Compositional Data Augmentation for Abstractive Conversation Summarization

Anonymous ACL submission

Abstract

Recent abstractive conversation summarization systems generally rely on large-scale annotated summaries. However, collecting conversations and annotating their corresponding summaries can be time-consuming and labor-intensive. To alleviate the data scarcity issue, in this work, we present a simple yet effective compositional data augmentation method, COMPO, for generating diverse and high-quality pairs of conversations and summaries. Specifically, we generate novel conversation and summary pairs through first extracting conversation snippets and summary sentences based on conversation stages and then randomly composing them constrained by the temporal relation and semantic similarities. To deal with the noises in the augmented data, we further utilize knowledge distillation to learn concise representation from a teacher model trained on high-quality data. Extensive experiments on benchmark datasets demonstrate that COMPO significantly outperforms prior state-of-the-art baselines in terms of both quantitative and qualitative evaluation, and exhibits reasonable level of interpretability.

Lucas: Hey! How was your date?
Demi: Hey there!
Demi: It was pretty fine, actually, thank you! } S_1

Demi: I just got promoted! :D
Lucas: Whoa! Great news!
Lucas: Congratulations! } S_2

Lucas: Such a success has to be celebrated.
Demi: I agree! :D
Lucas: Tonight at Death & Co.?
Demi: Sure!
Lucas: When will you be available?
Demi: See you there at 10 pm? } S_3

Lucas: Yeah! Looking forward to it! } S_4

(S_2) Demi got promoted. (S_3) She will celebrate that with Lucas at Death & Co at 10 pm.

Figure 1: A conversation and its paired summary. S_i stand for referred stage snippets, i.e., *Opening*, *Intention*, *Discussion* and *Conclusion*. The corresponding summary consists of two sentences, each sentence corresponds to one snippet S_i (illustrated by color).

1 Introduction

Abstractive conversation summarization, which aims to summarize unstructured conversations into short, concise and structured text, has benefited a lot from neural generative models trained on large-scale annotated data. Much attention has been paid to address various aspects in conversation summarization, such as modeling conversations in a hierarchical way (Zhao et al., 2019; Zhu et al., 2020), leveraging dialogue acts Goo and Chen (2018), using key phrases and entities Liu et al. (2019a); Narayan et al. (2021), utilizing topic segments (Liu et al., 2019b), stage components (Chen and Yang, 2020) and discourse relations (Chen and Yang, 2021b; Feng et al., 2020b). However, training these generative models often requires abundant high-quality data, i.e., conversation and its paired summary, which is usually time-consuming and labor-intensive to obtain. As a result, it is challenging to apply them to new settings or real-world situations where labeled summaries are limited.

A direct solution is to employ data augmentation techniques, which is popular in various areas across computer vision (Cubuk et al., 2018) and natural language processing (Sennrich et al., 2015; Feng et al., 2021a). Existing data augmentation methods can be categorized into token-level (Feng et al., 2020a; Shen et al., 2020), sentence-level (Yu et al., 2018), adversarial style (Miyato et al., 2016; Zeng et al., 2020) and augmentation in the hidden space (Cheng et al., 2020; Jiang et al., 2019). Different from plain context, augmentation for conversations is challenging as we have to take into account conversation structures such as speaker information, topic split, and conversation stages (Gritta et al., 2021; Shuster et al., 2021). Directly applying these augmentation methods into the context of conversations fail to consider any unique structures of conversations and might be limited in creating high-quality and diverse data pairs.

To fill in these gaps, in this work, we propose a simple and effective data augmentation method, COMPO, to improve the performances of abstractive conversation summarization in low-resourced settings by generating augmented data in a *compositional* way, where diverse conversations and summaries are generated from *composing different conversation snippets extracted based on conversation structures*. As a starting point, here we consider conversation stage, a prevalent pattern existing in almost all the context, since conversations often follow certain patterns to develop (e.g. *Opening*, *Intention*, *Discussion* and *Conclusion*) (Chen and Yang, 2020). People tend to summarize the conversation in an almost linear way with a strong temporal dependency (Wu et al., 2021), as illustrated in Figure 1. As a result, it is intuitive to first segment conversation into stages and match these stages with their corresponding summary sentences, and then reorganize them into novel paired conversations and summaries as shown in Figure 2. In this way, sub-components of conversations can be re-organized and re-composed to generate augmented pairs that might not be seen in the original corpus, resulting in more diverse training data.

