Extractive Dialogue Summarization Without Annotation Based on Distantly Supervised Machine Reading Comprehension in Customer Service

Bing Ma, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao

***Abstract*—Given a long dialogue, the dialogue summarization system aims to obtain a shorter highlight which retains the important information in the original text. For the customer service scenarios, the summaries of most dialogues between an agent and a user focus on several fixed key points, such as users' question, users' purpose, the agent's solution, and so on. Traditional extractive methods are difficult to extract all predefined key points exactly. Furthermore, there is a lack of large-scale and high-quality extractive summarization datasets containing the annotation for key points. Moreover, the speaker's role information is ignored or not fully utilized in previous work. In order to solve the above challenges, we propose a Distant Supervision based Machine Reading Comprehension model for extractive Summarization (DSMRC-S). DSMRC-S transforms the summarization task into the machine reading comprehension problem, to fetch key points from the original text exactly according to the predefined questions. In addition, a distant supervision method is proposed to alleviate the lack of eligible extractive summarization datasets. What's more, a speaker's role token and the solver classification task are proposed to make full use of speaker's role information. We conduct experiments on a real-world summarization dataset collected in customer service scenarios, and the results show that the proposed method outperforms the strong baseline methods by 6 percentage points on ROUGE_L.**

***Index Terms*—Extractive summarization, machine reading comprehension, distant supervision, customer service.**

I. Introduction

In THE customer service scenario, dialogue summarization is becoming more and more attractive to researchers. Given a long dialogue, the summarization system aims to obtain a shorter highlight which retains the important information in the original text. It is of great commercial value. For example, after an agent receives a complaint call, the summary system can provide a short summary that contains salient information. This will help the agent's colleagues get the highlights of the dialogue quickly and improve the efficiency of the subsequent steps. Such a requirement is universal in the customer service scenario. In other words, an excellent text summary system can save a lot of human resources.

Dialogue summary is very close to the text summary. Recently, text summarization methods are mainly divided into generative summarization and extractive summarization. The former encodes the entire document and then generates a summary word by word [1]–[5]. The extractive summarization methods [6]–[10] score semantic units of different granularity and put them together to obtain a summary. Because the generative summarization methods are easy to lose stability [2] and logicality [4], this paper focuses on the extractive summarization methods.

However, in some specific scenarios, there are still some major challenges when using extractive summarization methods:

1) For the customer service scenario, the summaries of most dialogues between an agent and a user focus on several fixed key points [4]. For example,[1] the dialogue in Table I can be summarized into four key points: background, purpose, key, and solution. Traditional extractive methods fetch the fragments containing salient information from the original text, but it is hard for them to extract all predefined key points exactly. Therefore, it is easy to lose important information in the output summary.
2) Furthermore, in the customer services scenario, there is a lack of large-scale and high-quality extractive summarization datasets containing the annotation for key points. It is very time-consuming and laborious to link every key point to the original dialogue. A simple method [11] is to calculate the rouge score between each sentence in the document and the gold summary, then the sentences with top-n scores is returned as the extractive ground-truth sentences. However, It is not practical in real scenarios. A serious concern is that the summary may be just closely related to a phrase in a long utterance, while selecting the whole utterance will introduce a lot of noise.

[1]This sample is translated from Chinese to English manually.

Manuscript received May 9, 2021; revised September 15, 2021 and October 30, 2021; accepted November 25, 2021. Date of publication December 10, 2021; date of current version December 24, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 62071067, 62171057, 62101064 and 62001054, and in part by the Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sakriani Sakti. *(Corresponding authors: Haifeng Sun, Jingyu Wang.)*

The authors are with the State Key Laboratory of Networking and Switching Techonology, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the EBUPT Information Technology Company, Ltd., Beijing 100191, China (e-mail: mabing@ebupt.com; sunhaifeng_1@ebupt.com; wangjingyu@ebupt.com; qiqi@ebupt.com; liaojianxin@ebupt.com).

Digital Object Identifier 10.1109/TASLP.2021.3133206

TABLE I
A REAL-WORLD SAMPLE IN THE CS-SUMMARY CORPUS. FOR THE PREDICTION OF THE FIRST KEY POINT "BACKGROUND," GIVEN A QUESTION "WHAT IS THE BACKGROUND OF THE DIALOGUE," OUR PROPOSED METHOD CAN EXTRACT THE UNDERLINED SEGMENT AS THE ANSWER

Dialogue	
Agent:	Hello, can I help you?
User:	Hello, my **shopping vouchers** have **expired**, but I bought them on your platform yesterday. This is bullying, you have to compensate me for the unused shopping vouchers.
Agent:	Lady, I see your order, the reason why the shopping vouchers are expired is that your membership card is invalid today.
Users:	there was no hint when I bought them, and this is the first time I placed an order on this APP.
Agent:	Hmm, I'd like to explain to you that the prerequisite for the validity of our shopping vouchers is that you are our member. Because you are a new user, we can apply for half of the cost of your unused shopping vouchers.
Users:	It's OK.
Agent:	Thank you for your understanding. Is there anything else I can help you with?
Users:	that is all, thanks.
Key point	**Sub-summary**
background	**User complains about her expired shopping vouchers**
purpose	Compensation for the unused shopping vouchers
key	Shopping vouchers are only valid for members
solution	Compensation for half of the cost of the unused shopping vouchers
gold summary	User complain about her expired shopping vouchers, asking for compensation for the unused shopping vouchers, the key factors of the question is that shopping vouchers are only valid for members, so our solution is compensation for half of the cost of the unused shopping vouchers.
model summary	Shopping vouchers have expired, asking for compensate me for the unused shopping vouchers, the key factor of the question is that the prerequisite for the validity of our shopping vouchers is that you are our member, so our solution is apply for half of the cost of your unused shopping vouchers.

3) Moreover, the speaker's role information is very important for dialogue summarization, which is ignored or not fully utilized in previous work [4], [12], [13]. For example, the utterance spoken by the user usually contains information about the purpose of the dialogue, and the solutions are mostly provided by the agent.

In order to solve the above challenges, we propose a **D**istant **S**upervision based **M**achine **R**eading **C**omprehension model for extractive dialogue **S**ummarization (DSMRC-S).

