7ZB-01

Leveraging Acoustic and Motion Signals for Detecting Topic Transitions in VR Meetings

Zhankun Liu†, Jiadong Chen†, Chenghao Gu†, Jiayi Zhang†, Shin'ichi Konomi‡

† *Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University*
‡ *Faculty of Arts and Science, Kyushu University*

I. Introduction

With the proliferation of consumer-grade VR devices, remote meetings within virtual environments are becoming increasingly popular. Meeting segmentation can swiftly offer users a valuable, advanced understanding of past meeting discourse while enhancing team communication efficiency in virtual environments. However, detecting topic transitions within VR-based meetings presents a unique set of challenges, primarily due to the limitations of non-verbal communication in virtual spaces. In subsequent research, we aim to investigate the correlation between topic transitions, acoustic characteristics, and changes in body posture within virtual environments. Furthermore, we propose a novel approach for dialogue topic segmentation that integrates both acoustic features and motion data.

II. Related Works

The multifaceted nature of topic segmentation extends beyond mere data partitioning; it encapsulates the intricate interplay among linguistic features, discourse structures, and semantic coherence. In this domain, supervised, unsupervised, and semi-supervised methodologies, each offering unique advantages and challenges in identifying thematic boundaries within textual and speech data.

In terms of textual models, there's a divide between supervised and unsupervised methods. Supervised approaches face challenges due to their reliance on extensive annotated data, leading to potential biases and difficulties with rare or domain-specific topics. Unsupervised models [1] alleviate some data annotation challenges but rely heavily on similarity measures, making them sensitive to nuances and variations in language and context.

Moreover, integrating non-textual features, specific acoustic characteristics [2] around speech segment boundaries exhibit recognizable patterns useful for topic segmentation, especially in the absence of lexical cues. These include prosodic changes and pauses. Although rhythm and acoustic-based models show promise for broad topic segmentation, the effective methods often combine acoustic, prosodic, and lexical features for optimal performance.

III. Experiment & Data collection

To investigate the relationship between acoustic feature data, motion data, and topic transitions, we established a pilot experiment. Initially, we recruited 24 participants (12 males and 12 females) and organized them into groups of four (2 males and 2 females)for task-based discussions within a virtual reality (VR) setting. We employed the same experimental setting as described in [3]. The main discussion lasted 20 minutes, focusing on a survival game scenario called "Lost at Sea". Each group was tasked with discussing and prioritizing a pre-defined list of 15 items based on their perceived importance for survival, culminating in a final ranked list. After the discussion concluded, each participant completed a survey based on their experience and the content of the experiment.

The questionnaire probed various aspects of discussion dynamics, including the frequency of deviating from the main topic, perceived impact of digressions on discussion efficiency, clarity of recall post-discussion, and the smoothness of transitions between topics. Responses were recorded using a Likert scale (see Table I), ranging from 1 (strongly disagree) to 5 (strongly agree).

Table I: Responses to Group Discussion Questionnaire

Questions	**1**	**2**	**3**	**4**	**5**
Frequencies of off-topic discussion	14	7	0	3	0
Effect of digressions on efficiency	1	5	7	8	3
Clarity of post-discussion recall	0	1	1	10	12
Smoothness of topic transitions	0	4	7	9	4

Notably, a small subset of participants perceived that deviations from the primary topic were prominent, indicating occasional but noticeable instances of digression. Despite the overall low frequency of such off-topic occurrences, the majority of participants believed that these instances adversely affected the discussion's efficiency to some extent. The smoothness of topic transitions received varied responses, suggesting variability in how different groups experienced the conversational flow. Some participants noted that transitions were relatively unnatural, marked by conspicuous silences or pauses.

IV. Modelling

We propose to utilize a specific variant of the sentence-transformer as the foundational text-based model, coupled with the NaturalConv [4] dataset for pre-training. This particular dataset is distinguished by its thematic conversation structure, compiled through engaging a wide array of participants in discussions centered on selected articles from newspapers and magazines. The dialogues captured in this dataset are rich in detailed explorations of various topics and exhibit natural transitions between multiple themes, making it an ideal choice for studying conversational dynamics and topic segmentation.

In the preparatory phase of the experiment, each group's dialogical exchange will be denoted as a document D = $\langle U, T \rangle$, where U = $\{u_1, ..., u_N\}$ represents a sequence of N consecutive utterances and T = $\varnothing$ signifies an initially empty topic set. We will identify n-1 intervals

between adjacent utterances, represented as V = $\{v_1$, ..., $v_N\}$. The segmentation algorithm will aim to predict segment boundaries B = $\{b_1$, ..., $b_k\}$, where k is the number of predicted boundaries, and each b_i represents a potential division point in the dialogue. For each interval, a relevance score r_i will be calculated, indicating the likelihood that adjacent segments belong to the same thematic block. Higher relevance scores suggest greater thematic continuity between segments. A segmentation technique, such as TextTiling or a derivative thereof, will be used to determine the final segment boundaries, dividing D into a series of distinct thematic sections T = $\{T_1$, ..., $T_{k+1}\}$.

