GENDEX: Generative Data Augmentation Strategy Leveraging External Data for Abstractive Dialogue Summarization

Sangwon Park[1], **Dongha Choi**[1], **Hongseok Choi**[2], and **Hyunju Lee**[1*]
[1]GIST Artificial Intelligence Graduate School, Gwangju, South Korea,
[2]Electronics and Telecommunications Research Institute, Daejeon
{psw, dongha528}@gm.gist.ac.kr, hongking9@etri.re.kr, hyunjulee@gist.ac.kr

Abstract

With the proliferation of digital communication, dialogue summarization has become increasingly important. However, it still faces a shortage of data. To address this issue, we developed **Gen**erative **D**ata Augmentation Strategy Leveraging **Ex**ternal Data for Abstractive Dialogue Summarization (**GENDEX**), which is based on the hypothetical foundation that texts containing people and their interpersonal interactions can potentially serve as summaries of corresponding dialogues. We filter short texts containing people and resolve coreferences for better contextual analysis. We then identify the semantic roles of words within the texts and filter them based on the patterns observed in the dialogue summarization datasets. Using these texts, we generate synthetic dialogues through a controlled generation method. To better leverage the augmented data, we utilize noise-tolerant training to fine-tune the summarization model. The experimental results demonstrate the effectiveness of our proposed method, showing its robust performance, generalizability, and scalability. Moreover, performance improvements by *GENDEX* were observed regardless of complexity of dialogues. The code is available at https://github.com/DMCB-GIST/GENDEX.

1 Introduction

Text summarization is to generate short texts containing essential information (Radev et al., 2002). In an era overflowing with web information, it can serve as a convenient tool. To date, many summarization studies have focused on document data. Recently, the need for dialogue summarization has emerged, following the growing usage of mobile communications and social network services. Dialogue summarization is not just about reducing content but also enhancing comprehension and the utility of conversations by making information more accessible, supporting business intelligence, and aiding decision making. Techniques that automatically summarize dialogues can be advantageous in various fields requiring interpersonal communications, such as customer service (Liu et al., 2019a), business meetings (Feng et al., 2020), and medical consultations (Joshi et al., 2020a).

However, there is a significant lack of dialogue summarization data, making it a more challenging task than document summarization (see to Appendix A for comparative results). There are some reasons for this. While document summarization datasets can be collected in an automated manner by utilizing titles, headlines, and abstracts as summaries, dialogue data require manual annotations (Feng et al., 2021). Moreover, there are privacy issues of real dialogues, which must be addressed to publicize them (Zhu et al., 2021). Further, dialogue data are structurally complex. Dialogues inherently consist of turns. This turn-based structure is complex compared to the straightforward structure of plaintext-based documents. These properties complicate the collection, organization, and publicization of dialogue data.

To address data shortage, various approaches have been explored. One way is transfer learning utilizing abundant document summarization data to benefit from adaptation (Yu et al., 2021; Zhang et al., 2021). Another way is a simple perturbation-based method at the token level, such as synonym replacement and random word deletion (Wei and Zou, 2019; Kumar et al., 2020; Kobayashi, 2018). Chen and Yang (2021a) introduced a simple augmentation method that slightly perturbs the original data at the utterance level, considering conversational characteristics. Gunasekara et al. (2021) used the original training data for reverse-training and synthesizing new dialogues.

Although these approaches have improved the dialogue summarization performance, several limitations still remain. Transfer learning methods may

*Hyunju Lee is the corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 3171–3185

Dialogue:
Lucas: Hey! How was your day?
Demi: Hey there! It was pretty fine, actually, thank you! I just got promoted! :D
Lucas: Whoa! Great news! Congratulations! Such a success has to be celebrated.
Demi: I agree! :D Tonight at Death & Co.?
Lucas: Sure! See you there at 10pm?
Demi: Yeah! See you there! :D

Summary: Demi got promoted. She will celebrate that with Lucas at Death & Co at 10 pm.

Figure 1: Example of the dialogue summarization data from SAMSum (Gliwa et al., 2019).

not fully reflect the characteristics of dialogue data. Moreover, as described by Zhang et al. (2021), pre-training on multiple datasets does not always guarantee further improvements. Perturbation-based approaches cannot significantly improve semantic diversity. In addition, they could harm performance because they modify the original data (Wei and Zou, 2019). The approach that uses the original data for reverse-training and generation has an upper limit on the amount of data that can be augmented. Further, there is a trade-off between generation performance and amount of data being augmented (Gunasekara et al., 2021).

Figure 1 shows an example of dialogue summarization data. Dialogue summaries generally include people who appear in the dialogues and their interpersonal interactions. We analyzed three popular dialogue summarization datasets and found out that 97 - 99% of summaries contain people (see Appendix B for more details). This finding emphasizes the critical role that person names play, suggesting their presence as a key element for capturing the context of dialogues. Motivated by these, we hypothesize that texts including people and their interpersonal interactions can potentially serve as summaries of the corresponding dialogues. Building on this insight, we propose **Gen**erative **D**ata Augmentation Strategy Leveraging **Ex**ternal Data for Abstractive Dialogue Summarization (**GENDEX**). To address the shortage of dialogue data, we utilize abundant external out-of-domain (OOD) data. We apply Named Entity Recognition (NER) to select texts containing person names, resolve coreferences for context, and determine semantic roles to filter texts aligning with dialogue summary patterns. These steps enable us to generate dialogues from curated texts. We then generate synthetic data using controlled generation. Lastly, we use two-stage noise-tolerant training to better utilize synthetic data. Experimental results demonstrate that *GENDEX* is an effective technique in various aspects. It not only performs well in both quantitative and qualitative evaluations but also exhibits generalizability and scalability. In addition, it improves performance regardless of complexity of dialogues.

2 Related Works

As dialogue summarization has recently received much attention, several datasets have been proposed in various domains, such as meetings (Zhong et al., 2021), chats (Chen et al., 2021; Gliwa et al., 2019), and customer services (Feigenblat et al., 2021). Several methods have been proposed for effective dialogue summarization. Most initial works directly used document summarization models for dialogue summarization (Gliwa et al., 2019). Subsequently, the structural and contextual information of the dialogue data was explored. Chen and Yang (2020) extracted conversational structures from different views and then incorporated them for better representation. Feng et al. (2020) enhanced the understanding of dialogues by introducing discourse relations using a relational graph encoder. Chen and Yang (2021b) incorporated discourse relations and action triples to better represent interactions. Zhang et al. (2021) explored the transfer learning between document and dialogue data. Zhong et al. (2022a) proposed a pre-training framework for long dialogue summarization task. However, many models still require large amounts of data to achieve cutting-edge performance (Yu et al., 2021). Several methods have been proposed to improve dialogue summarization performance in such low-resource environments. Yu et al. (2021) utilized abundant document datasets for adaptation. Additional techniques, such as data modification and augmentation have also been proposed. Chen and Yang (2021a) proposed a data augmentation method consisting of swapping, deletion, insertion, and substitution at utterance level. Other augmentation methods by replacing text sections in dialogue and summary using pre-trained language model (Liu et al., 2022; Ouyang et al., 2023) were also introduced. Gunasekara et al. (2021) proposed a generative augmentation technique using a portion of the training data.

