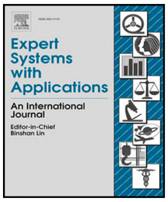
Expert Systems With Applications 237 (2024) 121364

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics

Ming Zhang [a,b], Chengzhang Li [a,b], Meilin Wan [c], Xuejun Zhang [a], Qingwei Zhao [a,*]

[a] *Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China*
[b] *School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China*
[c] *School of Microelectronics, Hubei University, Wuhan 430064, People's Republic of China*

ARTICLE INFO

Keywords:
Automatic summarization evaluation
Semantic similarity
Lexical similarity
Contrastive learning
Back-translation

ABSTRACT

With the development of pre-trained language models and large-scale datasets, automatic text summarization has attracted much attention from the community of natural language processing, but the progress of automatic summarization evaluation has stagnated. Although there have been efforts to improve automatic summarization evaluation, ROUGE has remained one of the most popular metrics for nearly 20 years due to its competitive evaluation performance. However, ROUGE is not perfect, there are studies have shown that it is suffering from inaccurate evaluation of abstractive summarization and limited diversity of generated summaries, both caused by lexical bias. To avoid the bias of lexical similarity, more and more meaningful embedding-based metrics have been proposed to evaluate summaries by measuring semantic similarity. Due to the challenge of accurately measuring semantic similarity, none of them can fully replace ROUGE as the default automatic evaluation toolkit for text summarization. To address the aforementioned problems, we propose a compromise evaluation framework (ROUGE-SEM) for improving ROUGE with semantic information, which compensates for the lack of semantic awareness through a semantic similarity module. According to the differences in semantic similarity and lexical similarity, summaries are classified into four categories for the first time, including good-summary, pearl-summary, glass-summary, and bad-summary. In particular, the back-translation technique is adopted to rewrite pearl-summary and glass-summary that are inaccurately evaluated by ROUGE to alleviate lexical bias. Through this pipeline framework, summaries are first classified by candidate summary classifier, then rewritten by categorized summary rewriter, and finally scored by rewritten summary scorer, which are efficiently evaluated in a manner consistent with human behavior. When measured using Pearson, Spearman, and Kendall rank coefficients, our proposal achieves comparable or higher correlations with human judgments than several state-of-the-art automatic summarization evaluation metrics in dimensions of coherence, consistency, fluency, and relevance. This also suggests that improving ROUGE with semantics is a promising direction for automatic summarization evaluation.

1. Introduction

As one of the areas of greatest interest in natural language processing (NLP), automatic text summarization (ATS) has been widely studied for decades (El-Kassas, Salama, Rafea, & Mohamed, 2021; Garg & Kumar, 2022; Xiao, He, & Jin, 2022). Especially in recent years, ATS has developed rapidly due to the introduction of large-scale datasets (Cohen, Kalinsky, Ziser, & Moschitti, 2021; Fabbri, Li, She, Li, & Radev, 2019) and the proposal of pre-trained language models (PLMs) (Ghadimi & Beigy, 2022; Mohd, Jan, & Shah, 2020; Xie, Bishop, Tiwari, & Ananiadou, 2022). In particular, an effective automatic summarization evaluation metric will be a great boon for ATS, which can not only liberate people from time-consuming and labor-intensive manual evaluation but also greatly promote the development of text summarization.

As mentioned in Koto, Baldwin, and Lau (2022), the mainstream evaluation of ATS employs ROUGE (Lin, 2004), a simple but useful evaluation metric that calculates the overlapping units between candidate summaries and reference summaries. However, the widely used ROUGE is not perfect for automatic summarization evaluation. ROUGE is popular for its intuitiveness, simplicity, and ease of calculation, but there are studies have pointed out that it is still flawed (Lin et al., 2022; Schluter, 2017; ShafieiBavani, Ebrahimi, Wong, & Chen, 2018).

