Contrastive Topic Prompt Learning for Dialogue Summarization

Anonymous ACL submission

Abstract

The prevalence of online meetings, such as Zoom and Microsoft Teams, has highlighted the necessity for effective dialogue summary. Nevertheless, relatively less work has been done on topic-guided dialogue summarization, which is critical considering the fact that a dialogue involves numerous speakers and conversational aspects. This study proposes Contrastive Topic Prompt Learning, a simple method that generates several aspect summaries based on a user-specified topic. Firstly, to guide the summary towards a specific topic, inspired by the success of prompts in guiding aspects in general summarization, we utilized topic prompt for guiding our dialogue summarization. Nevertheless, our preliminary experiment revealed that depending solely on the topic prompt frequently leads to repetitive summaries and hence calls for more sophisticated training methods. We propose the use of contrastive learning on topic prompt, which make use of positive and negative pairs to enforce the models to learn the similarities and differences of different topics. Experimental results showed that our model outperformed state-of-the-art (SOTA) models in the ROUGE score on the DialogSum dataset. Further analysis revealed that our models generate diverse summaries that are less repetitive.

Dialogue Example	
#Person1#:	Hey, How's it going?
#Person2#:	Not good. I lost my wallet.
#Person1#:	Oh, that's too bad. Was it stolen?
#Person2#:	No, I think it came out of my pocket when I was in the taxi.
#Person1#:	Is there anything I can do?
#Person2#:	Can I borrow some money?
#Person1#:	Sure, how much do you need?
#Person2#:	About 50 dollars.
#Person1#:	That's no problem.
#Person2#:	Thanks. I'll pay you back on Friday.
#Person1#:	That'll be fine. Here you are.
#Person2#:	What are you going to do now?
#Person1#:	I'm going to buy some books and then I'm going to the gas station.
#Person2#:	If you wait a minute I can go with you.
#Person1#:	OK. I'll wait for you.
Gold Summary1:	#Person2# lost the wallet and borrows some money from #Person1#.
Gold Summary2:	#Person2#'s wallet was lost and has to borrow some money from #Person1#.
Gold Summary3:	#Person1# lends money to #Person2# as #Person2# lost the wallet.
BART$_{large}$:	#Person2# lost the wallet and asks #Person1# to lend some money to buy some books and go to the gas station.
T 1:	#Person2# lost the wallet. #Person1# helps #Person2# borrow 50 dollars. They are going to buy some books and then go to the gas station.
T 2:	#Person2# lost #Person1#'s wallet and asks to borrow some money to buy some books and then go to the gas station.
T 3:	#Person2# lost the wallet. #Person1# lends #Person2# 50 dollars to buy some books and then go to the gas station.
T-L 1:	#Person2# lost the wallet and borrows 50 dollars from #Person1#.
T-L 2:	#Person2# lost the wallet and borrows 50 dollars from #Person1#.
T-L 3:	#Person2# lost the wallet and borrows 50 dollars from #Person1#.
T-L-CL (Ours) 1:	**#Person2# borrows 50 dollars from #Person1# to buy books.**
T-L-CL (Ours) 2:	**#Person2# lost the wallet and borrows 50 dollars from #Person1#.**
T-L-CL (Ours) 3:	**#Person2# lost the wallet and asks #Person1# for money.**

Figure 1: Typical pretrained model such as BART$_{large}$ produces generic single summary. Topic prompt (T) generates non-diverse summaries, each of which are slightly different thus prone to semantic inconsistencies. Topic prompt paired with length prompt (T-L) generates more consistent summaries as you constrain the length, but it produces very repetitive summaries. Our proposed technique (T-L-CL) generates diverse summaries relevant to the specified topic.

1 Introduction

Dialogue summarization condenses key information from a dialogue and presents it in a more concise form, enabling individuals to quickly grasp the essential points. Many approaches were proposed, namely utilizing existing pre-trained summarization models (Khalifa et al., 2021; Chen et al., 2021; Feng et al., 2021), graph-based approaches for understanding complex relationships (Zhao et al., 2020a; Chen and Yang, 2021), multi-encoders for understanding different views of dialogues (Chen and Yang, 2020), contrastive learning for understanding co-references or utterances of similar topics (Tang et al., 2021; Liu et al., 2021), and further pretraining the models using certain objectives such as masking (Zhong et al., 2022).

