IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 March 2024, accepted 2 May 2024, date of publication 8 May 2024, date of current version 15 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3398085

RESEARCH ARTICLE

Dual-Level Contrastive Learning for Improving Conciseness of Summarization

WEI PENG, HAN ZHANG, DAN JIANG, KEJING XIAO, AND YUXUAN LI
Beijing Institute of Graphic Communication, Beijing 102627, China

Corresponding author: Han Zhang (zhanghan@bigc.edu.cn)

This work was supported in part by Beijing Municipal Education Commission under Grant KM202110015003, in part by the Beijing Natural Science Foundation-Key Project of Beijing Municipal Education Commission under Grant KZ202010015021, in part by Joint Training Base for Graduate Students Majoring in Electronic Information under Grant 21090224002, and in part by the Research Foundation for Youth Scholars of Beijing Institute of Graphic Communication under Grant 27170123036 and Grant 27170123034.

ABSTRACT The task of text summarization aims to provide highly condensed summaries of long textual information, with the ideal summary being both precise and concise. In recent years, there has been extensive research on the brevity of summaries, but these methods still have significant room for improvement in ROUGE scores, especially when the beam width is increased. We propose a new model called the DC (Dual-Level Contrastive Learning), which combines contrastive learning and data augmentation, and design a new scoring function during the training phase to enhance accuracy and conciseness. Ultimately, our framework achieves excellent ROUGE scores, ensuring concise and readable output even with increased beam width. Experimental results on the CNN/DailyMail (47.82 ROUGE-1, 0.017 VAR) and XSum (47.31 ROUGE-1, 0.0052 VAR) datasets demonstrate that our approach can significantly enhance the accuracy and conciseness of the summaries. Some metrics have exceeded those of the current state-of-the-art model BRIO, promoting the state-of-the-art performance to a higher level.

INDEX TERMS Contrastive learning, data augmentation, conciseness, KL-divergence, text summarization.

I. INTRODUCTION

Text summarizations aim to extract crucial information from texts and documents, which are required to be accurate, concise, and easily comprehensible. A lot of effort has gone into making the summary concise, i.e. controlling the length of the output to match what is actually needed [2], [3]. Some researchers have employed sinusoidal positional encoders within neural encoder-decoder models to conduct length constraints [4]. Others have integrated the prediction of length into the encoder side and subsequently infused the projected length into the decoder side to generate final summarizations [5]. Although it has been demonstrated that these methodologies were simple and effective, they tend to produce lower-quality outputs with increased beam width [6]. There are still potential for improvement in terms of ROUGE [7] scores.

Besides conciseness and readability, achieving word-level accuracy is also essential. Summarizations are typically categorized into extracted and abstracted forms based on their generation methods. Abstractly generative methods stand as the dominant approach currently. Most notably, abstract text summarization has gained significant attention in recent years owing to the emergence of large-scale pre-training models like BART [8], PEGASUS [9], T5 [10], and others which exhibit remarkable performance. However, primarily based on the Transformer architecture, many of these state-of-the-art models have the risk of overlooking full-text semantics [11]. To address this limitation, several studies proposed to employ sentence-level data augmentation to develop denoising seq2seq models [12].

Moreover, the majority of current models solely rely on Maximum Likelihood Estimation (MLE). However, the exclusive use of MLE Loss introduces exposure bias which deteriorate model performance and yield subpar summarizations [13]. To counter exposure bias and empower models to select the optimal summary from multiple candidates, a model founded on the contrastive learning framework was proposed [1].

Based on the analysis mentioned above, we suppose that the integration of contrastive learning with data augmentation holds promise for enhancing both conciseness and accuracy

The associate editor coordinating the review of this manuscript and approving it for publication was Ioannis Schizas.

in text summarization. Motivated by this, we introduce a novel framework structure termed DC (Dual-level Contrastive Learning)[1]. DC aims to achieve excellent ROUGE scores while ensuring the concise and readable sentences generated, even at higher beam widths. As depicted in Fig.1, our model operates in two key stages. First, we leveraged sentence-level data augmentation [12] to enhance the denoising ability of the model. Specifically, we augmented the original documents(D) with sentence-level data and then input the augmented two documents(D_1, D_2) into a pre-training model, which generated multiple corresponding candidate summarizations. Second, we utilized contrastive learning to fine-tune the model to encourage the model to assign higher estimated probabilities to better candidate summaries. At this stage, we use symmetric KL loss to quantify the difference between two discrete data summaries ($Z_1 1, Z_2 1$).

Our main contributions are as follows. First, we created a new scoring function Eq.8, i.e., $F(S)$.in the Contrastive Learning phase. It is utilized in the loss function to guarantee that the generated summaries closely resemble the reference summary at the word level and aim to maintain the length of the generated summaries as consistent as possible with the reference summary. This function simultaneously focus on conciseness and accuracy which significantly enhances the quality of the summary. Second, we validated the effectiveness of utilizing symmetric KL Loss for discrete data in the context of text summarization tasks. Third, we conducted an in-depth exploration to determine the optimal combination of data augmentation techniques aimed at improving the summarization performance of the model. Finally, our proposed framework exhibits promising results across various metrics including ROUGE, BERTScore, VAR, and FKGL, exhibiting remarkable performance on the CNN/DM and XSum dataset. Some of these metrics, such as the ROUGE score, have surpassed the current SOTA model BRIO, driving the SOTA performance to a new level.

