From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization

Yue Fang [1]*, **Hainan Zhang** [2], **Hongshen Chen** [2], **Zhuoye Ding** [2], **Bo Long** [2], **Yanyan Lan** [2] and **Yanquan Zhou** [1]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]JD.com, Beijing, China
{fangyue, zhouyanquan}@bupt.edu.com,
zhanghainan1990@163.com,ac@chenhongshen.com,
lanyanyan@ict.ac.cn,{dingzhuoye, bo.long}@jd.com

Abstract

Due to the dialogue characteristics of unstructured contexts and multi-parties with first-person perspective, many successful text summarization works have failed when dealing with dialogue summarization. In dialogue summarization task, the input dialogue is usually spoken style with ellipsis and co-references but the output summaries are more formal and complete. Therefore, the dialogue summarization model should be able to complete the ellipsis content and co-reference information and then produce a suitable summary accordingly. However, the current state-of-the-art models pay more attention on the topic or structure of summary, rather than the consistency of dialogue summary with its input dialogue context, which may suffer from the personal and logical inconsistency problem. In this paper, we propose a new model, named ReWriteSum, to tackle this problem. Firstly, an utterance rewriter is conducted to complete the ellipsis content of dialogue content and then obtain the rewriting utterances. Then, the co-reference data augmentation mechanism is utilized to replace the referential person name with its specific name to enhance the personal information. Finally, the rewriting utterances and the co-reference replacement data are used in the standard BART model. Experimental results on both SAMSum and DialSum datasets show that our ReWriteSum significantly outperforms baseline models, in terms of both metric-based and human evaluations. Further analysis on multi-speakers also shows that ReWriteSum can obtain relatively higher improvement with more speakers, validating the correctness and property of ReWriteSum.

1 Introduction

Despite many existing text summarization works on single-speaker written documents, such as news and encyclopedia articles (Rush et al., 2015; Gehrmann et al., 2018), dialogue summarization has gain increasing attention (Zhang et al., 2021). One reason is that it has various promising applications in real world, such as customer services and doctor-patient interaction. More importantly, the dialogue summarization process is more difficult since there are more interactive participants with first-person perspective, and unstructured context to consider (Chen and Yang, 2021), which poses great challenges for researchers in this area.

For this task, it is clear that there is a big gap between the input spoken dialogue and the output formal summaries. That is, in dialogue, users tend to use many incomplete utterances, which al-

Example one	
Ann:	Hi, is the **laptop** still available?
Josh:	Yes **it** is.
Ann:	I can pay 200 dollars.
Josh:	The price is 250 and **it's** non-negotiable.
Ann:	Do you have a bag for **it**? Some other accessories?
Josh:	I have a bag and a small usb mouse.
Ann:	Sounds good, **I'll take it**, where can I pick **it** up?
Ground-Truth: Ann wants to buy Josh's laptop for $200. Josh doesn't want to negotiate the price. Ann will take it for $250 with accessories.	
BART Prediction: Ann will pay 200 dollars for the laptop and the price is non-negotiable. Ann will pick a bag from Josh.	
Example two	
Mike:	Dude, **Wendy** has grown prettier.
Dave:	I know right?
Mike:	Yeah, since she came from Houston, she looks like an angel.
Dave:	I'll have to hit on her soon.
Mike:	Haha, stay off, I hear **Jerry** is her lover.
Dave:	Since when?
Mike:	Haha, I don't know, but you can push your luck.
Dave:	Haha, I will.
Ground-Truth: Mike and Dave notice Wendy got prettier. Dave wants to hit on her, but she's with Jerry. he'll try anyway.	
BART Prediction: Wendy has grown prettier since she came from Houston. Mike will have to hit on her soon. Jerry is Wendy's lover.	

Table 1: Two personal and logical inconsistent examples from the state-of-the-art model in dialogue summarization. Green words in ground-truth indicate the dialogue facts. Red words in BART show the inconsistent content, results from ellipsis and co-reference.

*Work done during internship at JD.com.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3859 - 3869

ways omit or refer back to entities appeared in the history, called ellipsis and co-reference. But the summary is usually formal and written, which contains rich and complete salient information. Here we give two examples, as shown in Table 1. In the first example, the incomplete utterance "I will take it" omits "laptop" which can be seen in the first sentence, while the ground-truth summary contains the complete information "Ann will take it for $250 with accessories". We can see that the generated summary by BART confuses the accessories "bag" with the subject "laptop" and then generate a logic inconsistent summary "pick a bag". And in the second example, many people's names are in the contexts, which are more difficult for the summarization model to distinguish the co-reference relationship, i.e., "I'll have to hit on her" refers to "Dave" via "I". As a result, BART confuses "Mike" with "Dave", and then generates a personal inconsistent summary "Mike will have to hit on her". What's more, such factual inconsistencies have also been observed in previous studies (Cao et al., 2018; Kryściński et al., 2019, 2020). Therefore, it is critical to complete the omission and co-reference information in dialogue utterances for dialogue summarization task.

