A Multi-Task Learning Approach for Summarization of Dialogues

Saprativa Bhattacharjee
Department of Information Technology
Government Polytechnic Daman
India
saprativa.bhatt@gov.in

Kartik Shinde
Department of Civil Engineering
Indian Institute of Technology Patna
India
kartik_1901ce16@iitp.ac.in

Tirthankar Ghosal
Charles University
MFF, ÚFAL
Czech Republic
ghosal@ufal.mff.cuni.cz

Asif Ekbal
Department of Computer Science and Engineering
Indian Institute of Technology Patna
India
asif@iitp.ac.in

Abstract

We describe our multi-task learning based approach for summarization of real-life dialogues as part of the *DialogSum Challenge* shared task at INLG 2022. Our approach intends to improve the main task of abstractive summarization of dialogues through the auxiliary tasks of extractive summarization, novelty detection and language modeling. We conduct extensive experimentation with different combinations of tasks and compare the results. In addition, we also incorporate the topic information provided with the dataset to perform topic-aware summarization. We report the results of automatic evaluation of the generated summaries in terms of ROUGE and BERTScore.

1 Introduction

Much of the early works on summarization devoted attention to *monologues* such as news articles (Nallapati et al., 2016; Narayan et al., 2018), patents (Sharma et al., 2019), Wikipedia articles (Liu et al., 2018; Cohen et al., 2021), scientific research papers (Cohan et al., 2018), Government reports (Huang et al., 2021) and even court judgements (Gao et al., 2019). But more recently, the focus of the summarization community has started shifting from monologues to *dialogues* largely owing to the rising popularity of chatbots, personal assistants, instant messaging platforms and online meetings. While monologues are characterised by the fact that they are authored by a single person, a dialogue involves the utterances of more than one participant (which alone can make them inherently more difficult to summarize). However, the available dialogue summarization datasets (Gliwa et al., 2019; Zhu et al., 2021; Feigenblat et al., 2021) are fewer in number, limited in scale, domain-specific and sometimes even extremely noisy and semi-structured (Carletta et al., 2005; Janin et al., 2003) as compared to the datasets available for monologue texts.

To mitigate these issues a high-quality large-scale dialogue summarization dataset named *DialogSum* was released by Chen et al. (2021a). The dataset consists of a wide variety of task-oriented dialogues from daily-life conversations. One sample dialogue and its corresponding summary from DialogSum's training set is presented in Figure 1, which is a conversation between a doctor and his patient on the topic of getting a check-up. To further encourage research in dialogue summarization, the authors proposed a shared task named DialogSum Challenge (Chen et al., 2021b) as part of INLG 2022, and in this article, we describe our submission to the shared task as Team IITP-CUNI.

Specifically, we attempt to tackle the problem of abstractive dialogue summarization through the use of a mutli-task learning model (Ruder, 2017; Crawshaw, 2020; Vandenhende et al., 2020) based on Transformers (Vaswani et al., 2017). We intend to improve the main task of abstractive summarization of the dialogues through the auxiliary tasks

Dialogue:
#Person1#: Hi, Mr. Smith. I'm Doctor Hawkins. Why are you here today?
#Person2#: I found it would be a good idea to get a check-up.
#Person1#: Yes, well, you haven't had one for 5 years. You should have one every year.
#Person2#: I know. I figure as long as there is nothing wrong, why go see the doctor?
#Person1#: Well, the best way to avoid serious illnesses is to find out about them early. So try to come at least once a year for your own good.
#Person2#: Ok.
#Person1#: Let me see here. Your eyes and ears look fine. Take a deep breath, please. Do you smoke, Mr. Smith?
#Person2#: Yes.
#Person1#: Smoking is the leading cause of lung cancer and heart disease, you know. You really should quit.
#Person2#: I've tried hundreds of times, but I just can't seem to kick the habit.
#Person1#: Well, we have classes and some medications that might help. I'll give you more information before you leave.
#Person2#: Ok, thanks doctor.

Summary:
Mr. Smith's getting a check-up, and Doctor Hawkins advises him to have one every year. Hawkins'll give some information about their classes and medications to help Mr. Smith quit smoking.

Topic:
get a check-up

Figure 1: A sample dialogue-summary pair along with the topic information from the DialogSum dataset's training set.

of extractive summarization, novelty detection and language modeling. Additionally, we also explore the usefulness of topic-aware summarization, as in the DialogSum dataset, topics are provided along with the summaries (see Figures 1 and 2).