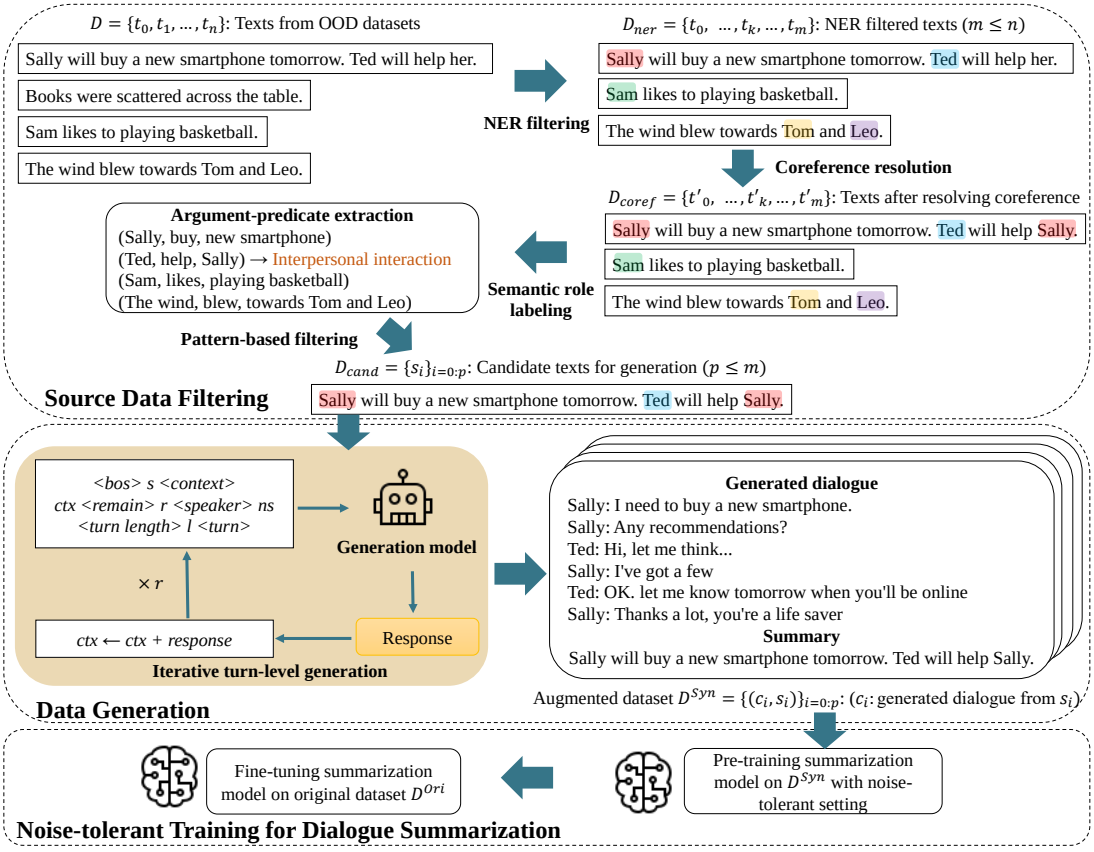


Figure 2: Illustration of the example-based overall pipeline of *GENDEX*

3 Method

In this section, we introduce *GENDEX*, an effective generative data augmentation strategy for abstractive dialogue summarization. Our approach begins with source data filtering process to ensure the quality of the seed texts for generation (Section 3.1). We then generate synthetic data (Section 3.2). Finally, we train the dialogue summarization model to better utilize the augmented data (Section 3.3). The overall pipeline of *GENDEX* is illustrated in Figure 2.

3.1 Source Data Filtering

NER Filtering Unlike document summarization, accurately matching the speakers with their utterances and actions is important when summarizing a dialogue. Thus, most dialogue summaries contain people who participate in the dialogues as shown in Figure 1. Thinking about this in reverse, we attempt to generate dialogues from short texts containing people and their interpersonal interactions. We analyzed summaries in three dialogue summarization datasets, and discovered that 98, 99, and 99% of the summaries in SAMSum, TweetSumm, and DialogSum include at least one person, respectively. Furthermore, summaries containing more than two people are 80, 98, and 94% of the total in each respective dataset (see Appendix B for more details). Considering the interpersonal interactions that constitute the content of dialogues, we filter texts involving at least two different people. We use NER to detect persons in texts. NER is an information extraction technique that identifies and classifies named entities in text into predefined categories, such as person, organization, location, etc (Nadeau and Sekine, 2007). We use the spaCy NER tool, which is a powerful, transformer-based English language processing tool by spaCy. We first collect large out-of-domain (OOD) text data $D = \{t_0, t_1, ..., t_n\}$. We then filter short texts containing person names from D and obtain NER-filtered texts $D_{ner} = \{t_0, ..., t_k, ..., t_m\}(m \leq n)$.

Coreference Resolution The next step is coreference resolution. Coreference resolution is the task identifying expressions that refer to the same entity within a text. For example, if we have a text "Sally will buy a new smartphone. Ted will help her.", this can be transformed to "Sally will

buy a new smartphone. Ted will help Sally." NER-filtered texts still have many personal pronouns. In this step, we resolve such coreferences in the texts from $D_{ner} = \{t_0, ..., t_k, ..., t_m\}$ and obtain $D_{coref} = \{t'_0, ..., t'_k, ..., t'_m\}$. By replacing personal pronouns with the corresponding people, we can analyze interactions more clearly. We use AllenNLP, which is an open-source platform built on PyTorch to solve the NLP task.

Semantic Role Labeling (SRL) and Pattern-Based Filtering Unlike document summarization stated in a narrative format of a single speaker, dialogue summarization involves interactions between multiple participants. Previous studies showed that actions represent explicit information between participants (Chen and Yang, 2021b) and event-oriented text can be an effective source for generation (Daniel et al., 2003). Therefore, we focus on filtering texts containing interpersonal interactions. To this end, we use SRL. SRL is identifying semantic roles of words in a sentence. These roles help explain the relationship between predicates and arguments. A predicate is typically a verb indicating an action, event, or state. An argument is typically an entity that participates in the action or event described by the predicate (Larionov et al., 2019). We analyzed the summaries in the dialogue summarization datasets using SRL to find frequent patterns focusing on predicate-argument structures and their positional relations (more detailed examples can be found in Appendix C). Based on this, we filter texts in $D_{coref} = \{t'_0, ..., t'_k, ..., t'_m\}$ and obtain $D_{cand} = \{s_i\}_{i=0:p}(p \leq m)$, which contains texts that match these patterns and can be used as source texts for dialogue generation.