II. RELATE WORK

A. CONTRASTIVE LEARNING

Contrastive learning, an unsupervised learning paradigm, offers the advantage of not requiring extensive labeled training data. It has robust generalization capabilities surpassing those of supervised learning which have been widely used in machine vision and natural language processing. The core idea is to minimize the distance of positive examples from the anchor examples and maximize the distance of negative examples from the anchor examples. There are two critical issues in designing a good Contrastive Learning framework. First, To construct positive and negative samples, positive samples typically represent the target of our task, whereas negative samples contrastingly represent non-target examples. By comparing these positive and negative samples, the model can learn the target features and subsequently perform classification, recognition, or other related tasks effectively. Second, Loss function selection, a reasonable contrast loss function, prevents the model from collapsing, i.e., all positive and negative instances are mapped to the same point on the hypersphere, and the model cannot learn any useful information from the data. These papers initially necessitated larger batch sizes to accommodate more negative samples [14], [15], [16]. However, subsequent advancements have shown promising results by achieving comparable performance using smaller batch sizes without the need for negative samples [17], [18]. Similarly, within text summarization tasks, some researches have achieved superior results employing Contrastive Learning without relying on large batch sizes or even without incorporating negative samples [19], [20].

B. DATA AUGMENTATION

Data augmentation was first widely used in machine vision [21]. Various operations, such as flipping, cropping, and sharpening, were applied to images through augmentation methods. For contrastive learning, data augmentation is utilized to create positive sample pairs, yielding favorable outcomes upon training. Chen et al [22] have conducted ablation experiments to ascertain the most effective image data augmentation methods from a plethora of options. This concept was extended into the realm of natural language processing, predominantly divided into sentence-level and word-level data augmentation. Currently, the prevailing approach involves sentence-level augmentation techniques such as random flipping, deletion, document rotation and random swap of the order of sentences. Notably, in the ESACL paper [12], it was demonstrated that document rotation is usually harmful to performance for text summarization tasks. Consequently, in our approach, we adopt Random Deletion and Random Swap techniques in sentence-level data augmentation.

C. REGULARIZATION LOSS

Regularization Loss plays a vital role in deep learning, and it is a technique used to improve the generalization of models and reduce model overfitting. KL-divergence (Kullback-Leibler divergence) is one of the regularization methods. KL-divergence represents the gap between discrete data, with lower values indicating that the probability distributions of data are more similar. Eq.1 is the standard KL-divergence formula, where $p(x)$ denotes the true distribution, $q(x)$ denotes the predictive distribution of the model, and X denotes the set of all possible values.

$$D_{KL}(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)} \tag{1}$$

Subsequently, a symmetric KL-divergence was introduced, exhibiting the ability to enhance the robustness of model and improve performance significantly across various NLP

[1] All code is publicly available at https://github.com/pengwei-ui/DC-Model.

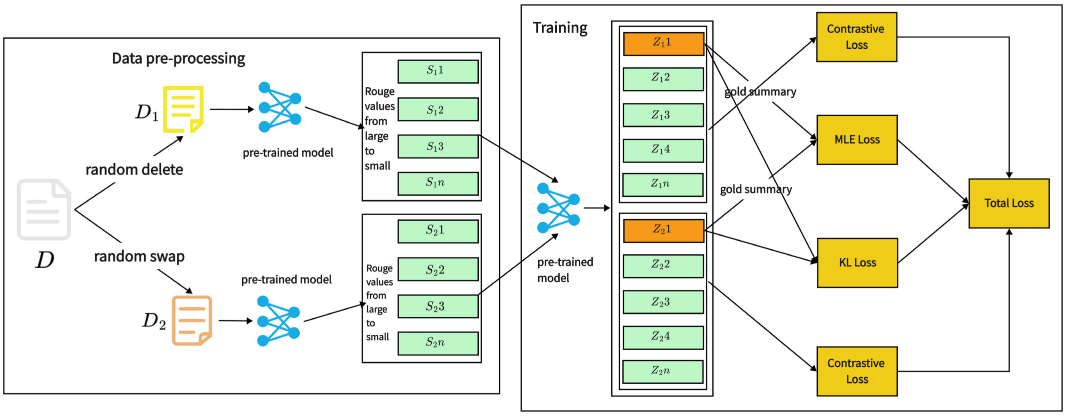

FIGURE 1. The Dual-level Contrastive Model framework involves a two-stage data pre-processing approach. Different pre-trained models are used for different datasets: Bart is used for the CNNDM dataset, and PEGASUS is used for the XSum dataset.

tasks [23]. This technique was further extended to text summarization tasks, yielding favorable outcomes as evidenced by studies such as [24] and [25].

$$L_{KL}(p,q) = D_{KL}(p||q) + D_{KL}(q||p) \tag{2}$$

III. OUR PROPOSED MODEL

In the context of training text summarization models, the primary objective is to ensure that given a document D, the model generates summaries S that closely approximate the reference summaries S^*. Our proposed model, DC, adopts a hybrid approach incorporating cross-entropy loss, KL-divergence, and contrastive loss mechanisms to achieve this objective effectively. The training and evaluation phases of the model and the general flow is shown in Fig.2. The training begins with a long text and n candidate summaries as input. Subsequently, using an encoder-decoder, the model ranks the n candidate summaries based on a range of scores from maximum to minimum. Through contrastive loss function training, the model reduces the distance between the highest-scoring summary and the reference summary while increasing the distance between the lower-scoring summary and the reference summary. The process also utilizes MLE loss and KL divergence to measure the disparity between the reference and candidate abstracts.

For this experiment, we used several pre-trained Transformer models as the basis for the text feature encoder. Specifically, we employed the BART[2] pre-trained model for the CNNDM dataset and the PEGASUS[3] pre-trained model for the XSum dataset.

