Simple Conversational Data Augmentation for Semi-supervised Abstractive Conversation Summarization

Jiaao Chen
School of Interactive Computing
Georgia Institute of Technology
jiaaochen@gatech.edu

Diyi Yang
School of Interactive Computing
Georgia Institute of Technology
dyang888@gatech.edu

Abstract

Abstractive conversation summarization has received growing attention while most current state-of-the-art summarization models heavily rely on human-annotated summaries. To reduce the dependence on labeled summaries, in this work, we present a simple yet effective set of **Co**nversational **D**ata **A**ugmentation (CODA) methods for semi-supervised abstractive conversation summarization, such as random swapping/deletion to perturb the discourse relations inside conversations, dialogue-acts-guided insertion to interrupt the development of conversations, and conditional-generation-based substitution to substitute utterances with their paraphrases generated based on the conversation context. To further utilize unlabeled conversations, we combine CODA with two-stage noisy self-training where we first pre-train the summarization model on unlabeled conversations with pseudo summaries and then fine-tune it on labeled conversations. Experiments conducted on the recent conversation summarization datasets demonstrate the effectiveness of our methods over several state-of-the-art data augmentation baselines. We have publicly released our code at https://github.com/GT-SALT/CODA.

1 Introduction

Abstractive conversation summarization, which targets at processing, organizing and distilling human interaction activities into short, concise and natural text (Murray et al., 2006; Wang and Cardie, 2013), is one of the most challenging and interesting problems in text summarization. Recently, neural abstractive conversation summarization has received growing attention and achieved remarkable performances by adapting document summarization pre-trained models and (Gliwa et al., 2019; Yu et al., 2021) and incorporating structural information (Chen and Yang, 2020; Feng et al., 2020c; Zhu et al., 2020a; Chen and Yang, 2021; Liu et al., 2019b). However, most of these models usually require *abundant human-annotated summaries* to yield the state-of-the-art performances (Gliwa et al., 2019), making them hard to be applied into real-world applications (e.g. summarizing counseling sessions) that lack labeled summaries.

Data augmentation, which perturbs input data to create additional augmented data, has been utilized to alleviate the need of labeled data in various NLP tasks, and can be categorized into three major classes: (1) manipulating words and phrases at the token-level like designed word replacement (Kobayashi, 2018; Niu and Bansal, 2018), word deletion/swapping/insertion (Wei and Zou, 2019; Feng et al., 2020a), token/span cutoff (Shen et al., 2020b); (2) paraphrasing the entire input text at the sentence-level through round-trip translation (Sennrich et al., 2015; Xie et al., 2019; Chen et al., 2020b) or syntactic manipulation (Iyyer et al., 2018; Chen et al., 2020c); and (3) adding adversarial perturbations to the original data which dramatically influences the model's predictions (Jia and Liang, 2017; Niu and Bansal, 2019; Zhang et al., 2019). Despite the huge success, the former two mainly perturbs sentences locally while ignoring the diverse structures and context information in dialogues to create high-quality augmented conversations for summarization. The third one might utilize context through additional backward passes, but often require significant amount of computational and memory overhead (Zhang et al., 2019; Zhu et al., 2019), especially for summarization tasks with long input.

To this end, we introduce simple and novel set of **Co**nversational **D**ata **A**ugmentation (**CODA**) techniques for conversation summarization guided by conversation structures and context, including: (1) **random swapping/deletion** randomly swap or delete utterances in conversations to perturb the discourse relations, (2) **dialogue-acts-guided insertion** randomly insert utterances based on the

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6605–6616
November 7–11, 2021.

