Corpus construction for topic-based summarization of multi-party conversation

Yuri Nakayama
Department of Artificial Intelligence
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka, Japan

Tsukasa Shiota
Graduate School of Computer Science
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka, Japan

Kazutaka Shimada
Department of Artificial Intelligence
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka, Japan
shimada@ai.kyutech.ac.jp

2021 International Conference on Asian Language Processing (IALP) | 978-1-6654-8311-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IALP54817.2021.9675229

***Abstract*—In this paper, we report corpus construction and topic-based summarization methods for multi-party conversation. We have already constructed reference summaries and a list of important utterances in each discussion. However, fine-grained summaries about topics in a discussion often are desired in many situations. Therefore, we construct topic-based summaries and propose an important utterance extraction method and two summarization processes using the extracted utterances; extractive and abstractive methods. For the important utterance extraction, we use SVMs with 12 types of features. We use mBART, which is a neural network-based model, as the abstractive method. In the experiment, the extractive method was superior in terms of "accuracy as a summary (relevance)," while the readability of the abstractive method was superior.**

***Index Terms*—Multi-party conversation, Topic-based summarization, Summarization Corpus**

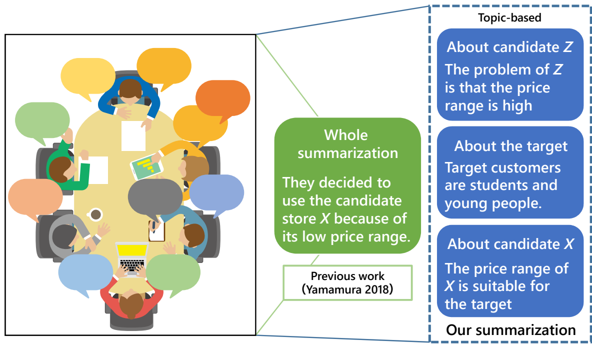
I. Introduction

Meetings are often held in laboratories and companies to exchange their opinions on a topic. For people who could not attend the conversation, a summary of the conversation has an important role in sharing information. However, manual generation of the summary needs much time and human effort. The focus of this paper is automatic summary generation to reduce human effort.

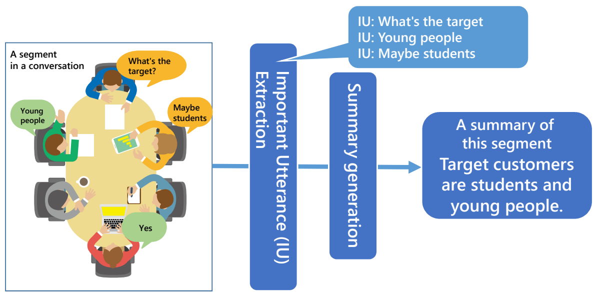
There are many studies about text summarization [1]. One approach to summarization is to extract important sentences/utterances from the target. Here, we focus on two types of issues about the summarization task; the scope of a summary and the method of the summarization.

Firstly, we consider the scope of a summary. Generally, the target of a summary is all utterances in a meeting. The information about the final decision is often contained as a summary. Yamamura et al. [2] have constructed reference summaries from all the utterances in each conversation of a corpus. Then, they have reported an extraction method to obtain the summary of each conversation [3]. However, it is often necessary to summarize not only information about the final decision but also various other information such as the ideas related to the final decision. In this paper, we divide one conversation into several topics. Figure 1(a) shows the difference between our study and the previous study. We construct topic-based reference summaries of the corpus.

The other issue in this paper is the summarization method. Mani [1] has summarized the definition of automatic summarization methods. Automatic summarization methods can be categorized into two types: extractive summarization and abstractive summarization. The merit of extractive summarization is that the summary contains the correct information because the summary can be created using only words that appear in the source text. The disadvantage of extractive summarization is that it tends to produce a non-readable summary because it is just a combination of extracted utterances. On the other hand, one advantage of abstractive summarization is that it can generate readable summaries, in contrast to extractive summarization. Figure 1(b) shows the outline of our method. In this paper, we compare the two types of summarization approaches by using a two-step approach consisting of an important utterance extraction process, namely extractive summarization, and a generation process, namely



(a) Our contribution. In this paper, we construct topic-based summaries and summarization for them.



