Meeting Decision Tracker: Making Meeting Minutes with De-Contextualized Utterances

Shumpei Inoue[1], **Hy Nguyen**[1], **Pham Viet Hoang**[1], **Tsungwei Liu**[1], **Minh-Tien Nguyen**[2,*]
[1]Cinnamon AI, 10th floor, Geleximco building, 36 Hoang Cau, Dong Da, Hanoi, Vietnam.
{sinoue, hy, hugo, tsungwei}@cinnamon.is
[2]Hung Yen University of Technology and Education, Hung Yen, Vietnam.
tiennm@utehy.edu.vn

Abstract

Meetings are a universal process to make decisions in business and project collaboration. The capability to automatically itemize the decisions in daily meetings allows for extensive tracking of past discussions. To that end, we developed Meeting Decision Tracker, a prototype system to construct decision items comprising decision utterance detector (DUD) and decision utterance rewriter (DUR). We show that DUR makes a sizable contribution to improving the user experience by dealing with utterance collapse in natural conversation. An introduction video of our system is also available at https://youtu.be/TG1pJJo0Iqo.

1 Introduction

Obtaining a brief description and salient contents of meetings is a functionality that can certainly help business operations. Although automatic speech recognition enables us to transcribe meeting records automatically, its transcription is possibly much more verbose, noisy, or collapsed, and is far from being utilized in its raw form. Previous research attempted to extract important information from dialogue, such as decision-making utterances, (Bak and Oh, 2018; Karan et al., 2021), extractive summaries of online forums (Tarnpradab et al., 2017; Khalman et al., 2021), or group chat threads (Wang et al., 2022). Another study, Lugini et al. (2020) presented a discussion tracker to facilitate collaborative argumentation in classroom discussion by visualizing discussion transcription.

However, extracted utterances are usually incomplete and difficult to understand due to ellipses and co-references in conversations (Su et al., 2019). Figure 1 (the right) shows an example of a partial dialogue ending with a decision-related utterance in our dataset. This shows that objects or indicatives in utterances in natural conversations are usually ambiguous, and the meaning of decision-related utterances has a strong dependency on context. Furthermore, especially in Japanese, the format of the spoken language is often far apart from the written language because of frank expressions and many filler phrases. This nature reduces user experience with the naive use of utterances extracted from dialogues. In response to this, Incomplete Utterance Restoration (IUR) (Pan et al., 2019; Su et al., 2019; Huang et al., 2021; Inoue et al., 2022) handles the problem where the model rewrites and restores incomplete utterances by considering the dialogue context with promising results. However, we have yet to see IUR models applied for practical use in actual business applications.

This paper presents *Meeting Decision Tracker (MDT)*, a system that automatically generates the itemized decision list from meeting transcription. Given the meeting transcription, MDT detects decision-making utterances and rewrites them to the *de-contextualized utterance*, i.e., the written form with omissions restoration and filler removal. Such a capability allows users to look back at the previous meeting contents quickly and have asynchronous communication with no effort from a minute taker. The system has three crucial characteristics.

- By combining modules for extracting and rewriting decision-related utterances, the system has a down-to-earth strategy to generate itemized decision lists from meeting transcription. The combination allows us to investigate the role of IUR in a bigger context with significant impact for real business applications.
- Besides the ordinary task of IUR, our rewriter handles the translation from the spoken language to written language by filtering filler phrases. It enables users to understand the decision item at a glance, which contributes to improving the user experience.
- Although our system is originally built for

*Corresponding Author.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations, pages 88–94
November 20–23, 2022.

