DIONYSUS: A Pre-trained Model for Low-Resource Dialogue Summarization

Yu Li[*†], **Baolin Peng**[‡], **Pengcheng He**[‡], **Michel Galley**[‡], **Zhou Yu**[†], **Jianfeng Gao**[‡]
[†]Columbia University, New York, NY
[‡]Microsoft Research, Redmond, WA
{yl5016, zy2461}@columbia.edu
{bapeng,penhe,mgalley,jfgao}@microsoft.com

Abstract

Dialogue summarization has recently garnered significant attention due to its wide range of applications. However, existing methods for summarizing dialogues have limitations because they do not take into account the inherent structure of dialogue and rely heavily on labeled data, which can lead to poor performance in new domains. In this work, we propose DIONYSUS (dynamic input optimization in pre-training for dialogue summarization), a pre-trained encoder-decoder model for summarizing dialogues in any new domain. To pre-train DIONYSUS, we create two pseudo summaries for each dialogue example: one from a fine-tuned summarization model and the other from important dialogue turns. We then choose one of these pseudo summaries based on information distribution differences in different types of dialogues. This selected pseudo summary serves as the objective for pre-training DIONYSUS using a self-supervised approach on a large dialogue corpus. Our experiments show that DIONYSUS outperforms existing methods on six datasets, as demonstrated by its ROUGE scores in zero-shot and few-shot settings.

1 Introduction

Text summarization aims to produce concise and accurate summaries of long texts. Recent research on pre-trained neural language models has shown success in summarizing monologues (Lewis et al., 2020; Raffel et al., 2022; Zhang et al., 2019; He et al., 2022), such as news articles (Lee et al., 2022; Ravaut et al., 2022) and scientific publications (Ibrahim Altmami and El Bachir Menai, 2022; Dong et al., 2021). However, dialogue summarization presents additional challenges due to the different information distribution in dialogues.

Self-supervised text summarization models (Zhang et al., 2019; Wan and Bansal, 2022; Phang et al., 2022) are typically pre-trained on free-form text data, with selected sentences as the pre-training objective. While this approach can be effective for monologues such as news articles, it is less successful at summarizing semistructured and multi-participant dialogues. As illustrated in Figure 1, in daily chats, dialogue information is often dispersed across various dialogue turns, making it difficult to extract all relevant information through a few selected turns. While a golden summary needs to accurately captures vital information throughout the entire conversation. Furthermore, real-world dialogue-summarization applications often have limited or even no labeled data, making it challenging to develop effective models. Therefore, it is crucial to develop dialogue summarization models that can perform well in zero-shot and few-shot

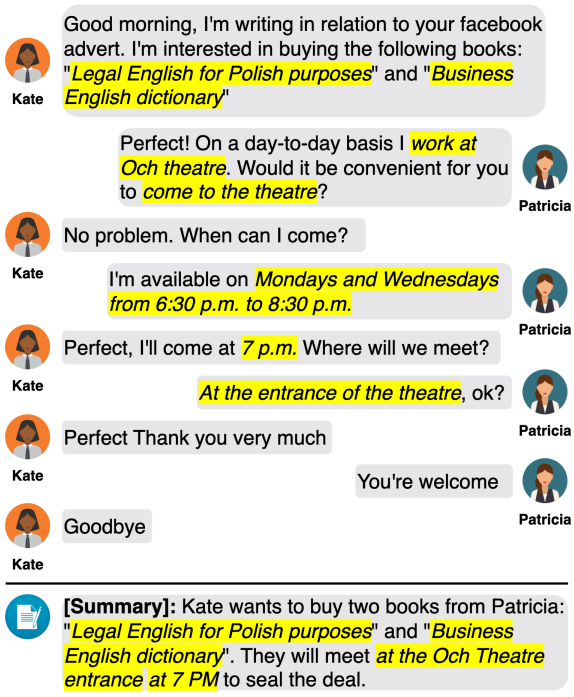

Figure 1: A summary of a dialogue in the SAMSum dataset, where the golden summary effectively compiles relevant information (in yellow) from the entire conversation.

[*]Work was done when Yu Li was interning at MSR

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1368–1386

settings for their practical use.

To address these challenges, we propose DIONYSUS, a pre-trained sequence-to-sequence model designed to summarize dialogues in any domain, even with a lack of labeled data. It uses pseudo summaries as its pre-training objective, which can be dynamically selected from two sources.

First, for daily chats where multiple dialogue turns are not sufficient to summarize the dialogue, we train a summary helper using high-quality dialogue summarization datasets to generate pseudo summaries for these types of dialogues. On the other hand, for dialogues like meeting minutes, interviews, and debates, which can be summarized through a selection of essential turns, we use a method inspired by the gap sentence generation (GSG) technique in PEGASUS to select these turns as pseudo summaries for training. For instance, choosing the final few turns in a conversation can effectively summarize meeting minutes. We have improved upon the GSG method by using the generated summaries from the summary helper as references during gap sentence selection, as they tend to have less noise compared to the full dialogue context. We refer to this source of pseudo summaries as "Principal" and refer to our improved method as GSG+. We find that our improved method outperforms previous methods in low-resource settings across different domains, such as daily chats, emails, and customer service dialogues. Additionally, we study different objective strategies for selecting the pseudo summary as a pre-training objective from the generated summary and the "Principal."

We evaluate DIONYSUS on six dialogue summarization datasets. Our best model trained on 19 dialogue corpora surpasses $\text{PEGASUS}_{\text{LARGE}}$ in a zero-shot setting across all domains. We also found that the best performance is achieved by selecting the source with the highest ROUGE score as the objective strategy. Our main contributions are:

- The development of DIONYSUS, a pre-trained sequence-to-sequence model for summarizing dialogues in any domain in a zero-shot or few-shot setting.
- The introduction of new self-supervised pre-training objectives for dialogue summarization using a summary helper and GSG+.
- The demonstration that DIONYSUS outperforms baselines on six domains in low-resource settings, and can be fine-tuned with only 10 training examples to outperform vanilla T5 (Raffel et al., 2022) fine-tuning with $1,000$ examples.

2 Approach

Figure 2 outlines the steps for constructing DIONYSUS: § 2.1 First, a summary helper is constructed using two high-quality dialogue summarization datasets. This helper generates a pseudo summary for each dialogue in our pre-training corpus. § 2.2 Next, the "Principal" is extracted using GSG+ as the other pseudo summary for the dialogue. § 2.3 Finally, various strategies are employed to select the best pseudo summaries from the first and second steps to serve as the objective for pre-training.

2.1 Summary Helper

In certain types of dialogue, such as daily chats, it can be challenging to gather all necessary information from just a few dialogue turns due to the dispersed nature of dialogue information. To address this problem, we have created a summary helper model that generates pseudo summaries for each training example in our pre-training corpus.

We build our summary helper upon the T5 (Raffel et al., 2022) model. To capture essential information in a dialogue, we have trained our helper on the MultiWoz dataset (Budzianowski et al., 2018; Eric et al., 2020) in DS2 (Shin et al., 2022), which contains summaries derived from dialogue states using templates. This allows us to capture essential information from each turn in the conversation. Additionally, we have continued training our helper on the DialogSum (Chen et al., 2021) dataset, a human-annotated dataset in the daily life domain. This allows us to overcome the fixed format of summaries introduced by templates in DS2 and produce more natural pseudo summaries.