Specifically, COMPO involves the following steps. Firstly, we construct a pool of candidate pairs for conversation stage and summary sentences as units for composition. Secondly, we sample units from the candidate pool according to some specified requirements to guarantee temporal relations and semantic similarities, and then perform easy-to-use deletion/insertion/replacement operations to both the conversation and summary to construct augmented data based on a given paired data. Theoretically we can generate *infinite* amount of data as we use online sampling during the training process. To alleviate the noise in the augmented data, we first train a teacher model on the original high-quality dataset, and then distill a generative model by mimicking the distribution produced by the teacher model on the augmented data (Hinton et al., 2015). Note that COMPO can be smoothly extended to other conversation-related tasks. To demonstrate the effectiveness of COMPO, we conduct experiments on two benchmark datasets, SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021). Both quantitative and qualitative evaluations show that COMPO surpasses prior state-of-the-art baselines by a large margin.

2 Related Work

2.1 Abstractive Conversation Summarization

Abstractive conversation summarization, as opposed to extraction summarization, requires generative models to have a strong ability in language understanding as the words in output may not appear in the input. Prior work on abstractive conversation summarization can be divided into two categories. One is to directly apply existing document summarization models to conversations (Shang et al., 2018; Gliwa et al., 2019). The other is to design conversation-tailored methods, for instance, modeling conversations in a hierarchical way (Zhao et al., 2019; Zhu et al., 2020). The rich structured information in conversations has also been leveraged. For example, Goo and Chen (2018) used dialogue acts; Liu et al. (2019a); Narayan et al. (2021) leveraged key phrases and entities. Topic segments (Liu et al., 2019b), stage components (Chen and Yang, 2020) and discourse relations (Chen and Yang, 2021b; Feng et al., 2020b) are also explored to understand conversation context for summarization. However, most approaches in the aforementioned categories focus on neural supervised methods and require abundant data to achieve the state-of-the-art performance, which is time-consuming and labor-intensive. In this work, we introduce conversation specific data augmentation methods to help address data scarcity on paired conversation and summaries.

2.2 Data Augmentation in NLP

Data augmentation is an effective approach to boost the performance of neural supervised models, and has been widely applied in various NLP tasks such as text classification (Wei and Zou, 2019; Zheng et al., 2020), machine reading comprehension (Yu et al., 2018), and machine translation (Sennrich et al., 2015). Only a few have made attempts in data augmentation for conversations (Chen and Yang, 2021a). Augmentation for conversations is quite different from traditional classification tasks as it requires models to consider conversation structures and speaker information. Commonly seen practices involve designed word/synonym replacement (Kobayashi, 2018; Niu and Bansal, 2018), word deletion/swapping/insertion (Wei and Zou, 2019), back translation (Sennrich et al., 2015; Xie et al., 2019) and compositional augmentation (Jia and Liang, 2016; Andreas, 2019). Specifically, compositional data augmentation leverages small fragments from the input and re-combine them to create augmented examples. Existing compositional data augmentation often requires carefully-designed rules (Chen et al., 2020b; Nye et al., 2020), and operates at the sentence level (Furrer et al., 2020). Motivated by these, we propose a compositional data augmentation method specific for conversations. Compared with previous work (Chen and Yang, 2021a), we augment conversation data in sub-structure level instead of utterance-level. Also, note that we are the first to augment paired data, i.e., *conversations and its paired summaries* in a compositional way.

3 Methodology

To generate diverse conversation-summary pairs to deal with the data scarcity issue, this section presents a simple and effective compositional data augmentation method COMPO for supervised abstractive conversation summarization.

3.1 Compositional Augmentation

Our compositional augmentation method COMPO operates at the sub-structure level of conversations. By extracting different sub-components of conversations and recombining them based on certain orderings, COMPO can produce novel and diverse conversation and its summaries that might not been seen in the original corpus. To get a reasonable granularity of conversation sub-parts, we choose conversation stages, building upon prior work on conversation structures (Althoff et al., 2016; Chen and Yang, 2020). Dialogues naturally develop following certain stages such as "Openings → Intention → Discussion → Conclusion" in daily chats. Sometimes, the human annotated summaries are also based on different stages in different sentences; sentences within a reference summary usually have very strong, linear temporal dependency (Wu et al., 2021), as shown in Figure 1. Thus we propose a compositional inductive approach through the composing different conversation stages and their corresponding summary sentences (Andreas, 2019).

Specifically, we construct a set of new conversation-summary pairs $\mathcal{D}_a = \{\langle C'_i, S'_i \rangle\}_{i=1}^{M}$ out of the original paired dataset $\mathcal{D}_p = \{\langle C_i, S_i \rangle\}_{i=1}^{N}$, where $M > N$ and C, S denote

Algorithm 1: Constructing Candidate Pairs

Input: A conversation stage $c_i \in C$, a summary S containing n sentences, sliding window size interval [a,b]

Output: Corresponding summary sentences S^i_{paired} for c_i

```
for w = a to b do
    for j = 1 to |S| − w] do
        cand=C_{j,j+w}
        r(j,w) ← ROUGE(cand, s_i)
        W ← W ∪ cand
        j ← j + w/2
    w ← w + 1
j_best, w_best ← argmax_{j,w} r(j,w)
S^i_paired ← C_{j_best,(j_best+w_best)}
```

the conversation and paired summary respectively through compositional augmentations. Our compositional augmentation approach involves two major steps as shown in 2 (a): 1) constructing candidate pairs of summary sentences and conversation snippets and 2) generating augmented conversation-summary samples out of the constructed pairs.