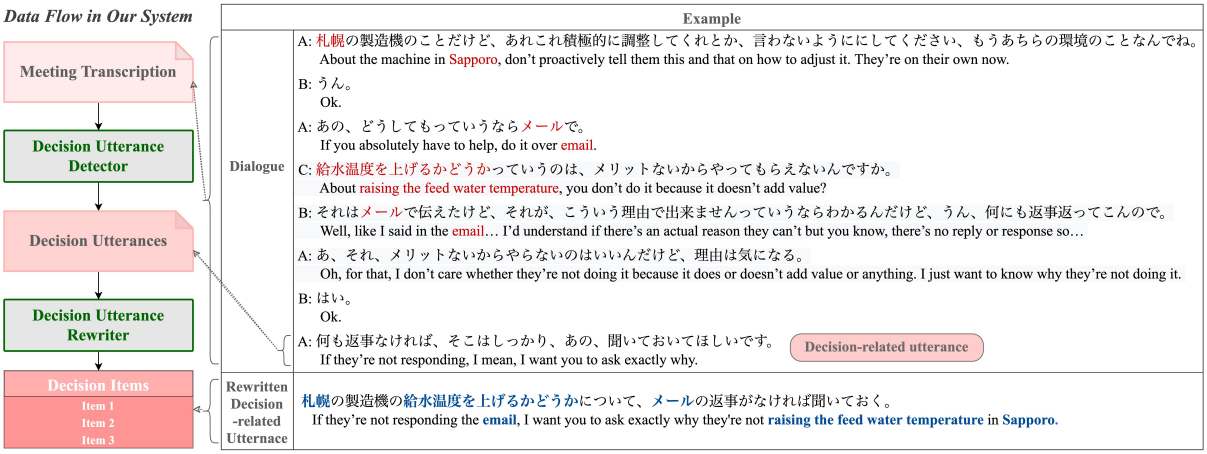


Figure 1: The data flow in our system and the conversation example. The red in the dialogue shows information omitted in the decision-related utterance. The blue shows information to be restored by Decision Utterance Rewriter.

decision utterance itemization, the proposed method can be applied as a general solution for information extraction from the dialogue.

2 System Design

The overall system architecture of Meeting Decision Tracker (MDT) is depicted in Figure 1 (the left). The main function of MDT is to generate decision items with de-contextualized representations from the transcription of daily business meetings. MDT comprises of two modules: Decision Utterance Detector (DUD) and Decision Utterance Rewriter (DUR). The detector extracts a decision list from meeting transcription and the rewriter translates (rewrites) the list to the written format. Figure 1 (the right) shows the example pair of the input and expected output for the system. The example indicates two points. First, the transcription contains decision-related utterances that can be used to summarize the content of the meeting. Second, the decision utterance itself is usually not self-consistent and comprehensible only after the utterance is restored by DUR. The next sections introduce the detector and rewriter.

2.1 Decision utterance detector

The first step of the detector is to detect decision-related utterances from transcription. We formulate the detection as a sequence labeling problem on the utterance level and describe the detector in two steps: input representation and classification.

Input representation The input uses the sequence of utterances $\{u_1, u_2, ..., u_w\}$ for the sequential classification, where w is the window size. Following Cohan et al. (2019), we used the input representation $\{[\text{CLS}], u_1, [\text{SEP}], u_2, [\text{SEP}], ..., u_w, [\text{SEP}]\}$, which contains the [CLS] token at the head of the whole input and [SEP] tokens at the tail for each utterance. Then the input was encoded by BERT (Devlin et al., 2019) for contextual representation. We set the window size as 5 empirically based on the observation of results.

Classification There are several studies have addressed the decision utterance detection as a classification. Fernández et al. (2008) defined the decision-making sub-dialogues as being composed of several dialogue act tags such as the introduction of issue, decision adopted/proposed/confirmed, agreement. Murray and Renals (2008) created abstract describing decisions, actions and problems of meeting and then associated the utterances used for abstract as the action item utterances. Chen and Hakkani-Tur (2016) classified action items in the token level following the semantic intent schema.

In this study, the task of decision-related utterance extraction was formulated as binary sequence labeling on the sentence level, different from Fernández et al. (2008). This is because we want to keep a simple setting to confirm the efficiency of IUR in actual cases. To take advantage of context, we followed Cohan et al. (2019) to jointly encode consecutive utterances. Preceding utterances leading to decision are essential because followed by Fernández et al. (2008), we hypothesize that the particular kinds of patterns of conversation co-occur with decision. Utterances following decision are also important since affirmative response by others supports the confidence of detection.

