UNSUPERVISED TOPIC MODELING FOR LEADER DETECTION IN SPOKEN DISCOURSE

Raia Hadsell Zsolt Kira Wen Wang Kristin Precoda

SRI International, Princeton, NJ and Menlo Park, CA, USA

{raia.hadsell,zsolt.kira}@sri.com, {wwang,precoda}@speech.sri.com

ABSTRACT

In this paper, we describe a method for leader detection in multi-party spoken discourse that relies on unsupervised topic modeling to segment the discourse automatically. Latent Dirichlet allocation is applied to sliding temporal windows of utterances, resulting in a topic model which captures the fluid transitions from topic to topic which occur in multi-party discourse. Further processing discretizes the continuous topic mixtures into sequential topic segments. Features are extracted from topic shift regions and used to train a binary role classifier. The added topic shift features significantly improve the baseline performance on two corpora, demonstrating both the value of the features and the robustness of the unsupervised segmentation. Furthermore, our classification results on the ICSI corpus, using automatically segmented topics, are better than the results using ground truth segmentations.

Index Terms— Social role classification, leader detection, speaker role labeling, latent Dirichlet allocation, topic shift analysis

1. INTRODUCTION

Social conversation is a deep and broad information source that may yield rich understanding of a group and its constituents when examined by trained analysts. Automatic analysis of such data is much more difficult, however, especially multi-party discourse that does not follow a well-defined structure. Many researchers have considered the problem of social role labeling in broadcast conversation and have developed models that rely on lexical cues, structural features, dialog markers, interaction patterns, and social network analysis [1, 2, 3]. We build on this research, but focus on automatic topic segmentation as a source of features for improving leader detection. Rather than broadcast conversation, we evaluate our approach on meeting transcript, specifically the ICSI and Nuclear Waste Technical Review Board (NWTRB) corpora.

Topic shift can be observed in any social discourse that is of sufficient length. In a broadcast show, the dominant topic might shift from pleasantries, to sound bites, to focused discussion with a guest. In a workplace meeting, the dominant topics might include administrative issues, focused agenda items, and presentations by individuals. In a social discourse setting, topics may shift continuously and quickly. Our approach seeks to capture topic shift in any of these domains by training a probabilistic topic model from unlabeled data. After training, the model gives a topic distribution for each utterance. Smoothing and aggregation are then applied to the topic mixtures to find discrete topic segments as well as identify mixed topic or transitional segments. Compared with the hand-labeled ground truth segmentations for the ICSI corpus, our purely unsupervised and unguided segmentations capture the essential flow and shift of the topics over time.

Our motivation for topic segmentation is to better understand and detect leaders. We hypothesize that leaders, hosts, and moderators are more likely to guide a meeting or dialog by beginning and ending discussions, and therefore are more likely to speak during topic transitions and at the onset of new topic segments. Automatically segmenting the transcript allows us to extract features from these inter-topic areas. The features can then be used to train a classifier, which can give statistical confirmation to our hypothesis of a correlation between leader role and topic shift. We find that segmentation-related features improve our f-score on ICSI meeting leader detection from 0.673 to 0.726, and improve our f-score on NWTRB meeting leader detection from 0.751 to 0.807. Extracting the same features from ground truth segmentations gives a lower f-score, possibly due to a more aggressive segmentation, giving further support to our unsupervised approach.

Other research has shown the effectiveness of topic modeling for segmentation in spoken discourse [4, 5], but has not applied the segmentation for leader detection. Many supervised and unsupervised methods for speaker role labeling have been proposed [6, 1, 2, 3], but none has used automatic topic segmentation.

In the remainder of this paper, we discuss the system overview and our procedures for topic modeling and segmentation (Section 3), topic-related feature extraction (Section 4), and classifier training and evaluation (Section 5). Results are given in Section 6.