1) In DSMRC-S, for extracting the content of different key points exactly, we use a typical machine reading comprehension (MRC) method to solve the dialogue summarization task. Specifically, a dialogue (context) and a question are input into the model, and the model extracts the answer from the dialogue (context). Finally, the extracted answer is regarded as a sub-summary. Take Table I as an example, given a predefined question *"What is the background of the dialogue"*, the MRC module fetches a key segment from the dialogue corresponding to the key point *"background"*.
2) For the second challenge, a distant supervision method is used to establish the connection between the key point and the dialogue without annotated samples. Specifically, we label the tokens appearing in the sub-summary as 1 (the red tokens in Table I) and others as 0, then the MRC module is trained to predict the probability of each token appearing in the sub-summary. Moreover, a density-based extraction strategy is proposed to extract the most appropriate span as the sub-summary. Instead of scoring each utterance, DSMRC-S extract spans flexibly not limited to sentence-level, which can avoid a lot of noise.
3) To make full use of speaker's role information, we insert a speaker's role token at the beginning of each utterance, which can not only model the alternation information of speakers like [14] but also incorporate the absolute role information. Besides, we propose a new task named **SO**lver **C**lassification task (SOC) to further enhance the absorption of the speaker's role information. SOC is a binary task to distinguish which speaker is the solver for the dialogue and the current question. Through this task, the model can obtain a priori knowledge: which speaker should be highly concerned for the dialogue and the current question. Moreover, the labels of SOC can be automatically obtained.

To verify the effectiveness of DSMRC-S, we collected a real-world Chinese summary dataset from the logs in a customer service platform, named CS-Summary corpus. Experiments conducted on the CS-Summary corpus indicate that the proposed DSMRC-S performs significantly better than the strong baselines.

The contributions of this paper are as follows:

- DSMRC-S transforms the summarization task as a MRC problem, to fetch the key segment from the original text exactly according to the predefined question.

- A distant supervision method is proposed to alleviate the lack of large-scale and high-quality extractive summarization dataset containing key points. Our method can be trained without annotated samples.
- A method to make full use of speaker's role information, which contains a speaker's role token and the proposed solver classification task.
- Our method outperforms strong baseline methods more than 6 percentage points on ROUGE_L on the CS-Summary corpus.

II. Related Works

A. Generative Summarization

Generative summarization methods encode the entire document and then generate a summary word by word [1]–[5], [15]. The neural encoder-decoder architecture is firstly applied in text summarization [1], [16]. Then a pointer-generator network [2] is proposed to enable the model to copy words from the original text, and a coverage mechanism is used to keep track of words that have been summarized. Celikyilmaz *et al.* divide the task of encoding a long text across multiple collaborating encoder agents [17]. A method that combines standard supervised word prediction and reinforcement learning is proposed in [18]. Gehrmann *et al.* [3] proposed a content selector to improve the copy attention distribution in the pointer-generator network. A Leader-Writer network [4] is proposed to generate more logical and integrated summaries in the customer service scenario. The leader network predicts the key point sequence and then the writer network generates a summary based on the predicted key point sequence by the leader network. A graph-based method [19] is proposed to model the dialogue as an interaction graph according to the topic word information and generate the summary relied on the graph-to-sequence framework and topic words. A saliency-aware topic model [20] is proposed to learn word-saliency correspondences in the dialogue, then a topic-informed attention mechanism is employed to pick out topic-relevant salient information.

B. Extractive Summarization

Extractive summarization methods score semantic units of different granularity and put them together to achieve a summary [6]–[10], [21]. The extractive summarization task is regarded as a sequence labeling problem in [11]. In this method, it is easy to result in redundancy because each sentence is predicted independently. Thus, an auto-regressive decoder [22], [23] is proposed to make the scoring of different sentences affect each other. A reinforcement learning-based system is proposed in [24] to be trained toward optimizing the ROUGE metric. Extractive summarization task is regarded as a latent variable inference problem [25], the latent model maximizes the likelihood of human summaries given selected sentences rather than the gold standard labels. Liu and Titov *et al.* [26] represented documents as multi-root dependency trees where each root node is a summary sentence, and the sentences whose content is covered by the summary sentence are the subtrees. A hierarchical document encoder and a recurrent neural network (RNN) based sentence extractor are proposed for extractive document summarization [9].

Except in sentence level, a discourse-aware neural summarization model [27] is proposed to generate a concise and informative summary on discourse unit level. Zhong and Liu *et al.* [28] formulated the extractive summarization task as a text-matching problem, and proposed a summary-level extractive method based on the powerful ability of semantic matching on BERT. The proposed DSMRC-S is an extractive summarization method, which regards the extractive summarization task as an MRC problem, and a distant supervision method is designed to alleviate the lack of eligible datasets.

C. Dialogue Reading Comprehension

Recently, some methods are proposed to improve the performance of the understanding of multi-turn dialogue. [29] proposed a sequential matching network that matched a response with each utterance and accumulates the matching information in chronological order through RNN, which started the fashion that performing utterance-response matching before aggregation of matching vectors. [30] paid attention to the different importance of each utterance in the context, and considered the relationship among utterances within a context. [31] used self-attention method to capture the intra word-level dependencies in utterance or response, and used cross-attention method to perform the utterance-response matching. [32] realized the deep-level utterance-response matching through 7 stacked interaction blocks. The pre-trained language models are first employed for the understanding of multi-turn dialogue [33]. A speaker-aware BERT is proposed to absorb the alternation information of speakers [14]. In their method, a speaker embedding is employed to add to each token in utterance. But in this paper, the proposed DSMRC-S can not only model the alternation information of speakers but also incorporate the absolute role information. Moreover, the solver classification task is proposed to make full use of speaker's role information.

D. Dialogue Summarization

Compared with document summarization, dialogue summarization is a more difficult problem due to its more complex structure. A simple method for dialogue summarization is to treat the dialogue as a document, but this will lose the information about the speaker and make it difficult to capture the interaction between utterances. Previous work [12] used a hierarchical encoder to catch the interactions. Some work adds auxiliary information to improve the performance of summarization. For example, the dialogue act is used to assist the summarization, where each utterance is assigned a dialogue act to label its effect on the interactions [13]. Liu *et al.* proposed a Leader-Writer network to use key point sequences as auxiliary labels [4].

