TCS_WITM_2022 @ DialogSum : Topic oriented Summarization using Transformer based Encoder Decoder Model

Vipul Chauhan, Prasenjeet Roy, Lipika Dey and **Tushar Goel**
TCS Research
New Delhi India
(chauhan.vipul, r.prasenjeet2, lipika.dey, t.goel)@tcs.com

Abstract

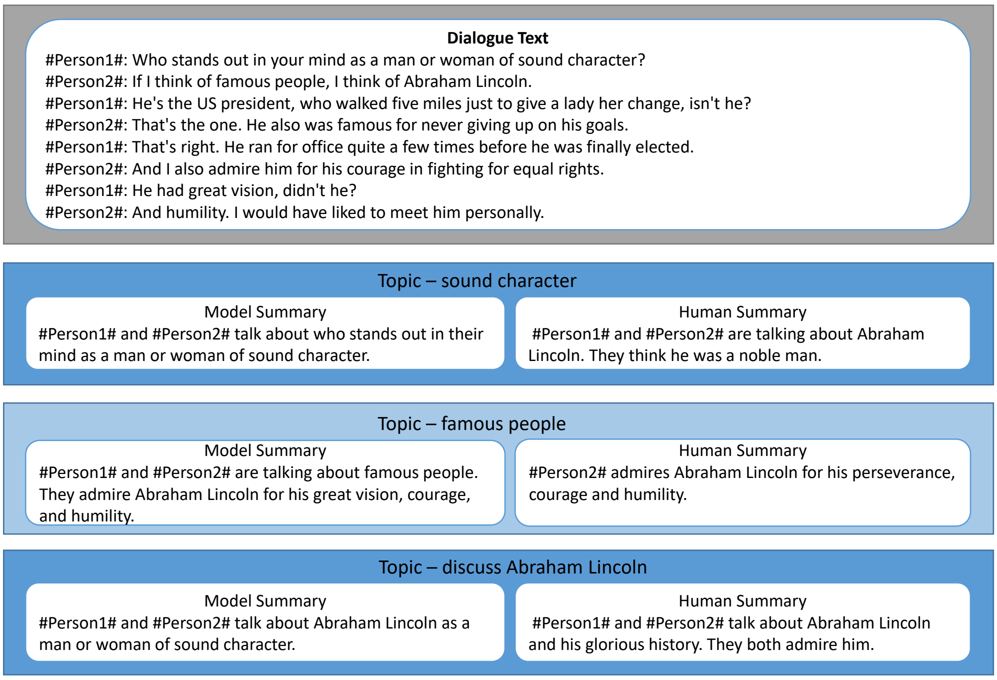
In this paper, we present our approach to the DialogSum challenge, which was proposed as a shared task aimed to summarize dialogues from real-life scenarios. The challenge was to design a system that can generate fluent and salient summaries of a multi-turn dialogue text. Dialogue summarization has many commercial applications as it can be used to summarize conversations between customers and service agents, meeting notes, conference proceedings etc. Appropriate dialogue summarization can enhance the experience of conversing with chatbots or personal digital assistants. We have proposed a topic-based abstractive summarization method, which is generated by fine-tuning PEGASUS[1], which is the state of the art abstractive summary generation model.We have compared different types of fine-tuning approaches that can lead to different types of summaries. We found that since conversations usually veer around a topic, using topics along with the dialoagues, helps to generate more human-like summaries. The topics in this case resemble user perspective, around which summaries are usually sought. The generated summary has been evaluated with ground truth summaries provided by the challenge owners. We use the py-rouge score and BERT-Score metrics to compare the results.

1 Introduction

Automatic text summarization is an important task in natural language processing, and it has been studied for decades. While extractive summarization focused on picking up the most important sentences from the text and create a summary, abstractive summarization generates new concise sentences with the important concepts. The task of abstractive summarization thus has two sub-tasks - identifying the important concepts within content and generating new sentences that are grammatically correct and can cover all important concepts sufficiently without repetition or redundancy. Both the summarization techniques have received attention from researchers of natural language processing. Some of the most cited works in the area of extractive summarization are (Erkan and Radev, 2004; Rai et al., 2021), and for abstractive summarization one may refer to (Lewis et al., 2019; Raffel et al., 2019; Zhang et al., 2020).