[*] Corresponding author.
E-mail addresses: zhangming@hccl.ioa.ac.cn (M. Zhang), lichengzhang@hccl.ioa.ac.cn (C. Li), wanml@hubu.edu.cn (M. Wan), zhangxuejun@hccl.ioa.ac.cn (X. Zhang), zhaoqingwei@hccl.ioa.ac.cn (Q. Zhao).

https://doi.org/10.1016/j.eswa.2023.121364
Received 1 April 2023; Received in revised form 10 August 2023; Accepted 27 August 2023
Available online 4 September 2023

Since ROUGE may exhibit lexical bias (Ng & Abrecht, 2015) by measuring the lexical similarity between candidate summaries and reference summaries, it has the following limitations in evaluating ATS. Firstly, ROUGE is generally considered unsuitable for evaluating abstractive summarization, as it limits the diversity of generated summaries. It is well known that the same source document can generate multiple summaries in different expressions for people with different knowledge or purposes. However, ROUGE limits the diversity of generated summaries by rewarding summaries with greater lexical similarity and penalizing summaries with less lexical similarity. Secondly, ROUGE with lexical bias cannot comprehensively evaluate candidate summaries. To thoroughly evaluate candidate summaries, manual evaluation usually takes many factors into consideration, including redundancy, informativeness, and readability, etc. However, ROUGE is inherently incapable of evaluating the text quality of candidate summaries, as it only considers the lexical similarity between candidate summaries and reference summaries. Specifically, ROUGE exhibits better correlations for coherence and fluency, but poorer correlations for consistency and relevance, which is a common problem for lexical similarity-based metrics. Finally, ROUGE, which has repeatedly been shown to correlate well with manual evaluation, still leaves much room for improvement due to these limitations.

To improve automatic summarization evaluation, many efforts have been made to address the above limitations of ROUGE. On the one hand, some studies have extended ROUGE with synonym replacement and paraphrase, such as ROUGE-WE (Ng & Abrecht, 2015), ROUGE 2.0 (Ganesan, 2018) and ROUGE-G (ShafieiBavani et al., 2018). On the other hand, some studies have considered the semantic relationship between words to alternate the standard ROUGE. Due to the limitations of exact word matching, more and more semantic embedding-based metrics have been proposed in recent years, which calculate the similarity between vector representations of two summaries. As early representatives of semantic embedding-based metrics, GM (Rus & Lintean, 2012), VE (Forgues, Pineau, Larchevêque, & Tremblay, 2014), and SMS (Clark, Celikyilmaz, & Smith, 2019) have played an active role in automatic summarization evaluation. Recently, Cao and Zhuge (2022) adopts the semantic link network to evaluate the fidelity, conciseness, and coherence of candidate summaries. Especially with the rapid development of PLMs, research on PLMs-based automatic summarization evaluation has attracted considerable attention, such as MoverScore (Zhao et al., 2019), BERTScore (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020), and BARTScore (Yuan, Neubig, & Liu, 2021). Recently, SPEED (Akula & Garibay, 2022) uses sentence-level embeddings that pre-trained specifically for sentence-pair tasks to calculate the semantic similarity of two texts. Sem-nCG (Akter, Bansal, & Santu, 2022) is a gain-based evaluation metric that is not only semantically aware, but also rewards summaries based on the ranking of sentences. In addition, ENMS (He, Jiang, Chen, Le, & Ding, 2022) leverages semantic information to enhance existing N-gram based evaluation metrics. Due to the difficulty of obtaining reference summaries, researchers have also proposed reference-free metrics for evaluating candidate summaries, such as SUPERT (Gao, Zhao, & Eger, 2020), SDC* (Liu, Jia, & Zhu, 2022) and Shannon (Egan, Vasilyev, & Bohannon, 2022). Despite ongoing efforts to improve automatic summarization evaluation, none of these metrics can fully replace ROUGE as the default automatic evaluation toolkit for text summarization, as it has been repeatedly shown to correlate well with human judgments across multiple dimensions.

In this paper, we propose a compromise approach to tackle the aforementioned limitations of ROUGE, since accurately measuring semantic similarity is challenging. Inspired by ShafieiBavani et al. (2018), we propose a pipeline framework (ROUGE-SEM) that uses ROUGE combined with semantic information for automatic summarization evaluation. Specifically, a Siamese-BERT network with contrastive learning is adopted as a semantic similarity module to compensate for the lack of semantic awareness. As shown in Fig. 1, the proposed evaluation framework is composed of a candidate summary classifier, a categorized summary rewriter, and a rewritten summary scorer. These individual components constitute a pipeline approach in line with human behavior, that is, first classifying candidate summaries using semantic and lexical similarity, then rewriting summaries that are difficult to evaluate, and finally rescoring summaries based on the results of classifying and rewriting.