The rest of the paper is organised as follows. Related work is presented in Section 2. The DialogSum Challenge is described in details in Section 3. Section 4 presents our system. Results and discussion are in Section 5. Finally, the conclusion is drawn in Section 6.

2 Related Work

In this section, we discuss some of the most recent works on dialogue summarization and multi-task learning strategies for abstractive summarization. For long dialogue summarization, Zhong et al. (2021) proposed a window-based pre-training strategy using five different types of dialogue-related noise – speaker mask, turn splitting, turn merging, text infilling and turn permutation. At first, the window is corrupted with noise, and then the model is tasked with de-noising and reconstructing the window. On the other hand, Zhang et al. (2022) utilize a multi-stage approach for dealing with long dialogues. In the preliminary stages, they segment the input and produce coarse summaries, while in the final stage, the coarse summaries are used to generate the final fine-grained summary. Zhang et al. (2021) studied the effectiveness of different strategies to deal with long dialogues and concluded that a retrieve-then-summarize pipeline model works better in comparison to Longformer (Beltagy et al., 2020) or HMNet (Zhu et al., 2020). However, in the case of DialogSum, as the input data is well within the limit of the popular pre-trained Transformer models such as BART (Lewis et al., 2020), we are not faced with any such issues. Moreover, Chen et al. (2021a) have shown that the larger version of BART performs better than others on DialogSum. We start our investigation with this strong baseline.

Another direction of work has been the incorporation of topic information to further improve the abstractive dialogue summarization. In this direction, Zou et al. (2021) proposed a novel topic-augmented two-stage dialogue summarizer (TDS) along with a saliency-aware neural topic model (SATM) to perform topic-aware summarization of customer service dialogues. Qi et al. (2021) fused the topic segmentation embedding along with positional embedding in the utterance-level encoder input of a hierarchical Transformer architecture. To capture the topic information of dialogues Liu et al. (2021) came up with two contrastive learning strategies, namely coherence detection and sub-summary generation. And all of them reported performance benefits of taking topic information into account while performing abstractive summarization. We too explore the topic-aware summarization as the DialogSum dataset provides topic information along with the summaries.

A slightly different but closely related task that deserves mention is that of automatic minuting of meeting transcripts. The first shared task on Automatic Minuting (AutoMin) (Ghosal et al., 2021a) at Interspeech 2021 and the SIGDial 2021 Special Session on Summarization of Dialogues and Multi-Party Meetings (SummDial) (Ghosal et al., 2021b) brought out a plethora of interesting works targeting the task such as the attempt to use BART for generation of readable minutes (Shinde et al.,

Dialogue:	**Summary 1:**	**Summary 2:**	**Summary 3:**
#Person1#: Ms. Dawson, I need you to take a dictation for me. #Person2#: Yes, sir… #Person1#: This should go out as an intra-office memorandum to all employees by this afternoon. Are you ready? #Person2#: Yes, sir. Go ahead. #Person1#: Attention all staff... Effective immediately, all office communications are restricted to email correspondence and official memos. The use of Instant Message programs by employees during working hours is strictly prohibited. #Person2#: Sir, does this apply to intra-office communications only? Or will it also restrict external communications? #Person1#: It should apply to all communications, not only in this office between employees, but also any outside communications. #Person2#: But sir, many employees use Instant Messaging to communicate with their clients. #Person1#: They will just have to change their communication methods. I don't want any - one using Instant Messaging in this office. It wastes too much time! Now, please continue with the memo. Where were we? #Person2#: This applies to internal and external communications. #Person1#: Yes. Any employee who persists in using Instant Messaging will first receive a warning and be placed on probation. At second offense, the employee will face termination. Any questions regarding this new policy may be directed to department heads. #Person2#: Is that all? #Person1#: Yes. Please get this memo typed up and distributed to all employees before 4 pm.	Ms. Dawson helps #Person1# to write a memo to inform every employee that they have to change the communication method and should not use Instant Messaging anymore.	In order to prevent employees from wasting time on Instant Message programs, #Person1# decides to terminate the use of those programs and asks Ms. Dawson to send out a memo to all employees by the afternoon.	Ms. Dawson takes a dictation for #Person1# about prohibiting the use of Instant Message programs in the office. They argue about its reasonability but #Person1# still insists.
	Topic 1: communication method	**Topic 2:** company policy	**Topic 3:** dictation

Figure 2: A sample from the DialogSum test set which contains one dialogue and the three reference summaries along with three topics corresponding to each summary.