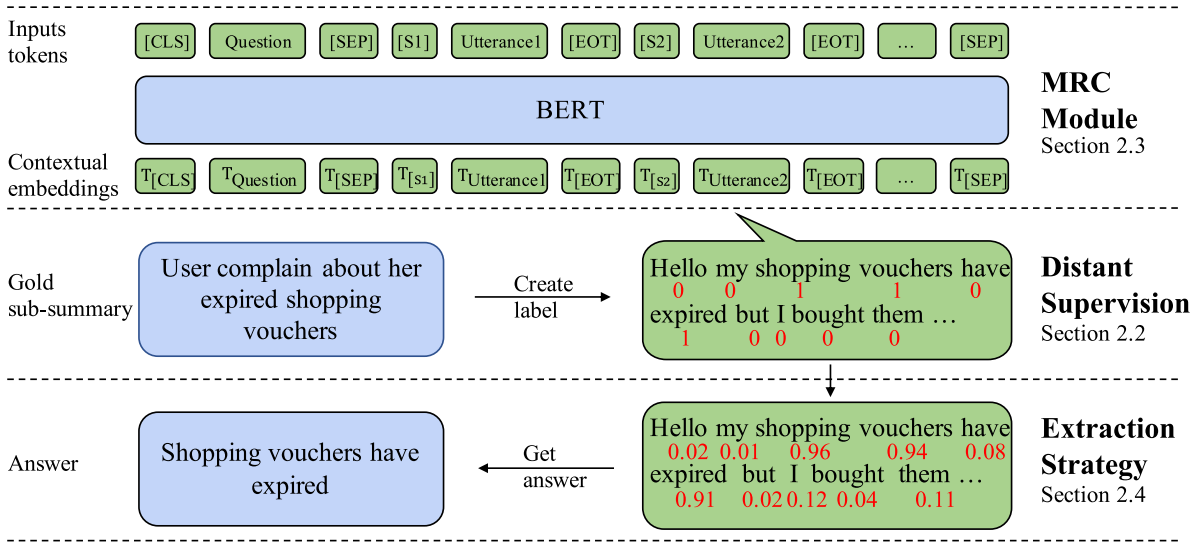


Fig. 1. The overall architecture of DSMRC-S. The key point "background" is taken as an example in this figure.

III. Method

A. Model Overview

DSMRC-S consists of a BERT-based [34] MRC module and a density-based extraction strategy, the overview of DSMRC-S is shown in Fig. 1. In preprocessing, we label the tokens in the dialogue according to each key point. Given a predefined question, the MRC module is trained to predict the probability of each token appearing in the corresponding key point. After that, the most appropriate span is extracted as the answer according to the density-based extraction strategy. In the inference, we repeat the above steps several times to extract all key points. Finally, all extracted key points will be combined into an output summary.

Let $G = \{u_1, u_2, \ldots, u_M\}$ be the dialogue with M utterances, where u_i is the i-th utterance in the dialogue. $S = \{k_1, k_2, \ldots, k_N\}$ is the corresponding summary where k_i is the i-th key point, N is the number of key points. Each key point is written by the agent manually and contains a list of word tokens. Each key point is assigned a corresponding predefined question. For example, the question "*What is the background of the dialogue*" is paired with the "*background*" key point.

B. Distant Supervision

In DSMRC-S, an MRC module is employed to extract key points from the dialogue. However, the key points in the summary are not quoted from the original dialogue text generally. It is difficult to locate the key point to the original text by string-match methods, so a distant supervision method is proposed to solve this problem.

As shown in Fig. 1, given a predefined question "*What is the background of the dialogue*," the answer is the corresponding gold sub-summary, i.e. "*User complain about her expired shopping vouchers*". In this paper, we call the token appearing in the answer as the gold token. Then we label each token in the dialogue according to the principle: **the gold tokens are labelled as 1, and the others are labelled as 0**[2], i.e., the dialogue "... my shopping vouchers have expired but..." is labelled as "... 0 1 1 0 1 0...". The MRC module is required to predict the probability of each token appearing in the sub-summary. Finally, DSMRC-S can extract the segment "*shopping vouchers have expired*" from the dialogue as the answer.

C. MRC Module

In this section, we introduce the MRC module in three parts: Token Classification with BERT, Speaker's role Token, and Solver Classification Task.

1) Token Classification With BERT: As shown in Fig. 1, the general architecture of BERT is employed in the MRC module. Question and dialogue are entered into BERT in pairs. Keeping the original BERT settings, [SEP] token is added at the end of question and dialogue to indicate the boundary, and [CLS] token is inserted at the beginning of the text as an aggregation vector containing the whole sequence information. Dong *et al.* [35] proves that segmentation tokens play an important role in multi-turn response selection, so we add a [EOT] token at the end of a turn to represent the alternation of turns in the dialogue. Each token in the input text is assigned to the sum of three kinds of embeddings: token embeddings, segment embeddings, and position embeddings. The input is composed of the sum of these three embeddings, and then it is fed to a Transformer with multiple layers. Due to the limited space, we omit more description of BERT and recommend readers to refer to [34].

After obtaining the hidden state $\mathbf{h}$ of the last layer, we compute the token classification loss as:

$$\mathbf{p} = softmax(ReLU(W_1\mathbf{h} + b_1)) \tag{1}$$

$$\mathcal{L}_t(\mathbf{y}, \mathbf{p}) = -\sum_j \mathbf{y}_j log \mathbf{p}_j \tag{2}$$

[2]The special tokens such as [CLS], [SEP], etc. are labelled as 0.

where W_1 and b_1 are the parameters updated during training, $\mathbf{y}$ is the labels created in Section III-B.

2) Speaker Role Token: In the customer service scenarios, the speaker's role information is very important. For example, the utterance spoken by the user usually contains information about the purpose of the dialogue, and the solutions are mostly provided by the agent. Therefore, in order to make BERT predict different key points more accurately, it is necessary for the model to perceive the role of the speaker. Furthermore, the alternation of speakers may represent the change of the conversation topic or personal style. A speaker-aware model is more powerful to understand the content of the multi-turn dialogue [14].

Therefore, we add an additional token ([S1] and [S2] in Fig. 1) to represent the speaker's role in front of each utterance, named speaker's role token. This token is used to indicate the role of the speaker of the utterance. In the process of training, the embedding of this token is randomly initialized and shares among all samples.

There are two differences between the proposed speaker's role token and speaker embeddings in SA-BERT [14]:

- Speaker's role token contains absolute role information like the agent or the user, not just model the the alternation of speakers.
- Speaker's role token is inserted at the beginning of each utterance, rather than a vector added to each token of the utterance. In this way, the speaker's role tokens can take interaction with the other tokens without affecting their original meanings.