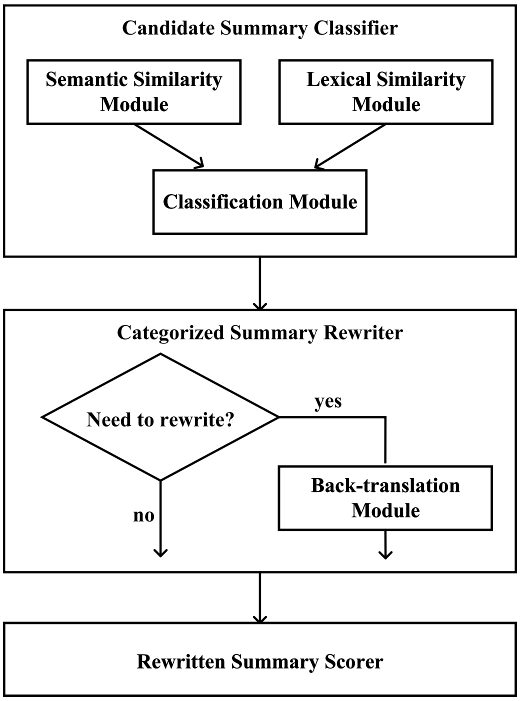
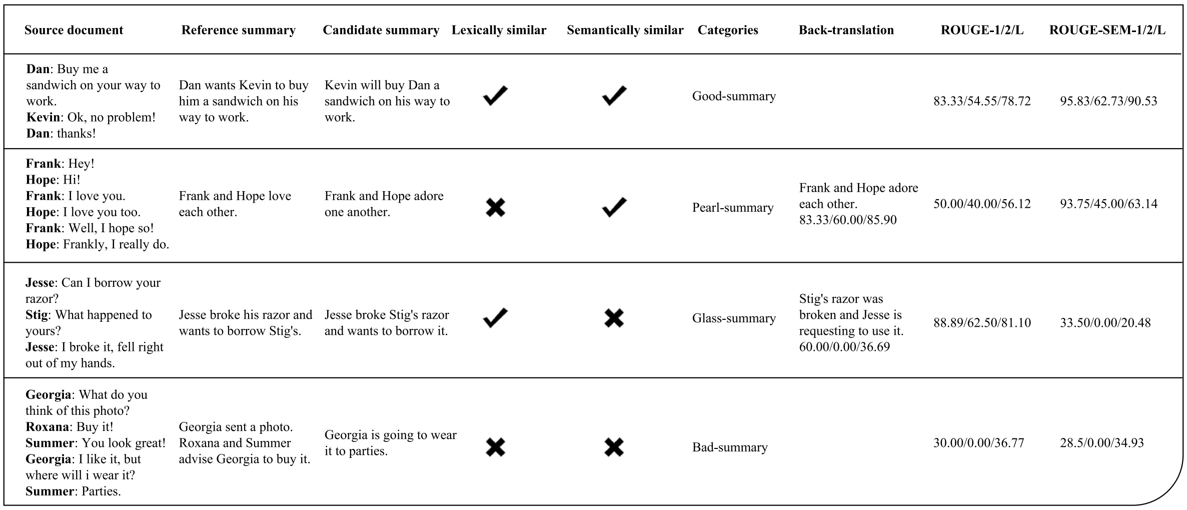

Candidate Summary Classifier
Semantic Similarity Module
Lexical Similarity Module
Classification Module
Categorized Summary Rewriter
Need to rewrite?
yes
no
Back-translation Module
Rewritten Summary Scorer

Fig. 1. Schema of the ROUGE-SEM framework.