(b) The outline of our approach. We discuss the summarization approaches.

Fig. 1. An overview of this paper.

abstractive summarization.

Our contributions in this paper are as follows:

- We construct a topic-based summarization dataset about Japanese conversation. In addition, we open the constructed data on the web[1].
- We compare the results about the corpus by using extractive summarization and abstractive summarization approaches.

II. Related work

Document summarization is one of the important tasks in natural language processing [1]. Recently, a number of generative methods based on neural networks have been proposed by [4]. However, the disadvantage is that it tends to convert the words and phrases of the source text. As a result, the content of the source text tends to disappear, as compared with extractive summarization approaches, although the generated summaries are fluent.

In this study, our target is to summarize a multi-party conversation. Xie et al. [5] have shown that features related to the topic of the sentence are effective in the task of extractive summarization in meetings, and proposed a method for resampling sentences based on saliency. Shang et al. [6] have proposed an unsupervised graph-based framework for meeting summarization, which was evaluated on the AMI and ICSI corpora. Li et al. [7] have proposed a meeting summarization method that abstracts from both video and audio and introduced the speaker ’ s attention as multi-modal features and the traditional sentence-related features. Yamamura et al. [8] have proposed an important utterance extraction method for multi-party conversations by using features on temporal information and text segmentation using LCseg [9]. In this study, on the basis of these related papers, we first extract important utterances on each segment grouped by a topic. Then, we generate summaries from the results by applying extractive and abstractive models.

III. Dataset

In this section, we explain the construction of our topic-based summarization dataset. The corpus itself for the construction is an existing Japanese conversation corpus, namely the Kyutech corpus by [2]. First, we describe the bare minimum information about the Kyutech corpus to understand our work in Section III-A. We also define the segments that are the units of the summaries in this study. Next, we describe the construction of reference summaries for each segment in Section III-B. Then, in Section III-C, we explain the reference summary’s mapping to the segment’s utterances (important utterance annotation). Finally, the agreement rates of the reference summaries and annotation results are examined in Section III-D.

[1]http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html#kc3

TABLE I
AN EXAMPLE OF THE KYUTECH CORPUS. THE ORIGINAL TEXT IS WRITTEN IN JAPANESE.

#ID	Main	Option	Utterances with tags
A	CandX		I see/ [なるほど]
C	CandS		and+ [で+]
C	CandS		the restaurant A and (F this) restaurant B (Q)/ [ふうじんと (F この) ボノパスタ (Q)]
A	CandS		yep (うん)/
C	CandS		I am not sure about selecting/ [この二つで迷ってて]
A	CandS		yes/ [うん]
C	CandZ	Vague	and (F umm)(?)+ [で (F えーっと)(?)]
C	CandZ	Vague	so/ [あっそう]
C	CandZ		About the restaurant B+ [結局俺はボノパスタ]
C	CandZ		I think that this is not good/ [これはほんともう三角]
C	CandZ		I think it is not changed/ [あんま変わらん]
A	CandZ		Uh-huh/ [ふんふん]
C	CandY	People	but+ [ただ]

A. Target Corpus

The Kyutech corpus contains nine conversations about a restaurant opening plan in a virtual shopping mall. One conversation consists of utterances by four participants who pretend to be the decision-makers of the virtual shopping mall. An example is shown in Table I. For each transcribed utterance, additional tags for the utterance (F tags for filler and Q tags for questions in Table I) and 28 topic tags (CandX and Vague in Table I) are annotated. The ”+” at the end of each utterance indicates that the utterance leads to the next utterance, and the ”/” indicates the end of one utterance. For more details on the tags, please refer to [2] because it is not our contribution.

This study aims to summarize the utterances in small segments. In this paper, we consider a switch of a main tag as a border of this segment. In Table I, the bold lines indicate the segment borders. There are 418 segments in all 9 conversations in the Kyutech corpus.

B. Topic-based reference summary

We construct topic-based reference summaries from each segment. First, as a pre-processing step, we removed some segments as the noise data. We remove segments of which the main topic tag are the following tags: “Chat,” “Vague,” and “Meeting.” As a result, we obtained 381 segments from 418 segments.