3.2 Controlled Dialogue Generation

Controlled text generation is generating texts whose attributes can be controlled by adding components to the input sequence (Prabhumoye et al., 2020). The controllable attributes can include style, content, and plot. To enhance the model's understanding of dialogue context, we process the original dialogue summarization datasets (i.e., SAMSum, TweetSumm, and DialogSum) to make turn-level inputs and train the model to learn the connectivity among conversational components. The model is trained using this processed data, with constraints imposed on syntactic, semantic, and length aspects. We use DialoGPT (Zhang et al., 2020b), which is a dialogue response generation model pre-trained on Reddit data. We make turn-level dialogue generation inputs containing control parameters by processing the original data. The following five elements are used as control parameters to generate dialogue turns. (1) Summary (*s*): the dialogue summary containing key contents. (2) Context (*ctx*): The dialogue history before the turn to be generated. Empty for the first turn. (3) The number of remaining turns to generate (*r*). (4) The speaker of the next turn (*ns*). (5) The length of the next turn (*l*). We use the following input representation to fine-tune our generation model: $< bos > s < context > ctx < remain > r < speaker > ns < turn\ length > l < turn > next\ turn < eos >$. Given a summary text $S = \{s_1, ..., s_n\}$, the number of remaining turns r, dialogue context $C = \{c_1, ..., c_m\}$ ($r+m = k, k$ is the total number of turns of dialogue), next speaker ns, and the length of the next turn l, the goal is to generate a target response $T = \{t_1, ..., t_l\}$. The conditional probability of $P(T|S, C, r, ns, l)$ can be written as the product of a series of conditional probabilities:

$$P(T|S, C, r, ns, l) = \prod_{n=1}^{l} p(t_n|t_1, ..., t_{n-1}, S, \{c_p\}_{p=1:k-r}, ns) \quad (1)$$

The next step is to generate synthetic data. We extract person names from the filtered text ft using NER and obtain $P = \{p_1, ..., p_n\}$. The elements in P act as participants in the generated dialogue and can be assigned to the value of ns in the input sequence. Then, we set the appropriate number of turns to be generated. We analyzed the ratio between the length of summary and the number of turns from the data of three dialogue summarization datasets and observed that the average ratio of $len(summary)$ / $\#turns$ is 2.5. These two variables have a positive correlation. If a dialogue contains much information, the summary that condenses it tends to be long, and vice versa. Based on these statistical observations, we set the appropriate number of turns, which will decide the number of iterations for turn-level generation. During the generation, we indicate the number of remaining turns using the control parameter r. r decreases by one for each generation of a turn and the generation is repeated until r is reduced to 1. We randomly select the next speaker from among the elements in P and the length of a turn l in a range

of 3-15 at each iteration. The generated response T is added to context C for the next iteration. In summary, we give the model the following input representation at inference time so that it can generate the next turn by text completion: $< bos > ft < context > C < remain > r < speaker > p_i \in P < turn\ length > l < turn >$. We generate dialogues turn-by-turn so that the model can generate a more context-relevant response, providing it gradually accumulating context information with conversational components.

3.3 Noise-tolerant training

Synthetic data may contain some noise compared to human-labeled original data. To effectively train the model on such noisy data, we train our summarization model using early stopping based two-stage noise-tolerant training setting. In the first stage, we train the model using synthetic data while monitoring its performance on the validation set from the original dataset. We then implement early stopping based on the validation performance to prevent overfitting on the noisy data. The objective function of the first stage is

$$L_1 = E_{(c',s') \in D^{Syn}} \log P(s'|c') \quad (2)$$

where D^{Syn} is the synthetic dataset, c' is the generated dialogue, and s' is filtered short text which will act as summary. In the second stage, we train the model on original data. The objective function is

$$L_2 = E_{(c,s) \in D^{Ori}} \log P(s|c) \quad (3)$$

where D^{Ori} is the original dataset, c is the original dialogue, and s is the original summary in D^{Ori}.

4 Experiments

4.1 Datasets and Metrics

We experimented on three public dialogue summarization datasets. The detailed dataset statistics are presented in Appendix D. SAMSum (Gliwa et al., 2019) contains messenger-style dialogues on daily topics. TweetSumm (Feigenblat et al., 2021) contains chat dialogues between agents and customers in customer service. DialogSum (Chen et al., 2021) contains spoken daily dialogues. We use three text corpora for synthetic data generation: BookCorpus, Wikipedia, and ROCStories. BookCorpus (Zhu et al., 2015) is a large collection of novel books. Wikipedia is a large collection of articles from the Wikipedia website. To handle articles containing people, we especially used Wikipedia-person[1], which is a filtered version with only pages about people. ROCStories (Mostafazadeh et al., 2016) is a dataset of five-sentence stories designed for story understanding and generation. We filtered 20K texts each from BookCorpus and Wikipedia, and 9K texts from ROCStories and used them as the source texts for dialogue generation. We adopted the ROUGE metric (Lin, 2004) for automatic evaluation. ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) are used to calculate the unigram overlap, bigram overlap, and longest common subsequence between the model output and reference summary, respectively. We use the F1-score for R-1, R-2, and R-L.

4.2 Methods for Comparison

To compare various aspects of different methods applied in low-resource environments, we selected methods used for transfer learning, general-text data augmentation, and dialogue data augmentation as baselines. **BART-base** (Lewis et al., 2020) is a popular pre-trained model often used for summarization tasks. We use it as a backbone model in our experiments. **AdaptSum**$_{XSUM}$ and **AdaptSum**$_{CNN/DM}$ (Yu et al., 2021; Zhang et al., 2021) are BART-base models pre-trained on XSUM (Narayan et al., 2018) and CNN/DailyMail (Hermann et al., 2015), respectively. These are two widely used news-domain document summarization datasets. These methods are used for transfer learning to improve the performance on low-resource data using abundant OOD data. **Synonym Replacement (SR)** (Kumar et al., 2020; Kobayashi, 2018; Wei and Zou, 2019) and **Token Cutoff (TC)** (Wei and Zou, 2019) are general token-level augmentation methods for text data. **SR** replaces random words while maintaining their semantic meanings. **TC** removes random tokens in order to give perturbations to data. **Summ grounded aug (SGA)** (Gunasekara et al., 2021) applies reverse-training and then augments data using a portion of the training data. **CODA** (Chen and Yang, 2021a) augments data by slightly perturbing the training data at utterance level. These two methods are used especially for dialogue data augmentation.

