Neural Processing Letters (2024) 56:160
https://doi.org/10.1007/s11063-024-11619-8

A Dialogues Summarization Algorithm Based on Multi-task Learning

Haowei Chen[1] · Chen Li[1] · Jiajing Liang[1] · Lihua Tian[1]

Accepted: 6 April 2024

Abstract
With the continuous advancement of social information, the number of texts in the form of dialogue between individuals has exponentially increased. However, it is very challenging to review the previous dialogue content before initiating a new conversation. In view of the above background, a new dialogue summarization algorithm based on multi-task learning is first proposed in the paper. Specifically, Minimum Risk Training is used as the loss function to alleviate the problem of inconsistent goals between the training phase and the testing phase. Then, in order to deal with the problem that the model cannot effectively distinguish gender pronouns, a gender pronoun discrimination auxiliary task based on contrast learning is designed to help the model learn to distinguish different gender pronouns. Finally, an auxiliary task of reducing exposure bias is introduced, which involves incorporating the summary generated during inference into another round of training to reduce the difference between the decoder inputs during the training and testing stages. Experimental results show that our model outperforms strong baselines on three public dialogue summarization datasets: SAMSUM, DialogSum, and CSDS.

Keywords Dialogues summarization · Multi-task · Contrast learning

1 Introduction

With the continuous advancement of social information, online dialogue has become an indispensable way of communication in our lives. Dialogue summarization has become a lively research area in recent years due to the dramatic increase in the number of conversations, which is meaningful for many applications, such as online customer service and medical dialogue [1].

✉ Chen Li
cclidd@xjtu.edu.cn

Haowei Chen
chenhaowei@stu.xjtu.edu.cn

Jiajing Liang
liangjiajing@stu.xjtu.edu.cn

Lihua Tian
lhtian@xjtu.edu.cn

[1] College of Software, Xi'an Jiaotong University, Changan District, Xian 710100, Shanxi, China

Published online: 02 May 2024

Dialogue summarization is a sub-task of text summarization, but dialogue text has its own unique characteristics, such as text with multiple perspectives and informal expression [2], all of which undoubtedly increase the difficulty of the task. The current research work mainly focuses on dialogue interaction modeling, which is suitable for the characteristics of dialogue text. Graph neural network is widely used as the encoder to model the dialog interaction. The dialog corresponds to the node of the graph, and the interaction behavior corresponds to the edge of the graph [3].

In the current summarization model, each word input comes from the real sample during training, but the current input uses the output of the previous word during reasoning, so there will be inconsistencies between the training and reasoning processes, which will make the model ineffective. This phenomenon is called exposure deviation. In order to solve this problem, an auxiliary task is proposed to reduce exposure deviation, which introduces inference into the training stage through contrastive learning [4] to explicitly reduce the score of low-quality summaries to solve this problem.

In addition, by analyzing the abstracts directly generated by BART, we find that 6% of the summarizations have inappropriate gender pronouns, which affects the factual consistency and fluency of the summarizations. Therefore, a gender pronoun discrimination auxiliary task [5] is designed to help the model learn this difference to improve the performance of the algorithm.

Finally, we find that the loss function during summarization task training is the maximum likelihood estimation loss. The goal of this loss function is to train a model to maximize the probability of the input training sample as much as possible. It is a local word-level objective function. The evaluation index of the abstract task is Rouge [6], which is used to compare the overall similarity between the generated abstract and the reference abstract. It is a sentence-level objective function based on the whole. Therefore if we set a likelihood function as the objective function of the training phase, the evaluation metric used in the training phase is not consistent with that used in the training phase and testing phase. This is also an exposure bias. In response to this problem, we propose to use the minimum risk training method as a loss function to alleviate the inconsistency between the goals of the two stages of training and testing.

To sum up, our contributions are as following:

(1) We propose to use Minimum Risk Training as the loss function of the dialogue summary task to alleviate the problem of this inconsistency between the training phase and the test phase goals.
(2) We design an exposure deviation reduction auxiliary task and a gender pro-noun discrimination auxiliary task to help the model generate more reasonable summaries.
(3) We conduct experiments on the large-scale conversation summary data set SAMSum, DialogSum and Chinese customer service dialogue summary data set CSDS, and demonstrate the effectiveness of our proposed method.

