Coreference-Aware Abstractive Dialogue Summarization

Arda Yüksel
Technical University of Munich
Munich, Germany
arda.yueksel@tum.de

Georgios Sigas
Technical University of Munich
Munich, Germany
georgios.sigas@tum.de

Abdullah Orhun Aksoy
Technical University of Munich
Munich, Germany
orhun.aksoy@tum.de

***Abstract*—There have been various studies on abstractive dialog summarization in the past years. These studies take place on different data sets from different domains and they also try to tackle the problem in different ways. The common feature of these is the recently emerged popular topic in the neural network area: transformers. In this report, we give a brief introduction about the general summarization task, introduce the related corpora, mention some of the other approaches that are interesting and lastly focus on one of these approaches, called coreference-aware abstractive dialog summarization to conduct our experiments. We mostly take advantage of the pre-trained BART models to save computing resources and time, compare the partial training of these pre-trained models as training only the encoder or decoder, include different data sets for the training and further compare the results with different training hyper-parameters such as different learning rates and hidden size. At last, we display our results and also mention the suitability of the common evaluation metrics for the abstractive summarization.**

I. Introduction

Text summarization in Natural Language Processing is the process of summarizing the information in large texts for quicker consumption. It is essential for the summary to be a fluent, continuous and depict the significant.

Text summarization has two types as extractive and abstractive. The difference between extractive and the abstractive summarization is that extractive aims to identify the crucial information, extract them and associate them for a succinct summary. In the other hand, abstractive summarization tries to understand the internal semantic representation of the text so it could generate a summary from it afterwards.

There have been lots of studies on extractive summarization with different approaches [1] [2] [3] [4] through past years and they successfully yield good results on different metrics. In more recent years, the focus of the research community is shifted more towards the abstractive text summarization and we saw lots of different approaches on different types of source texts.

Some of the popular data-sets for these different types of source texts are being Reddit [5] that is the collection of posts from Reddit, CNN/Daily Mail [6] is news stories from CNN and Daily Mail websites, NEWSROOM [7], another name is CORNELL NEWSROOM, is articles from major publications, and SAMSum [8] is chat dialogues. These data-sets will be referred as corpus in the rest of the paper.

Through recent years, after the famous transformers made its debut with [9], transformer based approaches for abstractive summarization gained lots of traction. Some of the best performing transformer based architectures are BART [10], PEGASUS [11], GLM-XXLarge [12] and ERNIE-GEN [13].

Most of those approaches focused on the document summarization where the language is formal and the text is well structured for the training. Abstractive dialogue summarization which is a sub-topic for the abstractive summarization has been ignored for some time. The source texts, dialogues, for this problem has multiple parties interacting through the conversation opposed to the document summarization.

[14] categorizes the difficulties of dialog summarization as multiple speakers, role shifting and ubiquitous referring: mentioning of a third party that is not participating the conversation. [15] further mentions about an another challenge that dialog summarization has, which is usage of pronouns that might be anaphora or cataphora.

Our work for this report goes over and briefly explains different approaches particularly for the dialogue summarization and demonstrates the experiment results that we have conducted.

This paper is structured in four other sections. In section II, we will cover the important details regarding historical development and modern methodologies for Abstractive Dialogue Summarization. Then in section III, our methodology will be explained. Our results and their importance will be discussed in Section IV. Finally, we will summarize our work and notable findings in Section V.

II. Previous Works

A. Corpora

We mentioned some of the data sets before that are being used in the training for the abstractive summarization tasks. In this section, we'll go through the corpora, particularly for the abstractive dialog summarization.

