Dialogue Summarization using BART

Conrad Lundberg* and **Leyre Sánchez Viñuela*** and **Siena Biales***
University of Tübingen
conrad.lundberg@student.uni-tuebingen.de
leyre.sanchez-vinuela@student.uni-tuebingen.de
siena.biales@student.uni-tuebingen.de

Abstract

This paper introduces the model and settings submitted to the INLG 2022 DialogSum Challenge, a shared task to generate summaries of real-life scenario dialogues between two people. In this paper, we explored using intermediate task transfer learning, reported speech, and the use of a supplementary dataset in addition to our base fine-tuned BART model. However, we did not use such a method in our final model, as none improved our results. Our final model for this dialogue task achieved scores only slightly below the top submission, with hidden test set scores of 49.62, 24.98, 46.25 and 91.54 for ROUGE-1, ROUGE-2, ROUGE-L and BERTSCORE respectively. The top submitted models will also receive human evaluation.

1 Introduction

Dialogue summarization is a variation of text summarization which aims to generate concise, coherent summaries of conversations. Dialogue summarization requires far deeper insight than summarizing a news article or similar documents, as is done in text summarization. When handling a dialogue, a model must address semantic roles, resolve definite pronouns and coreference, and handle various other complexities (Chen et al., 2021b). We investigate the best methods for summarizing a dialogue while retaining these difficult relations that do not present a problem when summarizing a simple text.

The INLG 2022 DialogSum Challenge is a shared task with the goal of generating summaries of real-life scenario dialogues between two people. In this paper, we will describe our approach to this task using a fine-tuned BART model. Additionally, we explore the effects of using intermediate task transfer learning, reported speech for this task. However, we did not use such a method in our final model, as none improved our results.

* All authors contributed equally.

2 Background

The field of text summarization has been in focus for decades. Research into automatic text summarization began as early as 1958 with the summarization of magazine articles and technical papers (Luhn, 1958). Text summarization proves challenging for many reasons. The model must be able to identify important topics and condense them in a way that is not redundant, but yet remains readable and cohesive (El-Kassas et al., 2021). Primarily, there are two approaches to text summarization: extractive and abstractive. Extractive summarization seeks to extract the most important information and present it as is. Abstractive summarization, in contrast, may use novel words to create a linguistically correct condensed representation (Zhang et al., 2020). Originally, research on extractive summarization was in the foreground (Murray et al., 2005) but the field is now moving towards abstractive summarization based on neural sequence-to-sequence encoder-decoder models (Sutskever et al., 2014). Top performing models to create summaries have also been based on transformers (Vaswani et al., 2017). Pointer-generator models (See et al., 2017) are another state-of-the-art summarization technique, combining extractive and abstractive methods.

Dialogue summarization is now emerging as a new interest in the field of natural language processing. As early as 2010, Higashinaka et al. were exploring methods of extractive summarization to summarize contact center dialogues using a hidden Markov model called Class Speaker HMM. Since then, more unique and effective methods have emerged. Yuan and Yu (2019) proposed a Scaffold Pointer Network (SPNet), which incorporated three types of semantic scaffolds found in dialogue: speaker role, semantic slot, and dialog domain. Chen and Yang (2020) introduced a multi-view sequence-to-sequence model, which utilized

conversational structures and topic segmentation to assist in better dialogue summarization.

BART is a sequence-to-sequence model that pre-trains by combining Bidirectional and Auto-Regressive Transformers, and achieves good results on a range of abstractive dialogue and summarization tasks (Lewis et al., 2019). Khalifa et al. (2021) found BART to be a viable base model for dialogue summarization and showed additional methods could improve results. For this reason, we selected BART as our base model.

3 System Overview and Methods

In this section, we discuss the setup and hyperparameters of our final model, as well as attempts to improve our results, which included using intermediate task transfer learning, reported speech, and an additional dataset.

3.1 Setup and Hyperparameter Tuning

Our model was made by fine tuning a BART model on 12460 dialogue/summary pairs in the DIALOGSUM dataset provided by the INLG 2022 DialogSum Challenge (Chen et al., 2021a). The training and validation datasets provided to us contain a dialogue, a gold summary, an identifier, and a topic. The dialogue is formatted such that each line represents one dialogue turn. The lines begin with either $\#Person1\#$: or $\#Person2\#$: to identify who is speaking. We pass the full dialogue to the model as input without any further preprocessing apart from randomization of the dataset and tokenization.

In the training dataset, there were 7434 unique topics provided. Some examples of the most common topics are "shopping", "job interview", or "phone call", but even these were only found in about 100 of the 12,460 training instances. The least common topics were only found on one instance and include "job losing", "look ill", "stop doing business", or "the language club". While the topic data could prove useful, we discarded the topic for the purposes of this task.

