DIALOGUE CONTEXT MODELLING FOR ACTION ITEM DETECTION: SOLUTION FOR ICASSP 2023 MUG CHALLENGE TRACK 5

Jie Huang, Xiachong Feng, Yangfan Ye, Liang Zhao, Xiaocheng Feng, Bing Qin, Ting Liu

Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

ABSTRACT

Action item detection aims at recognizing sentences containing information about actionable tasks, which can help people quickly grasp core tasks in the meeting without going through the redundant meeting contents. Therefore, in this paper, we thoroughly describe our carefully designed solution for the *Action Item Detection Track of the General Meeting Understanding and Generation (MUG) challenge in the ICASSP 2023 Signal Processing Grand Challenge*. Specifically, we systematically analyze the task instances provided by MUG and find that the key ingredient for successful action item detection is leveraging the dialogue context information into consideration. To this end, we design a simple and effective method for modelling context and utterance information concurrently. The experimental results show our method achieves remarkable improvements over baseline models, with an absolute increase of 0.62 of the F_1 score on the validation set. The stable generalizability of our method is further verified by our score on the final test set[1].

***Index Terms*—** Natural Language Processing, Meeting Understanding, Action Item Detection.

1. INTRODUCTION

Meetings are vital for improving productivity and become more and more frequent in our daily lives. With the development of communication technologies, meetings are widely being recorded and transcribed. Accordingly, there is a growing demand for the development of automatic techniques to analyze the key contents of meetings in various ways, such as topic detection, summarization, and action item detection.

Action items are one or more utterances in one meeting, which contain decisions made within the meeting that require post-meeting attention or manual labour. Giving quick access to actionable tasks in the meeting for participants can drastically reduce their efforts in processing verbose and redundant meeting contexts. To this end, track 5 of the MUG challenge, namely the action item detection (AID) track, requires participants to build a machine learning model to fulfill a binary classification task for every utterance in the meeting, where the positive samples being action items, as shown in Figure 1.

[1]https://modelscope.cn/competition/17/ranking

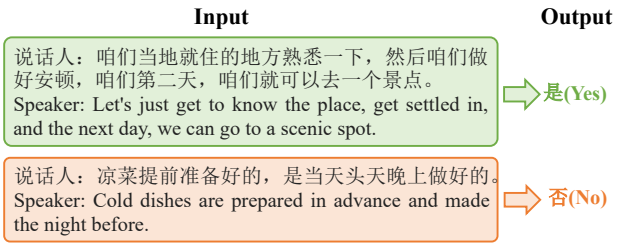

Fig. 1. Illustration for the action item detection task.

The crucial challenge of the action item detection task is that there is a lack of sufficient information given a single utterance, which means it is hard to judge whether one utterance is an action item or not if the background information of the meeting is completely unknown. Therefore, we first give a comprehensive analysis on hard action item samples, details are shown in section 2. Based on our observations, we conclude that it is necessary to take the dialogue context information, such as the tone and attitudes of other participants towards the statement of one participant, into consideration.

To this end, we propose a simple and effective method by modelling utterance and dialogue context information concurrently. Specifically, when detecting one utterance, we not only consider the current utterance but also incorporate the dialogue context information as well as the speaker information. We employ the StructBERT [1] as our backbone model and conduct extensive experiments. The experimental results on the one hand show the effectiveness of our method and on the other hand verify the strong generalizability of our model.

