GECSum: Generative Evaluation-Driven Sequence Level Contrastive Learning for Abstractive Summarization

Jiawen Xie[1], **Shaoting Zhang**[2], **Xiaofan Zhang**[1,2,*]
[1]Shanghai Jiao Tong University, China
[2]Shanghai AI Laboratory, China

Abstract

While dominant in abstractive summarization, transformer-based language models with the standard maximum likelihood estimation (MLE) training remain challenged by two discrepancies: the misalignment between token-level training and sequence-level evaluation, and the divergence between teacher-forcing training manner and auto-regressive generation behavior. Recent studies have shown that sequence-level contrastive learning, which utilizes the quality differences between multiple summaries as prior information, can effectively mitigate these issues. However, as certain evaluation metrics often determine the contrastive signals in existing methods, this leads to the model performance aligning with the preferences of these metrics being limited by the evaluation capabilities of these metrics. Inspired by prior works that treat the evaluation of generated text as a text generation problem, we propose a generative evaluation-driven contrastive learning framework, which leverages the semantic understanding capabilities of the abstractive model itself to evaluate summary in reference-based settings. In this way, our method establishes a connection between the model's reference-based evaluation and reference-free generation scenarios, allowing them to share the benefits of model capability enhancements. Extensive experiments on four summarization datasets demonstrate that our method outperforms the previous state-of-the-art regarding comprehensive performance. Various empirical analyses further substantiate the effectiveness of our method.

Keywords: abstractive summarization, transformer, contrastive learning, generative evaluation

1. Introduction

Abstractive text summarization (Gupta and Gupta, 2019) is a technique in natural language processing (NLP) that involves generating a summary of a source document by creating new sentences and phrases. Unlike extractive summarization, which selects salient sentences from the original document, abstractive summarization aims to produce a more human-like summary that maintains the context and coherence of the original document (Liu and Lapata, 2019; Raffel et al., 2020a). In recent years, pre-trained language models (PLMs) based on the transformer architecture have become the cornerstone of abstractive summarization systems (Bao et al., 2020). These models, such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020), can generate summaries more similar to those produced by humans in aspects such as coherence and readability, making them a promising tool for abstractive summarization.

However, it is well-known that the standard MLE training objective for abstractive summarization often suffers from two primary discrepancies between the training and inference stages. The first discrepancy lies in the incongruity between the token-level training objective (i.e., cross-entropy loss) and the sequence-level evaluation criteria like ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020). This divergence might lead to a scenario where a model excels in its training phase (evident by a minimal cross-entropy loss) yet underperforms when evaluated based on certain metrics (such as garnering a low ROUGE score). The second discrepancy is the discordance between the teacher-forcing training manner (Lamb et al., 2016) and auto-regressive generation behavior. This discrepancy can lead to compounding errors during inference because the model is not exposed to its errors during training, which is also known as *exposure bias* (Bengio et al., 2015; Ranzato et al., 2015).

To alleviate these discrepancies in abstractive summarization, various training methods beyond MLE from different aspects have been proposed. Among these, sentence-level training intends to optimize the abstractive model based on the sequence-level prior information about summary quality from certain evaluation criteria in reference-free or reference-based scenarios (Paulus et al., 2018; Stiennon et al., 2020). More recently, contrastive learning has been introduced into the training process of abstractive models, which substantially enhances the model's ability by requiring the model to differentiate the positive (good) and negative (bad) summaries (Liu et al., 2022; Xie et al., 2023). Nonetheless, in the vast majority of works, the quality of different summaries is often adjudicated by external evaluation metrics. This often results in the summaries produced by the abstractive models utilizing contrastive learning skewing towards the characteristics of these evaluation mea-

*Corresponding author. xiaofan.zhang@sjtu.edu.cn

LREC-COLING 2024, pages 7581–7595
20-25 May, 2024.

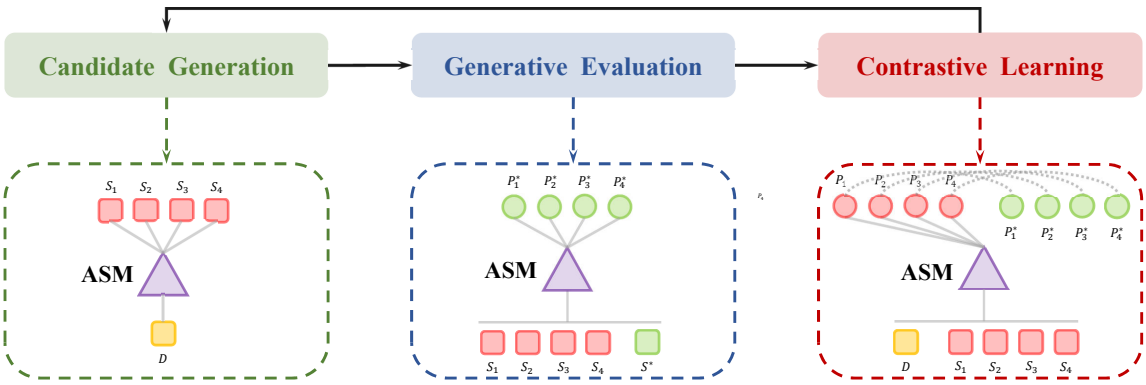

Figure 1: The iterative three stages of the GECSum framework. First, given a source document D, we use the abstractive summarization model (ASM) to collect diverse candidate summaries S_i. Then, through the ASM, we take the candidate summaries S_i as inputs and regard the output probabilities P_i^* assigned to the reference summary S^* as the quality scores of the candidate summaries S_i. Finally, based on P_i^*, we cast sequence-level contrastive learning into the ASM under reference-free conditions. Specifically, we take the source document D as input and require that the order of output probabilities P_i assigned to the candidate summaries S_i by the ASM is consistent with the order of quality scores P_i^*.

sures (Liu et al., 2023). Consequently, the genuine performance of the model gets tethered to the assessment capabilities of the metrics themselves.