For sequence labeling, the model uses the en-

coding of [SEP] tokens corresponding to each utterance and predicts tags (decision or not) by a feedforward network. Different from Cohan et al. (2019), we used only the prediction for the second utterance from the back in the input and slide the window with the stride of 1 over conversation to obtain the predictions for all utterances.

2.2 Decision utterance rewriter

After extracting decision-related utterances, the rewriter translates the extracted utterances from the spoken to written language to improve user experience. We describe the rewriter in two steps: input representation and rewriting.

Input representation The input of DUR comprises of utterances $\{u_1, u_2, ..., u_n\}$ where $u_1, ..., u_{n-1}$ is contextual dialogue and the tail utterance u_n is the decision-related utterance. For input representation, we followed Inoue et al. (2022) to use three types of special tokens, $[X1]$, $[X2]$ and $<\backslash s>$. We inserted $[X1]$ after each utterance in contextual dialogue u_i for $i = 1, .., n-1$, $[X2]$ after decision-related utterance u_n, and $<\backslash s>$ at the tail of whole input as the EOS token. For inference, DUR rewrites only the decision utterances detected by DUD. For each decision utterance, we used preceding utterances, including up to 360 tokens by the T5's tokenizer as the contextual dialogue.

Rewriting JET (Inoue et al., 2022) was adopted and fine-tuned on our dataset for utterance rewriting. JET uses T5 (Raffel et al., 2020) for the picker and writer which were jointly trained for picking important tokens and text generation. The picker picks up important tokens from dialogue context which contribute to rewriting. The two components are jointly optimized by sharing parameters of the T5's encoder, which allows the model to restore omitted information while keeping the capability of abstractive text generation to translate from the spoken to written form with fillers removal.

3 Evaluation

In this section, we first show data annotation for the detector and rewriter, and then describe the settings used for experiments. We finally report the results and discussion of the detector and rewriter.

3.1 Dataset

Decision utterance detector Our Japanese dataset was constructed based on multi-party conversations with various users' intents and decisions in real-world business scenarios. We recorded client meetings in a variety of fields, including banking, finance, and insurance, and accurately transcribed all speeches including fillers.

For decision detection annotation, as stated in Section 2.1, we adopted the schema of binary to decide whether an utterance is a decision (labeled by TD) or not (non-TD). With this simple schema, we aimed to extract decision-related utterances with high coverage and relied on rewriter to restore the contextual information involving decision.

To do the annotation, we asked three annotators who have at least N2 Japanese skills to give a label for each utterance whether it is a decision utterance or not. N2 Japanese members are those who have ability to understand Japanese used in everyday situations and in a variety of circumstances to a certain degree.[1] We combined three annotators to create three groups in which each group has two annotators. To reduce resources and avoid specific bias, each group was assigned a small part of the dataset for annotation. To maintain label quality, annotators prepared a list of the specific expressions frequently used in decision utterances such as "I decided to...", "I have to..." and shared it between them. It comes from the observation that utterances containing the specific expression tend to be decision-related utterances. Each utterance was tagged by two annotators and if the tags differed, the final tag was determined after reconsideration. The Cohen Kappa agreement computed over the three groups is 0.672, showing that the agreement is moderate. It is understandable because transcription is quite noisy compared to common data types, e.g., news. The annotated data was divided into training, validation, and testing sets by meeting units and contains 27006, 3030, and 1425 utterances. The dataset is highly imbalanced where decisions only account for 6% of the entire data, creating challenges for classifiers.

Decision utterance rewriter We created the dataset for DUR based on the DUD dataset. We selected 1120 utterances tagged by **TD** and extracted their preceding utterances containing up to 360 tokens. Two native Japanese annotators created the rewritten version of decision utterances. Annotators re-wrote decision utterances with three requirements: (i) restore omitted information extracted from preceding utterances, (ii) remove fillers, and (iii) convert from the spoken form to written form.

