ParaSum: Contrastive Paraphrasing for Low-Resource Extractive Text Summarization

Moming Tang[1], Chengyu Wang[2], Jianing Wang[1], Cen Chen[1(✉)], Ming Gao[1], and Weining Qian[1]

[1] School of Data Science and Engineering, East China Normal University, Shanghai, China
cecilia.cenchen@gmail.com
[2] Alibaba Group, Hangzhou, China

Abstract. Existing extractive summarization methods achieve state-of-the-art (SOTA) performance with pre-trained language models (PLMs) and sufficient training data. However, PLM-based methods are known to be data-hungry and often fail to deliver satisfactory results in low-resource scenarios. Constructing a high-quality summarization dataset with human-authored reference summaries is a prohibitively expensive task. To address these challenges, this paper proposes a novel paradigm for low-resource extractive summarization, called ParaSum. This paradigm reformulates text summarization as textual paraphrasing, aligning the text summarization task with the self-supervised Next Sentence Prediction (NSP) task of PLMs. This approach minimizes the training gap between the summarization model and PLMs, enabling a more effective probing of the knowledge encoded within PLMs and enhancing the summarization performance. Furthermore, to relax the requirement for large amounts of training data, we introduce a simple yet efficient model and align the training paradigm of summarization to textual paraphrasing to facilitate network-based transfer learning. Extensive experiments over two widely used benchmarks (i.e., CNN/DailyMail, Xsum) and a recent open-sourced high-quality Chinese benchmark (i.e., CNewSum) show that ParaSum consistently outperforms existing PLM-based summarization methods in all low-resource settings, demonstrating its effectiveness over different types of datasets.

Keywords: low-resource scenarios · extractive summarization · textual paraphrasing · transfer learning · pre-trained language model

1 Introduction

The exponential proliferation of information on the Internet has created an urgent need for industrial scenarios to extract knowledge from vast amounts of documents. Extractive summarization aims at reducing information acquisition costs while preserving the key information of the document, which leads to a significant surge in interest in text summarization from both academic

Z. Jin et al. (Eds.): KSEM 2023, LNAI 14119, pp. 106–119, 2023.
https://doi.org/10.1007/978-3-031-40289-0_9

and industrial communities [1,2]. Most extractive summarization methods are implemented in a supervised fashion, which can be categorized into two classes, i.e., ranking-based summarization and auto-regressive summarization. Ranking-based methods are designed to assign a numerical score to each sentence in a given document and subsequently select the top K sentences that achieve the highest scores to form a summary [1,3], while auto-regressive summarization methods usually employ the Seq2Seq architecture and extracts summaries on a sentence-by-sentence basis during the decoding phase, with each sentence being selected based on the summary that has been produced up to that point [4,5].

In recent times, pre-trained language models (PLMs) have emerged as an essential infrastructure for a wide range of downstream natural language understanding (NLU) tasks [6,7]. These models are fully pre-trained on large-scale corpora through well-designed self-supervised learning techniques. Recent methods have incorporated the PLM-based fine-tuning paradigm into either ranking-based [3,8] or auto-regressive summarization [9] approaches and successfully achieve much better summarization performance through the use of fully-supervised labeled data. However, the conventional fine-tuning frameworks heavily depend on the time-consuming and labor-intensive process of data annotation[1], which may be bothersome in real-world low-resource scenarios (e.g., having only 200 labeled samples for model training). In addition, there is a large gap between the pre-training objective of PLMs, i.e., NSP and the fine-tuning objective of extractive summarization, which hinders the transfer and adaptation of knowledge in PLMs to summarization tasks. Fortunately, PLMs have demonstrated remarkable ability in few-shot learning by redefining the downstream tasks' formats similar to their pre-training objectives [10]. To this end, the rich knowledge encoded in PLMs can be better probed through specific patterns to facilitate downstream NLP tasks even when there is a scarcity of labeled data available. Therefore, a natural question arises: *how can we employ text paraphrasing in PLMs to boost the model performance for low-resource extractive summarization?*

In order to effectively probe the knowledge of PLMs to improve extractive summarization, this paper presents a novel paradigm called ParaSum. The primary objective of this approach is to reformulate the extractive summarization task as text paraphrasing. Textual paraphrasing that determines whether given sentence pairs are paraphrases of one another, has a latent connection to extractive summarization which aims to extract a summary that paraphrases the gist of the document. To support this reformulation, we obey two principles:

a) The training paradigm gap between extractive summarization and PLMs should be minimized.
b) The model's backbone architecture should be simple enough to facilitate network-based transfer learning to relax the requisition of training data and achieve satisfactory performance.