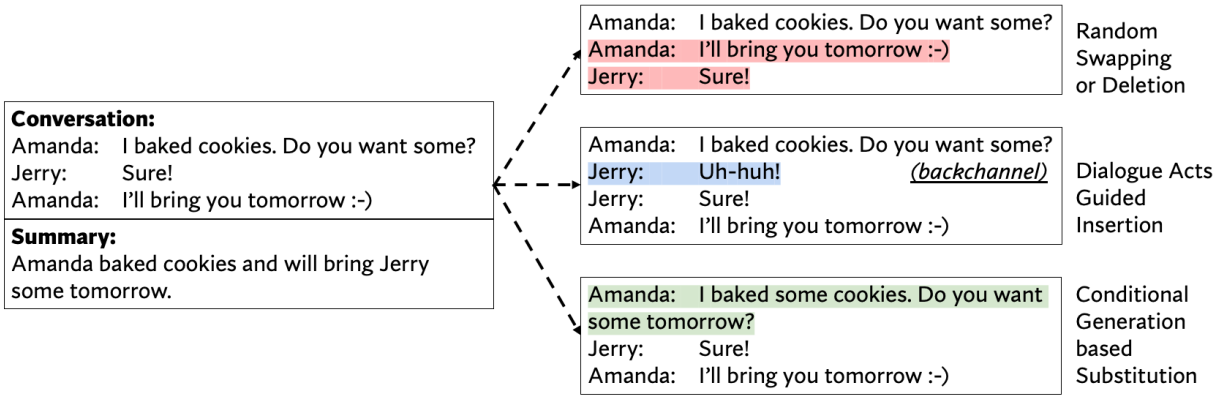

Figure 1: Examples of utilizing different CODA strategies to augment the given conversation including (1) random Swapping/Deletion where last two utterances are swapped (top), (2) dialogue-acts-based Insertion where a backchannel utterance is inserted after the first utterance (middle), and (3) conditional-generation-based substitution where the first utterance is substituted with a model-generated one (bottom).

dialogue acts like self-talk, repeating utterance and back-channel (Allen and Core, 1997; Sacks et al., 1978) to interrupt the conversations, and (3) **conditional-generation-based substitution** randomly substitute utterances in conversations based on pre-trained utterance generation models conditioned on the conversation context. Examples for operations in CODA are shown in Figure 1. To further enhance the performance when labeled summaries are limited, we extend CODA to semi-supervised settings, **Semi-CODA**, where we combine CODA with two-stage noisy self-training (Xie et al., 2020; He et al., 2020) to utilize conversations without annotated summaries. Specifically, we repeat the process where we first generate pseudo summaries for unlabeled conversations with the base summarization model, then we pre-train a new model on pseudo data points and fine-tune the model on labeled conversations to form the updated summarization model. To sum up, our contributions are:

- We propose simple yet effective data augmentation techniques for conversation summarization by considering the structures and context of conversations.
- We introduce a semi-supervised conversation summarization framework by combing CODA and two-stage noisy self-training.
- We demonstrate the effectiveness of our proposed methods through extensive experiments on two conversation summarization datasets , SAMSum (Gliwa et al., 2019) and ADSC (Misra et al., 2015).

2 Related Work

2.1 Abstractive Conversation Summarization

Abstractive conversation summarization has received much attention recently. Other than directly apply document summarization models to conversational settings (Gliwa et al., 2019), models tailored for conversation are designed to achieve the state-of-the-art performances such as modeling conversations in a hierarchical way (Zhao et al., 2019; Zhu et al., 2020b). The rich structured information in conversations are also explored and leveraged such as dialogue acts (Goo and Chen, 2018), key point/entity sequences (Liu et al., 2019a; Narayan et al., 2021), topic segments (Liu et al., 2019c; Li et al., 2019), stage developments (Chen and Yang, 2020), discourse relations (Chen and Yang, 2021; Feng et al., 2020b). External information like commonsense knowledge has also been incorporated to help understand the global conversation context as well (Feng et al., 2020c). However, current summarization models still heavily rely on abundant parallel data to achieve the state-of-the-art performances (Yu et al., 2021). Little work has focused on low-resourced settings where well-annotated summaries are limited or even unavailable. To fill this gap, in this work, we introduce a set of conversational data augmentation techniques to alleviate the dependence on labeled summaries.

2.2 Data Augmentation for NLP

Data augmentation is one of the most common approaches to mitigate the need for labeled data in various NLP tasks (Feng et al., 2021). The augmented data is usually generated by modify-

ing existing data points through transformations while keeping the semantic meaning unaffected like designed word/synonym replacement (Kobayashi, 2018; Niu and Bansal, 2018; Kumar et al., 2020), word deletion/swapping/insertion (Wei and Zou, 2019), token/span cutoff (Shen et al., 2020b), and paraphrasing through round-trip translation (Sennrich et al., 2015; Xie et al., 2019; Chen et al., 2020b). Even though they could be directly applied to conversation summarization settings, these prior techniques mainly modify the text *locally* and largely ignore the structure and context information in conversations to generate more effective and diverse augmented conversations. To this end, our CODA augmentation will perturb the conversation structures and substitute paraphrases by taking into account the conversation context.

2.3 Semi-supervised Learning Methods

Semi-supervised learning methods can further reduce the dependency on labeled data and enhance the models by using large amounts of unlabeled data (Chapelle et al., 2009; Gururangan et al., 2019; Chen et al., 2021). Unlabeled data is usually incorporated through consistency training (Xie et al., 2019; Chen et al., 2020b,a), co-training (Clark et al., 2018), variational auto encoders (Gururangan et al., 2019; Chen et al., 2018; Yang et al., 2017) or self-training (Scudder, 1965; Riloff and Wiebe, 2003; Xie et al., 2020). In this work, we focus on self-training, one of the most classic "pseudo-label" semi-supervised learning approaches (Yarowsky, 1995; Riloff and Wiebe, 2003). Self-training often iteratively incorporates unlabeled data by learning student models from pseudo labels assigned by teacher models. The teacher model could be the model trained on labeled data or the model from last iteration (Zhu and Goldberg, 2009). Recent work showed that combining self-training with better noise/augmentation techniques to perturb the input space greatly improve the performances on classification tasks (Rasmus et al., 2015; Laine and Aila, 2017; Miyato et al., 2019; Xie et al., 2020). However, their impact on language generation tasks like summarization is largely underexplored because, unlike classification tasks, the the pseudo summaries might be quite complicated and very different from human-annotated labels (He et al., 2020). Inspired by these previous self-training work, we will combine our CODA with the two-stage noisy self-training framework (He et al., 2020) for semi-supervised abstractive conversation summarization.

