A Comparative Study of Mixture Models for Automatic Topic Segmentation of Multiparty Dialogues

Maria Georgescul
ISSCO/TIM, ETI
University of Geneva
maria.georgescul@eti.unige.ch

Alexander Clark
Department of Computer Science
Royal Holloway University of London
alexc@cs.rhul.ac.uk

Susan Armstrong
ISSCO/TIM, ETI
University of Geneva
susan.armstrong@issco.unige.ch

Abstract

In this article we address the task of automatic text structuring into linear and non-overlapping thematic episodes at a coarse level of granularity. In particular, we deal with topic segmentation on multi-party meeting recording transcripts, which pose specific challenges for topic segmentation models. We present a comparative study of two probabilistic mixture models. Based on lexical features, we use these models in parallel in order to generate a low dimensional input representation for topic segmentation. Our experiments demonstrate that in this manner important information is captured from the data through less features.

1 Introduction

Some of the earliest research related to the problem of text segmentation into thematic episodes used the word distribution as an intrinsic feature of texts (Morris and Hirst, 1991). The studies of (Reynar, 1994; Hearst, 1997; Choi, 2000) continued in this vein. While having quite different emphasis at different levels of detail (basically from the point of view of the employed term weighting and/or the adopted inter-block similarity measure), these studies analyzed the word distribution inside the texts through the instrumentality of merely one feature, i.e. the one-dimensional inter-block similarity.

More recent work use techniques from graph theory (Malioutov and Barzilay, 2006) and machine learning (Galley et al., 2003; Georgescul et al., 2006; Purver et al., 2006) in order to find patterns in vocabulary use.

We investigate new approaches for topic segmentation on corpora containing multi-party dialogues, which currently represents a relatively less explored domain. Compared to other types of audio content (e.g. broadcast news recordings), meeting recordings are less structured, often exhibiting a high degree of participants spontaneity and there may be overlap in finishing one topic while introducing another. Moreover while ending the discussion on a certain topic, there can be numerous new attempts to introduce a new topic before it becomes the focus of the dialogue. Therefore, the task of automatic topic segmentation of meeting recordings is more difficult and requires a more refined analysis. (Galley et al., 2003; Georgescul et al., 2007) dealt with the problem of topic segmentation of multiparty dialogues by combining various features based on cue phrases, syntactic and prosodic information. In this article, our investigation is based on using merely lexical features.

We study mixture models in order to group the words co-occurring in texts into a small number of semantic concepts in an automatic unsupervised way. The intuition behind these models is that a text document has an underlying structure of "latent" topics, which is hidden. In order to reveal these latent topics, the basic assumption made is that words related to a semantic concept tend to occur in the proximity of each other. The notion of proximity between semantically related words can vary for various tasks. For instance, bigrams can be considered to capture correlation between words at a very

short distance. At the other extreme, in the domain of document classification, it is often assumed that the whole document is concerned with one specific topic and in this sense all words in a document are considered to be semantically related. We consider for our application that words occurring in the same thematic episode are semantically related.

In the following, the major issues we will discuss include the formulations of two probabilistic mixture approaches, their methodology, aspects of their implementation and the results obtained when applied in the topic segmentation context. Section 2 presents our approach on using probabilistic mixture models for topic segmentation and shows comparisons between these techniques. In Section 3 we discuss our empirical evaluation of these models for topic segmentation. Finally, some conclusions are drawn in Section 4.

2 Probabilistic Mixture Models

The probabilistic latent models described in the following exploit hierarchical Bayesian frameworks. Based on prior distributions of word rate variability acquired from a training corpus, we will compute a density function to further analyze the text content in order to perform topic segmentation at a coarse level of granularity. In this model, we will be working with 'blocks' of text which consist of a fixed number of consecutive utterances.

In the following two subsections, we use the following notation:

- We consider a text corpus $\mathcal{B} = \{b_1, b_2, ..., b_M\}$ containing M blocks of text with words from a vocabulary $\mathcal{W} = \{w_1, w_2, ..., w_N\}$. M is a constant scalar representing the number of blocks of text. N is a constant scalar representing the number of terms in vocabulary $\mathcal{W}$.

- We pre-process the data by eliminating content free words such as articles, prepositions and auxiliary verbs. Then, we proceed by lemmatizing the remaining words and by adopting a bag-of-words representation. Next, we summarize the data in a matrix $\mathcal{F} = (f(b_i, w_{i,j}))_{(i,j)\in M\times N}$, where $f(b_i, w_{i,j})$ denotes the $log.entropy$ weighted frequency of word $w_{i,j}$ in block b_i.

