Leveraging Non-dialogue Summaries for Dialogue Summarization

Seongmin Park **Dongchan Shin** **Jihwa Lee**
ActionPower
Seoul, Republic of Korea
{seongmin.park, dongchan.shin, jihwa.lee}@actionpower.kr

Abstract

To mitigate the lack of diverse dialogue summarization datasets in academia, we present methods to utilize non-dialogue summarization data for enhancing dialogue summarization systems. We apply transformations to document summarization data pairs to create training data that better befit dialogue summarization. The suggested transformations also retain desirable properties of non-dialogue datasets, such as improved faithfulness to the source text. We conduct extensive experiments across both English and Korean to verify our approach. Although absolute gains in ROUGE naturally plateau as more dialogue summarization samples are introduced, utilizing non-dialogue data for training significantly improves summarization performance in zero- and few-shot settings and enhances faithfulness across all training regimes.

1 Introduction

Dialogue summarization fundamentally differs from its non-dialogue counterparts in two ways: the presence of speaker information and the inherent abstractiveness that demands any dialogue summarization system to "read between the lines". Consequently, training a dialogue summarization model requires datasets appropriate for the dialogue domain, which often calls for different provisions than those commonly found in traditional, non-dialogue summarization datasets.

The bulk of research efforts in summarization, however, has historically been focused on written documents. As a result, the research community faces a shortage of diverse dialogue summarization data, in contrast to the abundance of non-dialogue summarization data (Feng et al., 2021; Tuggener et al., 2021). From such state of the literature, we identify a strong need for methods to utilize widely available non-dialogue summarization data in reinforcing dialogue summarization models.

In this work, we present recipes to transform non-dialogue data into formats that enable direct integration into dialogue summarization training. During the transformation process, we also inherit desirable properties that arise from the extractiveness of non-dialogue summarization datasets. Factual inconsistency and hallucination are major research problems in dialogue summarization (Maynez et al., 2020; Ladhak et al., 2021; Cao et al., 2018; Huang et al., 2021). Since extractive summaries naturally remain more faithful to the source text, we design our transformation schemes to retain such properties when adapting non-dialogue summary data to the dialogue domain.

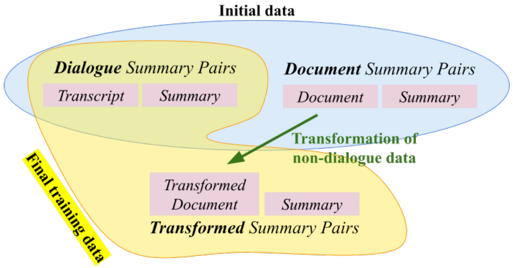
Initial data
Dialogue *Summary Pairs*
Transcript
Summary
Document *Summary Pairs*
Document
Summary
Transformation of non-dialogue data
Final training data
Transformed Document
Summary
Transformed *Summary Pairs*

Figure 1: Overview of our proposed method. We transform non-dialogue data into a format exploitable by dialogue summarization models.

Our contributions are as follows:

1. We present formulas to transform non-dialogue summarization datasets into patterns usable for dialogue summarization. Summarization models trained with the additional data produce summaries more similar to gold reference summaries.

2. We show that utilizing non-dialogue summarization data preserves faithfulness in otherwise factually-unchecked summaries.

3. We test our data manipulation scheme across two languages (English and Korean) and on document summary datasets with different levels of abstraction.

Proceedings of the 29th International Conference on Computational Linguistic, pages 1–7
October 12–17, 2022.

In Section 2, we first describe existing challenges in dialogue summarization. In Section 3, we describe our dataset adaptation methods in detail. In Section 4, we describe datasets, evaluation metrics, and experiments used to test our methods.

2 Related works

2.1 Non-dialogue data for dialogue summarization

Even in high-resource languages like English, diversely annotated dialogue summarization datasets are scarce (Feng et al., 2021; Tuggener et al., 2021; Zou et al., 2021). The need for a diverse collection of dialogue summarization datasets is further exacerbated by the fact that dialogue is recorded in many formats, such as meetings, chats, and spontaneous speech.

To appease such a need for more data, several attempts have been made to utilize non-dialogue data in dialogue summarization (Figure 1). (Zou et al., 2021) pre-trains a language model with BookCorpus (Zhu et al., 2015) to provide training samples across diverse domains. (Khalifa et al., 2021) pre-trains BART (Lewis et al., 2020) with unlabeled dialogue corpora and fine-tunes the language model with downstream summary tasks.