2. SYSTEM OVERVIEW

Figure 1 shows the overview of the learning system. In the unsupervised learning phase, a data corpus is input to the topic modeling algorithm, producing a probabilistic distribution over a set of topics for each utterance. Through margin-based aggregation, the topic distributions are converted into discrete segments, each dominated by a single topic. Features are calculated from the topic shift areas and added to the baseline system feature vector. Given these feature representations, a boosting classifier is trained to distinguish each speaker as a leader or non-leader. The baseline system without the topic features is described in detail in [7]. This paper focuses specifically on topic modeling and the extraction of features from the resulting topic segmentation.

3. TOPIC SEGMENTATION

Latent Dirichlet allocation (LDA) is a generative, probabilistic method for unsupervised topic modeling which has been extensively used for natural language research and analysis in recent years [8]. Applied to a set of documents, the LDA model assumes that the words in each document are drawn from an underlying mixture of fixed topics, where each topic is a multinomial distribution over all possible words.

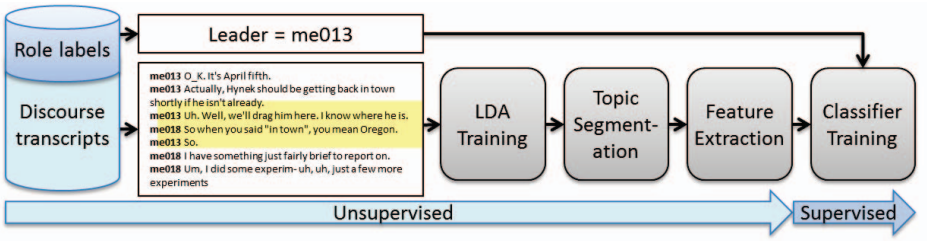

Fig. 1. Overview of the learning system. Transcript utterances are input to the LDA training system, producing topic segments. For each speaker, features from all utterances are combined with topic shift features, and the resulting feature vector is input to the classifier. During training, the leader label is used for supervision.

Given a conversation transcript composed of utterances from multiple participants, we attempt to identify discrete segments in which a single topic dominates. Since informal conversation is fluid, we expect that there will often be transitions between topics where no one topic is dominant. Therefore, instead of imposing a complete segmentation on the utterances as in other research [4], we first apply topic modeling to obtain topic mixtures for each utterance. Then, we aggregate dominant topics over temporal windows to build consistent topic segments.

Similarly to research by Purver et al. ([4]), we consider utterances as separate topic mixtures. However, we find that the topic model is more consistent if each utterance is contextualized by a window of utterances before and after. Thus, the training set $\mathcal{S}$ is a series of concatenated discourse transcripts which consist of N sequential utterances where U_i is the i^{th} utterance, containing p_i tokens. The context window k defines the total number of sequential utterances which are concatenated to form a single data instance, which is denoted D_i. In addition, utterances with fewer than p_{min} tokens were removed from the set of training instances. The number of topics N_t is also fixed. In practice, the approach was evaluated with $k = \{3, 5\}$, $p_{min} = \{5, 8\}$, and $N_t = \{5, 8, 12, 20\}$.

After training the LDA model, each data instance D_i has a topic distribution $\phi^{(i)}$, with the probability of topic t being $\phi_t^{(i)}$. This is smoothed temporally by applying a Gaussian kernel ($\sigma = .3$) over each topic. Next, topic segments are isolated by applying margin and segment length criteria. The margin criterion is a threshold on the minimum difference between the dominant topic and the next highest ranked topic. Data instances that do not meet these criteria are marked as transitions.

An example of topic modeling and topic segmentation is given in Figure 2. Changing topic distributions are shown for a full ICSI meeting, with the automatic segmentation indicated using black line segments. The hand-labeled topic segmentation (from [9]) is shown above the automatic segmentation.