2.2 Gap Sentence Generation Plus (GSG+)

Algorithm 1 GSG+

$P \leftarrow \emptyset$
for $j \leftarrow 1$ to m **do**
  $s_i := rouge(P \cup \{x_i\}, G), \forall i \; s.t. \; x_i \notin P$
  $k := \text{argmax}\{s_i\}_n$
  $P := P \cup \{x_k\}$
end for

Dialogues in certain settings, such as meetings and medical dialogues, often include summary

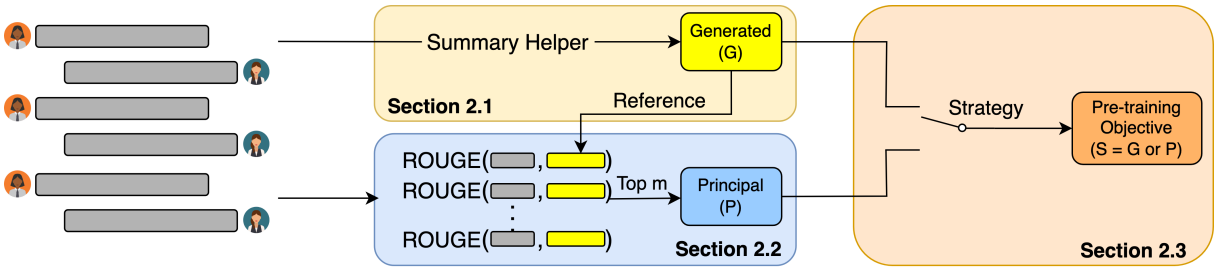

Figure 2: A diagram of pre-training in DIONYSUS: The summary helper (§ 2.1) generates a pseudo-summary (G) to select dialogue turns (§ 2.2) as the "Principal" (P) and using various strategies (§ 2.3) to choose between the generated summary and the principal as the pre-training objective.

turns that summarize the entire conversation. For example, a participant may summarize a meeting, or a doctor may explain the outcome. These summary turns can be used as a pre-training objective because they highlight the main points of the dialogue and provide a concise overview of the topic discussed. In order to make DIONYSUS more adaptable to these scenarios, we have improved the independent principal method in the GSG method (Zhang et al., 2019) by using it to select essential summary turns as pseudo summaries for training. Our new method, called Gap Sentence Selection Plus (GSG+), uses the ROUGE1-F1 score between each dialogue turn x_i and the generated summary G from the helper in Section 2.1 rather than the remaining text $D \setminus x_i$ to determine the importance of each turn. The generated summary eliminates much of the extraneous information from the dialogue and thus tends to have less noise than the full dialogue context, resulting in a less cluttered summary. This enables us to select the top-m-scored summary turns as the "Principal," which will provide a more comprehensive overview of the vital information in the dialogue. For instance, Using the summary helper to identify key points increases the likelihood of selecting the most important dialogue turns as the "Principal" summary when creating pseudo summaries for meeting minutes instead of randomly selecting dialogue turns.

Specifically, given a dialogue $D = \{x_i\}_n$, we use Algorithm 1 to obtain the pseudo-summary "Principal" P. The input for our training example is the remainder of the dialogue $D \setminus P$. In Appendix C, we explore the impact of the dialogue turns order on the formation of the "Principal". Using GSG+ can effectively identify essential summary turns and generate more accurate pseudo-summaries than with the original GSG method.

Algorithm 2 Better ROUGE

$S \leftarrow \emptyset$
$s_g := rouge(G, D \setminus \{P\})$
$s_p := rouge(P, D \setminus \{P\})$
if $s_g > s_p$ **then**
    $S := G$
else
    $S := P$
end if

2.3 Pre-training Objectives Strategy

To generate the final pseudo summary S for each specific dialogue training example, we consider three strategies. These strategies are based on the generated pseudo summary G and the extracted "Principal" P. These strategies serve as the pre-train objective for the dialogue training example.

All G $S = G$: We always select the generated summary from the summary helper as the pre-training objective.

All P $S = P$: We always select the "Principal" as the pre-training objective.

Better ROUGE We use either G or P based on the recall of information from the dialogue to determine the pre-training objective. We utilize Algorithm 2 to get the pre-training objective by calculating the ROUGE1-F1 score for the pseudo summaries and the dialogue, excluding the "Principal" $D \setminus P$. It is important to note that we use the same reference to ensure a fair comparison.

For pre-training with above strategies, if we choose G as the pseudo summary, we input the full dialogue. If we choose P, we input the dialogue, excluding the "Principal," $D \setminus P$ to create an abstract summary. However, we also include the "Principal" with a probability, using a copying mechanism to create an extractive summary.

More information about this copy mechanism can be found in Section 5.4. It is important to note that we do not combine these two pseudo summaries for a single training example. Each example in our pre-training corpus will have either G or P as its designated pseudo summary.

3 Training Corpus

To train DIONYSUS, we utilized 19 conversational corpora that do not come with pre-defined dialogue summaries. We employed a self-supervised approach by using pseudo-summaries as the pre-training objective.

Conversational Corpora We collect 19 available conversational corpora consisting of 1.7M examples after truncating for pre-training. Corpus information is listed in Table 1. We access these corpora through ConvoKit v2.5.3[1]. This helps us to ensure that DIONYSUS is well-equipped to handle a variety of conversational scenarios.

Corpora	# Dialogues
CaSiNo (Chawla et al., 2021)	1,030
Chromium (Meyers et al., 2018)	163,675
Gone Awry (CMV) (Zhang et al., 2018)	6,842
Gone Awry (Wiki) (Zhang et al., 2018)	4,188
Diplomacy (Peskov et al., 2020)	246
Friends (Zhou and Choi, 2018)	1,301
GAP (Braley and Murray, 2018)	28
IQ2 (Zhang et al., 2016)	108
Cornell Movie Dialogs[2]	83,097
Parliament (Zhang et al., 2017b)	216,894
PERSUASIONFORGOOD[3]	1,017
Reddit Coarse (Zhang et al., 2017a)	9,483
Reddit Corpus (small) [4]	8,286
Supreme Court [5]	7,700
Switchboard (Stolcke et al., 2000)	1,155
Tennis (Fu et al., 2016)	81,974
Wiki Deletion (Mayfield and Black, 2019)	383,918
Wiki Talk Pages[6]	125,292
Winning Arguments (Tan et al., 2016)	3,051

Table 1: Corpora we use to pre-train DIONYSUS.

We train our objective summary helper with a rule-based dialogue summarization dataset (DS2) and an abstractive summarization dataset (DialogSum).

[1]https://convokit.cornell.edu/

[2]Cornell Movie Dialogs Corpus is from Danescu-Niculescu-Mizil and Lee (2011)

[3]PERSUASIONFORGOOD is from Wang et al. (2019)

[4]https://convokit.cornell.edu/documentation/reddit-small.html

[5]https://convokit.cornell.edu/documentation/supreme.html

[6]Wikipedia Talk Pages is from Danescu-Niculescu-Mizil et al. (2012)

DS2 This dataset (Shin et al., 2022) creates dialogue summaries for the MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020) dataset by heuristic rules from the dialogue states. It includes 5 domains and $10,000$ dialogues.