The focus of such approaches lies in whetting a model to be more responsive to limited dialogue summarization data. We suggest a new line of research that directly manipulates the training data instead of steering a model's disposition directly.

2.2 Faithfulness in dialogue summarization

Factual incorrectness is a problem commonly observed in abstractive summarization systems (Cao et al., 2018; Huang et al., 2021; Tang et al., 2021). Tang et al. (2021) identifies categories of factual errors that dialogue summarization models may generate. To improve the factual consistencies of generated summaries, the authors corrupt dialogue transcripts to create negative samples in a contrastive-learning scheme.

We employ a similar noising approach. Negative sample generation in (Tang et al., 2021) requires accurate token-level operations, such as part-of-speech extraction and word negation. Our manipulation scheme forgoes such additional components, relying only on deterministic sentence-level edits.

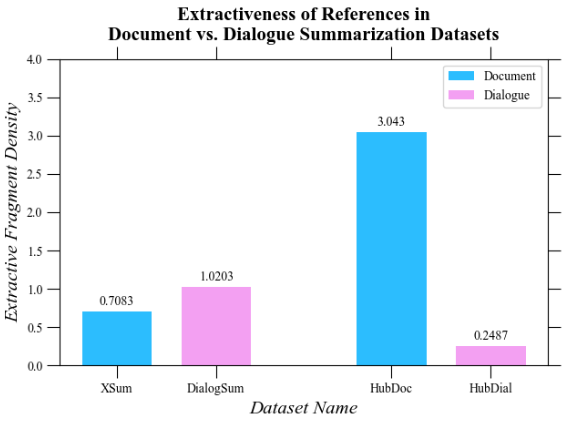

Figure 2: Extractiveness of reference summaries. Extractive Fragment Density (Grusky et al., 2018) is the longest extractive token span from the source data that matches the reference summary.

3 Proposed method

3.1 Preliminaries

Let

$$DocSet = \{(A_0, X_0), (A_1, X_1), ..., (A_i, X_i)\} \tag{1}$$

be a non-dialogue (document) summarization dataset, where A_i is the i-th document in the set, and X_i is the corresponding reference summary. A_i is a sequence of sentences $(a_0, a_1, ..., a_m)$, where m is the sentence count.

Similarly, we define a dialogue summarization dataset,

$$DialSet = \{(B_0, Y_0), (B_1, Y_1), ..., (B_j, Y_j)\}, \tag{2}$$

where B_j is the j-th dialogue transcript in the dataset, and Y_j is the corresponding dialogue summary. Like any A, B_j consists of ordered sentences $(b_0, b_1, ..., b_n)$.

We define $F = \{f_0, ..., f_k\}$, a set of *transformation functions* to be applied to each A_i in $DocSet$. A transformation function is a set of operations to transform non-dialogue text data into a pattern usable in dialogue summarization training.

We introduce three such transformation functions: **forcing plain text into dialogue format** (e.g. by inserting pseudo-speaker information), **shuffling sentence order**, and **omitting the sentence with highest extractive overlap** with the reference summary. Each suggested transformation function is formerly defined in succeeding sections.

Once f_k is applied to each A_i in $DocSet$, each transformed non-dialogue input text is paired with

Table 1: Evaluation metrics for full training. f_d (D) transformation is consistently effective in boosting match-based ROUGE. Even though marginal gain in ROUGE from our method naturally decreases as the size of $DialSet$ increases, incorporating document summarization data greatly improves summary faithfulness. R1, R2, RL, Prec., Rec., Faith. respectively stands for ROUGE-1, ROUGE-2, ROUGE-L, Precision, Recall, and Faithfulness. Underlined values are the highest in each column. Higher is better for all metrics.