Recent research has discovered that the evaluation of generated summaries can be naturally modeled as a text generation problem using pre-trained sequence-to-sequence (Seq2Seq) models. The underlying idea is that models, trained to transmute the generated text to/from a reference output or the source text, will procure superior scores when the generated text is better (Yuan et al., 2021). Inspired by these works, in this paper we introduce GECSum, a generative evaluation-driven sequence-level contrastive learning framework that establishes a connection between the abstractive model's reference-based evaluation and reference-free generation scenarios, allowing them to share the benefits of model capability enhancements. Figure 1 illustrates the three main operations in our approach: candidate generation, generative evaluation, and contrastive learning. In more detail, during the candidate summary generation phase, we employ the current abstractive summarization model to acquire diverse summaries (candidate summaries) for each source text in a reference-free context. Subsequently, we require certain evaluation metrics to measure the quality scores of these candidate summaries. Especially, instead of using external evaluation tools, in our experimental setup, we leverage the generative mechanism of the abstractive model itself to collect the sequence-level quality scores of these candidate summaries in a reference-based scenario. Finally, based on the differences in quality scores, we introduce sequence-level contrastive learning, requiring the order of probability scores estimated by the abstractive model in a reference-free scenario to be consistent with the order of prior quality scores. Additionally, to preserve the reasonable generation capability of the abstractive model and fully utilize available data resources for generalization, we supplement the MLE training objectives for reference summaries and high-quality candidate summaries.

Extensive experiments on the CNN/DailyMail, XSum, SAMSum, and MeQSum datasets show that GECSum can surpass the prior state-of-the-art in terms of comprehensive performance. Moreover, further qualitative and quantitative analyses substantiate that our method is capable of generating summaries of superior quality.[1]

2. Preliminary

Language Modeling In the training for text generation tasks, there are a series of token decisions in an auto-regressive manner. This process is expressed as a multiplication of decision probabilities that correspond to specific tokens. Given an input sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{|\boldsymbol{X}|})$ and its corresponding output $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{|\boldsymbol{Y}|})$, we model the following conditional probability:

$$P_\theta(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{|\boldsymbol{Y}|} p(\boldsymbol{y}_i|\boldsymbol{Y}_{<i}, \boldsymbol{X}; \theta), \quad (1)$$

where $\boldsymbol{Y}_{<i}$ represents the prefix-sequence in $\boldsymbol{Y}$ before $\boldsymbol{y}_i$, $\boldsymbol{Y}_{<1}$ is a special begin-of-sequence token, $|\boldsymbol{X}|$ and $|\boldsymbol{Y}|$ stand for the number of tokens of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, θ indicates the parameters of the auto-regressive model. Therefore, the training objective can be transformed to minimize the

[1]Our code is available at https://github.com/xjw-nlp/GECSum

following negative log-likelihood loss (NLL):

$$\begin{aligned}\mathcal{L}^{(h)}_{\boldsymbol{X},\boldsymbol{Y}}(\theta) &= -\frac{\log P_\theta(\boldsymbol{Y}|\boldsymbol{X})}{|\boldsymbol{Y}|} \\ &= -\frac{\sum_{i=1}^{|\boldsymbol{Y}|}\log p(\boldsymbol{y}_i|\boldsymbol{Y}_{<i},\boldsymbol{X};\theta)}{|\boldsymbol{Y}|},\end{aligned} \tag{2}$$

where the model is required to maximize the conditional log-likelihood of the tokens in a given reference output.

Label Smoothing As the standard MLE loss in Equation 2 tends to assign all probabilities to the over-confident outputs that are difficult to arise during inference, to regularize the model for generalization, we adjust the generative objective to consider further the conditional probabilities of all possible output $\hat{\boldsymbol{y}}$ in the vocabulary space:

$$\mathcal{L}^{(s)}_{\boldsymbol{X},\boldsymbol{Y}}(\theta) = -\frac{\sum_{i=1}^{|\boldsymbol{Y}|}\sum_{j=1}^{\mathcal{V}} w_{ij}\log p(\hat{\boldsymbol{y}}_{ij}|\boldsymbol{Y}_{<i},\boldsymbol{X};\theta)}{|\boldsymbol{Y}|}, \tag{3}$$

where $\mathcal{V}$ represents the vocabulary size. Specially, w_{ij} is the conditional weight for label smoothing:

$$w_{ij} = \begin{cases} 1-\alpha, & \hat{\boldsymbol{y}}_{ij} = \boldsymbol{y}_i; \\ \frac{\alpha}{\mathcal{V}-1}, & \hat{\boldsymbol{y}}_{ij} \neq \boldsymbol{y}_i, \end{cases} \tag{4}$$

where α adjusts the probabilities assigned to tokens not in $\boldsymbol{Y}$.

3. Methodology

For the abstractive text summarization, given the parameterized abstractive summarization model $\mathcal{G}_\theta$ that has undergone the task-specific fine-tuning, the source document $\boldsymbol{D}$ and corresponding reference summary $\boldsymbol{S}^*$, we first use the abstractive summarization model with diverse beam search (Vijayakumar et al., 2018) in an auto-regressive manner to collect a set of candidate summaries $\{\boldsymbol{S}^c_i\}^k_{i=1}$ including k candidate summaries.

$$\{\boldsymbol{S}^c_i\}^k_{i=1} \leftarrow \mathcal{G}_\theta(\boldsymbol{D}). \tag{5}$$

3.1. Generative Evaluation

Instead of using certain external evaluation metrics, we utilize the abstractive summarization model to gauge the quality scores of candidate summaries as prior information for the subsequent contrastive learning process. Specifically, similar to Yuan et al. (2021), we acquire the candidate summary quality scores $\mathcal{M}_{\mathcal{G}_\theta}$ given the reference summary $\boldsymbol{S}^*$ via Equation 1:

$$\mathcal{M}_{\mathcal{G}_\theta}(\boldsymbol{S}^c_i, \boldsymbol{S}^*) = P_{\mathcal{G}_\theta}(\boldsymbol{S}^*|\boldsymbol{S}^c_i), \tag{6}$$

where $\boldsymbol{S}^c_i$ is a generated summary in the candidate summary set $\{\boldsymbol{S}^c_i\}^k_{i=1}$.