2 Related Work

Early work on dialogue summarization focuses more on extraction techniques for meeting summaries than abstract techniques. In terms of meeting summarization, Shang et al. [7]proposed an unsupervised graph-based approach to sentence compression for meeting summaries at AMI and ICSI. Goo and Chen [8] utilize hidden representations

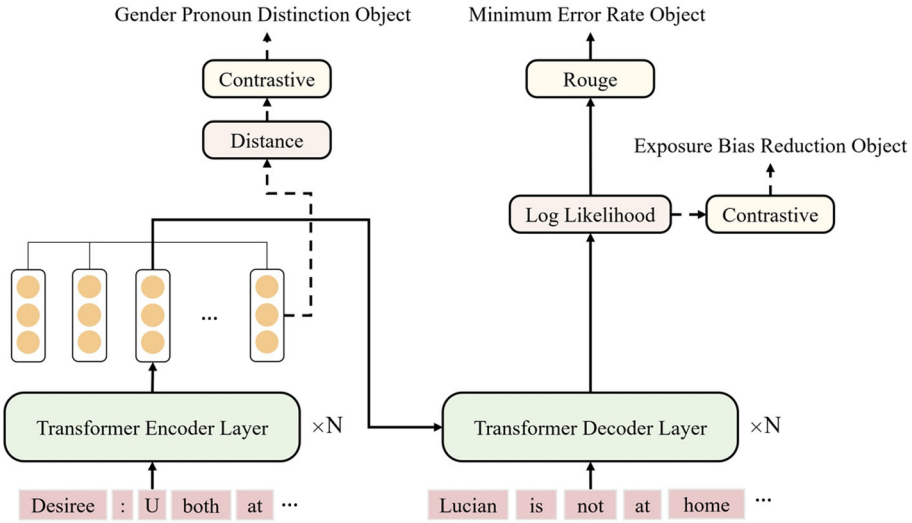

Fig. 1 Model structure with contrastive objectives

from conversation behavior classifiers to guide the abstract decoder through a gated attention mechanism.

Recently, Gliwa et al. [9] proposed SAMSum, which is a benchmark for abstract summaries of everyday dialogues. Zhao et al. [10] has modeled dialogues using a graph structure of words and utterances and generated summaries using a graph-to-sequence architecture. Chen and Yang [11] has proposed a multi-view summarization model in which views can include topics or stages. Zhao et al. [10] has coded speakers as labels into the model through a self-attention mechanism, implicitly modeling the complex relationship between participants and related personal pronouns. Liu et al. [12] has proposed using a co-referencing model to extract co-referencing information in conversations, and incorporating co-referencing information into the network through a graph neural network, enabling the model to better model dialogue context. Kim et al. [13] has introduced common sense information into summary models to help generate summaries. Liang et al. [14] has improved their model by identifying important subtopics and contexts in subtopics to guide the model to generate sound summaries.

These prior researches largely ignore inconsistencies in the training and testing phases of dialogues summarization models, such as inconsistent objective functions and inconsistent decoder inputs. To solve these problems, we propose a new model which introduces auxiliary tasks to help the model better generate summaries through multi-task learning.

3 Method

To generate abstractive and factual summaries from unstructured conversations, we propose first using Minimum Risk Training as a loss function for the dialogues summarization task (Sect. 3.1), then gender pronoun discrimination task (Sect. 3.2) and reducing exposure deviation tasks (Sect. 3.3) help the model better summarize a given conversation. The entire architecture of our method is shown in Fig. 1.

As shown in Fig. 1, the training process of the model is as follows: First, the dialogue is input into the encoder composed of the Transformer coding layer, the vector representation of its high-dimensional vector space is obtained, the distance of each vector in the vector space is calculated, and then the gender pronoun is optimized by the comparison loss function. At the same time, the encoded vector is input into the decoder composed of the Transformer decoding layer to obtain the output sequence. After the likelihood function is calculated, the goal of reducing exposure bias is optimized by comparing the loss function, and Rouge is calculated at the same time to optimize the goal of minimum risk training. Therefore, there are three goals in the training stage of the model, which correspond to the main task of dialogues summarization, the task of distinguishing gender pronouns, and the task of reducing exposure bias, respectively. The following subsections will expand and describe each task.

3.1 Dialogues Summarization Task

The dialogues summarization model uses maximum likelihood estimation as the loss function. The goal of this loss function is to train a model so that the input training sample appears as probable as possible:

$$\hat{\theta}_{MLE} = argmax_{\theta} \left\{ \log \sum_{n=1}^{N} P\left(y^{(n)} \mid x^{(n)}; \theta\right) \right\}. \tag{1}$$

where x denotes the input of model. y denotes the output of model. θ denotes model parameter. $P\left(y^{(n)} \mid x^{(n)}; \theta\right)$ denotes our model. N denotes the number of training samples. n denotes the n-th sample.