However, despite the advancements, relatively less work has been done on topic-guided dialogue summarization. This is relevant accounting for the fact that a dialogue involves multiple speakers with different perspectives, intents, and actions (Chen et al., 2021). In other words, it can be beneficial to allow users to generate a summary that is relevant to their interests. Indeed, there were few such

attempts, e.g., Amplayo et al. allowed users to control opinion summaries by specifying aspects; similarly, Xu and Lapata proposed query-based summarization for multi-document summarization, which summarizes multiple documents based on a given query. In any case, such a technique usually requires modification of model architectures or requires a query as input for training. In addition, little work has been done specifically on topic-guided dialogue summarization. As for recent years, the idea of prompt has attracted much interest due to its simplicity (e.g., does not require the modification of model architecture). For example, Zhang et al. achieved controllable summarization through prompts which use control signals (e.g., length of generated summaries, named entities that appear in summaries) during the model training phase. Nevertheless, the use of prompt remains underexplored in the area of dialogue summarization.

This study proposes Contrastive Topic Prompt Learning, a simple method that generates summaries based on user-specified topics. We chose DialogSum (Chen et al., 2021) as it closely represents real-world situations. First, we used topic prompts to guide our dialogue summarization. Nevertheless, our preliminary experiment revealed that depending solely on the topic prompt frequently leads to repetitive and occasionally hallucinated summaries (See Figure 1), which hence calls for more sophisticated training methods. Inspired by Xiong et al., we apply contrastive learning on the topic prompt. Specifically, it makes use of positive and negative pairs which helps enforce the model to better distinguish different topics. We found that contrastive learning is especially useful to learn multiple topics during the training phase, even when topic annotation is limited. For example, in the DialogSum, only a single topic is available for the training set, while the testing set contains three topics, which resemble real-world cases of scarce topic annotations.

Our extensive experimental results showed that our model outperformed other state-of-the-art models in the ROUGE score. Further analysis demonstrated reduced repetitiveness in produced summaries. We also further experimented other possible components such as length control and variations of negative examples.

The contributions to our work are as follows:

1. We propose Contrastive Topic Prompt Learning for dialogue summarization.
2. Our simple prompt-based approach outperformed other state-of-the-art models on the DialogSum dataset.
3. We have conducted extensive experiments and analysis comparing different topic prompt variants, yielding a few research insights.

2 Related Work

The difficulties in dialogue summarization stem from the interplay between multiple participants, changes in subjects, numerous co-references, varied interaction cues, and domain-specific language (Feng et al., 2021). The generation of a dialogue summary still faces issues with repetition, a lack of variation, incoherence, and lack of topic-guided summarization (Sun and Li, 2021).

BART (Lewis et al., 2019) is an encoder-decoder model that has been widely employed in dialogue summarization. Khalifa et al. discovered that BART performed better than UniLM and other conventional abstractive methods when tested on the SAMSum (Gliwa et al., 2019) dataset. On the DialogSum dataset which is highly abstractive and resemble real-life scenarios, Chen et al. found that BART performance on DialogSum is similar to that used by UniLM model.

Graph-based techniques were presented to address the intricate relationships in dialogue summarization. Zhao et al. proposed a graph attention based mechanism to encode long-distance relationships within the dialogue. Chen and Yang utilized structured graph to model "who do what" to input to graph attention network for better dialogue summarization. However, the use of a graph-based technique is often not suitable for parallelization and can be computationally demanding.

Typically, researchers have explored aspect-based or query-based summarization in the field of summarization. We are interested in examining this topic because to its high relevance to dialogue summarization. Amplayo et al. proposed the use of aspect controllers which pool the tokens, sentences, and documents that are most relevant to the user's specified aspect. Xu and Lapata proposed query-based summarization for multi-document summarization. Both works excel in enabling users to input certain aspects or queries that can direct the summary process. However, it is worth noting that a potential disadvantage is that these works necessitate either altering the architecture of the model or requiring a query as input during training.