To construct reliable reference summaries, we assign two workers for the reference summary generation process. Two workers read each segment and then write the summary of the segment. The writing rules are as follows:

- A reference summary is created with one or two sentences per segment.
- The content words in the summary must appear in the original utterance in the segment. Do not paraphrase in the summary writing.
- However, we allow using words that do not appear in the segment if the sentence cannot be completed using only

TABLE II
REFERENCE SUMMARIES ON EACH SEGMENT. "NO SUMMARY" DENOTES THAT THE WORKER CANNOT CREATE ANY REFERENCE SUMMARY.

Utterance	Summary
(F Umm) anyway+	Summarize good points about restaurant A and B
yes/	
when comparing restaurant A and B+	
yep/	
(F the)good points about them(Q))/	
I agree/	
and+	
(L let me summarize)/	
No, but, umm+	No summary
junior high students are/	
I'm ungraspable/	
on afternoon/	
(?)/	
(?)/	

the words in the segment, such as the main verb in the summary.

- If he/she cannot create any summary from the segment, "No summary" is allowed.

Some examples of reference summaries are shown in Table II[2]. The reference summaries created by the workers are A1 and A2. We also create merged reference summaries by us from summaries of A1 and A2. If the contents of A1 and A2 are almost the same, the summary of A1 is used as M.

C. Important utterance annotation

The next step is to map the reference summaries of each segment created in Section III-B to the utterances in the segment (important utterance annotation). In other words, for A1, A2, and M in Section III-B, we assign a value (1 or 0) to each utterance; whether it is an utterance related to the summary or not. The annotation rules are as follows:

- Label "1" if it is necessary to create the reference summary, otherwise "0".
- When there are similar utterances, the utterance in the earlier start time has higher priority than that in the latter one.
- If the reference summary is "No summary", all the utterances in the segment are labeled by "0".

Some examples of the annotation are shown in Table III. For example, for the first summary "Summarize good points about restaurant A and B", each worker judges whether each utterance is related to the summary. In this example, a worker judged that three utterances are needed to generate the summary (the labels are "1".)

As a preliminary experiment, we validated this annotation process with a small number of segments. As a result, we confirmed that the agreement between workers was high ($\kappa = 0.71$). Therefore, the data were generated by one worker for each of A1, A2, and M.

[2]As mentioned above, the target data are Japanese conversation. However, we explain our examples by English sentences due to limitations of space.

TABLE III
THE ANNOTATION OF IMPORTANT UTTERANCES FROM EACH REFERENCE SUMMARY.

Summary	Utterance	Label
Summarize good points about restaurant A and B	(F Umm) anyway+	0
	yes/	0
	when comparing restaurant A and B+	1
	yep/	0
	(F the)good points about them(Q))/	1
	I agree/	0
	and+	0
	(L let me summarize)/	1
No summary	no, but, umm+	0
	junior high students are/	0
	I'm ungraspable/	0
	on afternoon/	0
	(?)/	0
	(?)/	0

TABLE IV
THE ROUGE SCORE BETWEEN SUMMARIES.

	A1&A2	A1&M	A2&M
ROUGE-1	0.596	0.921	0.662

D. Dataset analysis

We verify the reference summary created in Section III-B and the important utterance data created in Section III-C. First, we checked whether the merged summaries "M" by the author were suitable as the merged data of A1 and A2. We evaluated the ROUGE-1 [10] score among them.

Table IV shows the result of ROUGE-1 for A1, A2, and M reference summaries. From the ROUGE-1 value between A1 and A2 in Table IV, we can see that A1 and A2 were similar; this Rouge score, 0.596, is usually regarded as "high". We also compared the ROUGE-1 values of A1 and M, and A2 and M. The high ROUGE-1 value between A1 and M indicates that the reference summaries of A1 and M are very similar. The reason why that we regarded A1 as M in the situation that A1 was almost the same as A2. The score between A2 and M was higher than that between A1 and A2. Hence, we conclude that the merged summaries are appropriate.