[1] Existing mainstream summarization datasets typically contain at least 100,000 news articles with corresponding human-authored reference summaries [11–13].

This paper proposes a novel paradigm for extractive summarization in low-resource scenarios, which is referred to as ParaSum. The approach involves reformulating extractive summarization as a textual paraphrasing task between the original document and candidate summaries. This paradigm is aligned with the self-supervised task (NSP) of BERT, thereby enabling the leveraging of knowledge from both textual paraphrasing and PLMs to guide the model in distinguishing semantically salient summaries. Extensive experiments were conducted in this study on two commonly used benchmarks (CNN/DailyMail, Xsum) and a recently released high-quality Chinese benchmark (CNewSum). The results indicate that ParaSum consistently outperforms the baseline models with scarcity available labeled data.

2 Related Work

Incorporating PLMs-based fine-tuning paradigm into extractive summarization methods significantly improves their results on large-scale, high-quality, open-source summarization datasets. However, as the parameter size of PLMs increases, the fine-tuning paradigm increasingly relies on a substantial amount of labor-intensive annotation. In many real-world application scenarios, only small-scale datasets are available, rendering existing data-hungry methods impractical to apply. While some researchers have been studying low-resource abstractive summarization [14–16], it is important to note that summaries generated by such methods may deviate from documents' main information. Therefore, this paper mainly focuses on extractive summarization. To the best of our knowledge, this study represents the first investigation of low-resource extractive summarization.

Another thread of related work is few-shot Natural Language Understanding (NLU). Here, we briefly summarize few-shot NLU in the following topics. i) Partial-parameter fine-tuning paradigms only tune a subset parameters of PLMs during training aiming to preserve most pre-trained knowledge of PLMs while reducing the magnitude of tunable parameters [17,18]. ii) Prompt Engineering [19] reformulates visual grounding and visual question answering as a "fill-in-blank" problem by hand-crafted prompts. iii) Adapter-based paradigms [20,21] utilize a lightweight adapters module to learn from downstream small-scale training datasets during training, while combining the knowledge of both PLMs and downstream tasks through residual functions during inference to complete downstream tasks [22]. iv) Transfer learning [23,24] is employed to overcome the constraint that training and testing data must be independent and identically distributed, thereby alleviating the scarcity of supervisory signals caused by small-scale training datasets. Drawing inspiration from prompt learning and transfer learning, this paper aims to alleviate the issue of sparse supervisory signals in low-resource scenarios by leveraging knowledge from relevant NLU tasks and PLMs for extractive summarization.

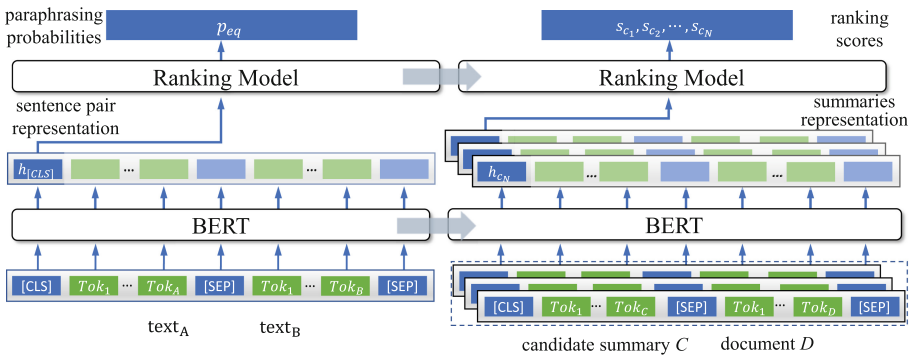

Fig. 1. Model architecture and transfer learning procedure of ParaSum.

3 Method

This paper dispenses with data-hungry summarization paradigms (ranking-based [3] and auto-regressive [4,9]). Since PLMs are few-shot learners [25], this paper aims to minimize the training gap between extractive summarization and PLM in order to leverage the knowledge of PLMs for downstream summarization. Additionally, network-based transfer learning is employed to alleviate the issue of sparse supervisory signals. To achieve these objectives, extractive summarization is reformulated as textual paraphrasing. Specifically, the method distinguishes the most salient summary from candidate summaries of the document. The model architecture and the transfer learning process are illustrated in Fig. 1.