- Each occurrence of a word in a block of text is considered as representing an observation $(w_{m,n}, b_m)$, i.e. a realization from an underlying sequence of random variables $(W_{m,n}, B_m)_{1\leq n\leq N}^{1\leq m\leq M}$. $w_{m,n}$ denotes the term indicator for the n-th word in the m-th block of text.

- Each pair $(w_{m,n}, b_m)$ is associated with a discrete hidden random variable $Z_{m,n}$ over some finite set $\mathcal{Z} = \{z_1, z_2, ..., z_K\}$. K is a constant scalar representing the number of mixture components to generate.

- We denote by $P(z_{m,n} = z_k)$ or simply by $P(z_k)$ the probability that the k-th topic has been sampled for the n-th word in the m-th block of text.

2.1 Aspect Model for Dyadic Data (AMDD)

In this section we describe how we apply latent modeling for dyadic data (Hofmann, 2001) to text representation for topic segmentation.

2.1.1 Model Setting

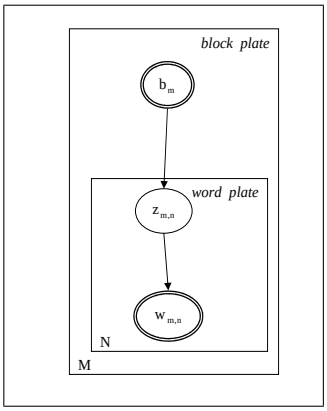

Figure 1: Graphical model representation of the aspect model.

We express the joint or conditional probability of words and blocks of text, by assuming that the choice of a word during the generation of a block of text is independent of the block itself, given some (unobserved) hidden variable, also called *latent* variable or *aspect*.

The graphical representation of the AMDD data generation process is illustrated in Figure 1 by using

the plate notation. That is, the ovals (i.e. the nodes of the graph) represent probabilistic variables. The double ovals around the variables $w_{m,n}$ and b_m denote observed variables. $z_{m,n}$ is the mixture indicator, the hidden variable, that chooses the topic for the n-th word in the m-th block of text. Arrows indicate conditional dependencies between variables. For instance, the $w_{m,n}$ variable in the word space and the b_m variable in the block space have no direct dependencies, i.e. it is assumed that the choice of words in the generation of a block of text is independent of the block given a hidden variable. The boxes represent "plates", i.e. replicates of sampling steps with the variable in the lower left corner referring to the number of samples. For instance, the "word plate" in Figure 1 illustrates N independently and identically distributed repeated trials of the random variable $w_{m,n}$.

According to the topology of the asymmetric AMDD Bayesian network from Figure 1, we can specify the joint distribution of a word $w_{m,n}$, a latent topic z_k and a block of text b_m: $P(w_{m,n}, z_k, b_m) = P(b_m) \cdot P(z_k|b_m) \cdot P(w_{m,n}|z_k)$. The joint distribution of a block of text b_m and a word $w_{m,n}$ is thus:

$$P(b_m, w_{m,n}) = \sum_{k=1}^{K} P(w_{m,n}, z_k, b_m) = P(b_m) \cdot \sum_{k=1}^{K} \underbrace{P(z_k|b_m)}_{mixing\ proportions} \cdot \underbrace{P(w_{m,n}|z_k)}_{mixture\ components} \quad (1)$$

Equation 1 describes a special case of a finite mixture model, i.e. it uses a convex combination of a set of component distributions to model the observed data. That is, each word in a block of text is seen as a sample from a mixture model, where mixture components are multinomials $P(w_{m,n}|z_k)$ and the mixing proportions are $P(z_k|b_m)$.

2.1.2 Inferring and Employing the AMDD Model

The *Expectation-Maximization (EM)* algorithm is the most popular method to estimate the parameters for mixture models to fit a training corpus. The EM algorithm for AMDD is based on iteratively maximizing the log-likelihood function: $\mathcal{L}_{PLSA} = \sum_{m=1}^{M} \sum_{n=1}^{N} f(b_m, w_{m,n}) \cdot logP(w_{m,n}, b_m)$. However, the EM algorithm for AMDD is prone to overfitting since the number of parameters to be estimated grows linearly with the number of blocks of text. In order to avoid this problem, we employed the tempered version of the EM algorithm that has been proposed by Hofmann (2001).