3 Methods on Semi-Supervised CODA

In order to generate more diverse and effective augmented data for conversation summarization and alleviate the reliance on human annotations, we propose a set of simple **Co**nversational **D**ata **A**ugmentation (**CODA**) to perturb conversations based on the conversation structures and global context (Section 3.1). We further introduce **Semi-CODA** under the self-training framework to utilize unlabeled conversations for semi-supervised conversation summarization (Section 3.2).

3.1 CODA

For a given conversation $c = \{u_0, ..., u_n\}$ with n utterances, CODA random performs one of the conversational perturbations described below to generate augmented conversation c' while preserving the semantic information of the global conversation.

Random Swapping or Deletion Utterances from different speakers in conversations usually follow Gricean Maxims (Dale and Reiter, 1995) to achieve effective communication in social situations, which requires utterances to be related to each other orderly under the context of discourse (Murray et al., 2006; Qin et al., 2017). From the perspective of perturbing discourse relations to create augmented conversations (Gui et al., 2021), we introduce two simple operations to perturb the discourse relations: (1) random swapping, which breaks the discourse relations by randomly swapping two utterances in one conversation to messes up the logic chain of utterance, and (2) random deletion, which goes against the discourse requirement by randomly deleting $K_r = \alpha_d \cdot n$ utterances to provide less information in the conversations, where n is the number of utterances in conversations and α_d is a hyper-parameter to control the strength of the deleting perturbation, as shown in Figure 1. In practice, for one conversation c, we combine these two strategies by randomly choosing one of them to generate the augmented conversation c'.

Dialogue Acts Guided Insertion Unlike structured documents, conversations have unique characteristics of interruptions (Allen and Core, 1997) such as repetitions, false-starts, reconfirmations, hesitations and backchanneling (Sacks et al., 1978), making it challenging for summarization models

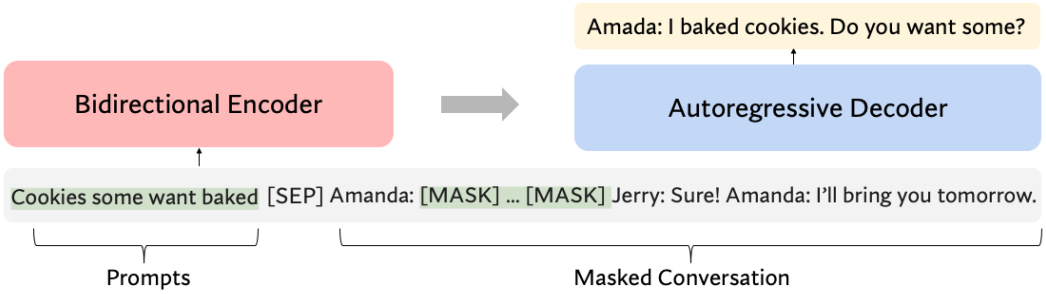

Figure 2: The model architecture for the conditional-generation-based. The randomly shuffled unique tokens in the utterance u_i are prepended to the masked conversation $c^{mask} = \{u_0, ..., u_{i-1}, u_i^{mask}, u_{i+1}, ..., u_n\}$ as input. A transformer-based sequence-to-sequence model is then applied to generate the corresponding utterance u.

to reason over conversations and extract key information. Inspired by these observations, we introduce a novel dialogue-acts-guided insertion to *interrupt* the conversations to generate augmented data. Specifically, for a given conversation c, we randomly insert repeated utterances or insert utterances whose dialogue acts (Jurafsky and Shriberg, 1997) are interruptions including acknowledge/backchannel (e.g., *"Uh-huh"*), response acknowledgement (e.g., *"Oh, okay"*), backchannel in question form (e.g., *"Is that right?"*), self-talk (e.g., *"What is the thing I am thinking of"*), or hedge (e.g., *"I don't know if I'm making any sense or not."*) to generate augmented c', as shown in Figure 1 where an utterance with backchannel act "*Uh-huh!*" is inserted.