3.1 PLM-Based Textual Paraphrasing

Given the textual paraphrasing sentence pairs, namely $\{(s_1^t, s_2^t), y_t\}_{t=1}^T$, where (s_1^t, s_2^t) indicates the t-th sentence pair, and y_t is the corresponding binary label. The sentence pair (s_1^t, s_2^t) is concatenated as: x_{in} = `[CLS]` s_1^t `[SEP]` s_2^t `[SEP]`. The PLM, parameterized by $\mathcal{M}$, takes x_{in} as input and maps x_{in} to a sequence of hidden representations. The representation located at position `[CLS]`, denoted as $h_{\texttt{[CLS]}}$, is used as input to a ranking model, which computes the probability p_{eq}. This probability indicates the extent to which s_1^t and s_2^t are paraphrases of each other. The ranking model typically consists of a single-layer feed-forward network (FFN) and a `sigmoid` function. The computation proceeds as follows:

$$h_{\texttt{[CLS]}} = \mathcal{M}(x_{in})_{\texttt{[CLS]}} \tag{1}$$

$$p_{eq} = \text{sigmoid}(W^\top h_{\texttt{[CLS]}} + b) \tag{2}$$

where W and b are learnable parameters of FFN.

3.2 Extractive Summarization as Textual Paraphrasing

PLM-based approaches for textual paraphrasing adhere to the NSP self-supervised task of PLMs [6] which transforms textual paraphrasing into a task that requests answers from PLMs, thereby reducing the need for large-scale training data. To enhance the performance of summarization models in low-resource scenarios by retrieving relevant knowledge from PLMs, we reformulated the summarization task as a series of textual paraphrasing operations between the source document and its corresponding candidate summaries.

Given a document D with L sentences, namely $D = \{s_i | i = 1 \cdots L\}$, and its reference summary R_D, where s_i denotes the i-th sentence of D. We construct candidate summaries of D, namely $C_D = \{c_j | j = 1 \cdots N\}$, where N denotes the number of candidate summaries. The candidate summary $c_j = \{s_k^j | k = 1 \cdots I \ , s_k^j \in D\}$ consists of I sentences from D, where s_k^j denotes the k-th sentence of c_j. Sentences in c_j are sorted according to their position in D. Candidate summaries in C_D are sorted in descending order according to their ROUGE score [26], namely $R(\cdot)$, with the reference summary R_D, in other words:

$$R(c_m) > R(c_n); \forall c_m, c_n \in C_D \ and \ m < n \tag{3}$$

The first step in our approach is concatenating the candidate summary c_j with the source document D, just as x_{c_j} = `[CLS]` c_j `[SEP]` D `[SEP]`. Next, we input x_{c_j} into PLM $\mathcal{M}$ and utilize the output representation at the `[CLS]` position to compute the ranking score s_{c_j} for c_j. This score is computed using the same ranking model employed in the textual paraphrasing step. Subsequently, we obtain ranking scores for the candidate summaries in C_D, which are represented as $S_D = \{s_{c_j} | j = 1 \cdots N\}$. Furthermore, we compute the ranking score of the reference summary R_D, denoted as s_{R_D}, using the same calculation procedure.

3.3 Training Paradigm

Supervised Training Paradigm. In order to ensure that the model assigns the highest probability to the candidate summary with the highest ROUGE score, denoted as c_1, we employ the cross-entropy loss (CE). Specifically, we set the ground-truth label for c_1 to 1, and 0 for the remaining summaries:

$$L_{CE} = \sum_{j=1}^{N} -l_j \cdot \log s_{c_j}; \quad l_1 = 1 \text{ and } \{l_j\}_{j=2}^{N} = 0 \tag{4}$$

where l_j denotes the ground-truth label for c_j. However, in the case of low-resource datasets, the supervised information provided by the CE loss is insufficient to train the model to achieve satisfactory performance. The sorting information of the candidate summaries can be employed by the contrastive learning paradigm to guide the model in distinguishing salient summaries. Contrastive learning allows for the definition of positive and negative samples based on the specific application scenarios, while ensuring that negative samples are kept distinct from positive samples [27]. Given any candidate summary pair $c_m, c_n \in C_D$

where $m < n$, we define c_m with a higher ROUGE score as a positive sample and c_n with lower ROUGE score as a negative sample. The contrastive loss for any summary pair of C_D is computed following [28,29], as:

$$L_{C_D} = \sum_{i=1}^{N} \sum_{j=i}^{N} \max(0, s_{c_j} - s_{c_i}) \tag{5}$$

Moreover, it is essential to make sure that the reference summary is ranked higher than all the candidate summaries:

$$L_{R_D} = \sum_{i=1}^{N} \max(0, s_{c_i} - s_{R_D}) \tag{6}$$

Finally, the loss function for extractive summarization, namely L_{ext}, is defined as follows:

$$L_{ext} = L_{C_D} + L_{R_D} + L_{CE} \tag{7}$$

Knowledge Transfer. Initially, the summarization model is pre-trained using a textual paraphrasing dataset. However, conventional PLM-based textual paraphrasing methods only employ the CE loss during the training phase. To facilitate better knowledge transfer, the training paradigm of textual paraphrasing must be aligned with that of extractive summarization. We define the calculated probability of the ground-truth paraphrasing label, denoted as p_g, as the positive sample, and $1 - p_g$ as the negative sample. We then use the contrastive loss to maximize the likelihood of the ground-truth sample. The contrastive loss, denoted as L_{CL}, the BCE loss, denoted as L_{BCE}, and the final loss function of textual paraphrasing, denoted as L_{para}, are defined as follows:

$$L_{CL} = \max(0, (1 - p_g) - p_g) \tag{8}$$

$$L_{BCE} = -l_t \log p_{eq} - (1 - l_t) \log(1 - p_{eq}) \tag{9}$$

$$L_{para} = L_{CL} + L_{BCE} \tag{10}$$

where l_t indicates the ground-truth label for the textual paraphrasing dataset. Subsequently, the pre-trained model is fine-tuned with the small-scale summarization dataset using the extractive loss, namely L_{ext}.

The whole training algorithm for ParaSum is listed in Algorithm 1. For a document D that contains L sentences, the number of its candidate summaries, denoted as N, is equal to $\binom{L}{I}$. This value is typically much larger than L, and computing ranking scores for all candidate summaries can be time-consuming. Therefore, it is necessary to exclude trivial candidate summaries in advance. To address this issue, we employ a straightforward yet effective heuristic approach [8] that utilizes a ranking-based method [3] to pre-select the top K sentences ($K \ll L$), and then enumerate all combinations of the K sentences according to I to generate candidate summaries. It should be noted that the ranking-based method [3] is trained using the same data as ParaSum.

Algorithm 1: Training Paradigm of ParaSum

Input: learning rate α; number of epochs for transfer learning and summarization M,N; number of training steps per epoch for transfer learning and summarization Q, S

Initialize parameters of the ranking model, namely, W and b;
for $i \leftarrow 1$ **to** M **do**
  for $q \leftarrow 1$ **to** Q **do**
    $W_{i,q},\ b_{i,q},\ \mathcal{M}_{i,q} \leftarrow$
    $W_{i,q-1} + \alpha_{i,q}\nabla_W L_{para},\ b_{i,q-1} + \alpha_{i,q}\nabla_b L_{para},\ \mathcal{M}_{i,q-1} + \alpha_{i,q}\nabla_{\mathcal{M}} L_{para}$
    update learning rate $\alpha_{i,q}$
  end
end
for $j \leftarrow 1$ **to** N **do**
  for $s \leftarrow 1$ **to** S **do**
    $W_{j,s},\ b_{j,s},\ \mathcal{M}_{j,s} \leftarrow$
    $W_{j,s-1} + \alpha_{j,s}\nabla_W L_{ext},\ b_{j,s-1} + \alpha_{j,s}\nabla_b L_{ext},\ \mathcal{M}_{j,s-1} + \alpha_{j,s}\nabla_{\mathcal{M}} L_{ext}$
    update learning rate $\alpha_{j,s}$
  end
end

4 Experiments

4.1 Experimental Settings

Summarization Datasets. Three widely-used datasets are adopted for evaluation, including two mainstream English news datasets and one Chinese news dataset:

- CNN/DailyMail [11] consists of news articles and corresponding multi sentences summaries from news stories in CNN and Daily Mail websites.
- Xsum [12] is more abstractive with only single-sentence summaries that answers the question "What is the article about?".
- CNewSum [30] is a high-quality Chinese news dataset collected from hundreds of thousands of news publishers with human-authored summaries.

Paraphrasing Datasets. The present study reports on the adoption of two commonly utilized textual paraphrasing datasets for the purpose of transfer-learning in the context of English and Chinese text summarization tasks:

- Quora Question Pairs (QQP) [31] consists of over 400,000 question pairs, and each question pair is annotated with a binary value indicating whether the two questions are paraphrases of each other.
- LCQMC [32] consists of Chinese question pairs with manual annotations.