Presented below (Table II) is a preliminary example of textual segmentation of group discussion dialogues, conducted using the all-mpnet-base-v2 variant of the sentence-transformer model. The segmentation approach adheres to the methodology of depth scoring and threshold setting as outlined in [5].

Table II: Dialogue Segmentation sample

Seg	Discussion Dialogue
...	...
37	Item 15 is definitely more important. Right, so I think item 15 should be ranked quite high up in priority.
38	After all, you can fish for food. The mosquito net can be used to catch fish. Yes, item 15.
39	I think item 15 can be ranked towards the front. The radio should also be prioritized towards the front.
40	But the radio is useless once the batteries run out. It's a portable radio. Some components inside the radio might still be useful. But you are a physics major, aren't you?
...	...

In the current text-based model's topic segmentation, a clear limitation is evident in its treatment of discussion elements as distinct topics when they might be perceived as part of a larger thematic context. Specifically, the model segments the conversation about determining an item's priority and its potential uses into two separate topics (Segments 37-38). However, this segmentation might not align with some users' perception, who may view these elements as facets of a broader, singular topic.

V. Towards VRMeeting Segmentation That Aligns With User's Perceptions

This issue we identified in the previous section subtly underscores the potential benefit of self-supervised learning. In contexts where topic boundaries are subjective and not explicitly defined, self-supervised learning can help the model develop a more nuanced understanding of thematic relationships and transitions based on the inherent structure of the dialogue.Additionally, the incorporation of multimodal models becomes critical, especially considering the inaccuracies in speech-to-text translation and the importance of non-verbal cues.

Regarding subsequent training methodologies, we plan to employ the self-supervised learning paradigm. Initially, the text-based pre-trained model will undertake preliminary topic segmentation. Subsequently, a panel of reviewers, comprising 5 to 6 individuals, will be assembled to evaluate the pertinence of each segmentation point, scoring them on a scale from 1 to 10. These evaluations will guide the construction of positive and negative samples for self-supervised learning tasks.

In the training phase of the multimodal model, the emphasis will be on harnessing a wide spectrum of motion and acoustic data to offer an alternative methodology for topic segmentation that does not rely exclusively on textual data. The motion data will include variables such as head orientation, nodding, tilting, hand and interactive gestures, and spatial positioning, each contributing to a richer understanding of the communicative context. Acoustic features will encompass elements like fundamental frequency, pitch, loudness, speech rate, pauses, and spectral qualities, as well as non-verbal cues like laughter and sighs [2], which are indicative of the speaker's emotional state and engagement.

VI. Conclusion

In this work, we described experimental data collection along with a questionnaire to examine the requirements of VR meeting segmentation, and discussed methods to segment meetings in ways that align with user's perceptions. As a next step, We plan to select an appropriate model to fulfill the predetermined experimental scheme and conduct comparative performance evaluations against existing topic segmentation models. Subsequently, We aim to compare methods that do not rely on textual models with those that do, in order to validate the efficacy of non-textual cues in making accurate predictions of transitions, even in the absence of semantic information.

We envision a practical application facilitating real-time summarization and topic transition prediction for VR meetings and assessing participant engagement levels with the current topic, thereby providing insights into whether a shift in the discussion focus might be beneficial. This application aims to enhance the efficacy and adaptability of VR meetings, ensuring that discussions remain relevant, engaging, and productive.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP20H00622,JP23H03507,JP23K02469.

References

[1] H. Gao, R. Wang, T.-E. Lin, Y. Wu, M. Yang, F. Huang, and Y. Li, "Unsupervised dialogue topic segmentation with topic-aware contrastive learning," in *Proc.46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2481–2485.

[2] C. Lai, M. Farrús, and J. D. Moore, "Integrating lexical and prosodic features for automatic paragraph segmentation," *Speech Communication*, 121, pp. 44–57, 2020.

[3] J. Chen, C. Gu, J. Zhang, Z. Liu, and S. Konomi, "Sensing the intentions to speak in vr group discussions," *Sensors*, 24, no. 2, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/2/362

[4] X. Wang, C. Li, J. Zhao, and D. Yu, "Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation," in *Proc.AAAI Conference on Artificial Intelligence*, 35(16), 2021, pp. 14 006–14 014.

[5] L. Xing and G. Carenini, "Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring," *Proc.22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 167–177, 2021.