We use the density estimation method in AMDD to reduce the dimension of the blocks-by-words space. Thus, instead of using the words as basic units for each block of text representation, we employ a "topic" basis, assuming that a few topics will capture more information than the entire huge amount of words in the vocabulary. Thus, the m-th block of text is represented by the vector $(P(z_1|b_m), P(z_2|b_m), ..., P(z_k|b_m))$. Then, we use these posterior probabilities as a threshold to identify the boundaries of thematic episodes via support vector classification (Georgescul et al., 2006). That is, we consider the topic segmentation task as a binary-classification problem, where each utterance should be classified as marking the presence or the absence of a topic shift in the dialogue.

2.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (Blei et al., 2003) can be seen as an extension of AMDD by defining a probabilistic mixture model that includes Dirichlet-distributed priors over the masses of the multinomials $P(w|z)$ and $P(z|b)$.

2.2.1 Model Setting

In order to describe the formal setting of LDA in our context, we use the following notation in addition to those given at the beginning of Section 2:

- $\vec{\theta}_m$ is a parameter notation for $P(z|b = b_m)$, the topic mixture proportion for the m-th block of text;
- $\vec{\alpha}$ is a hyperparameter (a vector of dimension K) on the mixing proportions $\vec{\theta}_m$;
- $\Theta = \left\{\vec{\theta}_m\right\}_{m=1}^{M}$ is a matrix (of dimension $M \times K$), composed by placing the vectors $\vec{\theta}_1, \vec{\theta}_2, ..., \vec{\theta}_M$ as column components;
- $\vec{\varphi}_k$ is a parameter notation for $P(w|z_k)$, the mixture component for topic k;
- $\vec{\beta}$ is a hyperparameter (a vector of dimension N) on the mixture components $\vec{\varphi}_k$;

- $\Phi = \{\vec{\varphi}_k\}_{k=1}^{K}$ is a matrix of dimension $K \times N$ composed by placing the vectors $\vec{\varphi}_1, \vec{\varphi}_2, ..., \vec{\varphi}_K$ as column components;
- N_m denotes the length of the m-th block of text and is modeled with a Poisson distribution with constant parameter ξ;

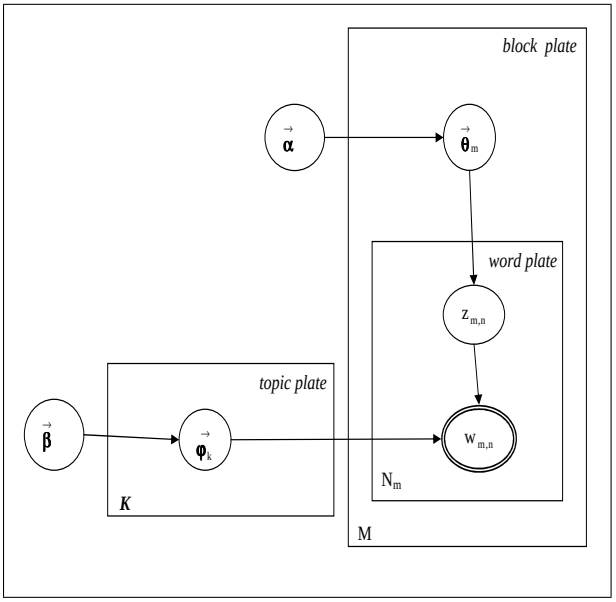

Figure 2: Graphical model representation of latent Dirichlet allocation.

LDA generates a stream of observable words $w_{m,n}$ partitioned into blocks of text $\vec{b}_m$ as shown by the graphical model in Figure 2. The Bayesian network can be interpreted as follows: the variables Φ, θ and z are the three sets of latent variables that we would like to infer. The plate surrounding $\vec{\varphi}_k$ illustrates the repeated sampling of word distributions for each topic z_k until K topics have been generated. The plate surrounding $\vec{\theta}_m$ illustrates the sampling of a distribution over topics for each block b for a total of M blocks of text. The inner plate over $z_{m,n}$ and $w_{m,n}$ illustrates the repeated sampling of topics and words until N_m words have been generated for a block $\vec{b}_m$.

