Optimal Summaries for Enabling a Smooth Handover in Chat-Oriented Dialogue

Sanae Yamashita **Ryuichiro Higashinaka**
Graduate School of Informatics, Nagoya University
yamashita.sanae.w7@s.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

Abstract

In dialogue systems, one option for creating a better dialogue experience for the user is to have a human operator take over the dialogue when the system runs into trouble communicating with the user. In this type of handover situation (we call it intervention), it is useful for the operator to have access to the dialogue summary. However, it is not clear exactly what type of summary would be the most useful for a smooth handover. In this study, we investigated the optimal type of summary through experiments in which interlocutors were presented with various summary types during interventions in order to examine their effects. Our findings showed that the best summaries were an abstractive summary plus one utterance immediately before the handover and an extractive summary consisting of five utterances immediately before the handover. From the viewpoint of computational cost, we recommend that extractive summaries consisting of the last five utterances be used.

1 Introduction

Dialogue systems are widely utilized in chatbots and call centers to respond automatically to users (Pappas et al., 2015; Sheehan et al., 2020). However, it is often difficult for such systems to deliver fully autonomous dialogue. To ensure a good dialogue experience, human operators sometimes need to intervene in a dialogue if communication difficulties arise. We call this process handover or intervention and define it as joining a dialogue in the middle to achieve the original objective of the dialogue.

In this study, we investigate which type of summary should be presented to the human operator in an intervention for a smooth handover. Specifically, we conducted a large-scale experiment focused on chat dialogues to investigate the most useful summary for handover among seven types of dialogue summaries consisting of abstractive, extractive, and keyword summaries. Our findings showed that the best summaries were an abstractive summary plus one utterance immediately before the handover and an extractive summary consisting of five utterances immediately before the handover. From the viewpoint of computational cost, we recommend that extractive summaries consisting of the last five utterances be used.

2 Related Work

The handover in dialogues from systems to human operators has been researched extensively in the context of call routing. In call routing, the dialogue is transferred to an appropriate operator and the system hands over the dialogue (Gorin et al., 1997; Walker et al., 2000). However, there has been little research on the actual type of information to be shown to an operator during call routing.

Various frameworks have been proposed in which a semi-autonomous dialogue system performs most of the dialogue and hands over to a human operator when necessary (Glas et al., 2012; Kawahara et al., 2021; Kawasaki and Ogawa, 2021; Kawai et al., 2022). However, it is not clear exactly what type of information or summary would be the most useful for a smooth handover.

Automatic summarization has long been studied (Mani, 2001; Rennard et al., 2022), and various datasets have been released (Carletta et al., 2006; Janin et al., 2003; Zhong et al., 2021b) and are currently in use (Goo and Chen, 2018; Zhong et al., 2021a). Recently, large-scale pre-trained language models have been utilized to generate abstractive summaries (Chen and Yang, 2020; Liu et al., 2021) using a large corpus of summaries (Gliwa et al., 2019; Chen et al., 2021; Liu and Chen, 2021). In this study, we examine what kind of summary is useful for a specific situation: handover.

Proceedings of the AACL-IJCNLP 2022 Student Research Workshop, pages 25–31

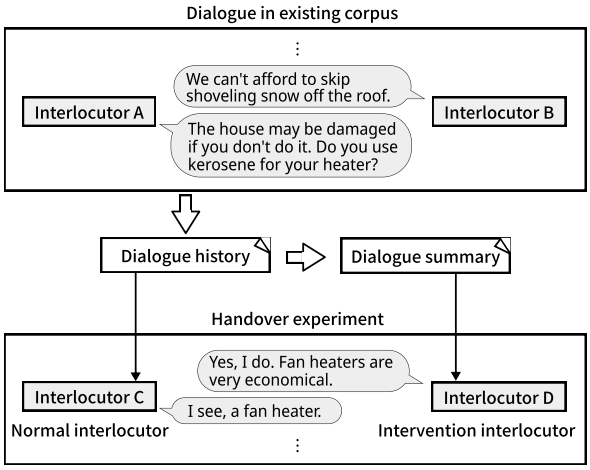

Figure 1: Overview of handover experiment.

3 Approach

To determine the type of summaries needed for a smooth handover in dialogue, our approach is to present a variety of summaries to the operator during the intervention process, examine the smoothness of the dialogue after the intervention, and quantify the effects of each summary type.