3) Solver Classification Task: Especially, giving an input consisting of a question and a dialogue as the same as Section III-C1, we obtained the hidden state $\mathbf{h}$ of the last layer in BERT. Then we compute the solver classification loss as:

$$\mathbf{p} = softmax(ReLU(W_2\mathbf{h} + b_2)) \tag{3}$$

$$\mathcal{L}_s(\bar{\mathbf{y}}, \mathbf{p}) = -\sum_{j \in S} \bar{\mathbf{y}}_j log \mathbf{p}_j \tag{4}$$

where W_2 and b_2 are the parameters updated during training, and S is the set containing indexes of all speaker's role tokens. $\bar{y}$ is the labels of SOC, which is obtained automatically according to the following method: the speaker whose utterances contain more gold tokens is the solver for the question. Thus, we count the gold tokens in each speaker's utterances. The speaker with more gold tokens is labelled as 1, and the other is labelled as 0.

There is a similar work in [13], their work also uses some auxiliary signals to obtain better performance of summarization. Here we list the differences between our method and their method: (1) Their method uses the signal of dialogue acts, and requires annotation for each utterance; While our method use the solver information, which can be labelled automatically. (2)Their method is an utterance level classification task, while our method is a token level classification task.

D. Extraction Strategy

As shown in Fig. 1, the extraction strategy is designed to extract answer from the dialogue based on output probability of the MRC module. Given the output probability $\mathbf{p} = \{p_1, p_2, \ldots, p_m\}$, for a span $\{x_{i:i+l}\}$ of the whole tokens sequence $\{x_1, x_2, \ldots, x_m\}$, we define the density of this span as :

$$d_i^{i+l} = (l)^{\alpha-1} \cdot \sum_{k=i}^{k=i+l} p_k \tag{5}$$

where α is a hyper-parameter in [0,1], l is the length of the span $\{x_{i:i+l}\}$. We go through all the possible spans in every utterance, and choose the span with the highest density as the answer. α indicates the preference of the algorithm for a longer span. When $\alpha = 0$, the token with the maximum probability will be returned, and the longest utterance will be selected when $\alpha = 1$. We choose an appropriate value of α in experiments. In this way, the segment like "shopping vouchers have expired" in Table I can be extracted, although not all tokens are labelled as 1 in Section III-B.

E. Training

The training process of DSMRC-S includes two stages: Domain Adaptation and Finetune. The details will be elaborated in the following sections.

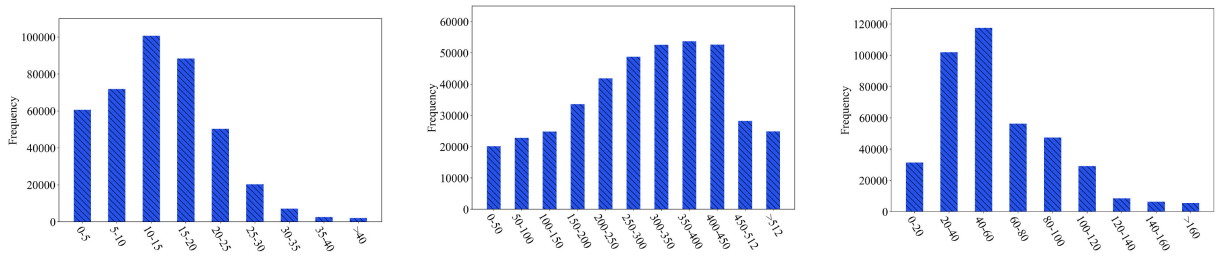
1) Domain Adaptation: The original BERT is trained to learn general knowledge on a large text corpus, and there is a gap between the source domain and the target domain. Therefore, we performed the domain adaptation to alleviate this issue on the target domain dataset. The domain adaptation is performed by optimizing a combination of two loss functions on the target domain: masked language model (MLM) loss, and next sentence prediction (NSP) loss [34].

- *MLM:* We mask 15% of all tokens in each sequence at random follow [34] and let the model restore them to the original tokens. Specifically, among these 15% tokens, 80% of tokens are replaced with [mask], 10% of tokens are replaced with a random token, and 10% of tokens keep unchanged.
- *NSP:* We input a sentence pair like {sentence A, sentence B} to let the model distinguish whether sentence B is the natural next sentence of sentence A. 50% of the time sentence B is the actual next sentence of sentence A and 50% of the time it is a random sentence from the training set. [CLS] is used to compute the binary classification loss. The speaker's role token can be pre-trained in this task.

2) Multi-Task Fine-Tuning: It has been proved by many works [36], [37] that knowledge sharing among tasks can be realized by the joint training of two related tasks in BERT, and then the improvements of both tasks can be achieved. Thus, we combined the token classification loss and the solver classification loss during the fine-tuning stage, the total objective function is:

$$\mathcal{L} = \mathcal{L}_t + \lambda \mathcal{L}_s. \tag{6}$$

Where $\mathcal{L}_t$ is the token classification loss described in Section III-C1, and $\mathcal{L}_s$ is the solver classification loss described in Section III-C3. λ is a trade-off parameter set to 0.1 according to our experimental results.



(a) Turns of dialogue (b) Length of dialogue (c) Length of summary

Fig. 2. (a) demonstrates the distribution of dialogue turns, where x-axis denotes the number of utterances in a dialogue; (b) demonstrates the distribution of dialogue length, where x-axis denotes the number of words in a dialogue; (c) demonstrates the distribution of summary length, where x-axis denotes the number of words in a summary. Y-axis in all subfigures denotes the frequency.

TABLE II
STATISTICS OF THE CS-SUMMARY CORPUS

Dataset	train	valid	test
#dialogue-summary pairs	404k	9k	9k
Avg #words per summay	41.61	42.39	42.97
Avg #turns per dialogue	18.89	18.96	18.97
Avg #words per utterance	23.76	23.81	23.83

IV. EXPERIMENTS

A. Dataset

Due to the lack of the extractive summary dataset containing annotation of key points, we collected a large-scale Chinese dialogue summary dataset from the logs in a customer service platform, called CS-Summary corpus. CS-Summary corpus contains more than 400 k dialogues. Each dialogue describes a conversation that takes place on the phone between customer services agents and the users. The role of the speaker is provided at the beginning of each utterance in the dataset, which is easy to obtain in real scenarios. Dialogue texts are obtained through Automatic Speech Recognition (ASR) [38] technology and careful manual correction. After each conversation, the customer services agent is asked to write four sub-summaries around four fixed key points, i.e., "What is the background of the call," "What is the purpose of the user's call," "What is the key of the question" and "What is your solution". The combination of these sub-summaries is considered as the summary of the dialogue. The statistics of the dataset are shown in Table II.

We demonstrated the distributions of the dialogue turns, dialogue length, and the summary length in Fig. 2. 46% of the dialogues have more than 15 utterances, 27% of the summaries have more than 80 tokens that indicates CS-Summary corpus is challenging.