2021). Singh et al. (2021) present an empirical analysis of the state-of-the-art summarization models for the task of generating meeting minutes and arrive at the conclusion that they are far from being satisfactory. A novel dataset of meetings in English and Czech (Nedoluzhko et al., 2022) is also being released to further encourage the research community to take up the challenging task.

Lee et al. (2021) claim to be the first ones to have applied multi-task learning to dialogue summarization task. Leveraging Part-of-Speech (PoS) information, they constructed a syntax-aware dialogue summarization model on SAMSum corpus (Gliwa et al., 2019). The main intuition behind their approach is that different speaker roles are characterised by different syntactic structures (voiceprints), which could be captured via POS information. More recently, for low-resource datasets Magooda et al. (2021) experimented with several combinations of auxiliary tasks for abstractive summarization in a multi-task setting. They concluded that a certain combination of tasks indeed improved the abstractive summarization results across different datasets and models. Prior to these, in the multi-task setting, the primary task of abstractive summarization has been combined and experimented with several other auxiliary tasks such as entailment generation (Pasunuru et al., 2017); question generation and entailment generation (Guo et al., 2018); extractive summarization (Chen et al., 2019); text categorization and syntax labeling (Lu et al., 2019); dialogue act classification and extractive summarization (Manakul et al., 2020); keyword extraction and key-sentence extraction (Xu et al., 2020). Very recently, Chen et al. (2022) formulated the five different tasks of dialogue understanding (DU) as a unified generation task. These tasks include dialogue summarization, dialogue completion, dialogue state tracking, slot filling and intent detection. Then they experimented with eight different multi-task training strategies and concluded that their proposed method achieves superior performance on both few-shot as well as zero-shot settings. These encouraging results of the multi-task learning strategies on abstractive summarization motivated us to apply the same to the DialogSum Challenge.

3 DialogSum Challenge

In this section, we give a brief overview of the DialogSum Challenge by first describing the dataset and then going through the task description.

3.1 Dataset Description

The DialogSum dataset consists of a total of 13,460 dialogue-summary pairs, out of which 12,460 (92.6%) are in the training set, 500 (3.7%) in the development set and 500 (3.7%) more in the test set, as depicted in Figure 3. The dialogue data has been collected from multiple sources, namely 58.22% from DailyDialogue dataset (Li et al., 2017), 16.94% from DREAM dataset (Sun et al.,

Split	#Dialogues	#Turns	Turn Len.	Dialogue Len.	Summary Len.	%-Compression
train	12460	9.49	20.10	191.37	29.36	83.72
dev	500	9.38	20.17	188.89	27.21	84.74
test	500	9.71	20.04	196.12	23.76	86.70
hidden	100	10.88	19.03	209.42	–	–

Table 1: DialogSum dataset split statistics. '#Dialogues' contains absolute values while rest of the columns report average values. 'Len.' stands for Length. 'hidden' is the hidden test set for which only the dialogues and topics have been released publicly and hence the Summary Length and %-Compression details are not available.

2019), 13.89% from MuTual dataset (Cui et al., 2020) and the rest have been crawled from English speaking practice websites. The dialogues revolve around real-life conversations on topics such as schooling, work, medication, shopping, leisure and travel. The data from these varied sources are cleaned and transformed into a unified format before being annotated.

Some statistics of interest for each split of the dataset are presented in Table 1. Although the training, development and test sets are quite similar in terms of the average number of turns and the average turn length, the test set average dialogue length is larger while the average summary length is smaller than the other two sets. This also gets reflected in the test set's marginally higher compression ratio. Moreover, the average dialogue length of the hidden test set is higher than all other sets, but this may be attributed to the smaller size of the hidden set. In training and development sets, for each dialogue, one human written summary is provided. Figure 1 shows an example dialogue-summary pair from the training set. In addition to the summary, the human annotators also provide the topic information. On the other hand, for each dialogue in the test set, three human written reference summaries are provided. Figure 2 shows an example dialogue from the test set and its three reference summaries. For each reference summary, its corresponding topic is also provided.