For inserting repeated utterances, we randomly select $K_r = \alpha_r \cdot n$ utterances from the input conversation and directly insert them back. For other types of dialogue-acts-guided insertion, we randomly insert $K_d = \alpha_d \cdot n$ utterances sampled from a pre-defined utterance set to random positions in the input conversation. The pre-defined set consists of 8,000 utterances: (1) utterances with desired dialogue acts from a human annotated Switchboard corpus (Jurafsky and Shriberg, 1997), and (2) utterances with desired dialogue acts from high confidence predictions using a state-of-the-art dialogue acts classifier (Raheja and Tetreault, 2019) (with 82.9% accuracy on Switchboard corpus) on SAMSum corpus (Gliwa et al., 2019).

Conditional Generation based Substitution Paraphrasing has been effective as data augmentation on sentence-level tasks like sentence classification (Xie et al., 2019; Chen et al., 2020b) as it could generate sentences with similar semantic meaning but with different word choices. However, when it comes to utterances in a dialogue, simple paraphrasing techniques like round-trip translation (Sennrich et al., 2015) might not be able to capture the context information in conversation, leading to limited diversity and low quality in its augmented utterances. To this end, we propose a conditional generation based method to generate new utterances and substitute the original utterances. , with its architecture shown in Figure 2.

We first pre-train the conditional generation model $\mathbf{g}(.;\theta)$ which could generate an utterance u_i with a masked conversation $c^{mask} = \{u_0, ..., u_{i-1}, u_i^{mask}, u_{i+1}, ..., u_n\}$ and a prompt p_i as input. Specifically, during the pre-training stage, utterance $u_i \in c$ is randomly sampled and substituted with *<MASK>*. The unique tokens in u_i are then randomly shuffled to form the prompt p_i. We initialize the generation model $\mathbf{g}(.;\theta)$ with BART-base (Lewis et al., 2020), and prepend the prompt p_i to the masked conversation c^{mask} as input. The pre-training objective is:

$$\mathcal{L} = -\sum \log P(u_i|\mathbf{g}(p_i, c^{mask}; \theta)) \quad (1)$$

During the augmentation stage, for a random utterance u_i in c, we construct the c^{mask} and p_i in the same way as the pre-training stage. We employ the random sampling strategy with a tunable temperature τ to generate u_i' and construct the augmented conversation c' by substituting u_i with u_i' in c; τ is a hyper-parameter to control the diversities (higher temperature would result in more diverse generations while injecting more noise). In practice, we randomly substitute $K_g = \alpha_g n$ utterances in c with generated utterances from $\mathbf{g}(.;\theta)$.

CODA for Conversation Summarization When training conversation summarization models $\mathbf{f}(.;\theta)$, for any input conversation c with summary

Dataset	Split	# Conv	# Users	# Turns	# Words (Conv)	# Words (Summary)
SAMSum	Full Train	14732	2.40	11.17	83.90	20.35
	Unlabeled	7366	2.41	11.57	84.93	-
	Val	818	2.39	10.83	83.26	20.14
	Test	819	2.36	11.25	83.87	20.43
ADSC	Full	45	2.00	7.51	672.00	150.75

Table 1: Statistics of SAMSum (daily chat) and ADSC (debate) datasets, including the total number of conversations (# Conv), the average number of participants (# Users), the number of turns, the number of words in conversations and summaries per data point.

Algorithm 1 Semi-supervised CODA

Input Labeled conversations $C^l = \{(c_i^l, s_i^l)\}_{i=1:n}$, unlabeled Conversations $C^u = \{(c_i^u)\}_{i=1:m}$, maximum iteration K

Output Conversation summarization model $\mathbf{f}(.)$

1: Train a base model $\mathbf{f}(.)$ on C^l with CODA
2: **for** $t = 1, \ldots, K$ **do**
3: Predict pseudo summaries for C^u with $\mathbf{f}(.)$ without CODA perturbations
4: Pre-train a new model $\mathbf{f}(.)$ on C^u with CODA perturbations
5: Fine-tune $\mathbf{f}(.)$ on C^l with CODA
6: **end for**

s in the training set C, we randomly choose and perform one of the above augmentations to generate c' in each epoch. The objective is:

$$\mathcal{L} = -\mathbb{E}_{(c,s)\sim C}\mathbb{E}_{c'\sim \text{CODA}(c)} \log P(s|\mathbf{f}(c';\theta)) \quad (2)$$

Note that our introduced CODA augmentation techniques can also be combined and performed in a sequential manner. CODA is agnostic to any conversation summarization models. In this work, we utilize the state-of-the-art summarization model, BART (Lewis et al., 2020), as our base model.

3.2 Semi-supervised CODA

To further improve the performance of learning with limited annotated conversations, we combine CODA with two-stage noisy self-training framework (Xie et al., 2020; He et al., 2020) for utilizing unlabeled conversations. The semi-supervised CODA algorithm is shown in Algorithm 1.