A. ABSTRACTIVE TEXT SUMMARIZATION

The primary objective of the abstract summarization task is to train a model g to produce a summary S that closely resembles the reference summary S^*. Maximum Likelihood Estimation (MLE) serves as the standard training algorithm, enhancing the probability of the model generating S^*.i.e.,

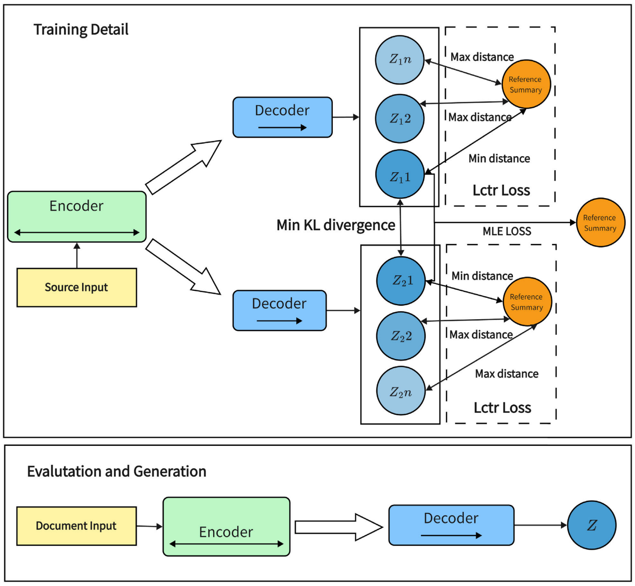

FIGURE 2. In the training phase, the two sets of data-augmentation long text and the corresponding n candidate summaries are used as the input to the encoder, and the decoder will sort the n candidate summaries [($Z_1 1, Z_1 2 \ldots Z_1 n$), ($Z_2 1, Z_2 2 \ldots Z_2 n$)].

$$\theta^* = argmax \sum_i logp_{g\theta}(S^{*(i)}|D^{(i)};\theta) \tag{3}$$

where θ denotes the parameters of g and $p_{g\theta}$ denotes the probability distribution entailed by these parameters,$\{D^{(i)}, S^{*(i)}\}$ is the i-th training sample.

Eq.4 is equivalent to minimizing the sum of negative logarithm of likelihoods of the tokens in the reference summary S^* whose length is l, which is the cross-entropy loss:

$$L_{MLE} = -\sum_{j=1}^{l}\sum_{s} p_{\text{true}}\left(s|D, S^*_{<j}\right)\log p_{g_\theta}\left(s|D, S^*_{<j};\theta\right) \tag{4}$$

[2] https://huggingface.co/facebook/bart-large-cnn
[3] https://huggingface.co/google/pegasus-xsum

where $S^*_{<j}$ denotes the partial reference sequence and s^*_0 is a pre-defined start token. P_{true} utilizing the regular One-hot distribution tends to result in model overfitting, and label smoothing is now widely used to overcome the shortcomings [26], which alleviates the effect of overfitting problem and improves model generalization. N is the size of the dictionary and β is the soft threshold. We set β to 0.1 in our experiments.

$$p_{true}(s|D, S^*_{<j}) = \begin{cases} 1-\beta & s = s^*_j \\ \dfrac{\beta}{N-1} & s \neq s^*_j \end{cases} \tag{5}$$

B. DC: DUAL-LEVEL CONTRASTIVE LEARNING

In our model, for the original document D, sentence-level data augmentation is first performed to generate D_1, D_2, and then the pre-trained model generates n candidate summaries by using beam search [27], and ROUGE computation is carried out. The ROUGE scores of the n candidate summaries from the two groups D_1 and D_2 are to be sorted from high to low respectively. Finally, we get the candidate abstracts which are grouped: $\{S_1 1, S_1 2 \ldots S_1 n\}$, $\{S_2 1, S_2 2 \ldots S_2 n\}$.

Input$[\{D_1\}, \{S_1 1, S_1 2 \ldots S_1 n\}]and[\{D_2\}, \{S_2 1, S_2 2 \ldots S_2 n\}]$ into the model, and let the model find the optimal summary from n candidate summaries by comparing losses. We consider $Z_1 1$ and $Z_2 1$ to be the best summaries of D_1 and D_2 generated by the model. Our contrastive learning takes the form of triplet loss [28]. This loss function allows to minimizing the vector distance between anchor examples(S^*) and positive examples($Z_1 1, Z_2 1$) while maximizing the vector distance between anchor examples(S^*) and negative examples($Z_1 2, Z_1 3 \ldots Z_2 2, Z_2 3 \ldots$).

$$L_{CTR} = \sum_i \sum_{j>i} max(0, f(S_j) - f(S_i) + \lambda_{ij}) \tag{6}$$

where S_i and S_j are two different candidate summaries and ROUGE(S_i, S^*) > ROUGE(S_j, S^*),[4] $\forall i, j, i < j, \lambda_{ij}$ is the margin multiplied by the difference in rank between the candidates, $\lambda_{ij} = \lambda \times (j - i)$, where λ is set to the appropriate value according to the different datasets. $f(S_i)$ and $f(S_j)$ is the length-normalized estimated log probability,[5] which is calculated as shown in Eq.7.

$$f(S) = \frac{\sum_{t=1}^{L} log_{g_\theta}(s_t|D, S_{<t}; \theta)}{|S|^\alpha} \tag{7}$$

Eq.7 is from BRIO, where α is the length penalty hyperparameter, D is the input long text, s_t is the t-th word of the summary, and its formula is used to estimate the probability of the next word s_t given the previously predicted sequence $S_{<t}$. However, this function focuses on the relationship between the words of generated summary and the candidate summary and does not closely relate to the gap between the lengths of the two sentences.