The data is preprocessed in the same way[3] as in previous work [4]: (1) The samples are normalized firstly. Specifically, the plate number, phone number, trip information and time are replaced with some special symbols such as PHONE, PLATE_NUM, TRIP, TIME, and so on. (2) We detected whether adjacent utterances have the same speaker, if so, these utterances are combined into one utterance. (3) Long dialogues will be truncated until they meet the input length limit of BERT. (4) The validation set and the test set are selected from the corpus randomly.

[3]We didn't truncate the utterance into no more than 75 tokens like in [4], because it is easy to lose important information.

B. Implementation Details and Metrics

We use the base version of BERT to implement our models. The total losses are minimized by an Adam optimizer with the settings $\{\beta_1 = 0.9, \beta_2 = 0.999, lr = 10^{-4}\}$. Model checkpoints are saved and evaluated on the validation set every 1000 steps. The top-3 checkpoints based on evaluation loss are selected, and the averaged results on the test set are reported. The batch size is set to 32, α in (5) is set to 0.4 and λ in (6) is set to 0.1. In the generative methods, the max decode length is set to 50, and the beam search method is used during decoding.

We employed the commonly used evaluation metrics: ROUGE [39] and BLEU [40], which analyze the co-occurrences of n-gram between the output summary and the reference. These two metrics are calculated on the Chinese characters to eliminate the influence of Chinese tokenization. Besides, Distinct [41] is used to measure the diversity of the output summary.

C. Overall Comparisons

Comparison Settings: We performed experiments with following methods:

1) BERTSUMEXT is a BERT-based extractive summarization method according to [42]. Since there is no corresponding label in the dataset, we calculated the ROUGE_L score of the gold sub-summary and each sentence. The sentence with the highest ROUGE_L is labelled as 1.
2) Seq2Seq+Att is a RNN-based Sequence-to-Sequence [43] model with attention mechanisms [44].
3) Seq2Seq+Att+Pointer adds the pointer mechanism [2], which decides whether to generate a token or copy a token from the original text.
4) Transformer+Att+Pointer replaces RNN with Transformer [45] by using only the attention mechanism, enabling the parallel computing models.

TABLE III
MODEL COMPARISON. +ATT MEANS TO USE ATTENTION MECHANISM, +POINTER MEANS TO USE THE POINTER MECHANISM, (W) MEANS TO PREDICT THE WHOLE SUMMARY AT ONE TIME (THE OTHERS PREDICT THE SUB-SUMMARY FIRST, AND THEN COMBINE THEM INTO THE SUMMARY). +SRT MEANS TO INSERT SPEAKER'S ROLE TOKEN AT THE BEGINNING OF EACH UTTERANCE, AND +SOC MEANS TO OPTIMIZE THE COMBINATION OF THE TOKEN CLASSIFICATION LOSS AND THE SOLVER CLASSIFICATION LOSS

Models	BLEU	ROUGE_L	ROUGE_1	ROUGE_2	Dist_1	Dist_2
BERTSUMEXT	9.72	23.21	31.45	15.76	55.23	66.78
Seq2Seq+Att	14.46	33.73	35.87	17.13	49.51	65.92
Seq2Seq+Att+Pointer (w)	14.63	34.11	36.14	17.67	48.23	65.76
TGDGA (w)	15.42	34.87	37.67	18.22	48.56	65.42
Seq2Seq+Att+Pointer	16.03	35.14	37.42	18.55	48.81	65.82
Transformer+Att+Pointer (w)	17.07	36.29	38.31	19.65	51.25	65.93
TDS+SATM (w)	16.21	36.45	38.84	20.46	51.56	66.12
TGDGA	18.32	37.19	38.77	20.64	51.03	65.84
Transformer+Att+Pointer	19.21	39.43	42.74	22.44	51.33	66.34
Leader+Writer	19.37	40.57	42.89	23.41	53.26	67.18
TDS+SATM	20.31	41.13	43.76	25.35	53.78	67.36
DSMRC-S	23.24	44.68	46.69	28.40	57.68	67.54
DSMRC-S+SRT	23.67	45.56	47.23	28.89	57.65	67.43
DSMRC-S+SRT+SOC	**24.32**	**46.83**	**48.28**	**29.85**	**57.72**	**67.61**

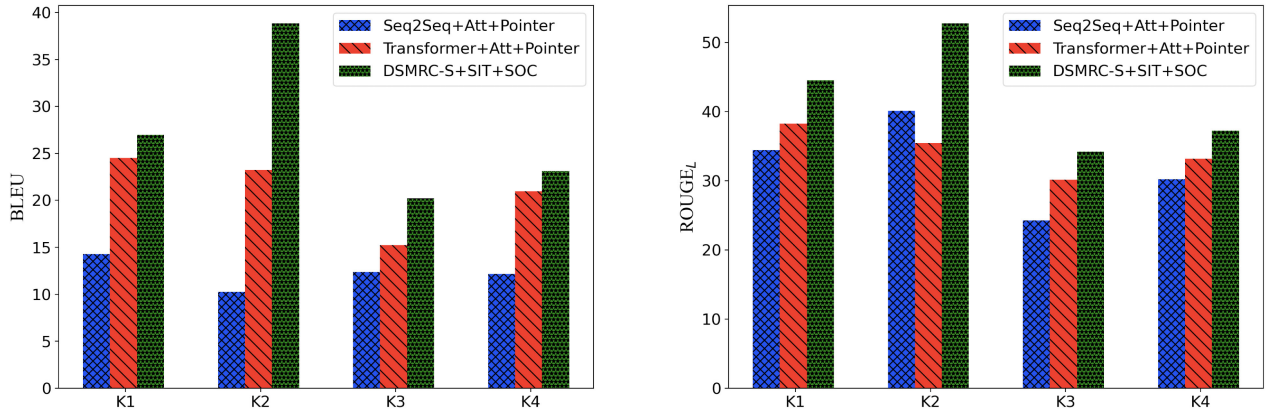
5) Leader+Writer[4] is a hierarchical transformer structure [4]. The Leader net predicts the key point sequence, and the Writer net predicts the summary guided by the prediction of the Leader net.
6) TGDGA constructs the whole dialogue as a graph and generates summaries relied on the graph-to-sequence framework and topic words [19].
7) TDS+SATM[5] proposes a saliency-aware topic model (SATM) to learn word-saliency correspondences in the dialogue. Then a topic-informed attention mechanism is employed to pick out topic-relevant salient information [20].
8) DSMRC-S is the base version of the proposed model in this paper, which solves the summarization task through a pipeline method including token classification with BERT described in Section III-C1 and the extraction strategy described in Section III-D.
9) DSMRC-S+SRT inserts the speaker's role token (SRT) at the beginning of each utterance, to perceive the role information of the speaker, which is described in Section III-C2.
10) DSMRC-S+SRT+SOC performs the multi-task learning by optimizing the combination of the token classification loss and the newly designed solver classification loss described in Section III-C3.