The BART model described in this paper was first fine-tuned on the CNN/Dailymail corpus (Hermann et al., 2015). We used an NVIDIA Tesla P100 16GB GPU to train our fine-tuned model.

When tuning our hyperparameters, we began with the most impactful settings and documented improvements on each training iteration. In initial training runs with a high learning rate, the model outputted only a few words repeatedly, and appeared overfitted. We opted to use the same learning rate as the task organizers documented in their hyperparameter settings (3e-5) for our final model.

Our best model used a batch size of 2. Other batch sizes (e.g. 3,4,8) were also tested, yet with our settings, using larger batch sizes did not improve results. We trained our model for 3 epochs on the full training dialogue dataset with no early stopping.

3.2 Post-processing

When decoding the generated summary, important adjustments included the minimum and maximum summary lengths, along with a length penalty parameter, which penalizes longer summaries. A very low value for the length penalty tells the model to generate shorter sequences. The perfect summary length is subjective, but these parameters helped to obtain results that were most similar to the target test set summaries. In our final model, we used a minimum length of 14, a maximum length of 64, and a length penalty of 0.04.

To reduce hallucinations in the transformer model, we preemptively replace any instances of speakers who did not appear in the initial dialogues, such as $\#Person3\#$ or $\#Person4\#$, to $\#Person1\#$ or $\#Person2\#$. In addition, we fixed any instances of duplicate labels, such as $\#Person1\#Person1\#$ or $\#Person2\#Person2\#$.

3.3 Intermediate Task Transfer Learning

We experimented with the use of intermediate task transfer learning for this task. Pruksachatkun et al. (2020) studied the effects of multiple intermediate tasks on a variety of target tasks trained on RoBERTa. Although none of the target tasks in the paper were related to text or dialogue summarization, there were some intermediate tasks (Cosmos QA, HellaSwag) that improved target task results across the board, regardless of the task. We decided to investigate the use of one of these generally successful intermediate tasks, HellaSwag, on the dialogue summarization task to see if we could observe any improvement.

The HellaSwag dataset (Zellers et al., 2019) is a natural language inference dataset modeled as multiple-choice questions, where there are four possible answers for continuing the scene set in the "question". This task is easy for humans to

determine the correct sentence continuation given the context in the initial sentence, but computers struggle to achieve the same success. In order to alter the HellaSwag question-answer dataset into a sequence-to-sequence problem that our model could solve, we opted to remove all the negative answers and treat the context sentence as the initial sequence, with the correct answer choice as the target sequence.

We trained our BART model for 1 epoch on 10% of the HellaSwag training split and then trained the same model on the DIALOGSUM training dataset exactly as described previously. Unfortunately, the ROUGE scores were all consistently lower using this technique. ROUGE-1, ROUGE-2, and ROUGE-L dropped by 1.2, 1.9, and 1.1 points respectively.

Although training with HellaSwag as an intermediate task did not yield positive results, we also attempted intermediate task transfer learning on a more similar task, namely, news article summarization. For this, we used a portion of the XSum dataset (Narayan et al., 2018). The XSum dataset contains a series of news articles along with one-sentence summaries of each article, making it already ideal for a sequence-to-sequence task with no preprocessing required. Similarly to the HellaSwag dataset, we first trained our BART model for 1 epoch on the XSum training split, and then used this to train on the DIALOGSUM training dataset. Unfortunately, this also resulted in consistently lower ROUGE scores. ROUGE-1, ROUGE-2, and ROUGE-L dropped by 1.5, 1.4, and 0.9 points respectively when using XSum for intermediate task transfer learning.

Our attempt at intermediate task transfer learning did not yield improved results and was not used in our final model, however it did provide valuable information in regard to the question of where intermediate task transfer learning can be applied. In further work, it may be beneficial to further optimize the hyperparameters, such as increasing the number of training epochs on the intermediate task or using larger training splits, before completely ruling out the potential uses of intermediate task transfer learning on the task of dialogue summarization.

3.4 Directed and Reported Speech

The dialogues used for this task and the news articles that the BART model was originally fine-tuned with contain quite different discursive and linguistic structures. The dialogues contain direct speech, using mainly the first and second person verbs conjugations, whereas the news articles have a more narrative style, with a higher use of the third person. We experimented with transforming the structure of our dialogues into reported speech without altering their content to make it more similar to the structure of the news, with the hope that fine-tuning BART with more similar data to what it had been originally fine-tuned with would yield better results.