3.1.1 Constructing Candidate Pairs

Following Althoff et al. (2016) and Chen and Yang (2020), we utilize Hidden Markov Model (HMM) to extract stages in conversations. We set the number of hidden stages as 4 (number of conversation stages) and the observations are initialized with representations from sentence-BERT (Reimers and Gurevych, 2019). The segmented conversation is denoted as $C = \{c_1, ..., c_4\}$ where c_i is the stage that contains several consecutive utterances. Then we split the summary into several sentences as $S = \{s_1, ...s_n\}$ where s_i is one sentence in the summary, n is the total number of sentences.

Building on these preprocessed segmented conversation stages and summary sentences, we then match the summary sentences S^i_{paired} to its corresponding conversation stage c_i. Note that this is not a one-to-one matching, a conversation stage can be matched with several consecutive summary sentences. Every conversation stage has its corresponding paired summary snippet. The detailed algorithm is shown in Algorithm 1.

3.1.2 Generating Augmented Pairs

Given the constructed pool of candidate pairs $P = \{\langle c_i, S^i_{paired} \rangle\}$, we then construct augmented data

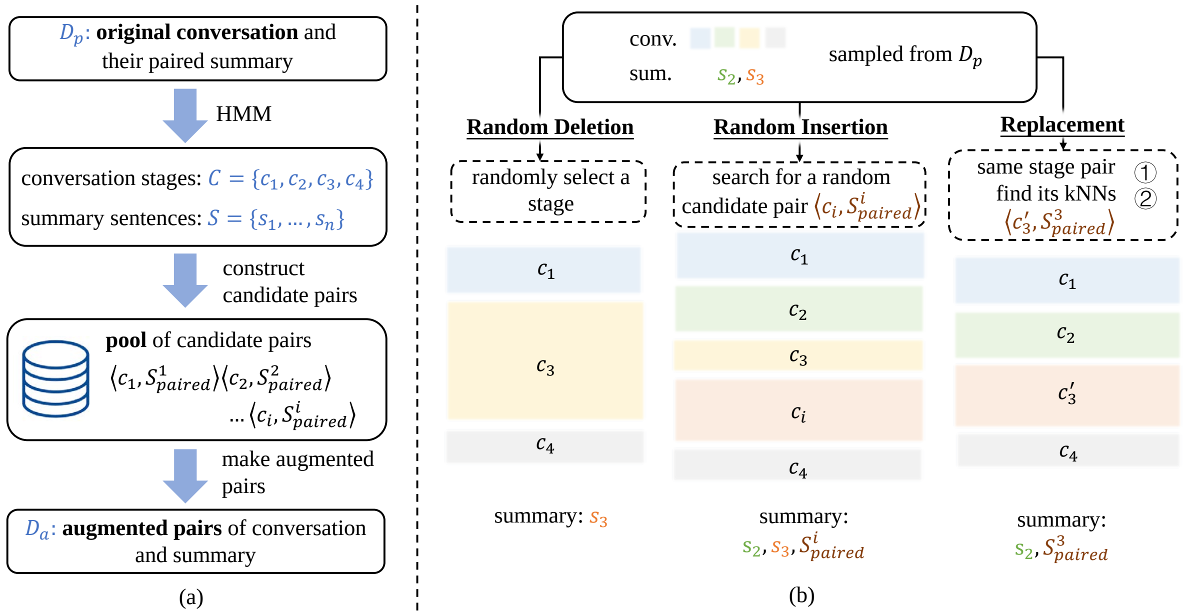

Figure 2: Framework of how we construct augmented pairs (a), and examples of utilizing compositional augmentation strategies to augment the given conversation and its paired summary (b). Given a conversation and its paired summary, we would randomly delete/insert one conversation stage and the corresponding summary sentences, or replace the original conversation stage of the same stage and semantic similarities.

pairs by re-combining the fragments (i.e., candidate pairs). For each sample, we randomly perform the operations described below to generate augmented conversation C_i' and its corresponding summary S_i'. Examples for these operations are shown in Figure 2 (b). Note that we adapt the speakers' names with string matching in all these operations.

Random Deletion and Insertion of Sub-Parts After the construction of candidate pairs, the context of each stage is relatively well summarized in its corresponding summary sentence. To perturb temporal relations to create paired augmented conversations and summaries, we introduce two simple operations: (1) randomly deleting one conversation stage and its corresponding summary sentences to provide less information in the conversation context, and (2) random insertion, which introduces new context by inserting one conversation stage c_i randomly selected from P into a random position of the original four stages. The paired sub-summary is placed in the corresponding position.