We designed a new score function Eq.8 in place of Eq.7, and BP is the length penalty in the BLEU(Bilingual Evaluation Understudy) [30] metric for machine translation, in which c is the length of the candidate summary, r is the length of the reference summaries. The design was originally intended to allow the model to generate summaries with words and lengths close to the reference summary. The model is trained using a contrastive loss function, which tends to generate shorter candidate summaries to minimize Eq.6.

$$F(S) = BP \times \frac{\sum_{t=1}^{L} log_{g_\theta}(s_t|D, S_{<t}; \theta)}{|S|^\alpha} \tag{8}$$

BLEU is a metric used to assess the quality of machine translation by comparing the similarity between the results of automatic translations and human-referenced translations. In machine translation, shorter sentences often receive higher n-gram matching scores, and the length penalty BP(Eq.9) in BLEU is intended to discourage the generation of excessively short sentences. The length penalty in the BLEU metrics inspired me to design a new scoring function Eq.8, which biases the model towards generating shorter summaries, enhancing the brevity of the summaries.

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-\frac{r}{c})} & if \quad c \leq r \end{cases} \tag{9}$$

The model identifies the two best summaries $S_1 i$ and $S_2 j$ from multiple candidate summaries where $i, j \in n$ by a new score function $F(S)$. Since $S_1 i$ and $S_2 j$ are generated by pre-trained models for semantically almost identical documents, $S_1 i$ should be close to $S_2 j$, We use Eq.2 to calculate the difference between two sentences.

Finally, combining the cross-entropy loss, contrastive loss, and KL loss, we can get the final objective in Eq.10, where α is the weight of the cross-entropy loss, β is the weight of the contrastive loss, γ is the weight KL Loss.

$$Loss = \alpha L_{MLE} + \beta L_{CTR} + \gamma L_{KL} \tag{10}$$

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASET

The CNN/DailyMail dataset [31] contains news articles and their associated highlights from the CNN and Daily Mail websites. The dataset contains 287,128 training articles, 13,368 validation articles, and 11,490 test articles. The training set documents contain an average of 760.5 words, and the reference summaries contain 52.59 words. The XSum dataset [32] is a highly abstractive dataset of articles from the British Broadcasting Corporation (BBC). The dataset contains 204,046 training articles, 11,332 validation articles, and 11,334 test articles. The training set documents contain an average of 429.2 words, and the reference summaries contain 23.3 words.

[4] In order to speed up model training, we used pypi package rouge to calculate ROUGE, its version is 1.0.1(https://github.com/pltrdy/rouge).

[5] length-normalize as it is standard in comparing hypotheses in neural sequence generation [29].

2) EVALUATION METRICS

(1) ROUGE: The ROUGE metrics include ROUGE-1, ROUGE-2, and ROUGE-L. These metrics measure the degree of overlap of a single word (ROUGE-1), the degree of overlap of a two-word phrase (ROUGE-2), and the longest common subsequence (ROUGE-L), respectively, in order to determine the degree of similarity between the generated text and the reference text. In the model evaluation phase, ROUGE scores are computed with the ROUGE-1.5.5.pl script.[6]

(2) BERTScore [33]: ROUGE is used to compute word-to-word similarity, while BERTScore is used to compute sentence-to-sentence similarity. We use its default version for English texts.[7]

(3) FKGL [34]: It is an index used to assess the readability of a text. It estimates the level of education required for a text by analyzing sentence length and word complexity. The lower the FKGL index, the better the readability of the text, i.e., the easier it is to understand. We use the calculations from the EASSE paper [35]. See Eq.11 for specific calculations. Word represent the number of words, sentence represent the number of sentences, and syllables represent the number of syllables.

$$FKGL = 0.39 \times (word/sentence) + 11.8 \times (syllables/word) - 15.59 \tag{11}$$

(4) VAR: To assess the quality of predicted lengths and length controllability. Reference [5], we also use the length variance(VAR): Eq.12,where y_i is the length of the generated summary and y_i^* is the length of the reference summary.

$$VAR = 0.001 \times \frac{1}{n}\sum_{i=1}^{n} |y_i - y_i^*| \tag{12}$$

3) TRAINING DETAIL

In the data pre-processing stage, for the CNN/DM dataset, we performed random deletion and random swapping on the original text and then the text which have been enhanced was fed into the pre-training model BART to generate the corresponding 16 candidate summaries. We combined the original document, the augmented document with the corresponding 16 candidate summaries, two by two(such as $[\{D_1\}, \{S_1 1, S_1 2 \ldots S_1 n\}]$), for training. More details are described in Appendix B. For the XSum dataset, we directly used the dataset provided in the BRIO paper, which has been used to generate multiple candidate summaries using PEGASUS. We used the pre-training model PEGASUS for training.

B. RESULTS

1) EXPERIMENTAL RESULTS ON CNN/DM DATASET

We compare the more mainstream and representative summarization models available nowadays. The BART [8] and PEGASUS [9] model is among the hottest pre-training language models available. ESACL [12] uses a variety of sentence-level data augmentation. BART-RD [25] employs KL divergence to quantify the dissimilarity between two discrete distributions (i.e., model-generated summaries and reference summaries) to mitigate model inconsistency between the training and inference phases induced by Dropout. LFPE [5] takes into account the gap between the length of the generated summary and the reference summary during the fine-tuning modeling phase. BRIO [1] is a two-stage model framework using contrastive learning and is currently a SOTA model.