DialogSum This dataset (Chen et al., 2021) collects human annotated summaries for daily-life dialogues from three datasets: DailyDialog (Li et al., 2017), DREAM (Sun et al., 2019), and MuTual (Cui et al., 2020), as well as dialogues from an English-speaking practice website. It has 13,460 dialogues in total.

4 Experiments

4.1 Downstream Tasks and Metrics

We evaluate our methods on three public dialogue summarization datasets or benchmarks: SAMSum (Gliwa et al., 2019), ConvoSumm (Fabbri et al., 2021), and TWEETSUMM (Feigenblat et al., 2021)

SAMSum This dataset contains over 16k natural messenger-like dialogues with manually annotated summaries by language experts.

ConvoSumm It is a benchmark of four domains: New York Times comment, StackExchange, W3C email, and Reddit. Dialogues are extracted from publicly available data, and each domain has 500 dialogues. They hire crowdsorce workers on Amazon Mechanical Turk to annotate dialogue summary.

TweetSumm This dataset contains 1,100 reconstructed real-world customer support dialogues from Tweet. Each dialogue has human annotated abstractive summaries and extractive summaries. We only use abstractive summaries in the dataset as references in our experiments.

We report ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004) to evaluate generated summaries against references.

4.2 Baselines

We compare our methods with three competitive baselines.

T5v1.1 It is an improved version of the original T5 model (Raffel et al., 2022). Since the original T5 model is pre-trained on downstream tasks in supervised learning, the test set of downstream tasks overlaps with the pre-training data. To make a fair comparison in a zero-shot setting, we choose T5v1.1 as it is pre-trained on C4 without mixing in the downstream tasks.

PEGASUS Zhang et al. (2019) propose this pre-trained model for abstractive summarization tasks. The pre-training objective is GSG, transforms any text into an abstractive summarization example by selecting important sentences as output summaries. We use the $\text{PEGASUS}_{\text{LARGE}}$ checkpoint[7] as there is no publicly available $\text{PEGASUS}_{\text{BASE}}$ checkpoint.

GSG* We use the independent principal strategy of GSG training objective in PEGASUS (Zhang et al., 2019) but pre-train DIONYSUS with our training corpora. We build this baseline to explore the performance gap between our pre-training objective and GSG.

5 Results and Analysis

We focus on low-resource dialogue summarization settings because it is difficult to collect enough training examples. We evaluate DIONYSUS with "All G", "All P", and "Better ROUGE" strategies in zero-shot and few-shot settings and compare it to the baselines.

5.1 Zero-Shot Results

In order to evaluate the effectiveness of DIONYSUS, we conduct a zero-shot test on $\text{DIONYSUS}_{\text{LARGE}}$ with all strategies and other baselines. We present the results in Table 2. The ROUGE1-F1, ROUGE2-F1, and ROUGEL-F1 scores are used as the standard evaluation measures for summarization tasks. Our models show impressive performance improvements over the baselines on all downstream datasets. Specifically, $\text{DIONYSUS}_{\text{LARGE}}$ with the "Better ROUGE" strategy performs the best overall across all downstream datasets (Average: ROUGE-1/2/L: $29.7/8.0/20.2$), indicating that it benefits from both generated and extractive pseudo summaries and can adapt to various domains. The "All P" strategy performs better than the GSG* baseline on most datasets, indicating that our Gap Sentence Selection Plus method can effectively select dialogue turns that provide an accurate dialogue summary. Additionally, the $\text{DIONYSUS}_{\text{LARGE}}$ with "All G" and "Better ROUGE" strategies demonstrate significant improvement compared to $\text{T5v1.1}_{\text{LARGE}}$ (Average ROUGE2: $+5.6/+6.1$) and $\text{PEGASUS}_{\text{LARGE}}$ (Average ROUGE2: $+2.2/+2.7$), indicating that pre-training with our summary helper is highly beneficial. However, the "All G" strategy only performs as well as the "Better ROUGE" strategy on the SAMSum dataset (ROUGE-1/2/L/: $41.3/16.1/30.6 \rightarrow 41.3/16.2/30.9$), suggesting that the improvement from the summary helper is more pronounced on this particular dataset. This may be due to the similarity between the datasets used to train the helper and the SAMSum dataset, which we discuss further in Sections 5.5 and 5.6. Overall, our models outperform previous methods, such as PEGASUS, in a zero-shot setting, demonstrating their effectiveness and potential for further development.

5.2 Few-Shot Results

We investigated reducing annotation labor in dialogue summarization tasks by using few-shot dialogue summarization. We report ROUGE1-F1, ROUGE2-F1, ROUGEL-F1, and ROUGELSum-F1 scores to evaluate model performance. Specifically, We fine-tune $\text{DIONYSUS}_{\text{LARGE}}$, $\text{PEGASUS}_{\text{LARGE}}$, and $\text{T5v1.1}_{\text{LARGE}}$ with the first $1/10/100/1K/10K$ training examples from the SAMSum dataset. We show the results of our experiments with varying training data sizes in Figure 3. We found that all models improved with more examples. Among these models, $\text{DIONYSUS}_{\text{LARGE}}$ consistently outperformes both $\text{PEGASUS}_{\text{LARGE}}$ and $\text{T5v1.1}_{\text{LARGE}}$ when trained with a dataset ranging from 0 to $10,000$ examples. This suggests that our pre-training process helps DIONYSUS adapt to downstream tasks more quickly. Additionally, we observed that $\text{PEGASUS}_{\text{LARGE}}$ outperformed $\text{T5v1.1}_{\text{LARGE}}$ due to its pre-training on summarization tasks. Figure 3 shows the gap between $\text{DIONYSUS}_{\text{LARGE}}$ and $\text{PEGASUS}_{\text{LARGE}}$ is particularly significant when using fewer than 100 training examples, indicating better recall capabilities in dialogue summarization for DIONYSUS. Even with only 10 training examples, $\text{DIONYSUS}_{\text{LARGE}}$ achieves higher ROUGE scores than the $\text{T5v1.1}_{\text{LARGE}}$ model trained with 1,000 examples, making it the best option for low-resource dialogue summarization.

5.3 Effect of Compression Ratio

In GSG+, we can choose a fixed number of turns in the dialogue as a training objective or select turns with a compression ratio. We investigate the compression ratio in a dialogue turn level as the number of selected turns over the number of totals turns in the dialogue ($N_{principal}/N_{dialogue}$). A

[7] https://huggingface.co/google/pegasus-large

Model	SAMSum	NYT	Reddit	Stack	Email	TweetSumm	Avg.
T5v1.1	9.6/1.6/8.6	11.6/1.4/8.7	12.3/1.7/9.2	15.6/2.4/11.0	14.9/2.7/11.1	6.0/1.4/5.1	11.7/1.9/9.0
PEGASUS	27.5/7.6/21.5	23.7/3.2/13.2	23.1/4.1/13.6	26.7/4.8/15.2	23.9/5.7/15.3	21.8/6.3/16.0	24.5/5.3/15.8
GSG*	13.3/3.5/12.0	17.1/2.4/12.9	16.0/2.1/12.5	21.2/3.5/15.1	21.0/4.2/15.9	15.4/2.8/13.1	17.3/3.1/13.6
Ours: G	**41.3**/16.1/30.6	21.7/3.7/14.8	23.5/4.3/15.7	26.3/5.4/16.8	26.4/7.1/17.2	29.4/8.4/22.1	28.1/7.5/19.5
Ours: P	23.5/7.5/18.6	19.8/2.7/12.9	20.0/2.9/12.7	24.5/4.3/15.0	24.3/5.5/15.8	22.1/6.7/17.6	22.4/4.9/15.4
Ours: BR	**41.3/16.2/30.9**	**24.1/4.0/15.4**	**24.8/4.4/15.9**	**28.5/5.6/17.6**	**28.9/7.7/18.0**	**30.7/10.1/23.4**	**29.7/8.0/20.2**

Table 2: The ROUGE-1/ROUGE-2/ROUGE-L scores of the DIONYSUS$_{\text{LARGE}}$ with strategy P: "All P", G: "All G", and BR: "Better ROUGE" and compared to T5v1.1$_{\text{LARGE}}$ and PEGASUS$_{\text{LARGE}}$ in a zero-shot setting on three datasets: SAMSum, ConvoSumm, and TweetSumm.