However, when measuring the effect of a model, different tasks and users have their own concerns and requirements, so researchers have proposed many evaluation indicators. For example, in machine translation, the main evaluation indicators used are BLEU, METOER, etc.; in information retrieval, there are mainly Recall, Precision, F-Score, etc. In dialogues summarization, Rouge-N and Rouge-L are usually used as evaluation indicators. Rouge-N is the sum of the number of matches between the reference digest and the model-generated digest after being split by N-gram, excluding the reference digest. The sum of the number after splitting by N-gram, Rouge-L is to calculate the coincidence rate of the longest common subsequence. Therefore, if the objective function of the training phase is set as the likelihood function, there will be a problem of inconsistency with the evaluation indicators used in the testing phase.

However, for this problem, Rouge can't be directly used as the training target in the training phase, because the calculation formula of Rouge is non-differentiable, and the gradient cannot be calculated, and then the model parameters can be updated.

This paper proposes to use Minimum Risk Training instead of maximum likelihood estimation to alleviate the problem of inconsistent goals in the training and testing phases. Minimum Risk Training's advantage is that it can introduce the non-differentiable evaluation index into the loss function. Its basic idea is to introduce the risk function $\triangle\left(y, y^{(n)}\right)$, which is the difference between model generation summary y and reference summary $y^{(n)}$. This loss function tries to find a set of parameters to minimize the expected loss of the model on the training set. Its training objective is:

$$\hat{\theta}_{MRT} = argmin_{\theta} \left\{ \log \sum_{n=1}^{N} \sum_{y \in Y\left(x^{(n)}\right)} P\left(y \mid x^{(n)}; \theta\right) \triangle\left(y, y^{(n)}\right) \right\}. \tag{2}$$

Rouge is the evaluation index of the testing phase. In order to avoid the inconsistency between the training objectives and the testing objectives, we chose Rouge as the risk function. Rouge-1 and Rouge-2 are equally important, and an increase in both is an increase in Rouge-L. We give Rouge-1 and Rouge-2 the same weight. In this paper, set $\triangle\left(y, y^{(n)}\right)$ to:

$$\triangle\left(y, y^{(n)}\right) = 1 - \frac{Rouge1 + Rouge2}{2}. \tag{3}$$

where Rouge1 denotes a score based on rouge-1. Rouge2 denotes a score based on rouge-2.

Therefore, the loss function of the dialogues summarization task is:

$$Loss_{MRT} = \log \sum_{n=1}^{N} \sum_{y \in Y\left(x^{(n)}\right)} \frac{P\left(y|x^{(n)};\theta\right)(2-Rouge1+Rouge2)}{2}. \tag{4}$$

where $x^{(n)}$ denotes the input. $Y\left(x^{(n)}\right)$ denotes the search space which represents the set of summaries of all possible outputs corresponding to input x.

3.2 Gender Pronoun Distinction Task

We analyze the abstracts generated by the BART model and found that 6% of the abstracts has inappropriate gender pronouns, which affected the factual consistency and fluency of the abstracts. Gender pronouns are particularly challenging in dialogue summarization. In dialogue, different gender pronouns have almost the same semantic meaning and may refer to many different characters. In complex conversations, the model may have difficulty distinguishing gender pronouns. On the other hand, gender pronouns usually refer to the person or object mentioned earlier. However, it is difficult to determine the pronunciation of gender pronouns when the context is ambiguous or the information about the previously mentioned characters is incomplete. Sometimes in conversation, some gender pronouns may be more commonly used to refer to a certain gender, but this is not always true. This stereotype also makes the task of dialogue summaries difficult. Therefore, this algorithm designs a gender pronoun distinction task based on contrastive learning to help the model learn the difference between different gender pronouns. The task network is shown in Fig. 2.

Figure 3 shows the reference summary and the summary which is generated by the BART model for the conversation. It can be seen that the gender pronoun used in the reference summary is her and the gender pronoun used in the summary generated by the BART model is his. For this conversation, the BART model does not correctly recognize the gender of the interlocutor.

Our model uses the original input as the anchor point and uses the vector representation obtained by the repeated-input encoder as a positive example. It makes the positive example semantically identical to the anchor, with slightly different representations. We transform the personal pronouns in the original input into the corresponding personal pronouns (e. g. he becomes she, his becomes her). Use the changed input as a negative example of the gender pronoun discrimination task. Calculate the distance between the anchor point and the positive and negative sample on the vector space, and calculate the gender pronoun distinction task loss after softmax, thereby updating the model parameters, adjust the vector distance, so that the model learns Distinguish between different gender pronouns.