The important aspect of this corpora is that the text is simply a conversation where two or more speaker is participating. ICSI [16] is the collection of audio records, transcripts and the information about the participants of the meetings that took place in the International Computer Science Institute from 2000 to 2003. It provides a corpus for from speech recognition

TABLE I
CORPORA FOR ABSTRACTIVE DIALOG SUMMARIZATION

Name	Domain	Year
ICSI [16]	Meeting	2003
AMI [17]	Meeting	2005
SAMSum [8]	Chat	2019
TWEETSUMM [18]	Customer Service	2021
EMAILSUM [19]	Email	2021
MEDIASUM [20]	Interview	2021
DIALOGSUM [21]	Spoken	2021
XSUM* [22]	Articles	2018

to dialog modeling. AMI [17], stands for Augmented Multi-party Interaction, is the data set that consists of hundred hour of meetings. Together with real meetings, this corpus also has scenario driven meetings that is designed specially to extract natural behaviours in conversations.

TWEETSUMM [18] includes real-world customer care dialogues together with extractive and abstractive summaries. EMAILSUM [19] is a corpus that is created for the email thread summarization task and it had over 2500 email threads that are categorized as short and long in different topics.

MEDIASUM [20] is a fairly large corpus that includes more than 463.000 transcripts of media interviews with abstractive summaries. The source of these transcripts are NPR and CNN and they contain complex multi party dialogs from different domains.

DIALOGSUM [21] is another large scale data set that has approximately 13.000 face to face conversations under different topics such as school, medication, work and travel. These conversations happens in between friends, colleagues and customers.

SAMSum [8] is the corpus that contains natural conversations created by linguists that happens in digital environments such as messenger applications. There are different types of conversations from formal to informal which include slang phrases, emojis and typing mistakes. It has over 16.000 dialogues

It's not a corpus that is for abstractive dialogue summarization but we include the definition of the XSUM [22] data set as well since it's mentioned in our experiments and chosen method at last. It consist of more than 226.000 news articles from BBC accompanied with the abstractive one-sentence summaries which answers the question of "What the article is about?".

B. Transformer

Word embedding models developed to model languages to increase performances in many tasks in the fields of Natural Language Understanding (NLU) and Natural Language Generation (NLG). Before the current advances in transformer architectures, classical statistical approaches, [23], and generic RNN/CNN based models, [24], are used in the task of abstractive summarization interchangeably.

Currently most of the NLP tasks utilize attention mechanism, [9]. The transformer models allow models to better generalize key concepts within the texts as they can better attend past information. One of the earlier works developed to be BERT, [25], (Bidirectional Encoder Representations from Transformers) which revolutionized semantic modelling. After the introduction of BERT, many researchers altered transformer architectures to better tackle the issues of NLG tasks such as machine translation and text summarization. For these purposes, a sequence-to-sequence autoencoder model that encapsulates transformer encoder-decoder structure is developed. This denoising autoencoder, BART [10], is now fully utilized in the domains of abstractive summarization due to its significant performance increase. Aside from BART, many focused on improving capabilities of the models via transfer learning approach to obtain a versatile model that can achieve high performances in many tasks at once. This aim led to development of T5, [26] and many more.

C. Modern Approaches

Our work through the semester while looking for different approaches on the abstractive dialog summarization, we encountered with some interesting methods and we would like to mention them in here.

[27] demonstrates a different approach for abstractive dialog summarization where the two challenges, varying topic through the conversation and scattered key information to different speakers are focused on. Two contrastive learning objectives, one for each, are introduced as auxiliary tasks for the primary dialog summarization task to tackle these problems. First of these tasks, named coherence detection, captures the change in the topic and the second one, named sub-summary generation, handles the dispersed information in sentences one by one.

Dialogue state tracking is about defining a complete view of what the user wants at that particular point in the dialog, including the target constraints, the set of requested slots, and the user's dialog behavior whenever the dialog changes. In [28], this method is used to summarize dialogues where they assume dialogue summaries are simply unstructured dialogue states. First, they train a language model with a set of dialogues based on a synthetic model generated by a set of dialog state rules. Then, it's possible restore the dialog states by reversing the summary generation rules.

