Detecting speaker roles and topic changes in multiparty conversations using latent topic models

Ashtosh Sapru[1,2] *and Hervé Bourlard*[1,2]

[1]Idiap Research Institute, 1920, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
ashtosh.sapru@idiap.ch,herve.bourlard@idiap.ch

Abstract

Accessing and browsing archives of multiparty conversations can be significantly facilitated by labeling them in terms of high level information. In this paper, we investigate automatic labeling of speaker roles and topic changes in professional meetings. Using the framework of unsupervised topic modeling we express speaker utterances as mixture of latent variables, each of which is governed by a multinomial distribution. The generated latent topic distributions are then used as features for predicting role and topic changes. Experiments performed on several hours of meeting data selected from AMI corpus reveal that latent topic features are effective predictors of speaker roles and topic changes. Furthermore, experiments also reveal an improvement in performance when latent topic information is combined with other multistream features.
Index Terms: speaker role labeling, latent topic models, topic boundary detection

1. Introduction

Automatic processing of spoken audio documents is a research domain with important applications in data indexing, summarization and information retrieval. Common approaches for automatically structuring and accessing spoken documents are typically based on Automatic Speech Recognition (ASR) techniques, keyword spotting and speaker segmentation. Building on those techniques that extract the verbal content or the speaker identity, recent research has aimed at extracting high level information from spoken audio, such as speaker roles and topic segmentation. Roles and topics can enhance audio browsing by grouping speakers based on their characteristic behavior and identifying topically coherent segments in multiparty discourse.

Speaker role recognition and topic change detection has been widely studied in the case of Broadcast News (BN) recordings [1, 2]. Typical roles considered in BN data are formal roles, i.e., roles imposed from the news format and related to the task each speaker performs in the show, such as anchorman, journalists and interviewees. Statistical classifiers model various cue phrases used by anchors, speaker introductions and compact structure of speaker interactions in news programs to recognize roles [1, 3]. Earlier methods for BN story segmentation have exploited the fact that news segments tend to be lexically cohesive [4]. Other methods have used classifiers like decision trees to identify topic boundaries by combining lexical and acoustic indicators like long pauses, variations in fundamental frequency and speaking rate [5].

In recent years, large collection of spontaneous multiparty conversations, e.g., meetings, have been recorded and the problem of role recognition and topic change detection has been investigated in these settings. Contrary to BN, conversations do not have precise and well defined topic boundaries, speakers frequently take turn in the discussions (speaker turns are short), they overlap each other and speech is often disfluent. Furthermore, speaker segmentation and ASR systems used for extracting relevant features produce significantly higher errors due to the previously mentioned phenomena making automatic labeling of meetings a challenging task.

Even with these limitations, automatic labeling of multiparty conversations in terms of role and topics have been studied on several meeting recordings like the CMU corpus [6], ICSI corpus [7, 8] and AMI corpus [9, 10]. Most of these studies follow a supervised approach where statistical classifiers are trained on a combination of lexical, structural and prosodic features [11]. In [6], authors investigated role recognition and meeting segmentation using simple speech activity based features. However, this study considered meeting segmentation in terms of discourse states like presentation, discussion, etc., rather than topic segments. In [12], a lexical cohesion algorithm was presented that uses cosine similarity metric to identify topic segments in meetings. The algorithm can choose the number of topic boundaries by itself, or perform topic segmentation with known number of boundaries. A HMM based approach was used to find topic boundaries in ICSI meetings [8]. The background topics were modeled as hidden states of the HMM and state emissions were trained using a n-gram language model. More recent studies [13, 11] have also shown that speaker roles along with vocalization events like pauses and overlaps can be informative for detecting topic changes.