2. METHOD

Pre-trained models such as BERT, GPT-3 and ChatGPT have advanced various NLP tasks. Therefore, we draw support from these powerful models and finally decide to use StructBERT [1] as our backbone model due to its efficiency in dealing with broken sentences. After our preliminary experiments, we find that the baseline model is skilled at *identifying an action item, which contains task-related information and will be carried out at a specific time in the future*. However, it fails to have a more comprehensive understanding of complex action items: **(a) An action item must be approved**

Input Formats	Precision	Recall	F_1
Single Setting			
X	71.77	67.57	69.61
Base Concat Setting			
X [SEP] Y	68.81	62.61	65.57
X [SEP] 同一个人说：Y	69.23	68.92	69.07
Z [SEP] X	70.28	67.11	68.66
X [SEP] Y [SEP] Y'	72.33	67.12	69.63
Speaker-aware Concat Setting			
X [SEP] 同一个人说：Y X [SEP] 另一个人说：Y	68.10	71.17	69.60
Z [SEP] 同一个人继续说：X Z [SEP] 另一个人说：X	69.03	70.27	69.64
X [SEP] 同一个人说：Y X [SEP] 其他人说：Y	70.81	66.67	68.68
X [SEP] 同一个人继续说：Y X [SEP] 另一个人说：Y	**72.60**	**68.02**	**70.23**

Table 1. Experimental results on the validation set. X means the current utterance, Y and Y' represents the next two following utterances respectively, Z represents the utterance before X. For the Speaker-aware Concat setting, we supply different prefixes for two utterances according to whether the speakers of them are the same one.

by other participants. An utterance should not be detected as an action item if the action or the time to carry out the action is modified in the following meeting. **(b) An action item must include complete and workable action.** Thus, an utterance labelled with an action item is often a complete and self-contained sentence. **(c) An action item needs to be a certain decision, rather than an uncertain suggestion.** Thus, if an utterance is a question, only if it is approved by others in the following meeting, then it can become an action item. Accordingly, we conclude that the critical ingredient for successful action item detection is not only modelling the single utterance but also considering meeting context information concurrently to complement the comprehensive understanding of the current utterance. To this end, we propose a simple and effective method by concatenating both the context of the utterance as well as the speaker information to the utterance that needs to be detected, as shown in Table 1.

3. EXPERIMENT

3.1. AliMeeting4MUG Benchmark

We evaluate our method on the standard AliMeeting4MUG Corpus (AMC) [2]. Specifically, the AMC consists of 654 recorded meetings with high-quality manual annotations for the action item detection task. In detail, the train set contains 295 meetings, with 213,235 utterances in total. Note that there are barely 1,014 action item utterances in the train set, which makes this task of great challenge and leads to the severe class imbalance problem in machine learning. Additionally, the validation set has the same distribution with 65 meetings, 45,869 sentences and only 222 action items.

3.2. Implementation Details

We employ the base version of StructBERT. The learning rate is 2e-5, with a linear learning rate scheduler. We train this StructBERT for 5 epochs in total, the batch size is set to 32.

3.3. Experimental Results

Table 1 shows the experimental results. We can find that our method achieves the best performance, which shows the effectiveness of our design. We also find that the information of the speaker is necessary, which benefits the recall of the model. We attribute this key factor to the fact that the backbone model is trained using the next sentence prediction task during the pretraining stage, which makes it confused about the relationship between the two utterances we give to it in our experiments. Besides, our experiments show that adding more contextual information does not bring more benefits. Additionally, the final results on the test set clearly reveal the strong generalizability of our method, which achieve the first prize.

4. CONCLUSION

In this paper, we carefully present our analyses for the action item detection task and find several challenges that are essential for solving this challenging problem. Based on our observations, we intuitively propose a simple and effective method for modelling both utterance and dialogue context information, which on the one hand completes crucial evidence for identifying action items and on the other hand does good for utterance representation learning. Final experimental results show the effectiveness of our method that achieves the best results on the AMC benchmark, while also indicating the strong generalizability of our method.

5. REFERENCES

[1] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si, "Structbert: Incorporating language structures into pre-training for deep language understanding," *arXiv preprint 1908.04577*, 2019.

[2] Jiaqing Liu Hai Yu Qian Chen Wen Wang Zhijie Yan Jinglin Liu Yi Ren Zhou Zhao Qinglin Zhang, Chong Deng, "Mug: A general meeting understanding and generation benchmark," *ICASSP*, 2023.