Next, we compare our dataset with the summary data of the previous work [3]. The data were based on a reference summary written by a human, and then the utterances considered to be related to the reference summary are labeled as important utterances. This process is essentially the same as our process. The difference is the scope of reference summaries: the whole conversation or topic-based segments. The percentage of utterances labeled as "important" is shown in Table V. The percentage of important utterances in the

TABLE V
THE SUMMARIZATION RATIO OF EACH REFERENCE SUMMARY AND THE PREVIOUS WORK.

A1	A2	M	Yamamura [3]
26.90	26.35	27.61	22.98

previous study was the lowest, namely the high summary rate. The result denotes that the number of important utterances in the segment created in this study was 3-4% higher than in the previous study. Our data of this study is a highly comprehensive summary based on the topic, while the previous study was a summary to follow the whole conversation. The result shows that the summary data are different in granularity.

IV. Proposed Method

In this section, we explain our proposed method of important utterance extraction and summarization using the output.

A. Important utterance extraction

We use Support Vector Machine (SVM) [11] for the important utterance extraction. Here we use the annotated data for the merged data M only because they were appropriate as the merged data from A1 and A2.

The features for SVMs are based on the method of Tokunaga et al [12]. 12 types of features were used to vectorize the utterances. We use a morphological analyzer MeCab [3] for the vectorization.

- Utterance length: A feature based on the number of characters in an utterance. However, parentheses and symbols are not counted.
- Presence of frequent words: A binary value (1 or 0) that indicates whether a segment contains or does not contain one of the top 50 nouns, verbs, or adjectives.
- Presence of specific topic tags: a binary value (1 or 0) indicating whether the tag ”Chat” or ”Vague” is included or not.
- Presence of demonstratives: binary (1 or 0).
- Presence of conjunctions: binary (1 or 0).
- Presence of fillers: binary (1 or 0).
- Presence of nouns: binary (1 or 0).
- Presence of verbs: binary (1 or 0).
- Presence of adjectives: binary (1 or 0).
- Previous utterance: A binary value (1 or 0) indicating whether the previous utterance contained or did not contain the (Q) tag.
- Speaker continuance: A binary value (1 or 0) indicating whether the speaker of the current utterance is the same or different from the speaker of the previous utterance. However, the first utterance of a dialogue is always treated as ”not continuance”.
- Position of the utterance: This indicates the position of the target utterance in the segment. For example, the first utterance in the segment is ”1”, and the second one is ”2”.

B. Summarization methods

In this section, we describe two methods for generating summaries using the results of important utterance extraction. First, in Section IV-B1, we explain the extractive method that generates a summary by concatenating important utterances.

[3]http://taku910.github.io/mecab/

TABLE VI
THE OUTPUT BY THE EXTRACTIVE METHOD.

Utterance	Predict	Generated extractive summary
At first+	0	Japanese restaurant is closed so the sales decrease
(D the)+	1	
Japanese restaurant is closed/	1	
yep/	0	
the restaurant is/	0	
so (L closed)+	1	
and, the reason for closed+	0	
the sales decrease +	1	
d/	0	
after the decrease /	0	
the (D under)/	0	
(D under)/	0	
just underperforming /	0	
(L underperforming)/	0	
(L yes)/	0	

TABLE VII
THE OUTPUT BY THE ABSTRACTIVE METHOD

Utterance	Predict	Generated abstractive summary
and+	0	The reason for the closing is that it is not popular with women. One of the reasons for the closing is the lack of menu items.
This is (F so)a common to the restaurant C, not enough about menu items/	1	
I see/	0	
and basically+	0	
the restaurant C is+	0	
(F I mean)+	1	
(F so) the reason for closing is not popular with women+	1	
ya-ya-/	0	
isn’t it?(Q)/	1	
if anything(D this)the restaurant really+	1	
(F so)free refill, so suitable for men(Q)/	1	
yes/	0	

Then, in Section IV-B2, we explain the abstractive method that generates a summary using a neural network based model, mBART.

1) Extractive method: In this method, the important utterances extracted in Section IV-A are concatenated in chronological order by applying the following rules.

- Generate a list of the utterances predicted to be important (”1”).
- Remove any tags such as “(F)” and symbols such as “+” at the end of the utterance.
- Concatenate all the important utterances in chronological order.