TABLE 1. ROUGE evaluation on CNN/DM dataset, * results reported in the original papers,† results from our own evaluation script,R-1/2/L are the ROUGE-1/2/L F_1 scores.

Model	R-1	R-2	R-L
$BART^*$	44.16	21.28	40.90
$PEGASUS^*$	44.17	21.47	41.11
$ESACL^*$	44.24	21.06	41.20
$BART + RD^*$	44.51	21.58	41.24
$LFPE^*$	45.93	22.30	42.44
$BRIO^*$	47.78	23.55	44.57
$BART^\dagger$	44.33	21.15	41.06
$BRIO^\dagger$	47.78	23.55	44.56
DC	**47.82**	**23.59**	**44.63**

TABLE 2. Experimental results on CNN/DM. * results reported in the original papers,† results from our own evaluation script,R-1/2 are the ROUGE-1/2 F_1 scores, BS denotes BERTScore,FG denotes FKGL.

Model	R-1	R-2	BS	VAR↓	FG↓
$BART^\dagger$	44.33	21.15	87.99	0.022	8.18
$BRIO^\dagger$	47.78	23.55	**89.07**	0.02	7.17
$LFPE^*$	45.93	22.30	-	0.03	-
DC	**47.82**	**23.59**	88.83	**0.017**	**6.90**

The ROUGE evaluation results on the CNN/DM dataset are presented in Table.1. Our proposed DC model exhibits superior performance compared to most baseline models. Notably, our DC model significantly outperforms both BART and PEGASUS. Furthermore, when compared to ESACL using similar data augmentation techniques, our model exhibits better performance. In a scenario where both models utilize KL Loss to enhance denoising effects, our model surpasses BART-RD. Additionally, our model further improves the ROUGE score and simplicity compared to LFPE and BRIO.

The model's conciseness evaluation results on the CNN/DM dataset are depicted in Table.2. Our DC model exhibits significantly superior performance in terms of simplicity and readability when compared to existing benchmark models.

2) EXPERIMENTAL RESULT ON XSUM DATASET

For this experiment conducted on the XSum dataset, we deliberately refrained from employing data augmentation due to the inherently concise and abstract nature of the

[6] with -c 95 -r 1000 -n 2 -m arguments.

[7] https://github.com/Tiiiger/bert_score

TABLE 3. Experimental results on XSum. * results reported in the original papers,† results from our own evaluation script,R-1/2/L are the ROUGE-1/2/L F_1 scores,FG denotes FKGL.

Model	R-1	R-2	R-L	VAR↓	FG↓
*BART**	45.14	22.27	37.25	-	-
*PEGASUS**	47.21	24.56	39.25	-	-
*BRIO**	49.07	25.59	40.40	-	-
*ESACL**	44.64	21.62	36.73	-	-
LFPE	-	-	-	-	-
PEGASUS†	47.38	24.54	39.41	0.0054	9.345
BRIO†	**49.07**	**25.59**	**40.47**	**0.0049**	**8.866**
DC	47.75	24.86	39.72	0.0052	8.944

TABLE 4. Calculate the length of the summaries in the CNNDM and XSum training datasets. Ref-L denotes the average length of the reference summaries, Cand-L denotes the average length interval of the candidate summaries sorted by ROUGE scores, and $\frac{ref}{cand}$ denotes the average ratio interval between the length of the reference summaries and the length of the candidate summaries after ROUGE scores have sorted the candidate summaries.

Dataset	Ref-L	Cand-L	$\frac{ref}{cand}$
XSum	23.19	(15.87,24.53)	(0.96,1.57)
CNN/DM	52.58	(50.16,64.47)	(0.81,1.04)

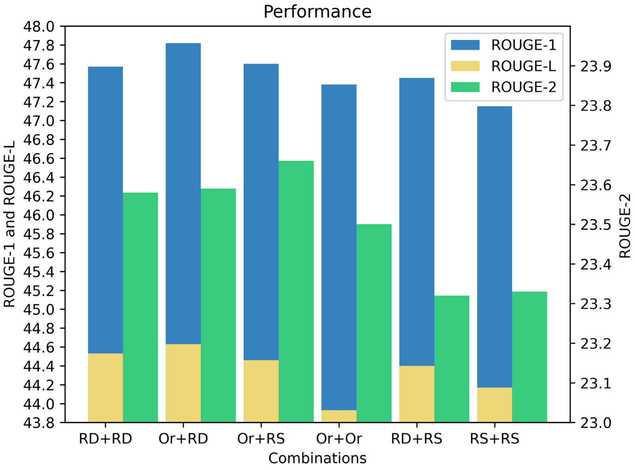

FIGURE 3. Performance on CNN/DM under different combinations: RD denotes Random Deletion, RS denotes Random Swap, Or denotes Original Document. Parameter settings for experimental results: α is 10, β is 0.1, γ is 0.001.

XSum dataset. Previous experiment [12] revealed that the model performance was deteriorated when data augmentation techniques were applied to this dataset.

The ROUGE evaluation results on the XSum dataset are presented in Table.3. Our DC model displays superior performance when compared to most baseline models. Notably, following the fine-tuning of the pre-training model PEGASUS, our model outperforms both BART and PEGASUS on this dataset. and slightly worse than the best-performing model BRIO.