Specifically, for a parallel conversation dataset $C^l = \{(c_i^l, s_i^l)\}_{i=1:n}$ where c_i^l is the conversation and s_i^l is the annotated summary, and a large unlabeled dataset $C^u = \{(c_i^u)\}_{i=1:m}$, where $m >> n$. In semi-CODA, a teacher conversation summarization model $f(.;\theta^*)$ is first trained on C^l where CODA perturbations are utilized to inject noise. Then semi-CODA iteratively (1) apply the teacher model $f(.;\theta^*)$ to predict pseudo summaries on unlabeled conversations C^u without any noise injected, (2) pre-train a new summarization model $f(.;\theta)$ on C^u with CODA being applied, (3) fine-tune $f(.;\theta)$ on labeled data C^l with CODA being applied and update the teacher model $f(.;\theta^*)$. The objective function of semi-CODA for annotated conversation is the same as Equation 2, while the objective function for unlabeled conversation is:

$$\mathcal{L}_u = -\mathbb{E}_{c\sim C^u}\mathbb{E}_{c'\sim \text{CODA}(c)} \log P(\mathbf{f}(c;\theta^*)|\mathbf{f}(c';\theta)) \quad (3)$$

Here, θ^* is the parameter from the teacher model (from last iteration) and fixed within the current iteration. In practice, after step (1) in semi-CODA, we apply BERT-score (Zhang* et al., 2020) to calculate the semantic relevance between generated summaries and the unlabeled conversation, and select a subset of C^u with the BERT-score higher than a threshold T for the following steps.

4 Experiments

4.1 Datasets

To demonstrate the effectiveness of our CODA methods on a human-annotated dialogue dataset, we chose SAMSum (Gliwa et al., 2019) that contains open-domain daily-chat conversations such as arranging meetings, planning travels and chit-chat. We use the original validation and test set as our validation and test set. To construct a low-resourced setting, we randomly selected 1% (147) and 5% (735) conversations in the original training set as our training set, and 50% conversations (7366) as unlabeled conversation. We also evaluated the generalizability of our methods on Argumentative Dialogue Summary Corpus (ADSC) (Misra et al., 2015) about summarizing debates. The data statics are shown in the Table 1. During

Model	Unlabeled Data	ROUGE-1			ROUGE-2			ROUGE-L		
		F	P	R	F	P	R	F	P	R
BART-base	no	41.00	52.34	36.35	17.18	22.77	15.29	37.70	48.21	33.43
AdaptSum	no	40.88	52.48	35.13	16.75	22.08	14.35	36.44	46.53	32.51
Token Cutoff	no	41.49	51.18	37.60	16.78	21.43	15.24	37.86	48.76	34.31
Span Cutoff	no	41.32	51.05	37.62	16.88	22.81	15.39	37.77	47.73	33.57
Round-trip Translation	no	41.38	52.91	36.17	17.02	22.35	15.29	37.92	49.47	33.08
Ran. Swapping/Deletion†	no	41.53	51.71	37.46	17.20	22.28	15.45	38.26	47.45	34.34
Dia. Insertion†	no	41.34	50.09	38.34	17.09	21.32	15.77	38.48	46.47	35.50
Cond. Substitution†	no	41.95	51.58	38.24	17.21	22.04	15.65	38.38	47.41	34.99
CODA†	no	**42.16**	52.18	38.14	**17.82**	22.84	16.19	**38.89**	48.16	35.19
Semi. Token Cutoff	yes	43.25	49.23	41.52	18.13	21.27	17.55	39.89	45.49	38.27
Semi. Span Cutoff	yes	43.20	49.20	41.35	18.22	21.56	17.56	40.32	46.00	38.59
Semi. Round-trip Translation	yes	43.49	50.53	41.01	18.70	22.52	17.60	40.37	46.95	38.05
Semi. Ran. Swapping/Deletion†	yes	43.73	50.55	41.23	18.72	21.94	17.81	40.68	46.98	38.68
Semi. Dia. Insertion†	yes	43.37	49.95	41.14	18.56	21.74	17.31	40.29	46.38	38.26
Semi. Cond. Substitution†	yes	43.83	49.97	41.97	18.87	22.05	18.27	40.88	46.58	39.17
Semi-CODA†	yes	**44.34**	50.67	42.32	**19.22**	22.33	18.69	**41.16**	47.03	39.32

Table 2: ROUGE-1, ROUGE-2 and ROUGE-L scores for different methods on the SAMSum Corpus test set with 1% (147) conversations where summaries are used for training. † means our methods.

pre-processing, we separate every utterance in conversations with a special separator ("</s><s>") and truncate the input conversation into 800 tokens.

