Identifying Discourse Boundaries in Group Discussions using a Multimodal Embedding Space

Ken Tomiyama
Seikei University
Musashino, Tokyo 180-8633 Japan
dm176207@cc.seikei.ac.jp

Fumio Nihei
Seikei University
Musashino, Tokyo 180-8633 Japan
dd166201@st.seikei.ac.jp

Yukiko I. Nakano
Seikei University
Musashino, Tokyo 180-8633 Japan
y.nakano@st.seikei.ac.jp

Yutaka Takase
Seikei University
Musashino, Tokyo 180-8633 Japan
yutaka-takase@st.seikei.ac.jp

ABSTRACT
In group discussion, it is not always easy for the participants to effectively control the discussion to make it fruitful. With the goal of contributing to facilitating group discussions, this study proposes a method of segmenting a discussion. Predicted discussion boundaries may be useful for tracking the discussion topics, analyzing the discussion structure, and determining a timing for intervention. We created a multimodal embedding space using an autoencoder, and represented each multimodal utterance data in the embedding space. Then, a simple unsupervised approach was used to detect the discussion boundary. In a preliminary analysis, we found that the proposed method can generate discussion segments that are comprehensible for analyzing a discourse structure. But, the performance in the discourse segmentation task should be improved as future work.

Author Keywords
Group discussion; discourse segmentation; autoencoder; multimodal embedding.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Group discussion is widely used for decision-making and idea generation. However, it is not always easy for the participants to effectively control the discussion by themselves. A facilitator is a person who helps the participants establish common understanding and reach consensus during the conversation. In order to make an effective contribution, the facilitator needs to choose a right timing for intervening to the discussion while observing the discussion. Thus, for the purpose of exploiting information technology in supporting a discussion, tracking a discussion is one of the basic function for a computer system to facilitate the discussion.

There were many previous studies for topic tracking and discourse segmentation. There were mainly two approaches in this research area. Unsupervised approach is based on lexical cohesion, such as identical words, synonyms, and hypernyms [1, 2]. Discourse boundary is determined by the concise similarity between word vectors. The other approach is the supervised approach, where a set of features are calculated and a classifier is learned to decide a boundary or non-boundary [3]. While the motivation of these previous studies is to use discourse boundaries to identify more informative segments, retrieve specific information more accurately, and generate a summary of the discourse. The purpose of discourse segmentation in this study is slightly different. We aim to identify discussion boundaries, each of which is a kind of shift in the discussion and may be an appropriate intervention timing for facilitation. Thus, each discourse segment divided by a boundary should be a coherent discourse.

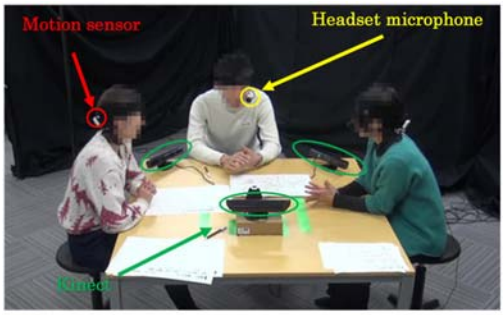

Figure 1. Snapshot of experiment

Moreover, group discussion is not well-structured compared to texts, and discussion segmentation would be more difficult than text segmentation. The discussion sometimes does not go straightforwardly, and the same topic may be discussed multiple times. As more closely related work, [4, 5] proposed a discourse segmentation model by employing a feature-based supervised classification approach.

However, feature selection is a painful process. In this study, we employ an autoencoder to learn multimodal embedding space to represent each utterance as a vector. The advantage of this approach is that feature selection is not necessary. Then, we employ unsupervised approach to decide a discourse boundary by calculating cosine similarity between the vectors.

SymCollab '18, March 11, Tokyo, Japan.

GROUP DISCUSSION CORPUS

Task and Subjects

We recruited 30 subjects (10 groups of 3 people), who were native Japanese speakers. They participated in a group discussion for 30 minutes to create a one-day travel plan for foreigners. The group of participants cooperatively filled in a work sheet in which they described (1) the country of the expected travelers, (2) the catchphrase and (3) the details of the sightseeing course, and (4) its selling points. The participants were instructed to discuss four themes (1) to (4) in this order. In order to enhance the motivation to be engaged in the task, they were also instructed that their plan would be evaluated later (e.g. the number of sightseeing spots included in the plan).