[1] https://www.jlpt.jp/e/about/levelsummary.html

To prepare a consistent dataset, annotators reused the original words in contextual dialogue for rewriting as much as possible, rather than creating new phrases. Annotators also checked rewriting each other every 100 samples to align the quality.

3.2 Experimental settings

For the detector, we used pretrained BERT (Cohan et al., 2019) (cl-tohoku/bert-base-japanese) and fine-tuned MLP (dimensions 512, 400, 5) by AdamW in 20 epochs with drop-out of 0.2, the batch size of 16, and the learning rate of $5e-5$. For rewriter, we trained JET with pretrained t5-base-japanese T5 by AdamW with weight decay of 0.01 in 70 epochs with the batch size of 6, the leaning rate of $2e-5$, and the beam size of 5. All models were trained on a single Tesla P100 GPU.

3.3 Results and discussion

Decision utterance detector We compared the BERT model with two different task formulations: sequential sentence labeling (SL) and sentence classification (SC). For sequence labeling, we used the same model described in Section 2.1. For sentence classification, we trained the model by using BERT to predict the tag of the second utterance from the back given the input utterances $\{u_1,u_2,...,u_w\}$. It follows input representation in Section 2.1 and uses the [CLS] tokens for binary classification. To deal with the imbalanced dataset, we also tested the model with **back translation** (BT), a technique to augment the data by translating original text data into another language and then back into the original language. We augmented the positive samples[2] by seven times using seven languages.[3]

Table 1: Results of the Decision utterance detector.

Method	Precision	Recall	F1
BERT (SC)	0.32	**0.59**	0.42
BERT (SC) + BT	0.33	0.58	0.42
BERT (SL)	**0.48**	0.55	**0.51**
BERT (SL) + BT	0.44	0.55	0.49

Table 1 shows the performance comparison. As we can observe, sequence labeling (BERT (SL)) without using back translation is the best. BERT (SL) achieves better performance than BERT (SC) in general. This suggests that the knowledge of jointly predicting tags helps to better understand the dependencies between utterances. So it leads to improving the performance. Binary sentence classification does not show high F-scores even though the model uses context by using concatenation. It suggests more sophisticated combinations for improving the performance of binary sentence classification. Back translation does not help to improve the quality of the detector. This is because utterances are quite broken in terms of writing and contain fillers. It suggests other data augmentation methods for conversation.

Decision utterance rewriter For the writing part, we compared JET to T5 (Raffel et al., 2020) and s2s-ft (Bao et al., 2021) due to its efficiency for the IUR task. **T5** uses a text-to-text framework pretrained on data-rich tasks with transformer encoder-decoder. **s2s-ft** applies attention masks with fine-tuning methods for the generation task. We did not report the results of ProphetNet (Qi et al., 2020) and UniLM (Dong et al., 2019) due to no pre-trained models for Japanese; SARG (Huang et al., 2021) and RUN-BERT (Liu et al., 2020) due to its low accuracy for IUR (Inoue et al., 2022).

For evaluation, we followed Pan et al. (2019) to use ROUGE, BLEU and f-scores.[4] All methods used the beam width of 5. To obtain the reliable comparison, we also report the human evaluation by using **Text Flow** and **Understandability** (Kiyoumarsi, 2015). **Text Flow** shows how the rewritten utterance is correct grammatically and easy to understand. **Understandability** shows how much the prediction is similar to reference semantically. Three annotators (who are at least N2 Japanese skills) involved the judgement and each annotator gave a score (1: bad; 2: acceptable; 3: good) to each rewritten utterance. The three evaluators scored for each 190 testing samples and the final scores were calculated by the average of scores from the evaluators.

Results in Tables 2 and 3 show that JET is the best for both automatic and human evaluation. This is because the model was empowered by T5 and the picker, that picks up important tokens for rewriting. T5 is the second best due to the strong pre-trained model for Japanese. s2s-ft does not show competitive performance compared to model with text-to-text pre-training framework.