[1] https://huggingface.co/datasets/rcds/wikipedia-persons-masked

Model \ Dataset	SAMSum			TweetSumm			DialogSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
BART	$45.80_{\pm.21}$	$22.51_{\pm.29}$	$38.71_{\pm.31}$	$33.87_{\pm.36}$	$16.17_{\pm.27}$	$29.96_{\pm.25}$	$40.08_{\pm.11}$	$15.93_{\pm.16}$	$33.47_{\pm.12}$
Transfer Learning									
$\text{AdaptSum}_{\text{XSUM}}$	$46.63_{\pm.24}$	$23.09_{\pm.35}$	$39.15_{\pm.22}$	$34.12_{\pm.40}$	$16.07_{\pm.39}$	$30.09_{\pm.28}$	$40.33_{\pm.14}$	$15.91_{\pm.16}$	$33.69_{\pm.12}$
$\text{AdaptSum}_{\text{CNN/DM}}$	$46.73_{\pm.31}$	$23.67_{\pm.27}$	$39.33_{\pm.30}$	$34.28_{\pm.19}$	$16.12_{\pm.16}$	$30.10_{\pm.17}$	$40.16_{\pm.13}$	$16.13_{\pm.11}$	$33.54_{\pm.13}$
Text Data Augmentation									
Synonym Replacement	$46.89_{\pm.16}$	$23.71_{\pm.13}$	$39.70_{\pm.12}$	$34.29_{\pm.16}$	$16.29_{\pm.03}$	$30.20_{\pm.10}$	$40.13_{\pm.12}$	$16.15_{\pm.14}$	$33.59_{\pm.06}$
Token Cutoff	$46.75_{\pm.21}$	$23.64_{\pm.20}$	$39.61_{\pm.19}$	$34.23_{\pm.29}$	$16.28_{\pm.21}$	$30.19_{\pm.21}$	$40.11_{\pm.10}$	$16.13_{\pm.11}$	$33.55_{\pm.15}$
Dialogue Data Augmentation									
Summ grounded aug	$46.22_{\pm.28}$	$23.28_{\pm.24}$	$39.38_{\pm.21}$	$34.17_{\pm.23}$	$16.19_{\pm.21}$	$30.33_{\pm.22}$	$40.10_{\pm.21}$	$16.20_{\pm.16}$	$33.61_{\pm.22}$
CODA	$46.99_{\pm.29}$	$23.80_{\pm.37}$	$39.94_{\pm.34}$	$34.42_{\pm.35}$	$16.31_{\pm.27}$	$30.23_{\pm.22}$	$40.15_{\pm.28}$	$16.16_{\pm.29}$	$33.62_{\pm.31}$
Ours									
$\text{GENDEX}_{\text{BC}}$	$\mathbf{47.83}_{\pm.21}$	$24.25_{\pm.24}$	$40.28_{\pm.26}$	$35.55_{\pm.20}$	$16.80_{\pm.29}$	$31.18_{\pm.32}$	$40.78_{\pm.27}$	$16.69_{\pm.23}$	$34.21_{\pm.28}$
$\text{GENDEX}_{\text{WIKI}}$	$47.79_{\pm.24}$	$24.43_{\pm.31}$	$40.43_{\pm.30}$	$\mathbf{35.91}_{\pm.26}$	$\mathbf{16.98}_{\pm.32}$	$\mathbf{31.31}_{\pm.22}$	$\mathbf{40.98}_{\pm.25}$	$\mathbf{17.03}_{\pm.18}$	$\mathbf{34.65}_{\pm.24}$
$\text{GENDEX}_{\text{ROC}}$	$47.77_{\pm.29}$	$\mathbf{24.45}_{\pm.30}$	$\mathbf{40.55}_{\pm.24}$	$34.85_{\pm.25}$	$16.68_{\pm.29}$	$30.69_{\pm.32}$	$40.56_{\pm.26}$	$16.58_{\pm.21}$	$34.16_{\pm.28}$

Table 1: Evaluation results on three dialogue summarization datasets: SAMSum, TweetSumm, and DialogSum. 'R' denotes the ROUGE metric. BC, WIKI, and ROC denote the source data, Bookcorpus, Wikipedia, and ROCStories, respectively. The subscripted numbers represent the standard deviation.

5 Results and Discussion

5.1 Main Result

Table 1 shows the test results on the three dialogue summarization datasets. Pre-training on document summarization datasets, such as $AdaptSum_{XSUM}$ and $AdaptSum_{CNN/DM}$, helps in dialogue summarization. $AdaptSum_{CNN/DM}$ shows slightly better performance than $AdaptSum_{XSUM}$ on average. This could be because CNN/DailyMail is larger than XSUM (approximately 312 K and 226 K samples, respectively). For R-1, R-2, and R-L, $AdaptSum_{XSUM}$ and $AdaptSum_{CNN/DM}$ improved BART's performance across all datasets by averages of (1.1%, 0.6%, 0.7%) and (1.1%, 2%, 0.8%), respectively. Such simple pre-training on document datasets generally improves the performance on dialogue datasets. *SR* and *TC* improved performance by (1.2%, 2.5%, 1.2%) and (1.1%, 2.3%, 1.1%) on average, respectively. These results surpass $AdaptSum_{XSUM}$ and $AdaptSum_{CNN/DM}$. In other words, improving robustness by perturbing data through *SR* and *TC* works better than simply pre-training on document data. *CODA* performs better than these methods. This is because *CODA* was specifically designed for dialogue data considering its distinct features. *CODA* also outperforms *SGA*. On average, *CODA* and *SGA* improved BART's R-1, R-2, and R-L performance by (1.5%, 2.7%, 1.5%) and (0.6%, 1.7%, 1.1%), respectively. Our method, *GENDEX*, improved the performance by the largest margin, with average increases of 3.6%, 5.9%, and 3.6% for R-1, R-2, and R-L, respectively.

Comparison with Transfer Learning Dialogue summarization datasets are usually small so it is effective to introduce external knowledge (Zhang et al., 2021). Therefore, pre-training on abundant document datasets has been used. However, obvious structural and contextual gaps exist between document and dialogue data. Thus, pre-training on different data may sometimes hurt the performance because of differences between source and target domains (Zhang et al., 2021). Therefore, simply pre-training on document data may not be the best for dialogue summarization. Meanwhile, our method can generate dialogue-formatted data from OOD short texts. Thus, data that demonstrate conversational characteristics can be used for training. Based on these observations, we can conclude that not only external knowledge but also task similarity is important, and *GENDEX* can satisfy both of them. Moreover, our method achieved better performance despite using less data. *GENDEX* used at most 20 K samples for training, whereas $AdaptSum_{CNN/DM}$ and $AdaptSum_{XSUM}$ used 312 K and 226 K samples, respectively.

Comparison with Text Data Augmentation *SR* and *TC* improved dialogue summarization performance. However, there exist possibilities to hurt performance because they perturb the original data. Replacing words can change the identity of a sentence, and deleting words can make a sentence unintelligible (Wei and Zou, 2019). Therefore, determining an appropriate augmentation ratio is important. However, our method does not alter the original data. Instead, it leverages external data,