Data	DialogSum						HubDial					
	R1	*R2*	*RL*	*Prec.*	*Rec.*	*Faith.*	*R1*	*R2*	*RL*	*Prec.*	*Rec.*	*Faith.*
Original	39.57	15.43	32.97	-3.8962	-4.3175	-4.1101	35.42	16.90	31.11	-9.9774	-9.7934	-8.9965
Naive	39.36	14.89	32.56	-2.9108	-2.9933	-2.5351	35.97	17.68	31.13	-7.8063	-7.6203	-7.9189
D	**40.47**	**16.41**	**33.89**	**-2.9085**	-2.8702	-2.4615	**36.32**	17.61	**31.79**	**-7.7715**	-7.6190	-7.9202
S	39.94	15.70	33.31	-2.9310	-2.8966	-2.4261	36.08	**18.13**	31.41	-7.8159	**-7.5791**	**-7.8610**
O	39.80	15.97	33.32	-2.9211	**-2.8253**	**-2.4072**	35.96	17.51	31.32	-7.7975	-7.6251	-7.8954
D + S	39.87	15.73	33.44	-2.9235	-2.9224	-2.4970	36.03	**17.55**	**31.93**	-7.8179	-7.6066	**-7.9104**
D + O	**40.66**	**16.77**	**34.15**	**-2.9044**	**-2.8073**	**-2.4196**	36.11	17.52	31.92	**-7.7776**	-7.6245	-7.9395
S + O	40.34	16.33	33.82	-2.9077	-2.8797	-2.4376	35.52	17.21	31.32	-7.8456	-7.6149	-7.9244
D + S + O	39.97	16.00	33.56	-2.9402	-2.8678	-2.4278	**36.26**	17.29	31.29	-7.8105	**-7.5956**	-7.9371

its corresponding reference X_i to form a new training set:

$$NewDocSet = \{(f_k(A), X) \mid (A, X) \in DocSet\}. \quad (3)$$

$NewDocSet$ can be used as additional training data for dialogue summarization models.

3.2 Arranging text into dialogue format (f_d)

Given a plain document, we convert its contents into transcript format by segmenting the document into sentences and appending a psuedo speaker:

$$f_d(A) = (concatenate(\text{“}\mathit{Speaker\ 1:}\text{”},\ a))_{a \in A}. \quad (4)$$

This operation serves two purposes: we prime our model to be more receptive of dialogue-formatted data through prompting (Liu et al., 2021). We also remove the gap between data patterns in training and inference by standardizing diverse non-dialogue document data into the dialogue domain.

The prompt *"Speaker 1"* was chosen empirically: multiple configurations, such as varying speaker numbers and inserting real names, were tested. Such complex configurations led to only marginal increases in evaluation metrics and introduced additional roadblocks in reliable reproduction (upper bound in speaker number has to be arbitrarily selected; a dictionary with realistic names has to be distributed). Both English and Korean datasets used *"Speaker 1"*.

3.3 Shuffling sentence order (f_d)

To combat lead bias commonly observed in traditional summarization datasets (Grenander et al., 2019; Zhu et al., 2021), we shuffle the order of sentences in A:

$$f_s(A) = shuffle(A). \quad (5)$$

Previous research has shown sentence shuffling helps in reducing read bias (Grenander et al., 2019). Since information in dialogues is often dispersed across multiple utterances, we find sentence shuffling to be more impactful when dealing with dialogues, compared to documents.

3.4 Omitting the most extractive sentence (f_o)

Among all sentences in a document, we delete the sentence with the most extractive overlap with the reference summary. The degree of overlap is calculated by the number of shared character 3-grams between a single sentence from the source document and the whole reference.

$$f_o(A_i) = A_i\{a_{ex}\}, \quad (6)$$

where a_{ex} in A_i has the highest 3-gram overlap with X_i. By removing the most extractive sentence, we aim to make $DocSet$ more abstractive and reduce copying behavior.

Table 2: Datasets used in the experiment. “Dial." and “Doc." stand for “dialogue" and “document".

Name	**Lang.**	**Type**	**Size**	**Abstractive?**
DialogSum	English	Dial.	15,600	Yes
XSum	English	Doc.	204,045	Yes
HubDial	Korean	Dial.	16,000	Yes
HubDoc	Korean	Doc.	334,160	No

4 Experiments

4.1 Experiment setup

We conduct comprehensive experiments that apply transformation functions defined in Section 3.

4.1.1 Our models

First, we create different variants of $NewDocSet$ by applying functions in $F = \{f_d, f_s, f_o\}$ both individually and in combination. Such application results in 7 different variations of $NewDocSet$: D, O, S, $D+O$, $D+S$, $S+O$, $D+S+O$, where, for example, $D + O = \{f_d \circ f_o(A) \mid A \in DocSet\}$.

With newly acquired training data, we train a BART-base (Lewis et al., 2020; Wolf et al., 2019) summarizer under three different configurations:

1. *Zero-shot*: $NewDocSet$ is the training set.
2. *Few-shot*: Training data consists of $NewDocSet$ and 100 or 1000 samples from $DialSet$.
3. *Full training*: Training data consists of $NewDocSet + DialSet$.