To illustrate the proposed ROUGE-SEM more intuitively, we provide some typical examples from the DialSummEval dataset. As shown in Fig. 2, the source document, reference summary, and candidate summary are presented in the first three columns, respectively. The fourth and fifth columns assess whether the candidate summary is lexically or semantically similar to the reference summary, respectively. Then, the category of candidate summary is shown in the sixth column. The seventh column presents the results of back-translation. Finally, the last two columns show the standard ROUGE-1/2/L score and the proposed ROUGE-SEM-1/2/L score, respectively. From Fig. 2, we observe that candidate summaries are divided into four categories, including good-summary, pearl-summary, glass-summary, and bad-summary, based on the differences in semantic and lexical similarity. Due to lexical bias, it is difficult for ROUGE to accurately evaluate semantically related but dissimilar pearl-summary and semantically unrelated but similar glass-summary. By rewriting the above-mentioned summaries with the back-translation technique, we can alleviate its bias towards lexical similarity through more diverse synonymous expressions. In this way, the underestimated pearl-summary has a high probability of higher scores and the overestimated glass-summary has a high chance of lower scores. That is why ROUGE-SEM is a more effective evaluation metric than traditional ROUGE, which significantly improves the evaluation performance of pearl-summary and glass-summary by addressing the problem of lexical bias.

To validate our proposed evaluation metric, extensive experiments on SummEval (Fabbri, Kryściński, McCann, Xiong, Socher, & Radev, 2021) and DialSummEval (Gao & Wan, 2022) are conducted. In particular, the Pearson, Spearman, and Kendall correlation coefficients are used to measure the evaluation performance in terms of coherence, consistency, fluency, and relevance. The experimental results demonstrate that ROUGE-SEM outperforms or performs comparably to



Source document	Reference summary	Candidate summary	Lexically similar	Semantically similar	Categories	Back-translation	ROUGE-1/2/L	ROUGE-SEM-1/2/L
Dan: Buy me a sandwich on your way to work. **Kevin**: Ok, no problem! **Dan**: thanks!	Dan wants Kevin to buy him a sandwich on his way to work.	Kevin will buy Dan a sandwich on his way to work.	✓	✓	Good-summary		83.33/54.55/78.72	95.83/62.73/90.53
Frank: Hey! **Hope**: Hi! **Frank**: I love you. **Hope**: I love you too. **Frank**: Well, I hope so! **Hope**: Frankly, I really do.	Frank and Hope love each other.	Frank and Hope adore one another.	✖	✓	Pearl-summary	Frank and Hope adore each other. 83.33/60.00/85.90	50.00/40.00/56.12	93.75/45.00/63.14
Jesse: Can I borrow your razor? **Stig**: What happened to yours? **Jesse**: I broke it, fell right out of my hands.	Jesse broke his razor and wants to borrow Stig's.	Jesse broke Stig's razor and wants to borrow it.	✓	✖	Glass-summary	Stig's razor was broken and Jesse is requesting to use it. 60.00/0.00/36.69	88.89/62.50/81.10	33.50/0.00/20.48
Georgia: What do you think of this photo? **Roxana**: Buy it! **Summer**: You look great! **Georgia**: I like it, but where will i wear it? **Summer**: Parties.	Georgia sent a photo. Roxana and Summer advise Georgia to buy it.	Georgia is going to wear it to parties.	✖	✖	Bad-summary		30.00/0.00/36.77	28.5/0.00/34.93

Fig. 2. Typical examples of ROUGE-SEM on the DialSummEval dataset.

several state-of-the-art summarization evaluation metrics. Compared to the well-established ROUGE metric, the proposed evaluation metric has shown much higher and more consistent correlations with human judgments in terms of four dimensions, regardless of the correlation measures used. These exciting results confirm the effectiveness of using semantics to enhance ROUGE, suggesting this is a promising direction for automatic summarization evaluation.

Our main contributions of this work are summarized as follows.