To this end, it would be best to collect dialogues in a situation where one interlocutor changes to another in the middle of the dialogue. However, such experimentation would be extremely costly. Therefore, in this paper, we simulate the dialogue handover. Specifically, instead of performing the handover in real-time, we present the dialogue history of an existing dialogue to one of the interlocutors and a summary of that dialogue to the other interlocutor and have them continue the dialogue. We call this experiment a handover experiment.

Figure 1 shows an overview of the handover experiment. As a dialogue history, we prepared a chat dialogue between two interlocutors, A and B, from an existing corpus. The participants in the handover experiment are interlocutors C and D. Interlocutor C is given the dialogue history and interlocutor D is given the dialogue summary created from the dialogue history, and they continue the dialogue on the basis of the information given to each. In this paper, we refer to interlocutor C as the "normal" interlocutor and interlocutor D as the "intervention" interlocutor.

We investigated the usefulness of a summary by utilizing various summary types and analyzing their effects. The questionnaire responses of interlocutors and the number of dialogue breakdowns (Higashinaka et al., 2016) (a state in which dialogue cannot be continued smoothly) after the handover were used for the quantification of the effects. To cover typical summaries, we focused on the following summary types.

- **Abstractive summary**
  An abstractive summary is created by reconstructing important information from a document (Zhong et al., 2021a; Liu and Chen, 2021).

- **Extractive summary**
  An extractive summary is created by extracting important sentences from a document (Nallapati et al., 2016; See et al., 2017).

- **Keyword summary**
  A keyword summary is created by extracting important keywords from a document (Kawahara et al., 2021; Kawasaki and Ogawa, 2021).

4 Handover Experiment

4.1 Existing Corpus

We randomly selected 20 dialogues from an existing chat corpus[1] (Higashinaka et al., 2020). These dialogues are text chats between two people for a total of at least 20 utterances, each of which is about 50 characters in length. In the dialogues, the interlocutors chat freely on any topic of their choice. The dialogues are in Japanese.

4.2 Preparing Dialogue History

We define dialogue history as the past utterance logs from the beginning of the dialogue to a certain point in time. To ensure variations in the progress of the dialogue, we prepared short dialogue histories (from the beginning to the 9th or 11th utterance) and long dialogue histories (from the beginning to the 15th or 17th utterance). For each of the 20 extracted dialogues, we prepared dialogue histories of two different lengths, for a total of 40 dialogue histories.

4.3 Preparing Dialogue Summary

As detailed below, we prepared two types of abstractive summary, three types of extractive summary, and one type of keyword summary. Also, as a control condition, the entire dialogue history was used as one type of summary.

[1] https://github.com/dsbook/dsbook/blob/master/dialogue_data.zip

(i) Abstractive summary (Abs) Abstractive summary manually created from the whole dialogue history. We manually prepared the summaries because we are interested in how the types of summaries affect the handover; if we use automatically generated summaries, we thought that the noise might make it difficult to evaluate the exact effect of this summary type. We recruited 30 workers through crowdsourcing[2] and had them create 40 summaries corresponding to the 40 dialogue histories. The quality of these summaries was verified by a separate crowd sourcing experiment in which we confirmed that the quality was adequate (average rating: 4.4 on a 5-point Likert scale).

(ii) Abstractive summary + last utterance (AbsLast1) Abstractive summary manually created from the dialogue history except for the last utterance plus the last utterance; the last utterance was included to facilitate the handover. Abstractive summaries were created manually in the same way as Abs.

(iii) Keyword + last utterance (KeyLast1) A list of keywords (proper nouns) in the dialogue history plus the last utterance. To extract proper nouns, we used MeCab[3] (version 0.996) with the NEologd dictionary[4] (Release 20200827-01), which covers an extensive amount of proper nouns extracted from the Internet. The last utterance was included to facilitate the handover.

(iv)–(vi) Extractive summary consisting of last few utterances (ExtLast1, ExtLast3, ExtLast5) Extractive summary created by extracting the last one, three, or five utterances immediately before the handover. We utilized a LEAD-like method (Wasson, 1998; Grenander et al., 2019) focusing on the last utterances of the dialogue history, which should contain important information for a handover.

(vii) Dialogue history (control condition) Entire dialogue history as a summary.