Experimental Environment

Figure 1 shows a snapshot of the experiment. Three people were seated at a table, and each of them wore a head set microphone (Audio-technica HYP-190H) to record speech data. Inertial Motion Unit (IMU, ATR-Promotions: WAA-010) were attached to the back of each participant's head. These sensors measured head acceleration, angular velocity, and terrestrial magnetism in the x, y, and z coordinates at 20 fps. A Kinect sensor placed on the other side of each participant was used to collect face tracking data individually[1]. In addition, two video cameras were set to record the overview of the communication. Speech data were manually transcribed.

MULTIMODAL EMBEDDING SPACE

From the speech audio[2], we obtained 7052 utterances, for each of which we calculated following verbal and nonverbal features.

Features

(1) The number of new/already used nouns: Nouns were extracted from speech transcription using the Mecab morphological tagger. Then, each of the extracted nouns was categorized as a new noun or a used noun. If the noun had already been used in the conversation, it was categorized as a used noun, If not, it was categorized as a new noun. The number of new/already used nouns was counted for each utterance.

(2) The number of nous in common/different between the current and the previous utterance: We counted the number of nouns that were shown in both the current and the previous utterance (hereafter "nouns in common"). We also counted the number of nouns that were shown in the current utterance but not in the previous one (hereafter "different nouns").

(3) The number of verbs in common/different between the current and the previous utterance: The number of verbs in common and the number of different verbs were counted in the same way as in (2).

(4) Utterance length (time duration and the number of morphemes): We used two types of measures for utterance length. One is the time duration of utterance. The other is the number of morphemes contained in the utterance.

(5) Utterance overlap: If a given utterance was overlapped with other one, the length of overlapping time was measured. If the utterance was overlapped with other two utterances (three people were speaking at the same time), both overlapping time intervals were added up.

(6) Speech intensity Speech intensity (db) was measured every 10 ms using the Praat audio analysis tool, and maximum, minimum, average, and variance were calculated for each utterance.

(7) Head rotation: Head rotation in the y-axis was measured every 20 ms from the Kinect face tracking data. Then, maximum, minimum, average, and variance were calculated for each utterance.

(8) Composite head acceleration: IMUs were attached to the back of each participant's head, and the acceleration was measured at 20 frames per second (fps). The composite acceleration for x, y, and z axes was computed for each time frame *i* using the following equation;

$$HA_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$$

Then, maximum, minimum, average, and variance were computed for each participant per utterance.

(9) Wavelet features for the composite head acceleration: This feature is used for measuring the synchrony of the head motions between discussion participants. Multiresolution analysis with Daubechies wavelets [6] was applied to the composite acceleration calculated in (8). Then, maximum, minimum, average, and variance were computed for a wavelet at the highest resolution.

(10) Doc2Vec features: A Doc2Vec [6] model, which was trained by using Wikipedia articles written in Japanese, was applied to each utterance, and a 200-dimensional vector was obtained. All elements of the vector were used as features.

LEARNING A MULTIMODAL EMBEDDING SPACE

All the features described in the previous section were concatenated, and each utterance was represented as a 214-dimentional vector, including 12-dimentions for Wavelet analysis, 4-dimentions for speech intensity, 4-dimentions for head rotation, and 200-dimentions for Doc2Vec features. We

[1] Kinect data was not used in this work.

[2] One group was excluded from the analysis because the speech audio was not recorded by mistake.

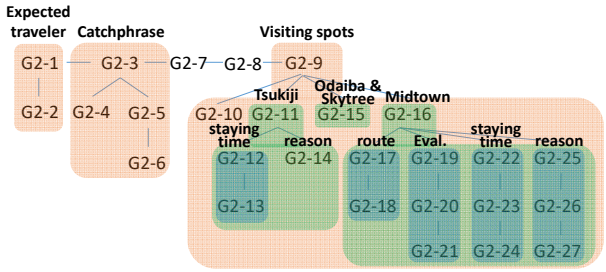

Figure 2. Interpretation of the structure of a discussion

used this 214-dimentional vectors as the input to an autoencoder.

We built an autoencoder consisting of one input layer, one hidden layer, and one output layer. We used ReLu as the activation function in the hidden layer, and a linear function for output layer. Minimum square error was used in the cost function. The 241-dimensionnal data in the input layer was reduced to150-dimensions in the hidden layer. The data from 7 out of 9 groups (4044 utterances) were used for training, and the data from the remaining two groups (1124 utterances) were used for testing.

ANALYSIS

The test data obtained from two groups were used in the following analysis. Each utterance was represented as an output vector from the autoencoder. Then, the cosine similarity values were calculated by pairing the current utterance with the previous three utterances, and the average of three similarity values was calculated. If the average similarity with the recent three utterances was lower than 0.75, the current utterance was identified as a discussion boundary.