Replacement of Sub-Parts Replacement can be seen as a refined version for paraphrasing (Sennrich et al., 2015) in compositional conversation augmentation. In order to preserve the conversation structure of the augmented data, we substitute the same conversation stage, e.g., we only substitute the *Opening* stage by another *Opening* pair sampled from the pool. To guarantee similar semantic meanings to avoid noise as much as possible, we select the candidate pair with k-nearest neighbors (kNNs). The motivation here is that kNNs may contain the same entity words as the original sentences and words, but in different contexts and forms (Chen et al., 2020a). In practice, we map the summary sentences for all the candidate pairs of the same stage (pre-specified) into a hidden space, and then collect each sentence's kNNs using l^2 distance. We fetch the candidate pair that has the nearest summary sentences as the substitute.

When creating augmented conversation and summary pairs, we conduct a *online sampling* approach, which means that we can generate an infinite amount of labeled data theoretically.

3.2 Model Distillation

A straight-forward way to improve a generative model with the augmented data is to directly merge the original data. However, this naive approach may lead to sub-optimal performance as it may bring much noise. Therefore, we apply model distillation in the training process to learn more concise representation with clean signals.

For a generative model, it captures the distribution of a summary sequence S given the conversation context C, i.e., $\mathcal{P}_\theta(S|C)$. This can be formalized as follows:

Dataset	Split	*Number of Participants*			*Number of Turns*			*Reference Length*		
		Mean	Std	Interval	Mean	Std	Interval	Mean	Std	Interval
SAMSum	Train 14732	2.40	0.83	[1,14]	11.17	6.45	[1,46]	23.44	12.72	[2,73]
	Dev 818	2.39	0.84	[2,12]	10.83	6.37	[3,30]	23.42	12.71	[4,68]
	Test 819	2.36	0.83	[2,11]	11.25	6.35	[3,30]	23.12	12.20	[4,71]
DialogSum	Train 12460	2.01	0.13	[2,7]	9.49	4.16	[2,65]	22.87	10.71	[5,153]
	Dev 500	2.01	0.13	[2,4]	9.38	3.99	[2,29]	20.91	9.76	[6,56]
	Test 500	2.01	0.27	[2,3]	9.71	4.99	[2,65]	19.09	9.20	[6,84]

Table 1: Statistics of the used datasets. *Interval* denotes the minimum and maximum range.

$$\mathcal{P}_\theta(S|C) = \Pi_{i=1}^{|S|} P_\theta(s_i|s_{<i}, C), \tag{1}$$

where $|S|$ is the length of S, $s_{<i} = s_1...s_{i-1}$ is the token sequence before s_i. The model parameters θ can be learned by optimizing the NLL loss:

$$\mathcal{L}_{nll}(\theta) = -\sum_{i=1}^{|S|} logP_\theta(s_i|s_{<i}, C) \tag{2}$$

In this work, we parameterize the summary generation model using the Transformer based encoder-decoder framework (Vaswani et al., 2017). To perform model distillation, we first train a teacher model $P_{\theta_t}(S|C)$ by optimizing the NLL loss on the original dataset. After the training process is completed, the teacher model is then fixed and used to compute a knowledge distillation (KD) (Kim and Rush, 2016) loss as:

$$\mathcal{L}_{kd}(\theta) = -\sum_{i=1}^{|S|}\sum_{j=1}^{|\mathcal{V}|} P_{\theta_t}(s_i = j|s_{<i}, C) \times logP_\theta(s_i = j|s_{<i}, C), \tag{3}$$

where $|\mathcal{V}|$ denotes the size of the vocabulary and θ_t is the parameter of the teacher model. The final training objective of the summarization model is:

$$\mathcal{L}_G(\theta) = \mathcal{L}_{nll}(\theta) + \alpha\mathcal{L}_{kd}(\theta), \tag{4}$$

Here, $\mathcal{L}_G(\theta)$ is evaluated on the augmented dataset. α is the weight used to balance these two losses.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our proposed framework, we conduct experiments on two benchmarks of conversation summarization: SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021). More detailed data statistics are shown in Table 1.

SAMSum contains open-domain daily-chat conversations in English written by linguists, each of which is annotated with summary by language experts. The topics contain arranging meetings, planning travels, chit-chat and so on. There are 14,732 dialogue-summary pairs for training, 818 and 819 instances for validation and test, respectively.

DialogSum is a large-scale dataset for real-life scenario conversations, and contains diverse task-oriented conversations. Specifically, speakers in DialogSum are denoted with $\#Person\_1\#$ and $\#Person\_2\#$. The public dataset consists of 12,460 training samples. The validation and test set have equal instances of 500.