[2]positive sample refers the consecutive utterances $u_1,...,u_w$ with the decision tags for u_{w-2}.

[3]We used Google Translate API with 7 languages, "vi", "en", "zh-CN", "zh-TW", "fr", "de", "ko"

[4]We used sumeval for ROUGE and BLEU scores (https://github.com/chakki-works/sumeval) and f-scores are based on n-grams with the MeCab tokenizer.

Table 2: Results of Decision utterance rewriter. RG is ROUGE and BL stands for BLEU.

Method	RG-1	RG-2	BL	f1	f2
JET	**56.71**	**36.60**	**25.97**	**36.81**	**21.52**
T5	54.91	35.10	24.48	36.61	21.42
s2s-ft	47.71	29.52	19.91	27.41	15.74

Table 3: Human Evaluation

Method	Text Flow	Understandability
JET	**2.53**	**1.90**
T5	2.41	1.79
s2s-ft	2.16	1.55

Effectiveness of utterance rewriter A human evaluation was conducted to see how the rewriter contributes to improving the quality of decision items. A good rewriter requires (i) to keep the original contents before writing and (ii) to enrich the content by supplementing omitted information. Given the pair of the original decision utterance (ODU) and the rewritten decision utterance (RDU), we defined the scoring criteria in the range of 1 to 5 as the following.

1. RDU completely lost meaning of ODU.
2. RDU somewhat lost meaning of ODU.
3. RDU keep meaning of ODU but no additional information.
4. RDU keep meaning of ODU with a few additional information.
5. RDU keep meaning of ODU with sufficient additional information.

As far as the RDU lost the meaning of ODU, the score would be 1 or 2 even there was any additional information. In accordance with this criteria, we collected the scores from the three evaluators by using the utterances before and after the rewriting on test data from the DUR dataset. These three evaluators are annotators who also worked to construct the RDU dataset (Section 3.1).

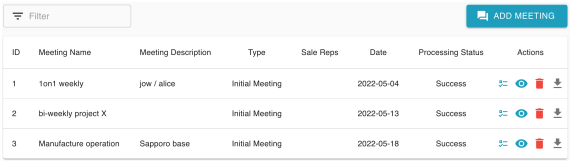
Table 4: Effectiveness evaluation.

score	1	2	3	4	5
ratio	3.76%	7.04%	20.7%	27.7%	40.8%

Table 4 shows the result of evaluation with the ratio of each score. It indicates 10.8% of samples decrease in quality (score $\leq$ 2) while 68.5% of samples increase in quality (score $\geq$ 4). The average score from the evaluators was 3.948, higher than 3, showing that the quality of decision items increases by our rewriter in general. These results show our rewriter certainly contributes to better user experience when displaying decision items (Figure 2b).

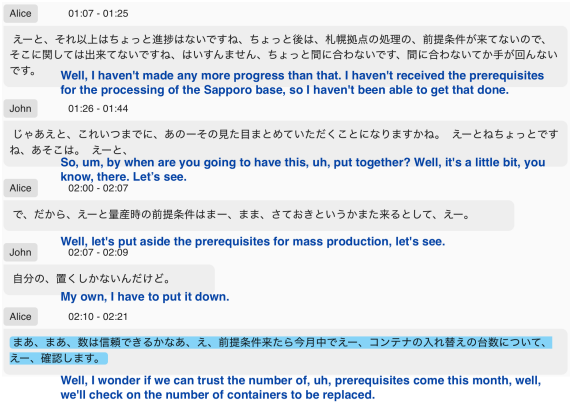
4 Demonstration Scenario

We provide a UI[5] that allows users to look back at decision-making items in past meetings at a glance. Especially in business settings, the accumulation of daily meetings can be compactly stored as itemized decisions to be accessed easily and support project progress management and sharing.



(a) The uploaded meeting list.

(b) The decision items.