III. METHODOLOGY

A. Transformer Model

As we noticed in the previous section most of the papers work with transformers. In order to find our base model, the one that we will start build on, we compare many variations of BART, as it was the most popular among the aforementioned papers, but in addition we tested a t5 model as well. As you will see in the result tables BART had the best results.

BART Bidirectional and Auto-Regressive Transformer (BART) is a sequence-to-sequence model with a bidirectional encoder similar to BERT's and a left to right autoregressive decoder similar to GPT's.

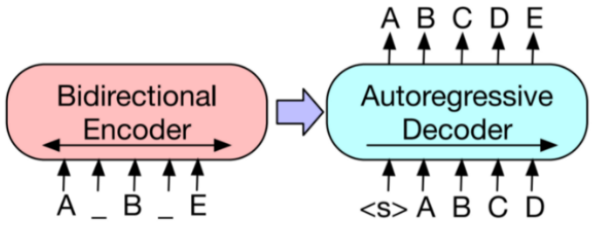

Fig. 1. BART Model Encoder - Decoder Structure

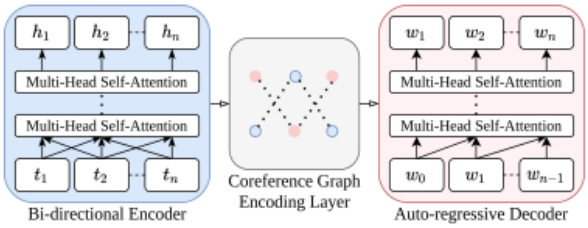

Fig. 2. Coreference Aware Transformer Encoder - Decoder Structure

B. Transfer Learning

As BART is a very large model with millions of parameters, its training would be difficult and time consuming. Therefore, our first thought was that we must use all the knowledge that a pretrained BART model has and manipulate it in order to be used in our task. Transfer learning is utilized in complex models in literature which tackles the problem of adapting pretrained structures to perform in similar but not entirely the same tasks.

At first we search and choose the most accurate BART models trained on dialog summarization or on simple summarization tasks and then we would further train them on the data sets of our task. However, the training could not be on the whole model as we mentioned before and that is why we used the method of transfer learning. This method, allow us to keep all the knowledge that a model has from previous training and use it for our own purpose just by modifying a small part of it. Specifically, we managed to freeze a part of the model, at the training phase and just train a small proportion of parameters,the unfrozen part, focused mostly on the parameters of the encoder and the decoder.

In our transfer learning models, we utilized abstractive summarization models that are not necessarily trained for dialogue summarization tasks. Since the task is branch of the abstractive summarization but the samples are drawn from another distribution, transfer learning can be utilized in later layers of transformer models such as BART or T5.

C. Coreference-Aware Dialogue Summarization

At last we tried to find a way to boost our model even further. What we did was to add a coreference graph between the encoder and decoder parts.The coreference graph is basically an encoding layer used to model the coreference connections between all mentions. So for example all the pronouns would match to a subject and as a result we will not have any abstract words in the input text.

In Figure 2, one can see the updated structure for generic transformer Encoder-Decoder schema. In between BART's encoder and Auto-regressive Decoder, a graph encoding layer is appended. This layer allows model to make more accurate connections between pronouns and understand referrals to prior text in addition to already advanced attention mechanism.

IV. Results

In the result section, our study examined multiple variants for generic BART Large models and our Graph Based Coreference Aware BART Large models. In addition demonstrate the results clearly, we divided them according to the stages of development. In training, one can see the progression of loss with respect to number of training samples for BART Large and T5 models for multiple configurations. In validation, we indicated the results of both Graph Coreference Models and Normal Transformer models. In the test subsection, we showed Rouge Scores for the best performing models obtained after the analysis of the training and validation performances. For the Dialogue subsection, we also provided sample dialogues outside of our corpora in order to show capabilities of our model.

A. Training

In the Figure 3, one can see the progression of loss for multiple BART Large models and T5 model. In these configurations, the effect of pretrained corpora selection, model selection and frozen layers are tested. Unless it is stated otherwise, the encoder layer is frozen. For the linear models, both the encoder and decoder are frozen.