4.2 Evaluation Metrics and Baselines

Evaluation Metrics We use the standard ROUGE metric (Lin, 2004) as automatic evaluation metrics, including ROUGE-1, ROUGE-2, and ROUGE-L. For SAMSum, following previous work (Gliwa et al., 2019), we use pyROUGE [1] library with stemming. For DialogSum, we use py-rouge[2] following Chen et al. (2021). Note that the ROUGE scores might vary with different tookits.

Baselines in literature On SAMSum dataset, we select the baseline models reported in (Gliwa et al., 2019): **Longest-3** is a commonly-used extractive summarization baseline which takes the top three longest sentences as summary. The **pointer generator** (See et al., 2017) is RNN-based with copy-attention mechanism or policy gradient. The **Transformer** (Vaswani et al., 2017) is a random-initialized self-attention architecture with multi-head attention. **D-HGN** (Feng et al., 2021b) incorporated commonsense knowledge from ConceptNet for conversation summarization. **UniLMv2**

[1] pypi.org/project/pyROUGE/.
[2] pypi.org/project/py-rouge/

(Bao et al., 2020) is used which is a pretrained language model for autoencoding and partially autoregressive language modeling. **BART** (Lewis et al., 2020) is trained by corrupting text with anarbitrary noising function and learning to reconstruct the original text. On DIlogSum dataset, we compare our model with baselines in (Chen et al., 2021).

Baselines with different augmentation strategy To demonstrate the superiority of our proposed compositional augmentation over traditional data augmentation methods, we conduct experiments on SAMSum with different representative data augmentation methods at different granularity including token-level, sentence-level and context-level: (1) *Synonym Replacement (SR)* (Kobayashi, 2018; Kumar et al., 2020) is a token-level approach, which keeps the semantic meaning unaffected by replacing a random word in the conversation with its synonyms. (2) *Back Translation (BT)* (Xie et al., 2019) is a utterance-level method, which firstly translates an selected utterance in the conversation into an intermediate language, and then translates it back to the original language. (3) *Utterance Swapping (US)* is a context-level manner, which randomly selects two utterances in the conversation at first, and then swaps them, leaving the total information unchanged.

4.3 Implementation Details

During training process, the encoder an decoder share the same set of parameters, which are initialized using a pre-trained BART (Lewis et al., 2020). The teacher model uses the same architecture and it is fine-tuned using the original paired dataset $\mathcal{D}_p$ for 8 epochs on the NLL loss (Eq. 2). The final generative conversation summarization model is firstly initialized using the pre-trained BART weights and fine-tuned using the loss in Eq. 4 for another 8 epochs on $\mathcal{D}_p \cup \mathcal{D}_a$ with learning rate set to 3e-5 and a total 16 of batchsize. The value of α in Eq. 4 is set to 1. We generate 1 augmented pair per data sample. It takes around 2 hours to train on a single NVIDIA TITAN RTX 2080Ti GPU.

4.4 Results

Table 2 and Table 3 show the results on SAMSum and DialogSum[3] benchmark datasets. We observe that, (1) Our proposed method obtains substantial gains over the competitive baselines on both the datasets, notably 50.66 for ROUGE-L score on SAMSum test set, which demonstrates the effectiveness of COMPO. (2) Compared with other augmentation methods, our proposed compositoinal augmentation technique works significantly better. This further demonstrates that data generated by COMPO could provide more diverse and effective information used for summarization. (3) Training the generative models on the merged data $\mathcal{D}_p \cup \mathcal{D}_a$ without distillation (i.e., w/KD) brings little or no performance improvements compared to directly training on $\mathcal{D}_p$ (i.e., BART). This verifies the effectiveness of distillation to get rid of noise in the augmented data. (4) With BART$_{base}$ as the pretraining model, our method even outperforms the performance of BART$_{large}$ baseline on SAMSum, indicating that the proposed method is effective in conversation summarization. (5) Our model also performs well on DialogSum, which is a more abstractive, open-domain and spoken analogous (Chen et al., 2021). We can infer that COMPO has great summarization ability as it comes to more challenging tasks.

Model	R-1	R-2	R-L
In literature			
Longest-3*	32.46	10.27	29.92
Pointer Generator*	37.27	14.42	24.26
Transformer*	42.37	18.44	39.27
D-HGN	42.03	18.07	39.56
UniLM*	47.85	24.23	46.67
BART$_{base}$	51.74	26.46	48.72
BART$_{large}$†	53.12	27.95	49.15
SR + KD	51.94	26.69	49.21
BT + KD	52.14	26.83	49.43
UR + KD	52.18	26.91	49.50
COMPO$_{base}$	53.32	27.78	50.66
w/o KD	51.79	26.54	48.70
COMPO$_{large}$	**54.03**	**28.42**	**50.87**
w/o KD	53.21	27.89	49.23

Table 2: Results on SAMSum test. * and † indicate that the results are taken from Gliwa et al. (2019) and Chen et al. (2021) respectively. COMPO$_{base}$ and LARGE$_{large}$ denotes COMPO with BART$_{base}$ and BART$_{large}$.