Other approaches have also been proposed. Chen and Yang proposed a multi-view decoder models that takes in hidden states from multiple encoders that encodes different views, and the decoder will decide the attention weights which view it should focus on, to produce the final summary. Zhong et al. proposed a pre-trained methodology using masking techniques for dialogue summarization. In any case, both works do not consider topic-guided summarization.

Contrastive learning was proposed as a means to gain a deeper understanding of facets discussed in a dialogue. CONFIT (Tang et al., 2021) incorporated contrastive loss to mitigate the issues of missing information and incorrect references in conversation summarization tasks. Xiong et al. utilized contrastive learning as a means to decrease repetition in the context of scientific summarization. Tan and Sun has shown that the implementation of contrastive learning may enhance the quality of the output summary by allowing the model to distinguish training falls resulting from false negative samples. Liu et al. make use of contrastive learning by enforcing the models to contrast positive and negative samples, where positive samples are defined based on a specified window utterance sizes, allowing the decoder to capture salient intent information. Regardless, none of this work specifically focuses on the production of topic-guided summarization. We are inspired by the achievement of these methods and aim to employ contrastive learning for the purpose of topic-guided dialogue summarization.

Recently, there has been a growing interest in the use of prompts to regulate summarization, mostly because of its simplicity. They have the ability to manipulate the characteristics of produced summaries and potentially enhance the quality of summarising. For example, Zhang et al. achieved controllable summarization through prompts. They used control signals (e.g., length of generated summaries, named entities that appear in summaries) during the model training phase. Based on its simplicity, we begin to explore the possible combination of prompt-based approach together with contrastive learning for topic-guided dialogue summarization.

3 Methodology

We propose a contrastive topic prompt learning, a method that enables summary generation based on a specified topic. Here, we chose DialogSum (Chen et al., 2021) as it closely represents real-world situations. Specifically, DialogSum comprises a triple of document, topic and summary $\{(D, T, S)\}$ where a document is coupled with a topic and a summary in the training set, while in testing, a document is coupled with a set of topics $T = \{t_1, t_2, t_3\}$ along with its respective summaries $S = \{s_1, s_2, s_3\}$.

Due to the limited topic annotation in the training set, we further enhanced our topic prompt using contrastive learning, which makes use of positive and negative examples. Specifically, the actual topic (i.e., specified by the dataset) serves as positive and its synonym and random topic words serve as negative samples to enforce the models to learn the similarities and differences of summaries in different topics. Note that although it is more intuitive to use synonyms as positive samples, but due to the very small difference between synonyms and the actual topics, synonyms are promising candidates to serve as *hard* negative samples, similar to the discussion of *hard negative mining* discussed in Kurowski et al.

In addition, inspired the work of Wang et al.; Wu et al.; Yoo and Lee, we experimented the addition of length control to constrain the model to produce concise summary. Note that there could be possible tradeoffs between recall and precision. By incorporating length control, it is expected that precision is increased, since the denominator of the precision equation is shortened. But length control can also leads to repetitive summaries. Conversely, when topic prompt is used without length control, recall can increased since the model can generate as much as it can, to increase the recall score. We envision that the ideal scenario is when both recall and precision are balanced so that the produced summary is both concise while exhibiting small diversity.

Finally, given the input, the objective is to minimize two losses namely, the contrastive loss and the negative log-likelihood to generate output summary.