The loss function based on triple loss function:

$$Loss_{gender} = \max\left\{d\left(x, x^{+}\right) - d\left(x, x^{-}\right) + margin, 0\right\} \tag{5}$$

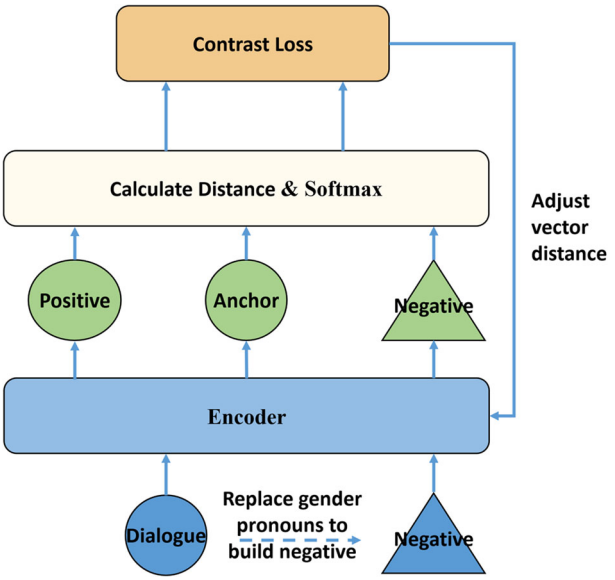

Fig. 2 Gender pronoun distinction task network

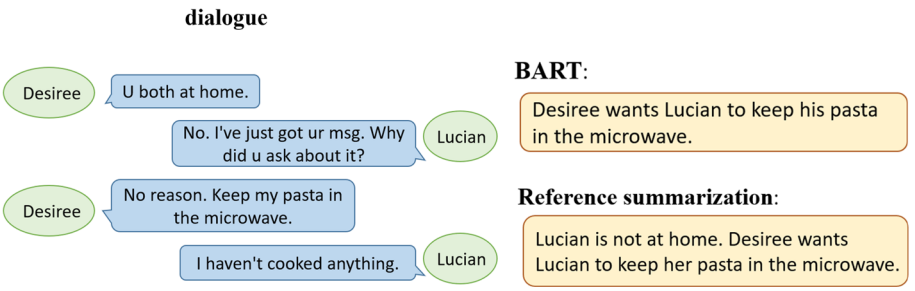

Fig. 3 An example of a summary generated by the BART model and our model

where x denotes an anchor point in contrast learning. Our model uses the original inputs as anchor points. x^+ denotes a positive sample in contrast learning. Our model uses vector representations obtained from repeated input encoders as positive samples. $d\left(x, x^+\right)$ denotes the distance between the anchor point and the positive sample. $d\left(x, x^-\right)$ denotes the distance between the anchor point and the negative sample. $margin$ denotes a constant greater than 1. It controls the distancing of positive and negative samples. The smaller the margin, the closer the positive and negative samples are. The larger the margin, the more distant the positive and negative samples are.

3.3 Reducing Exposure Deviation Task

The dialogues summarization model is guided by the reference summary to generate the abstract during the training phase, and the output of the current step is guided by the input of the previous step, but in the testing phase, there is no reference summary, and the output of the current step is guided by the output of the previous step, so there may be a situation

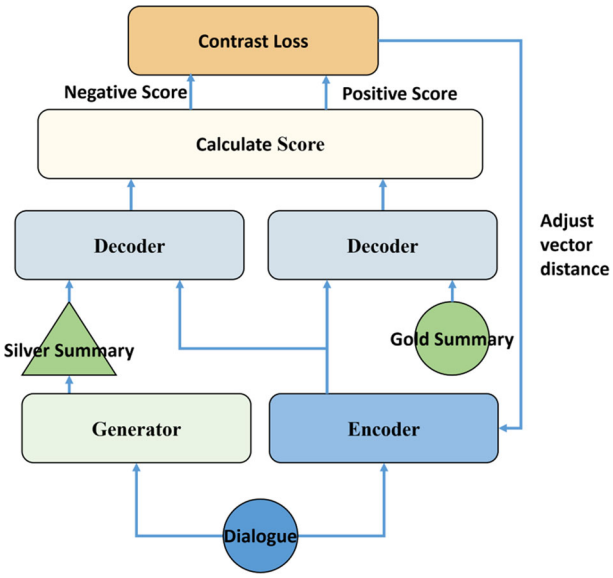

Fig. 4 Reducing exposure deviation task network

where a low-quality candidate summary gets a higher score. Researchers are accustomed to calling the reference summary the gold summary, and in order to distinguish it from the gold summary, the summary generated by the model is called the silver summary. The problem that low-quality silver summaries get higher scores is due to the fact that the model is able to observe gold summaries when training, and when reasoning, it is necessary to evaluate a large number of alternatives that have not been seen before and make choices from them., then it is possible to choose a low-quality candidate abstract, a situation known as exposure deviation.