As it can be seen from Figure 3, the BART Large models with frozen encoder setups perform better for training process. This conclusion can also be derived for the longer epochs. Figure 4 demonstrates the comparison of T5 and BART models more drastically.

B. Validation

For the validation setup, graph and normal transformer models are analyzed separately. Their Rouge Scores are given in Tables III and II respectively. For the generic transformer architectures, we analyzed the effect of choice of transformer, pretrained corpus, training data set, frozen layers. For the graph based coreference aware models, we tested for the multiple configurations according to work of [15]. In addition to that, the analysis of different model properties such as but not limited to learning rate and hidden size are completed.

Table II shows the importance of the model selection. BART Large performs better compared to T5. Thus, for the

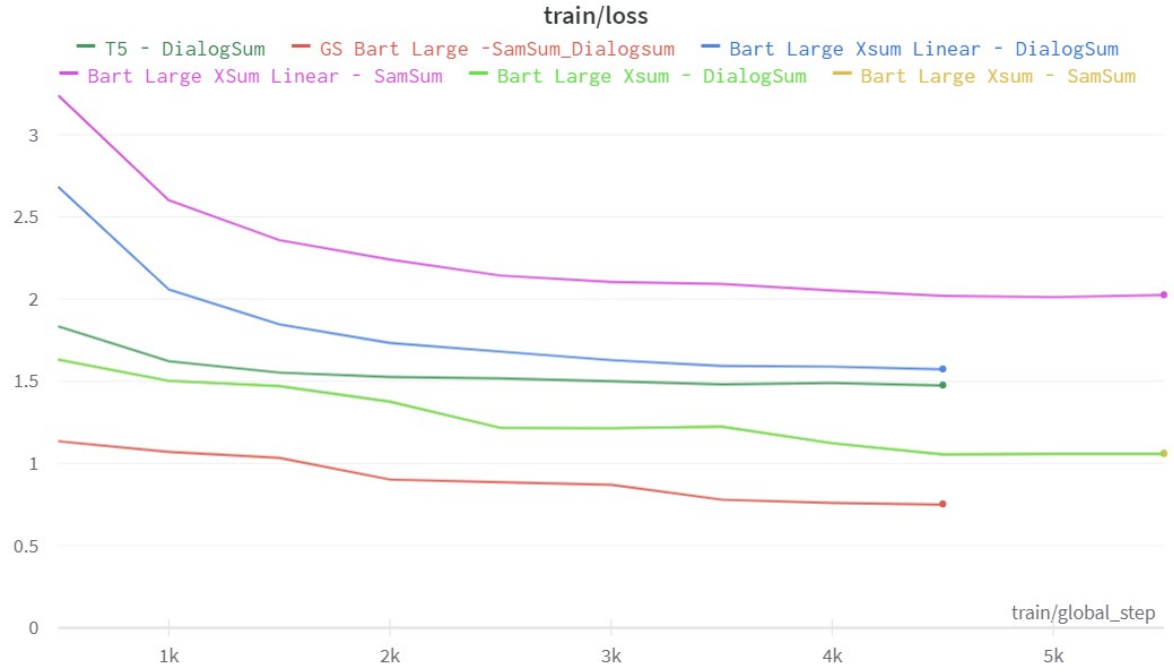

Fig. 3. Training Loss for Small Epoch

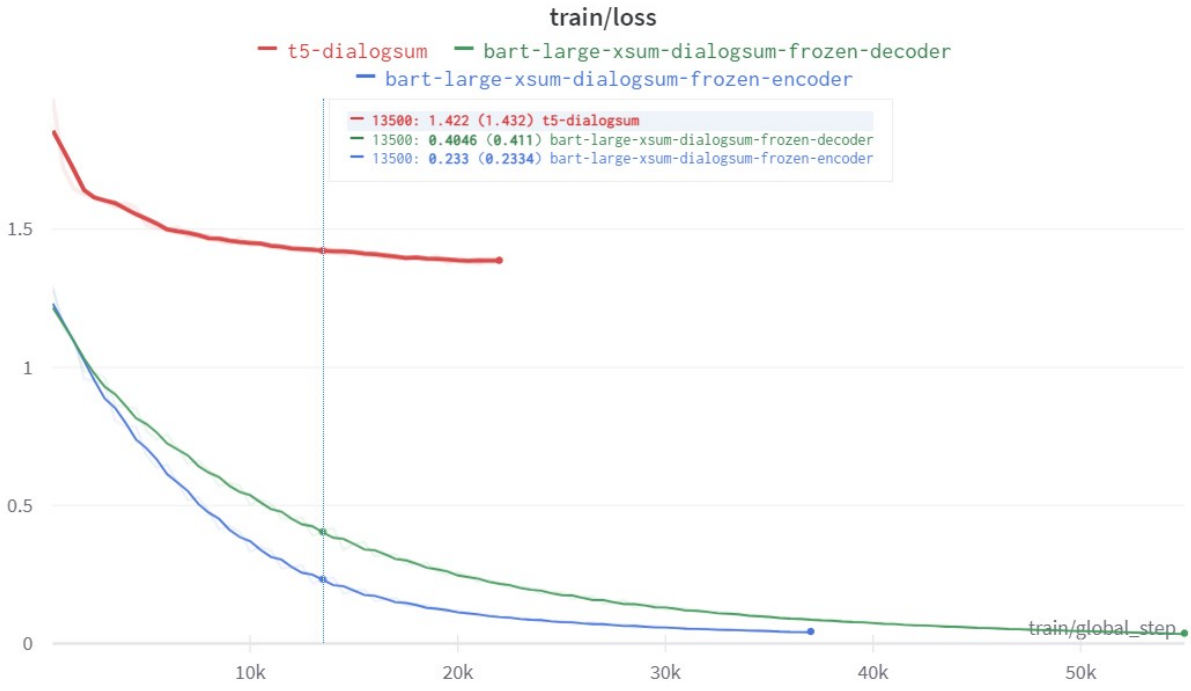

Fig. 4. Training Loss for Longer Epoch

coreference oriented development, BART Large is selected as the basis. In addition to that, freezing decoder layer resulted in significant performance drop. After training the same BART Large model pretrained on XSUM data set, we concluded that SAMSUM models generate better Rouge performance. Considering all these remarks, we investigated graph based architectures.