- We propose a novel summarization evaluation metric (ROUGE-SEM), which improves the traditional ROUGE by making up for the lack of semantic awareness through a Siamese-BERT network with contrastive learning. The proposed evaluation metric consists of three individual components, including candidate summary classifier, categorized summary rewriter, and rewritten summary scorer. Through this pipeline framework, summaries are first classified, then rewritten, and finally scored, which are efficiently evaluated in a manner consistent with human behavior.
- According to the differences of lexical similarity and semantic similarity between candidate summaries and reference summaries, we introduce the classification of candidate summaries. It includes good-summary which is semantically related and lexically similar, pearl-summary which is semantically related but lexically dissimilar, glass-summary which is semantically unrelated but lexically similar, and bad-summary which is semantically unrelated and lexically dissimilar. We believe that this will benefit the progress of automatic summarization evaluation, especially offering the potential for improvements of lexical overlap-based metrics.
- We conduct experiments on two benchmark datasets to verify the effectiveness of ROUGE-SEM. Experimental results show that our proposed metric outperforms or performs comparably to several state-of-the-art summarization evaluation metrics on SummEval and DialSummEval datasets, suggesting that this is a promising direction for automatic summarization evaluation. We also share the proposed ROUGE-SEM to facilitate future work on text summarization systems.

The remainder of this manuscript is structured as follows. Related work and preliminaries are presented in Sections 2 and 3, respectively. We describe the details of the proposed method in Section 4. Then, the experimental setups and results are discussed in Sections 5 and 6, respectively. Finally, we conclude this work and make plans for future work in Section 7.

2. Related work

Since manual evaluation of text summarization is not practical for large-scale datasets, automatic summarization evaluation has attracted much attention from researchers (Deutsch, Dror, & Roth, 2021; Shapira, Pasunuru, Ronen, Bansal, Amsterdamer, & Dagan, 2021; Wang, Otmakhova, DeYoung, Truong, Kuehl, Bransom, & Wallace, 2023; Zhao & Lui, 2022). Up to now, a wide range of metrics have been used to measure the performance of text summarization systems. An overview of automatic summarization evaluation metrics proposed in recent years is shown in Table 1. This section describes related work on ATS evaluation, which is divided into two categories: extrinsic evaluation and intrinsic evaluation.

2.1. Extrinsic evaluation

Extrinsic evaluation judges the quality of summaries based on their performance on external tasks (Lloret & Palomar, 2012), such as question answering (QA) tasks, text classification tasks and other NLP tasks.

For **QA** tasks, SummaQA (Scialom et al., 2019) carries out an extrinsic evaluation of the impact of summarization in the task of QA. Specifically, source documents are used to generate questions, and then candidate summaries are used to answer these questions. Unlike SummaQA, FEQA (Durmus et al., 2020) employs a BERT-based QA paradigm that uses source documents to answer questions generated from candidate summaries. QuestEval (Scialom et al., 2021) unifies the precision and recall-based QA metrics and extends them with a weighted component for a more robust metric. In addition, QAFactEval (Fabbri et al., 2022) is also an improved QA-based factual consistency metric, where question generation and answerability detection are key components. For **text classification** tasks, FactCC (Kryscinski et al., 2020) measures the factness of a candidate summary by performing the text classification task on whether it belongs to the same category as the source document. For **other NLP** tasks, BLANC (Vasilyev et al., 2020) measures the informativeness of a candidate summary by judging how helpful the summary is to understanding the source document. In addition, CTRLEval (Ke et al., 2022) is an unsupervised reference-free metric based on text infilling tasks, which evaluates candidate summaries from different quality dimensions.

Extrinsic evaluations have the advantage of assessing the utility of summarization in a task, so they are of great practical value for summarization (Crystal et al., 2005). However, they can only reflect the performance of certain aspects of summaries. For example, SummaQA and FEQA can only measure the factual consistency of summaries. Notably,

Table 1
An overview of automatic summarization evaluations proposed in recent decades. The first column presents extrinsic evaluation and intrinsic evaluation; the second column shows the categories of extrinsic evaluation ("Question answering", "Text classification", and "Other NLP tasks") and intrinsic evaluation ("Text content", "Text quality", and "PLMs-based"); finally, the third column evaluates whether semantic similarity ("Only Sem."), lexical similarity ("Only Lex."), or both ("With Both.") are used.