However, the current models pay more attention on introducing intrinsic information, such as dialogue acts (Goo and Chen, 2018), key point sequence (Liu et al., 2019a) and co-reference information (Liu et al., 2021b). They demonstrate that the introduction of intrinsic information and human annotation is effective in improving the quality of summary generation. However, dialogue acts and key point sequence require a lot of human effort, so they can not be widely used in applications. The co-reference chain is integrated by GNN, which only pays attention to the referencing information of entities but not supplement and restore the referred and omitted pronouns in the dialogue utterances, resulting in the misunderstanding of omitted contents. More importantly, they all ignore the consistency between the dialogue summary and its source dialogue, which may lead to the personal and logical inconsistency problem caused by multi-speakers.

In this paper, we propose a new model, namely ReWriteSum, to tackle this problem. The core idea is to use the utterance rewriting mechanism to complete the omitted content and utilize the data augmentation strategy to enhance the co-reference information. Specifically, we first use the utterance rewriter to complete the ellipsis content in dialogue contexts, and then obtain the rewritten utterances dataset. Then, we use the co-reference data augmentation mechanism to replace the referential person name with its specific name with a certain probability to enhance the personal information. Finally, we use both the rewritten utterances and the co-reference replacement data as input, and utilize the state-of-the-art model BART to generate the corresponding summary.

In our experiments, we use two public datasets to evaluate our proposed models, i.e. SAMSum and DialSum. The results show that ReWriteSum has the ability to produce more consistent and suitable summary than traditional summarization models. Besides, we conduct an analysis on multi-speakers, and the results show that the ReWriteSum obtains relatively higher improvement with more speakers, which indicates that the incomplete utterance rewriting and co-reference data augmentation mechanism by our model are reasonable.

2 Related Work

2.1 Document Summarization

The aim of automatic document summarization is to convert a well-structured document into short text containing salient information. It has received widespread attention in recent literature, especially abstractive document summarization. For example, Rush et al. (2015) introduce an attention-based sequence-to-sequence model for abstractive document summarization. To solve out-of-vocabulary and content repeat issues, See et al. (2017) propose a pointer-generator network with copy and coverage mechanism. Chen and Bansal (2018) leverage reinforcement learning to extract salient sentences in document and then generate summary. Recent studies have focused on the pre-trained models. Liu and Lapata (2019) take use of pre-trained language model BERT (Kenton and Toutanova, 2019) in extractive summarization and abstractive summarization. Lewis et al. (2020) propose BART which combined bi-directional encoder from BERT and auto-regressive decoder from GPT (Radford et al., 2018) to obtain the results of language generation.

2.2 Dialogue Summarization

Compared with document summarization, dialogue summarization aims at generating condensed text from the dialogue contexts among multiple speakers. For instance, Shang et al. (2018) propose an

unsupervised multi-sentence compression method to generate meeting summaries. Zhao et al. (2019) employ a hierarchical encoder and a reinforced decoder based on sequence-to-sequence model to generate meeting summaries.

Some studies have focused on employing conversational analysis for dialogue summarization. Goo and Chen (2018) use sentence-gated mechanism to apply dialogue act in the generation process. Liu et al. (2019a) design a key point sequence as auxiliary information to describe the logic of the abstract. Liu et al. (2019c) and Li et al. (2019) introduce topic information for dialogue summarization. However, their methods need a large amount of human annotation. To avoid this issue, Chen and Yang (2020) use diverse conversational structures like topic segments and conversational stages to design a multi-view summarizer. Recent works often introduce intrinsic information to better model the dialogue process. Liu et al. (2021b) use the graph neural network to employ co-reference information to generate summaries. Feng et al. (2020) introduce the dialogue discourse information, and design Meeting Graph to describe them. Lei et al. (2021) introduce speaker information to improve the generation performance in the context with multi-speakers.