3.2. Sequence-level Contrastive Learning

To simplify the subsequent formula expression, we assume $\mathcal{M}_{\mathcal{G}_\theta}(\boldsymbol{S}^c_i, \boldsymbol{S}^*) > \mathcal{M}_{\mathcal{G}_\theta}(\boldsymbol{S}^c_j, \boldsymbol{S}^*), \forall \boldsymbol{S}^c_i, \boldsymbol{S}^c_j \in \{\boldsymbol{S}^c_i\}^k_{i=1}, i < j$. In the reference-free scenario, we introduce a modification to Equation 1. Given the abstractive text summarization model $\mathcal{G}_\theta$, the model-predicted probabilities $\hat{\mathcal{M}}_{\mathcal{G}_\theta}$ of the candidate summary $\boldsymbol{S}^c_i$ with respect to the source document $\boldsymbol{D}$ can be defined as follows:

$$\hat{\mathcal{M}}_{\mathcal{G}_\theta}(\boldsymbol{D}, \boldsymbol{S}^c_i) = \frac{\log P_{\mathcal{G}_\theta}(\boldsymbol{S}^c_i|\boldsymbol{D})}{|\boldsymbol{S}^c_i|^\beta}, \tag{7}$$

where the hyper-parameter β controls the degree of length penalty (Wu et al., 2016). Accordingly, we note that the reference-free model-predicted probability ranges from $-\infty$ to 0. Finally, we formulate the contrastive learning objective below:

$$\begin{aligned}\mathcal{L}_{ctr} = \sum_{i=1}^{k}\sum_{j=i+1}^{k} \max\{&\hat{\mathcal{M}}_{\mathcal{G}_\theta}(\boldsymbol{D}, \boldsymbol{S}^c_j) - \\ &\hat{\mathcal{M}}_{\mathcal{G}_\theta}(\boldsymbol{D}, \boldsymbol{S}^c_i) + (j-i)\times\lambda, 0\}\end{aligned} \tag{8}$$

where λ is the threshold judging whether the difference of model-predicted probabilities of diverse candidate summaries engages in backpropagation.

3.3. Pseudo-Summary for Weighted Generative Objective

Theoretically, through the generative evaluation and sequence-level contrastive learning above, the reference-free predictive probability and reference-based evaluated quality score could approximately present the following relationship:

$$P_{\mathcal{G}_\theta}(\boldsymbol{S}^c_i|\boldsymbol{D})^{\frac{1}{|\boldsymbol{S}^c_i|^\beta}} \propto P_{\mathcal{G}_\theta}(\boldsymbol{S}^*|\boldsymbol{S}^c_i). \tag{9}$$

Considering that previous work has demonstrated the importance of not only the order but also the magnitude of predictive probabilities in contrastive learning, we additionally use the pseudo-summary S^c_1 with the highest evaluated quality score for the generative optimization. Through this operation, we simultaneously maintain the probabilities of high-quality summaries at a relatively elevated level and expand the model's perception of states during the training phase, thereby enhancing the model's generalization capabilities.

Comprehensively, the generative objectives of the reference summary and pseudo-summary are combined with the contrastive learning objective into a universe loss function (Edunov et al., 2018):

$$\mathcal{L}_{all} = \mathcal{L}^{(s)}_{\boldsymbol{D},\boldsymbol{S}^*} + P_{\mathcal{G}_\theta}(\boldsymbol{S}^*|\boldsymbol{S}^c_1)^\eta \mathcal{L}^{(h)}_{\boldsymbol{D},\boldsymbol{S}^c_1} + \gamma\mathcal{L}_{ctr}, \tag{10}$$

where the hyper-parameter η adjusts the degree to which the quality score $P_{\mathcal{G}_\theta}(\boldsymbol{S}^*|\boldsymbol{S}^c_1)$ of the pseudo-summary $\boldsymbol{S}^c_1$ affects its generative objective.

Dataset	Train	Dev	Test	Length
CNNDM	287,084	13.367	11,489	766.1/58.2
XSum	203,028	11,273	11,332	430.2/23.3
SAMSum	14,732	818	819	156.7/25.5
MeQSum	400	100	500	60.8/10.1

Table 1: Statistics of four abstractive text summarization datasets. Length indicates the average number of tokens.

4. Experiment

In this section, we elaborate on the datasets, baselines, implementation details, and evaluation methods in our experiments.

4.1. Datasets

In our settings, we conduct the comparison experiments on four single document summarization datasets including CNNDM (Hermann et al., 2015; Nallapati et al., 2016), XSum (Narayan et al., 2018), SAMSum (Gliwa et al., 2019), and MeQSum (Ben Abacha and Demner-Fushman, 2019).

CNNDM[2] is a widely-used abstractive text summarization dataset consisting of the news articles and associated highlights as summaries.

XSum[3] is an abstractive dataset within the realm of the news domain, containing one-sentence summaries for assessing abstractive single-document summarization systems.

SAMSum[4] comprises approximately messenger-style English conversations and corresponding summaries that provide a succinct third-person overview of the conversation's content.

MeQSum[5] is a medical question summarization dataset. The summaries in MeqSum are written by medical experts in a formal style.

More detailed information about these datasets is shown in Table 1.

4.2. Baselines

We aim to compare our experimental results with prior studies that have demonstrated exceptional performance. Specifically, **BART** (Lewis et al., 2020) is a highly regarded large-scale pre-trained language model that excels in sequence generation tasks. In contrast, **PEGASUS** (Zhang et al., 2020) is designed with a unique pre-training objective tailored for abstractive text summarization. The **GSum** framework (Dou et al., 2021) stands out for its ability to effectively integrate various forms of external guidance. **ConSum** (Sun and Li, 2021) tackles the problem of *exposure bias* by decreasing the likelihood of low-quality summaries and increasing that of reference summaries. **SeqCo** (Xu et al., 2022) views the document, its reference summary, and its candidate summaries as different perspectives of the same mean representation, and maximizes their similarities during training. **GOLD** (Pang and He, 2021) utilizes off-policy learning from demonstrations for generation. **SimCLS** (Liu and Liu, 2021)improves the text generation process with a two-stage method rooted in contrastive learning. **SummaReranker** (Ravaut et al., 2022) l learns to choose a high-quality summary from a pool of candidate summaries by applying reranking to a second-stage model. **BRIO** (Liu et al., 2022) presents a novel paradigm that assumes non-deterministic distributions instead of the deterministic distribution of the gold summary. **SimMCS** (Xie et al., 2023) introduces the new contrastive signal and tailored attention mechanism to alleviate *exposure bias* further. **D-HGN** (Xiachong et al., 2021) injects the utterance and commonsense knowledge to facilitate dialogue understanding and summary generation. **S-BART** (Chen and Yang, 2021) utilizes structured information from discourse relation graphs and action graphs. **ProphetNet** (Yadav et al., 2021) acquires two rewards from question-type identification and question-focus recognition to optimize a reinforcement learning-based model. **MTL-DA** (Mrini et al., 2021) uses a novel multi-task learning method with data augmentation for medical questions. **QFCL** (Zhang et al., 2022) generates hard negative samples based on the question focus and exploits contrastive learning to obtain better sentence-level representations.