Extrinsic evaluation	Question answering	Text classification	Other NLP tasks	With Lex. or Sem.
SummaQA (Scialom, Lamprier, Piwowarski, & Staiano, 2019)	✓			Only Sem.
FEQA (Durmus, He, & Diab, 2020)	✓			Only Sem.
FactCC (Kryscinski, McCann, Xiong, & Socher, 2020)		✓		Only Sem.
BLANC (Vasilyev, Dharnidharka, & Bohannon, 2020)			✓	Only Sem.
QuestEval (Scialom et al., 2021)	✓			Only Sem.
QAFactEval (Fabbri, Wu, Liu, & Xiong, 2022)	✓			Only Sem.
CTRLEval (Ke et al., 2022)			✓	Only Sem.
Intrinsic evaluation	**Text content**	**Text quality**	**PLMs-based**	**With Lex. or Sem.**
BLEU (Papineni, Roukos, Ward, & Zhu, 2002)	✓			Only Lex.
ROUGE (Lin, 2004)	✓			Only Lex.
METEOR (Banerjee & Lavie, 2005)	✓			Only Lex.
GM (Rus & Lintean, 2012)		✓		Only Sem.
VE (Forgues et al., 2014)		✓		Only Sem.
ROUGE-WE (Ng & Abrecht, 2015)		✓		Only Sem.
S^3 (Peyrard, Botschen, & Gurevych, 2017)	✓	✓		With Both.
ROUGE 2.0 (Ganesan, 2018)	✓			Only Lex.
ROUGE-G (ShafieiBavani et al., 2018)	✓	✓		With Both.
SMS (Clark et al., 2019)		✓		Only Sem.
MoverScore (Zhao et al., 2019)		✓	✓	Only Sem.
BERTScore (Zhang et al., 2020)		✓	✓	Only Sem.
SUPERT (Gao et al., 2020)		✓	✓	Only Sem.
BARTScore (Yuan et al., 2021)		✓	✓	Only Sem.
SummScore (Lin et al., 2022)		✓	✓	Only Sem.
SDC* (Liu, Jia, & Zhu, 2022)		✓	✓	Only Sem.
Shannon (Egan et al., 2022)		✓	✓	Only Sem.
SPEED (Akula & Garibay, 2022)		✓	✓	Only Sem.
SLN-based (Cao & Zhuge, 2022)		✓		Only Sem.
SEM-nCG (Akter et al., 2022)		✓	✓	Only Sem.
ENMS (He et al., 2022)	✓	✓	✓	With Both.

the experimental results are relatively expensive due to the careful design and extensive training of external tasks. For these reasons, our work falls into intrinsic evaluation.

2.2. Intrinsic evaluation

Intrinsic evaluation measures the quality of a summary by directly analyzing the summary itself (Mani, 2001). As mentioned in Ermakova, Cossu, and Mothe (2019), intrinsic evaluation can be broadly divided into two groups: text content evaluation and text quality evaluation. In particular, intrinsic evaluation methods using PLMs are highlighted in Table 1.

Text content evaluation. As the most dominant branch of intrinsic evaluation, text content evaluation is a method of calculating lexical overlap between candidate summaries and reference summaries. Papineni et al. (2002) proposed the mainstream metric BLEU in machine translation tasks, which calculates the n-gram overlap between candidate utterances and reference utterances. Similar to BLEU, as one of the most widely used metrics in text summarization tasks, ROUGE (Lin, 2004) matches exactly n-gram overlap between two summaries. As one of the early representatives, METEOR (Banerjee & Lavie, 2005) is based on the generalized concept of unigram matching between two texts. In addition, ROUGE 2.0 (Ganesan, 2018) uses the WordNet (Miller, 1995) to discover synonyms to facilitate the matching between two summaries.

Text quality evaluation. There is no doubt that text quality evaluation based on content understanding is more difficult than text content evaluation. To consider more meaningful semantic units instead of n-gram overlap, semantic embedding-based approaches are further proposed. As a pioneering work, Landauer and Dumais (1997) first proposed an embedding-based approach that computes the cosine similarity between the embeddings of two texts. In addition, a greedy matching algorithm is adopted in GM (Rus & Lintean, 2012) to measure the similarity of word embeddings. VE (Forgues et al., 2014) measures the semantic similarity of two summaries with sentence embeddings rather than word embeddings. As one of the important variants of ROUGE, ROUGE-WE (Ng & Abrecht, 2015) measures the soft semantic similarity of words rather than exact lexical matching. In particular, S^3 (Peyrard et al., 2017) models the results of ROUGE and ROUGE-WE to evaluate candidate summaries. SMS (Clark et al., 2019) employs word and sentence embeddings from the same representation space to calculate the semantic similarity of two summaries. Furthermore, ROUGE-G (ShafieiBavani et al., 2018) adopts a graph-based approach to enhance ROUGE using lexical and semantic similarity. Recently, Cao and Zhuge (2022) proposes a SLN-based metric that automatically evaluates the fidelity, conciseness, and coherence of summaries by transforming them into semantic link network (SLN).