Evaluation Metrics. The model performance on English datasets i.e., Xsum and CNN/DailyMail, are evaluated on `ROUGE` [26,33]. We compute the ROUGE

Table 1. ROUGE evaluations and their average on CNN/DailyMail dataset with 200, 500, 1000, 2000 training samples. Here, R-1/2/L stands for ROUGE-1/2/L, ParaSum^{-p} for ParaSum w/o para. Best results are in bold, and second best results are underlined.

Model	200			500			1000			2000		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
rnn-ext+rl [4]	40.13	17.73	36.36	40.20	17.76	36.42	40.23	17.76	36.46	40.30	17.77	36.51
MemSum [5]	35.30	15.10	32.44	37.05	16.10	33.87	37.11	16.12	33.92	39.14	17.45	35.65
PNBert [9]	40.18	17.58	36.32	40.42	17.95	36.60	40.54	17.96	36.69	40.56	18.06	36.80
BertSumExt [3]	40.34	17.41	36.57	40.45	17.51	36.68	40.48	17.59	36.72	40.49	17.56	36.72
MatchSum [8]	39.75	17.09	36.04	40.50	17.79	36.79	40.98	18.10	37.23	41.45	18.56	37.68
ParaSum^{-p}	40.49	17.53	36.56	41.10	18.26	37.22	41.56	18.44	37.63	41.58	18.51	37.62
ParaSum	**40.81**	**17.78**	**36.94**	**41.28**	**18.31**	**37.42**	**41.76**	**18.70**	**37.86**	**41.86**	**18.62**	**37.91**

scores using the standard pyrouge package, and report the F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L. These metrics measure the uniform, bigram, and longest common subsequence overlapping between the candidate summaries generated by different methods and the corresponding reference summaries.

Baselines: rnn-ext+rl [4] is a reinforcement learning (RL) based auto-regressive summarization method; **PNBert** [9] is a RL-based extractive summarization method that utilizes BERT as its base model, with ROUGE-based rewards; **BertSumExt** [3] is a ranking-based method that performs binary classification on each sentence based on its contextual representation from BERT; **MatchSum** [8] is the current state-of-the-art method that formulates summarization as a text-matching task, and employs a two-stage approach based on BertSumExt [3]. **MemSum** [5] is a recently proposed method that is specifically designed for summarizing long documents. This approach views extractive summarization as a multi-step episodic Markov Decision Process (MDP) that is aware of the extraction history.

Implementation Details: For the CNewSum dataset, ParaSum, PNBert, BertSumExt, and MatchSum employ the "bert-base-chinese" version of BERT, while rnn-ext+rl utilizes Chinese word embeddings [34]. We set $K = 5$ and $I = 2, 3$ for CNN/DailyMail, $I = 1$ for Xsum, and $I = 2$ for CNewSum. The reported best experimental results are obtained through multiple rounds of experiments, with concurrent grid search for the number of QQP sentence pairs used in textual paraphrasing transfer learning.

4.2 Experimental Results

Results in Different Low-Resource Settings: We evaluate the performance of ParaSum and baseline methods under various low-resource settings. Table 1 presents ROUGE evaluations of ParaSum and baselines trained with 200, 500, 1000, and 2000 samples of CNN/DailyMail. ParaSum^{-p} refers to ParaSum without textual paraphrasing transfer learning, and it consistently outperforms the baselines on all ROUGE metrics when trained with 500 and 1000 samples. Additionally, it outperforms the baselines on the ROUGE-1 metric when trained with

Table 2. Ablation study on CNN/DailyMail dataset with 1000 training samples.

Model	R-1	R-2	R-L
Full implement.	41.76	18.70	37.86
w/o. Para	41.56 (−0.20)	18.44 (−0.26)	37.63 (−0.23)
w/o. CE	41.65 (−0.11)	18.59 (−0.11)	37.76 (−0.10)
w/o. CL	40.18 (−1.58)	17.45 (−1.25)	36.21 (−1.60)

200 and 2000 samples. The results of this study indicate that conventional summarization paradigms are optimized for scenarios where there is an abundance of training instances. Consequently, these paradigms may demonstrate suboptimal performance in low-resource settings. In contrast, the ParaSum approach reframes summarization tasks as textual paraphrasing, which shares similarities with BERT's NSP self-supervised task. As a result, ParaSum is better equipped to leverage pertinent knowledge from BERT, thereby enhancing its performance in the context of summarization. ParaSum exhibits further enhancements over ParaSum^{-p}, and it significantly outperforms the baseline models across all low-resource scenarios. These findings provide compelling evidence that the knowledge obtained from textual paraphrasing can effectively assist the model in distinguishing between salient and trivial summaries, while simultaneously reducing the method's reliance on a large number of training instances.