Table 1 lists the average number of characters in each dialogue summary. Note that they are in varying lengths; we did not control the lengths of the summary deliberately because we wanted to first verify the types of summary for optimal handover in dialogue.

	No. of characters in summary
(i) Abs	46.8
(ii) AbsLast1	88.9
(iii) KeyLast1	76.4
(iv) ExtLast1	44.7
(v) ExtLast3	130.5
(vi) ExtLast5	209.5
(vii) Dialogue history	485.9

Table 1: Average number of characters in a dialogue summary.

Questionnaire item
Contextual appropriateness
Inconsistency (normal interlocutor only)
Speech style (normal interlocutor only)
Confidence (intervention interlocutor only)
Informativeness
Motivation to utter
Semantic comprehension
Naturalness
Continuity

Table 2: Questionnaire items used in this study.

4.4 Questionnaire

A questionnaire (Table 2) was administered to both the normal and intervention interlocutors to evaluate whether the handover dialogue was a success. The normal interlocutors evaluated the utterance quality of the intervention interlocutors, while the intervention interlocutors evaluated their own utterances, as we were interested in the intervention interlocutors' utterances to better understand the process and difficulty of intervention. We referenced the work of Finch and Choi (2020) here. Specifically, we utilized the questionnaire items focusing on coherence (inconsistency, speech style, contextual appropriateness, and semantic comprehension) and informativeness from their work and added items on the motivation to utter and the confidence of the utterance of the intervention interlocutor. To determine overall dialogue satisfaction, we also added an item for naturalness, which is commonly used in dialogue system evaluations (Hung et al., 2009). We also added an item for continuity, since it is important that an intervention interlocutor be able to continue a dialogue adequately in the handover experiment.

[2] https://crowdworks.jp
[3] https://taku910.github.io/mecab
[4] https://github.com/neologd/mecab-ipadic-neologd

Item	Abs	AbsLast1	KeyLast1	ExtLast1	ExtLast3	ExtLast5	Dialogue history
Contextual appropriateness	4.19	*4.54*	4.29	4.34	*4.54*	**4.66***	4.66
Consistency	4.24	*4.40*	4.16	4.34	4.41	**4.64***	4.38
Speech style	*4.62*	*4.62*	4.59	**4.69**	4.58	4.60	4.62
Informativeness	4.28	4.31	4.28	*4.39*	4.30	**4.47**	4.35
Motivation to utter	4.39	*4.44*	4.39	4.41	**4.46**	4.39	4.40
Semantic comprehension	4.75	*4.81*	4.78	4.74	4.68	**4.89**	4.81
Naturalness	4.51	*4.64*	4.59	4.50	4.61	**4.72**	4.62
Continuity	4.29	*4.51*	4.39	4.47	4.42	**4.60**	4.41

Table 3: Questionnaire results for normal interlocutors. The highest score for each item is shown in bold and the second highest in italics with an underline except for dialogue history. Consistency scores are calculated by 6 – inconsistency score. * denotes a significant difference at the 5% level over Abs or KeyLast1.

Item	Abs	AbsLast1	KeyLast1	ExtLast1	ExtLast3	ExtLast5	Dialogue history
Contextual appropriateness	4.39	4.41	*4.45*	4.18	**4.46**	4.44	4.45
Confidence	*4.25*	4.19	4.15	4.05	**4.30**	4.21	4.32
Informativeness	*3.89*	3.85	3.86	3.65	**3.94**	**3.94**	3.99
Motivation to utter	4.08	3.89	*4.09*	3.98	3.95	**4.15**	4.08
Semantic comprehension	4.72	4.64	4.66	4.69	*4.74*	**4.76**	4.72
Naturalness	**4.59**	4.40	4.42	4.38	4.53	*4.58*	4.47
Continuity	**4.39**	4.29	4.34	4.25	**4.39**	*4.35*	4.41

Table 4: Questionnaire results for intervention interlocutors. The highest score for each item is shown in bold and the second highest in italics with an underline except for dialogue history.

4.5 Conducting Handover Experiment

We combined the seven types of summary and 40 dialogue histories to create a total of 280 dialogue-summary patterns. To cover them, we recruited 280 pairs of interlocutors (560 interlocutors in total) through crowdsourcing and collected a total of 560 dialogues by having each pair conduct a dialogue twice. Eighty dialogues were collected per summary type.