An example of this method is shown in Table VI. As described in the previous rule, the predicted label of 1 is extracted, and tags (“(D the)” and “+” at the end) are removed.

2) Abstractive method: We use mBART as the abstractive method. mBART is a neural translation model that is trained by applying BART of Lewis et al.’s [13], to a large corpus of multiple languages. Here, we regard summarization as a machine translation task with a length constraint within a single language. In a similar way to the extractive method, the abstractive method also removes tags from the important

TABLE VIII
HYPER-PARAMETERS ON EACH CROSS-VALIDATION.

Data ID	Parameters		
	C	gamma	kernel
20150313_C1	10	0.0001	rbf
20150320_C1	1	0.0001	rbf
20150320_C4	10	0.0001	rbf
20150323_C3	1	0.0001	rbf
20150326_C1	1	0.0001	rbf
20150326_C2	1	0.0001	rbf
20150326_C4	100	0.0001	rbf
20150327_C2	1	0.0001	rbf
20150327_C3	1	0.0001	rbf

TABLE IX
THE RESULT ABOUT THE IMPORTANT UTTERANCE EXTRACTION.

Accuracy	Precision	Recall	F1
0.708	0.507	0.749	0.596

TABLE X
EVALUATION ON RELEVANCE AND READABILITY.

Method	Ext	Gen	Relevance	Readability
Baseline	No	mBART	1.54	2.28
Extractive	Yes	Concatenation	**2.44**	1.40
Abstractive	Yes	mBART	1.30	**2.32**

utterances. Then we apply concatenated utterances as the input to mBART. In other words, formatting the output of the extractive method by mBART with a length constraint is the abstractive method.

Table VII shows an example of a summary generated by the abstractive method. The output of the summary is fluent as a summary because it is a sentence generated by a neural network model, mBART.

V. EXPERIMENTS

In this section, we describe the experiments of the proposed method described in the previous section. The first experiment is the important utterance extraction by SVMs. The output from SVMs is the input of summarization methods. In the second experiment, we compare the results of two summarization methods and one baseline method.

A. Experiment about important utterance extraction

In this experiment, we evaluated our method with the dataset M by nine-conversation cross-validation although we constructed three datasets (A1, A2, and M) because the dataset M was appropriate as the merged data of A1 and A2. The hyper-parameters were optimized by Grid Search on the training data (Table VIII). Since the number of important and non-important utterances in the dataset was imbalanced, we used a class weight to SVMs: “1”: “0” = 3:1.

Table IX shows the accuracy, precision, recall, and F1 value of the important utterance extraction. The precision rate was not high (0.507). Since this process is, however, the pre-processing of summary generation, a high recall rate is desirable. The recall rate in Table IX shows the effectiveness of our important utterance extraction method.

From the analysis of the results, long utterances and utterances with frequent words tend to be correctly extracted. On the other hand, long utterances consisting of many nods tend to be incorrectly extracted as important utterances. Therefore, we need to consider other features to improve the accuracy. As another approach, the utilization of neural network-based methods such as SummaRuNNer [14] is important future work.

B. Experiment about summary generation

In this section, we evaluate the summary generated in Section IV-B. In order to verify the effectiveness of important utterance extraction as the preprocessing, we prepared a baseline method that generates a summary from all the utterances in a segment by mBART[4].

The mBART for the baseline and the abstractive method was trained on approximately 10,000 three-line summaries[5] from livedoor. For the summarization process, we need to set a limited output length. The limit on the number of output length tokens was set to 100.

The evaluation was conducted by three human evaluators on 100 randomly sampled segments. Each segment was evaluated for each of the three summaries (Baseline, Extractive, and Abstractive). There are two evaluation scales: one is the relevance, and the other is the readability of a summary.

The relevance denotes the facts in the summary consistent with the segment, namely accuracy as a summary. We evaluate the relevance of the generated summaries from three methods on a four-point scale from 0 to 3.

3 : The summary correctly contains the content of the segment.
2 : The summary basically contains the content of the segment.
1 : The summary contains a certain level of the content of the segment.
0 : The summary does not contain the sufficient content of the segment.