4.3. Implementation Details

Our Implementation is based on *PyTorch* (Paszke et al., 2019) and *Transformers* (Wolf et al., 2020) libraries. We conducted all the comparison experiments using 4 NVIDIA RTX 3090 GPUs.

Backbone Settings For a fair comparison, we uniformly use the full-attention PLM $\text{BART}_{\text{large}}$[6] with 12 layers each for the encoder and decoder as our backbone on the CNNDM, SAMSum, and MeQSum datasets. For the XSum dataset, we employ $\text{PEGASUS}_{\text{large}}$[7] as our backbone, which has 16 encoder layers and 16 decoder layers. In particular, the hidden size of both of them is 1024, which is converted into 16 attention heads with a hidden unit size of 64 for multi-head attention.

Task-specific Fine-tuning As we initialize some

[2] https://cs.nyu.edu/ kcho/DMQA/

[3] https://github.com/EdinburghNLP/XSum

[4] https://huggingface.co/datasets/samsum

[5] https://github.com/abachaa/MeQSum

[6] The checkpoint of BART is "facebook/bart-large" containing around 406M parameters, whose maximum encoding length is 1024.

[7] The checkpoint of PEGASUS is "google/pegasus-xsum" fine-tuned with XSum containing around 568M parameters, whose maximum encoding length is 512.

Model	R-1	R-2	R-L	BS	BaS
BART	44.16	21.28	40.90	87.95	-3.91
PEGASUS	44.17	21.47	41.11	88.13	-3.83
GSum	45.94	22.32	42.48	-	-
ConSum	44.53	21.54	41.57	-	-
SeqCo	45.02	21.80	41.75	-	-
GOLD-p	45.40	22.01	42.25	-	-
GOLD-s	44.82	22.09	41.81	-	-
SimCLS	46.67	22.15	43.54	66.14	-
SummaReranker	47.16	22.55	43.87	87.74	-
BRIO-Ctr	47.28	22.93	44.15	-	-
BRIO-Mul	47.78	23.55	44.57	-	-
SimMCS	48.38	24.17	44.79	89.31	-3.50
GECSum	**48.46**	**24.48**	**45.11**†	**90.23**†	**-2.97**†

Table 2: Average results on CNNDM test set. R-1/2/L indicates the ROUGE-1/2/L F1 score. BS and BaS refer to the neural model-based metrics BERTScore and BARTScore respectively. †: significantly better than the baseline model ($p < 0.05$). The best results are bolded.

Model	R-1	R-2	R-L	BS	BaS
BART	45.14	22.27	37.25	88.17	-4.05
PEGASUS	47.21	24.56	39.25	89.68	-3.89
GSum	45.40	21.89	36.67	-	-
ConSum	47.34	24.67	39.40	-	-
SeqCo	45.65	22.41	37.04	-	-
GOLD-p	45.75	22.26	37.30	-	-
GOLD-s	45.85	22.58	37.65	-	-
SimCLS	47.61	24.57	39.44	69.81	-
SummaReranker	48.12	24.95	40.00	**92.14**	-
BRIO-Ctr	48.13	25.13	39.84	-	-
BRIO-Mul	49.07	25.59	40.40	-	-
SimMCS	**49.48**	25.77	40.52	90.31	-3.73
GECSum	48.98	**25.91**	**41.50**†	91.56	**-2.80**†

Table 3: Average results on XSum test set. R-1/2/L indicates the ROUGE-1/2/L F1 score. BS and BaS refer to the neural model-based metrics BERTScore and BARTScore respectively. †: significantly better than the baseline model ($p < 0.05$). The best results are bolded.

model parameters utilizing a pre-trained backbone that has not been subjected to fine-tuning, before the contrastive learning, we necessitate the fine-tuning of the model to ensure its adaptability to the specific distribution of a downstream dataset.

Candidate Summary Preparation During the candidate summary generation phase, we resample 3200 samples from the training set without replacement (for the MeQSum dataset, we utilize all training samples). Then, based on the source text in each sample, the fine-tuned abstractive model generates 16 candidate summaries using a diverse beam search. Subsequently, we employ the design in Equation 6 to obtain the quality scores for these candidate summaries.

Contrastive and Generative Training At the training stage, for each source text, we leverage its 16 candidate summaries for sequence-level contrastive learning. In addition, to maintain the reasonable generation ability of our abstractive model, we employ the best pseudo-summary and reference summary for generative training. All PLMs are trained using the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, along with a tailored learning rate scheduler using warmup.

Auto-regressive Inference During inference, the abstractive summarization model is a standard auto-regressive generation system and produces summaries using beam search (Wiseman and Rush, 2016) in an auto-regressive manner given the source text.

4.4. Evaluation

Following most previous works, for abstractive summarization tasks, we measure the quality of generated summaries using the popular metric ROUGE (Lin, 2004). On the test set of CNNDM, XSum, SAMSum, and MeQSum, we report full-length F1-based ROUGE-1, ROUGE-2, and ROUGE-L scores compared between the system outputs and the reference summaries with the standard ROUGE Perl package. Moreover, We also use two popular model-based semantic metrics BERTScore (Zhang* et al., 2020) and BARTScore (Yuan et al., 2021) to demonstrate the superiority of our approaches more comprehensively. As shown in Figure 2, ROUGE-1 and ROUGE-L are strongly correlated (Pearson correlation score of 0.951). In particular, the low Pearson correlation coefficient (< 0.7) between BARTScore and other metrics indicates a significant difference in preferences between this and others.