In each pair of participants, one was randomly assigned as a normal interlocutor and the other as an intervention interlocutor. First, participants had sufficient time (three minutes) to read the dialogue history or summary presented on the screen. Each pair then conducted a text chat based on the information presented. The utterances were alternated between the intervention interlocutor and the normal interlocutor, in that order. As we wanted the intervention interlocutors to keep the conversation going for some time, each pair performed a total of 20 utterances after the intervention point. Each pair conducted two dialogues within one hour. After each dialogue, participants indicated their degree of agreement with the questionnaire items (Table 2) on a 5-point Likert scale.

4.6 Questionnaire Results

Table 3 shows the questionnaire results for the normal interlocutors. Overall, AbsLast1 and ExtLast5 had higher scores for all items. The scores for Abs and KeyLast1 tended to be low. We conducted Wilcoxon rank sum tests (with Bonferroni correction) between these two and ExtLast5, which had the highest score, and found a significant difference at the 5% level between ExtLast5 and Abs in terms of contextual appropriateness and between ExtLast5 and KeyLast1 in terms of consistency. These findings indicate that ExtLast5 is the most useful for handover in terms of contextual appropriateness and consistency. No significant differences were found between the other questionnaire items.

Table 4 shows the questionnaire results for the intervention interlocutors. No significant differences were found in any of the questionnaire items. Throughout, the scores for KeyLast1 and ExtLast1 were low. It seems that uttering based on keywords was difficult because it was unclear how the keywords were used in the dialogue history, making it difficult to continue the dialogue. Note that, although dialogue history should show the highest score with no information lost, it was not the case; this was probably because of the high cognitive load needed to comprehend the whole dialogue, although we thought we provided the interlocutors with ample time to read through the materials.

To summarize: the questionnaire results indicate that AbsLast1 and ExtLast5 are useful for handover, while Abs, KeyLast1, and ExtLast1 are unsuitable.

Error type	Abs	AbsLast1	KeyLast1	ExtLast1	ExtLast3	ExtLast5	Dialogue history
Context (102)							
Unclear intention	**7**	0	**7**	0	2	1	0
Topic transition error	**8**	0	2	0	6	0	3
Lack of information	1	0	**3**	0	0	0	0
Self-contradiction	2	3	**6**	0	1	1	0
Contradiction	5	0	5	3	3	2	**6**
Repetition	**16**	0	1	3	3	2	0
Response (49)							
Ignore question	**4**	1	3	0	1	0	1
Ignore expectation	**16**	3	12	0	4	2	2
Utterance (4)							
Grammatical error	0	0	0	0	**2**	0	0
Wrong information	0	0	0	0	**1**	**1**	0
Total	**59**	7	39	6	23	9	12

Table 5: Annotated error types causing dialogue breakdown. The number in parentheses represents the total number for that error scope. The largest number for each error type is shown in bold.

5 Analysis

In this section, we investigate why AbsLast1 and ExtLast5 received the highest scores in the questionnaire. Specifically, we first identified utterances in which dialogue breakdown occurred and classified them according to an existing taxonomy of errors in chat-oriented dialogue systems (Higashinaka et al., 2021), and then clarified the types of errors most common for each summary type. The specific procedures are described below.

5.1 Identification of Failure Utterances

First, for analysis, we took ten dialogue samples of the handover dialogue for each type of dialogue summary, resulting in 70 dialogue samples.

To identify which utterances were causing dialogue breakdowns, we annotated utterances of intervention interlocutors (700 utterances in all) as to whether they presented any discomfort. The annotators were provided with the dialogue history, the dialogue during the intervention, and each sampled utterance and then asked to specify whether they felt uncomfortable with the sampled utterances on a 4-point scale (1: not uncomfortable, 2: slightly uncomfortable, 3: uncomfortable, 4: clearly uncomfortable). Thirty annotators were recruited through crowdsourcing and utterances for which at least half of them (15 or more) responded that they felt at least a little uncomfortable (2 or more on the 4-point scale) were considered problematic. These utterances were determined as the cause of dialogue breakdown.

5.2 Annotation of Error Types

As the taxonomy of errors, we used the taxonomy consisting of 17 error types for chat-oriented dialogue proposed by Higashinaka et al. (2021).