In addition to the above, the organizers have also released a hidden test set consisting of 100 dialogues. Only the dialogues and topic information are provided for this hidden set, while the summaries have not been made public. The organizers will use this set for evaluation of the submitted models.

3.2 Task Description

The shared task participants need to design a model which will take as input the dialogue text and produce the corresponding abstractive summary. For automatic evaluation, each system-generated summary will be evaluated against the three human written reference summaries and the average ROUGE scores (Lin, 2004) and BERTScore (Zhang et al., 2020) will be used to determine the position on the DialogSum Challenge's leaderboard. Out of these two metrics, ROUGE (R1, R2 and RL) will be used as the primary metric, while BERTScore will be used as a supplementary metric. Additionally, the generated summaries will also be evaluated against the human-written summaries of the hidden test set. The lowest, highest and averaged scores will be reported for both the multi-reference test sets.

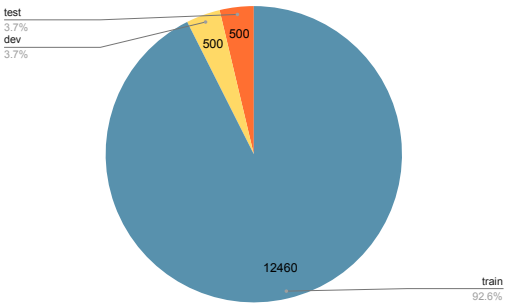

Figure 3: DialogSum dataset distribution.

For human evaluation, the submitted summaries will be judged on the following parameters: (i) fluency, consistency, relevance and coherence; (ii) co-reference information; (iii) intent identification; (iv) discourse relation; and (v) objective description. For more details about these parameters, we would like to refer the readers to the shared task paper (Chen et al., 2021b).

4 Our System

We employ a multi-task learning approach for the DialogSum Challenge. In multi-task learning, a machine learning model is trained simultaneously on more than one related task (Crawshaw, 2020).

Usually, there is a main task and one or more auxiliary tasks. In our case, the main task is abstractive summarization and the auxiliary tasks are extractive summarization, novelty detection and language modeling. There are many variants of multi-task learning. In this work, we employ a hard parameter sharing (Ruder, 2017) Transformers-based architecture in which all tasks share the same encoder layers but have task-specific decoder and/or LM head(s). The multi-task model architecture is depicted in Figure 4. It consists of a single BART encoder which is shared amongst all the tasks. The BART decoder is used for the main task of abstractive summarization, while task-specific heads are used for each of the respective auxiliary tasks. We now describe each of the tasks of our model one-by-one:

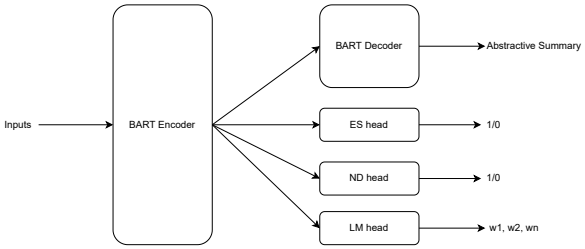

Figure 4: The multi-task learning model based on BART. AS: abstractive summarization; ES: extractive summarization; ND: novelty detection; LM: language modeling.

Abstractive Summarization (AS): For the main task of abstractive summarization, the transcripts are given as input to the BART encoder and the abstractive summaries are obtained as output from the BART decoder. This is a sequence-to-sequence task accomplished with the encoder-decoder architecture. In cases where we want to run only the single task for establishing the baseline, only this task is undertaken while keeping all other auxiliary tasks inactive through the training parameters.

Extractive Summarization (ES): The task of extractive summarization is formulated as a classification task where the goal is to classify a given sentence as either belonging to or not belonging to the extractive summary. The inputs are given in the format `[CLS] SW1, SW2, ..., SWn [SEP] CW1, CW2, ..., CWm`. Here, `[CLS]` is the start token, `[SEP]` is the separator token, `SW1...SWn` is the sentence to be classified as belonging to the extractive summary or not and `CW1...CWm` is the context around the sentence `SW1...SWn`. The sentence and the context around it are chosen in such a way that the maximum combined length does not exceed 1024 tokens.