However, most of the above-mentioned works has focused on single-speaker documents such as news (See et al., 2017; Nallapati et al., 2016), scientific publications (Nikolov et al., 2018) etc. The documents considered also were short and assumed to contain a limited number of concepts around which summaries were to be generated. On demand summarization based on user queries, summarization of multi-section large reports are some of the problems that are currently being explored in the above area. Content generated through interaction between two or more speakers is known as a dialogue. Dialogues are important forms of communication, which contain lot of information about ideas exchanged and nature of the participants. Dialogue summarization aims to condense a piece of content generated by multiple participants into a short passage. Dialogues are difficult to summarize since the underlying data contains diverse interactive patterns between speakers as well as inherent topic drifts (Feng et al., 2020). Human summarization sometimes focuses only on the content. sometimes gives more attention to the nature of interaction, while at others may be considering both. For example, while summarizing an argument it may be needed to capture the key points made by both the speakers separately and highlight it in the summary. For other scenarios like a customer communication it may be more important to detect dissents, agreements and the topics around which they occur. The difficulty of dialogue summarization stems from the heterogeneity of the

[1] https://huggingface.co/google/pegasus-large



Dialogue Text

#Person1#: Who stands out in your mind as a man or woman of sound character?
#Person2#: If I think of famous people, I think of Abraham Lincoln.
#Person1#: He's the US president, who walked five miles just to give a lady her change, isn't he?
#Person2#: That's the one. He also was famous for never giving up on his goals.
#Person1#: That's right. He ran for office quite a few times before he was finally elected.
#Person2#: And I also admire him for his courage in fighting for equal rights.
#Person1#: He had great vision, didn't he?
#Person2#: And humility. I would have liked to meet him personally.

Topic – sound character

Model Summary
#Person1# and #Person2# talk about who stands out in their mind as a man or woman of sound character.

Human Summary
#Person1# and #Person2# are talking about Abraham Lincoln. They think he was a noble man.

Topic – famous people

Model Summary
#Person1# and #Person2# are talking about famous people. They admire Abraham Lincoln for his great vision, courage, and humility.

Human Summary
#Person2# admires Abraham Lincoln for his perseverance, courage and humility.

Topic – discuss Abraham Lincoln

Model Summary
#Person1# and #Person2# talk about Abraham Lincoln as a man or woman of sound character.

Human Summary
#Person1# and #Person2# talk about Abraham Lincoln and his glorious history. They both admire him.

Figure 1: An example of topic focused summarization

underlying content. Dialogue summarization is an important problem that can be further classified into various sub-areas depending on the nature of input considered such as speech summarization, meeting summarization, chat summarization, email thread summarization and so on. A detailed survey on abstractive summarization is presented in (Zhong et al., 2021; Feng et al., 2021).

In recent times, masked language models using transformers that are basically multi-headed attention-based encoder-decoder models, have created a remarkable impact in the area of of text generation (Choi et al., 2019), and consequently tasks like abstractive summarization which are heavily dependent on it. PEGASUS is a transformer based model trained on large **C4** corpora introduced in (Raffel et al., 2020) containing 350M Web-pages and **HugeNews** dataset which consists of 1.5B articles from news-like website(2013-2019). Its pre-training objectives were set as Gap Sentence Generation (GSG), which was more aligned to the downstream task of summarization, and the model thereby is found to achieve much better and faster performance for abstractive summarization tasks, after some fine-tuning. In GSG, top m principal sentences, which are found to be most similar to the other sentences in the document according to ROUGE-F1 score, are initially masked while feeding the document to the model. These sentences are concatenated into a psuedo-summary, and the model is trained to generate these using a sequence to sequence generation task. This pre-training objective has pushed forward state of the art model on 12 diverse summarization datasets. It is found to perform exceptionally well on summarization tasks even when very few training samples are available for fine-tuning (Zhang et al., 2020)

In this challenge, a dialogue is found to contain information related to multiple topics. For example, “#Person2# arrives late because of traffic Jam. #Person1# suggests #Person2# quitting driving and taking public transport” contains two topics- ‘reason for being late’ and ‘benefit of public transport.’ Since the pre-trained PEGASUS model includes the salient information from the input text irrespective of user perspective, it can’t generate a topic-driven or user-perspective driven summary. The novelty of the proposed approach lies in proposing a new fine-tuning task in which a topic is passed as an input along with the dialogue text, to reformulate the task of dialogue summarization. The incorporation of the topic along with the input and a target summary during training allows for additional training of the model to generate topic-focused summaries. This enhances the quality of summary generated by PEGASUS in two ways - it learns to focus on different text segments that are centered around a given topic, and then use those portions to pick up the principal sentences. In the current context, the model learnt to focus on text segments that contained the parts of the conversation that were more relevant to the user-perspectives and thereby generated a topical summary. The significance of the proposed model is that the same text can be summarized differently based on the topics given, by focusing on different portions of the text. Fig 1 shows a sample dialogue from the test set, human-generated summaries around different topics and the outputs generated by our system for each of the given topics.

This paper is organized as follows: Section 2 gives the details of the shared task and the dataset provided. Section 3 provides a detailed description of the proposed methodology. Section 4 gives the details of baseline models and training parameters. Results are discussed in the Section 5 which is followed by the conclusion in Section 6.