The experimental results for the XSum dataset do not surpass those of BRIO in terms of model performance. We believe this is due to the use of Eq.8, which biases the model towards selecting shorter sentences as output. After conducting data preprocessing on the dataset, we calculated the average length of reference abstracts, the average length range of candidate abstracts ordered by ROUGE scores, and the range of the average ratio between the length of reference abstracts and the length of candidate abstracts ordered by ROUGE scores. as shown in Table.4. The average length of the candidate summaries in XSum is at least 15.87, while that of the reference summaries is 23.19,

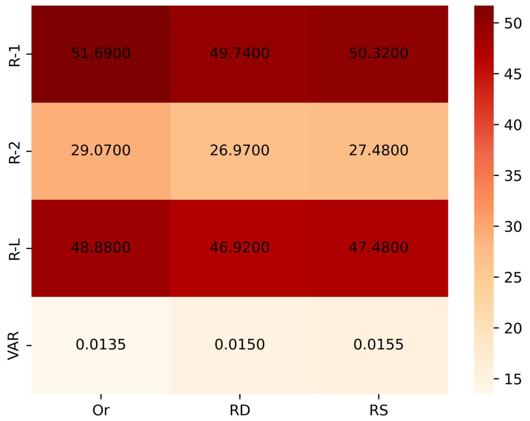

FIGURE 4. Optimal candidate summarization results under different operations. R-1/2/L are the ROUGE-1/2/L F_1 scores.Or denotes Original Document, RD denotes Random Deletion, RS denotes Random Swap.

TABLE 5. Experimental result on CNN/DM with different combinations: RD denotes Random Deletion, RS denotes Random Swap, Or denotes original document. Parameter settings for experimental results: α is 10, β is 0.1, γ is 0.001.R-1/2/L are the ROUGE-1/2/L F_1 scores,FG denotes FKGL.

	R-1	R-2	R-L	BS	VAR↓	FG↓
RD+RD	47.57	23.58	44.53	88.76	0.017	6.80
Or+RD	**47.82**	23.59	**44.63**	88.83	0.017	6.90
Or+RS	47.60	**23.66**	44.46	88.83	**0.014**	6.93
Or+Or	47.38	23.50	43.93	**89.06**	**0.014**	6.85
RD+RS	47.45	23.32	44.40	88.83	0.018	6.68
RS+RS	47.15	23.33	44.17	88.83	0.017	**6.43**

TABLE 6. Model performance with different γ Coefficient, Parameter settings for experimental results: α is 10, β is 0.1.R-1/2/L are the ROUGE-1/2/L F_1 scores.

Coefficient(γ)	R-1	R-2	R-L	VAR↓
0.001(Or+Or)	47.38	**23.50**	43.93	0.014
0 (Or+Rd)	47.72	**23.69**	44.62	**0.015**
0.001 (Or+Rd)	**47.82**	23.59	**44.63**	0.017

TABLE 7. Result on CNN/DM with different beam widths used in beam search.BW denotes beam-width, R-1/2 are the ROUGE-1/2 F_1 scores.

BW	BART			DC		
	R-1	R-2	VAR↓	R-1	R-2	VAR↓
4	**44.33**	**21.15**	**0.022**	47.82	23.59	**0.017**
10	43.74	20.60	0.022	47.99	23.86	**0.017**
16	43.48	20.40	0.023	**48.09**	**23.91**	**0.017**

TABLE 8. Results of different models on CNN/DM.

Model	Summary
Reference	Band of extreme weather predicted to hit an area including Chicago, Detroit and St. Louis, as well as Memphis, Tennessee, and Little Rock, Arkansas on Thursday .Storm Prediction Center estimates that 57 million people live in an area with an 'enhanced risk' of hail, damaging winds and tornadoes .Severe thunderstorms packing 80mph winds and large hail already made their way across central Missouri on Wednesday .Areas that don't see strong storms on Thursday could see heavy rain instead .
BART	Storm Prediction Center upgraded to its second-highest advisory level - a moderate risk - while stressing that a significant tornado or two could form in a narrow stretch from northern Oklahoma to central Missouri. Strong storms swamped Indianapolis, Cincinnati and Charleston, West Virginia, at midday Wednesday and forecasters said more severe weather could form as far away as the plains of West Texas. Heat, humidity and the approach of a cold front and a jet stream could cause severe weather on Thursday, too.
BRIO	Strong storms swamped Indianapolis, Cincinnati and Charleston, West Virginia, at midday Wednesday and forecasters said more severe weather could form as far away as the plains of West Texas. The Storm Prediction Center upgraded to its second-highest advisory level - a moderate risk - of a significant tornado or two in a narrow stretch from northern Oklahoma to central Missouri. 57 million people were at an 'enhanced risk' of seeing storms Wednesday and into Thursday.
DC	Strong storms swamped Indianapolis, Cincinnati and Charleston, West Virginia at midday Wednesday. Storm Prediction Center upgraded to a moderate risk of tornadoes from northern Oklahoma to central Missouri. 57 million people were at an 'enhanced risk' of seeing storms nearby. More severe weather could form as far away as West Texas.

TABLE 9. Presentation of results using different data augmentation, Or denotes Original Document, RD denotes Random Deletion, RS denotes Random Swap.