2.3 Incomplete Utterance Rewriting

Incomplete utterance rewriting has received extensive research attention. In question answering, Kumar and Joshi (2016) propose non-sentential utterance resolution based on sequence-to-sequence model for utterance rewriting. To resolve incomplete follow-up questions, retrieval-based sequence-to-sequence model (Kumar and Joshi, 2017) and copy-based sequence-to-sequence model (Elgohary et al., 2019; Quan et al., 2019) are proposed, which can generate complete questions. Liu et al. (2019b) take use of question structures to rewrite utterance in conversational semantic parsing. Pan et al. (2019) leverage BERT to select words, and use these words to generate rewritten utterance. Su et al. (2019) distinguish the weights of context utterances for utterance rewriting. Liu et al. (2020) employ edit-based text generation and semantic similarity measurement for utterance rewriting.

3 Model

In this section, we will describe our ReWriteSum model in detail, with architecture shown in Figure 1.

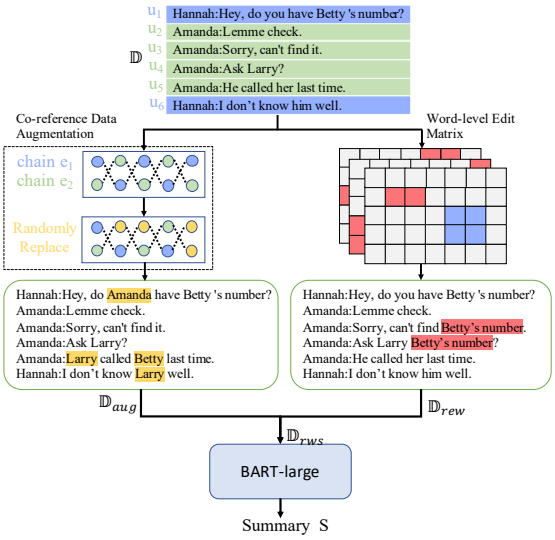

Figure 1: The model architecture of ReWriteSum. the left shows the co-reference data augmentation module and the right shows the incomplete utterance rewriting module.

ReWriteSum consists of an incomplete utterance rewriter, a co-reference data augmentation and a transformer-based BART summarization model.

For incomplete utterance rewriting, we establish a word-level edit matrix (Liu et al., 2020), whose element determines three editing operations: substitute, insert and none. To obtain the edit matrix, we conduct three neural networks, as shown in Figure 2: a BiLSTM-based context layer to obtain the word representation, an encoder layer to model the local information, and a segmentation layer to model the global information.

For co-reference data augmentation, we replace the co-reference word with its specific name entity through a co-reference resolution model (Joshi et al., 2020) and then augment the dataset with these replacement data. Specifically, we firstly use the co-reference resolution model to obtain the co-reference chain of the entire dialogue, then replace the pronoun in the co-reference chain with the specific name entity based on a certain probability. Finally, we utilize these replacement data to augment the personal information for dialogue summarization task.

3.1 Problem Formulation

Given the dialogue content set $D = \{u_1, \ldots, u_{|D|}\} \in \mathbb{D}$, each utterance in D is represented as $u_j = \{x_1^{(j)}, \ldots, x_L^{(j)}\}$, where $x_k^{(j)}$ represents the k^{th} word in utterance u_j. The corresponding summary of D is represented as

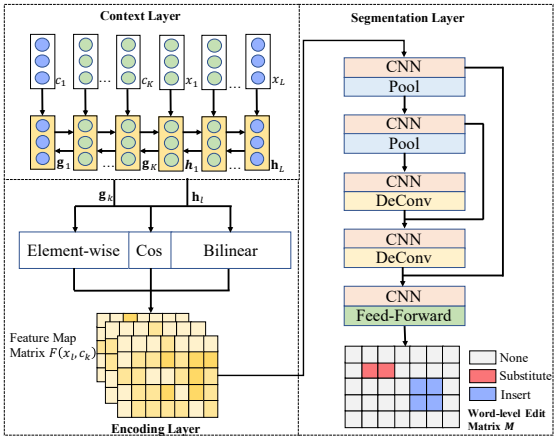

Figure 2: Architecture of incomplete utterance rewriter.

$S = \{y_1, \ldots, y_{L_S}\}$, where y_j represents the j^{th} word in summary S.

We adopt a neural model for abstractive dialogue summarization. In detail, given the dialogue D as input, we firstly utilize the utterance rewriting system and the co-reference resolution system to generate the new complete rewriting dialogue D'. And then, we use the rewriting dialogue D' as input, instead of dialogue D, to generate the dialogue summary.