3.1 Prompt Template

Here we introduce our prompt template that guides the generation using topic. Specifically, we frame our input as `Topic of Summary:`$\{t\}$`, Dialogue:` $\{d\}$ where t denote topic and d is our dialogue context. To train our model using contrastive learning, the topic t serves as positive (t_p) and its synonym and random topic word serve as negative samples

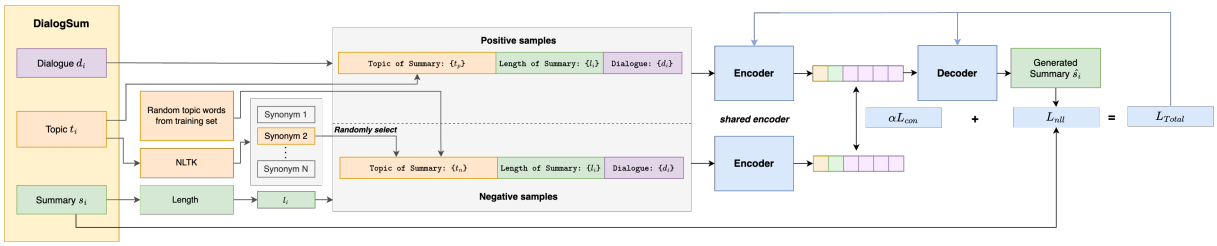

Figure 2: Overview of our framework

(t_n). Here the ratio between synonym and random is kept equal.

As for the synonym replacement, we employ wordnet of the NLTK library. For the given topic t, we obtain a set of synonyms $Syn_t = \{syn_1, syn_2, ..., syn_{|Syn_t|}\}$ in which we randomly select one as the replacement of the topic and represent as negative sample. On the other hand, for random word topics, we randomly select a topic word given in the training dataset.

Note that for the length control, we additionally included `Length of Summary:`$\{l\}$ as a part of our prompt template. Here, l denote the length of a summary used during the training phase which is simply a number of summary words defined by space (i.e., `string.split`). During interference, users can specify any desired length.

Hence, our final prompt template becomes, `Topic of Summary: Length of Summary` $\{l\}$. `Dialogue:` $\{d\}$, where t denote topic, l denote length and d is our dialogue context.

3.2 Contrastive Learning

Our framework incorporated contrastive learning to assist the learning. In particular, it makes use of positive and negative pairs to enforce the models to learn the similarities and differences of summaries in different topics. Specifically, we obtain the last hidden state of the encoder of positive and negative topic prompts and employ cosine similarity as our similarity measure. Below describes our contrastive loss computation.

$$\mathcal{L}_{con} = max(0, cos(h_1, h_2) - margin) \quad (1)$$

where h_1 denote the last hidden state of positive samples, while h_2 denote the last hidden state of negative samples and the margin is set to 0.5. Here the aim is to maximize the cosine similarity between the original and the comparison.

3.3 Dialogue Summarization

To generate dialogue summary, we perform fine-tuning on the pretrained model. Given the input, the objective is to minimize a joint loss namely the contrastive learning and the standard cross entropy losses of generating the output summary $s = \{s_1, s_2, ..., s_{|s|}\}$. The stand cross entropy loss is defined as negative log-likelihood (NLL) and is as follows:

$$\mathcal{L}_{nll} = -\sum_{i=1}^{|s|} f\left(s_i|D, s_{<i}\right) \quad (2)$$

where $f\left(s_i|D, s_{<i}\right)$ is the log-likelihood of the ith token of the reference summary. During training on the dataset using the NLL loss $\mathcal{L}_{nll}$, the score of the gold summary is expected to increase. This increase makes it more probable for the gold summary to be produced as one of the candidate summaries. Consequently, the model produces a generalized summary that covers several topics in a wider context.

Hence, total loss becomes,

$$\mathcal{L}_{total} = \mathcal{L}_{nll} + \alpha\mathcal{L}_{con} \quad (3)$$

Where $\mathcal{L}_{nll}$ is negative log-likelihood (NLL) and $\mathcal{L}_{con}$ is contrastive learning loss, it integrates NLL with loss by multiplying with alpha, which defaults to 0.5.

4 Experiments

4.1 Datasets

In DialogSum, a training sample comprises a document coupled with a topic and a summary. In testing, a document is coupled with a set of 3 topics and its respective summary. This limitation in topic annotation makes DialogSum challenging and resembles real-world scenarios (Chauhan et al., 2022). Specifically, the dataset is collected from various sources, including DailyDialog, DREAM,

and MuTual, and consists of 13,460 daily conversations, which are divided into three subsets: 12,460 for training, 500 for validation, and 500 for testing.