After fine-tuning BART with these dialogues in their reported-speech form, we had lower ROUGE scores than with the original ones, so we discarded this preprocessing step in our final model. This could be due to the poor quality of our rule-based reported speech transformation algorithm, which results in an excessive use of the verb "says" and some problems in the pronouns reference resolution, but this direct-to-reported-speech task could indeed be interesting to further explore.

3.5 Data Augmentation

Finally, we attempted augmenting our training data by adding a supplementary dataset with similar data to that found in DIALOGSUM. We used the SAMSum dataset (Gliwa et al., 2019), presented as a human-annotated dialogue dataset for abstractive summarization. This dataset presents 16k messenger-like conversations written by linguists fluent in English, together with their summaries.

After merging both datasets, we fine-tuned BART with them, however, we once again achieved results inferior to training on the original dataset alone. This could be due to the shorter length of the SAMSum dialogues and summaries compared to those in DIALOGSUM. It could also be attributed to the different linguistic features between the datasets; the SAMSum dialogues are in a written format, whereas the DIALOGSUM dialogues emulate spoken conversations.

4 Results

Many of the generated summaries produced were close matches to the target summaries. Sometimes generated summaries seemed as though they were a good summarization of the dialogue, but nonetheless had low ROUGE scores. In some cases, this was due to length discrepancies. In other cases, our model generated novel word choices which varied from the gold standard. Examples of a high scoring

TARGET	*#Person1# tells Kate that Masha and Hero get divorced. Kate is surprised because she thought they are perfect couple.*
GENERATED	*#Person1# tells Kate Masha and Hero are getting divorced. Kate is surprised because she thought they are the perfect couple.*
TARGET	*#Person1# and Mike are discussing what kind of emotion should be expressed by Mike in this play. They have different understandings.*
GENERATED	*#Person1# thinks Mike is acting hurt and sad because that's not how his character would act in this situation, but #Person2# thinks Jason and Laura had been together for 3 years so his reaction would be one of both anger and sadness.*

Table 1: Examples of a generated summary close to the target summary (above) and a less ideal generated summary (below)

and low scoring summary can be found in Table 1.

The results were evaluated on ROUGE-1, ROUGE-2, ROUGE-L and BERTSCORE. ROUGE scores measure the *n*-grams shared between the generated and target summaries. ROUGE-L measures the longest shared n-gram. BERTSCORE looks at contextual embeddings instead of exact matches to give a similarity score (Zhang et al., 2019). Our model performed comparable to current leaderboard results on the public test set, and also shows what seem to be respectable results on the hidden test set. Our scores can be found in Table 2.

Our attempts utilizing intermediate task transfer learning, reported speech, and additional datasets all proved unsuccessful. We hypothesize this is a result of insufficient hyperparameter tuning or training. When more complexity is introduced in a model, it often requires specific hyperparameter tuning to result in success, and we suspect this may be one reason our attempts failed.

5 Conclusion

In this paper, we have described our attempt at the INLG 2022 DialogSum Challenge shared task, aimed at generating summaries of real-life scenario dialogues. We utilized a fine-tuned BART model trained on the DIALOGSUM dataset provided to us to achieve our best results.

We explored utilizing intermediate task transfer learning to improve our model, however we speculate that this failed due to a domain mismatch in the datasets, or perhaps due to insufficient hyperparameter tuning and training. Future work could explore intermediate task transfer learning with an intermediate dataset that is better suited for dialogue summarization. Our attempts at altering our data from direct to reported speech, to reflect the dataset that our BART model was fine-tuned with did not work in our favor. We assume this was due to the quality of the reported speech transformation algorithm. Utilizing an additional dataset to increase our number of training samples also did not give desired results. This could be due to differences in the datasets, such as domain, length of texts and summaries, or other factors.

Our results show that it is possible to achieve relatively successful dialogue summarization results using only a basic BART model and fine-tuning on this dataset. In the future, we would further explore the methods we described above.

	R1	**R2**	**RL**	**BERTSCORE**
Public	47.29	21.65	45.92	92.26
Hidden	49.75	25.15	46.50	91.76

Table 2: Scores achieved using the model described in this paper, on both the public and hidden test sets

References

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

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.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text sum-

marization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

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*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Takahashi, and Genichiro Kikui. 2010. Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues. In *Coling 2010: Posters*, pages 400–408.

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

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*.

HP Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, pages 159–165.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. pages 593—-596.

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. *arXiv preprint arXiv:1808.08745*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lin Yuan and Zhou Yu. 2019. Abstractive dialog summarization with semantic scaffolds. *arXiv preprint arXiv:1910.00825*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.