In this study, we investigate the problem of identifying speaker roles and topic changes in AMI corpus meetings. We follow an unsupervised approach for extracting relevant features from unlabeled data using the framework of latent topic models. Previous studies [14, 15] have applied latent topic models for discourse segmentation, but this is first work to the best of our knowledge that applies latent topic modeling to recognize formal roles in professional meetings. The probabilistic topic models are trained on speech utterances and we use the estimated latent topic distributions to infer speaker roles and topic changes. In comparison to [9], which is based on lexical choice and social network analysis, we show that a much smaller set of representative features can be used to achieve better performance for recognizing speaker roles. In the remainder of this paper, Section 2 describes the corpus and the role annotation, Section 3 and Section 4 provide details of unsupervised feature extraction and classification methodology, respectively. The experiments and results are presented in Section 5. The paper is then concluded in Section 6.

2. Meeting Corpus

The AMI Meeting Corpus [16] is a collection of meetings captured in specially instrumented meeting rooms, which record the audio and video for each meeting participant. The corpus contains both scenario and natural meetings. The corpus is manually transcribed at different levels (roles, speaking time, words, dialog act and topics). For the purpose of this study, we consider scenario meetings which are annotated in terms of both role and topic labels. All the scenario meetings have four participants that play the role of a design team composed of Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial Designer (ID), who are tasked with designing a new remote control. These roles are fixed for a participant for the entire duration of the meeting. Each meeting follows an agenda which is supervised by the Project Manager and all participants are given a task that they have to accomplish.

Annotators were given a standard set of topic descriptions that were used as labels for identifying topic segments. Three categories of topic description were defined: top level topics reflect the meeting agenda (e.g., presentation, discussion), sub topics describe parts of top level topics and functional topics (e.g., opening, closing). The scenario meetings have on an average eight top level topic segments and three more sub topic segments. The intercoder agreement was found to be 0.66 for top level topics and 0.59 for sub topic level [17].

3. Unsupervised feature extraction

Topic models are probabilistic generative models that have been extensively used for natural language processing. In Latent Dirichlet allocation (LDA) [18], the corpus is generated from a fixed underlying mixture of K topics, and each topic is modeled as a multinomial distribution over all the possible words. The latent topics discover patterns based on the co-occurrence of words in the documents.

Let D be the set of documents in the corpus and V be the set of unique words. Each document d is represented by a bag of N_d words chosen from V. We also assume a fixed number of topics K for the entire corpus. In LDA, the word distribution for a given document is represented as a mixture of K topics $P(w) = \sum_{k=1}^{K} P(w|z_w = k)P(z_w = k)$, where z_w is the latent variable from which the word w is drawn. The distribution of words conditioned on z_w is given by a multinomial $P(w|z_w = k) = \phi_{z_w}^{(k)}$ and the latent variable z_w for a document d is also sampled from a multinomial distribution $p(z_w = k) = \theta_k^{(d)}$.

In this study, all the speech transcripts were generated using output of AMI-ASR system [19] which has a word error rate of less than 30%. For each participant in a given meeting, we extracted the spoken words and their timing information. Speaker roles are fixed for the participant during the meeting, while topic boundaries mark changes in content of meeting discourse. Given the difference in the nature of detection tasks, the speech transcripts were processed at different scales. For the task of speaker role recognition, each document d in the training set is represented by all the words uttered by a single speaker S during a meeting. Every participant in a meeting has to perform a separate function defined by its role. We hypothesize that the function of a role influences the distribution of latent topics used by the speaker to generate the spoken content. We extract the underlying latent topic distribution $\Theta(S) = \{\theta_k^S\}$ by applying LDA on the speaker level documents in the corpus.

For the the task of topic change detection, we process transcripts at the level of speech utterances. Following [17], we consider talk spurts, i.e., regions of continuous speech activity separated by pauses no longer than 0.5 seconds, as the unit of discourse segmentation. Using the speech activity information of all the meeting participants, the entire meeting is segmented into a sequence of talk spurts. We observed that some talk spurts correspond to back channels and others have only few word tokens. Most of these were removed during processing. Furthermore, we also applied a window that concatenates surrounding talk spurts. The documents in the corpus were represented as a bag of words corresponding to a given talk spurt. The training set for LDA modeling is formed by concatenating the talk spurt level documents for all the meetings.

