View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Edinburgh Research Archive

AUTOMATIC TOPIC SEGMENTATION AND LABELING IN MULTIPARTY DIALOGUE

Pei-Yun Hsueh and Johanna D. Moore

School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, GB

ABSTRACT

This study concerns how to segment a scenario-driven multiparty dialogue and how to label these segments automatically. We apply approaches that have been proposed for identifying topic boundaries at a coarser level to the problem of identifying agenda-based topic boundaries in scenario-based meetings. We also develop conditional models to classify segments into topic classes. Experiments in topic segmentation show that a supervised classification approach that combines lexical and conversational features outperforms the unsupervised lexical chain-based approach, achieving 20% and 12% improvement on segmentating top-level and sub-topic segments respectively. Experiments in topic classification suggest that it is possible to automatically categorize segments into appropriate topic classes given only the transcripts. Training with features selected using the Log Likelihood ratio improves the results by 13.3%.

1. INTRODUCTION

This study concerns the problem of segmenting a conversation record into a number of smaller segments and that of classifying each locally coherent segment into topic classes. Our interest in the problem is two-fold: First, topic segmentation and labeling provides the right level of detail for users to interpret what has transpired and locate relevant information in a multiparty dialogue. For example, upper management can efficiently locate critical deicsions made in a product design meeting by browsing the topic hierarchies. Second, it can lend support to the development of computer supported collaborative work applications, where group meeting records are automatically processed in order to extract information for summarization, question answering and providing thumbnail views on mobile devices.

2. RELATED WORK

Past research has explored the effect of a variety of features on characterizing topic boundaries. For example, [1] has studied lexical cohesion and proposed the TextTiling algorithm, an unsupervised approach that hypothesizes boundaries as points where the lexical cohesion score changes significantly. [2] and [3] have also used lexical cohesion to hypothesize segment boundaries in broadcast news transcripts and spontaneous speech. Recent advances in statistical text classification have inspired researches to cast the segmentation task as a binary classification task. Various combinations of features have been proposed to train the classification models, e.g., prosodic cues [4, 5], lexical features (N-grams) and discourse cues [6], lexical cohesion and conversational features [3].

[3] has applied a supervised classification approach that combines knowledge from various sources to identify top-level boundaries in meetings of the ICSI corpus. [7] has studied the problem of predicting topic boundaries at different levels of granularity and showed that the supervised classification approach performs better on predicting a coarser level of topic segmentation. As we would like to understand whether this finding is generalizable to agenda-based topic segmentations, this study applies these approaches to the problem of identifying topic boundaries in the scenario-driven meetings of the AMI corpus.

The task of topic labeling is a task complementary to that of topic segmentation. Prior research has proposed modeling topics explicitly using generative models, in which a collection of mutually independent observations are probabilistically generated by a hidden topic variable [8, 9]. Generative topic models can also be used to hypothesize segment boundaries where the value of the topic variable for the next observation changes. Other research has proposed merging similar utterances into topic clusters using unsupervised clustering approaches that minimize inter-cluster similarity and maximize intra-cluster similarity [10, 11].

3. METHODOLOGY

3.1. Topic Segmentation

In this study, we compare two segmentation approaches: (1) an unsupervised lexical cohesion-based algorithm (LCseg) using solely lexical cohesion information, and (2) a supervised classification approach that trains decision trees (C4.5) on a combination of lexical cohesion and conversational features.

The first approach, LCSeg, hypothesizes that a major topic

shift is likely to occur where strong term repetitions start and end. The algorithm works with two adjacent analysis windows, each of a fixed size which is empirically determined. For each utterance boundary, LCSeg calculates a lexical cohesion score by computing the cosine similarity at the transition between the two windows.

The second approach employs the supervised classification framework, in which each potential topic boundary is labelled as either boundary (POS) or non-boundary (NEG). Our objective here is to train decision trees (c4.5) to learn the most predictive combinations of features that can characterize topic boundaries. This study uses the features described in [7], including the amount of overlapping speech, the amount of silence between speaker segments, the level of similarity of speaker activity and the number of surrounding cue phrases. To study the effect of lexical cohesion on the performance of the combined model, this study also includes the lexical cohesion score, the estimated posterior probability, and the prediction of LCseg in the feature set.

