model; a topic model that had the same number of segments but with fixed evenly spread segmentation boundaries; an equiv-

alent with randomly placed segmentation boundaries; and the HMM. The other 10 lists contained random samples of 10 words from the other 40 lists. Results are shown in Figure 3D, with the topic segmentation model producing the most coherent topics and the HMM model and random words scoring less well. Interestingly, using an even distribution of boundaries but allowing the topic model to infer topics performs similarly well with even segmentation, but badly with random segmentation – topic quality is thus not very susceptible to the precise segmentation of the text, but does require some reasonable approximation (on ICSI data, an even segmentation gives a P_k of about 50%, while random segmentations can do much worse). However, note that the full topic segmentation model is able to identify meaningful segmentation boundaries *at the same time* as inferring topics.

4.3 Experiment 2: Dialogue robustness

Meetings often include off-topic dialogue, in particular at the beginning and end, where informal chat and meta-dialogue are common. Galley et al. (2003) annotated these sections explicitly, together with the ICSI "digit-task" sections (participants read sequences of digits to provide data for speech recognition experiments), and removed them from their data, as did we in Experiment 1 above. While this seems reasonable for the purposes of investigating ideal algorithm performance, in real situations we will be faced with such off-topic dialogue, and would obviously prefer segmentation performance not to be badly affected (and ideally, enabling segmentation of the off-topic sections from the meeting proper). One might suspect that an unsupervised generative model such as ours might not be robust in the presence of numerous off-topic words, as spurious topics might be inferred and used in the mixture model throughout. In order to investigate this, we therefore also tested on the full dataset without removing these sections (806,026 word tokens in total), and added the section boundaries as further desired gold-standard segmentation boundaries. Table 2 shows the results: performance is not significantly affected, and again is very similar for both our model and *LCSeg*.

4.4 Experiment 3: Speech recognition

The experiments so far have all used manual word transcriptions. Of course, in real meeting processing systems, we will have to deal with speech recognition (ASR) errors. We therefore also tested on 1-best ASR output provided by ICSI, and results are shown in Table 2. The "off-topic" and "digits" sections were removed in this test, so results are comparable with Experiment 1. Segmentation accuracy seems extremely robust; interestingly, *LCSeg*'s results are less robust (the drop in performance is higher), especially when the number of segments in a meeting is unknown.

Experiment	Model	known		unknown	
		P_k	W_D	P_k	W_D
2	$T = 10$	.296	.342	.325	.366
(off-topic data)	*LCSeg*	.307	.338	.322	.386
3	$T = 10$	.266	.306	.291	.331
(ASR data)	*LCSeg*	.289	.339	.378	.472

Table 2: Results for Experiments 2 & 3: robustness to off-topic and ASR data.

It is surprising to notice that the segmentation accuracy in this experiment was actually slightly higher than achieved in Experiment 1 (especially given that ASR word error rates were generally above 20%). This may simply be a smoothing effect: differences in vocabulary and its distribution can effectively change the prior towards sparsity instantiated in the Dirichlet distributions.

5 Summary and Future Work

We have presented an unsupervised generative model which allows topic segmentation and identification from unlabelled data. Performance on the ICSI corpus of multi-party meetings is comparable with the previous unsupervised segmentation results, and the extracted topics are rated well by human judges. Segmentation accuracy is robust in the face of noise, both in the form of off-topic discussion and speech recognition hypotheses.

Future Work Spoken discourse exhibits several features not derived from the words themselves but which seem intuitively useful for segmentation, e.g. speaker changes, speaker identities and roles, silences, overlaps, prosody and so on. As shown by (Galley et al., 2003), some of these features can be combined with lexical information to improve segmentation performance (although in a supervised manner), and (Maskey and Hirschberg, 2003) show some success in broadcast news segmentation using *only* these kinds of non-lexical features. We are currently investigating the addition of non-lexical features as observed outputs in

our unsupervised generative model.

We are also investigating improvements into the lexical model as presented here, firstly via simple techniques such as word stemming and replacement of named entities by generic class tokens (Barzilay and Lee, 2004); but also via the use of multiple ASR hypotheses by incorporating word confusion networks into our model. We expect that this will allow improved segmentation and identification performance with ASR data.

Acknowledgements

This work was supported by the CALO project (DARPA grant NBCH-D-03-0010). We thank Elizabeth Shriberg and Andreas Stolcke for providing automatic speech recognition data for the ICSI corpus and for their helpful advice; John Niekrasz and Alex Gruenstein for help with the NOMOS corpus annotation tool; and Michel Galley for discussion of his approach and results.

References

Satanjeev Banerjee and Alex Rudnicky. 2004. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of the 8th International Conference on Spoken Language Processing*.

Satanjeev Banerjee, Carolyn Rosé, and Alex 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*.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Proceedings of the Main Conference*, pages 113–120.

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

David Blei and Pedro Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval*, pages 343–348.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Alfred Dielmann and Steve Renals. 2004. Dynamic Bayesian Networks for meeting structuring. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Suffolk.

Thomas Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proc. 32nd Meeting of the Association for Computational Linguistics*, Los Cruces, NM, June.

Thomas Hofmann. 1999. Probablistic latent semantic indexing. In *Proceedings of the 22nd Annual SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.

Toru Imai, Richard Schwartz, Francis Kubala, and Long Nguyen. 1997. Improved topic discrimination of broadcast news using a model of multiple simultaneous topics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 727–730.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 364–367.

Agnes Lisowska, Andrei Popescu-Belis, and Susan Armstrong. 2004. User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Sameer R. Maskey and Julia Hirschberg. 2003. Automatic summarization of broadcast news using structural features. In *Eurospeech 2003*, Geneva, Switzerland.

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

Stehpan Reiter and Gerhard Rigoll. 2004. Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming. In *Proceedings of the International Conference on Pattern Recognition*.

Jeffrey Reynar. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 357–364.