4.2 Experimental setting

Here we describe the experimental setting of our work. Our implementation is based on BART$_{large}$ model with has 406M parameters. Here all input was truncated to 1024 and output is set to 128 tokens. For the fine-tuning, the learning rate is set to 5e-05 and the model was trained for 15 epochs at batch size 4 with the min and max output lengths of 1 and 128 respectively. Additionally, we adopt AdamW as our optimizer. At inference time, beam size of 4 is selected with the min and max output lengths are kept the same as fine-tuning. Note that all our language models were taken from HuggingFace and can run each experiment on one A6000 GPU.

As for the evaluation metric, ROUGE scores (Lin, 2004) including ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L) were selected.

5 Results

We experimented (1) the prompt design which includes the comparison with two baselines - pre-trained BART$_{large}$ (Lewis et al., 2019) and the current SOTA for DialogSum, i.e., LA-BART (Wang et al., 2022) - with topic prompt (`T`), with topic prompt + length control (`T-L`), with topic prompt + contrastive learning (`T-CL`), and with topic prompt + length control + contrastive learning (`T-L-CL`), (2) negative samples selection for contrastive learning, where we experimented using random words, synonyms, and combined in an equal ratio as negative samples. Note that random words here refer to words randomly picked from the list of topic words in the training set.

5.1 Prompt Design

We compared different topic prompt designs in Table 1. The result showed that our proposed method outperformed previous work across all F1 and precision scores, while our recall maintained competitiveness with the SOTA baseline. Particularly, our `T-L` was the best performer in terms of F1 scores, followed by `T-L-CL` and `T-L`, with baseline being the lowest performers. Note that though `T-L` was the best performer in terms of F1 scores, `T-L-CL` achieved a better balance between precision, recall and F1 scores.

5.2 Contrastive learning

We designed an experiment to explore the use of synonym and random words as our negative samples to assist contrastive learning. Note that random words are words randomly picked from the list of topic words in the training set. Specifically, we compared synonyms only, random topic words only and a combination of both in equal ratio as our negative samples. Table 2 showed that using a combination of both yielded the highest results in terms of F1 scores, while using random topic words only yielded the highest recall scores and synonyms only yielded the highest precision scores.

6 Discussion

6.1 Prompt Designs

We found that incorporating length control increases precision (i.e., produces concise summary) and that incorporating contrastive learning increases recall (i.e., produces diverse summary based on specified topic). In addition, by combining the two, it resulted into a balance between recall and precision. In effect, it produces concise summary while also offering diverse summary based on specified topic. Here we further analyze the possible reasons.

Couple of key observations can be made in these results: (1) incorporating solely the topic prompt `T` already provides competitive results with baselines, highlighting its usefulness despite its simplicity; (2) incorporating length control `T-L` greatly increases precision. This is expected as discussed in the Methodology, where by constraining the length, it produces more concise summary. Mathematically, the denominator of the precision equation got shortened; (3) incorporating contrastive learning `T-CL` enhances the model's ability to generate diverse summary based on different topics, hence increases recall. To clarify, note that in the testing set of DialogSum contains three summaries of different topics, thus by incorporating contrastive learning, it can generate more diverse summaries that has higher chance of hitting common words with the three reference summaries; (4) combining length control and contrastive learning `T-L-CL` balances the model's ability to produce concise summary while providing diverse summary for different topics. In other words, our view is that the ideal scenario is met recall and precision are balanced.

Prompt	R-1			R-2			R-L			Len. Δ
	P	R	F1	P	R	F1	P	R	F1	
Baseline	44.55	53.26	47.10	19.94	23.46	20.87	42.51	49.31	44.72	6.97
Baseline (LA-BART)	48.03	50.89	48.95	21.73	22.86	22.07	45.84	47.95	46.56	3.56
Ours (T)	44.30	54.53	47.39	19.89	23.85	20.98	42.22	50.15	44.85	7.84
Ours (T-CL)	43.31	**54.89**	46.99	19.25	23.87	20.61	41.36	**50.31**	44.45	8.17
Ours (T-L)	48.98	52.33	**50.22**	22.62	**23.97**	**23.09**	46.65	49.16	47.62	3.22
Ours (T-L-CL)	**50.36**	50.73	50.10	**23.17**	23.19	22.97	**47.92**	48.08	**47.70**	**2.95**