4.5 Human Evaluation

We conduct human annotations to evaluate the quality of augmented data and summaries generated by our proposed COMPO. Each generated sample is

[3] Since there are three reference summaries on DialogSum test set, the results here are the average of three scores.

Model	R-1	R-2	R-L
In literature			
Transformer*	35.91	8.74	33.50
BART$_{base}$	45.86	19.75	44.33
UniLMv2*	47.04	21.13	45.04
BART$_{large}$ *	47.28	21.18	44.83
SR + KD	45.81	19.84	44.39
BT + KD	46.32	20.03	44.57
UR + KD	46.22	20.26	44.53
COMPO$_{base}$	47.19	20.85	44.91
w/o KD	45.95	19.84	44.30
COMPO$_{large}$	**48.02**	**21.96**	**45.63**
w/o KD	47.26	21.23	44.87

Table 3: Results on DialogSum test split. * indicates that the results are taken from Chen et al. (2021)

Model	Fac.	Suc.	Inf.
Ground Truth	**4.01**	4.15	**3.97**
BART	3.69	3.95	3.71
COMPO	3.92	**4.23**	3.88

Table 4: Human evaluation for the quality of generated summaries in terms of **Fac**tualness, **Suc**cinctness, and **Inf**ormativeness.

Model	R-1	R-2	R-L
COMPO	53.32	27.78	50.66
w/o insertion	53.12	27.46	50.24
w/o deletion	52.83	27.36	49.65
w/o replacement	52.44	27.13	49.05
w/o kNNs	52.71	27.13	49.96

Table 5: Ablation results for different strategies and semantic similarity when making augmented pairs on the test set of SAMSum dataset.

annotated by three workers with English major and linguistic background. The inter-rater agreement among annotators is measured using the Fleiss's kappa $\mathcal{K}$ (Randolph, 2005).

Quality of Augmented data $\mathcal{D}_a$ We ask the annotators to rate a set of randomly sampled 50 pairs from $\mathcal{D}_a$ in terms of 1) *Fluency*: whether the augmented pairs are fluent; 2) *Coherency*: whether the summary is coherent with the conversation so that they make a plausible pair. Each metric is scored with scale 0 (worst) to 2 (best). The *Fluency* score for $\mathcal{D}_p$ and $\mathcal{D}_a$ is 1.82 and 1.78 with $\mathcal{K} = 0.61$ (substantial agreement), while the *Coherency* score is 1.59 and 1.51 with $\mathcal{K} = 0.43$ (moderate agreement). This indicates that the generated data is plausible. Some of the generated conversation examples and their summaries can be found in Appendix A.

Quality of Generated Summaries For summaries evaluation, we ask the annotators to rate a set of randomly sampled 100 generated summaries from ground-truth, BART and COMPO in terms of 1) *Factualness*: whether the generated summary is actual or based on fact; 2) *Succinctness*: whether the summary contain redundant information; 3) *Informativeness*: whether the generated summary contains the most important information. Each metric is scored with scale 1 (worst) to 5 (best). The $\mathcal{K}$ value for *factualness*, *succinctness* and *informativeness* is 0.46, 0.58, and 0.52 respectively, indicating moderate agreement (Koo and Li, 2016). As shown in Table 4, COMPO can generate significantly better summaries with respect to factualness, succinctness, and informativeness than baseline model. This might because that the incorporation of compositional augmented data enables the model to be better aware of the relations between summary sentences and its corresponding conversation snippets, thus improving the factualness over baseline. Also, the model is trained with more diverse data, requiring it to focus on the most salient parts in conversations, which further improves the succinctness and informativenes.

5 Ablation Studies

5.1 Different Augmentation Strategies

To investigate the different strategies used in compositional augmentation, we conduct an ablation study to explore the effect of random deletion, insertion and replacement mentioned in Section 3.1.2. We also provide the experiment results removing kNNs to see the effect of semantic similarity, i.e., we randomly select a pair to replace for the original one. The results are given in Table 5.

We can see that all the three strategies contribute to the performance, as removing any one of them causes a performance drop on ROUGE scores. Especially, the metrics drop by a great margin as we remove the replacement strategy, which shows that the replacement strategy is crucial in generating diverse and effective data pairs. In addition, when we randomly replace pairs of conversation stage and

Model	1-gram	2-gram	3-gram	4-gram
human ($\mathcal{D}_p$)	0.195	14.53	57.92	79.81
SR	0.208	13.91	58.09	78.13
BT	0.191	13.56	57.19	75.58
COMPO ($\mathcal{D}_a$)	0.229	15.71	60.21	80.13

Table 6: Experiment result for the quality of augmented pairs in terms of Distinct n-grams. Since Utterance Swapping has identical statistics as $\mathcal{D}_p$, we left it out.