PLMs-based evaluation. With the maturity of pre-trained language models, research on the summarization evaluation using PLMs has attracted considerable attention (Barbella, Risi, & Tortora, 2021). Specifically, MoverScore (Zhao et al., 2019) uses Word Mover's Distance (Kusner, Sun, Kolkin, & Weinberger, 2015) to measure the semantic similarity of two summaries with contextualized embeddings. In addition, the contextualized embeddings trained with BERT are used to measure the token similarity of two texts in BERTScore (Zhang et al., 2020). BARTScore (Yuan et al., 2021) uses pre-trained sequence-to-sequence models to measure how similar the candidate summary is to the reference summary in a generative way. Recently, Lin et al. (2022) proposes SummScore, a comprehensive metric for summary quality evaluation based on Cross-Encoder. With the help of pre-trained Cross-Encoder, it can effectively capture the semantic differences between candidate summaries and reference summaries. SPEED (Akula & Garibay, 2022) uses sentence-level embeddings that pre-trained specifically for sentence-pair tasks to calculate the semantic similarity of two texts. To address the lack of semantic understanding of the traditional ROUGE, Akter et al. (2022) proposes an alternative gain-based evaluation metric Sem-nCG for evaluating extractive summarization tasks. Specifically, it is not only semantically aware, but also rewards summaries based on the ranking of sentences. Besides, ENMS (He et al., 2022) leverages semantic information to enhance existing N-gram based evaluation metrics, replacing the unreasonable evaluation of hard matching. It should be note that SUPERT (Gao et al., 2020) is a reference-free metric that measures the semantic similarity of two summaries by aligning BERT-based contextualized embeddings at the token level. Similar to SUPERT, SDC* (Liu, Jia, & Zhu, 2022) is also a reference-free metric that evaluates the semantic distribution correlation and compression ratio between source documents and summaries. In addition, Shannon (Egan et al., 2022) is also a reference-free summarization evaluation metric that uses GPT-2 (Radford, Narasimhan, Salimans, Sutskever, et al., 2018) to estimate the information content shared between source documents and summaries.

Although more meaningful semantic units have been proposed and more PLMs have been applied to summarization evaluation, ROUGE is still the most widely used automatic evaluation metric, which is attributed to its competitive evaluation performance. As discussed in Section 6, we observe that text quality-based metrics perform better than text content-based metrics in terms of consistency and relevance, but worse in coherence and fluency. It is important to note that none of the semantic embedding-based metrics can fully replace ROUGE as the standard automatic evaluation protocol.

As shown in Table 1, we observe that most previous studies tend to evaluate summaries by measuring semantic similarity rather than lexical similarity. One possible explanation for this is the inevitable bias due to the exact lexical matching between candidate summaries and reference summaries. To the best of our knowledge, few studies have combined semantic similarity and lexical similarity for automatic summarization evaluation. ROUGE-G is one of the studies related to our work, which adopts a graph-based approach to obtain semantic similarity and combines it with lexical similarity to evaluate candidate summaries. Specifically, the Personalized PageRank algorithm (Haveliwala, 2003) is used to repetitive random walks on WordNet to measure semantic similarity and the standard ROUGE is used to calculate lexical similarity. Finally, ROUGE-G outputs the weighted sum of semantic similarity and lexical similarity. Another work close to this paper is ENMS, which enhances existing N-gram based evaluation metrics with semantic information. In particular, vector representation of N-grams is used in ENMS, while other studies focus on the granularity of word embeddings. In addition, by explicitly modeling the text content-based ROUGE metric and the text quality-based ROUGE-WE metric, S^3 does not show the expected good correlations with human judgments. However, none of these metrics consider the semantic and lexical similarity of candidate and reference summaries in a way consistent with human behavior.