4.2 Baselines

We compared CODA with several state-of-the-art augmentation techniques and baselines:

- **BART** (Lewis et al., 2020) is the state-of-the-art pre-trained models for summarization. We used BART-base [1] as our base model for all the methods. We also tested **AdaptSum** (Yu et al., 2021) by initializing the summarization model with BART-base pre-trained on XSUM (Narayan et al., 2018) summarization task.
- **Token Cutoff** (Wei and Zou, 2019; Shen et al., 2020a) randomly removes tokens from the input to create perturbed conversation.
- **Span Cutoff** (Shen et al., 2020a) randomly eases a contiguous span of text in conversations to lead to harder perturbed conversation.
- **Round-trip Translation** (Xie et al., 2019; Chen et al., 2020b) generate paraphrases by first translating them to an intermediate language like Romance and then translating them back. This work utilized pre-trained Marian translation model [2] to generate paraphrases.

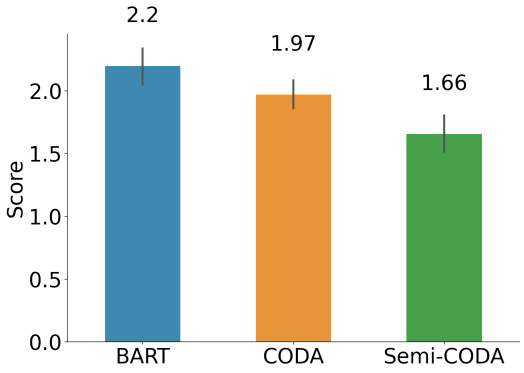

Figure 3: The average ranking every method receives from human evaluation (lower is better).

4.3 Model Settings

For the dialogue acts classifier, we directly followed the settings in Raheja and Tetreault (2019) and applied the trained classifier to predict dialogue acts of utterances in SAMSum corpus. We initialized our conditional generation model with BART-base (Lewis et al., 2020) and trained the model on SAMSum corpus. During augmentation, the sampling temperature is 0.7. α in CODA was selected from $\{0.1, 0.2, 0.3, 0.5\}$. We utilized RoBERTa-large [3] to initialize the BERT-score (rescale with baseline) (Zhang* et al., 2020) and set the filtering threshold $T = 0.25$. The maximum iteration for semi-CODA was set 5. For all the methods, we used BART-base to initialize the conversation summarization model. During training, we used a batch size of 12 for 10 iterations with a 3e-5 learning rate. We used Adam optimizer with momentum

[1] https://huggingface.co/transformers/model_doc/bart.html

[2] https://huggingface.co/transformers/model_doc/marian.html

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

Model	Unlabeled Data	ROUGE-1			ROUGE-2			ROUGE-L		
		F	P	R	F	P	R	F	P	R
BART-base	no	44.56	53.47	41.14	19.90	23.97	18.69	41.29	48.96	38.39
AdaptSum	no	44.60	54.01	40.62	19.88	24.02	18.32	41.38	48.73	38.68
Token Cutoff	no	43.98	52.84	40.57	19.77	24.46	18.33	40.5	48.62	37.42
Span Cutoff	no	44.73	51.24	42.69	20.01	23.38	19.07	40.74	47.69	38.76
Round-trip Translation	no	44.63	53.55	41.16	19.72	24.03	18.33	40.98	48.61	38.04
Ran. Swapping/Deletion†	no	45.14	52.2	42.53	20.3	24.11	19.18	41.54	48.03	39.18
Dia. Insertion†	no	44.72	52.68	41.38	19.78	23.89	18.87	41.38	48.73	38.92
Cond. Substitution†	no	44.69	53.11	41.05	20.10	24.91	19.03	41.69	49.44	38.32
CODA†	no	**45.23**	52.89	42.59	**20.42**	24.76	19.51	**42.02**	49.02	39.23
Semi. Token Cutoff	yes	45.32	51.80	43.05	20.31	23.79	19.34	41.77	47.78	39.71
Semi. Span Cutoff	yes	45.37	52.02	43.19	20.38	24.06	19.52	42.03	48.24	40.08
Semi. Round-trip Translation	yes	45.4	53.14	42.52	20.35	24.57	19.02	42.13	49.34	39.46
Semi. Ran. Swapping/Deletion†	yes	45.78	51.67	44.02	**21.08**	24.41	20.28	42.81	48.35	41.15
Semi. Dia. Insertion†	yes	45.46	53.26	42.68	20.43	24.62	19.01	42.25	48.26	40.33
Semi. Cond. Substitution†	yes	45.86	52.11	43.84	20.52	23.93	19.68	42.51	48.28	40.67
Semi-CODA†	yes	**46.21**	52.86	44.09	21.02	24.73	20.12	**42.85**	48.86	41.07

Table 3: ROUGE-1, ROUGE-2 and ROUGE-L scores for different methods on the SAMSum Corpus test set where 5% (735) conversations with summaries are used for training. † means our methods.