3.2. Topic Labeling

The topic labels in the AMI scenario-driven meetings are selected from a standardized set of topic descriptions. Therefore, the task of automatic topic labeling can be cast as a task similar to text classification, in which each segment is assigned to appropriate descriptions given the transcript of the speech. To ease the burden on classifiers, we convert the multi-class problem to multiple binary classification tasks: For each topic class, we compile the transcripts of speech in the segments that have been labeled as belonging to this topic class as its training data. Then each segment is represented as a vector space of N-grams. Finally conditional Maximum Entropy (MaxEnt) models are trained from the training data to classify an unseen topic segment.

The aim of this study is two-fold: first, we want to show whether it is possible to classify topics given only the lexical features extracted from the transcript. This is studied by examining the accumulated effect of all N-grams on topic classification accuracy. Second, we want to understand whether it is possible to attribute the classification accuracy to a subset of features that are indicative of the target topic class. This is studied by exploring different feature selection criteria to measure lexical discriminability, which is defined as the association strength between the occurrence of a given N-gram and that of a topic class. In particular, this study applies four measures, Log Likelihood (LL), Chi-Squared (X2), Point-wise Mutual Information (PMI) and Dice Coefficient (DICE), to assess the lexical discriminability of each N-gram.

The LL and X2 measures capture the association strength by summing over the amount of variation between the observed frequencies (O) and expected frequencies (E)[1] in the 2x2 contingency table into a single-valued parameter.[2] It is posited that if an N-gram occurs significantly more often in a topic class than expected by chance, the N-gram can be viewed as associated more strongly with this topic class. In contrast, the information theoretic PMI and DICE measures capture the association strength by measuring the mutual dependence of discrete events, that is, the correlation coefficient between the occurrence of the N-gram and that of the target topic class, estimated by the observed frequency of the N-gram in the target topic class ($O(a)$) and its expected frequency ($E(a)$), the occurrence counts of the N-gram ($O(ng)$), and the total number of N-grams in the topic class ($O(TOPIC)$).

$$LL(ng) = \sum O \log(O/E) \quad (1)$$

$$X2(ng) = \sum ((O-E)/E)^2 \quad (2)$$

$$PMI(ng) = \log(O(a)/E(a)) \quad (3)$$

$$DICE(ng) = 2O(a)/(O(TOPIC) + O(ng)) \quad (4)$$

4. EXPERIMENT

4.1. Annotation

Topic segmentation and labels have been annotated for the AMI meeting corpus. 138 out of 170 AMI meetings are driven by a scenario, wherein four participants play the roles of project manager, marketing expert, industrial designer, and user interface designer in a design team, taking a design project from kick-off to completion. Annotators have the freedom to mark a topic as subordinated (down to two levels) wherever appropriate. As participants follow a predetermined agenda in these scenario meetings, annotators are expected to find that most of the topics recur. Therefore, they are given a standard set of topic descriptions that can be used as labels for each identified topic segment. Annotators will only add a new label if they cannot find a match in the standard set. The standard set of topic descriptions has been divided to three categories:

- Top-Level Topics refer to topics whose content largely reflects the meeting agenda (e.g., presentation, discussion, evaluation) and the key issues of the design task (e.g., project specs, target group).
- Sub-Topics refer to parts of the top-level topics (e.g., project budget, look and usability, trend watching, components, materials and energy sources).
- Functional Topics are the parts of the meeting that refer to the meeting process (e.g., opening, closing, agenda), or are simply irrelevant (e.g., chitchat).

[1] The expected frequency in each cell is computed as if the occurrences of the N-grams and the topic classes were expected by chance.

[2] In the 2x2 contingency table, the values of the four cells correspond to the frequency of a given N-gram (ng) in the target topic class (a) and that in all the other non-target topic classes (b), and the frequency of all the other N-grams in the target topic class (c = the total number of N-grams in the topic class O(TOPIC) - a) and that in the non-target topic classes (d = the total number of N-grams in the non-target topic classes - b).