(c) The original transcription.

Figure 2: The screenshots from the system.

Figure 2 shows an example processed by our system. The original decision-related utterance is highlighted in blue in Figure 2c. Its content *"Well, I wonder if we can trust the number of, uh, prerequisites come this month, well, we'll check on the number of containers to be replaced."* is rewritten and displayed in the first line of the decision list in Figure 2b as the de-contextualized form, *"Once the prerequisites for processing at the Sapporo site*

[5]The system: https://bit.ly/3sH6193; user and pwd are `Guest123@MDT.com`. Please skip verification when login.

arrive, the number of containers to be replaced should be confirmed..".

The view of the first run is the list of the meetings uploaded (Figure 2a). When users click on the meeting from the list they intend to go back, so decision items for corresponding meeting are unfolded (Figure 2b). Here we display *de-contextualized* decision by DUR instead of original decision-related utterances. Since the DUR module makes decision-related utterances self-contained in the written language format, displayed decision items are straightforward and user-friendly to quickly understand them. To allow users to see the context of the discussion, users can click on the decision item and view the original transcription with a scrolling position where the corresponding decision-related utterance is at the bottom (Figure 2c).

5 Conclusion and Future Work

In this paper, we presented *Meeting Decision Tracker*, a system to automatically itemise the decision-making in daily meetings as well as the tracking of past discussions. We showed the effective adaptation of IUR for decision-item tracking in the context of actual business scenarios. MDT not only displays itemized decision-utterances with an easy-to-understand format, but also allows users to go back and review the contextual dialogue deriving for the decision. Future work will firstly improve the quality of the detector and rewriter. Other potential directions will incorporate ASR into MDT to create an end-to-end system and add functions to remind users of detected decisions and to search for past meetings.

Acknowledgement

We would like to thank Nguyen Duy Anh for the discussion and support of the decision utterance detector. We also thank anonymous reviewers who gave constructive comments on our paper.

References

JinYeong Bak and Alice Oh. 2018. Conversational decision-making model for predicting the king's decision in the annals of the joseon dynasty. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 956–961.

Hangbo Bao, Li Dong, Wenhui Wang, Nan Yang, and Furu Wei. 2021. s2s-ft: Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. *arXiv preprint arXiv:2110.13640*.

Yun-Nung Chen and Dilek Hakkani-Tur. 2016. Aimu: Actionable items for meeting understanding. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 739–743.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163, Columbus, Ohio. Association for Computational Linguistics.

Mengzuo Huang, Feng Li, Wuhe Zou, and Weidong Zhang. 2021. Sarg: A novel semi autoregressive generator for multi-turn incomplete utterance restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14, pp. 13055-13063*.

Shumpei Inoue, Tsungwei Liu, Nguyen Hong Son, and Minh-Tien Nguyen. 2022. Enhance incomplete utterance restoration by joint learning token extraction and text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3149–3158.

Mladen Karan, Prashant Khare, Patrick Healey, and Matthew Purver. 2021. Mitigating topic bias when detecting decisions in dialogue. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 542–547.

Misha Khalman, Yao Zhao, and Mohammad Saleh. 2021. Forumsum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4592–4599.

Farshad Kiyoumarsi. 2015. Evaluation of automatic text summarizations based on human summaries. *Procedia - Social and Behavioral Sciences*, 192:83–91.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing, pp. 2846–2857.

Luca Lugini, Christopher Olshefski, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. Discussion tracker: Supporting teacher learning about students' collaborative argumentation in high school classrooms. In *Conference Proceedings of the 28th International Conference on Computational Linguistics*.

Gabriel Murray and Steve Renals. 2008. Detecting action items in meetings. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 208–213. Springer.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1824-1833*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31.

Sansiri Tarnpradab, Fei Liu, and Kien A Hua. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The Thirtieth International Flairs Conference*.

Dakuo Wang, Ming Tan, Chuang Gan, and Haoyu Wang. 2022. Summarization of group chat threads. US Patent 11,238,236.