Model \ Dataset	SAMSum			TweetSumm			DialogSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
T5-small (60M)	$44.03_{\pm.19}$	$20.42_{\pm.28}$	$36.65_{\pm.15}$	$35.01_{\pm.18}$	$15.93_{\pm.12}$	$30.35_{\pm.19}$	$37.80_{\pm.23}$	$14.15_{\pm.22}$	$31.82_{\pm.14}$
+ GENDEX	$45.20_{\pm.23}$	$21.31_{\pm.18}$	$37.78_{\pm.17}$	$35.53_{\pm.16}$	$16.38_{\pm.18}$	$30.94_{\pm.21}$	$38.61_{\pm.18}$	$14.75_{\pm.23}$	$32.46_{\pm.25}$
DialogLED-base (162M)	$47.79_{\pm.15}$	$23.49_{\pm.18}$	$39.86_{\pm.20}$	$44.37_{\pm.24}$	$21.17_{\pm.27}$	$37.88_{\pm.23}$	$43.28_{\pm.27}$	$17.52_{\pm.21}$	$35.08_{\pm.26}$
+ GENDEX	$49.06_{\pm.22}$	$24.52_{\pm.20}$	$40.72_{\pm.19}$	$45.88_{\pm.21}$	$22.47_{\pm.22}$	$39.12_{\pm.28}$	$43.94_{\pm.24}$	$18.30_{\pm.29}$	$35.75_{\pm.30}$
PEGASUS-large (568M)	$51.64_{\pm.16}$	$27.47_{\pm.21}$	$43.16_{\pm.23}$	$45.09_{\pm.20}$	$21.88_{\pm.23}$	$38.48_{\pm.26}$	$45.65_{\pm.19}$	$19.97_{\pm.25}$	$37.40_{\pm.20}$
+ GENDEX	$52.72_{\pm.24}$	$27.91_{\pm.22}$	$43.51_{\pm.30}$	$46.12_{\pm.22}$	$22.61_{\pm.19}$	$38.84_{\pm.20}$	$45.99_{\pm.23}$	$20.42_{\pm.21}$	$37.96_{\pm.26}$

Table 2: Results of applying GENDEX to other language models

ensuring that the knowledge of original data is preserved. In Table 1, *GENDEX* shows better performance than *SR* and *TC*.

Comparison with Dialogue Data Augmentation *SGA* uses a portion of the training set to augment data. Therefore, it has an upper bound to the amount of data that can be augmented. Also, as mentioned by Gunasekara et al. (2021), there is a trade-off between the generation performance and amount of augmented data. However, our method leverages external data for augmentation and can augment large data without such limitations. *CODA* augments dialogue data by slightly perturbing original data at the utterance level and provides robustness. *GENDEX* introduces external knowledge by leveraging OOD data. In this experiment, the external knowledge introduced by our method appears to contribute more to the performance (see Appendix E for more details).

Evaluation on Different Datasets As shown in Table 1, all models show lower performance on TweetSumm and DialogSum compared to SAMSum. This is because these two datasets present more challenging tasks than SAMSum. Our method works well on the three dialogue summarization datasets. In particular, it significantly improves performance on TweetSumm compared to other methods. TweetSumm features a longer dialogue length and smaller amount of data than the other two datasets (see Appendix D). It has 20% of the data volume of SAMSum and 23% of that of DialogSum. Moreover, it was annotated under strict conditions (Feigenblat et al., 2021). *GENDEX* also performs well on DialogSum, which is more abstractive, open-domain, and spoken analogous (Chen et al., 2021). These results suggest that *GENDEX* can perform well on more challenging tasks.

	Model	Coh.	Con.	Flu.	Rel.	overall
SAMSum	BART	0.914	0.895	0.922	0.782	0.878
	$\text{GENDEX}_{\text{BC}}$	**0.920**	**0.901**	0.928	**0.803**	**0.888**
	$\text{GENDEX}_{\text{WIKI}}$	0.917	0.896	**0.929**	0.796	0.885
	$\text{GENDEX}_{\text{ROC}}$	0.919	0.896	0.928	0.797	0.885
TweetSumm	BART	0.837	0.831	0.790	0.642	0.775
	$\text{GENDEX}_{\text{BC}}$	**0.857**	**0.844**	0.770	0.664	0.784
	$\text{GENDEX}_{\text{WIKI}}$	0.847	0.839	**0.801**	0.673	**0.790**
	$\text{GENDEX}_{\text{ROC}}$	0.822	0.835	0.800	**0.674**	0.783
DialogSum	BART	0.928	0.910	0.921	0.839	0.900
	$\text{GENDEX}_{\text{BC}}$	0.936	0.918	0.931	0.851	0.909
	$\text{GENDEX}_{\text{WIKI}}$	0.942	**0.924**	**0.932**	0.859	0.914
	$\text{GENDEX}_{\text{ROC}}$	**0.943**	0.920	**0.932**	**0.865**	**0.915**

Table 3: Quality analysis results of summaries

Different Source Data We used three OOD data sources to filter the texts: Bookcorpus, Wikipedia, and ROCStories. As shown in Table 1, they showed slightly different performance, indicating the injection of different types of external knowledge. Using texts filtered from these corpora improved performance and outperformed the baselines. $\textit{GENDEX}_{WIKI}$ achieved the most significant improvement on average. $\textit{GENDEX}_{ROC}$ demonstrated slightly better results on SAMSum compared to using other OOD data, despite having the smallest amount of data.

5.2 Generalizability and Scalability

We experimented on other language models to verify the generalizability and scalability of *GENDEX*. The Wikipedia corpus was used as the source dataset in this experiment. We applied our method to T5 (Raffel et al., 2020), DialogLED (Zhong et al., 2022a), and PEGASUS (Zhang et al., 2020a). We used T5-small (60M), DialogLED-base (162M), and PEGASUS-large (568M). As shown in Table 2, *GENDEX* improved the performance of all models when tested on the three datasets. It per-

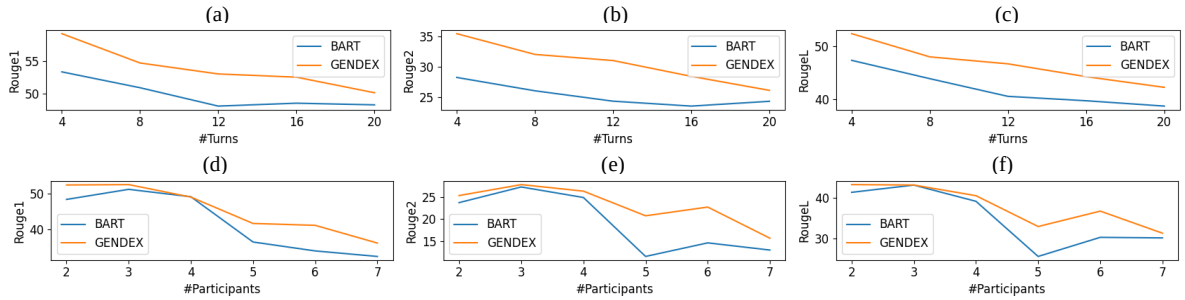


Figure 3: Test results based on the number of turns (a, b, c) and participants (d, e, f)

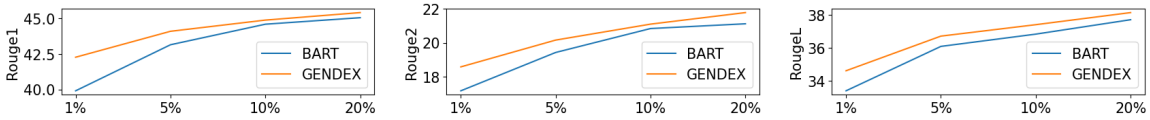


Figure 4: Test results in limited data setting

forms well on models that are both smaller and larger than BART-base (140M). These findings suggest that *GENDEX* is not only applicable to other types of models but also exhibits scalability. Moreover, although DialogLED is already pre-trained on dialogue-domain data, our method can further improve its performance.