2 Shared Task Details and Dataset

The DialogSum Challenge (Chen et al., 2021b) is focused on summarizing real-life dialogues. The task is to generate a fluent, concise, and coherent summary of the multi-turn dialogue text. The DialogSum dataset (Chen et al., 2021a) consists of $13,460$ dialogue conversations collected from three datasets viz Dailydialog (Li et al., 2017), DREAM (Sun et al., 2019), MuTual (Cui et al., 2020), and a few dialogues from English-speaking practice websites. This aggregated dataset [2] consists of a training set of 12460 dialogues, development set of 500 dialogues, and test set of 500 dialogues, where each dialogue was of average length 120 words. Both the training set, and the development set included a topic which usually spans over one to three words, and a human summary whose average length is 19 words. Each dialogue in the test set however had three topics and corresponding topic-focused human-generated summaries, which could be used for evaluating the model. A hidden test set with 100 dialogues and one topic each was provided as the actual challenge task.

3 The Proposed Method

For a given dialogue text $d = d_1, d_2, ...d_n$ of n words and the topic t of the conversation where $t = t_1, t_2, ..., t_k$ consists of k words, the task is to generate a dialogue summary $y = y_1, y_2, ...y_m$ containing m words. The end goal is to find the summary of a dialogue y^* that maximizes the probability $p(y|t, d)$. In order to achieve this objective, we adopt the state of the art pre-trained PEGASUS model [3], which was further fine-tuned on the downstream summarization task using the CNN/Daily News dataset (Nallapati et al., 2016). The target fine-tuning task was designed to generate the News highlights from the text.

The proposed framework used by us is shown in Figure 2 (b), while the standard one is shown in (a). We have fed the topic along with the dialogue text, where the two are separated by a special character. The target summary was a human input that came as a part of the data-set. It was observed that the topics represented human conceptualization of the content succinctly without borrowing key-words from the dialogue itself, unless necessary.

The motivation to use the topic to fine-tune the model was derived from the fact that the test dataset came with three different human summaries, formed around different topics for each dialogue. One such example dialogue with three target summaries are shown in figure1. This clearly indicated that the same conversation could be viewed from different perspectives and hence summarized differently. Though humans inherently tend to map any piece of text to topics, a human summarization tends to occur around these topics. In this dataset, the human annotation contained both the topic and the summary, which we could use to train our model in order to obtain better summaries than default PEGASUS. The idea was that using the topics as input for fine-tuning will be able to generate more topic-oriented summaries, by guiding the model towards sentences that are important for the topic and not by default ROUGE F1 similarity. Since the final hidden dataset also had a topic given, the task could clearly be modeled as one of topic-oriented summarization.

However, not all possible summarization scenarios may come with the topics explicitly mentioned, though the need may still be to do topic-focused summarization. The model in that case may be enhanced to identify the key topics first and then use them for summarization. The present dataset may serve as a good source for training a model to identify topics.