Operation	Result
Or	By .Richard Spillett .**An electronic cigarette that exploded while charging in a car has prompted fresh safety fears over the product .**No one was inside at the time , but the back seat was completely melted and the mother-of-two was horrified at the thought she and her family could have been injured .**Scroll down for video .**Safety fear : Kim Taylor and her son Jake , 11 , had to extinguish a fire in her car after an e-cigarette exploded .I 've gone to sleep with it plugged in .before , she said .ate to think what would have happened if it had .**exploded next to me .** ....
RD	By .Richard Spillett . ~~An electronic cigarette that exploded while charging in a car has prompted fresh safety fears over the product .~~ No one was inside at the time , but the back seat was completely melted and the mother-of-two was horrified at the thought she and her family could have been injured .~~Scroll down for video .~~Safety fear : Kim Taylor and her son Jake , 11 , had to extinguish a fire in her car after an e-cigarette exploded .I 've gone to sleep with it plugged in .before , she said .ate to think what would have happened if it had .~~exploded next to me .~~ ....
RS	By .**An electronic cigarette that exploded while charging in a car has prompted fresh safety fears over the product .**Richard Spillett .**Scroll down for video .** No one was inside at the time , but the back seat was completely melted and the mother-of-two was horrified at the thought she and her family could have been injured .Safety fear : Kim Taylor and her son Jake , 11 , had to extinguish a fire in her car after an e-cigarette exploded .**exploded next to me .** I 've gone to sleep with it plugged in .before , she said .ate to think what would have happened if it had . ....

representing a significant discrepancy between the two. This disparity prevents the resulting metrics VAR and ROUGE from reaching the desired values.

Experimental results on both datasets show that the designed loss function, within a DC framework, allows for a few of improvements in the accuracy and conciseness of the summarization. The DC model outperforms many benchmark models, outperforms the current SOTA model on the CNN/DM dataset, and performs similarly to the SOTA model on the XSum dataset.

C. ABLATION STUDY

We conducted ablation experiments to explore the impact of various document combinations and KL Loss on the model's performance.

The results, illustrated in the Fig.3, indicate minimal variations in the effectiveness of different combinations in the text summarization task. Notably, employing different data augmentations generally yields superior results compared to a single data augmentation strategy. Surprisingly, combinations involving random swapping tend to yield less favorable outcomes in comparison. The performance of the different combinations was evaluated to explain this phenomenon. We computed the highest ROUGE scores and VAR for multiple candidate summaries utilizing various data augmentations. The results presented in the Fig.4 illustrate that ROUGE scores are higher for random swapping than for random deletion. However, the combination involving random swapping displays relatively poorer performance, which we attribute to the model's inclination towards abstracts with lower VAR values, influenced by our new scoring function Eq.8. In our experiments, we observed that the combination of the Original Document with Random Deletion yielded the most favorable ROUGE scores. We extensively measured and compared the performance across various combinations using ROUGE, BERTScore, VAR, and FKGL metrics. Refer to Table. 5 for a comprehensive overview.

The model exhibits relatively poorer performance when trained using the same combination of documents. We attribute this behavior to the influence of KL loss during model training, as evidenced by the fact that when identical documents are input as a combination, their resulting KL Loss is 0. Our investigation reveals that the magnitude of the KL coefficient has no discernible effect. To validate the impact of KL Loss on the model, we conducted experiments setting the parameter γ of KL Loss to 0 and 0.01. We compared the performance between the combination of the Original Document with itself and the combination of the Original Document with Random Deletion. The comprehensive results are showcased in Table.6. Our comparative analysis indicates that KL Loss significantly contributes to enhancing model performance.

It is worth noting that when the KL Loss coefficient is 0.01, the KL loss value is always 0 for the original document and the combination of the original document computation, which does not affect the training, and the settings are almost the same except that the scoring function is not the same as that of the BRIO model, which results in a final model with negligible differences in ROUGE, but with a significant increase in simplicity. The side shows that the scoring function we designed positively impacts the text summary's simplicity.

As shown in Table.5, the readability of the DC model on the CNN/DM dataset has also been somewhat improved. From the FKGL calculation method (Eq.11), we believe that this is due to the improved simplicity, which has led to a significant reduction in the value of the word parameter in the calculation equation, thus allowing for improved readability of the summary.

D. ANALYSIS

To ascertain the model's ability to discern the quality among multiple candidate summaries, we conducted experiments using different beam widths, with specific results outlined in Table.7. Our experiments revealed that as the beam width increases, the DC model exhibits an enhanced capability akin to the BRIO model in distinguishing the quality of multiple candidate summaries, showing that the ROUGE score improves with increasing beam size and the VAR score decreases with increasing beam size. Conversely, the BART model is unable to discern among multiple candidate summaries. This limitation in the capability of the BART model is due to the use of beam search techniques [27], where increasing the beam width introduces lower-quality candidate summaries [6], and their generators may not be able to distinguish them from high-quality candidates, resulting in lower quality summaries. Consequently, the BART model struggles to differentiate between these summaries effectively.

Table.8 show that the DC model generates more concise summaries that closely approximate the reference summaries compared to BART and BRIO. In conclusion, our findings highlight several significant advancements: Firstly, our model exhibits substantial enhancement in conciseness compared to the benchmark models while maintaining semantic proximity to the reference summaries. Secondly, the devised contrastive learning scoring function effectively improves the conciseness of the text summaries. Lastly, leveraging the randomly deleted data augmentation approach alongside the KL loss function contributes significantly to bolstering the model's robustness.

V. CONCLUSION AND FUTURE WORK

We designed the DC model, a two-stage modeling framework founded on contrastive learning and aimed at enhancing the conciseness and readability of generated summaries without compromising their quality. Our exploration of data augmentation reveals that Random Deletion significantly alters the original document's semantics, impacting the model's performance. To prioritize conciseness, we consider integrating a dedicated conciseness loss function. Additionally, for targeting specific metrics, we contemplate including task specific loss functions during training. Future endeavors might involve adopting data augmentation techniques involving reverse translation for improved outcomes. Notably, the model's enhanced readability is likely attributed to the succinct nature of the generated sentences. To further improve readability in future iterations, we plan to modify the ranking phase of candidate summaries, incorporating the FKGL metrics as part of our assessment and refinement process.