We allowed the evaluators to assign equal scores into summaries. For example, it’s possible to contain 3 points for all the summaries from three methods. The readability is a measure of easy-to-read in Japanese. This is a ranking score. The evaluators judged the rankings of three outputs. Here, the 1st rank is 3 points, the 2nd is 2 points, and the 3rd is 1 point.

The average of the three evaluator scores is shown in Table X. “Ext” denotes the presence of the important utterance extraction. “Gen” denotes the method for the summary generation.

[4]We also removed tags and symbols for the baseline.
[5]https://github.com/KodairaTomonori/ThreeLineSummaryDataset

It can be seen from Table X that the extractive method correctly contains the original information. This is due to the fact that the summary of the extractive method consists of words contained in the target segment. In addition, the recall rate of the important utterance extraction in Section V-A was relatively high. On the other hand, abstractive and baseline methods generate a summary. Therefore, the relevance tends to decrease essentially although the readability becomes high. The difference between abstractive and baseline methods is the use of the important utterance extraction as input. Hence, we expected the abstractive method to improve not only the readability but also the relevance, as compared with the baseline. However, the improvement in readability was slight. The reason was the unnaturalness as the input for mBART. The outputs from the important utterance extraction often contained an ill-formed sentence because it was just a concatenation of original utterances that were extracted automatically. The unnatural inputs led to the generation of unnatural summaries.

Overall, the extractive method was superior in terms of the relevance, namely accuracy as a summary, and the abstractive method was superior in terms of the readability. To ensure high relevance and readability, it is necessary to introduce a framework that retains some information of the source text, such as the Pointer-Generator [15].

VI. Conclusions

In this study, we constructed a new annotated corpus for topic-based summarization on multi-party conversations. We also compared two summarization methods: extractive and abstractive. First, we created topic-based reference summaries for the Kyutech corpus. We created two reference summaries by two workers (A1 and A2) and the merged summaries (M). For each of them, we constructed annotated data of important utterances by using the reference summaries.

After that, we built a model for SVM-based important utterance extraction. Two types of summarization methods, extractive and abstractive, were applied to the summary generation from the important utterances. The extractive method was better in terms of the relevance, namely accuracy as a summary, while the abstractive method was better in terms of the readability.

For future work, we need to improve the accuracy of the important utterance extraction. We applied SVMs to the task. It was based on our previous work [12]. However, the study was carried out before the deep learning era. Applying neural network-based approaches is one solution for this future work. The summarization method is also an important factor for the quality of summary generation. Although we used mBART as the method, there are alternatives for the summarization method such as [16]. Comparison with other methods is interesting future work. In addition, although we focused on topic-based summarization based on topic tags, there are many other types of summarization, such as summarization of important utterances by a speaker or by a target in the discussion.

Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research 20K12110.

References

[1] Inderjeet Mani. *Automatic Summarization*. 2001.

[2] Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. The Kyutech Corpus and Topic Segmentation Using a Combined Method. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 95–104, 2016.

[3] Takashi Yamamura and Kazutaka Shimada. Annotation and Analysis of Extractive Summaries for the Kyutech Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3216–3220, 2018.

[4] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, 2015.

[5] Shasha Xie, Yang Liu, Hui Lin. Evaluating the Effectiveness of Features and Sampling in Extractive Meeting Summarization. In *Proceedings of IEEE Spoken Language Technology Workshop*, pages 157–160, 2008.

[6] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorre. Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 664–674, 2018.

[7] Manling Li, Lingyu Zhang, Heng Ji, Richard J. Radke. Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, 2019.

[8] Takashi Yamamura, Yo Tokunaga, and Kazutaka Shimada. Multi-party conversation summarization using time information and text segmentation. In *Technical report of IEICE SIG-NLC*, pages 43–48, 2015.

[9] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse Segmentation of Multi-Party Conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, 2003.

[10] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[11] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[12] Yo Tokunaga and Kazutaka Shimada. Multi-party conversation summarization based on sentence selection using verbal and nonverbalinformation. In *2nd International Conference on Smart Computing and Artificial Intelligence*, pages 464–469, 2014.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and C omprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[14] Ramesh Nallapati, Feifei Zhai, Bowen Zhou. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, pages 3075–3081, 2017.

[15] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083, 2017.

[16] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.