This paper proposes an auxiliary task based on Contrastive learning, which introduces inference into training through Contrastive learning, expands the state space perceived by the model during the training phase, and explicitly reduces the score of low-quality silver abstracts to alleviate the problem of exposure deviation. Contrastive learning can make high-quality silver abstracts and low-quality silver abstracts more sparse in the vector space, and guide the model selection of high-quality silver abstracts. The task network is shown in Fig. 4.

We refer to the summaries generated by the model based on the inputs as silver summaries. The input is routed through the encoder to get a vector representation. This vector is fed into the decoder along with the gold summary to get the positive case score. This vector is fed into the decoder along with the silver summary to get a negative example score. The formula for calculating the score is as follows:

$$S\left(\hat{Y} \mid X\right) = \frac{1}{m^{\beta}} \sum_{i=1}^{m} f\left(\hat{y}_l \mid X, \widehat{y < l}\right) \tag{6}$$

where $f\left(\hat{y}_l \mid X, \widehat{y < l}\right)$ denotes the probability of the i-th word in the generation sequence. $\widehat{y < l}$ denotes the word produced before the word y. β the penalty coefficient of the sequence length. For the summary generation task, β should be less than 1 to avoid generating redundant information.

Contrastive learning loss function makes the model increase the probability of positive examples as much as possible, and its loss function is similar to the gender pronoun distinction task:

$$Loss_{deviation} = \max\left\{S\left(\hat{Y} \mid X\right) - S\left(Y \mid X\right) + margin, 0\right\} \tag{7}$$

where $S\left(\hat{Y} \mid X\right)$ denotes negative score. $S\left(Y \mid X\right)$ denotes positive score. $margin$ denotes a constant greater than 1.

3.4 Multi-task Learning

There are two options to combine the primary and auxiliary tasks: (1) summing the three objectives as a single one and update the model parameters using the summation loss; (2) alternatively update the model parameters using one of three objectives at each time. The empirical studies (Sect. 4.3) show that the alternating updating strategy performs better. Thus, in this work, we adopt the alternating parameter updating strategy. For a certain batch of dialogue-summary pairs, three objectives are adopted to update parameters in sequence.

The three objectives share the same learning rate α. Since the main focus is to geneate better dialogue summaries with the help of auxiliary contrastive objectives, we give more attentions to the primary task. In order to drive the auxiliary tasks to contribute to the primary one yet not to be dominate, we also introduce task-wise coefficients to each task, denoted as w_{MRT}, w_{gender} and $w_{deviation}$, individually. Following experiments demonstrate the effectiveness of the alternating strategy and the introduced task-wise coefficients.

4 Experiment

4.1 Datasets

SAMSum [9] contains a total of 16,369 dialogue records, and high-quality manual reference summaries are added to each dialogue record by English-proficient linguists. There are 14,732 dialogue-summary pairs for training, 818 and 819 instances for validation and test, respectively.

DialogSum [15] contains multi-round conversations of real scenes collected from three conversation corpora. Unlike the SAMsum data set, its conversations are more focused on the spoken domain than the written chat domain, and its conversations are more long than those in SAMsum conversation.

CSDS [16] is a Chinese customer service dialogues summarization data set, which contains conversations between users and merchants on pre-sales and after-sales topics in real e-commerce scenarios.

4.2 Implementation Details

The Transformer model is adopted as our backbone architecture, implemented using Fairseqtoolkit4. To be specific, our model is initialized with a pre-trained model, i.e., BART. Thus they share the same architectures, 12-layer encoder–decoder Transformer for BARTLARGE. Each layer in BARTLARGE has 16 attention heads, and the hidden size and feed-forward filter size is 1024 and 4096, respectively, resulting in 400M trainable parameters. The dropout

rates for all layers are set to 0.1. The optimizer is Adam with warmup. The learning rate α for SAMSum is $1e-5$, $5e-5$ for DialogSum and CSDS. The maximum number of tokens for a certain batch is 800 for SAMSum, 1000 for DialogSum, and 2000 for CSDS dividually. The w_{gender} and $w_{deviation}$ are set to 175 and 55 dividually.

4.3 Evaluation

To evaluate our models, we utilize the ROUGE [17] to measure the quality of summary output generated by different models. We adopt the files2rouge package based on the official ROUGE-1.5.5.pl perl script to get full-length ROUGE-1, ROUGE-2 and ROUGE-L F-measure scores. For simplicity, we use R-1, R-2 and R-L to refer to ROUGE-1, ROUGE-2 and ROUGE-L, respectively. And we evaluated the model-generated summaries by inputting both the gold summaries and the model-generated summaries into the BARTScore [18] model. There are four ways to use BARTScore based on different generation directions. We have chosen BARTScore-r-h as our evaluation criterion. BARTScore-r-h uses the arithmetic mean of precision and recall and can be widely used to evaluate the semantic overlap between reference and generated texts. The formula for BARTScore-r-h is as follows:

$$BARTScore = \sum_{t=1}^{m} \omega_t \log p(y_t | y_{<t}, x, \theta) \tag{8}$$

where θ is the trainable parameter. x is the source sequence. m is the length of the target token. y is the generation probability conditional on x. Since the generation probability is in the interval [0, 1]. The closer the value of BARTScore is to 0, the higher the quality of the generated summaries.