Coherence

In order to test the coherence of each discussion segment, we calculated the lexical similarity between the segments. First, a word vector was generated for each segment by extracting nouns and verbs from the transcription. Then, the cosine similarity was calculated for all the pairs of segments. The cosine similarity was generally very low. In more than 90% of the pairs, the cosine similarity is lower than 0.2. While in 0.7% of the pairs, the similarity was over 0.5, the content of these segments was quite similar to each other (e.g. discussing the same place). These results suggest that each discussion segment had enough lexical coherence.

As a qualitative analysis, we visualized the structure of the discussion based on the segments obtained. Figure 2 visualizes the structure of a discussion of Group2. Labels, such as G2-1, indicate a discussion segment. As shown in the diagram, the main stream of the discussion can be easily interpreted: starting from determining the expected travelers, followed by the discussion about the catchphrase and the visiting spots. In addition, it was also possible to assign sub-

Window size	Precision	Recall	F-measure
4	0.52	0.58	0.55
5	0.60	0.64	0.62
6	0.67	0.70	0.68

Table 1. Agreement of discussion boundary judgment with a human annotator. (autoencoder)

Window size	Precision	Recall	F-measure
4	0.47	0.52	0.50
5	0.57	0.62	0.60
6	0.65	0.70	0.67

Table 2. Agreement of discussion boundary judgment with a human annotator. (doc2vec)

topics for some segments. For instance the topic “Midtown” (G2-16) has four sub-topics: “route”, “evaluation”, “staying time”, and “reason.” This suggests that the results of automatic segmentation is comprehensible for a human analyzer, and there is a good possibility that such segmentation is useful for supporting a human facilitator.

Agreement with the segmentation by a human annotator

As a preliminary analysis, we compared the result of automatic segmentation with the segmentation by a human annotator. For the last part of the group work, the participants mainly worked on filling out the task sheet, and the interaction is very different from other parts. Thus, we did not use the data for this part. So we used 404 utterances from the Group2 discussion, and 453 from the Group7 discussion. The model detected 58 boundaries for the Group2 discussion, and 79 for the Group7 discussion, while the human annotator detected 56 and 55 respectively. In order to permit near miss judgment, we set a tolerance window of size *n,* and judged that the model prediction was correct if there was a boundary or no boundary within the window for both model prediction and human judgement. With this tolerated agreement measure, we calculated precision, recall, and F-measure. Table 1shows the evaluation results of the proposed model, and Table 2 shows the evaluation results of a model only using Doc2Vec features. These two models were created to compare a language-based model and a multimodal model. As the average segment length in the human annotation was 7.8, we assume that window size n=4 (half the average segment size) may be reasonable. Although the model performance should definitely be improved, the multimodal model outperformed the language-based unimodal model for all window sizes. As our final goal is not finding discourse boundaries, but identifying a good timing for intervention,

we need to propose more appropriate evaluation metrics as future work.

CONCLUSION

This study proposed a method for finding discussion boundaries based on the similarity of utterance vectors in a multimodal embedding space which was created by using an autoencoder. Although the performance in the discourse segmentation task is not good enough, the proposed method can generate segments that are comprehensible for interpreting a discourse structure.

As future directions, we will test the model in terms of determining a timing for intervention. In addition, it is necessary to improve the model for tracking the topics of discussion.

ACKNOWLEDGMENTS

This work was supported by CREST, JST.

REFERENCES

1. Hearst, M.A., *Multi-paragraph segmentation of expository text*, in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. 1994, Association for Computational Linguistics: Las Cruces, New Mexico. p. 9-16.
2. Choi, F.Y.Y., *Advances in domain independent linear text segmentation*, in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. 2000, Association for Computational Linguistics: Seattle, Washington. p. 26-33.
3. Beeferman, D., A. Berger, and J. Lafferty, *Statistical Models for Text Segmentation.* Mach. Learn., 1999. **34**(1-3): p. 177-210.
4. Galley, M., et al., *Discourse segmentation of multi-party conversation*, in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. 2003, Association for Computational Linguistics: Sapporo, Japan. p. 562-569.
5. Hsueh, P.-Y. and J.D. Moore. *Automatic topic segmentation and labeling in multiparty dialogue*. in *IEEE Spoken Language Technology Workshop*. 2006.
6. Le, Q. and T. Mikolov, *Distributed representations of sentences and documents*, in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. 2014, JMLR.org: Beijing, China. p. II-1188-II-1196.