Model	R-1	R-2	R-L
COMPO	53.32	27.78	50.66
w/k-consecutive	51.49	26.77	48.76
w/extractive	52.44	26.51	49.02

Table 7: Results on the SAMSum test set, when we apply different methods on conversation segmentation in constructing candidate pairs.

summary sentences, the performance drops. The could be the introduction of irrelevant topic and context, which may bring noise for summarization.

5.2 Diversity of Augmented Data

Inspired by Zhang et al. (2020), we also evaluate the diversity of augmented pairs for conversation and summary with automatic metric *Distinct* (Li et al., 2015), which measures the proportion of unique n-grams in the augmented dialogue pairs ($n = 1, 2, 3, 4$). A higher score denotes that the data sample is more diverse. As shown in Table 6, our augmented data pairs are more diverse compared with $\mathcal{D}_p$, consistent across distinct n-grams.

5.3 Effect of Different Strategies When Constructing Candidate Pairs

There are many other ways of segment the conversation and match the components with summary sentences. One way is to directly search for k consecutive utterances in the conversation for each summary sentence. Other line of work uses the extractive approach (Wu et al., 2021). Suppose we have summary S with $|S|$ sentences within, and conversation C. We divide the conversation into $|S|$ parts, each corresponding to one summary sentence. The difference between these two methods is that the former allows overlap between the separated conversation snippets. Experiment results for the aforementioned methods are shown in Table 7.

We notice that segmenting with conversation stages and then matching with summary sentences led to the best performance. This is intuitive as conversation stage contains the information of potential conversation patterns and temporal information compared to other methods. Directly searching for best k consecutive utterances for each summary sentence almost has no improvement over BART, even degraded a bit. This sheds light on how to carefully deal with information overlap when constructing candidate pairs.

Model	R-1	R-2	R-L
COMPO	53.32	27.78	50.66
w/ jointly-train	51.79	26.54	48.70
w/ two-stage	52.91	27.15	50.10

Table 8: Results when using different strategies combining $\mathcal{D}_p$ and $\mathcal{D}_a$ for training on the SAMSum test set.

5.4 Strategies for Combining $\mathcal{D}_p$ and $\mathcal{D}_a$

Except the knowledge distillation discussed in Section 3.2 in the training process, we also experiment with another two strategies combining $\mathcal{D}_p$ and $\mathcal{D}_a$. (1) merge $\mathcal{D}_p$ and $\mathcal{D}_a$ directly and train models on them *jointly* (Edunov et al., 2018). (2) the *two-stage* method, which firstly fine-tune the pre-trained BART model on the augmented Data $\mathcal{D}_a$ and then fine-tune on $\mathcal{D}_p$ with the NLL loss. As shown in Table 8, when model-level knowledge distillation is employed, the performance is significantly better than using the other two strategies.

6 Conclusion

In this paper, we introduced a simple and effective compositional data augmentation method for conversation summarization, which is composed of the following processes, i.e., 1) constructing candidate pairs of conversation snippet and summary sentence based on conversation stages and 2) organizing the candidate pairs into newly augmented data with various operations. There is also a model distillation process to get rid of the noise introduced by the augmented data. Extensive experiments on benchmark datasets demonstrate that COMPO significantly outperforms prior state-of-the-art baselines in terms of both quantitative and qualitative evaluation, through generating compositional and diverse augmented data. Our method has key implications for designing augmentation techniques for low-resource dialogue related tasks.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Jacob Andreas. 2019. Good-enough compositional data augmentation. *arXiv preprint arXiv:1904.09545*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020a. Local additivity based data augmentation for semi-supervised ner. *arXiv preprint arXiv:2010.01677*.

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

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021b. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. 2020b. Compositional generalization via neural-symbolic stack machines. *arXiv preprint arXiv:2008.06662*.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. *arXiv preprint arXiv:2006.11834*.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020a. Genaug: Data augmentation for finetuning text generators. *arXiv preprint arXiv:2010.01794*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021a. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021b. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *China National Conference on Chinese Computational Linguistics*, pages 127–142. Springer.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020b. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. *arXiv preprint arXiv:2012.03502*.

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. 2020. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures. *arXiv preprint arXiv:2007.08970*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2021. Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management. *Transactions of the Association for Computational Linguistics*, 9:36–52.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. *arXiv preprint arXiv:1606.03622*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

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*.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

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.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization. *arXiv preprint arXiv:2104.07606*.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. *arXiv preprint arXiv:1809.02079*.