5.3 Evaluation Based on Difficulty

Turns and participants are the two essential components of dialogue data. As the number of turns and participants increases, the dialogue becomes more complex. Therefore, the performance generally decreases as they increase (Chen and Yang, 2020). The performance shown in Table 1 may not ensure satisfactory performance across all difficulty levels. Thus, we evaluated performance according to the number of turns and participants using the SAMSum test set. As shown in Figure 3, the performance generally decreases as the number of turns and participants increases, indicating increasing task difficulty. Our method improved the performance across all difficulty levels. We can infer that *GENDEX* has a good ability on summarizing complex dialogue data effectively.

5.4 Quality Analysis

Similarity-based metrics such as ROUGE may not fully capture aspects such as fluency, consistency, and coherence. Therefore, we evaluated the quality of the summaries using a multi-dimensional evaluator (Zhong et al., 2022b) on three datasets. The summaries were evaluated in terms of the following four aspects: coherence, consistency, fluency, and relevance. These metrics assess the following: whether all sentences form a coherent whole, whether there is factual alignment between the summary and source text, the quality of individual sentences, and whether the summary encapsulates the important information from the source text. As shown in Table 3, *GENDEX* generally improves performance in all four perspectives. Especially, there was the largest increase in relevance. This result indicates that our method enables the model to accurately capture and reflect the important contents of dialogue, ensuring semantic alignment with the source dialogue.

5.5 Limited Data Setting

As mentioned in Section 5.1, TweetSumm comprises 20% and 23% of the data volume of SAMSum and DialogSum, respectively. However, to test our method under stricter environments, we experimented in limited data settings. We randomly sampled 1%, 5%, 10%, and 20% training data from SAMSum and applied our method to BART using these subsets. For a fair comparison, we used the same entire test set in all settings. As shown in Figure 4, *GENDEX* improved performance under all settings. In particular, it significantly improves performance when there is less data.

5.6 Ablation Studies

To observe the effect of source data filtering in the *GENDEX* framework, we conducted ablation studies. As shown in Figure 2, there are several pre-steps for pattern-based filtering. Coreference resolution is a pre-step for semantic role labeling.

Model	SAMSum			TweetSumm			DialogSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
$\text{GENDEX}_{\text{BC}}$	47.83	24.25	40.28	35.55	16.80	31.18	40.78	16.69	34.21
- Patten-based filtering	46.84	23.43	39.61	34.63	15.90	30.45	39.44	15.55	33.28
$\text{GENDEX}_{\text{WIKI}}$	47.79	24.43	40.43	35.91	16.98	31.31	40.98	17.03	34.65
- Patten-based filtering	46.40	23.08	39.19	34.14	16.30	30.01	39.63	15.50	33.23
$\text{GENDEX}_{\text{ROC}}$	47.77	24.45	40.55	34.85	16.68	30.69	40.56	16.58	34.16
- Patten-based filtering	46.30	23.34	39.36	34.21	16.02	30.09	39.37	15.37	33.16

Table 4: Ablation studies on SAMSum, TweetSumm, and DialogSum.

It clarifies the text and helps in contextual analysis by replacing personal pronouns with corresponding names before applying semantic role labeling. Therefore, coreference resolution cannot be eliminated alone. Similarly, semantic role labeling is a pre-step for pattern-based filtering. By analyzing text using semantic role labeling, we obtain a structured representation of the predicate-argument relationships within the text. These results are used for pattern-based filtering. Thus, semantic role labeling cannot be eliminated alone. Finally, we conducted ablation studies to eliminate pattern-based filtering, which entails a chain of elimination of these three consecutive steps.

Table 4 shows the results of the ablation studies. We verified the effect of source data filtering for three OOD datasets by eliminating pattern-based filtering from $\textit{GENDEX}_{BC}$, $\textit{GENDEX}_{WIKI}$, and $\textit{GENDEX}_{ROC}$. We tested the performance on three dialogue summarization datasets: SAMSum, TweetSumm, and DialogSum. As shown in Table 4, eliminating pattern-based filtering degrades the performance. It drops R-1, R-2, and R-L scores by an average of 3%, 6%, and 3%, respectively. Additionally, it drops the performance by up to 9% in the most extreme case. These results imply that our proposed pattern-based filtering method contributes to the performance. We conducted additional experiments to further verify *GENDEX*'s effectiveness. We compared the performance with that of training the model on a combined dataset of the three dialogue summarization datasets, and *GENDEX* showed better performance despite using less data for training (see Appendix F for more details).

6 Conclusion

In this paper, we proposed a generative data augmentation strategy leveraging external data for abstractive dialogue summarization, called *GENDEX*. We filter OOD texts by focusing on the presence of person and interpersonal interactions in the texts. We generate synthetic dialogues from these filtered texts and augment dialogue summarization data. This approach enhances data diversity in terms of semantics and introduces external knowledge by leveraging OOD data and reconstructing it in a dialogue format. The experimental results show that *GENDEX* can improve the performance of dialogue summarization in both quantitative and qualitative aspects, outperforming previous state-of-the-art methods. They also demonstrate the effectiveness of *GENDEX*, highlighting its robust performance, generalizability, and scalability. Furthermore, *GENDEX* improved the performance of dialogue summarization regardless of the complexity of dialogue data.

7 Limitations

One of the limitations of our approach is its dependency on NLP tools and pre-trained language models. These tools are not entirely perfect and may cause error propagation. For example, when using the NER tool, names that overlap with common nouns, pet names that sound like human names, and names that are underrepresented in training data may not be recognized well. Another limitation is the dependency on external data. While external data are much more abundant than dialogue data, careful selection of OOD data that include the presence of people and their interpersonal interactions is needed. Additionally, the experiments were primarily conducted on English dialogue summarization datasets, which may affect the applicability of our proposed method to other languages.

Acknowledgements

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT)

(2021R1A2C2006268) and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) [No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)].

References

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.

Jiaao Chen and Diyi Yang. 2021a. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616.

Jiaao Chen and Diyi Yang. 2021b. 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.

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.

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

Naomi Daniel, Dragomir Radev, and Timothy Allison. 2003. Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 9–16.

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

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

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

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *arXiv preprint arXiv:2012.03502*.

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.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Sachindra Joshi, and David Konopnicki. 2021. Summary grounded conversation generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3748–3756.

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

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020a. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020b. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

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.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Daniil Larionov, Artem Shelmanov, Elena Chistova, and Ivan Smirnov. 2019. Semantic role labeling with pretrained language models for known and unknown predicates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 619–628.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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*, pages 1957–1965.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

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.

Siru Ouyang, Jiaao Chen, Jiawei Han, and Diyi Yang. 2023. Compositional data augmentation for abstractive conversation summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1471–1488.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14.