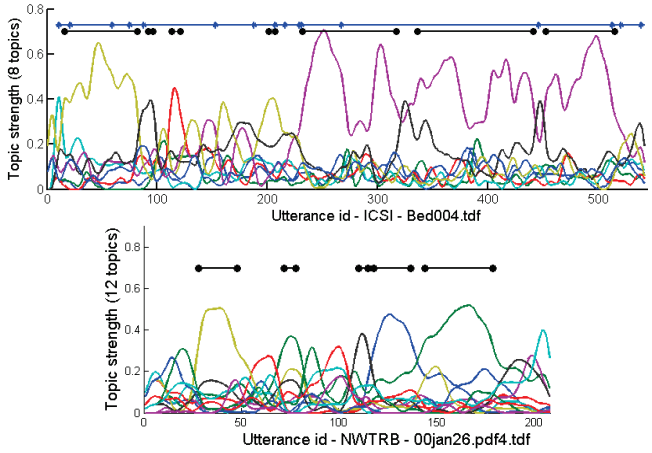

Fig. 2. Topic modeling and topic segmentation is shown for a full ICSI meeting (top), and an excerpt from a NWTRB meeting (bottom), with the automatic segmentation indicated using black line segments. The hand-labeled topic segments are shown in blue for the ICSI meeting, above the automatic segmentation.

4. FEATURE EXTRACTION

The underlying premise of this work is that correlations to predict leaders exist around topic shifts. To find these correlations, count-based features are accumulated for each speaker over utterances surrounding topic breaks.

For each topic shift, a window is designated to calculate the statistics. This window is defined both in terms of a minimum number of utterances around the shift as well as a temporal window around the shift. Both methods are used since topic shifts tend to occur during utterances that vary widely in length, and hence a temporal window alone does not suffice. Note that since the output of unsupervised topic segmentation results in transition regions, the window is centered at the onset of a new segment rather than during the transition.

Eighteen features, listed in Table 4, were extracted for each speaker. The first two consist of simple counts and times for all utterances that were included in the window around topic shifts. The subsequent twelve features consist of variations of counts and times for four different types of events that can occur around the topic shift. For example, the first count is the number of times that the speaker actually caused the shift; i.e. the utterance for the speaker was the one that began a new topic. The next three measure when the utterance followed a previous topic shift (i.e. the person spoke right after the previous shift), the current shift, and the next shift, respectively. These features capture the different ways in which the speaker was involved in the shift. Features 3–6 represents counts normalized by the total number of utterances the speaker had around topic shifts. Features 7–10 are the same counts, but normalized by the number of total topic shifts that occurred in the excerpt. Finally, features 11–14 represent the amount of time the speaker was involved in the various events.

Feature	Description
1–2	Number sentences and time spoken
3–6	Number of times the speaker caused a shift, or was around previous, current, or next shift, respectively (normalized by number of utterances)
7–10	Number times the speaker caused a shift, or was around previous, current, or next shift, respectively (normalized by number of topic shifts)
11–14	Total utterance time of the speaker when he/she caused a shift, or was around previous, current, or next shift (normalized by total utterance time)
15	Average number of interruptions per turn which was produced by this speaker
16	Count of this speaker's utterances that are interrupted by someone else divided by count of all speakers' utterances that are interrupted by someone
17	Median length of an overlap by this speaker within a turn by another speaker (or 0 if no such overlaps). Calculated from the timings output by the speech recognizer.
18	Percent of this speaker's utterances that contain an utterance initiating a (dis)agreement opportunity (labeled as an initiating utterance by the (dis)agreement annotations

Table 1. Features extracted around topic shifts that augmented the baseline set of features used by the classifier.

The last four features were found in [7] to be discriminative in general, so we included them to be measured specifically around topic shifts. Three measure various aspects of interruption, with the idea that interruptions caused during topic shifts may be exhibited by leaders. The final feature measures the initiation of disagreements around topic shifts.

5. TRAINING AND EVALUATION

In order to validate the use of topic modeling for speaker role labeling, we conducted several experiments on two meeting-based corpora (ICSI and NWTRB). For the ICSI dataset, we trained the LDA model (using the Mallet implementation [10]) on 58 meeting excerpts, with context windows of $k = \{2, 4\}$ and number of topics $N_t = \{5, 8, 12\}$. We centered the number of topics around eight because this was the number of dominant topics found in related work for this dataset [9]. In this case, we did not use n-gram features and hence there were 80 baseline features (see [7] for details). For the NWTRB dataset, we used 186 excerpts with context windows of $k = \{2, 4\}$ and N_t topics, $N_t = \{8, 12, 20\}$.