For the fake of fairness, in (2)-(6), we predicted the sub-summary one by one through adding the corresponding predefined question at the end of the dialogue like [46]. Finally, the summary combined with multiple output sub-summaries will be used for comparison.

Comparison Results: The overall results are reported in Table III. These results support the following statements:

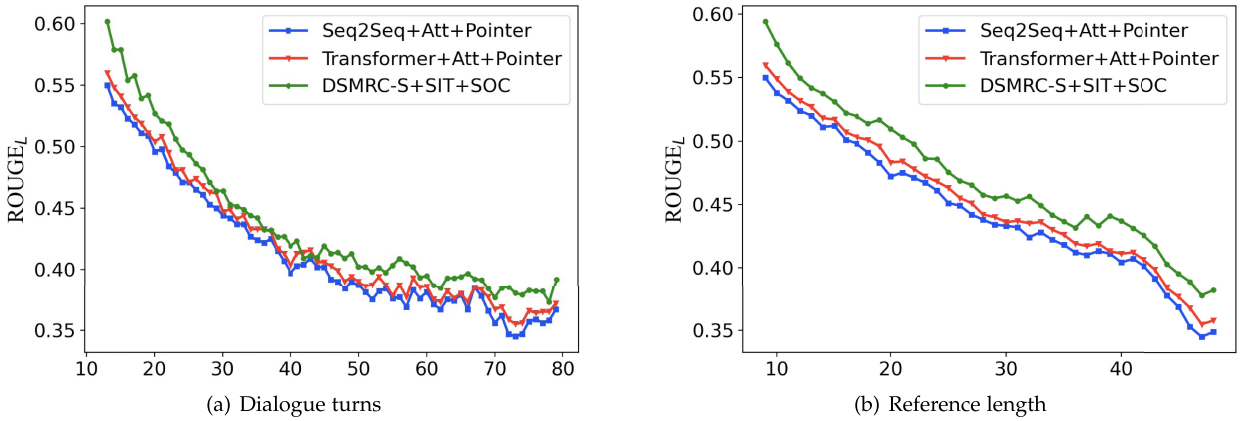
1) Our DSMRC-S+SRT+SOC achieves the highest performance and outperforms the best baseline method by about 6 percentage points on ROUGE_L on the CS-Summary corpus. This improvement may come from two aspects: (a). We realized an extractive summarization approach without the annoatations, based on distantly supervised machine reading comprehension (MRC), and the customer service scenarios are more suitable for the extractive summarization approach. (b). The way of using MRC to solve the summarization problem can take advantage of the popular pre-training language model with strong language understanding ability.
2) Predicting each key point separately is significantly higher than the corresponding version of predicting the whole summary (marked as "(w)" in Table III), Transformer+Att+Pointer outperforms Transformer+Att+Pointer (w) obviously (+3.14% ROUGE_L), which indicates that it is necessary to predict each sub-summary separately in the customer service scenario.
3) The performance of BERTSUMEXT is very poor, because the extractive output in sentence-level will introduce a lot of noise. In other words, the sub-summary is just closely related to a phrase in a long utterance, which is very common in the real customer service scenarios, but BERTSUMEXT outputs the long utterance. Unlike BERTSUMEXT, DSMRC can extract a fragment from a long utterance flexibly even without manual annotation due to the proposed density-based extraction strategy described in Section III-D.
4) Inserting the speaker's role token at the beginning of each utterance is an effective way to incorporate the speaker's role information (+0.88% ROUGE_L). Furthermore, the proposed solver classification task can be used to enhance the speaker's role information (1.27% ROUGE_L).

[4]To make Leader+Writer net comparable with our model, we set the key point sequences to be known.
[5][Online]. Available: https://github.com/RowitZou/topic-dialog-summ



(a) BLEU on each key point (b) ROUGE$_L$ on each key point

Fig. 3. The performance for prediction of different key points. K1, K2, K3 and K4 are the corresponding four key points in the CS-Summary corpus. To be specific, K1: background K2: purpose, K3: key, K4: solution.



(a) Dialogue turns (b) Reference length

Fig. 4. The performance of different dialogue turns and reference length.

5) From the perspective of diversity, compared with generative methods, extractive methods such as BERTSUMEXT and our method have obvious advantages to avoid repetition. However, unlike BERTSUMEXT, which is limited by the inflexible output in sentence-level and manual annotation, our method achieves significantly higher performance than the best baseline (+4.46% Dist$_1$).

D. Performance on Different Key Points

To explore the advantages of our method more clearly, we recorded the performance for the prediction of the different key points. As shown in Fig. 3, DSMRC-S+SRT+SOC outperforms other baseline methods consistently in all key points, which demonstrates that the combination of the MRC model and density-based extraction strategy is better at grasping the characteristics of key points. More specifically, our method performs significantly better in predicting users' purposes, which may be due to the fact that the users' purpose written by the agents is more frequently quoted from the original dialogue text.

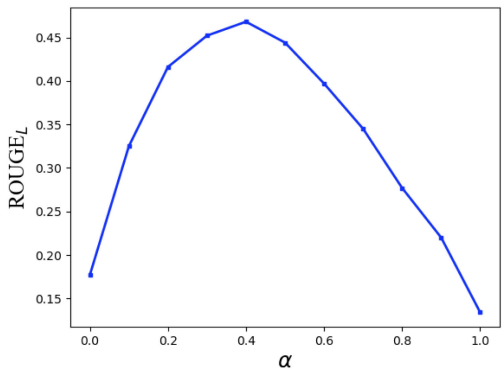


Fig. 5. The performance of the different hyper-parameter α.