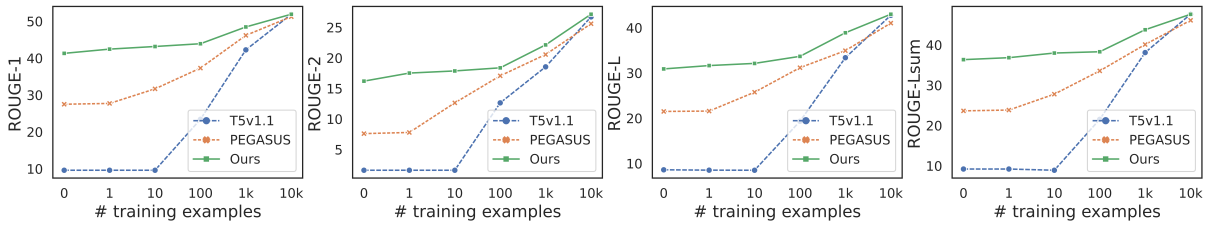

Figure 3: Comparison of T5v1.1$_{\text{LARGE}}$, PEGASUS$_{\text{LARGE}}$, and DIONYSUS$_{\text{LARGE}}$, fine-tuned with limited training examples on the SAMSum dataset. The training data is within 10,000 examples. The results show that DIONYSUS outperforms both PEGASUS and T5v1.1 on all four metrics.

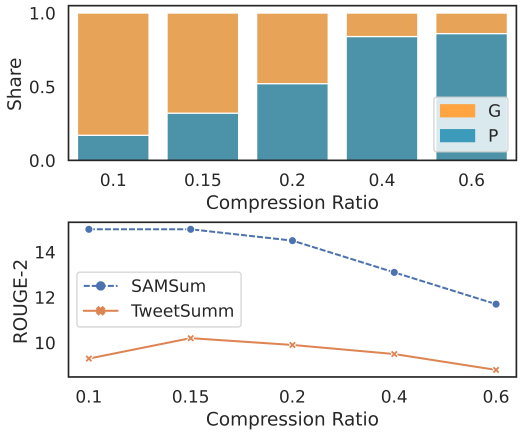

Figure 4: Comparison of compression ratios in DIONYSUS$_{\text{BASE}}$ using "Better ROUGE" strategy. The upper figure reflects the percentage of generated summaries (G) and "Princial" (P) at different compression ratios. The performance is measured using the ROUGE2-F1 metric on the SamSum and TweetSumm development sets.

low compression ratio will select fewer turns in the dialogue as the objective, making pre-training less challenging. However, it tends to have a lower ROUGE1-F1 score with the remaining dialogue turns, meaning the "Better ROUGE" strategy selects more generated summaries as the objective. While choosing a high compression ratio will make the pre-training more challenging. Nevertheless, it has a higher ROUGE score compared to generated summaries, leading to more principal under the "Better ROUGE" strategy. We show the zero-shot performance on development sets of the SAMSum dataset and TweetSumm dataset with compression rates from 10% to 60% in Figure 4. It shows that the model with 15% compression ratio achieves the highest ROUGE-2 score.

5.4 Effect of Copying Mechanism

ROUGE-1/2/L	All P	w/o copying
SAMSum	**25.8/8.5/19.7**	17.7/5.7/15.7
NYT	**21.3/2.7/13.5**	17.4/2.2/13.4
Reddit	**22.3/3.4/13.8**	16.3/2.6/13.1
Stack	**25.9/4.5/15.8**	20.3/3.4/15.1
Email	**26.6/6.1/16.8**	20.0/3.5/14.7
TweetSumm	**24.1/8.5/19.0**	19.4/3.8/16.3

Table 3: ROUGE-1/2/L scores of zero-shot setting for DIONYSUS$_{\text{BASE}}$ with "All P" strategy and "All P" without copying mechanism on SAMSum, ConvoSumm, and TweetSum.

The copying mechanism is important for dialogues like meetings and medical dialogues because it allows for summarization of entire dialogue through several turns. As shown in Table 3, we compare the performance of the "All P" strategy to a scenario where 50% of the selected dialogue turns are retained in the input rather than being removed. In this case, the input for each pre-training example includes the entire dialogue D, rather than $D \setminus P$. This leads the model to focus on extractive summarization. We observed that adding a random copy mechanism significantly improved the overall performance. Additionally, we

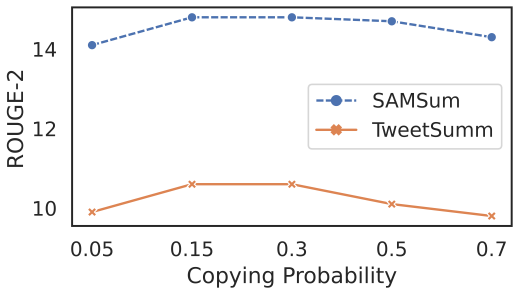

Figure 5: Comparing probabilities of copying selected sentences in the input of the "Principal" using the "Better ROUGE" strategy. Evaluating performance using the ROUGE2-F1 metric on SamSum and TweetSumm development datasets.

also evaluate the "Better ROUGE" strategy with different copying probabilities ranging from 0.15 to 0.7. In these experiments, we choose top-2 dialogue turns as principal, which results in 51.9% of pre-training objectives being the principal, and the rest is the generated summary. Figure 5 shows that leaving 15% of dialogue turns in the principal best enhances the overall quality of dialogue summarization.

5.5 Comparison Between All G and Summary Helper

ROUGE-1/2/L	All G	Helper
SAMSum	**41.3/16.1/30.6**	35.8/13.5/27.9
NYT	**21.7**/3.7/14.8	21.2/4.0/15.2
Reddit	**23.5/4.3/15.7**	20.2/3.5/14.4
Stack	**26.3/5.4/16.8**	25.1/5.0/16.0
Email	**26.4/7.1/17.2**	22.9/5.6/15.2
TweetSumm	**29.4/8.4/22.1**	26.8/6.2/20.8

Table 4: ROUGE-1/2/L scores of zero-shot setting for $\text{DIONYSUS}_{\text{BASE}}$ with "All G" strategy and the summary helper on SAMSum, ConvoSumm, and TweetSum.

Since the summary helper model provides the generated summary as an objective candidate and has shown strong capabilities in zero-shot dialogue summarization. As shown in Table 4, we compare the helper model to our "All G" model in a zero-shot setting. The difference is that we train the "All G" model on the pre-training corpora annotated by the helper. We found that the helper model is not on par with our model. While the helper model may have performed well on a particular task (NYT), its overall performance is not as strong as our model. This is because DIONYSUS has been extensively trained on various dialogue datasets, which makes it consistently perform well in a wide range of tasks and scenarios.