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
BART-base	49.1	54.38	48.42	24.29	27.46	24.08	45.76	50.68	45.15
Token Cutoff	49.16	54.34	48.57	24.09	27.25	23.98	45.73	50.55	45.22
Span Cutoff	49.52	54.77	49.15	24.38	27.75	23.78	46.01	50.85	45.75
Round-trip Translation	49.50	54.18	49.23	24.22	27.11	24.13	46.06	50.44	45.82
Ran. Swapping/Deletion†	49.74	54.72	49.3	24.6	27.65	24.56	46.27	50.89	45.99
Dia. Insertion†	49.61	54.88	48.52	24.54	27.63	24.43	46.18	50.65	45.38
Cond. Substitution†	49.66	55.00	48.86	24.41	27.57	24.07	46.25	51.20	45.56
CODA†	**50.08**	55.18	49.45	**24.62**	27.68	24.55	**46.89**	51.27	46.03

Table 4: ROUGE-1, ROUGE-2 and ROUGE-L scores for different methods on the SAMSum Corpus test set where all (14732) the conversations with summaries are used for training. † means our methods.

Model	Ulbl.	R-1	R-2	R-L
BART-base	no	23.74	4.99	22.21
Token Cutoff	no	24.28	5.03	22.17
Span Cutoff	no	24.46	5.12	22.35
Round-trip Translation	no	23.34	4.74	21.41
CODA†	no	**26.35**	**5.49**	**23.98**
Token Cutoff	yes	26.94	5.57	24.73
Span Cutoff	yes	27.01	5.88	25.32
Round-trip Translation	yes	24.87	4.77	22.02
Semi-CODA†	yes	**28.97**	**6.99**	**27.00**

Table 5: ROUGE scores for different methods on the out-of-domain ADSC Corpus where 1% (147) labeled conversations in SAMSum are used for training.

$\beta_1 = 0.9, \beta_2 = 0.998$. During the decoding stage, we used beam search with a beam size of 4.

4.4 Results

Using Limited Labeled Summaries We varied the number of conversations with summaries for training in both fully-supervised and semi-supervised settings. The ROUGE scores using the *rouge* package [4], were shown in Table 2 (1% (147) labeled data was used) and Table 3 (5% (735) labeled data was used). Compared to *BART-base* by pre-training on a news summarization corpus XSUM (Narayan et al., 2018), AdaptSum (Yu et al., 2021) shows similar performances, probably due to the large differences between news and daily chats. When applying *Cutoff* based augmentations or *Round-trip Translations* to generate new conversations, performances boosted compared to *BART-base* as more data was used in the training. Through perturbing conversation structures to generate harder conversations via randomly swapping/deleting utterances and inserting interruption utterances, *Random Swapping/Deletion* and *Dialogue-acts-guided Insertion* outperformed the baseline augmentation methods. Substituting utterances with more context-aware paraphrases from *Conditional-generation-based Substitution* also consistently improved *Round-trip Translations*. By combining all the conversational augmentation techniques, *CODA* achieved the best scores (e.g., with an increase of 2.8% on ROUGE-1, 3.7% on ROUGE-2 and 3.2% on ROUGE-L compared to

[4] https://github.com/pltrdy/rouge

BART-base when 1% labeled data was used).

After incorporating unlabeled conversations through two-stage noisy self-training framework, all the augmentation methods showed large performance improvements over our base model *BART*. Compared to previous state-of-the-art data augmentations (*Cutoff* and *Round-trip Translation*), our proposed conversational augmentation techniques worked better when combined with noisy self-training as they could provide more effective perturbations. Consistently, our *Semi-CODA* achieved the significantly better performances especially when there are less labeled data (e.g., with an increase of 8.1% on ROUGE-1, 11.9% on ROUGE-2 and 9.2% on ROUGE-L compared to *BART-base* when 1% labeled data was used).

Using All Labeled Summaries Table 4 summarized performances on the full setting where all the labeled data was utilized for training. *CODA* still showed performance gains compared to all baselines, suggesting that our proposed conversational data augmentation methods work well for conversation summarization even when a large number of labeled conversations is available for training.

Human Evaluation We conducted human annotations to evaluate summaries generated by different models trained with 1% (147) conversations from SAMSum. Specifically, we asked annotators from Amazon Mechanical Turk [5] to rank summaries via a 1 (the most preferred) to 3 (the least preferred) scale, generated from *BART*, *CODA* and *Semi-CODA* for randomly sampled 150 conversations. Workers were paid 0.15$ for each ranking task. Every summary triples were ranked by three workers. The rank for every summary was aggregated by majority voting. The Intra-Class Correlation (*ICC1k*) was 0.561, indicating moderate agreement (Koo and Li, 2016)). As shown in Figure 3, our *CODA* and *Semi-CODA* received lower average rankings, which further demonstrated the effectiveness of CODA and Semi-CODA.