Table 1: Precision, Recall and F1-score in ROUGE metric on different prompt. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L recall respectively. The highest scores are bolded. Here the performance of our designs are compared against two baselines - pretrained BART$_{large}$ (Lewis et al., 2019) and the current SOTA for DialogSum, i.e., LA-BART. The designs include topic prompt (`T`), topic prompt + length control (`T-L`), topic prompt + contrastive learning (`T-CL`), and topic prompt + length control + contrastive learning (`T-L-CL`)

Prompt	Positive	Negative	R-1			R-2			R-L			Len. Δ
			P	R	F1	P	R	F1	P	R	F1	
T-L-CL	Actual Topic	Random	48.37	**52.75**	49.96	22.23	**23.98**	22.83	46.15	**49.44**	47.41	3.60
T-L-CL	Actual Topic	Synonym	**50.88**	49.5	49.73	**23.18**	22.38	22.56	**48.1**	46.9	47.2	**1.65**
T-L-CL	Actual Topic	Random, Synonym	50.36	50.73	**50.1**	23.17	23.19	**22.97**	47.92	48.08	**47.7**	2.95

Table 2: Precision, Recall and F1-score in ROUGE metric on different types of negative samples. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L Recall respectively. The highest scores are bolded. Here the performance of our proposed topic prompt with length prompt, and contrastive learning (T-L-CL) using both random topic words and synonym as negative samples is compared against one with random topic words only and synonym only as negative samples to assist contrastive learning.

6.2 Contrastive Learning

Our results showed that a combination of synonym and random topic words achieved the best performance in terms of F1 scores, while random topic words alone and synonym alone achieved the best performance in terms of recall and precision scores respectively.

A possible reason is when random topic words were used as negative samples, due to the large difference between random topic words and the actual topic, the range of possible generated summaries are widened, leading to an increase in recall (i.e., since recall does not restrict the length of the generated summary). Conversely, when synonym was used as negative samples, due to the small difference between synonym and the actual topic, the range of generated summaries are constrained within a smaller domain, thus a increased precision is observed. At the end, by combining both synonyms and random words, it helps balance the recall (i.e., diverse) and precision (i.e., concise), delivering a diverse and concise summary.

6.3 Prompt Tagging

We also experimented with adding a simple tag where the prompt looks like this: Tagging was widely used in past works in NLP areas (Passali and Tsoumakas, 2022). It is assumed by past literature that it provide model's ability to better understand the context.

Table 3 shows the results. It was found that by adding special tag to our prompt (`T-ST`, `T-L-ST`), a drop in recall and F1 scores were observed, when compared with non-tagging variants (`T`, `T-L`). However, this is not true when learned in a contrastive learning manner (`T-ST-CL`, `T-L-ST-CL`). In fact, adding a simple prompt tagging is able to outperform our best performer (`T-L-CL`) in many metrics. This helps enhance the model ability to focus on the topic which encourages the model to generate diverse summaries.

A possible reason of explanation is that tagging assists contrastive learning in identifying the point of interests, similar to past work (Passali and Tsoumakas, 2022). On the other hand, using only tagging might cause the model to overly focus on certain aspects of the prompt and thus leading the overall decreased performance. Further future work may be required to fully scrutinized the mechanisms of tagging.

7 Conclusion and Future Work

We propose Contrastive Topic Prompt Learning, a simple method that generates summaries tailored