As we are interested in comparing the segmentation algorithms on predicting topic boundaries at different levels of granularity, this research flattens the subtopic structure and considers only two levels of segmentation–top-level topics and all subtopics. The scenario-driven meetings in the AMI corpus have on average eight top-level topic segments, which either describe items in the agenda or serve functional purposes. In addition, the AMI meetings have on average three more sub-topic segments that form parts of the top-level Topics. Compared to the ICSI corpus, AMI meetings are shorter and with relatively shallower hierarchies.[3] To establish the reliability of the procedure, we have calculated the intercoder agreement (kappa) on the two meetings that have multiple codings, achieving in average 0.66 and 0.59 at the top-level and subtopic level.

4.2. Evaluation Metrics

To evaluate the performance of segmentation models, we use metrics that have proven useful in the fields of text segmentation (P_k and W_d), designed to overcome limitations in the use of precision and recall.[4] To evaluate the performance of topic models, classification accuracy is calculated as follows. We loop over each topic in the standardized set and compute the precision and recall as the total number of segments that have been assigned correctly to the topic class divided by the total number of reference segments and hypothesized segments of the topic class respectively.

5. RESULTS

5.1. Experiment 1: Predicting top-level and subtopic segment boundaries

This experiment aims to explore whether approaches previously proposed can be applied to identify functional segments and agenda-based top-level and subtopic segments in the AMI scenario meetings. In this experiment, we perform a five-fold cross validation. In each fold, we train models on 6 series of 4 meetings each and test on one unseen series of meetings. All of the results are reported on the test set. Table 1 shows the performance of the LCSeg algorithm and the feature-based classification models (CM) integrating the lexical cohesion and conversational features discussed in Section 3.1. Results show that CM performs better than LCSeg on predicting topic boundaries in the scenario meetings when the number of segments is unknown. CM performs better on the task of predicting top-level agenda-based topic boundaries, achieving 20% improvement over the performance of predicting top-level and subtopic boundaries by LCSeg. The results are consistent with previous findings that feature-based approaches are preferred for finding topic boundaries at a coarser level.[5]

Error Rate (Pk/Wd)	LCSeg (k)	LCSeg (unk)	CM (c4.5)
ICSI (TOP)	0.26/0.29	0.36/0.47	0.28/0.30
ICSI (SUB)	0.32/0.36	0.32/0.38	0.37/0.39
AMI (TOP)	0.33/0.49	0.41/0.51	0.33/0.33
AMI (SUB)	0.36/0.47	0.41/0.49	0.36/0.36

Table 1. *Performance comparison of probabilistic models at the two levels of topic granularity: Top-Level Topics (TOP) and Sub-Topics (SUB). The parameter of LCSeg specifies whether the number of segments is known (k) or not (unk).*

5.2. Experiment 2: Classifying topics using only lexical features

In this study, the task of automatic topic labeling is cast as multiple binary classification tasks. Again, we performed a five-fold cross validation. The first column of Table2 suggests that simply using the uni-gram features can automatically classify some of the Functional segments (e.g, agenda, closing) and agenda-based Top-Level topics (e.g., project spec, target group) and Sub-Topics (e.g., budget, trend watching). Results show that unigram features have distributions different enough between topic classes to be used to train the models for classification. We also trained models using bi-gram and tri-gram features. However, none of these models work better than models using unigram features alone.

Accuracy (F1)	1gram ALL	1gram LL-Q1	1gram DICE-Q1
FUNCTIONAL (average)	0.57	0.62	0.54
FUNCTIONAL (Closing)	0.56	0.67	0.53
FUNCTIONAL (Agenda)	0.58	0.58	0.55
TOP-LEVEL (average)	0.45	0.53	0.48
TOP-LEVEL (Target Group)	0.36	0.63	0.38
TOP-LEVEL (Project Spec)	9.52	0.44	0.50
TOP-LEVEL (Evlauation)	0	.67	0
SUB-TOPIC (average)	0.40	0.44	0.40
SUB-TOPIC (Budget)	0.50	0.71	0.57
SUB-TOPIC (Trend)	0.50	0.55	0.50

Table 2. *Effect of N-gram features on the accuracy of classification models.*

[3]The AMI scenario meetings last approximately 30 minutes, whereas the ICSI meetings last an hour in average. The meetings in the ICSI corpus have on average seven top-level topic segments and ten more subtopic segments.

[4]The P_k measure is the probability that a randomly drawn pair of utterances are incorrectly predicted as from the same segment. The WindowDiff (W_d) measure is the probability that the number of hypothesized and reference boundaries in a given window frame are different.