3.2 Incomplete Utterance Rewriting

Given the whole dialogue $D = \{u_1, \ldots, u_{|D|}\}$, we define the context as $C = \{u_1, \ldots, u_{t-1}\}$ and the incomplete utterance as $u_t (t \leq |D|)$. Incomplete utterance rewriting aims at rewriting u_t to u_t^* through the context C. After rewriting, u_t^* should not only have the same meaning as u_t, but also can be understood separately. Specifically, we concatenate all the contextual utterances C into a K-length word sequence $\mathbf{c} = (c_1, \ldots c_K)$. At the same time, the incomplete utterance is represented as $u_t = \{x_1, \ldots x_L\}$, where L is the length of u_t. And then, the rewritten utterance u_t^* can be obtained by editing the incomplete dialogue u_t using the words in $\mathbf{c}$.

In order to determine the editing operation, we define a word-level edit matrix $\mathbf{M}$ (Liu et al., 2020), where each element m_{kl} represents the editing type between c_k and x_l. There are three editing types: substitute, insert and None. The substitute operation means replacing the word x_l with the context word c_k. The insert operation means inserting a word c_k before or after a certain token x_l. And None means no operation. Following Liu et al. (2020), we establish a word-level edit matrix through three neural layers: a context layer, an encoding layer and a subsequent segmentation layer, as shown in Figure 2, and then generate rewritten utterance based on this word-level edit matrix.

3.2.1 Context Layer

Given the contextual word sequence $\mathbf{c}$ and the incomplete utterance u_t, we firstly concatenate the $\mathbf{c}$ and u_t as input, and employ Glove (Pennington et al., 2014) to initialize the word embedding. And then, we use BiLSTM (Schuster and Paliwal, 1997) with both the left-to-right and right-to-left text representations to obtain the contextual information:

$$BiLSTM(\mathbf{c}; u_t) = (\mathbf{g}_{1,\ldots,K}; \mathbf{h}_{1,\ldots,L}),$$

where $\mathbf{g}_k$ is the hidden state of contextual word c_k in $\mathbf{c}$ and $\mathbf{h}_l$ is the hidden state of the word x_l in u_t.

3.2.2 Encoding Layer

After obtaining the context-aware hidden states $\mathbf{g}$ and $\mathbf{h}$, we use three similarity functions to calculate the word-level relevance between context and incomplete utterance. Specifically, for each word c_k and x_l, a D-dimensional vector $\mathbf{F}(x_l, c_k)$ is set to indicate the relevance:

$$\mathbf{F}(x_l, c_k) = [\mathbf{h}_l \odot \mathbf{g}_k; \cos(\mathbf{h}_l, \mathbf{g}_k); \mathbf{h}_l \mathbf{W}_{Bi} \mathbf{g}_k], \quad (1)$$

where $\odot$ is the element multiplication operation to obtain the element-wise similarity, $\cos(.,.)$ is the cosine similarity, and $\mathbf{W}_{Bi}$ is a learned parameter in learned bi-linear similarity. Finally, we obtain the feature map matrix $\mathbf{F} \in \mathbb{R}^{L \times K \times D}$.

Similarity function is used to describe word-to-word relevance from various aspects, which is a necessary condition for the edit type. However, the encoder layer can only obtain local information, which is not enough for incomplete utterance rewriting. Therefore, we conduct a segmentation layer to introduce the global information.

3.2.3 Segmentation Layer

Given the feature map matrix $\mathbf{F} \in \mathbb{R}^{L \times K \times D}$ in Equation 1, we use the segmentation layer to calculate the word-level edit matrix $\mathbf{M} \in \mathbb{R}^{L \times K}$. The segmentation layer is inspired by UNet (Ronneberger et al., 2015), consisting of five convolutional neural network(CNN) with skip-connection mechanism, which is used to extract the global con-

textual editing information, as shown in Figure 2:

$$\mathbf{F}^{'} = \text{CNN}(\mathbf{F}),$$
$$\mathbf{F}^{''} = \text{CNN}(\text{Pool}(\mathbf{F}^{'})),$$
$$\mathbf{F}^{'''} = \text{DeConv}(\text{CNN}(\text{Pool}(\mathbf{F}^{''}))),$$
$$\mathbf{F}^{''''} = \text{DeConv}(\text{CNN}(\mathbf{F}^{'''}, \mathbf{F}^{''})),$$
$$\mathbf{M} = \text{FeedForward}(\text{CNN}(\mathbf{F}^{''''}, \mathbf{F}^{'})),$$

where $\mathbf{CNN}(.)$ is the two layers of convolutional modules, $\mathbf{Pool}(.)$ is the MaxPooling operation, $\mathbf{DeConv}(.)$ is the deconvloution neural network, and $\mathbf{FeedForward}(.)$ is the feedforward layer.