Ablation Studies: To evaluate the potential advantages of utilizing knowledge from textual paraphrasing for text summarization, we remove the textual paraphrasing transfer learning stage from ParaSum (referred to as ParaSum w/o. Para) and train the model from scratch using randomly initialized parameters. To further investigate the effectiveness of the contrastive learning paradigm in providing high-quality supervision signals, we conducted an additional experiment in which we removed either the cross-entropy loss (referred to as ParaSum w/o. CE) or the contrastive loss (referred to as ParaSum w/o. CL) from ParaSum during the training phase. Table 2 presents the ROUGE evaluations of ParaSum and its ablations, trained with 1000 samples of CNN/DailyMail. Our experimental results indicate that removing the contrastive loss from ParaSum (ParaSum w/o. CL) results in a significant decrease in performance. This finding aligns with the theoretical foundations of contrastive learning, which posit that it can effectively guide the model to differentiate between the most important summary among any given pair of candidate summaries, thus providing more supervised signals to the model. In contrast, removing the cross-entropy loss from ParaSum (ParaSum w/o. CE) leads to a modest reduction in model performance. This is likely because the cross-entropy loss encourages the model to assign a high probability to the most optimal summary, thereby enhancing the overall quality of the summarization output. Furthermore, removing the textual paraphrasing pre-training stage from ParaSum (ParaSum w/o. Para) results in a significant decrease in model performance. This finding provides compelling evidence that the knowledge acquired through textual paraphrasing can effectively aid ParaSum in distinguishing between crucial and non-essential summaries.

Fig. 2. Evaluations of ParaSum pre-trained with different amount of QQP sentence pairs and trained with 1000 samples of CNN/DailyMail. Avg stands for average results.

Impact of the Quantity of QQP Data: In this study, we carried out an evaluation to examine the effect of the number of QQP sentence pairs used in the transfer learning stage on ParaSum. Following the transfer learning stage, we trained ParaSum using 1000 samples of CNN/DailyMail. The results concerning the influence of varying the number of QQP sentence pairs on model performance are presented in Fig. 2. The experimental results reveal that ParaSum attains optimal performance when pre-trained using 10,000 Quora Question Pairs (QQP) sentence pairs in transfer learning stage. The performance of ParaSum tends to degrade as the number of QQP sentence pairs increases. This is likely due to the fact that an excessive number of QQP pairs can cause ParaSum to deviate from extractive summarization and overfit to the QQP dataset. Conversely, decreasing the quantity of QQP sentence pairs also leads to a decrease in performance, as the model may not acquire sufficient knowledge to improve extractive summarization.

Generalization Studies: To evaluate the generalizability of ParaSum, we conducted an assessment on two additional datasets: Xsum, which is an abstractive summarization dataset, and CNewSum, which is a high-quality Chinese summarization dataset. For this evaluation, we utilized a training set comprising of 1000 samples for each dataset. Table 3 presents the results of our evaluation of ParaSum and the baseline models on Xsum and CNewSum. Our results demonstrate that although ParaSum outperforms the baseline models on the Xsum dataset, it only exhibits a minimal improvement over MatchSum, and slightly underperforms MemSum on the ROUGE-2 evaluation metric. This finding may

Table 3. ROUGE evaluations on Xsum and CNewSum.

Model	Xsum			CNewSum		
	R-1	R-2	R-L	R-1	R-2	R-L
rnn-ext+rl [4]	19.70	2.41	14.85	30.12	17.28	25.02
MemSum [5]	20.32	**3.26**	15.84	29.26	14.99	23.91
PNBert [9]	21.06	3.06	15.87	33.18	18.22	27.85
BertSumExt [3]	20.01	2.61	15.16	31.17	17.00	26.25
MatchSum [8]	21.12	3.03	15.74	32.48	18.22	27.05
ParaSum^{-p}	20.97	3.04	15.72	33.09	18.46	27.47
ParaSum	**21.15**	3.08	**15.91**	**33.76**	**19.22**	**28.26**

Table 4. T-Test of ParaSum and baselines on Xsum.