Baselines PTGen [19] extends sequence-to-sequence model with copy and coverage mechanisms. BART [20] is a pre-trained encoder–decoder Transformer model, with two versions BARTBASE and BARTLARGE. Our method mainly compares with BARTLARGE. Multiview BART [11] divides the dialogue into multiple stages and perspectives, and generates multi-perspective dialogues summarization by fusing dialogue topic and status information. CODS [21] first generates sketches for conversations, extracts intents and key phrases, and then generates summaries based on sketches. SICK [13] introduces common sense information into the summary model to help generate summaries. GLC [14] helps the model by identifying important subtopics and contexts in subtopics to guide the model to generate sound summaries. RAC [22] for role modeling, capture role interactions, generate summaries. ReWriteSum [23] rewrites the discourse, completing the omitted content of the dialog and focusing more on the consistency of the dialog summary with its input dialog context. DialogBART [24] provides a set of sentences that summarize the entire conversation chronologically to generate a summary. Ctrlsum [25] proposes a topic-oriented dialog summary aims to generate a summary that covers the main elements of a given topic in a dialog. LNSDS [26] presents formulas for converting non-conversational summarization datasets into models that can be used for conversational summarization. Summarization models trained with additional data generate summaries that are more similar to golden reference summaries. ATM [27] proposes an effective dual-stream model by combining the temporal and speaker streams of a conversation. Maximize the conversation information and generate more accurate conversation summaries.

Results on SAMSum The results on SAMSum dataset are listed in Table 1. As we can see that, according to ROUGE script our model MultiDigSum significantly outperforms all other models. Rouge-1 and Rouge-2 scores are higher than other methods of comparison,

Table 1 Results on SAMSum test split

Model	R-1	R-2	R-L
PGN [19]	40.08	15.28	36.63
BART [20]	52.79	27.67	43.50
CODS [21]	52.65	27.84	50.79
DialogBART [24]	52.91	28.39	43.90
SICK [13]	53.73	28.81	49.50
GLC [14]	53.74	28.83	49.62
ReWriteSum [23]	54.20	27.10	50.10
Ours	54.18	29.01	45.15

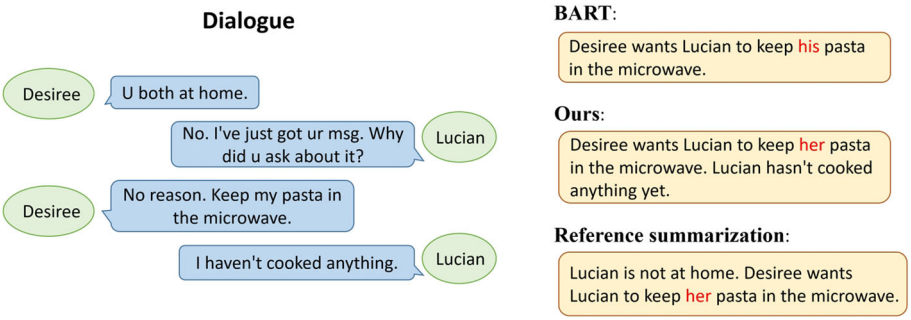

Fig. 5 Abstracts generated by BART and the paper model for the same conversation

and Rouge-L scores are slightly lower than some models. This indicator pays more attention to the similarity between the generated abstract as a whole and the reference abstract, and has sequential requirements. The improvement points in this paper are not explicitly improved for this part, but implicitly improved through other improvements, so the improvement of this indicator is not as large as that of other models.

An example of the abstract generated by BART and our model for the same conversation is shown in Fig. 5. It can be seen that the gender pronouns used in the abstract generated by this model are consistent with the reference abstract, which indicates that the model has learned the differences between gender pronouns and can accurately select the correct gender pronouns for the dialogue abstract, thereby improving the overall quality of abstract generation.