Given the word-level edit matrix $\mathbf{M}$, for each word c_k in contextual utterances and x_l in incomplete utterance, the element M_{kl} determines one of three editing operations: substitute, insert and none. Specifically, when M_{kl} is close to 0, the corresponding operation is none. When close to 1, the operation is substitution, 2 is inserting before and 3 is inserting after. After that, we can rewrite every utterance u_t in D as u_t^* based on M. Finally, we use all the rewritten utterances u^* to replace D as D', and obtain the rewriting dataset $\mathbb{D}_{rew} = \{(D'_1, S_1), (D'_2, S_2), \ldots, (D'_N, S_N)\}$.

3.3 Co-reference Data Augmentation

Taking into account that there are a large number of names and referential relations in the dialogue process, we propose to use the data augmentation mechanism to enhance the personal information for dialogue summarization task.

Given a dialogue content D, we utilize a co-reference resolution system (Joshi et al., 2020) to obtain its corresponding co-referential chain set $E = \{e_1, e_2, \ldots, e_{|E|}\}$, where $e_i = \{x_{i1}, x_{i2}, \ldots, x_{i|e_i|}\}$ is represented as the i^{th} co-referential chain in dialogue D and x_{ij} denotes the j^{th} word in co-referential chain e_i. Take the example two in Table 1 as an example, the $E = \{\{Mike_0, \ldots, I_{64}\}, \{Wendy_4, \ldots, her_{52}\}, \{Dave_9, \ldots, I_{79}\}, \{Jerry_{50}\}\}$, where the $word_{idx}$ is the idx^{th} word in $\mathbf{c}$.

Then, we refer to all the pronouns in the whole dialogue D and replace it with its corresponding person name x_{name_i} based on a certain probability: when the length of pronouns $|e_i^{(pron)}| >= 5$, if the output probability of co-reference system $P(x_{ij}) >= 0.5$, then replace x_{ij} with x_{name_i}, otherwise, no replacement; when $0 < |e_i^{(pron)}| < 5$, if $P(x_{ij}) >= 0.8$, then replace; when $|e_i^{(pron)}| = 0$, remove this example.

Finally, after the person's name replacement, we obtain an additional dialogue dataset $\mathbb{D}_{aug} = \{(D''_1, S_1), \ldots (D''_G, S_G)\}$, where G is the number of dialogue-summary pairs after removing.

3.4 Summary Generation

Given $\mathbb{D}_{rew}$ and $\mathbb{D}_{aug}$, we combine them to obtain our rewriting dataset $\mathbb{D}_{rws}$. To generate the summary, we utilize the state-of-the-art model BART (Lewis et al., 2020) to encode the dialogue content D and decode the summary S step by step.

We use maximum likelihood estimation to train our model. Given a pair of dialogue D and summary $S = \{y_1, \ldots, y_{L_S}\}$ from $\mathbb{D}_{rws}$, we minimize the negative log-likelihood of the target sequence:

$$\mathcal{L} = -\sum_{\mathbb{D}_{rws}} \sum_{t=1}^{|L_S|} \text{logP}(y_t | y_1 ... y_{t-1}, D; \theta_{BART\_large}).$$

4 Experiments

In this section, we conduct experiments on two English dialogue summarization datasets SAMSum (Gliwa et al., 2019) and DialSum (Chen et al., 2021) to evaluate our proposed method.

4.1 Experimental Settings

We first introduce some empirical settings, i.e., datasets, baselines, and evaluation measures.

4.1.1 Datasets

We use two public dialogue summarization datasets. SAMSum contains everyday English message-like dialogues and annotated summary. We randomly split the SAMSum data to training, validation, and testing sets, which contains 14,732, 818 and 819 pairs, respectively. DialSum[1] contains English speaking practice dialogue and annotated summary, which has been cleaned and pre-processed by publisher, including deleting non-English characters, correcting spelling errors and grammatical errors. We randomly split the DialSum data to training, validation, and testing sets, which contains 12,460, 500 and 500 pairs, respectively.

4.1.2 Baselines and Parameters Setting

Seven baseline models are used for comparison on SAMSum, and four baseline models on DialSum. Lead3 (See et al., 2017) model extracts

[1]https://github.com/cylnlp/DialogSum

the first three leading sentences in the article as the summary. LONGEST (Gliwa et al., 2019) model selects the top N longest sentences as the summary. PTGen (See et al., 2017) model introduces copy and coverage mechanisms into the basic sequence-to-sequence model. FastAbs-RL (Chen and Bansal, 2018) model firstly selects salient sentences and then generates abstractive summaries through reinforcement learning. DynamicConv + GPT-2/News (Wu et al., 2018) model replaces the attention mechanism with a lightweight dynamic in transformer. BART (Lewis et al., 2020) is a pre-trained model, which uses the noise function to destroy text, and then reconstructs the original text, including two versions, BART(base) and BART(large). Multiview BART (Chen and Yang, 2020) extracts different views of dialogue features, and then uses a multi-view decoder to combine these features to generate summaries.