Dragomir Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification 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 6382–6388.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022b. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

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

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 *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Volume Comparison of Summarization Datasets

Figure 5 illustrates the volume of document and dialogue summarization datasets. Arxiv/PubMed (Cohan et al., 2018) consists of scientific papers, focusing on the abstract generation of scientific articles. CNN/DailyMail (Hermann et al., 2015) is a collection of news articles from CNN and Daily Mail. XSUM (Narayan et al., 2018) is also a news-domain summarization dataset containing BBC articles covering various topics. These are widely-used document summarization datasets. DialogSum (Chen et al., 2021) contains spoken daily dialogues. SAMSum (Gliwa et al., 2019) contains messenger-style dialogue on daily topics. TweetSumm (Feigenblat et al., 2021) contains chat dialogues between agents and customers in customer service. These are popular dialogue summarization datasets. As illustrated in Figure 5, the gap

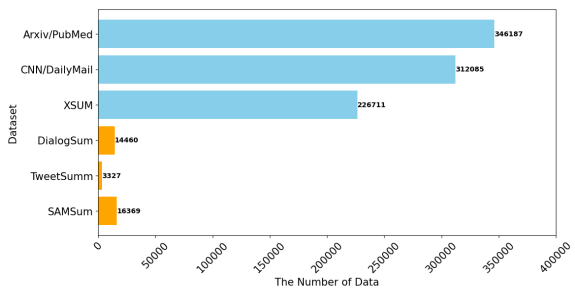


Figure 5: Comparative statistics of the volume of the document (blue) and dialogue (orange) summarization datasets.

in their volume is evident. Document datasets are significantly larger, showing a vast amount of data reaching into the hundreds of thousands. In contrast, dialogue datasets are considerably smaller in size. This visual representation highlights the discrepancy in data availability between document and dialogue summarization datasets, indicating the challenges faced in dialogue summarization due to the limited data resources.

B Dialogue Summary Analysis

To assess the occurrence of people within dialogue summaries, we analyzed the summary content across three distinct dialogue summarization datasets: SAMSum, DialogSum, and TweetSumm. Table 5 represents the number of summaries based on the number of people in each dataset. For instance, the SAMSum dataset comprises 8,403 summaries containing two people. A substantial majority of summaries—98% in SAMSum, 99% in TweetSumm, and 99% in DialogSum—include at least one person. Moreover, the proportions of summaries containing more than two persons are 80% in SAMSum, 98% in TweetSumm, and 94% in DialogSum. Dialogue data features multiple speakers, which is the most significant difference from document data. The content of dialogue data is constituted by the exchange of turns among these speakers. In this context, these statistical results suggest the importance of human names in the construction of dialogue summaries, indicating that their inclusion is vital for effectively capturing the context of the dialogues.

C Examples of SRL Pattern

Using Semantic Role Labeling (SRL), we analyzed the structure of summaries in three dialogue summarization datasets: SAMSum, TweetSumm, and DialogSum. Table 6 presents frequently occurring argument-predicate patterns in these summaries. ARG represents a range of semantic roles associated with the predicate within a text. ARG0 typically represents the agent or speaker, who is the doer or the subject of the action or state. ARG1 usually denotes the direct object or patient, which is the entity affected by the action or state. ARG2 indicates an indirect object, instrument, or beneficiary. ARG3 is used to express a starting point, path, or direction. ARG4 denotes an endpoint or goal, used to express the final position or target of the action. Additionally, there are arguments

#Person / Dataset	0	1	2	3	4	5	6	7	8	9	10	11
SAMSum	375	2949	8403	3225	1018	272	72	28	13	11	1	2
DialogSum	9	798	11547	1743	314	40	7	2				
TweetSumm	17	49	3230	30	1							

Table 5: Statistical result of the number of people in summaries of three dialogue summarization datasets.

Patterns
ARG0 V ARG1
ARG1 V ARG2
ARG0 V ARG2 ARG1
ARG0 V ARG1 ARG2
ARG0 V ARG1 ARGM-TMP
ARG0 ARGM-MOD V ARG1
ARG0 ARGM-NEG V ARG1
ARG0 V ARG2
ARG0 V ARG1 ARGM-PRP
ARG0 ARGM-MOD V ARG1 ARGM-TMP
ARG0 V ARG1 ARGM-LOC
ARG1 V ARG2 ARGM-TMP
ARG0 V ARG1 ARGM-CAU
ARG0 V ARG1 ARGM-ADV
ARG0 V ARG1 ARGM-MNR
ARG1 V ARGM-TMP
ARGM-TMP ARG0 V ARG1
ARG0 ARGM-TMP V ARG1
ARG1 V ARGM-NEG ARG2
ARG1 ARGM-MOD V ARG2
ARG0 ARGM-ADV V ARG1
ARG0 V ARG4
ARG0 ARGM-MOD ARGM-NEG V ARG1
ARG1 ARG0 V
ARG1 V ARG2 ARGM-CAU
V ARG2
ARG0 ARGM-MOD V ARG1 ARG2
ARG0 ARGM-MOD V ARG2 ARG1
ARG1 ARGM-MOD V ARG2 ARGM-TMP

Table 6: Examples of frequent Argument-Predicate patterns appearing in the summaries of SAMSum, TweetSumm, and DialogSum.

playing adjunct roles starting with ARGM-, such as ARGM-TMP (time), ARGM-LOC (location), ARGM-MNR (manner), and ARGM-CAU (cause). ARMG-MOD represents a modal in a text, indicating possibility, permission, obligation, etc. ARGM-NEG indicates negation word, such as not or never. Based on these frequently occurring patterns, we analyzed the positional relationship between argument and predicate when people are included in the argument. In these patterns, arguments containing people appeared significantly before and after, and after the predicate (96.19% in SAMSum, 83.14% in TweetSumm, 96.18% in DialogSum). Based on this observation, we used a pattern-based matching method to filter the text.

D Statistics of Dialogue Summarization Datasets

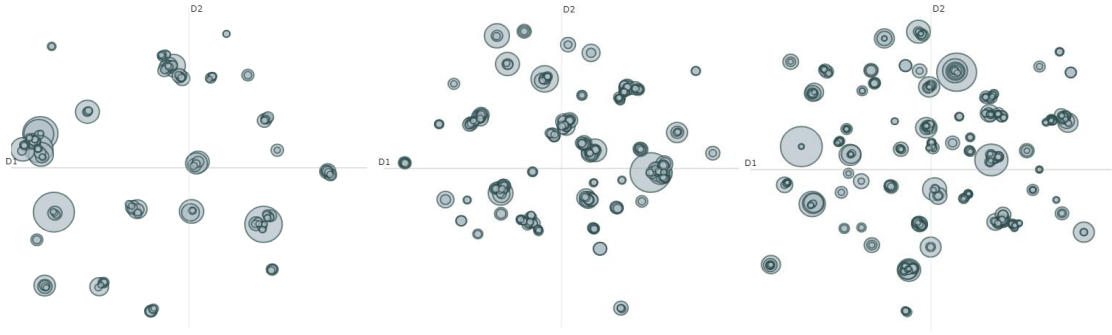
Table 7 shows statistics of three popular dialogue summarization datasets: SAMSum, TweetSumm, and DialogSum. SAMSum (Gliwa et al., 2019) contains over 16,000 chat dialogues with manually annotated summaries. It mainly focuses on creating abstractive summaries from messenger-style conversations covering various daily topics. It consists of 14,732 training samples, 818 validation samples, and 819 test samples. TweetSumm (Feigenblat et al., 2021) includes chat dialogues from a customer service context with manually annotated summaries, featuring conversations between agents and customers. It offers both extractive and abstractive summaries, and we used the abstractive ones. It consists of 2,629 training samples, 356 validation samples, and 342 test samples. DialogSum (Chen et al., 2021) contains spoken daily dialogues with manually annotated summaries and emphasizes real-life scenarios. It consists of 12,460 training samples and 500 validation samples. For test samples, there are two versions of the publicly available DialogSum dataset: one with 500 samples and another with 1,500 samples. The latter version includes slightly varied summaries from three distinct annotators for each dialogue sample, enabling a more strict and comprehensive assessment. We used the version with 1,500 samples.