5. Discussion

5.1. Results

We report the model performance of the baselines, previous works from the literature, and our proposed approaches on four abstractive summarization datasets. We make the following observations: (1) Table 2 and Table 3 report the results over the test sets of CNNDM and XSum. We have discerned that casting the backbone $\text{BART}_{\text{large}}$ into our framework GECSum facilitates an all-encompassing transcendence over antecedent strong baselines within the CNNDM test set, as gauged by the five evaluation criteria. On the XSum test set with the backbone $\text{PEGASUS}_{\text{large}}$, apart from ROUGE-1, we have perceived patterns in the model's performance fluctuations across various metrics that bear

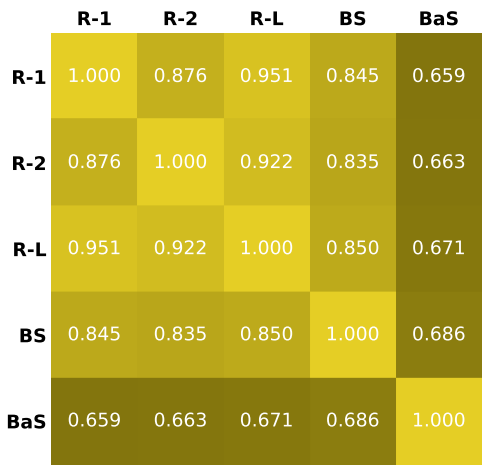

Figure 2: Pearson correlation coefficient between the five evaluation metrics (i.e., R-1, R-2, R-L, BS, BaS) for our GECSum with beam search on CNNDM test set. R-1/2/L denotes ROUGE-1/2/L. BS and BaS denote BERTScore and BARTScore.

Model	R-1	R-2	R-L	BS	BaS
BART	45.15	21.66	44.46	85.64	-3.88
D-HGN	42.03	18.07	39.56	-	-
S-BART	46.07	22.60	45.00	-	-
GECSum	**47.37**†	**23.42**†	**46.30**†	**88.10**†	**-3.25**†

Table 4: Average results on SAMSum test set. R-1/2/L indicates the ROUGE-1/2/L F1 score. BS and BaS refer to the neural model-based metrics BERTScore and BARTScore respectively. †: significantly better than the baseline model ($p < 0.05$). The best results are bolded.

a resemblance to those on the CNNDM test set. Particularly, we note that both BRIO and SimMCS utilize ROUGE-1 to procure sequence-level prior information for contrastive learning, thereby rendering their optimization more concentrated on ROUGE-1. Contrarily, our candidate summary quality assessment abstains from employing certain external metrics, while it exhibits superior performance in a comprehensive view, thereby underscoring the merits of our framework.

(2) Table 4 and Table 5 elucidate the performance outcomes of antecedent robust baselines juxtaposed with our approach on the test sets of SAMSum and MeQSum respectively. We observed that when employing the same backbone, our contrastive learning model outperforms earlier methods that boost performance through the incorporation of external knowledge. This not only underscores the superior performance of our model but also confirms its exceptional capacity for generalization. Hence, it's suitable for abstractive summarization tasks spanning a variety of domains.

(3) Astonishingly, we have discerned that GECSum

Model	R-1	R-2	R-L	BS	BaS
BART	46.17	29.50	44.80	87.23	-3.43
ProphetNet	45.52	27.54	48.19	-	-
MTL-DA	49.20	29.50	44.80	-	-
QFCL	51.48	34.16	49.08	-	-
GECSum	**52.28**†	**35.46**†	**50.02**†	**89.42**†	**-2.98**†

Table 5: Average results on MeQSum test set. R-1/2/L is the ROUGE-1/2/L F1 score. BS and BaS are the neural metrics BERTScore and BARTScore respectively. †: significantly better than the baseline model ($p < 0.05$). The best results are bolded.

achieves a substantial performance enhancement on the BARTScore relative to prior methods across all our summarization datasets. We postulate that this could potentially be attributed to the fact that the generative evaluation procedure inherent in our framework closely mirrors the evaluation mechanism of BARTScore, thereby aligning more congruently with its pattern of semantic comprehension.

5.2. Analyses

We further analyze the properties of our approach to gain more insight.

Ablation Study GECSum incorporates additional sequence-level contrastive learning (Ctr) and pseudo-summary MLE training (PST) procedures into the MLE training of the reference summary. To investigate the contributions of each component, we independently removed each one. Moreover, we concurrently eliminated both of them, thereby regressing it to the state of standard auto-regressive training with reference summary, to probe the efficacy of the combination of them. As demonstrated in Table 6, the performance experienced a significant decline when either Ctr or PST was removed from GECSum, underscoring their effectiveness.

Comparing w/o All and w/o Ctr, we find that adding PST to the MLE training of reference summary has a very slight impact on model performance. Nonetheless, subsequent to the incorporation of contrastive learning, comparing the outcomes of GECSum and w/o PST, we discerned that the addition of PST can markedly enhance the performance of the model. We conjecture that this may be the result of two factors working together. The first factor is that contrastive learning improves the quality of the pseudo-summary generated by the model. The second factor is that PST takes into consideration the magnitude of predictive probabilities in contrastive learning, thereby circumventing potential degradation in the model's generative capability as a consequence of contrastive learning.

Low-Resource Evaluation In real-world applications, the quantity of training samples available for downstream tasks is frequently quite constrained.

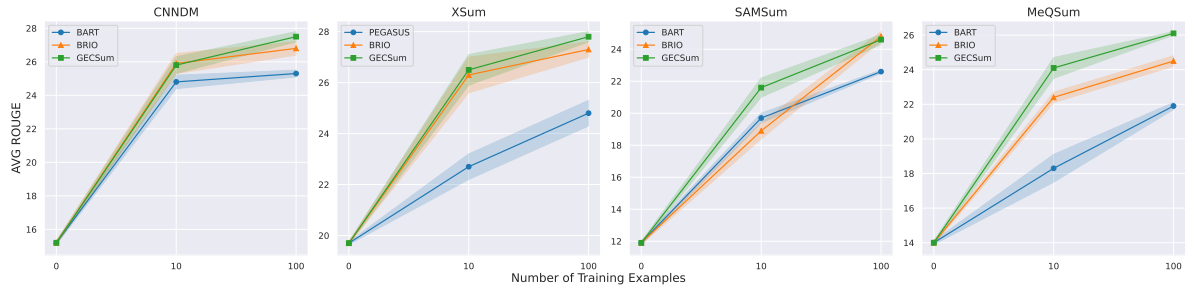

Figure 3: The AVG ROUGE scores (R-1, R-2, and R-L) of the pre-trained models with 0, 10, and 100 training examples with variance. All results are obtained by the average of 5 random runs with different seeds.