Our model uses a pre-trained model BART(large)[2] for initialization. In detail, BART (large) has 12 layers of encoder-decoder Transformer structure. Each layer has 16 attention heads. The hidden size and feed forward filter size are 1024 and 4096, respectively. It contains a total of 400M trainable parameters. The dropout rates for all layers are set to 0.1. The optimizer uses Adam (Kingma and Ba, 2015) with 200 warmup. The learning rates of SAMSum and DialSum are both 3e-5, and the maximum tokens for a certain batch are 800 and 1000, respectively. We run our models on a Tesla V100 GPU card with Pytorch.

4.1.3 Evaluation Measures

To evaluate our models, we utilize both quantitative metrics and human evaluation in our experiment. In detail, we use ROUGE-1, ROUGE-2 and ROUGE-L as quantitative metrics, which is widely used in NLP and summary tasks (Liu et al., 2021a,b; Chen and Yang, 2020). For human evaluation, we randomly select 100 dialogue-summary pairs from the test set of SAMSum and DialSum, respectively. Five annotators(all CS majored students studying NLP) are demanded to give the comparison between our model and baseline models. They are not told which summaries are derived from the baseline model and which summaries are derived from our model. They are required to evaluate the generated summary from three aspects: whether the generation is fluent, whether it has omitted content, and whether it has factual inconsistent errors. The evaluation results are represented as win, loss and tie, respectively indicating that the quality of generated summary by ReWriteSum is better, weaker or equal to baselines.

[2]https://huggingface.co/facebook/bart-large

SAMSum Dataset			
Model	R-1	R-2	R-L
Lead3	31.4	8.7	29.4
PTGen	40.1	15.3	36.6
DynamicConv+GPT-2	41.8	16.4	37.6
FastAbs-RL	42.0	18.1	39.2
DynamicCov+News	45.4	20.7	41.5
Multiview BART	52.2	**27.4**	49.9
BART(large)	50.9	25.0	47.1
ReWriteSum(ours)	**54.2**	27.1	**50.1**
DialSum Dataset			
Model	R-1	R-2	R-L
Lead3	27.5	6.8	27.3
LONGEST	24.1	6.2	22.7
BART(base)	33.7	13.8	30.9
BART(large)	34.1	13.7	31.2
ReWriteSum(ours)	**35.1**	**14.6**	**32.1**

Table 2: Metric-based evaluations of ReWriteSum and baselines on SAMSum and DialSum. R-1, R-2, R-L denote ROUGE-1, ROUGE-2, ROUGE-L, respectively.

SAMSum Dataset			
Model	ReWriteSum vs.		
	win(%)	loss(%)	tie(%)
Multi-view	48.5	6.9	44.6
BART	52.6	5.1	42.3
DialSum Dataset			
Model	ReWriteSum vs.		
	win(%)	loss(%)	tie(%)
Multi-view	42.3	8.1	49.6
BART	46.8	7.3	45.9

Table 3: Human evaluations on SAMSum and DialSum.

4.2 Experimental Results

In this section, we demonstrate our experiment results on SAMSum and DialSum datasets.

4.2.1 Metric-based Evaluation

The quantitative evaluation results on SAMSum and DialSum datasets are shown in Table 2. For SAMSum dataset, we refer to (Gliwa et al., 2019) to show the results of Lead3, PTGen, DynamicConv+GPT-2/News, and FastAbs-RL. From the results, we can see that the pre-trained models, such as BART and Multiview BART, outperform the traditional summarization models, showing the effectiveness of pre-training

Dialogue Example 1	Dialogue Example 2
Mia: could anybody **help me to buy a flight ticket?** ... Mia: I don't have a credit card at the moment. ... Tom: You can use mine help Mia to buy a flight ticket! Mia: Should I send you the link to buy a flight ticket? Tom: Just send me the flight, company and your personal data that I may need to buy a flight ticket. Mia: Great, so nice of you, thanks Tom.	Maria: Who's gonna be at imf lecture tomorrow? ... Alexander: On Saturday Alexander already meet for another. So my option is Friday afternoon or tomorrow. Sarah: Tomorrow and on Friday Sarah available ... Sarah: So can we meet **tomorrow evening? 17:15?** Alexander: It is fine by me. **Lawrence**: Lawrence will be late, but you can start **without me**.
BART Prediction: Mia doesn't have a credit card at the moment. Tom will use his card to buy a flight ticket for himself. Tom needs the flight, company and personal data.	**BART Prediction:** Alexander, Martha, Sarah, Lawrence and Sarah will meet tomorrow evening at 17:15 to discuss the imf lecture.
ReWriteSum Prediction: Mia doesn't have a credit card at the moment. Tom will use his card to buy a flight ticket for her.	**ReWriteSum Prediction:** Maria, Sarah, Alexander, and Martha will meet tomorrow evening at 17:15 to discuss the imf lecture. Lawrence will be late.