Method Comparison	Xsum P-Value		
	R-1	R-2	R-L
ParaSum & rnn-ext+rl	1.22E-09	7.33E-11	4.87E-12
ParaSum & MemSum	4.92E-11	3.04E-08	1.50E-02
ParaSum & PNBert	8.61E-06	1.00E-03	6.00E-03
ParaSum & BertSumExt	5.94E-12	9.26E-09	1.66E-10
ParaSum & MatchSum	2.70E-02	2.01E-05	8.05E-07

be attributed to the fact that Xsum imposes strict limitations on the methods, allowing them to select only a single sentence to create the summary. As a result, the potential performance gains offered by ParaSum are limited. To further analyze the performance of ParaSum and the baseline models on Xsum, we conducted T-tests with a confidence level of 0.95, and the results are presented in Table 4. The results of the T-tests reveal that the P-Values of ParaSum and the baseline models on the ROUGE-1, ROUGE-2, and ROUGE-L metrics are below the threshold of 0.05. This suggests that the observed differences in performance are statistically significant, thereby providing further support for the superior performance of ParaSum over the baseline models. In contrast, evaluations on CNewSum dataset demonstrated that ParaSum significantly outperformed the baseline models. This finding emphasizes the importance of utilizing knowledge acquired from textual paraphrasing and pre-trained language models in low-resource scenarios, especially when summarizing a document whose reference summary comprising multiple sentences.

Case Study: Table 5 illustrates a use case of ParaSum and various baselines, with models trained using 1000 CNN/DailyMail samples. The first and second sentences extracted by BertExtSum contain similar content, whereas the third summary sentence deviates from the reference summary by providing additional details. In contrast, the third sentence extracted by MatchSum presents content conveyed in the first summary sentence, thus demonstrating a level of semantic redundancy. Based on the subjective analysis, the summary extracted by ParaSum is superior to the summaries generated by the baseline methods.

Table 5. Use case study of ParaSum and baselines, all of which are trained using 1000 samples from CNN/DailyMail.

Ref. Summary	1. american women look to celebrities for hair inspiration, often uneducated about the potential dangers of beauty procedures 2. many celebrities who wear weaves, such as beyonce, selena gomez and paris hilton, could be doing serious damage to their hair 3. jennifer aniston, sandra bullock and jennifer lopez were revealed as having the three most popular celebrity hairstyles
ParaSum	1. one in five american women are willing to undergo dangerous beauty treatments in order to achieve the ideal look, despite the risks that these procedures pose to their health. 2. the survey, conducted by beauty research organization lqs and associates, looked at the lengths 1,000 american women go to in order to enhance their appearances or copy a celebrity, and the potentially disastrous consequences they might face in doing so, including hair loss, skin swelling, and overly painful procedures.
BertExtSum	1. one in five american women are willing to undergo dangerous beauty treatments in order to achieve the ideal look, despite the risks that these procedures pose to their health. 2. according to a new study, while just over half of women worry about the long term damage of beauty treatments, nearly a fifth would still pursue a treatment to get the right look - even if it proved hazardous to their health. 3. seven per cent, meanwhile, have actually had allergic reactions.
MatchSum	1. according to a new study, while just over half of women worry about the long term damage of beauty treatments, nearly a fifth would still pursue a treatment to get the right look - even if it proved hazardous to their health. 2. the survey, conducted by beauty research organization lqs and associates, looked at the lengths 1,000 american women go to in order to enhance their appearances or copy a celebrity, and the potentially disastrous consequences they might face in doing so, including hair loss, skin swelling, and overly painful procedures. 3. the cost of beauty: women often do n't realize the dangers of salon treatments before sitting in the styling chair

5 Conclusion

To address the issue of limited supervised signals arising from small-scale training datasets, this paper proposes a novel paradigm for extractive summarization called ParaSum, which is tailored for low-resource scenarios. ParaSum reframes extractive summarization as textual paraphrasing between the candidate summaries and the document. This approach helps to reduce the training gap between the summarization model and PLMs, thereby enabling the effective retrieval of relevant knowledge from PLMs to enhance summarization performance. In addition, ParaSum utilizes the knowledge acquired from textual paraphrasing to guide the summarization model in distinguishing high-quality summaries among the candidate summaries. Furthermore, ParaSum takes advantage

of contrastive learning to provide additional supervised signals for model training. The experimental results indicate that ParaSum consistently outperforms conventional summarization paradigms in low-resource scenarios.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 62202170 and Alibaba Group through the Alibaba Innovation Research Program.