Model	SAMSum				MeQSum			
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
GECSum	48.21	24.31	47.08	88.96	53.20	36.85	50.77	90.50
w/o Ctr	46.36	21.33	45.32	86.40	47.04	30.13	45.40	87.97
w/o PST	47.71	24.05	46.97	88.62	52.88	36.70	50.53	90.07
w/o All	46.05	21.48	45.02	86.36	46.78	29.89	45.11	87.80
Δ_{Ctr}	$\uparrow$ 03.99%	$\uparrow$ 13.97%	$\uparrow$ 03.88%	$\uparrow$ 02.96%	$\uparrow$ 13.10%	$\uparrow$ 22.30%	$\uparrow$ 11.83%	$\uparrow$ 02.88%
Δ_{PST}	$\uparrow$ 01.05%	$\uparrow$ 01.08%	$\uparrow$ 00.23%	$\uparrow$ 00.38%	$\uparrow$ 00.61%	$\uparrow$ 00.41%	$\uparrow$ 00.47%	$\uparrow$ 00.48%
Δ_{All}	$\uparrow$ 04.69%	$\uparrow$ 13.18%	$\uparrow$ 04.58%	$\uparrow$ 03.01%	$\uparrow$ 13.72%	$\uparrow$ 23.29%	$\uparrow$ 12.55%	$\uparrow$ 03.08%

Table 6: Ablation study results on the development sets of SAMSum and MeQSum. Performance changes compared with the full model GECSum are reported. R-1/2/L is the ROUGE-1/2/L F1 score. BS and BaS refer to neural model-based semantic metrics BERTScore and BARTScore respectively.

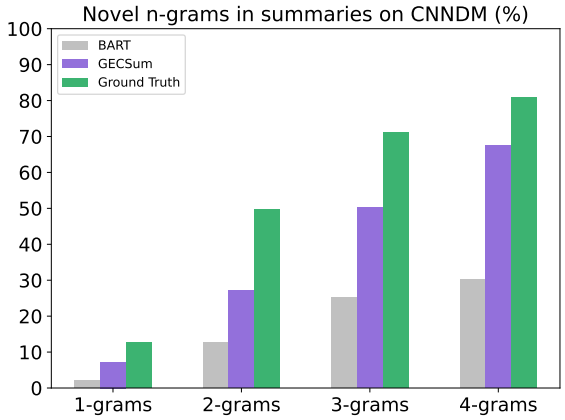

Figure 4: Novel n-grams with BART on CNNDM dataset. We use the beam search to generate summaries during inference.

Beams	BART		GECSum	
	R-L	BS	R-L	BS
4	**40.90**	**87.95**	45.11	90.23
10	40.76	87.84	45.23	90.32
20	40.79	87.80	45.55	90.42
50	40.31	87.65	45.70	90.48
100	40.10	87.38	**45.82**	**90.50**

Table 7: Results on CNNDM test set with different beam widths (the number of beams) used in beam search. The default beam width is 4. R-L and BS are the ROUGE-L F1 score and BERTScore F1 score respectively.

Consequently, the performance of the model under low-resource conditions is of paramount importance. In our setup, we randomly sample a few (10 and 100) training examples from specific datasets (CNNDM, XSum, SAMSum, and MeQSum) to adapt the PLMs for corresponding data distributions. The results in Figure 3 indicate that our model demonstrates superior sample efficiency in low-resource scenarios compared to previous robust baselines.

Increasing the Beam Width Intuitively, we often believe that in the case of auto-regressive generative models using beam search for inference, a larger beam width (i.e., the number of candidate sequences maintained during beam search) would enable more candidates for next token prediction, thus leading to better quality of the final generated sequence. However, since the probability scores assigned to sequences by auto-regressive models trained on MLE loss do not have a clear alignment with their evaluation metrics, the sequence that the model assigns the highest probability score is often

System	Summary
Reference (Case 1)	juan arango escaped punishment from the referee for biting jesus zavela. he could face a retrospective punishment for the incident. arango had earlier scored a free kick in his team 's 4 - 3 defeat.
BART	club tijuana star juan arango conjured memories luis suarez in his team 's 4 - 3 defeat by monterrey in the mexican league. the venezuelan icon arango sank his teeth into the shoulder of jesus zavela as his temper flared in the defeat. arango had earlier curled in a magnificent free kick for his team to bring them level after falling 2 - 0 down early on.
GECSum	juan arango bites the shoulder of jesus zavela in tijuana 's 4 - 3 defeat to monterrey in the mexican league. arango earlier scored a magnificent free kick to bring his side level. the tijuana star could face a retrospective ban for the bite.
Reference (Case 2)	new pictures show raheem sterling and jordon ibe with shisha pipes. the liverpool pair are dressed in casual clothing and have a pipe each. pictures emerged last week of liverpool star sterling smoking shisha. footage also emerged of him inhaling nitrous oxide from a balloon. the pictures create a fresh problem for liverpool boss brendan rodgers. the images will be a concern for any potential suitors of the liverpool star.
BART	raheem sterling and jordon ibe have been pictured smoking from a shisha pipe in the past few days. the images will be a concern to clubs considering parting with £ 40million to sign the star. becoming as good as lionel messi or cristiano ronaldo — gareth bale even — requires more than ability alone.
GECSum	liverpool players raheem sterling and jordon ibe have been pictured smoking a shisha pipe. sterling and ibe were also pictured drinking fanta. the images will be a concern to clubs considering £ 40million to sign the star. sterling was pictured smoking from a pipe in the past few days.
Reference (Case 3)	barcelona temporarily opened up a seven - point lead over real madrid at the top of la liga. argentina superstar lionel messi opened the scoring with a trademark curled finish after 33minutes. second - placed real madrid play rayo vallecano later on wednesday evening to close the seven - point gap. luis suarez doubled the catalan 's lead with a similarly curling left - footed stunner after the interval. barca defender marc bartra netted the third with a far post header from xavi 's whipped in cross. former liverpool striker suarez tapped in a late goal from pedro 's cross in injury time to complete the rout.
BART	lionel messi opened the scoring in the 33rd minute with a trademark curling strike. luis suarez doubled the lead with a left - footed strike in the 55th minute. defender marc bartra added a third and suarez scored his second with the last kick of the game. barcelona remain seven points clear at the top of la liga with eight games to play.
GECSum	barcelona beat almeria 4 - 0 at the nou camp to go seven points clear at the top of la liga. leo messi opened the scoring in the 33rd minute with a curling left - foot strike. luis suarez doubled the lead in the second half. defender marc bartra added a third and suarez scored his second in the final minute.