Table 4: Generated summaries from different models on SAMSum. Red words show the inconsistent content. Green words show the factul content. Blue words show the supplemented part by our model. Orange words show the name replacement by our model.

language model for dialogue summarization task. Our ReWriteSum model performs the best. Take the ROUGE-1 and ROUGE-L score for example, our ReWriteSum model obtains 54.2 and 50.1, respectively, which obviously outperforms Multiview BART model, i.e., 52.2 and 49.9.

From the results on DialSum in Table 2, we can see that our model also obtains the best performance. Take the ROUGE-1 and ROUGE-L score for example, our ReWriteSum obtains 35.1 and 32.1, respectively, which obviously outperforms BART(large), i.e., 34.1 and 31.2. However, the performance increment on DialSum is not significant as comparation on SAMSum. The reason is that utterances in DialSum are relatively more complete and the interactive speakers are fewer than SAMSum. According to statistics, there are only 13 sentences with more than 4 speakers in DialSum, which leads to relatively few errors caused by multi-speakers. We have conducted the significant test, and the result shows that the improvements of our model are significant on both datasets, i.e., p-value < 0.01.

In conclusion, our ReWriteSum model has the ability to generate a more complete and accurate summary than baselines.

4.2.2 Human Evaluation

Human evaluation results are shown in Table 3. The percentages of win, loss and tie, as compared with the baselines, are given to evaluate the fluency, completeness and consistency of generated summary by ReWriteSum. From the results, we can see that the proportion of evaluators who think our model better is the largest, surpassing other models. Take SAMSum dataset for example, ReWriteSum model obtains preference gains (win subtract loss) 41.6%, 47.5%, respectively.

4.2.3 Case Study

To further understand our proposed model, we give some generated cases in Table 4. According to the result, we can notice that ReWriteSum model performs better than baseline models. Take example1 in Table 4 as an example, BART model generates that "Tom will buy a flight ticket for himself", but in the dialogue content, the dialogue fact is "Tom will buy a flight ticket for Mia'. The reason is that the dialogue content tends to be omitted in daily dialogues. From example1, we can see that, in the entire dialogue, only Mia mentions "help me to buy a flight ticket" at the beginning, and this sentence is omitted in the subsequent utterances, which makes BART unable to correctly understand who the "ticket" will be bought for. When we rewrite the incomplete dialogue (in blue font), "help Mia to buy a flight ticket" is added to the end of some utterances, so that our model can generate a more accurate and logical consistent summary.

From example2, due to the complex references in this dialogue, BART misunderstood "Lawrence will be late" as "Lawrence will meet tomorrow evening at 17:15". When co-reference data augmentation is carried out, it strengthens the connection between "you" and "Alexander, Martha, Sarah" in the sentence "Lawrence: I will be late, but you can start without me", so as to avoid this personal inconsistency error.

4.3 Analysis

In order to confirm whether the improvement is related to incomplete utterance rewriting(IUR) and co-reference data augmentation(CDA), a further analysis is conducted, containing ablation study, the impact of participants, and the error analysis.

Model	R-1	R-2	R-L
ReWriteSum	54.2	27.1	50.1
- w/o CDA	51.1	25.1	47.5
- w/o IUR	52.3	25.1	48.1

Table 5: Ablation experiment results on SAMSum.

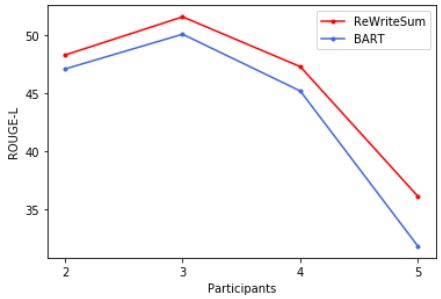

Figure 3: Rouge-L scores of ReWriteSum and BART with different number of speakers.