According to Table III, faster convergence is attained with

TABLE II
COMMON TRANSFORMER MODELS VALIDATION SCORES FOR DIFFERENT CORPUS AND SETUPS. ALL BART MODELS ARE SELECTED AS BART LARGE. CORPUS NAMES ARE SAM FOR SAMSUM AND DIALOG FOR DIALOGSUM. FOR THE FROZEN LAYER PART, ALL INDICATES FROZEN AND DECODER LAYERS. ROUGE SCORES ARE REFERRED AS R-1, R-2 AND R-L.

Model	**Corpus**	**Frozen**	R - 1	R - 2	R - L
T5	Dialog	Encoder	34.99	12.18	29.14
BART SAM	Dialog	Encoder	48.43	24.06	40.23
BART XSUM	Dialog	All	37.58	14.86	31.76
BART XSUM	SAM	All	38.92	15.49	31.6
BART XSUM	Dialog	Encoder	47.78	23.25	39.65
BART XSUM	SAM	Encoder	**52.25**	**26.44**	**41.91**

TABLE III
BART LARGE XSUM GRAPH MODEL VALIDATION SCORES IN SAMSUM DATASET

Model	ROUGE 1	ROUGE 2	ROUGE L
Different Learning Rate	**50.90**	23.83	46.19
Coreference Head	50.79	23.34	45.47
Replace Coref Head	50.14	22.87	45.02
Graph Only Training	50.31	24.20	46.69
Larger Hidden Size	49.30	23.66	45.86
Frozen Encoder	50.31	**24.66**	**46.73**

different learning rate which is expected due to size of SAMSum data set. Different configurations of [15] did not provide significant increase. For the validation, Frozen Encoder and Frozen Encoder-Decoder converged to similar results and as it can be seen from the Table, they are overall the best performing candidates. Hence, for the testing they are selected.