Each block of text is first generated by drawing a topic proportion $\vec{\theta}_m$, i.e. by picking a distribution over topics from a Dirichlet distribution. For each word $w_{m,n}$ from a block of text $\vec{b}_m$, a topic indicator k is sampled for $z_{m,n}$ according to the block-specific mixture proportion $\vec{\theta}_m$. That is, $\vec{\theta}_m$ determines $P(z_{m,n})$. The topic probabilities $\vec{\varphi}_k$ are also sampled from a Dirichlet distribution. The words in each block of text are then generated by using the corresponding topic-specific term distribution $\vec{\varphi}_{z_{m,n}}$.

Given the graphical representation of LDA illustrated in Figure 2, we can write the joint distribution of a word $w_{m,n}$ and a topic z_k as:
$P(w_{m,n}, z_k|\vec{\theta}_m, \Phi) = P(z_k|\vec{\theta}_m) \cdot P(w_{m,n}|\vec{\varphi}_k)$.
Summing over k, we obtain the marginal distribution:
$P(w_{m,n}|\vec{\theta}_m, \Phi) = \sum_{k=1}^{K} \left(\underbrace{P(z_k|\vec{\theta}_m)}_{mixture\ proportion} \cdot \underbrace{P(w_{m,n}|\vec{\varphi}_k)}_{mixture\ component} \right)$.
Hence, similarly to AMDD (see Equation 1), the LDA model assumes that a word $w_{m,n}$ is generated from a random mixture over topics. Topic probabilities are conditioned on the block of text a word belongs to. Moreover LDA leaves flexibility to assign a different topic to every observed word and a different proportion of topics for every block of text.

The joint distribution of a block of text $\vec{b}_m$ and the latent variables of the model $\vec{z}_m$, $\vec{\theta}_m$, Φ, given the hyperparameters $\vec{\alpha}$, $\vec{\beta}$ is further specified by: $P(\vec{b}_m, \vec{z}_m, \vec{\theta}_m, \Phi|\vec{\alpha}, \vec{\beta}) = \overbrace{P(\Phi|\vec{\beta})}^{topic\ plate} \cdot \underbrace{P(\vec{\theta}_m|\vec{\alpha}) \cdot \prod_{n=1}^{N_m} \overbrace{P(z_{m,n}|\vec{\theta}_m) \cdot P(w_{m,n}|\vec{\varphi}_{z_{m,n}})}^{word\ plate}}_{block\ plate}$.

Therefore, the likelihood of a block $\vec{b}_m$ is derived as the marginal distribution obtained by summing over the $z_{m,n}$ and integrating out the distributions $\vec{\theta}_m$ and Φ.

2.2.2 Inferring and Employing the LDA Model

Since the integral involved in computing the likelihood of a block $\vec{b}_m$ is computationally intractable, several methods for approximating this posterior have been proposed, including variational expectation maximization (Blei et al., 2003) and Markov chain Monte Carlo methods (Griffiths and Steyvers, 2004).

We follow an approach based on Gibbs sampling as proposed in (Griffiths and Steyvers, 2004). As the convergence criteria for the Markov chain, we

check how well the parameters cluster semantically related blocks of text in a training corpus and then we use these values as estimates for comparable settings.

The LDA model provides a soft clustering of the blocks of text, by associating them to topics. We exploit this clustering information, by using the distribution of topics over blocks of text to further measure the inter-blocks similarity. As in Section 2.1.2, the last step of our system consists in employing binary support vector classification to identify the boundaries of thematic episodes in the text. That is, we consider as input features for support vector learning the component values of the vector $(\theta_{m,z_1}, \theta_{m,z_2}, ..., \theta_{m,z_k})$.

3 Experiments

In order to evaluate the performance of AMDD and LDA for our task of topic segmentation, in our experiments we used the transcripts of ICSI-MR corpus (Janin et al., 2004), which consists of 75 meeting recordings. A subset of 25 meetings, which are transcribed by humans and annotated with thematic boundaries (Galley et al., 2003), has been kept for testing purposes and support vector machine training. The transcripts of the remaining 50 meetings have been used for the unsupervised inference of our latent models. The fitting phase of the mixture models rely on the same data set that have been pre-processed by tokenization, elimination of stopwords and lemmatization.

Once the models' parameters are learned, the input data representation is projected into the lower dimension latent semantic space. The evaluation phase consists in checking the performance of each model for predicting thematic boundaries. That is, we check the performance of the models for predicting thematic boundaries on the same test set. The size of a block of text during the testing phase has been set to one, i.e. each utterance has been considered as a block of text.