E. Performance on Different Dialogue Turns and Reference Length

To explore the effects of dialogue turns and reference length, we recorded the performance for different dialogue turns and reference lengths. As shown in Fig. 4, with the increase of dialogue turns and reference length, the performance of all

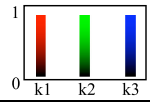
Dialogue	Agent：您好请问有什么可以帮您(Hello, may I help you) User：我点了一个订单已经严重超时了…(I placed an order and it's seriously over time…) Agent：好的，正在帮您处理… (OK, I will handle it for you…)这里看到原因是餐家出餐太慢导致的超时。不好意思，那您现在是想要退款吗？(I see that the reason for the overtime is that the restaurant is too slow to prepare the food. Sorry for it, do you want a refund now?) User：是的，把这个单子退款了(Yes I want the refund of this order) Agent：好的，退款的金额会在一到三个工作日退回您原支付方式 …(OK, the refund will be returned in one to three working days according to your original payment path…)另外因为耽误您用餐，我向平台给您申请了一个十五元的心意红包作为补偿…(In addition, due to the delay in your meal, I applied to the platform for a 15 yuan red envelope as a compensation…) User：嗯好的，谢谢(Well, thank you) 
Gold sub-summary	(k1) background：订单严重超时(The order is seriously over time) (k2) purpose：把这个订单退款(Want the refund of this order) (k3) key：餐家出餐慢导致超时(The food preparation in the restaurant is too slow, resulting in overtime) (k4) solution：给用户退款，补偿心意红包到账户(Refund for the user, and give a red envelope as a compensation)
Transformer +Att+Pointer	background：订单超时订单一个小时送达两个多小时了还没送过来(The order is overtime the order hasn't been delivered for two hours, although it was shown that delivered in one hour) purpose：用户来电退款订单要退款(The user calls to say that he want the refund of the order) key：餐家出餐骑手还没有把餐送过去(The waiter hasn't delivered the food yet) solution：补偿红包红包(Give red envelope as a compensation)
Transformer +Att+Pointer (w)	有(have) 订单(order) 一个(one) 小时(hour) 送达(arrived) 超时(overtime) 订单(order) 退款(refund) 三个(three) 工作日(working days) 退款(refund) 补偿(compensation) 十五元(15 yuan) 心意(heart) 红包(red envelope) 补偿(compensation)
Our method	background：订单已经严重超时(The order is seriously over time) purpose：把这个单子退款(Want to refund this order) key：餐家出餐太慢导致的超时(The food preparation in the restaurant is too slow, resulting in overtime) solution：心意红包作为补偿(A red envelope as a compensation)

Fig. 6. A real sample in the test set of CS-Summary corpus and the prediction results for this sample. The RGB value of each token in the dialogue is used to represent the output probability of the MRC Module for key point "background," "purpose" and "key" respectively. For example, the tokens 订单 (order) in the dialogue is yellow because these two tokens have obtained high probabilities for predicting "background"(Red) and "purpose"(Green).

methods decreases, which indicates an increase in difficulty for summarization. However, our method is consistently better than other baseline methods, which implies that our method has the advantage of regarding the summary problem as an MRC problem for multiple key points in the customer service scenarios.

F. Performance Varies With the Hyper-Parameter α

As the hyper-parameter α in (5) has a great impact on performance, we studied how the performance of DSMRC-S varies with α. Without loss of generality, we used different values of α to conduct the experiment, while other hyper-parameters keep unchanged as in Section IV-B. The performances of different α are reported in Fig. 5.

As shown in Fig. 5, when α is equal to 0, (5) degenerates to the average probability of a span and the model will output a token with the maximum probability; When α is equal to 1, (5) degenerates to the sum of probability of a span and the model will select an utterance with the maximum sum of probabilities (usually a very long utterance). As a result, whether a is equal to 0 or 1, the model gets very poor performance. With the increase of α, the model tends to choose a longer span. Through the experiment, the best performance is obtained when α is equal to 0.4.

G. Case Study

To demonstrate why our method works, we showed a real sample in the test set of CS-Summary corpus in Fig. 6.

The output probability of the MRC module: As shown in Fig. 6, the RGB value of each token in the dialogue is used to represent the output probability of the MRC module for key point "background," "purpose" and "key" respectively. It can be seen that: (1) The probability of most of the tokens is close to 0, while a few of that often appear in the key points have a high probability, that is consistent with the reality. (2) Our method can obtain diverse distribution for different key points, rather than just picking out some important tokens. (3) The method of the MRC module combined with a well-designed extraction strategy can effectively fetch the key segment from the dialogue.

The prediction results: We also showed the prediction results of several baseline methods and our method, some observations are supported by these results in Fig. 6: (1) The Generative summarization methods are prone to grammatical problems, such as the repetition of 订单 (order) and 红包 (red envelope) in the result of Transformer+Att+Pointer. (2) The method Transformer+Att+Pointer(w) outputs a text that is not fluent but contains some important words, which demonstrates the importance of predicting each key point separately.

We further conducted a case study to analyze the causes of bad cases, the results are shown in Table IV. 100 samples are

TABLE IV
100 RANDOMLY SELECTED SAMPLES ARE DIVIDED INTO SIX TYPES. FOR EACH TYPE, WE PROVIDE AN EXAMPLE AND THE PROPORTION

Types	Gold Sub-summary	Predict Result	Percentage
Satisfactory	无法用支付宝支付 (Unable to pay it through Alipay)	用不了支付宝支付 (Can't pay it through Alipay)	28%
Acceptable but hurt performance	退款并告知在3个工作日到账 (Refund and tell him that the money will arrive in 3 working days)	钱将会在3个工作日内原路返回到支付的账户 (The money will be returned to your payment account in 3 working days)	21%
Multiple Spans	骑手提前点击送达，准时宝未赔付 (The deliveryman clicked "delivered" button in advance, and the platform did not pay for overtime)	骑手提前点击送达 (The deliveryman clicked "delivered" button in advance)	9%
Nonexistent Span	安抚用户，告知加急配送 (Appease customer and tell him it will be expedited.)	加急配送 (Expedited delivery)	7%
Error caused by spoken style or ASR	投诉骑手 (Complain about the deliveryman)	投诉这骑手我要投诉这骑手 (Complain about this deliveryman, I want to complain about this deliveryman)	16%
Unacceptable but no exact reason	商家在线未营业，出不了餐 (The restaurant is not open now, and can not provide meals)	商家是有个营业执照 (The restaurant has a business license)	19%

randomly selected from the test set of CS-Summary corpus. We divided all prediction results into six types: (1) **Satisfactory**, where the predicted summary is Satisfactory. (2) **Acceptable but hurt performance**, where the predicted summary is semantically similar with gold summary but the Rouge_L score is less than 0.5. (3) **Multiple Spans**, where a sub-summary is consisit of several spans, but DSMRC-S can only extract one span as the sub-summary. (4) **Nonexistent Span**, where the span in gold sub-summary does not mentioned in the dialogue. (5) **Error caused by spoken style or ASR** (Automatic Speech Recognition), where the errors are caused by ASR errors or the spoken style such as repetition and stutter. (6) **Unacceptable but no exact reason**, where the predict summary is not acceptable but no exact reasons can be found.