Results on DialogSum The results on DialogSum dataset are listed in Table 2. The results also verify the effectiveness of the method proposed in this paper in the dialogues summarization task. The Rouge-1 and Rouge-2 scores of the model in this paper are higher than those of other methods compared. Rouge-L is second only to the SICK model. The SICK model introduces common sense information. It is more helpful for the model to generate a more reasonable summary as a whole, so the Rouge-L index is higher. We not only show how our model compares to other models on the ROUGE score, but also give an evaluation of our model on the BARTScore. We can see that our model performs well on the BARTScore scores on DialogSum. Our methods can predict words with up to 70% accuracy. Our model has equally good performance under different evaluation metrics.

Results on CSDS The results on CSDS dataset are listed in Table 3. The results show that the Rouge-1 and Rouge-2 scores of our model MultiDigSum are higher than those of other

Table 2 Results on DialogSum test split

Model	R-1	R-2	R-L	BARTScore
ReWriteSum [23]	35.10	14.60	32.10	–
LNSDS [26]	38.85	13.55	31.62	2.43
BART [20]	45.15	19.78	36.57	–
CODS [21]	44.27	17.90	36.98	–
Ctrlsum [25]	45.28	20.47	37.62	–
SICK [13]	46.26	20.95	41.05	–
DialogBART [24]	46.68	21.46	38.32	–
ATM [27]	46.49	21.12	41.56	2.09
Ours	46.38	21.80	38.64	0.20

Table 3 Results on CSDS test split

Model	R-1	R-2	R-L
PGN [19]	50.20	35.12	47.59
BART [20]	53.89	40.24	50.85
RAC [22]	54.14	40.51	52.64
GLC [14]	54.59	40.32	51.10
Ours	56.91	40.54	48.97

Table 4 Ablation Study on SAMSum test split

Model	R-1	R-2	R-L
Gender-2	53.12	27.73	43.65
Bias	53.85	27.98	43.06
Ours-all	54.05	28.56	45.03
Ours-turn	54.18	29.01	45.15

methods of comparison, and the Rouge-L score is slightly lower than that of some models. Therefore, it can be concluded that the model is still applicable to the Chinese dialogues summarization task., and the effect is better than other models compared. The model in is also suitable for the Chinese dialogues summarization task, because the problems solved by the improvement points proposed by the model in this paper also exist in the Chinese dialogues summarization task. For example, there is also misuse of gender pronouns in the Chinese dialogues summarization task. Our model is improved for these problems, so the performance will also be improved.

Ablation Study We verify the contribution of each improvement point to the model by ablating the experimental control variables, using the SAMSum data set as the experimental test data set. The impact of each improvement point on the overall algorithm is shown in Table 4. Among them (1) MRT-1 is the loss function that uses Minimum Risk Training as the dialogues summarization task, the risk function is $1-(Rouge1+Rouge2)/2$, and no auxiliary tasks are introduced; (2) MRT-2 is the loss function that uses Minimum Risk Training as the dialogues summarization task, the risk function is $1-(Rouge1+Rouge2+RougeL)/3$, and no auxiliary tasks are introduced; (3) Gene-1 is the task of introducing gender pronouns based on Contrastive learning, and Minimum Risk Training is not used As the loss function of the dialogues summarization task, the exposure deviation reduction task is also not introduced;

(4) Gene-2 is to introduce a supervised gender pronoun distinction task, which is regarded as a text classification task, and a fully connected layer is added after the encoder to obtain the classification results. Others are the same as (3); (5) Bias is to introduce the exposure deviation reduction task without other changes; (6) Ours-all is to introduce all improvements, and the three goals are combined into one goal, and the sum of losses is used to update the model parameter; (7) Ours-turn To introduce all improvements, use one of the three targets to update the model parameters at a time.

The results show that the introduction of auxiliary tasks and the use of Minimum Risk Training can effectively improve the summarization ability of the model. Comparing the experimental results of MRT-1 and MRT-2, it can be seen that the risk function used by Minimum Risk Training has a certain impact on the indicators. Considering the three indicators, this chapter selects $1-(Rouge1+Rouge2)/2$ as the final risk function. Comparing the experimental results of Gender-1 and Gender-2, it can be seen that the gender pronoun discrimination task based on Contrastive learning improves the main task of dialogues summarization higher than the supervised gender pronoun discrimination task. Comparing the experimental results of Ours-all and Ours-turn, it can be seen that the way parameters are updated in multi-task learning will also affect the final effect of the model to a certain extent. In the model designed in this paper, multiple tasks update model parameters in turn to achieve better results.

5 Conclusion

In this paper, we investigate how to improve the performance of the dialogues summarization model through multi-task learning. We design two auxiliary tasks, namely gender pronoun distinction task and reducing rxposure deviation task, to work in tandem with the main summary task during training. The primary and secondary tasks are coordinated using an alternating parameter update strategy. Experiments on three baseline datasets demonstrate the effectiveness of the proposed model. The current models are mostly autoregressive models, and the speed of generating summaries for dialogues is slow. How to improve the summary generation speed while ensuring the summary generation effect in the future is a challenging task.