We choose the BART architecture due to its widespread use and proven track record in summarization (Fabbri et al., 2021; Akiyama et al., 2021; Zhao et al., 2021).

4.1.2 Baselines

We compare our trained models with two baselines:

1. *Original*: $DialSet$ is the training set.
2. *Naive*: Training data consists of $DialSet$ and $DocSet$ (i.e. $f_{naive}(A) = A$).

4.2 Datasets

For English, we use DialogSum (Chen et al., 2021) as $DialSet$ and XSum (Narayan et al., 2018) as $DocSet$. For Korean, we use AIHub Dialogue Summarization Dataset[1] (HubDial) as $DialSet$ and AIHub Document Summarization Dataset[2] (HubDoc) as $DocSet$. Table 2 contains a brief description of each dataset.

Transformations f_s and f_o hinge on the assumption that non-dialogue summarization datasets typically display considerable lead bias and are more extractive than dialogue summarization datasets. To gauge how extractiveness of non-dialogue data effects final summary generation performance, we conduct experiments on both highly extractive (HubDoc) and extremely abstractive (XSum) document summarization datasets (Figure 2).

4.3 Evaluation metrics

Performance of our model is measured as the similarity between model summaries and reference summaries, calculated with standard ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L) (Lin, 2004). We also measure the faithfulness of the output summaries to input dialogues with BartScore (Yuan et al., 2021). BartScore is a state-of-the-art evaluation metric for factual consistency and faithfulness in text generation.

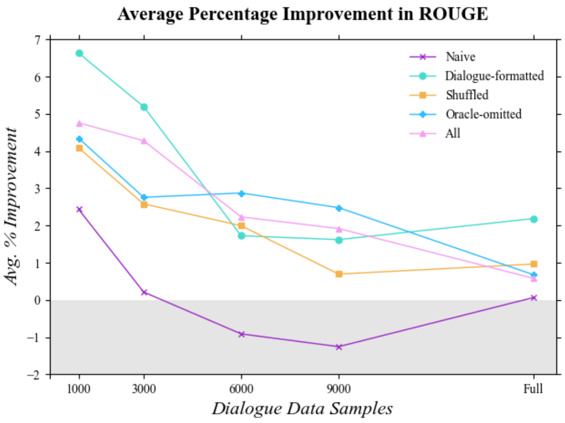

Figure 3: Averaged percentage ROUGE-1, ROUGE-2, and ROUGE-L improvements over dialogue-only training on HubDial test set. Shaded regions indicate configurations that underperform dialogue-only training.

5 Results

5.1 Full training

Both English and Korean summarization models benefit from additional data curated by our transformation functions. Only naive application of non-dialogue data fails to improve ROUGE scores compared to dialogue-only training. While marginal increase in ROUGE saturates as more dialogue summaization training samples are added, the addition of document data significantly enhances factual consistency of summaries (Table 1).

5.1.1 Abstractive document summary dataset

In terms of ROUGE, models trained with abstractive document summarization data (XSum) are most affected by f_d (D) transformations. Highest scoring data transformation combinations mostly

[1] https://aihub.or.kr/aidata/30714

[2] https://aihub.or.kr/aidata/8054

Table 3: Few-shot results on English DialogSum. Since XSum is already highly abstractive, f_d (D) transformation is the most effective. Almost all maximum values in each category involve a f_d transformation. Notations are the same as in Table 1.

	Zero-shot				100-shot				1000-shot			
	R1	*R2*	*RL*	*Faith.*	*R1*	*R2*	*RL*	*Faith.*	*R1*	*R2*	*RL*	*Faith.*
Original	-	-	-	-	31.05	10.55	26.58	-4.4925	35.12	12.23	29.20	-4.7314
Naive	13.64	2.71	11.21	-2.9081	31.05	9.31	25.81	-4.5829	37.97	13.22	31.11	-2.4799
D	15.46	3.18	13.05	**-2.7045**	**34.93**	**10.74**	**28.21**	-4.4414	38.28	13.23	31.08	-2.4319
S	14.35	**3.51**	12.11	-2.8926	32.83	09.82	26.80	**-2.4675**	**38.33**	13.62	**31.45**	**-2.4146**
O	**16.41**	2.80	**13.66**	-2.9846	32.89	09.53	27.24	-2.8102	38.28	**13.57**	31.12	-2.4495
D + S	**17.47**	**4.27**	**14.56**	**-2.4899**	34.51	**10.96**	27.96	-2.7527	38.26	13.27	31.21	-2.4494
D + O	14.73	2.88	12.07	-2.9530	34.40	10.76	28.18	**-2.6696**	**38.85**	**13.55**	**31.62**	**-2.3949**
S + O	16.69	3.42	13.96	-3.1872	33.84	10.27	27.83	-3.0668	38.55	13.44	31.19	-2.4154
D + S + O	16.36	3.84	13.76	-2.5456	**34.65**	10.82	**28.25**	-2.7152	36.80	13.03	30.42	-2.4381