5.6 Test-Set Overlap with Pre-Training Corpora

Threshold	ConvoKit	DS2	DialogSum
≥ 1.0	0%	0%	0%
≥ 0.8	0%	0%	0%
≥ 0.6	0%	0%	1%
≥ 0.4	5%	0%	3%

Table 5: Percentage of overlap between the SAMSum test set and the datasets used for pre-training. The ConvoKit corpora were comprised of a randomly selected 10% of the total datafor calculating the similarity.

In order to ensure a fair comparison, we check for overlap between pre-training and downstream test datasets. This is done by calculating the similarity between all pairs of test set targets in the SAMSum dataset and pre-training documents using the ROUGE2-recall measure, which is calculated as the number of overlapping bigrams divided by the total number of bigrams in the test target. We then count the number of test set examples that have a similarity to any pre-training example above a certain threshold. As shown in Table 5, the overlap between the SAMSum dataset and the datasets used for training the helper and the pre-training datasets is low when the similarity threshold is set between 0.4 and 1.0. This suggests that there is not significant similarity between our test set and the pre-training datasets. It indicates that the improvement in DIONYSUS is due to the pre-training process rather than potential test data leakage.

5.7 Human Evaluation

	Ratings
$\text{T5v1.1}_{\text{LARGE}}$	3.54^{**}
$\text{PEGASUS}_{\text{LARGE}}$	3.90^{*}
$\text{DIONYSUS}_{\text{LARGE}}$	4.04
Human-written	4.08

Table 6: Human evaluation results of zero-shot generation. We test the T5v1.1 baseline and the PEGASUS model against DIONYSUS with **p < 0.01, *p < 0.05.

We evaluate the performance of DIONYSUS by conducting human evaluation experiments on Amazon Mechanical Turk. We randomly select 100 examples from the SAMSum dataset to compare summaries generated by our model with those written by humans in the dataset. We choose DIONYSUS trained with the "Better ROUGE" strategy and

generate summaries in a zero-shot setting. Participants are asked to rate the summaries on a scale of 1 to 5, with higher scores indicating better quality. We collect the scores from three participants for each example and report the average scores in Table 6. A paired t-test is conducted to determine if scores are significantly different between our model and other models. Our results show that DIONYSUS could generate summaries of similar quality to human-written summaries without any training data. DIONYSUS also gets better ratings than the vanilla T5 and PEGASUS models, which aligns with the results obtained from the automatic evaluation. More information on the human evaluation process can be found in Appendix F.

6 Related Work

Dialogue summarization is a rapidly growing area of research that focuses on automatically generating concise and informative summaries of conversations (Feng et al., 2022). Unlike research on traditional documents like news articles (Fabbri et al., 2019; Ahuja et al., 2022) or scientific papers (Lu et al., 2020; Ibrahim Altmami and El Bachir Menai, 2022), dialogue summarization is particularly relevant in multi-party interactions, such as emails (Zhang et al., 2021), meetings (Carletta et al., 2005), medical dialogues (Zeng et al., 2020), and daily chats (Chen et al., 2021). However, many existing methods for dialogue summarization require a large training dataset with annotated summaries. This can be a major barrier to applying these methods in real-world scenarios, particularly in cases with limited or no annotated data available. Our study examines the use of dialogue summarization in low-resource settings to make the process more practical and effortless in various contexts.

Pre-trained Transformer-based (Vaswani et al., 2017) language models (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019) have become increasingly popular in natural language processing tasks for tackling the data shortage problem. However, many of these models have limitations when it comes to dialogue summarization. Zhang et al. (2019) propose PEGASUS, which masks multiple whole sentences and pre-trains sequence-to-sequence models to reconstruct the original text. Built on that, Wan and Bansal (2022) improve the sentence selection strategy and add modules for ensuring factuality during fine-tuning to address the problem of factuality in summarization. Phang et al. (2022) extend PEGASUS with a modified architecture and long-sequence pre-training to tackle long-input summarization. He et al. (2022) propose ZCode++, a pre-trained language model optimized for abstractive summarization with improved encoder. However, all these methods rely on the Gap Sentence Selection method, which has limitations for dialogue summarization. In contrast, our approach uses pseudo-summary construction as the pre-training objective, making it possible for zero-shot dialogue summarization.

Another line of work focuses on pre-trained models for dialogues. DialoGPT (Zhang et al., 2020) and PLATO (Bao et al., 2020), which are pre-trained on large-scale conversation datasets such as Reddit. For dialogue summarization, Jia et al. (2022) post-train pre-trained language models to rephrase dialogues into narratives and then fine-tunes them for summarization. In contrast, our approach follows the T5 model's unified text-to-text format for both pre-training and fine-tuning. Zhong et al. (2022) train UNILM (Dong et al., 2019) with a window-based denoising framework for long dialogue understanding and summarization but do not focus on low-resource settings. Zou et al. (2021) propose a pre-training paradigm that pre-trains the encoder and decoder separately in a supervised manner. While our method uses a self-supervised pre-training approach that applies to any dialogue dataset, making it easier to extend to larger pre-training corpora for further improvement.

7 Conclusion and Future Work

We present DIONYSUS, a pre-trained encoder-decoder model for zero-shot dialogue summarization in any new domain. We pre-train using a self-supervised approach that generates pseudo-summaries for large dialogue corpora as the pre-training objective. We investigate the impact of various pre-training objective strategies and model sizes on dialogue summarization performance. Our experiments show that DIONYSUS outperforms state-of-the-art models on six datasets in a zero-shot setting. Furthermore, DIONYSUS can be fine-tuned with only 10 examples to outperform vanilla T5 fine-tuning with 1,000 examples. This makes dialogue summarization more practical and easier to use in various contexts with minimal effort. We plan to extend this method to abstractive summarization tasks to develop a general zero-shot summarization model.

8 Limitations

Training Data Our pre-training data is sourced from 19 existing dialogue datasets. However, it's important to note that these datasets may contain noise, such as harmful content, irrelevant file names, and URL links. Despite utilizing multiple automatic tools to filter out this content during pre-processing, there is still a chance that some noise may be present in our pre-training data. This could potentially impact the performance of DIONYSUS, making it important to monitor and improve the pre-processing steps continuously.

We also know the potential drawbacks of constructing pseudo summaries using the GSG method, which may lead to unnatural summaries for dialogue data. To mitigate this, we introduced the Summary Helper in Section 2.1, which is specifically trained on two dialogue summarization datasets containing natural summaries. This approach enables more realistic pseudo-summaries and enhances zero-shot performance. Although we employ top-m turns as an additional source of pseudo summaries, Figure 4 illustrates that GSG+ contributes a minor portion of the pseudo summary, with a 0.7 to 0.3 ratio between generated and top-m turns. Our method thus minimizes referent and pronoun confusion, ensuring better coherence than solely employing the standard GSG technique.

Training Resource To improve our model's performance, we employ the "Better ROUGE" strategy, which calculates the ROUGE score for both candidates and selects the best one as the final training objective. This data pre-processing process can be pretty time-consuming, taking approximately one day to complete for our pre-training data when utilizing 100 threads. Additionally, we utilize 16 Nvidia V100 GPUs to train our models, which may not be accessible or reproducible for all researchers. This could present a significant obstacle for those looking to replicate or build upon our work.