Novelty Detection (ND): Novelty detection in NLP refers to the identification of novel text, i.e., text containing new information (Ghosal et al., 2022). This task is also formulated as a classification task. For this task, we use data from three different sources: (i) Quora Question Pair (QQP) dataset[1] consisting of more than 400 thousand question pairs. Each such pair is annotated with a binary value which indicates whether or not the questions in the pair are duplicates of each other. (ii) Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) is a corpus consisting of 5,801 sentence pairs from news articles where each pair is annotated by humans as being either a paraphrase or not and (iii) data created from the three reference summaries given in the public test set of DialogSum. We assume that the three reference summaries are paraphrases (non-novel) of each other. Since there are 500 dialogues, each with three reference summaries, we obtain 1,500 non-novel samples. We also extract a similar number of novel samples by taking summaries from two different dialogues, as shown in Table 2. The input is given in the form `[CLS] source text [SEP] target text`, and the task of the model is to classify the pair as either novel or non-novel (duplicates).

Source	Target	Novel
Ref. Summary 1	Ref. Summary 2	0
Ref. Summary 2	Ref. Summary 3	0
Ref. Summary 1	Ref. Summary 3	0
Ref. Summary (Dn)	Ref. Summary (Dm)	1

Table 2: Novelty dataset created from the three reference summaries provided in the public test set of DialogSum. Ref. Summary (Dn) & Ref. Summary (Dm) denotes reference summaries from different dialogues.

Language Modeling (LM): We perform masked language modeling on the gold summaries from the training set as per the training strategy adopted by Devlin et al. (2019). For this, 15% of the input tokens are masked and out of this, 80% are replaced by special tokens, 10% with random words and the remaining 10% are left unchanged.

[1] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

5 Results and Discussion

In this section, we first describe the experimental setup used and then present the results. Finally, we analyse the summaries generated by our best-performing model.

5.1 Experimental Setup

We run all the experiments on two NVIDIA A100-PCIE-40GB GPUs using a batch size of 4 for both training and evaluation and mostly use the default values for hyperparameters. The BART model is initialized with `facebook/bart-large`[2] and then finetuned using task-specific datasets. Mixed-precision training using fp16 is utilized for faster training and lesser memory footprint. We make use of the summarization script released by Hugging Face[3] and the multi-task learning ideas introduced by Magooda et al. (2021). The ROUGE evaluations are done using `py-rouge`[4] and BERTScore evaluations using `bert_score`[5] as suggested by the organizers of DialogSum Challenge.

5.2 Results

We provide all the results from our experiments in Table 3. The reported performance is the average of the scores of system-generated summaries with respect to the three reference summaries provided in the public test set. We consider the single-task setting where only abstractive summarization (AS) is done without any auxiliary tasks as the baseline. For the topic-aware abstractive summarization (AS[T]), we supply the topic information by prepending it to the input dialogue to the BART encoder as `[CLS] TOPIC [SEP] Dialogue`. We observe a marginal improvement in the scores using this strategy.

In the multi-task setting, we experiment with different combinations of tasks as well as data. The best ROUGE scores are obtained when abstractive summarization is done along with extractive summarization (ES), while the best BERTScore is obtained when abstractive summarization is combined with novelty detection (ND). Since extractive summaries were not provided with the Dialgosum dataset, we used `bert-extractive-summarizer`[6] to obtain the same. Alongside the newly created extractive data from DialogSum, we also experiment with the extractive summary data from AMI (Carletta et al., 2005). Results show that the model trained with auxiliary task of extractive summarization (from AMI) outperforms all others. To explain such a performance, we analyze the outputs and test other configurations with both extractive datasets. However, in our observation, there are no apparent reasons for the model to perform in such a manner on AMI data. Finally, we account this to the fact that AMI is a dataset of meeting transcript and summaries, in which the information is widely dispersed throughout the discourse of the transcript, which have a lot of redundancies. While, dialogues from the DialogSum dataset are relatively shorter, with lesser redundant texts. Moreover, most of the lines from these dialogues (even those that are coherent with parts of summary), have a generic fashion of day-to-day speech. Hence, the BART model learns better from the extractive data from AMI.