4. Classification Approach

Both automatic speaker role recognition and topic change detection are examples of machine learning problems that consist of finding a stochastic mapping from a set of features to a set of class labels. For the task of role recognition, the classes correspond to roles in the set $\mathcal{R} = \{PM, ME, UI, ID\}$. For topic change detection, each potential talk spurt corresponds to labels in the topic boundary set $\mathcal{B} = \{0, 1\}$.

The feature vector for automatic role recognition is the latent topic distribution Θ for each speaker in the meeting. Given $X(S) = \Theta(S)$, our approach is to find a speaker to role mapping $\phi(S) \mapsto R$, such that the following equation is satisfied:

$$\tilde{R} = \arg\max_{R \in \mathcal{R}} P(\phi(S) = R|X(S)) \quad (1)$$

where $\tilde{R}$ is a role in the set $\mathcal{R}$.

For the task of topic change detection each talk spurt instance ts_i is a possible location for topic change. The unsupervised features are given by the latent topic distribution $\Theta(ts_i)$ corresponding to this instance. However, we use the prior knowledge that neighboring talk spurts ts_i and ts_j, when they belong to the same topic should have similar values for $\Theta(ts_i)$ and $\Theta(ts_j)$, while $\Theta(ts_i)$ and $\Theta(ts_j)$ values for talk spurts on either side of the boundary should be different. We incorporated this information by putting a window of length L around the current instance and concatenating all the features within the window to represent the feature vector $X(ts_i) = [\Theta(ts_{i-\frac{L}{2}})...\Theta(ts_i)...\Theta(ts_{i+\frac{L}{2}})]$.

Given feature vector $X(ts_i)$, the mapping $\psi(ts_i) \mapsto b$ is given as,

$$\tilde{b} = \arg\max_{b \in \mathcal{B}} P(\psi(ts_i) = b|X(ts_i)) \quad (2)$$

where $\tilde{b}$ belongs to the set $\mathcal{B}$.

We employed a support vector machine (SVM) classifier to estimate the mapping functions ϕ and ψ from the training data. SVM considers each feature vector as a point in multidimensional feature space and the algorithm works by constructing a separating hyperplane between two classes. For the multiclass classification required for role recognition a one on one strategy was used and each binary classifier was trained using libsvm [20]. A sigmoid scaling function was used to convert the SVM decision scores into posterior probabilities.

5. Experiments

For evaluation of proposed methods, experiments were conducted using repeated crossvalidation (CV). One set of meetings (all but four meetings that have the same set of speaker

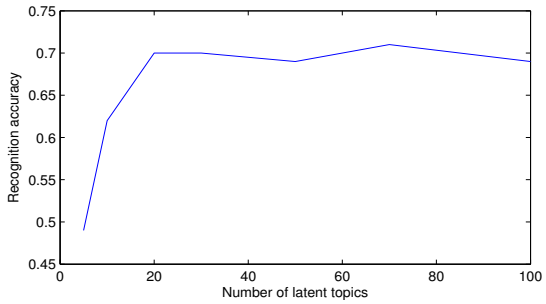

Figure 1: Variation of role recognition accuracy as the number of latent topics K in LDA is varied.

identities) was kept for training/tuning the model parameters, while a distinct set (remaining four meetings) was used for evaluation. We used the rbf kernel for SVM classifier and the model parameters were selected from a subset of training data.

The posterior distribution over documents and topics as well as the hyperparameters of the LDA model were estimated using Gibbs sampling (a Markov chain Monte Carlo method). After a burn in period, the sampling procedure ultimately results in a stationary distribution which corresponds to the topic distribution. For our experiments, we used the mallet implementation [21] of LDA with symmetric Dirichlet priors and hyper parameters initialized to small values (< 1).