Test Data Another potential concern is the test datasets used to evaluate DIONYSUS. The test set size is relatively small, which may not fully represent the breadth of dialogue types that a general dialogue summarization model should be able to handle. This could lead to the model performing well on the test set but not generalizing to other unseen dialogue types. Further, our analysis did not include the assessment of long dialogue summarization, such as lengthy meetings (Carletta et al., 2005; Zhong et al., 2021; Janin et al., 2003) or screenplays (Chen et al., 2022). However, our study's approach has the potential to handle these scenarios, even though it was not specifically designed for them. By incorporating LongT5 (Guo et al., 2022) or DialogLM (Zhong et al., 2022), which are known for their ability to process extended input sequences, we expect that they could efficiently tackle this task.

9 Acknowledgement

Our gratitude goes out to Microsoft Research for providing us with computational resources. We would also like to thank Kun Qian for valuable discussions and the Columbia NLP and Microsoft Deep Learning Group members for their feedback and discussions. Additionally, we thank the Mechanical Turk workers for conducting the human evaluation.

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

McKenzie Braley and Gabriel Murray. 2018. The group affect and performance (gap) corpus. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA. Association for Computing Machinery.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference*

on Machine Learning for Multimodal Interaction, MLMI'05, page 28–39, Berlin, Heidelberg. Springer-Verlag.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

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

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5453–5460. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. In *Proceedings of the IJCAI workshop on NLP meets Journalism*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong

Kong, China. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, et al. 2022. Z-code++: A pre-trained language model optimized for abstractive summarization. *arXiv preprint arXiv:2208.09770*.

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1011–1028.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Zhu. 2022. Post-training dialogue summarization using pseudo-paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1660–1669, Seattle, United States. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Elijah Mayfield and Alan W. Black. 2019. Analyzing wikipedia deletion debates with a group decision-making forecast model. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Benjamin S. Meyers, Nuthan Munaiah, Emily Prud'hommeaux, Andrew Meneely, Josephine Wolff, Cecilia Ovesdotter Alm, and Pradeep Murukannaiah. 2018. A dataset for identifying actionable feedback in collaborative software development. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 126–131, Melbourne, Australia. Association for Computational Linguistics.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *ArXiv*, abs/2208.04347.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. 2022. Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3824–3846, Dublin, Ireland. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017a. Characterizing online discussion using coarse discourse sequences. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):357–366.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017b. Asking too much? the rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1572, Copenhagen, Denmark. Association for Computational Linguistics.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained

model for long dialogue understanding and summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11765–11773.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 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, Online. Association for Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

A Implementation Details

Following Raffel et al. (2022) and Zhang et al. (2019) to save time and computation, we first conduct ablation experiments on a reduced-size T5v1.1$_{\text{BASE}}$ model with $250M$ parameters. Then we scale up with the best settings to the final T5v1.1$_{\text{LARGE}}$ model with $800M$ parameters. We use heuristics to clean up our pre-training corpora. First, we remove dialogues with less than two dialogue turns since they are too short to summarize. Then we remove URLs and emojis in the text. DIONYSUS is implemented with Huggingface Pytorch Transformers[8] (Wolf et al., 2020). We split dialogue turns with line breakers in pre-training input and add a "[Summary]" prefix. For pseudo summary creation, we use a compression ratio of 0.15 for the "Principal." This means that for a dialogue with l turns, we select $0.15l$ turns as "Principal." We explore the effect of different compression ratios in Section 5.3. We use Adam (Kingma and Ba, 2014) with weight decay for pre-training. We truncate dialogue training examples to ensure a maximum length of 512. Models are pre-trained with batch size 8 and learning rate 0.00001 on 16 Nvidia V100 GPUs until we observe no progress on validation data or up to 5 epochs. For few-shot experiments in Section 5.2, we fine-tune models up to 20 epochs with batch size 8 and learning rate 0.00005, and pick the checkpoint with the best validation performance.

[8]https://github.com/huggingface/transformers is licensed under the Apache License 2.0

B Additional Base Model Results

Table 7 presents the results of DIONYSUS$_{\text{BASE}}$ in a zero-shot setting, and Figure 6 compares the few-shot results of DIONYSUS$_{\text{BASE}}$ with those of the T5 base model. These initial results demonstrate the potential for further analysis and optimization of DIONYSUS. Upon comparison with other baselines, it is clear that DIONYSUS performs better under both zero-shot and few-shot conditions, outperforming the GSG* model. These results provide valuable insight into the capabilities of DIONYSUS and can inform the development of larger models.

C Effect of the Dialogue Turns Order in Principal

We could use two possible orders to align the dialogue turns in the principal. The first order is to align the text with the ROUGE1-F1 score. The second order is to align the principal with the order in the original dialogue. This means that the principal will be arranged in the same order as in the original dialogue, without rearrangement. This option helps preserve the original flow and structure of the dialogue. We compare these two orders of principal in the GSG* baseline. As shown in Table 8, the results suggest that keeping the order in the original dialogue helps improve zero-shot performance as it provides a more nuanced understanding of the dialogue. We choose this order for all our models.

D Pre-training Steps

To evaluate the performance of DIONYSUS during pre-training, we measured the ROUGE1-F1, ROUGE2-F1, ROUGEL-F1, and ROUGELSum-F1 scores on the SAMSum dataset in Figure 7. We keep track of the model's progress by logging its performance every 1,000 training steps. This allows us to monitor the model's improvements over time and confirm that it is learning effectively.

E Example Model Outputs

In order to evaluate the performance of DIONYSUS, we randomly selected model output examples

Model	SAMSum	NYT	Reddit	Stack	Email	TweetSumm
T5v1.1$_{\text{BASE}}$	9.7/1.2/8.6	5.8/0.7/4.9	8.9/1.2/7.3	11.5/1.7/8.9	8.4/1.6/7.2	6.8/1.0/6.2
GSG*	13.7/4.0/12.6	17.9/2.4/13.9	15.8/2.2/12.7	20.7/3.4/15.5	20.8/3.8/15.9	17.0/3.2/14.5
All G	39.2/15.2/29.5	20.0/3.1/13.7	21.4/3.6/14.7	24.1/4.9/16.0	24.1/6.5/16.0	28.3/9.0/22.1
All P	25.8/8.5/19.7	21.3/2.7/13.5	22.3/3.4/13.8	25.9/4.5/15.8	26.6/6.1/16.8	24.1/8.5/19.0
Better ROUGE	**39.6/15.4/30.1**	**23.1/3.7/15.0**	**23.1/4.0/15.1**	**27.3/5.6/17.1**	**27.0/6.9/17.6**	**30.3/9.8/23.2**

Table 7: The ROUGE-1/ROUGE-2/ROUGE-L scores of the DIONYSUS$_{\text{BASE}}$ when implemented with different strategies and compared to T5v1.1$_{\text{BASE}}$ in a zero-shot setting on three datasets: SAMSum, ConvoSumm, and TweetSumm.