We then performed topic segmentation, as described earlier, using the learned model and extracted features for each speaker around the resulting topic shifts. A context window of $k = \{2\}$ and temporal window of 10 seconds around the topics shifts were used. These features, in addition to the baseline set of features described in [7], were used to train a supervised Adaboost classifier (the BoostTexter implementation was used [11]) to classify leaders and non-leaders. We compared the resulting classification accuracy, as measured by precision, recall, and f-score, to the baseline system, which did not use topic-related features. In the case of ICSI, we also had ground-truth topic segmentations that we used as a comparison [9].

Dataset	Configuration	Precision	Recall	F1
ICSI	Original	0.712	0.638	0.673
	GT	0.740	0.638	0.685
	k=3,t=5	0.500	0.431	0.463
	k=3,t=8	0.745	0.707	**0.726**
	k=3,t=12	0.540	0.466	0.500
	k=5,t=5	0.709	0.672	0.690
	k=5,t=8	0.731	0.655	0.691
	k=5,t=12	0.694	0.586	0.636
NWTRB	Original	0.804	0.704	0.751
	k=3,t=8	0.798	0.720	0.757
	k=3,t=12	0.821	0.742	0.780
	k=3,t=20	0.818	0.699	0.754
	k=5,t=8	0.823	0.726	0.771
	k=5,t=12	0.870	0.753	**0.807**
	k=5,t=20	0.812	0.720	0.764

Table 2. Results of leader detection on two corpora across several parameter sets. The addition of topic-related features resulted in a significant increase in performance from an f-score of 0.638 to 0.726 for ICSI and from 0.751 to 0.807 for NWTRB. The system also performed favorably when compared to using ground truth topic segmentation for ICSI (GT).

All results represent accumulated performance over n-fold cross-validation, with $n = 58$ for ICSI and $n = 186$ for NWTRB.

6. RESULTS

Table 2 gives a summary of our experimental results. For the ICSI data, the original baseline system (excluding n-gram features) achieved an f-score of 0.673. When using ground truth topic segmentation and adding the features calculated around topic shifts, this increased to 0.685. When unsupervised topic segmentation was used, the best result achieved an f-score of 0.726, a significant improvement. Interestingly, the unsupervised segmentation performed significantly better than ground-truth topic segmentation. This may be because the human annotation of topics was more aggressive and annotated a larger amount of topic shifts, as can be seen in Figure 2. This can add significant noise to the features since some of these annotations did not represent shifts in high-level topics that a leader would tend to be involved in. For the NWTRB dataset, we again saw a significant improvement in f-score from 0.751 to 0.807. In this case, a context window of five resulted in better performance, although it is interesting to note that the same pattern of performance can be seen for both context window sizes as the number of topics is varied. This suggests that there may be an intrinsic number of agenda-level topics in the conversations.

Figure 3 is presented as an qualitative example of our approach. The topic mixtures and resulting segmentation of 150 utterances from an ICSI meeting is plotted, and two utterance snippets are displayed. These utterances were copied from the topic transition areas, and they give an example of the leader's active role at the onset of new topics. In the transition to Topic 1, the leader (me013) gives the go-ahead for me018 to introduce a new topic. In the transition to Topic 2, the leader solicits new topics: "So what else?". This type of leader activity is seen at or near topic transitions throughout the ICSI and NWTRB meetings.