Author Contributions HW Chen carried out the design of the experiment, the analysis of the experimental data and the writing of the first draft. C li carried out the preparation of the experimental equipment and the review of the first draft. JJ Liang carried out the writing of the manuscript. LH Tian is responsible for the preparation of the experimental equipment and the supervision of experiment.

Declarations

Competing interests The authors declare no competing interests.

References

1. Liu Z, Chen N (2019) Reading turn by turn: hierarchical attention architecture for spoken dialogue comprehension. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp 5460–5466
2. Gao S, Chen X, Ren Z, Zhao D, Yan R (2020) From standard summarization to new tasks and beyond: summarization with manifold information. arXiv:2005.04684
3. Park S, Lee J (2022) Unsupervised abstractive dialogue summarization with word graphs and POV conversion. arXiv:2205.13108
4. Liu Y, Liu P (2021) Simcls: A simple framework for contrastive learning of abstractive summarization. arXiv:2106.01890
5. Lee S, Behpour S, Eaton E (2021) Sharing less is more: lifelong learning in deep networks with selective layer transfer. In: International conference on machine learning. PMLR, pp 6065–6075
6. Tay W, Joshi A, Zhang XJ, Karimi S, Wan S (2019) Red-faced rouge: examining the suitability of rouge for opinion summary evaluation. In: Proceedings of the The 17th annual workshop of the Australasian language technology association, pp 52–60
7. Shang G, Ding W, Zhang Z, Tixier AJ-P, Meladianos P, Vazirgiannis M, Lorré J-P (2018) Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. arXiv:1805.05271
8. Goo C-W, Chen Y-N (2018) Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In: 2018 IEEE Spoken language technology workshop (SLT). IEEE, pp 735–742
9. Gliwa B, Mochol I, Biesek M, Wawer A (2019) SAMSum corpus: a human-annotated dialogue dataset for abstractive summarization. arXiv:1911.12237
10. Zhao L, Xu W, Guo J (2020) Improving abstractive dialogue summarization with graph structures and topic words. In: Proceedings of the 28th international conference on computational linguistics, pp 437–449
11. Chen J, Yang D (2020) Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. arXiv:2010.01672
12. Liu Z, Shi K, Chen NF (2021) Coreference-aware dialogue summarization. arXiv:2106.08556
13. Kim S, Joo SJ, Chae H, Kim C, Hwang S-w, Yeo J (2022) Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization. arXiv:2209.00930
14. Liang X, Wu S, Cui C, Bai J, Bian C, Li Z (2023) Enhancing dialogue summarization with topic-aware global-and local-level centrality. arXiv:2301.12376
15. Chen Y, Liu Y, Chen L, Zhang Y (2021) DialogSum: a real-life scenario dialogue summarization dataset. arXiv:2105.06762
16. Lin H, Ma L, Zhu J, Xiang L, Zhou Y, Zhang J, Zong C (2021) CSDS: a fine-grained Chinese dataset for customer service dialogue summarization. arXiv:2108.13139
17. Lin C-Y, Och FJ (2004) Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04), pp 605–612
18. Yuan W, Neubig G, Liu P (2021) Bartscore: evaluating generated text as text generation. Adv Neural Inf Process Syst 34:27263–27277
19. See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. arXiv:1704.04368
20. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461
21. Wu C-S, Liu L, Liu W, Stenetorp P, Xiong C (2021) Controllable abstractive dialogue summarization with sketch supervision. arXiv:2105.14064
22. Liang X, Bian C, Wu S, Li Z (2022) Towards modeling role-aware centrality for dialogue summarization. In: Proceedings of the 2nd conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing, pp 43–50
23. Fang Y, Zhang H, Chen H, Ding Z, Long B, Lan Y, Zhou Y (2022) From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 3859–3869
24. Asi A, Wang S, Eisenstadt R, Geckt D, Kuper Y, Mao Y, Ronen R (2022) An end-to-end dialogue summarization system for sales calls. arXiv:2204.12951
25. Lin H, Zhu J, Xiang L, Zhai F, Zhou Y, Zhang J, Zong C (2023) Topic-oriented dialogue summarization. IEEE/ACM Trans Audio Speech Language Process 31:1797–1810

26. Park S, Shin D, Lee J (2022) Leveraging non-dialogue summaries for dialogue summarization. arXiv:2210.09474
27. Xie K, He D, Zhuang J, Lu S, Wang Z (2022) View dialogue in 2D: a two-stream model in time-speaker perspective for dialogue summarization and beyond. In: Proceedings of the 29th international conference on computational linguistics, pp 6075–6088