[5]Note that because the procedure of obtaining the results on the ICSI and AMI corpus are different, the results reported here are not directly comparable across these two corpus.

5.3. Experiment 3: Selecting discriminative features for the task of topic classification

Having established that it is possible to classify topic classes given the unigrams in the transcript, we then assess the effect of lexical discriminability measures on classification accuracy. We perform the following procedure: We first apply each of the four measures to calculate the lexical discriminability of all N-grams in the topic model and then sort N-gram features according to their computed lexical discriminability scores. Then we train classification models using the 25% most discriminative (Q1), the 25% mildly discriminative (Q2), the 25% mildly indiscriminative (Q3) and the 25% least discriminative (Q4) of these sorted N-gram features. Finally, we examine the effect of features at different levels of lexical discriminability on classification accuracy.

We posit that if the lexical discriminability measure works well, the performance of the models trained using Q1 features should outperform the models trained using other subsets of less discriminative features. Table 3 suggests for models that are trained with the LL, or DICE measure, Q1 features are the best predictors, followed by Q2, Q3, and Q4 features. In other words, the LL and DICE measures correspond well to the association strength between an N-gram feature and a topic class. The second column of Table 2 suggests that using discriminative features selected by LL achieves the best performance in the task of topic classification, improving the results of classifying Functional, Top-Level Topics and Sub-Topics by 8.1%, 17.8%, and 10% respectively.

	Q1	Q2	Q3	Q4
LL	0.58	0.26	0.21	0.08
X2	0.51	0.39	0.41	0.08
DICE	0.55	0.24	0.00	0.00
PMI	0.00	0.09	0.28	0.29

Table 3. *Effect of feature selection methods on average classification accuracy (F1) of models trained with uni-gram features. Q1, Q2, Q3 and Q4 features refer to features selected at different levels of lexical discriminability.*

6. CONCLUSIONS AND FUTURE WORK

In this study, we have quantitatively assessed the effectiveness of approaches for both the task of topic segmentation and that of topic labeling. Experiments show the feature-based approach previously proposed for finding coarse-level topic boundaries can be applied to segment scenario meetings. Experiments also show that the topic modeling approach that uses only the lexical features extracted from the transcript works well on classifying the functional segments and the agenda-based segments which involve discussions of prominent topics. Furthermore, this study develops lexical discriminability measures to select a subset of topic-indicative features in order to further reduce the vector space required for representation and to improve the classification accuracy.

A natural next step is to develop probabilistic models that can perform the task of topic segmentation and labeling simultaneously. To systematically incorporate contextual dependencies, we will train conditional random fields geared towards labeling sequential data. Finally, as we do not want to assume perfect human transcripts are always available, we will assess these models directly on ASR output.

7. REFERENCES

[1] M. Hearst, "Texttiling: Segmenting text into multiparagraph subtopic passages," *Computational Linguistics*, vol. 25(3), pp. 527–571, 1997.

[2] N. Stokes, J. Carthy, and A.F. Smeaton, "Select: a lexical cohesion based news story segmentation system," *AI Communications*, vol. 17(1), pp. 3–12, Jan. 2004.

[3] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of the 41st ACL Annual Meeting*, 2003.

[4] B. Grosz and J. Hirschberg, "Some intonational characteristics of discourse structure," in *Proceedings of the ICSLP*, 1992.

[5] D.J. Litman and R.J. Passoneau, "Combining multiple knowledge sources for discourse segmentation," in *Proceedings of the ACL*, 1995.

[6] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, pp. 177–210, 1999.

[7] P-Y. Hsueh, J. Moore, and S. Renals, "Automatic segmentation of multiparty dialogue," in *the Proceedings of the 11th Conference of EACL*, 2006.

[8] P. van Mulbregt, J. Carp, L. Gillick, S. Lowe, and J. Yamron, "Segmentation of automatically transcribed broadcast news text," in *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[9] D. M. Blei and P. J. Moreno, "Topic segmentation with an aspect hidden Markov model," in *Proceedings of the 24th ACM SIGIR Conference*, 2001.

[10] J. M. Ponte and W. B. Croft, "Text segmentation by topic," in *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, 1997.

[11] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *Proceedings of the 28th Annual Meeting of the ACL*, 2001.