C. Test

Graph based models compared in Table IV for SAMSum data set. As it can be seen from the table, Frozen Encoder configuration is the better candidate. Though they have similar validation scores, the test Rouge Scores clearly indicates the difference.

TABLE IV
GRAPH BASED MODELS SAMSUM TEST ROUGE SCORES

Model	ROUGE 1	ROUGE 2	ROUGE L
Frozen Encoder	49.58	23.51	46.05
Graph Only Training	48.16	22.68	44.85

For the generic BART Large based architectures, we also showed Rouge Scores for SAMSum and DialogSum in Tables V and VI. BART Large model pretrained on SAMSum with frozen encoder is trained on DialogSum once more. The resulting model, according to our overall results, is the most consistent BART Large configuration.

TABLE V
BART MODELS TEST ROUGE SCORES FOR SAMSUM

Pretrained	**Trained**	R - 1	R - 2	R - LSUM
XSum	SAMSum	**51.01**	**25.41**	**46.52**
XSum	DialogSum	45.95	19.26	40.38
SAMSum	DialogSum	47.41	20.88	42.31

TABLE VI
BART MODELS TEST ROUGE SCORES FOR DIALOGSUM

Pretrained	**Trained**	R - 1	R - 2	R - LSUM
XSum	SAMSum	35.35	13.71	31.57
XSum	DialogSum	44.52	19.01	39.12
SAMSum	DialogSum	**44.98**	**19.62**	**39.92**

D. Dialogue

TABLE VII
SAMPLE DIALOGUE FROM *Harry Potter Chamber of Secrets* MOVIE AND ITS SUMMARY GENERATED BY OUR BART LARGE BASED IMPLEMENTATION

Dialogue	**Harry**: I couldn't help noticing certain things. Certain similarities. Between Tom Riddle and me. **Dumbledore**: Unless I'm much mistaken, he transformed some of his own powers to you the night he gave you that scar. **Harry**: Voldemort put a bit of himself in me? **Dumbledore**: Not intentionally, but... yes. **Harry**: So the Sorting Hat was right. I should be a Slytherin. **Dumbledore**: It's true, Harry. You do possess many of the qualities Voldemort himself prizes. Yet the Sorting Hat placed you in Gryffindor. **Harry**: Only because I asked it to. **Dumbledore**: Exactly. Which makes you very different from Voldemort. It's not our abilities that show what we truly are, Harry. IT'S OUR CHOICES.
Summary	Dumbledore tells Harry that the Sorting Hat placed him in Gryffindor because he asked it to. Harry thinks he should be a Slytherin but Dumbledore says he is different from Voldemort.

To demonstrate the validity of our models, we provided a sample dialog from a scene in a famous movie. This sample is not present in both of our training corpora. Our model obtained key points of the dialogue and added its own abstractive summary. Our result demonstrates that the model can summarize the dialogue clearly and correctly. In terms of abstractive summarization, or more specifically abstractive dialogue summarization, it is more important to

demonstrate summarization example according to the authors of this paper. ROUGE scores utilize matching ngrams in a given summary with respect to ground truth, but the aim of the abstractive summarization is for model to possess the capability of paraphrasing the sentences as well as extracting key information.

V. Conclusion

Abstractive summarization techniques advanced with the introduction of transformer based architectures. Availability of different types of corpora allowed researchers study dialogues for the abstractive approaches. In this study, the variants of transformer models are discussed for Abstractive Dialogue Summarization. Our work tests the validity of existing state of the art models such as but not limited BART.