Model	R1	R2	RL	BERTScore
Single-Task				
AS	46.15	20.41	43.93	92.40
AS[T]	46.91	20.28	44.26	92.38
Multi-Task				
AS+ES	46.24	19.42	43.54	92.40
AS+ES(AMI)	**47.26**	**21.18**	**45.17**	92.60
AS+ND(QQP)	46.62	20.12	44.09	**92.72**
AS+LM	45.11	18.92	43.08	92.30
AS+ES+ND(MRPC)	46.85	19.96	44.43	92.57
AS+ES(AMI)+ND	46.60	19.90	44.03	92.40
AS+ES(AMI)+ND(QQP)	46.73	20.30	44.44	92.43
AS+ES+LM	45.51	19.73	43.90	92.52
AS+ND(MRPC)+LM	45.14	19.60	43.20	92.26
AS+ES+ND(MRPC)+LM	45.62	19.80	44.10	92.60

Table 3: Results of single-task and multi-task models on the public test set of the DialogSum dataset. AS: abstractive summarization; ES: extractive summarization; ND: novelty detection; LM: language modeling; AS[T]: topic-aware abstractive summarization; ES(AMI): extractive summarization with AMI data; ND(MRPC): novelty detection with MRPC data; ND(QQP): novelty detection with Quora Question Pair data.

5.3 Analysis

We take our best performing model and manually analyse the summaries generated by it. Figure 5 and Figure 6 present the worst three and best

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

[3] https://github.com/huggingface/transformers/tree/main/examples/pytorch/summarization

[4] https://pypi.org/project/py-rouge/

[5] https://github.com/Tiiiger/bert_score

[6] https://pypi.org/project/bert-extractive-summarizer/

R1	Model Generated Summary	Reference Summaries
0.19	Person1 warns Person2 Person2 will be arrested if Person2 calls Person1 again.	Person1 is angry about the crank calls.
		Person1 gets a crank call and is angry about it.
		Person1 receives a phone call but no one speaks.
0.21	Person1 and Person2 meet each other for the first time. Person1 finds out they have met before. Person2 has to go.	Person1 thinks that she knows Person2 somewhere, but Person2 denies it.
		Person1 thinks she has met Person2 somewhere, but Person2 thinks it's a mistake.
		Person1 keeps asking where Person2's from because she thinks she knows Person2 but Person2 denies it.
0.21	Person1 tells Tony that everything has been going wrong lately in the toy department of the shopping center. Person1 thinks Christmas does not mean much now except more work and more headaches.	Person1 complains to Tony that Christmas has made Person1 busier.
		Person1 works as a toy salesperson and feels so tired recently because Christmas is coming, and everyone's shopping for presents.
		Person1 thinks selling gifts for kids is such an unpleasant job before Christmas.

Figure 5: The worst three model-generated summaries in terms of ROUGE-1.

R1	Model Generated Summary	Reference Summaries
0.89	Person1 congratulates Mr. Stuart on his winning the city marathon.	Person1 congratulates Mr. Stuart on winning a marathon.
		Person1 congratulates Mr. Stuart on winning the city marathon.
		Person1 congratulates Mr. Stuart on winning the city marathon.
0.83	Mr. Lee gives Mrs. Word a lift home.	Mr. Lee gives Mrs. Word a lift home.
		Mr. Lee gives Mrs. Word a lift home on a rainy night.
		Mr. Lee offers to give Mrs. Word a lift home on a terrible night.
0.81	Person2 shows Person1 the way to the central department stall and the national bank.	Person1 gets lost and asks Person2 where the central department stall and the national bank are. Person2 directs Person1.
		Person2 shows Person1 the ways to the central department stall and the national bank.
		Person1 asks Person2 the way to the central department stall and the national bank.

Figure 6: The best three model-generated summaries in terms of ROUGE-1.

three summaries generated by the model in terms of ROUGE-1, respectively. It is to be kept in mind that the ROUGE scores reported are the average of the generated summary with respect to the three reference summaries. Let us first consider the case of the three worst summaries shown in Figure 5. In the case of the first system-generated summary, we can see that it is longer than each one of the three reference summaries and the content is quite different. In the second case, our model is unable to figure out that Person1 "thinks" she met/knows Person2. Rather the model generates the phrase "finds out". Moreover, the last line, "Person2 has to go" is totally unnecessary for the summary. In the case of the third summary, although the system-generated summary conveys the same message as the reference summaries, yet the same is not reflected in terms of ROUGE-1 mainly because of the different set of unigrams used.