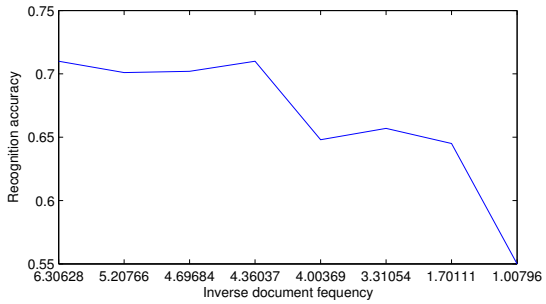

Figure 2: Effect of stop words on role recognition accuracy. Horizontal axis shows different values of IDF.

For evaluation of automatic role recognition, since each CV fold has the same distribution of classes (their being one to one mapping between speakers and roles for each meeting), we use the recognition accuracy as the metric of recognition performance. Our first experiment evaluates the influence of number of latent topics K on the extracted unsupervised features. Figure 1 shows the variation in accuracy for different choices of $K = \{5, 10, 20, 50, 70, 100\}$. The models with fewer number of latent topics are not able to capture all the role information. However, we observe a significant increase in performance with only $K = 20$ topics. Increasing the size of feature set after this does not reveal any significant increase in performance.

We also considered whether removing stop words during training and evaluation can yield unsupervised features that better capture the functional content of roles. A list of stop words was prepared based on their inverse document frequency (IDF) scores. We removed the words with low IDF scores and trained the LDA models on the processed documents. The influence of stop words on the role recognition accuracy is shown in Figure 2. The plot reveals that removing stop words (for moderate IDF scores) does not significantly increase the performance

Table 1: Per role accuracy obtained in recognizing roles for various classification models

	Overall and per-role Accuracy				
Model	PM	ME	UI	ID	All
LDA	0.84	0.72	0.64	0.55	0.69
lex+struct [10]	0.88	0.72	0.62	0.57	0.70
LDA+lex+struct	0.90	0.74	0.69	0.59	0.73
SNA+lex [9]	0.84	0.70	0.38	0.50	0.68

of the models. However, performance drops as IDF scores increase, showing that there exists a limiting size of vocabulary that is needed to express the functions of roles in the corpus.

In Table 1, we report the accuracy of individual roles and compare the performance of unsupervised feature extraction with comparable approaches in literature [9, 10]. In terms of performance, we observe that classifier trained using only few unsupervised features (K=20) performs almost as well as completely supervised models trained using full set of lexical and structural features (> 3000) [10]. Furthermore, we also noticed that when the two feature sets are combined the performance increases, revealing that LDA modeling captures some complementary information in data. We can also observe that performance across roles is not uniform. Compared to other roles, the role of PM is recognized much better by all the models.

Table 2 shows the latent topics which are most correlated with various speaker roles. Since the latent topics correspond to a multinomial distribution over the spoken words, we represent them using "top words" that have the highest frequency for that topic. The analysis of LDA topics and different roles suggests that top words in these topics capture role functions. This shows that unsupervised feature selection using LDA is effective in automatically clustering word patterns for different roles.

An analysis of errors by the classifier revealed that there is a systematic difference in performance that is related to the length of speakers utterances. We observed that average number of words spoken by PM in most meetings was greater than other participants. On the other hand, the design related roles UI and ID tend to speak less number of words when averaged across the meetings. In Figure 3, we plot the observed errors as a function of number of spoken utterances by different speakers. The plot reveals a clear pattern with error decreasing as the length of conversation increases.

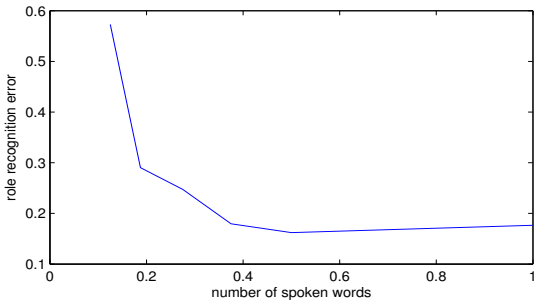

Figure 3: Analysis of role recognition errors with respect to number of spoken words in a document.