5.3 Analysis of Dialogue Breakdowns

A total of 35 utterances were determined to be dialogue-breakdown-causing. Since dialogue breakdowns are unlikely to occur in human-human dialogue, failure utterances are rare and worthy of analysis. Five annotators were recruited through crowdsourcing and asked to label the error types for these utterances in a multi-labeling manner. The total number of error types annotated for the utterances by the five annotators was 155.

Table 5 lists the number of error types for each dialogue summary type. In terms of the total number of error types, AbsLast1, ExtLast1, and ExtLast5 had the fewest (7, 6, and 9, respectively), indicating that they were non-problematic for the handover. In contrast, there were many errors when Abs, KeyLast1, and ExtLast3 were presented (59, 39, and 23, respectively), indicating that they were unsuitable.

When Abs was presented, there were eight topic transition errors and 16 repetitions. Dialogue summaries other than Abs contained one utterance immediately before the handover, and when those dialogue summaries were presented, there were fewer topic transition errors and repetitions, confirming that presenting the last utterance was helpful.

When KeyLast1 was presented, there were three instances of lack of information and six of self-contradiction. These errors occurred more than twice as often as when the other kinds of summaries were presented. We presume that many of these dialogue breakdowns occurred because the meaning of the keywords was not clear. This suggests the importance of surrounding context (not only a keyword) for sufficiently understanding the

content of dialogue for smooth handover.

6 Conclusion

In this study, we conducted a large-scale experiment to determine which summaries are the most useful for handing over a chat dialogue from seven types of dialogue summary consisting of abstractive, extractive, and keyword summaries. Our findings showed that the best summaries were an abstractive summary plus one utterance immediately before the handover and an extractive summary consisting of five utterances immediately before the handover. From the viewpoint of computational cost, summaries that do not require learning, such as keyword summary and extractive summary, are useful. Considering the results of the questionnaire and the analysis of dialogue breakdowns, we conclude that presenting the extractive summary consisting of the last five utterances is currently the most useful.

As future work, it will be necessary to perform experiments to verify the effects of summary lengths. We also want to perform similar experiments with automatically generated summaries so that we can grasp the utility of abstractive summaries in actual handover situations. In addition, we would like to verify the actual usefulness of the summaries by conducting real-time handover experiments. Although we targeted chat dialogues in this study, useful summaries for dialogues other than chat, such as task-oriented dialogue, should also be investigated.

References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Proceedings of the Second International workshop on Machine Learning for Multimodal Interaction*, pages 28–39.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4106–4118.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245.

Dylan F Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Teleoperation of multiple social robots. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(3):530–544.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop*, pages 735–742.

Allen L Gorin, Giuseppe Riccardi, and Jeremy H Wright. 1997. How may I help you? *Speech communication*, 23(1):113–127.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6019–6024.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3146–3150.

Ryuichiro Higashinaka, Michimasa Inaba, and Masahiro Mizukami. 2020. *Creating a Dialogue System with Python*. Ohmsha. (In Japanese).

Victor Hung, Miguel Elvir, Avelino Gonzalez, and Ronald DeMara. 2009. Towards a method for evaluating naturalness in conversational dialog systems. In *Proceedings of 2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 1236–1241.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–I.

Tatsuya Kawahara, Naoyuki Muramatsu, Kenta Yamamoto, Divesh Lala, and Koji Inoue. 2021. Semi-autonomous avatar enabling unconstrained parallel conversations—seamless hybrid of woz and autonomous dialogue systems—. *Advanced Robotics*, 35(11):657–663.

Haruki Kawai, Yusuke Muraki, Kenta Yamamoto, Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2022. Simultaneous job interview system using multiple semi-autonomous agents. In *Proceedings of the SIGdial 2022 Conference*, pages 107–110.

Kazuyoshi Kawasaki and Kohei Ogawa. 2021. Development of simultaneous summarizing technology of multiple sites context for multiple agent teleoperation system. In *Proceedings of the 36th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 2F1–GS–9–04. (In Japanese).

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2015. Anger detection in call center dialogues. In *Proceedings of 2015 6th IEEE International Conference on Cognitive Infocommunications*, pages 139–144.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2022. Abstractive meeting summarization: A survey. *Computing Research Repository*, arXiv:2208.04163.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115:14–24.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 210–217.

Mark Wasson. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 2, pages 1364–1368.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *Computing Research Repository*, arXiv:2109.02492.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021b. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.