Out-of-domain Evaluation We then directly evaluated models trained with 1% (147) conversations with summaries from SAMSum on the debate summarization dataset ADSC (Misra et al., 2015), to investigate the generalization abilities brought by different augmentation methods and unlabeled conversations. As shown in Table 5, consistent with in-domain evaluations, our introduced *CODA* and *Semi-CODA* achieved significantly better out-of-domain ROUGE scores than all the baselines, demonstrating the effectiveness of our designed conversational augmentation methods and the ways to incorporate unlabeled conversations.

[5] https://www.mturk.com/

Model	R-1	R-2	R-L
BART-base	41.00	17.18	37.70
Iteration 0	42.16	17.82	38.89
Iteration 1	42.32	18.22	39.54
Iteration 2	43.89	18.86	40.68
Iteration 3	**44.34**	**19.22**	**41.16**
Iteration 4	43.97	18.79	40.82

Table 6: ROUGE scores for different iterations in Semi-CODA on the SAMSum Corpus test set where 1% (147) labeled conversations are used for training.

Model	R-1	R-2	R-L
BART-base	41.00	17.18	37.70
Jointly-training	42.36	17.29	38.58
Two-stage	**44.34**	**19.22**	**41.16**

Table 7: ROUGE scores for different training strategies in Semi-CODA on the SAMSum Corpus test set where 1% (147) labeled conversations are used for training.

4.5 Ablation Studies

Number of Iterations in Semi-CODA Here we showed the effects of iterative training in Semi-CODA. For all the iterations in Semi-CODA, we adopted the same hyperparameters. As shown in Table 6, ROUGE scores kept improving and achieved the best performance at iteration 3, and then started to converge. This indicates the effectiveness of iterative training in Semi-CODA by continually updating the teacher model to generate better pseudo summaries.

Two-stage Self-training vs. Joint Self-training One alternative in self-training is to merge the labeled conversation and conversations with pseudo summaries and train new models on them *jointly* (Edunov et al., 2018). We compared our *two-stage* training strategy in Semi-CODA with the *jointly-training* with the same set of hyperparameters in Table 7. We found that *two-stage* training outperformed *jointly training*, indicating that our two-stage strategy in *Semi-CODA* could effectively mitigate the noise from pseudo summaries.

5 Conclusion

In this work, we introduced a simple yet effective set of conversational data augmentation methods CODA, for improving conversation summarization in low-resourced settings. To further utilize unlabeled conversations, we proposed Semi-CODA that utilizes a two-stage noisy self-training framework. Experiments on both in-domain and out-of-domain evaluations demonstrated that our CODA augmented conversations better compared to previous state-of-the-art augmentation methods. In the future, we plan to examine diverse conversation structures for conversation augmentation and work on zero-shot conversation summarization tasks.

Acknowledgements

We would like to thank the anonymous reviewers and the members of Georgia Tech SALT Lab for their feedback. This work is supported in part by grants from Cisco and Amazon.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

O. Chapelle, B. Scholkopf, and Eds A. Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *arXiv preprint arXiv:2106.07499*.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020a. Local additivity based data augmentation for semi-supervised ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020b. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel. 2018. Variational sequential labelers for semi-supervised learning. In *Proc. of EMNLP*.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2020c. Controllable paraphrasing and translation with a syntactic exemplar. *ArXiv*, abs/2010.05856.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020a. Genaug: Data augmentation for finetuning text generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 29–42.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020b. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020c. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.

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.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *2018 IEEE Spoken Language Technology Workshop (SLT)*.

Tao Gui, Xiao Wang, Qi Zhang, Qin Liu, Yicheng Zou, Xin Zhou, Rui Zheng, Chong Zhang, Qinzhuo Wu, Jiacheng Ye, Zexiong Pang, Yongxin Zhang, Zhengyan Li, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Bolin Zhu, Shan Qin, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. 2019. Variational pretraining for semi-supervised text classification. *CoRR*, abs/1906.02242.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Dan Jurafsky and E. Shriberg. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. 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 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD19, page 1957–1965, New York, NY, USA. Association for Computing Machinery.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, et al. 2019b. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 24–31.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019c. Topic-aware pointer-generator networks for summarizing spoken conversations. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 367–374, New York City, USA. 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.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Tong Niu and Mohit Bansal. 2019. Automatically learning data augmentation policies for dialogue tasks. 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 1317–1323, Hong Kong, China. Association for Computational Linguistics.

Kechen Qin, Lu Wang, and Joseph Kim. 2017. Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 974–984.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. 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 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

H. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *Computer Science*.

Dinghan Shen, M. Zheng, Y. Shen, Yanru Qu, and W. Chen. 2020a. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *ArXiv*, abs/2009.13818.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020b. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405.

Jason W Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. *CoRR*, abs/1702.08139.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2019. Adversarial attacks on deep learning models in natural language processing: A survey.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, WWW '19, page 3455–3461, New York, NY, USA. Association for Computing Machinery.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

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

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020b. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.