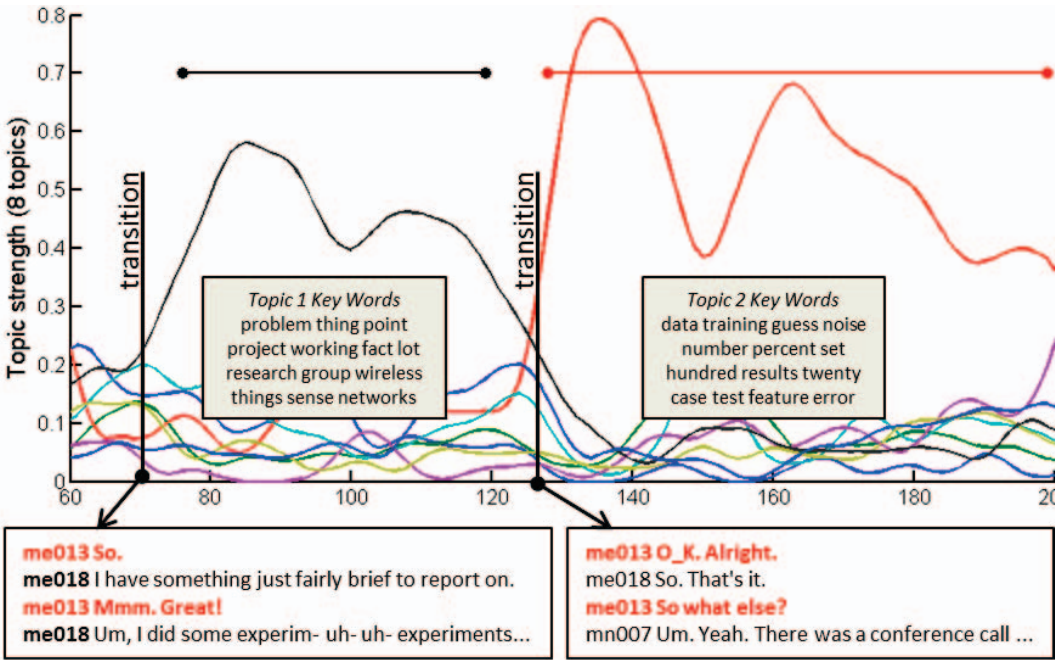

Fig. 3. This figure shows the topic modeling, topic segmentation, and selected utterances from 150 utterances near the beginning of ICSI meeting Bro014. The segmentation is shown by black line segments. The utterances are taken from the onset of new topics, and they show that the leader (bold/red utterances) is active at each transition. The key words for the two topics are also given.

7. SUMMARY

In this paper, we hypothesized that topics shifts occurring during multi-party spoken discourse could lend important clues towards leader detection. We demonstrated that LDA, in an unsupervised manner, can successfully determine topic shifts and that useful features can be computed around these shifts to improve leader detection accuracy. Building on this, we believe that tracking the shifts from topic to topic as conversations progress can yield important information not just for role labeling, but the detection of other group dynamics as well. For example, certain topics could correlate with different characteristics of a speaker or different frequencies of topic shifts could correlate with group characteristics such as cohesiveness or productivity.

Acknowledgments This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through Army Research Laboratory (ARL) contract number W911NF-09-C-0089. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

8. REFERENCES

[1] S. Yaman, D. Hakkani-Tür, and G. Tür, "Social role discovery from spoken language using dynamic bayesian networks," in *INTERSPEECH*, 2010, pp. 2870–2873.

[2] W. Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *ICASSP*, 2011, pp. 5556–5559.

[3] B. Hutchinson, B. Zhang, and M. Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *ICASSP*, 2010, pp. 5322–5325.

[4] M. Purver, K. P. Körding, T. L. Griffiths, and J. B. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *ACL*, 2006.

[5] M. Dowman, V. Savova, T. L. Griffiths, K. P. Körding, J. B. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 7, pp. 1238–1248, 2008.

[6] T. Strzalkowski, G. A. Broadwell, J. Stromer-Galley, S. Shaikh, S. Taylor, and N. Webb, "Modeling socio-cultural phenomena in discourse," in *COLING*, 2010, pp. 1038–1046.

[7] W. Wang, K. Precoda, R. Hadsell, Z. Kira, C. Richey, and G. Jiva, "Detecting leadership and cohesion in spoken interactions," in *Submitted to ICASSP*, 2012.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] A. Gruenstein, J. Niekrasz, and M. Purver, "Meeting structure annotation: Data and tools," in *SIGdial Workshop on Discourse and Dialogue*, 2005, pp. 117–127.

[10] A. McCallum, "Mallet: a machine learning for language toolkit," `http://mallet.cs.umass.edu`, 2002.

[11] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.