Figure 5: Examples of summaries generated by GECSum trained on CNNDM. The sentence in green is included in the GECSum summary, while the one in red is discarded.

Model	Base			GECSum		
	Info	Coher	Fact	Info	Coher	Fact
CNNDM	23.6	22.0	24.1	26.4	28.0	25.9
XSum	21.7	20.9	20.0	28.3	29.1	30.0
SAMSum	25.9	21.5	23.9	24.1	28.5	26.1
MeQSum	23.4	19.8	24.3	26.6	30.2	25.7

Table 8: Human evaluation of the summaries from GECSum and corresponding backbone (Base) respectively. The metric **Info** (informativeness) reflects the degree to which the summary encapsulates the source text, **Coher** (coherence) assesses whether the semantics of the summaries are coherent, and **Fact** indicates whether the statements in the summary can be found in the source text.

not optimal according to these metrics. Compared with the vanilla MLE training, introducing sequence-level contrastive learning often enables a better alignment between the model's predicted probabilities and evaluation metric scores. In Table 7, we test our abstractive summarization model with beam widths 4, 10, 20, 50, and 100. We observe a declining trend in the performance of BART as the beam width increases. In contrast, the performance of GECSum shows an improvement with the increase in beam width.

Abstractiveness We assess the abstractiveness of generated summaries by calculating the percentage of novel n-grams, which are defined as those that appear in the summary but not in the associated source document. As shown in Figure 4, our state-of-the-art model generates more abstractive summaries than the base model BART in terms of all used n-grams metrics. Additionally, Figure 5 demonstrates that the summaries generated by our model effectively convey salient information and are closer to the reference summaries.

Human Evaluation Besides the measurement of automatic evaluation metrics, we also quantitatively compare our method with its backbone through human preference. To elaborate, we randomly selected 50 source documents without replacement from the test set of each of the four datasets. For the same source document, we had both models generate summaries in the same auto-regressive manner, forming a pair. We recruited three graduates to select their preferred model output from a pair (unordered), based on each of the aspects of informativeness, coherence, and factuality. The average results in Table 8 indicate that the summaries generated by GECSum are more in line with human preferences.

Case Study on CNNDM To more clearly show the quality of generated summaries, we list three cases to compare our model with BART in Figure 5. In Case 1, the summary of GECSum almost covers all the key points in the reference summary. While the summary of BART also contains these points, it introduces additional information not present in the reference summary, such as the comparison to Luis Suarez and the detail about Tijuana falling 2-0 down early on. In Case 2, the summary of BART introduces additional information not present in the Reference Summary, such as the comparison to Lionel Messi, Cristiano Ronaldo, and Gareth Bale. In Case 3, the summary of BART lacks the detail about Barcelona beating Almeria 4-0, which is mentioned in the reference summary.

6. Related Works

Training Method of Sequence Generation Models Sequence generation tasks, such as machine translation (Stahlberg, 2020), text summarization (Widyassari et al., 2022), and dialogue systems (Ni et al., 2023), have been a central focus in the field of natural language processing. Various training methods have been proposed and developed to address these tasks. Traditionally, sequence generation models have been trained using standard maximum likelihood estimation (MLE) and its variants. This kind of method involves training a model to maximize the probability of the next token in the sequence given the previous tokens. However, it has some discrepancies between the training and inference to affect model performance severely. First, the MLE loss assumes a deterministic target distribution where an ideal model will assign all the probability mass to the reference summary, while during inference, the model needs to compare several system-generated summaries that have deviated from the reference summary (i.e., *exposure bias*). Second, to improve training efficiency, *teacher forcing* (Williams and Zipser, 1989) is usually applied to token-level cross-entropy loss, which mismatches the sequence-level evaluation criteria like ROUGE. To enhance the training method to be more suitable for inference situations, the Gumbel-Softmax trick (Jang et al., 2017) is another method used for training sequence generation models, This method allows for differentiable sampling, which makes it possible to backpropagate through the sampling operation. Adversarial training methods (Yu et al., 2017) involve training a generator network to fool a discriminator network. These methods have been used to encourage the generator to produce more realistic sequences. More recently, reinforcement learning (Shen et al., 2016; Ramamurthy et al., 2023; Ouyang et al., 2022) methods have been applied to sequence generation tasks. These methods involve training a model to maximize a reward signal, which can be designed to encourage desirable properties in the generated sequences.

Pre-trained Language Model for Abstractive Summarization In recent times, large sequence-to-sequence transformers, based on either encoder-decoder transformer or decoder-only transformer architectures, have shown promising performance in the field of natural language processing, including text summarization (Raffel et al., 2020a; Xie et al., 2024). These models are pre-trained using a variety of self-supervised objectives and fine-tuned with structured losses in downstream tasks. For example, BART (Lewis et al., 2020), a denoising auto-encoder, is pre-trained to reconstruct original text spans corrupted with an arbitrary noising function such as text infilling. PEGASUS (Zhang et al., 2020) is distinguished by its specifically tailored self-supervised pre-training objective for the summarization task. In PEGASUS, salient text spans are removed or masked from the original document, and the model aims to restore the remaining text spans to their original form. We use these models as backbones in our work.

Contrastive Learning Contrastive learning has been widely recognized as an effective method to enhance model performance by enabling the model to differentiate between the quality of various samples (Chuang et al., 2020; Cheng et al., 2023). Recently, this method has demonstrated promising performance in natural language generation tasks such as text summarization (Cao and Wang, 2021) and machine translation (Yang et al., 2019; Pan et al., 2021). These contrastive learning examples can be constructed using either rule-based or model-based methods. The latter is capable of producing text examples that are more akin to those generated by humans, thereby creating more natural contrastive schemes. On the other hand, contrastive learning can be performed in either latent or discrete space. For instance, Gao et al. (2021) introduces a contrastive learning framework into the representation of sentence embeddings and greatly advances state-of-the-art results. Liu et al. (2022) adopts the discriminative re-ranking over generated summaries in discrete space like other works (Shen et al., 2004; Och et al., 2004; Mizumoto and Matsumoto, 2016; Lee et al., 2021).