Let us now consider the best three summaries generated by our model as shown in Figure 6. In all three cases, it can be seen that the generated summary matches almost exactly to one of the three reference summaries. The second system-generated summary matches word-to-word with its first reference summary, while the first and third system-generated summaries differ with their respective best matches on only a single word. The

higher score of the first summary can be attributed to the fact that two out of the three reference summaries in this case turn out to be exactly the same, which takes the average score up.

6 Conclusion

In this paper, we describe our submission to the shared task on dialogue summarization named DialogSum Challenge at INLG 2022. DialogSum consists of 13,460 real-life scenario dialogues. We employ a multi-task learning approach for the task and achieve considerable improvement over the single-task baseline. Our best performing model is the multi-task combination of abstractive summarization as the main task and extractive summarization as the auxiliary task. We also incorporate the topic information supplied alongside the summaries to gain marginal improvement in performance over the baseline. In future work, we would like to experiment with other tasks to find the optimal combination. We would also like to explore methods other than multi-task learning for improving the abstractive summarization of dialogues.

Acknowledgements

Tirthankar Ghosal would like to acknowledge the support from the grant 19-26934X (NEUREM3) of the Czech Science Foundation and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 for the project European Live Translator (ELITR).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Yangbin Chen, Yun Ma, Xudong Mao, and Qing Li. 2019. Multi-task learning for abstractive and extractive summarization. *Data Sci. Eng.*, 4(1):14–23.

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, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Zhi Chen, Lu Chen, Bei Chen, Libo Qin, Yuncong Liu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022. Unidu: Towards A unified generative dialogue understanding framework. *CoRR*, abs/2204.04637.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. WikiSum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 212–219, Online. Association for Computational Linguistics.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *CoRR*, abs/2009.09796.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shen Gao, Xiuying Chen, Piji Li, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2019. How to write summaries with patterns? learning towards abstractive summarization through prototype editing. 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 3741–3751, Hong Kong, China. Association for Computational Linguistics.

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021a. Overview of the first shared task on automatic minuting (automin) at interspeech 2021. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–25.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, 48(1):77–117.

Tirthankar Ghosal, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2021b. Report on the SIGDial 2021 special session on summarization of dialogues and multi-party meetings (summdial). *ACM SIGIR Forum*, December 2021:1–17.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, pages 364–367. IEEE.

Seolhwa Lee, Kisu Yang, Chanjun Park, João Sedoc, and Heuiseok Lim. 2021. Who speaks like a style of vitamin: Towards syntax-aware dialogue summarization using multi-task learning. *IEEE Access*, 9:168889–168898.

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, Online. Association for Computational Linguistics.

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, Barcelona, Spain. Association for Computational Linguistics.

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

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yao Lu, Linqing Liu, Zhile Jiang, Min Yang, and Randy Goebel. 2019. A multi-task learning framework for abstractive text summarization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9987–9988. AAAI Press.

Ahmed Magooda, Diane Litman, and Mohamed Elaraby. 2021. Exploring multitask learning for low-resource abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*,

pages 1652–1661, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Potsawee Manakul, Mark J. F. Gales, and Linlin Wang. 2020. Abstractive spoken document summarization using hierarchical model with multi-stage attention diversity optimization. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4248–4252. ISCA.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR Minuting Corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of The 13th Language Resources and Evaluation Conference*, page To Appear.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2017. Towards improving abstractive summarization via entailment generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 27–32, Copenhagen, Denmark. Association for Computational Linguistics.

MengNan Qi, Hao Liu, YuZhuo Fu, and Ting Liu. 2021. Improving abstractive dialogue summarization with hierarchical pretraining and topic segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1121–1130, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team abc @ automin 2021: Generating readable minutes with a bart-based automatic minuting approach. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–8.

Muskaan Singh, Tirthankar Ghosal, and Ondrej Bojar. 2021. An empirical performance analysis of state-of-the-art summarization models for automatic minuting. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 50–60, Shanghai, China. Association for Computational Linguistics.

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.

Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. 2020. Revisiting multi-task learning in the deep learning era. *CoRR*, abs/2004.13379.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Weiran Xu, Chenliang Li, Minghao Lee, and Chi Zhang. 2020. Multi-task learning for abstractive text summarization with key information guide network. *EURASIP J. Adv. Signal Process.*, 2020(1):16.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *CoRR*, abs/2109.02492.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14665–14673. AAAI Press.