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

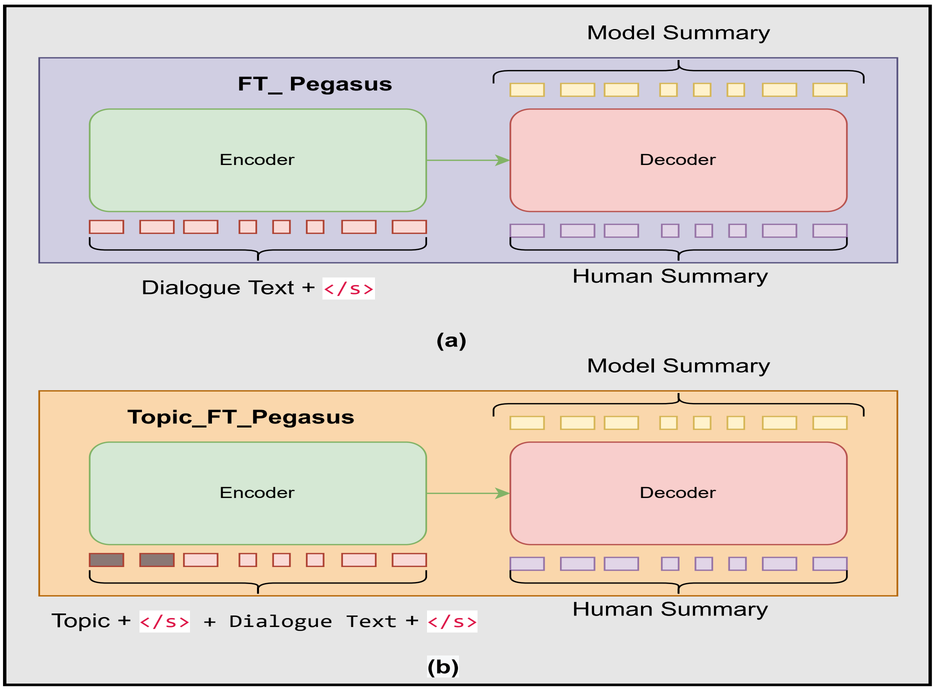


Figure 2: Proposed framework architecture

Model	Average Score				Best Score			
	R1	R2	RL	B-S	R1	R2	RL	B-S
PT_PEGASUS	25.99	6.41	20.97	87.77	37.63	9.63	26.48	88.15
FT_PEGASUS	43.36	18.36	36.23	92.19	51.59	26.58	45.54	92.64
Topic_FT_PEGASUS	**49.42**	**21.81**	**40.85**	**92.22**	**54.53**	**32.00**	**51.47**	**93.22**

Table 1: Evaluated Results over the public Dataset. R1, R2 , RL and BS stands for Rouge-1, Rouge 2, Rouge L and BERT Score respectively.

4 Experiments

This section describes the different baselines we used for comparison, followed by the training parameters used in these experiments.

4.1 Baselines

Following are the models we considered for our baselines:

1. PT_PEGASUS - In this setup, the pre-trained PEGASUS-LARGE model is adopted to generate the summary using the dialogue text as an input.

2. FT_PEGASUS - Here, the pre-trained PEGASUS-LARGE model uses the DialogSum train and development datasets. Only the dialogue text is used as an input.

4.2 Training Parameters

To fine-tune the PEGASUS model on the DialogSum dataset, training epochs is set to 10 with early stopping criteria. Since the PEGASUS is a heavy model and consumes 4 times more memory than the simple BERT model, batch_size is kept at 2 to avoid memory exhaustion. Warm-up steps are chosen at 500 with a $2e - 5$ learning rate and weight decay of 0.01.

4.3 Evaluation Metric

The results of our proposed approach and other baselines are shown in Table 1. We have reported the recall of ROUGE (Recall Oriented Understudy for Gisting Evaluation) (Lin, 2004) score. It automatically measures the quality of generated summary by counting the overlapping units like n-grams with reference summary. ROUGE-1,

[3] https://huggingface.co/google/pegasus-large

ROUGE-2 and ROUGE-L [4] have been used for the evaluation. Since rouge scores don't consider semantic similarity, hence BERTScore[5] has also been used as an evaluation metric. It leverages the pre-trained contextual embeddings from BERT and matches the conceptual similarity between the model-generated and human summaries. Since public test set contains three topics and corresponding three human summaries for each dialogue text, hence, we have generated three model summaries corresponding to each topic and reported the average and best scores among the three. It should be noted that the best score is based on the best RL score among the three human summaries.

5 Results and Discussion

When we compared our proposed approach with the baselines, we found that our model outperformed the baselines with significant improvement. ROUGE-L has increased by 4.62% compared to FT_PEGASUS. The improvements indicate that fine-tuning of the PEGASUS model on the DialogSum dataset and topic relevance helped the model in extracting the essential information from the dialogues. We also computed the average length difference between our outputs and ground-truth summaries as recall depends on the length of generated summary. The average length of our model-generated summaries is 22.28 words, which is comparable to the ground-truth summaries, whose average length was 19.99 words.

6 Conclusion

As part of the DialogSum shared task on learning to generate a concise, fluent and topic-oriented summary of dialogues picked up from real-life scenarious, we have enhanced the performance of a pre-trained abstractive summarizer model by incorporating the topic along with the input text, to generate a topic-oriented summary. We have shown that the SOTA pre-trained transformer-based encoder-decoder model PEGASUS can be fine-tuned using the proposed methodology, to generate more human-like summaries of dialogues. Our model performed better in comparison to the baselines. In future, we plan to improve the method further by incorporating nuances of dialogue, speech act theory etc. The model can also be trained to learn the topic before generating a summary.

[4] https://github.com/cylnlp/dialogsum/blob/main/Baseline/rouge.py

[5] https://huggingface.co/metrics/bertscore

References

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

Yulong Chen, Yang Liu, and Yue Zhang. 2021b. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

Hyungtak Choi, Lohith Ravuru, Tomasz Dryjański, Sunghan Rye, Donghyun Lee, Hojung Lee, and Inchul Hwang. 2019. Vae-pgn based abstractive model in multi-stage architecture for text summarization. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 510–515.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

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

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *arXiv preprint arXiv:2012.03502*.

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*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. 2018. Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Akshara Rai, Suyash Sangwan, Tushar Goel, Ishan Verma, and Lipika Dey. 2021. Query specific focused summarization of biomedical journal articles. In *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 91–100. IEEE.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.