The second set of experiments were performed for the task of topic change detection. Similar to task of automatic role recognition we evaluated the performance of models using cross validation. While AMI corpus is annotated hierarchically in terms of topic and sub topic information, we consider a flat-

Table 2: Top words in latent topics that are most correlated with role labels.

PM	ME	UI	ID
okay meeting design yes will minutes project not what going all your would work uhhuh new but yeah one two three	like spongy fruit remote control important fancy shape banana fashion look feel easy trends maybe innovative remote and people five they	buttons then but channel yellow need use see recognition rubber just functions all use shape television one should yes colour volume channel	chip components which infrared titanium signal energy will button design source battery interface circuit working basically power rubber scroll button curved plastic

tened segmentation structure for our experiments. The classification model performance was first evaluated against the random baseline. The baseline scores were obtained by taking the talkspurt sequence and randomly marking them as boundary candidates. However, to make effective comparison the boundary marks were made proportional to average number of topic changes in the training set. Evaluation scores were obtained by 100 iterations of this procedure during testing.

In comparison to role recognition experiments we do report the performance in terms of accuracy, since the classes are highly unbalanced. Also traditional metrics like F-measure, recall and precision do not differentiate near misses in placement of boundary candidates from false positives. Following [11], we use segmentation metrics P_k and WindowDiff to report our results. P_k calculates the probability of segmentation error [22], such that two talkspurts drawn randomly from the data are incorrectly identified as belonging to same topic segment. WindowDiff (WD) [23] calculates the error rate by moving a sliding window across the transcript and counting the number of times the hypothesized and reference segment boundaries are different.

Initial experiments revealed that the classification model tended to predict hypothesized topic boundaries in clusters due to overlap between features across neighboring talk spurts. To compensate for this effect we included a further processing stage where we retained a single hypothesized candidate from the cluster with the highest posterior probability. We experimented with different values of $K = \{6, 12, 24, 32, 40\}$ during LDA training. Our experiments revealed that for higher values of K the model tended to over predict the number of hypothesized boundary candidates. In Figure 4, we report the performance of the classifier as a function of size of feature window with K fixed to 12. The shape of the plot reveals that adding some surrounding context to the talk spurt features improves performance when measured using P_k score. However, for large value of the window size the classifier seems to overfit as the P_k score again increase.

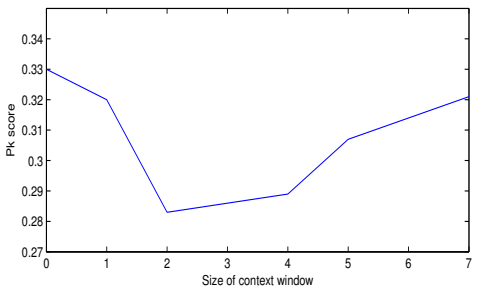

Figure 4: Variation in topic segmentation performance as a function of window size.

Table 3 shows the performance of random baseline and proposed models for topic segmentation. Table numbers reveal that

Table 3: Topic change detection performance for baseline and proposed models.

Model	P_k	WD
Baseline	0.46	0.55
LDA	0.29	0.37

LDA modeling results in a statistically significant increase in performance compared to baseline (paired t-tests, $p < 0.01$). We also compared the performance of LDA extracted features with other methods which use a completely supervised approach for modeling a combination of lexical, prosodic, structural and motion features [11]. Since the reported results in [11] are at top topic and sub topic level, we retrained the classifier for the same. Table 4 reveals that LDA modeling is better than other approaches when P_k scores are compared and reaches a comparable performance to the completely supervised approach for WD scores. We further observed that errors increase when evaluation is performed at sub topic level. However, the proposed method perform much better than LCSeg approach at both top topic and sub topic levels.

Table 4: Comparion with other methods evaluated on AMI corpus for ASR generated transcripts.