7. Conclusion

In this paper, we introduce a generative evaluation-driven contrastive learning framework for abstractive summarization, called GECSum. In contrast to previous contrastive learning methods, our framework does not rely on third-party evaluation metrics. Instead, it leverages the generative mechanism inherent in the abstractive model itself to obtain sequence-level signals of summary quality for contrastive training. Through this design, we have established a connection between reference-based evaluation and reference-free generation processes based on the same abstractive model, allowing both to share the benefits of enhanced model capabilities. Substantial experiment results and analyses demonstrate the satisfying effectiveness of GECSum. Besides, our method does not depend on any particular tasks or models, which have good generalization ability for various application scenarios.

8. Acknowledgement

We would like to express our deep gratitude to the anonymous reviewers for their insightful feedback and constructive suggestions.

9. Limitations

Since our sequence-level contrastive learning framework requires the model-predicted probability scores of candidate summaries given a source document during the training phase, there is an extremely large GPU memory footprint even if the batch size is small, which limits the scale of contrastive data and suppresses the potential of our method. Moreover, due to the alternating process of candidate summary updates and model training, the actual training time for our model tends to be relatively long. Meanwhile, due to the large search space of hyper-parameter combinations in the experiment, it is difficult to find the hyper-parameter setting that provides optimal model performance.

On the other hand, Like most existing work, our method primarily focuses on the performance of the abstractive summarization model, without paying excessive attention to controllable generation and hallucination of the model. This implies that the summaries generated by our model might face issues like information omission, redundancy, or the presentation of counterfactual information.

10. Ethical Considerations

While our work presents minimal risk, similar to established abstractive summarization systems, we can't ensure absolute factual consistency or complete absence of hallucination in the generated summaries. Hence, vigilance is essential when integrating our system into real-world projects.

11. References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775. Curran Associates, Inc.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long Papers), pages 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

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

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. 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 7250–7264, Online. 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.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023. On learning to summarize with large language models as references.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive 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 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Tomoya Mizumoto and Yuji Matsumoto. 2016. Discriminative reranking for grammatical error correction with statistical machine translation. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1133–1138, San Diego, California. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021. A gradually soft multi-task and data-augmented approach to medical question understanding. 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 1505–1515, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168, Boston, Massachusetts, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. 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 244–258, Online. Association for Computational Linguistics.

Richard Yuanzhe Pang and He He. 2021. Text generation by learning from demonstrations. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

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.

Chenze Shao, Xilin Chen, and Yang Feng. 2018. Greedy search with probabilistic n-gram matching for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4778–4784, Brussels, Belgium. Association for Computational Linguistics.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Shichao Sun and Wenjie Li. 2021. Alleviating exposure bias via contrastive learning for abstractive text summarization.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. 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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Feng Xiachong, Feng Xiaocheng, and Qin Bing. 2021. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 964–975, Huhhot, China. Chinese Information Processing Society of China.

Jiawen Xie, Pengyu Cheng, Xiao Liang, Yong Dai, and nan du. 2024. Chunk, align, select: A simple long-sequence processing method for transformers.

Jiawen Xie, Qi Su, Shaoting Zhang, and Xiaofan Zhang. 2023. Alleviating exposure bias via multi-level contrastive learning and deviation simulation in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9732–9747, Toronto, Canada. Association for Computational Linguistics.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11556–11565.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255, Online. Association for Computational Linguistics.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Ming Zhang, Shuai Dou, Ziyang Wang, and Yunfang Wu. 2022. Focus-driven contrastive learning for medical question summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6176–6186, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

12. Appendix

12.1. More Experimental Details

In this paper, for the training of GECSum on various datasets we uniformly employ the Adam optimizer with the following dynamic learning rate:

$$lr = 2 * 10^{-5} * \min\{(\frac{warmup}{step})^{0.5}, \frac{step}{warmup}\},$$

where *warmup* indicates the warmup steps, *step* is the number of updating steps. Through this design, the maximum learning rate in our scheduler is limited to $2 * 10^{-5}$ and it does not change with the variation of warmup steps

Dataset	Model	Batch Size	Max Epoch	Warmup Steps	α	β	η	γ	λ
CNNDM	$\text{BART}_{\text{large}}$	1	10	10000	0.1	2.0	2.0	100	0.001
XSum	$\text{PEGASUS}_{\text{large}}$	1, 2	10	6400, 10000	0.1	0.6, 1.0	2.0	100	0.001, 0.01
SAMSum	$\text{BART}_{\text{large}}$	1, 2, 4	100	6400, 10000	0.1	0.6, 1.0	2.0	100	0.001, 0.01
MeQSum	$\text{BART}_{\text{large}}$	1, 2, 4	1000	2500, 6400	0.1	0.6, 1.0	2.0	100	0.001, 0.01

Table 9: Hyper-parameter grid for downstream task fine-tuning. We use Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-6$) for all datasets.

Dataset	Generation Parameters
CNNDM	beam: 4, max_len: 150, min_len: 60, no_repeat_ngrams: 3
XSum	beam: 4, max_len: 65, min_len: 15, no_repeat_ngrams: 3
SAMSum	beam: 4, max_len: 70, min_len: 15, no_repeat_ngrams: 3
MeQSum	beam: 4, max_len: 50, no_repeat_ngrams: 3

Table 10: Hyper-parameter settings during inference for each dataset.

12.2. Hyper-parameters & Packages

Table 9 and Table 10 delineate the hyper-parameter settings at the training and inference phases of the experiment, respectively. For evaluation metrics, we used the following packages:

- For the uniform data processing, before the evaluation, all the reference summaries and model-generated summaries are converted to the lowercase and tokenized using the PTB tokenizer: https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/process/PTBTokenizer.html.
- For ROUGE metrics, we used the public rouge-score Perl package provided by the authors: https://github.com/summanlp/evaluation/tree/master/ROUGE-RELEASE-1.5.5.
- For BERTScore (Zhang* et al., 2020), we used the public bert-score package shared by the authors: https://github.com/Tiiiger/bert_score.
- For BARTScore (Yuan et al., 2021), we used the public code shared by the authors: https://github.com/neulab/BARTScore.