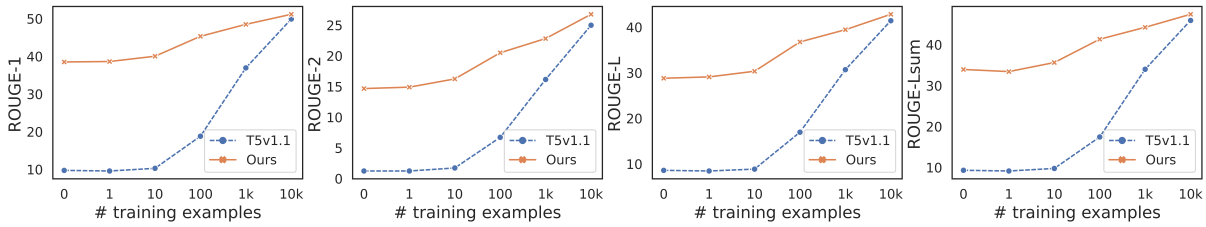

Figure 6: The ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum scores of low resource dialogue summarization with our best model and T5v1.1. Within 10,000 examples, DIONYSUS$_{\text{BASE}}$ beats T5v1.1 on all metrics on SAMSum dataset.

ROUGE-1/2/L	GSG* (Dialogue)	GSG* (ROUGE)
SAMSum	**13.7/4.0/12.6**	13.1/3.7/12.2
NYT	**17.9/2.4/13.9**	17.6/2.2/13.7
Reddit	**15.8**/2.2/**12.7**	15.3/2.2/12.5
Stack	**20.7/3.4/15.5**	20.1/3.1/15.2
Email	**20.8/3.8/15.9**	19.8/3.6/15.1
TweetSumm	**17.0/3.2/14.5**	15.1/2.7/12.8

Table 8: ROUGE-1/2/L scores of zero-shot setting for DIONYSUS$_{\text{BASE}}$ with GSG* and unordered GSG* on SAMSum, ConvoSumm, and TweetSum.

from both the SAMSum dataset and the TweetSumm dataset. We report these examples with their corresponding gold summaries in Tables 9 and 10. The gold summaries served as a benchmark for our model's output, allowing us to compare and estimate the quality of the generated summaries. We found that DIONYSUS could generate zero-shot summaries on par with those written by humans. However, we also identified factual errors in the generated summaries, such as misunderstandings of the subject matter. These errors suggest room for improvement in DIONYSUS, and we plan to address this issue in future work.

F Human Evaluation Details

In our human evaluation experiments, we utilized the task template shown in Figure 8. Mechanical workers were instructed to rate four summaries for a given dialogue on a scale of 1 (poor) to 5 (excellent). To minimize bias, we provided a dialogue with its corresponding gold summary as an example of a high-quality summary. The summaries were presented in a randomized order for each task to prevent order bias. Three different workers independently completed each task, and the median score across all workers was retained for each summary. Participants were compensated with 0.3 USD per task, and we implemented the following qualifications for worker selection to ensure a high level of quality: (1) HIT approval rate for all requesters' HITs is greater than 90%. (2) Location is one of AU, NZ, GB, and US. (3) Number of HITs approved is greater than 100.

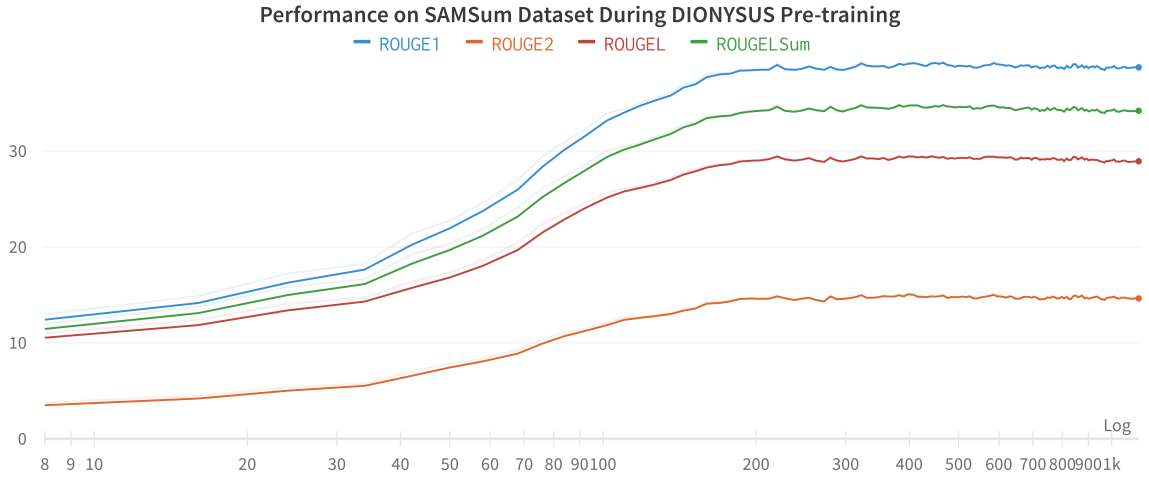

Figure 7: Performance of DIONYSUS on the SAMSum dataset during pre-training process.

Example	SAMSum
Dialogue#1	Dzuka: Until further notice, the staff meeting will be held at 8:30 instead of 8:00. Please change the calendar for everyone. Thanks. Anna: No problem. Why the change Dzuka: We had a few that never make it on time. I'm hoping this will encourage more participation. Anna: Could be just the opposite! Dzuka: We'll give it a try. Anna: Sure, no problem. Dzuka: I'll let you know if it changes again. Thanks. Anna: NP
Gold	The stuff meeting is postponed from 8.00 to 8.30 to encourage more participation. Dzuka will inform Anna if it changes again.
DIONYSUS	The staff meeting will be held at 8:30 instead of 8:00. Dzuka hopes this will encourage more participation and will let Anna know if it changes again.
Dialogue#2	Jane: Hello Vegano Resto: Hello, how may I help you today? Jane: I would like to make a reservation for 6 people, tonight around 20:00 Vegano Resto: Let me just check. Ah, I'm afraid that there is no room at 20:00. However, I could offer you a table for six at 18:30 or at 21:00. Would either of those times suit you? Jane: Oh dear. Let me just ask my friends. Vegano Resto: No problem. Jane: 21:00 will be ok. Vegano Resto: Perfect. So tonight at 21:00 for six people under your name. Jane: great, thank you!
Gold	Jane made a 9 PM reservation for 6 people tonight at Vegano Resto.
DIONYSUS	The restaurant has no room for six people at 20:00 and offers Jane a table for six at 18:30 or 21:00. Jane asks her friends to make a reservation at 21:00.
Dialogue#3	Mia: Hi Dad! I need a hand with repairing the bathroom door. William: Hi! What happened? Mia: Nothing. I can't open/close it properly. It's sort of sagging. William: I see. I'll drop by after work and take a look. Mia: Thank you so much! Love you! William: I love you too.
Gold	Mia's dad William will come to her place after work to repair her bathroom door.
DIONYSUS	The bathroom door is sagging. William will drop by after work and take a look.

Table 9: Example zero-shot outputs on the SAMSum dataset.