Model	Top topic		Sub topic	
	P_k	WD	P_k	WD
LCSeg [11]	0.45	0.58	0.40	0.47
Maxent [11]	0.31	0.34	0.34	0.37
LDA	0.28	0.34	0.29	0.37

6. Conclusions

In this work, we demonstrated the effectiveness of unsupervised feature modeling for detecting speaker roles and topic changes in AMI meetings. By applying LDA on speech transcripts we extracted unsupervised features from unlabeled data that were found to be informative for the two classification tasks. Our experiments on topic segmentation show that LDA features perform better than LCSeg and reached performance comparable to a completely supervised approach. The proposed method was able to perform non trivial classification of four speaker roles, reaching a recognition accuracy of 69%. Moreover, the accuracy increases to 73% when the proposed features where combined with other multistream features.

7. Acknowledgement

This work was funded by the Hasler Stiftung under SESAME grant, the EU NoE SSPNet, and the Swiss National Science Foundation NCCR IM2.

8. References

[1] Barzilay R., Collins M., Hirschberg J., and Whittaker S., "The rules behind roles: Identifying speaker role in radio broadcasts," *Proceedings of AAAI*, 2000.

[2] "Topic detection and tracking evaluation," `http://www.nist.gov/speech/tests/tdt/`.

[3] Salamin H. et al., "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, November 2009.

[4] Marti A. Hearst, "Multi-paragraph segmentation of expository text," in *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*. 1994, pp. 9–16, Association for Computational Linguistics.

[5] Gökhan Tür, Andreas Stolcke, Dilek Hakkani-Tür, and Elizabeth Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comput. Linguist.*, pp. 31–57, 2001.

[6] Banerjee S. and Rudnick A., "Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants," *Proceedings of ICSLP*, 2004.

[7] Laskowski K., Ostendorf M., and Schultz T., "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008.

[8] Melissa Sherman and Yang Liu, "Using hidden markov models for topic segmentation of meeting transcripts.," in *SLT*. 2008, pp. 185–188, IEEE.

[9] Garg N., Favre S., Hakkani-Tur D., and Vinciarelli A., "Role recognition for meeting participants: an approach based on lexical information and social network analysis," *Proceedings of the ACM Multimedia*, 2008.

[10] Sapru A. and Valente F., "Automatic speaker role labeling in AMI meetings: recognition of formal and social roles," *Proceedings of Icassp*, 2012.

[11] Pei yun Hsueh and Johanna D. Moore, "Combining multiple knowledge sources for dialogue segmentation in multimedia archives.," in *ACL*, John A. Carroll, Antal van den Bosch, and Annie Zaenen, Eds. 2007, The Association for Computational Linguistics.

[12] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. 2003, pp. 562–569, Association for Computational Linguistics.

[13] Saturnino Luz and Jing Su, "The relevance of timing, pauses and overlaps in dialogues: detecting topic changes in scenario based meetings.," in *INTERSPEECH*. 2010, pp. 1369–1372, ISCA.

[14] Mike Dowman, Virginia Savova, Thomas L. Griffiths, Konrad P. Krding, Joshua B. Tenenbaum, and Matthew Purver, "A probabilistic model of meetings that combines words and discourse features.," *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[15] Matthew Purver, Konrad P. Krding, and Thomas L. Griffiths, "Unsupervised topic modelling for multi-party spoken discourse," in *In COLING-ACL 2006*, 2006, pp. 17–24.

[16] Carletta J., "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[17] Pei yun Hsueh and Johanna D. Moore, "Automatic topic segmentation and labeling in multiparty dialogue.," in *SLT*, 2006, pp. 98–101.

[18] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 2003.

[19] Hain T., Wan V., Burget L., Karafiat M., J. Dines, Vepa J., Garau G., and Lincoln M., "The AMI System for the Transcription of Speech in Meetings.," *Proceedings of Icassp*, 2007.

[20] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[21] Andrew Kachites. McCallum, "Mallet: A machine learning for language toolkit.," *http://mallet.cs.umass.edu*, 2002.

[22] Doug Beeferman, Adam Berger, and John Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1-3, pp. 177–210, Feb. 1999.

[23] Lev Pevzner and Marti A. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Comput. Linguist.*, vol. 28, no. 1, pp. 19–36, Mar. 2002.