4.3.1 Ablation Study

To confirm the effectiveness of our IUR and CDA module, we conduct ablation experiments on SAMSum dataset. The results are shown in Table 6. ReWriteSum w/o IUR means that ReWriteSum model removes IUR module and only with CDA to generate summaries. From the results, we can see that when only CDA is applied, the ROUGE score is still larger than the baseline, but smaller than the ReWriteSum model. ReWriteSum w/o CDA means that our ReWriteSum model removes the CDA module and only with IUR to generate summaries. We can see that the ROUGE has decreased as compared with our ReWriteSum model, but it is still higher than baseline models. Therefore, we think that both incomplete utterance rewriting and co-reference data augmentation have positive effects for dialogue summarization.

Not only that, we also notice that the ROUGE score using IUR alone is higher than using CDA alone, indicating that IUR contributes more to the dialogue summarization task.

4.3.2 Impact of Participants

We conduct an experimental analysis with different number of participants, by calculating the ROUGE score for baselines and our ReWriteSum model on SAMSum. From the Figure 3 , we can see that with the increase of participants, the rouge score of our model decreases more slowly, because: (1) with the increase of participants, the omitted information will also increase, but our incomplete utterance rewriting module has the ability to reduce the impact of too much omitted information in the summary; (2) our co-reference data augmentation module can reduce the impact of complex referencing caused by too many participants.

Model	Missing Information	Wrong Reference	Incorrect Reasoning
BART(large)	36	24	19
ReWriteSum	14	6	8
- w/o CDA	17	19	9
- w/o IUR	29	7	11

Table 6: Percentage of typical errors in summaries generated by BART(large) and our ReWriteSum model.

4.3.3 Error Analysis

To further study the impact of IUR and CDA on the quality of generated summaries, we count the following 3 kinds of errors that appear in the summaries generated by the baseline model and our model: **Missing Information**: content information that appears in gold summaries is missing from generated summaries. **Wrong Reference**: content in the generated summaries, such as the person's actions or name, does not match what is described in the source dialogue. **Incorrect reasoning**: the conclusions drawn by the generated summaries are inconsistent with the facts in the source dialogue.

We randomly select 100 dialogues and their generations from SAMSum and count the error categories, as shown in Table 5. In terms of missing information, our model outperforms the baseline model because IUR can effectively prevent the model from missing information. According to the wrong reference numbers, our model performs better than the baselines because CDA can enhance the model's understanding of referential information. Errors occur in incorrect reasoning are also reduced as our model complements default information and enhances understanding of referential information.

5 Conclusion

In this work, we propose a new dialogue summarization model, namely ReWriteSum, which leverages incomplete utterance rewriting and co-reference data augmentation mechanism to generate summaries for dialogue. Our motivation comes from the fact that there are a lot of ellipsis and demonstrative pronouns in the dialogue, which seriously affects the quality of dialogue summary generation. Our core idea is to utilize the incomplete utterance rewriting module to complete the ellipsis information in the dialogue content and enhance the personal entities with the co-reference data augmentation mechanism. We conduct exper-

iments on both SAMSum and DialSum datasets, and the results on both quantitative and qualitative analysis verify the effectiveness of our proposed model. Therefore, we obtain the conclusion that the incomplete utterance rewriting and co-reference data augmentation are effective for improving the quality of generation for dialogue summarization.

6 Ethical Considerations

The abstractive summarization dialogue system proposed in this work can be applied to dialogue scenarios. It can quickly process a lengthy dialogue into a short content containing the core idea of the dialogue. Such features can be applied to meetings, customer service, and medical scenarios to facilitate people's life. The datasets SAMSum and DialSum used in this work are publishable and for research purposes only. There may be some biased content in the datasets, which should be viewed carefully.

Acknowledgement

Hainan Zhang and Yanquan Zhou are the corresponding authors. We would like to thank anonymous reviewers for their thoughtful comments and suggestions.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686.

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

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. *Can You Unpack That? Learning to Rewrite Questions-in-Context*.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *EMNLP*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Vineet Kumar and Sachindra Joshi. 2016. Non-sentential question resolution using sequence to sequence learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2022–2031.

Vineet Kumar and Sachindra Joshi. 2017. Incomplete follow-up question resolution using retrieval based sequence to sequence learning. In *Proceedings of*

the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 705–714.

Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7823–7827. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Manling Li, Lingyu Zhang, Richard J Radke, and Heng Ji. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *57th Conference of the Association for Computational Linguistics*.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021a. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243.

Qian Liu, Bei Chen, Jian-Guang Lou, Ge Jin, and Dongmei Zhang. 2019b. Fanda: A novel approach to perform follow-up query analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6770–6777.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2846–2857.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019c. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

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

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

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

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674.

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

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2018. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The world wide web conference*, pages 3455–3461.