Maxwell I Nye, Armando Solar-Lezama, Joshua B Tenenbaum, and Brenden M Lake. 2020. Learning compositional rules via neural program synthesis. *arXiv preprint arXiv:2003.05562*.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Jean-Pierre Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. *arXiv preprint arXiv:1805.05271*.

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

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2020. Openattack: An open-source textual adversarial attack toolkit. *arXiv preprint arXiv:2009.09191*.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiaoxi Mao, Yadong Xi, and Minlie Huang. 2020. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. *arXiv preprint arXiv:2009.09427*.

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*, pages 3455–3461.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. *arXiv preprint arXiv:2004.02016*.

A Sampled Data from $\mathcal{D}_a$

In this section, we display several augmented data pairs sampled from $\mathcal{D}_a$ generated with different strategies as shown in Figure 3.

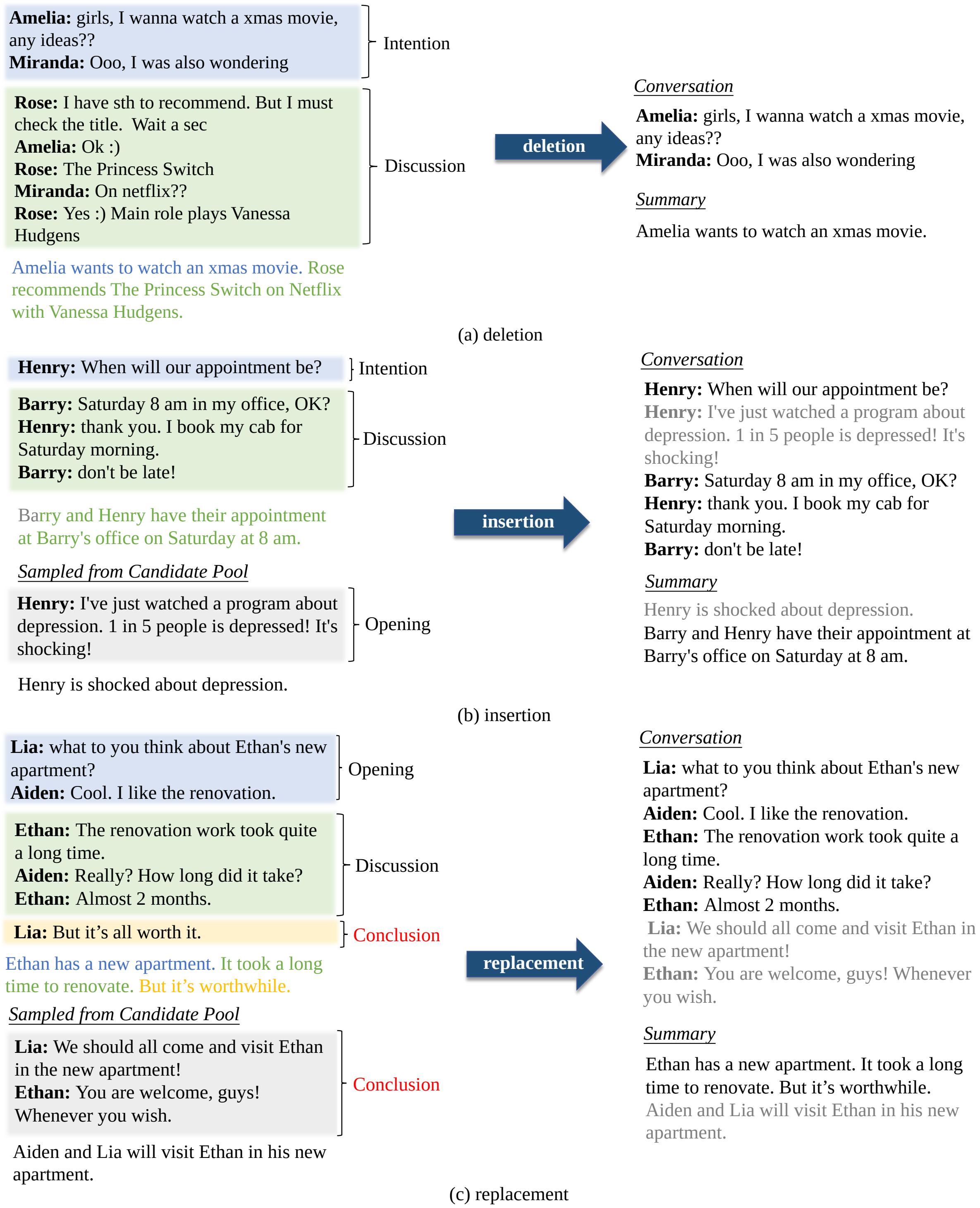

Figure 3: Data pairs sampled from $\mathcal{D}_a$ generated with different strategies. Words in grey indicate the newly introduced sub-parts.