Table 4: Few-shot results on Korean HubDial. Compared to less extractive English summarization, we see f_s (S) and f_o (O) transformations resulting in greater marginal increase in ROUGE. Notations are the same as in Table 1.

	Zero-shot				100-shot				1000-shot			
	R1	*R2*	*RL*	*Faith.*	*R1*	*R2*	*RL*	*Faith.*	*R1*	*R2*	*RL*	*Faith.*
Original	-	-	-	-	3.41	1.35	3.03	-10.2144	31.42	13.64	26.69	-7.9586
Naive	20.72	8.94	18.34	-7.4637	27.98	12.56	24.24	-7.7524	32.17	14.74	27.34	-7.8893
D	**26.34**	**11.74**	**22.97**	-7.6731	28.48	13.03	24.79	**-7.6451**	**33.05**	**15.16**	**28.46**	**-7.7255**
S	21.38	9.43	19.01	-7.3379	28.12	12.42	24.19	-7.9053	32.68	14.91	27.78	-7.8162
O	22.09	9.82	19.73	**-7.1363**	**29.50**	**13.26**	**25.01**	-7.8737	32.26	14.77	27.85	-7.8357
D + S	24.21	11.31	21.48	-7.8747	28.50	13.06	24.84	-7.8565	31.78	14.62	27.18	-7.7687
D + O	**24.81**	11.20	**22.04**	-7.7556	29.71	13.17	**25.68**	**-7.8058**	31.67	14.66	27.33	**-7.7651**
S + O	20.38	9.25	18.50	**-7.4731**	**29.79**	**13.47**	25.62	-7.9142	31.92	14.53	27.16	-7.7673
D + S + O	24.17	**11.37**	21.46	-7.8234	28.50	13.31	24.48	-7.8908	**32.77**	**15.25**	**27.96**	-7.7934

involve f_d. In terms of factual consistency and faithfulness, f_o transformations consistently score the highest. This is in line with our intention to introduce an additional in-comprehension understanding objective to the model that simple dialogue formatting cannot provide.

5.1.2 Extractive document summary dataset

f_s (S) and f_o (O) transformations are more influential when used to transform extractive data (HubDoc). Factual consistency is correlated the most with f_s, because of lead bias present in HubDoc.

5.2 Zero- and few-shot training

In zero- and full-shot training, we see significant improvements in both ROUGE and factual consistency (Tables 3, 4). Figure 3 shows improvements in ROUGE over $DialSet$-only training at different dialogue $DialSet$ sizes. Naively training with non-dialogue summarization data yields results no better than training with only dialogue data. In contrast, our suggested transformations provide significant gains in both span match and consistency measures in low-shot training regimes.

Comparative influence of each transformation function (f_d, f_s, and f_o) show trends similar to those observed in full training, with f_d proving the most dominant for already abstractive $DocSet$ (XSum) and f_s and f_o being more influential in comparatively extractive $DocSet$ (HubDoc).

6 Conclusion

We present simple but immediately effective methods to utilize abundant non-dialogue summarization data to improve dialogue summarization systems. We evaluate performance gains in similarity to reference summaries as well as in factual consistency to original transcript input. We find that our method is especially impactful in low-resource dialogue summarization.

Our research hints at two possible avenues for further investigation: reinforcing the three presented transformation recipes with a more methodical generation of prompts (Ghazvininejad et al., 2021), or introducing new transformations that better capture the unique properties of dialogue summarization datasets.

References

Kazuki Akiyama, Akihiro Tamura, and Takashi Ninomiya. 2021. Hie-BART: Document summarization with hierarchical BART. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 159–165, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. 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.

Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *ACL-IJCNLP 2021*.

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

Marjan Ghazvininejad, Vladimir Karpukhin, and Asli Celikyilmaz. 2021. Discourse-aware prompt design for text generation. *CoRR*, abs/2112.05717.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2021. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. *arXiv preprint arXiv:2108.13684*.

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.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. 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*, Brussels, Belgium.

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

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021*. ACM.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.