APPENDIX A
DATASET

As shown in the Table.9, we performed data augmentation following the method mentioned in the ESACL paper by randomly modifying 3 sentences of the original document. A pre-training model BART was used to generate 16 candidate summaries from the data augmented and original documents.

APPENDIX B
TRAINING DETAIL

Model For the CNN/DM dataset, We use huggingface's Facebook/bart-large-cnn model, which has 406M parameters. For the XSum dataset, We use huggingface's google/pegasus-xsum model, which has 568M parameters.

Optimizer We use the Adam optimizer with a warm-up learning rate and a warm-up step of 5000.The learning rate scheduling formula:

$$lr = 0.002 * min(step\_num^{-0.5}, \\ step\_num * warmup\_steps^{-1.5}) \tag{13}$$

Training Detail The batch-size of our model in the training phase is 1, We use the Adam optimizer with the learning rate set to 2×10^{-4}. For more detailed details on training, see our GitHub above for the code.:

ACKNOWLEDGMENT

The authors extend their heartfelt gratitude to the anonymous reviewers for their invaluable comments and constructive feedback, which significantly contributed to enhancing the quality of this work. They also thank Yixin Liu, the author of BRIO, for generously sharing insights and addressing their queries.

REFERENCES

[1] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing order to abstractive summarization," 2022, *arXiv:2203.16804*.

[2] I. Saito, K. Nishida, K. Nishida, A. Otsuka, H. Asano, J. Tomita, H. Shindo, and Y. Matsumoto, "Length-controllable abstractive summarization by guiding with summary prototype," 2020, *arXiv:2001.07331*.

[3] Z. Yu, Z. Wu, H. Zheng, Z. XuanYuan, J. Fong, and W. Su, "LenAtten: An effective length controlling unit for text summarization," 2021, *arXiv:2106.00316*.

[4] S. Takase and N. Okazaki, "Positional encoding to control output sequence length," 2019, *arXiv:1904.07418*.

[5] J. Kwon, H. Kamigaito, and M. Okumura, "Abstractive document summarization with summary-length prediction," in *Proc. Findings Assoc. Comput. Linguistics (EACL)*, 2023, pp. 606–612.

[6] F. Stahlberg and B. Byrne, "On NMT search errors and model errors: Cat got your tongue?" 2019, *arXiv:1908.10090*.

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[9] A. Sahu and S. G. Sanjeevi, "Better fine-tuning with extracted important sentences for abstractive summarization," in *Proc. Int. Conf. Commun., Control Inf. Sci. (ICCISc)*, Jun. 2021, pp. 11328–11339.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[11] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xie, "CERT: Contrastive self-supervised learning for language understanding," 2020, *arXiv:2005.12766*.

[12] C. Zheng, K. Zhang, H. J. Wang, L. Fan, and Z. Wang, "Enhanced Seq2Seq autoencoder via contrastive learning for abstractive text summarization," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 1764–1771.

[13] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1171–1179.

[14] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[15] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," 2021, *arXiv:2104.08821*.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, and B. Piot, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[18] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[19] S. Xu, X. Zhang, Y. Wu, and F. Wei, "Sequence level contrastive learning for text summarization," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 10, pp. 11556–11565.

[20] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," 2021, *arXiv:2106.01890*.

[21] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: When to warp?" in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–6.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.

[23] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," 2019, *arXiv:1911.03437*.

[24] A. Aghajanyan, A. Shrivastava, A. Gupta, N. Goyal, L. Zettlemoyer, and S. Gupta, "Better fine-tuning by reducing representational collapse," 2020, *arXiv:2008.03156*.

[25] L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, and T.-Y. Liu, "R-Drop: Regularized dropout for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10890–10905.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[27] A. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search for improved description of complex scenes," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7371–7379.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[29] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl. (SSST)*, D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://aclanthology.org/W14-4012

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[31] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1693–1701.

[32] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization," 2018, *arXiv:1808.08745*.

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.

[34] J. P. Kincaid, R. P. Fishburne Jr., R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel," Inst. Simul. Training, Univ. Central Florida, Res. Branch Rep. 8-75, 1975.

[35] F. Alva-Manchego, L. Martin, C. Scarton, and L. Specia, "EASSE: Easier automatic sentence simplification evaluation," in *Proc. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), Syst. Demonstrations*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 49–54. [Online]. Available: https://aclanthology.org/D19-3009

WEI PENG received the B.S. degree from Beijing Institute of Graphic Communication, in 2020, where he is currently pursuing the M.S. degree in electronic information. His research interests include text categorization and text summarization.

HAN ZHANG received the B.S. degree in automation from Taiyuan University of Science and Technology, China, and the M.S. and Ph.D. degrees in control science and engineering from the University of Science and Technology Beijing, China. She is currently with the School of Information Engineering, Beijing Institute of Graphic Communication. Her current research interests include artificial intelligence, deep learning, natural language processing, and smart education.

DAN JIANG received the B.S. degree in automatic control and information technology from Beihang University, China, and the M.S. and Ph.D. degrees in software engineering from Beijing University of Posts and Telecommunications, China. She is currently with the School of Information Engineering, Beijing Institute of Graphic Communication. Her current research interests include artificial intelligence, deep learning, and natural language processing.

KEJING XIAO received the M.S. degree in management science and engineering from Beijing Technology and Business University, China, and the Ph.D. degree from the School of Information, Renmin University of China, China. She is currently with the School of Information Engineering, Beijing Institute of Graphic Communication. Her current research interests include artificial intelligence, deep learning, natural language processing, and text mining.

YUXUAN LI is currently pursuing the master's degree in electronic information with Beijing Institute of Graphic Communication. Her research interests include text summarization and text simplification.