E Data Diversity

Data diversity is an important factor in machine learning. In this section, we investigate data diversity through topic modeling. Topic modeling can be useful when evaluating the semantic diversity of texts. We used BERTopic (Grootendorst, 2022), a popular text clustering method. We applied topic modeling to the dialogues from SAMSum, the synthetic dataset generated using *GENDEX*, and the

	# Dialogue	# Speakers		# Turns		Dialog len.		Summary len.	
		mean	interval	mean	interval	mean	interval	mean	interval
SAMSum	16,369	$2.40_{\pm 0.83}$	[2, 14]	$11.15_{\pm 6.44}$	[3, 46]	$84.71_{\pm 76.91}$	[5, 4340]	$20.49_{\pm 11.17}$	[1, 64]
TweetSumm	3,327	$2.26_{\pm 0.92}$	[2, 13]	$10.49_{\pm 2.86}$	[8, 25]	$199.58_{\pm 74.63}$	[72, 595]	$35.25_{\pm 12.53}$	[10, 104]
DialogSum	14,460	$2.01_{\pm 0.13}$	[2, 7]	$9.51_{\pm 4.25}$	[2, 65]	$124.00_{\pm 68.97}$	[32, 953]	$22.37_{\pm 10.56}$	[4, 153]

Table 7: Statistics of SAMSum, TweetSumm, and DialogSum.



(a) (b) (c)

Figure 6: Results of topic modeling: (a) SAMSum, (b) synthetic data generated using GENDEX, and (c) combined dataset of these two.

combined dataset of these two. Figure 6 presents the inter-topic distance map for each dataset. Each circle represents a topic, and its size represents the amount of data belonging to that topic. Thus, the number of circles and their distribution reflect the semantic diversity of the data. Topic circles are sparsely distributed in a few regions in SAMSum, whereas they are widely spread out in the synthetic data. SAMSum and the synthetic data resulted in 148 and 289 topics, respectively. The combined dataset of SAMSum and synthetic data generated using *GENDEX* contains 381 topics, which is slightly smaller than the sum of 148 and 289. These results suggest that our method augments the data for topics in SAMSum and also generates synthetic data for topics not present in the original dialogue summarization dataset. Such semantic diversity can be obtained by leveraging a large amount of OOD data. Plain-text data is relatively abundant compared to dialogue data, so it can cover a wide range of topics and synthetic dialogues generated from these texts can introduce external domain knowledge to the model. In this context, OOD data refers not only to data outside the dialogue domain (i.e., non-conversational data) but also to data that belongs to other topics, which the original dataset does not have.

F Comparison with Combined Dataset

We trained the BART model on combined datasets of the three dialogue summarization datasets (SAMSum, TweetSumm, and DialogSum) and tested it on each dataset. Then, we compared the performance with that of *GENDEX*. In Table 8, 'BART trained on (S + T + D)' denotes the BART model trained on the combined dataset. As shown in Table 8, using multiple datasets can improve performance. However, *GENDEX* enhances performance more than training the model on the combined dataset. In addition, a slight performance drop on DialogSum was observed when trained the model on the combined dataset. As described by Zhang et al. (2021), training a model on multiple datasets does not always guarantee performance improvement. This could be attributed to different characteristics, such as dialogue styles and topics. However, *GENDEX* can generate synthetic dialogues close to the original data by training the dialogue generation model to learn their style. Moreover, *GENDEX* shows better performance despite using less data for training. $GENDEX_{BC}$, $GENDEX_{WIKI}$, and $GENDEX_{ROC}$ used 20K, 20K, and 9K data for training, respectively, while 'BART trained on (S + T + D)' used 30K data.

Model \ Dataset	SAMSum			TweetSumm			DialogSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
BART	45.80	22.51	38.71	33.87	16.17	29.96	40.08	15.93	33.47
$\text{GENDEX}_{\text{BC}}$	**47.83**	24.25	40.28	35.55	16.80	31.18	40.78	16.69	34.21
$\text{GENDEX}_{\text{WIKI}}$	47.79	24.43	40.43	**35.91**	**16.98**	**31.31**	**40.98**	**17.03**	**34.65**
$\text{GENDEX}_{\text{ROC}}$	47.77	**24.45**	**40.55**	34.85	16.68	30.69	40.56	16.58	34.16
BART trained on (S + T + D)	46.23	23.02	39.29	34.10	16.20	30.19	39.98	15.92	33.48

Table 8: Comparative result with training on combined dataset. 'BART trained on (S + T + D)' denotes the BART model trained on the combined dataset (i.e., SAMSum + TweetSumm + DialogSum).

G Implementation Details

Named Entity Recognition (NER) We used spaCy toolkit which is an open-source software library for natural language processing. We selected *en_core_web_trf* model provided by spaCy. It is based on the transformer architecture, specifically utilizing the RoBERTa (Liu et al., 2019b) model. *en_core_web_trf* supports a wide range of NLP tasks, including tokenization, part-of-speech tagging, named entity recognition, dependency parsing, and more. It has a good ability especially in understanding the context and semantics of text.

Coreference Resolution For coreference resolution, we used AllenNLP, which is an open-source NLP research library built on PyTorch, developed by the Allen Institute for AI (AI2). We selected the *coref-spanbert-large* model provided by AllenNLP, which specifically utilizes the SpanBERT (Joshi et al., 2020b) model.

Semantic Role Labeling (SRL) We used AllenNLP for SRL. We selected the *structured-prediction-srl-bert* model provided by AllenNLP pre-trained model cards, which is a BERT (Devlin et al., 2019) based model with some modifications.

Dialogue Generation We utilized the DialoGPT-large (Zhang et al., 2020b) model for dialogue generation. For model training, we set the maximum input length to 512, learning rate to 5e-5, weight decay to 0.01, batch size to 3, and epochs to 10. For generation, we set the maximum generation length to 1000. The model was trained using an NVIDIA RTX 3090 GPU.

Dialogue Summarization We primarily used the BART-base (Lewis et al., 2020) model for dialogue summarization. For pre-training on synthetic data, we set the maximum input length to 1024, maximum target length to 256, learning rate to 2e-5, weight decay to 0.01, gradient accumulation steps to 2, and batch size to 4. Based on the two-stage noise-tolerant training setting, we utilized an early stopping-based strategy. For fine-tuning on original dialogue summarization data, all other settings remained the same, and the epoch was set to 3. The model was trained using an NVIDIA RTX 3090 GPU.