Because DSMRC-S is a extractive summarization method, it is impossible to extract what is not mentioned in the original dialogue. Fortunately, in the customer service scenarios, such cases are rare (7% in CS-summary corpus). In other words, extractive summarization may be more suitable for the customer service scenarios than generative summarization. "Error caused by ASR" and "Multiple Spans" are two types of error, which are the space (25% in CS-summary corpus) for our method to improve.

H. Extensibility and Limitations

In this section, we analyzed the extensibility and limitations of DSMRC-S.

Extensibility: Although we verified our method based on a special customer service dataset which contains four predefined key points, our method can be easily extended to other datasets as long as they meet the following requirements: (1) The summary consists of multiple key points; (2) There are some sub-summaries corresponding to key points in the dataset. However, in the customer service scenario, these two conditions are easy satisfied.

Limitations: However, DSMRC-S also has two limitations: (1) Because DSMRC-S is an extractive method, it has no ability to extract what is not mentioned in the dialogue. (2) DSMRC-S only supports extracting one span as a sub-summary currently. In fact, DSMRC-S can easily obtain the top k non-overlapping spans with the highest density, but we failed to find a suitable method to combine multiple spans into a fluent sub-summary. So we left it for future work. In the CS-Summary corpus, the influence of these two limitations is not serious. As shown in Table IV, about 16% cases are affected.

V. CONCLUSION

In the customer service scenario, most dialogues between an agent and a user can be summarized into several key points. Thus, in this paper, we transformed the summarization task as an MRC problem for extracting the content of different key points exactly. In addition, a distant supervision method is proposed to establish the connection between the key sub-summaries and the dialogue without annotated samples. Furthermore, a speaker's role token and the solver classification task are proposed to make full use of speaker's role information. We verified our method in real scenarios, a large-scale Chinese dialogue summary dataset is collected. Extensive experiments are conducted to demonstrate that our method outperforms the best baseline method significantly.

REFERENCES

[1] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389.
[2] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.
[3] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4098–4109.
[4] C. Liu, P. Wang, J. Xu, Z. Li, and J. Ye, "Automatic dialogue summary generation for customer service," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 1957–1965.
[5] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 93–98.

[6] Y. Miao and P. Blunsom, "Language as a latent variable: Discrete generative models for sentence compression," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 319–328.

[7] D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, "Heterogeneous graph neural networks for extractive document summarization," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 6209–6219.

[8] M. Zhong, D. Wang, P. Liu, X. Qiu, and X. Huang, "A closer look at data bias in neural extractive summarization models," 2019, *arXiv: abs/1909.13705*.

[9] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "Neural document summarization by jointly learning to score and select sentences," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 654–663.

[10] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proc. Assoc. Comput. Linguistics*, 2016.

[11] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 3075–3081.

[12] H. Pan, J. Zhou, Z. Zhao, Y. Liu, D. Cai, and M. Yang, "Dial2desc: End-to-end dialogue description generation," 2018, *arXiv:1811.00185*.

[13] C. Goo and Y. Chen, "Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 735–742.

[14] J. Gu *et al.*, "Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 2041–2044.

[15] K. Filippova, E. Alfonseca, C. A. Colmenares, L. Kaiser, and O. Vinyals, "Sentence compression by deletion with lstms," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 360–368.

[16] R. Nallapati, B. Zhou, C. N. dos Santos, Ç. Gülçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNS and beyond," in *CoNLL*, Berlin, Germany, 2016, pp. 280–290.

[17] A. Celikyilmaz, A. Bosselut, X. He, and Y. Choi, "Deep communicating agents for abstractive summarization," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2018, pp. 1662–1675.

[18] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[19] L. Zhao, W. Xu, and J. Guo, "Improving abstractive dialogue summarization with graph structures and topic words," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 437–449.

[20] Y. Zou *et al.*, "Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14 665–14673.

[21] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "A joint sentence scoring and selection framework for neural extractive document summarization," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 671–681, Jan. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8950377

[22] Y. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 675–686.

[23] A. Jadhav and V. Rajan, "Extractive summarization with SWAP-NET: Sentences and words from alternating pointer networks," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 142–151.

[24] S. Narayan, S. B. Cohen, and M. Lapata, "Ranking sentences for extractive summarization with reinforcement learning," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1747–1759.

[25] X. Zhang, M. Lapata, F. Wei, and M. Zhou, "Neural latent extractive document summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 779–784.

[26] Y. Liu, I. Titov, and M. Lapata, "Single document summarization as tree induction," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 1745–1755.

[27] J. Xu, Z. Gan, Y. Cheng, and J. Liu, "Discourse-aware neural extractive text summarization," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 5021–5031.

[28] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive summarization as text matching," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 6197–6208.

[29] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 496–505.

[30] Z. Zhang, J. Li, P. Zhu, H. Zhao, and G. Liu, "Modeling multi-turn conversation with deep utterance aggregation," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 3740–3752.

[31] X. Zhou *et al.*, "Multi-turn response selection for chatbots with deep attention matching network," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 1118–1127.

[32] C. Tao, W. Wu, C. Xu, W. Hu, D. Zhao, and R. Yan, "One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 1–11.

[33] M. Henderson *et al.*, "Training neural response selection for task-oriented dialogue systems," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 5392–5404.

[34] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[35] J. Dong and J. Huang, "Enhance word representation for out-of-vocabulary on ubuntu dialogue corpus," 2018, *arXiv:1802.02614*.

[36] Q. Chen, Z. Zhuo, and W. Wang, "BERT for joint intent classification and slot filling," 2019, *arXiv:1902.10909*.

[37] L. Song, K. Xu, Y. Zhang, J. Chen, and D. Yu, "ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 5429–5434.

[38] S. Chuang, A. H. Liu, T. Sung, and H. Lee, "Improving automatic speech recognition and speech translation via word embedding prediction," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 93–105, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9257188

[39] C.-Y. Lin, "ROUGE: A. package for automatic evaluation of summaries," in *Workshop Text Summarization Branches Out, Post-Conf. Workshop ACL*, Barcelona, Spain, Jul. 2004, pp. 74–81.

[40] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[41] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 110–119.

[42] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proc. Conf. Empirical Methods Natural Lang. Process.-Int. Joint Conf. Natural Lang. Process*, 2019, pp. 3728–3738.

[43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2015.

[45] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[46] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.