Prompt	R-1			R-2			R-L			Len. Δ
	P	R	F1	P	R	F1	P	R	F1	
Baseline	44.55	53.26	47.10	19.94	23.46	20.87	42.51	49.31	44.72	6.97
Baseline (LA-BART)	48.03	50.89	48.95	21.73	22.86	22.07	45.84	47.95	46.56	3.56
Ours (T)	44.30	54.53	47.39	19.89	23.85	20.98	42.22	50.15	44.85	7.84
Ours (T-CL)	43.31	**54.89**	46.99	19.25	23.87	20.61	41.36	**50.31**	44.45	8.17
Ours (T-L)	48.98	52.33	50.22	22.62	23.97	23.09	46.65	49.16	47.62	3.22
Ours (T-L-CL)	**50.36**	50.73	50.10	**23.17**	23.19	22.97	**47.92**	48.08	47.70	**2.95**
Ours (T-ST)	44.51	53.85	47.24	20.36	23.97	21.29	42.67	49.92	45.00	7.65
Ours (T-ST-CL)	43.13	54.77	46.74	19.20	23.86	20.55	41.13	50.21	44.21	8.45
Ours (T-L-ST)	49.51	51.55	49.95	22.72	23.40	22.78	46.91	48.42	47.29	3.81
Ours (T-L-ST-CL)	49.12	52.23	**50.30**	22.83	**24.07**	**23.27**	46.79	49.13	**47.72**	3.11

Table 3: Precision, Recall and F1-score in ROUGE metric on special tag. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L recall respectively. The highest scores are bolded. Here the performance of our proposed topic prompt with length prompt, special tag and contrastive learning (T-L-ST-CL) is compared against its variations and two baselines (i.e. no prompt and baseline (LA-BART)). The variations include topic prompt with length prompt (T-L), topic prompt with length prompt and special tag (T-L-ST), topic prompt with length prompt with contrastive learning (T-L-CL).

to user's specified topic. Specifically, to guide the summary towards a specific topic, topic prompt is utilized. Additionally, we propose contrastive learning on topic prompt which allows the model to generate less repetitive (i.e., more diverse) summaries on different topics. In other words, without contrastive learning, even with different topics, the summary looks the same. The experimental results showed that our model outperformed state-of-the-art models in ROUGE scores on the DialogSum dataset. Future work includes the inclusion of other variants of contrastive learning, prompt variations, and negative samples mining.

8 Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable suggestion and feedback.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593.

Vipul Chauhan, Prasenjeet Roy, Lipika Dey, and Tushar Goel. 2022. Tcs_witm_2022@ dialogsum: Topic oriented summarization using transformer based encoder decoder model. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 104–109.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. *arXiv preprint arXiv:2109.08232*.

Mariusz Kurowski, Andrzej Sroczyński, Georgis Bogdanis, and Andrzej Czyżewski. 2021. An automated method for biometric handwritten signature authentication employing neural networks. *Electronics*, 10(4):456.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. *arXiv preprint arXiv:2109.04994*.

Tatiana Passali and Grigorios Tsoumakas. 2022. Topic-aware evaluation and transformer methods for topic-controllable summarization. *arXiv preprint arXiv:2206.04317*.

Shichao Sun and Wenjie Li. 2021. Alleviating exposure bias via contrastive learning for abstractive text summarization. *arXiv preprint arXiv:2108.11846*.

Caidong Tan and Xiao Sun. 2023. Colrp: A contrastive learning abstractive text summarization method with rouge penalty. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713*.

Bin Wang, Chen Zhang, Chengwei Wei, and Haizhou Li. 2022. A focused study on sequence length for dialogue summarization. *arXiv preprint arXiv:2209.11910*.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Jing-Wen Xiong, Xian-Ling Mao, Yizhe Yang, and Heyan Huang. 2023. Cplr-sfs: Contrastive prompt learning to reduce redundancy for scientific faceted summarization. In *Journal of Physics: Conference Series*, volume 2506, page 012006. IOP Publishing.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*, pages 3632–3645.

Yumo Xu and Mirella Lapata. 2022. Document summarization with latent queries. *Transactions of the Association for Computational Linguistics*, 10:623–638.

Chongjae Yoo and Hwanhee Lee. 2023. Improving abstractive dialogue summarization using keyword extraction. *Applied Sciences*, 13(17):9771.

Yubo Zhang, Xingxing Zhang, Xun Wang, Si-qing Chen, and Furu Wei. 2022. Latent prompt tuning for text summarization. *arXiv preprint arXiv:2211.01837*.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020a. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020b. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11765–11773.