Example	TWEETSUMM
Dialogue#1	@549761: My BITS service resets it's Startup type from disabled to automatic. It leeches on to my bandwidth like crazy. Please provide a solution. @MicrosoftHelps: Hi. We hear you. We'd like to check what happened prior to this issue? What Windows version are you using? Let us know. @549761: I am using Windows 10 Home Single Language. Nothing specific happened prior to this issue. Just the service used to leech on to bandwidth (it claims to use idle network but doesn't). I want it to stop from resetting Startup type from disabled to automatic. @MicrosoftHelps: Thanks for the info. For us to isolate your concern, let's try the troubleshooting steps 1/2 https://t.co/3qcAsLFkaY listed in this link: https://t.co/IBZ1MaTm1l. Kindly check the post of Jesinta Rozario. @MicrosoftHelps: Hi, Pratik. How's it going? Please let us know if you need further assistance. We're here for you. @549761: Hi. The service still becomes running after disabling(after a few days). What can be the reason for the service switching it's startup type? @MicrosoftHelps: In that case, we suggest contacting Answer Desk: https://t.co/9Ouw33YVZI to further assist you with your concern. Let us know how it goes. @MicrosoftHelps: Hello, Pratik! Were we able to resolve your concern? If no, we're just one tweet away if you have other concerns. If yes, please send us your feedback about your experience with our support here: https://t.co/CczzJgTng1.
Gold	Customer is complaining about the BITS service for resetting startup type from disabled mode to automatic. Agent suggests to try out some troubleshooting steps by following the shared URL and reach out Answer desk team for further assistance.
DIONYSUS	The BITS service leeches on to the bandwidth like crazy. Pratik wants it to stop from resetting Startup type from disabled to automatic. MicrosoftHelps suggests checking the post of Jesinta Rozario.
Dialogue#2	@471404: Please bring security back to the Hall Green store. @471404: The store is getting a more an more uncomfortable vibe, not alone on this either! @Tesco: Hi there, sorry to be a pain but can you confirm which Hall Green store this is? TY - Reece @471404: It's the Hall Green store right next to the train station. Hoping you haven't removed security from the others too now... @Tesco: Hi, can you please confirm what you mean by "uncomfortable vibe"? - Nick @471404: Well there's pretty obvious shop lifters regularly, and today we had a man clearly intoxicated screaming and randomly asking people things. @Tesco: Yes the express store! Thanks aswell. I'd review the CCTV from when security were removed. If customers can see the changes you will too! @Tesco: Hi there. I have spoken to the store. They have had a few problems recently and are looking into improving security. Thanks - Ian @471404: Thank you again. I often worry for the staff as it is becoming a hot spot for undesirables. The homeless aren't the issue to save confusion! @Tesco: Hi there, thank you for bringing this to our attention the last thing we want is our customers to feel unsafe. Thank you - Brooke @471404: No thank you for taking it seriously here's hoping the store gets back to normal soon! @Tesco: Hi there, I'm glad one of my colleagues has dealt with the issue. Enjoy the rest of your weekend - Rian
Gold	The customer is complaining that he facing some uncomfortable vibe. The agent confronted the customer saying that they had a few problems recently and they are looking into improving security.
DIONYSUS	The store is getting a more an more uncomfortable vibe. Nick asks Tesco to bring security back to the Hall Green store and confirms the location. Nick also tells Tesco the Express store has had some problems recently and is looking into improving security.

Table 10: Example zero-shot outputs on the TWEETSUMM dataset.

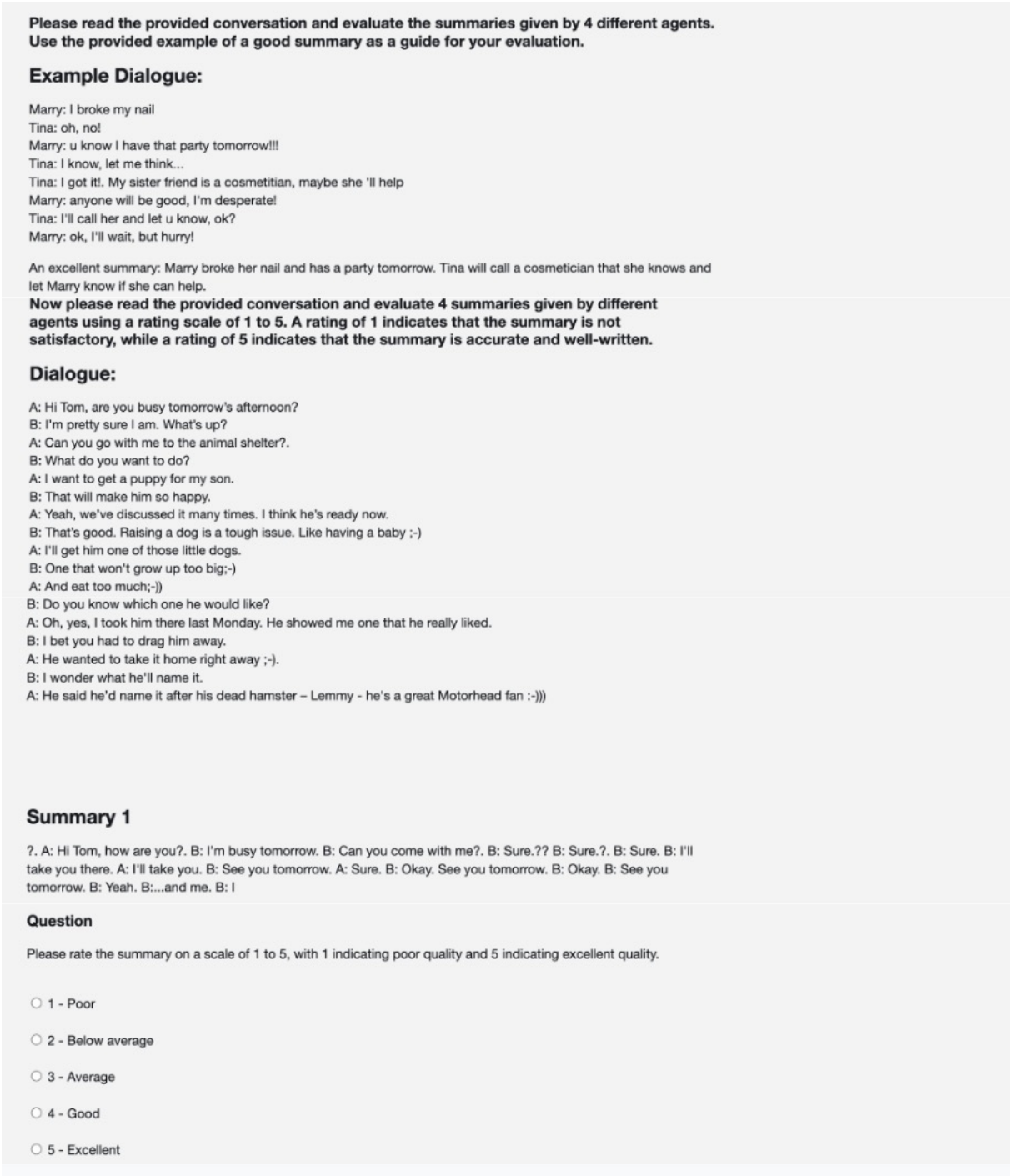

Figure 8: A screenshot of the human evaluation on Amazon Mechanical Turk.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 1

☑ B1. Did you cite the creators of artifacts you used?
Section 1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix B

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 1

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. It is discussed in the original artifacts I use.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. It is discussed in the original artifacts I use.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.7

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix F

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix F

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix F

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. It is in the Amazon Mechanical Turk user agreement protocal.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. It is in the Amazon Mechanical Turk user agreement protocal.