Figure 3 compares the performance obtained for various k values, i.e. various dimensions of the latent semantic space, or equivalently different numbers of latent topics. We have chosen k=$\{50, ...400\}$ using incremental steps of 50.

The performance of each latent model is measured by the accuracy $Acc = 1 - P_k$, where P_k denotes the error measure proposed by (Beeferman et al., 1999). Note that the P_k error allows for a slight variation in where the hypothesized thematic boundaries are placed. That is, wrong hypothesized thematic boundaries occurring in the proximity of a reference boundary (i.e. in a fixed-size interval of text) are tolerated. As proposed by (Beeferman et al., 1999), we set up the size of this interval to half of the average number of words per segment in the gold standard segmentation.

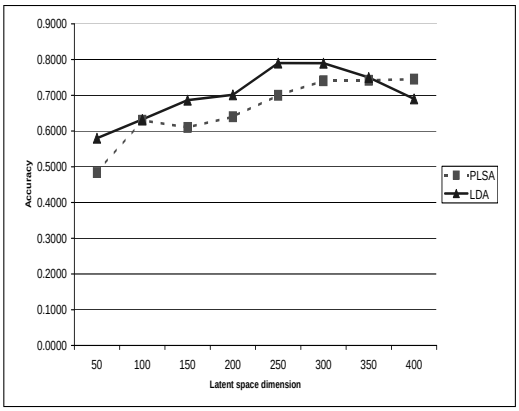

Figure 3: Results of applying the mixture models for topic segmentation.

As we observe from Figure 3, LDA and AMDD achieved rather comparable thematic segmentation accuracy. While LDA steadily outperformed AMDD, the results do not show a notable advantage of LDA over AMDD. In contrast, AMDD has better performances for less dimensionality reduction. That is, the LDA performance curve goes down when the number of latent topics exceeds over 300.

	LDA	*LCSeg*	*SVMs*
P_k error rate	21%	32 %	22%

Table 1: Comparative performance results.

In Table 1, we provide the best results obtained on ICSI data via LDA modeling. We also reproduce the results reported on in the literature by (Galley et al., 2003) and (Georgescul et al., 2006), when the evaluation of their systems was also done on ICSI data. The *LCSeg* system proposed by (Galley et al., 2003) is based on exploiting merely lexical features. Improved performance results have

been obtained by (Galley et al., 2003) when extra non-lexical features have been adopted in a decision tree classifier. The system proposed by (Georgescul et al., 2006) is based on support vector machines (SVMs) and is labeled in the table as *SVMs*. We observe from the table that our approach based on combining LDA modeling with SVM classification outperforms *LCSeg* and performs comparably to the system of Georgescul et al. (2006). Thus, our experiments show that the LDA word density estimation approach does capture important information from the data through 90% less features than a bag-of-words representation.

4 Conclusions

With the goal of performing linear topic segmentation by exploiting word distributions in the input text, the focus of this article was on both comparing theoretical aspects and experimental results of two probabilistic mixture models. The algorithms are applied to a meeting transcription data set and are found to provide an appropriate method for reducing the size of the data representation, by performing comparably to previous state-of-the-art methods for topic segmentation.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34:177–210. Special Issue on Natural Language Learning.

David M. Blei, Andrew Y. Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pages 993–1022.

Freddy Choi. 2000. Advances in Domain Independent Linear Text Segmentation. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, USA.

Michael Galley, Kathleen McKeown, Eric Fosler-Luissier, and Hongyan Jing. 2003. Discourse Segmentation of Multi-Party Conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 562–569, Sapporo, Japan.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2006. Word Distributions for Thematic Segmentation in a Support Vector Machine Approach. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, pages 101–108, New York City, USA.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2007. Exploiting Structural Meeting-Specific Features for Topic Segmentation. In *Actes de la 14ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 15–24, Toulouse, France.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5228–5235.

Marti Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Macias-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. The ICSI Meeting Project: Resources and Research. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Meeting Recognition Workshop*, Montreal, Quebec, Canada.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 25–32, Sydney, Australia.

Jane Morris and Graeme Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

Matthew Purver, Konrad P. Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised Topic Modelling for Multi-Party Spoken Discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 17–24, Sydney, Australia.

Jeffrey Reynar. 1994. An Automatic Method of Finding Topic Boundaries. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 331–333, Las Cruces, New Mexico, USA.