Coreference Aware summarization schema utilized graph structures with existing BART models to increase natural language understanding capabilities. In this research, we analyzed the multiple configurations for both plane BART models and coreference aware models for the most notable corpora for dialogue summarization. Our developed models proved that abstractive dialogue summarization is a field that can demonstrate abilities of existing transformer architectures in Natural Language Generation tasks for data drawn from various fields including mainstream shows.

In our study, we realized importance of the metric for summarization model evaluation. ROUGE scores which are the common metric for summarization model does not indicate the power of abstraction. The aim is not only obtain important information but also allow model to have creativity in summarization process. Ngram based approaches validate key feature matching from the passage but for the abstractive summarization different metrics should be discussed. For the future work, we note that Graph based models can be utilized further with more in depth analysis and computational power to ensure generalization of long dialogues and passages. The usage of coreference aware models will increase abstractive power of the summarization models.

References

[1] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2017, pp. 1–6.

[2] A. Sinha, A. Yadav, and A. Gahlot, "Extractive text summarization using neural networks," *arXiv preprint arXiv:1802.10137*, 2018.

[3] J. Xu, Z. Gan, Y. Cheng, and J. Liu, "Discourse-aware neural extractive text summarization," *arXiv preprint arXiv:1910.14142*, 2019.

[4] C. Fang, D. Mu, Z. Deng, and Z. Wu, "Word-sentence co-ranking for automatic extractive text summarization," *Expert Systems with Applications*, vol. 72, pp. 189–195, 2017.

[5] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[6] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[7] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," *arXiv preprint arXiv:1804.11283*, 2018.

[8] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization," *arXiv preprint arXiv:1911.12237*, 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[11] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[12] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335.

[13] D. Xiao, H. Zhang, Y. Li, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-gen: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation," *arXiv preprint arXiv:2001.11314*, 2020.

[14] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition."

[15] Z. Liu, K. Shi, and N. F. Chen, "Coreference-aware dialogue summarization," *arXiv preprint arXiv:2106.08556*, 2021.

[16] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[17] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th international conference on methods and techniques in behavioral research*, vol. 88. Citeseer, 2005, p. 100.

[18] G. Feigenblat, C. Gunasekara, B. Sznajder, S. Joshi, D. Konopnicki, and R. Aharonov, "Tweetsumm–a dialog summarization dataset for customer service," *arXiv preprint arXiv:2111.11894*, 2021.

[19] S. Zhang, A. Celikyilmaz, J. Gao, and M. Bansal, "Emailsum: Abstractive email thread summarization," *arXiv preprint arXiv:2107.14691*, 2021.

[20] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "Mediasum: A large-scale media interview dataset for dialogue summarization," *arXiv preprint arXiv:2103.06410*, 2021.

[21] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "Dialogsum: A real-life scenario dialogue summarization dataset," *arXiv preprint arXiv:2105.06762*, 2021.

[22] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *arXiv preprint arXiv:1808.08745*, 2018.

[23] M. Banko, V. O. Mittal, and M. J. Witbrock, "Headline generation based on statistical translation," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ser. ACL '00. USA: Association for Computational Linguistics, 2000, p. 318–325. [Online]. Available: https://doi.org/10.3115/1075218.1075259

[24] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 93–98. [Online]. Available: https://aclanthology.org/N16-1012

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019. [Online]. Available: https://arxiv.org/abs/1910.10683

[27] J. Liu, Y. Zou, H. Zhang, H. Chen, Z. Ding, C. Yuan, and X. Wang, "Topic-aware contrastive learning for abstractive dialogue summarization," *arXiv preprint arXiv:2109.04994*, 2021.

[28] J. Shin, H. Yu, H. Moon, A. Madotto, and J. Park, "Dialogue summaries as dialogue states (ds2), template-guided summarization for few-shot dialogue state tracking," *arXiv preprint arXiv:2203.01552*, 2022.