References

1. Xu, J., Gan, Z., Cheng, Y., Liu, J.: Discourse-aware neural extractive text summarization. In: ACL (2020)
2. Quatra, M., Cagliero, L.: End-to-end training for financial report summarization. In: COLING, pp. 118–123 (2020)
3. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: EMNLP-IJCNLP, pp. 3730–3740 (2019)
4. Chen, Y.-C., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. In: ACL (2018)
5. Gu, N., Ash, E., Hahnloser, R.: MemSum: extractive summarization of long documents using multi-step episodic Markov decision processes. In: ACL, Ireland, Dublin, pp. 6507–6522 (2022)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2019)
7. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: a robustly optimized BERT pretraining approach, arXiv, vol. abs/1907.11692 (2019)
8. Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., Huang, X.: Extractive summarization as text matching. In: ACL, pp. 6197–6208 (2020)
9. Zhong, M., Liu, P., Wang, D., Qiu, X., Huang, X.: Searching for effective neural extractive summarization: what works and what's next. In: ACL, pp. 1049–1058 (2019)
10. Schick, T., Schütze, H.: It's not just size that matters: small language models are also few-shot learners. In: NAACL, pp. 2339–2352 (2021)
11. Hermann, K.M., et al.: Teaching machines to read and comprehend. In: NeuralIPS, pp. 1693–1701 (2015)
12. Narayan, S., Cohen, S.B., Lapata, M.: Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In: EMNLP (2018)
13. Chen, K., Fu, G., Chen, Q., Hu, B.: A large-scale Chinese long-text extractive summarization corpus. In: ICASSP, pp. 7828–7832 (2021)
14. Shafiq, N., et al.: Abstractive text summarization of low-resourced languages using deep learning. PeerJ Comput. Sci. **9**, e1176 (2023)
15. Chen, Y.-S., Song, Y.-Z., Shuai, H.-H.: SPEC: summary preference decomposition for low-resource abstractive summarization. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 603–618 (2022)
16. Huh, T., Ko, Y.: Lightweight meta-learning for low-resource abstractive summarization. In: SIGIR, pp. 2629–2633 (2022)
17. Zaken, E.B., Ravfogel, S., Goldberg, Y.: BitFit: simple parameter-efficient fine-tuning for transformer-based masked language-models. In: ACL, pp. 1–9 (2022)

18. Song, H., Dong, L., Zhang, W., Liu, T., Wei, F.: CLIP models are few-shot learners: empirical studies on VQA and visual entailment. In: ACL, pp. 6088–6100 (2022)
19. Wang, S., Fang, H., Khabsa, M., Mao, H., Ma, H.: Entailment as few-shot learner, CoRR (2021)
20. Gao, P., et al.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters, arXiv (2021)
21. Zhang, R., et al.: Tip-adapter: training-free adaption of CLIP for few-shot classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13695, pp. 493–510. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19833-5_29
22. Houlsby, N., et al.: Parameter-efficient transfer learning for NLP. In: ICML, pp. 2790–2799 (2019)
23. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**, 2096-2030 (2016)
24. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NeurIPS (2014)
25. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: ACL, pp. 3816–3830 (2021)
26. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
27. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)
28. Liu, Y., Liu, P.: SimCLS: a simple framework for contrastive learning of abstractive summarization. In: ACL, pp. 1065–1072 (2021)
29. Liu, Y., Liu, P., Radev, D., Neubig, G.: BRIO: bringing order to abstractive summarization. In: ACL, pp. 2890–2903 (2022)
30. Wang, D., Chen, J., Wu, X., Zhou, H., Li, L.: CNewSum: a large-scale summarization dataset with human-annotated adequacy and deducibility level. In: Wang, L., Feng, Y., Hong, Yu., He, R. (eds.) NLPCC 2021. LNCS (LNAI), vol. 13028, pp. 389–400. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88480-2_31
31. Sharma, L., Graesser, L., Nangia, N., Evci, U.: Natural language understanding with the quora question pairs dataset, arXiv (2019)
32. Liu, X., et al.: LCQMC: a large-scale Chinese question matching corpus. In: COLING, pp. 1952–1962 (2018)
33. Hu, B., Chen, Q., Zhu, F.: LCSTS: a large scale Chinese short text summarization dataset. In: EMNLP, pp. 1967–1972 (2015)
34. Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., Du, X.: Analogical reasoning on Chinese morphological and semantic relations. In: ACL, pp. 138–143 (2018)