Different from previous studies, the evaluation framework proposed in this paper is a compromise between text content evaluation and text quality evaluation. On the one hand, evaluating the text content of candidate summaries is straightforward, but exact lexical matching leads to inevitable lexical bias. On the other hand, evaluating the text quality of candidate summaries is notoriously difficult, since accurately measuring semantic similarity is challenging. To address these issues, we propose an evaluation framework (ROUGE-SEM) for improving ROUGE with semantic information, which circumvents the lexical bias of ROUGE while preserving its advantages. Specifically, the proposed framework compensates for the lack of semantic awareness in traditional ROUGE through a semantic similarity module. Through this pipeline framework, summaries are first classified by candidate summary classifier, then rewritten by categorized summary rewriter, and finally scored by rewritten summary scorer, which are efficiently evaluated in a human-intuitive way. To sum up, our work differs in proposing a better summarization evaluation framework that considers the semantic and lexical similarity of candidate and reference summaries in line with human behavior.

3. Preliminaries

This section presents the preliminaries, which are necessary to explain the proposed evaluation framework.

3.1. Problem formulation

In this paper, our goal is to measure the performance of candidate summaries from relevant reference summaries and source documents. Following Bhandari, Gour, Ashfaq, Liu, and Neubig (2020), we formally define the problem of summarization evaluation as follows.

Let C represent a set of candidate summaries, R denote a set of reference summaries, and D be a set of source documents. Each individual source document d of D can be represented as a sequence of m words as $d = \{w_{d,1}, w_{d,2}, \dots, w_{d,m}\}$, where each component refers to the word in the source document. Similarly, the corresponding reference summary $r \in R$ can be represented as $r = \{w_{r,1}, w_{r,2}, \dots, w_{r,n}\}$, and n is the number of words. Given an input document and a reference summary, a summarizer aims to generate a candidate summary. We refer to $c = \{w_{c,1}, w_{c,2}, \dots, w_{c,k}\}$ as a candidate summary $c \in C$ with k words. Hence, the process of summarization evaluation is formalized as follows.

$$Score(c) = Evaluator(c, r, [d]) \tag{1}$$

where $Score(c)$ is the evaluation result of candidate summary c. In most cases, the summarization evaluation does not consider the source document d, so $[d]$ is an optional variable. The non-linear evaluation function $Evaluator(\cdot)$ calculates the correlation of input texts $c, r, [d]$ to obtain the evaluation score $Score(c)$.

3.2. Lexical similarity and semantic similarity

As mentioned in Section 2, numerous methods of evaluation have been proposed to evaluate the quality of candidate summaries. Essentially, these methods are nothing more than calculating the lexical similarity or semantic similarity between candidate summaries and reference summaries.

Lexical similarity. The lexical similarity can be calculated by the n-gram overlaps between the candidate summary c and the reference summary r. In other words, the higher the similarity of word expressions between two texts, the higher the lexical similarity. Following Eq. (1), the lexical similarity is defined as:

$$Score_{lex}(c) = Evaluator_{lex}(c, r, [d]) = ROUGE(c, r) \tag{2}$$

where $Score_{lex}(c)$ is the lexical similarity score of the candidate summary c. Specifically, we adopt the widely used ROUGE as the evaluation function $Evaluator_{lex}$ for lexical similarity.

Semantic similarity. Contrary to the lexical similarity, the semantic similarity measures the semantic-level correlation between c and r. In this way, we are no longer limited to surface lexicographic matching, but capture the intrinsic semantic relevance of candidate and reference summaries. Hence, the semantic similarity $Score_{sem}$ can be calculated as:

$$Score_{sem}(c) = Evaluator_{sem}(c, r, [d]) = Siamese\text{-}BERT(c, r) \tag{3}$$

where $Score_{sem}(c)$ is the semantic similarity score of the candidate summary c. The Siamese-BERT network with contrastive learning is the core of the semantic similarity module, which is used to measure semantic similarity between two summaries.

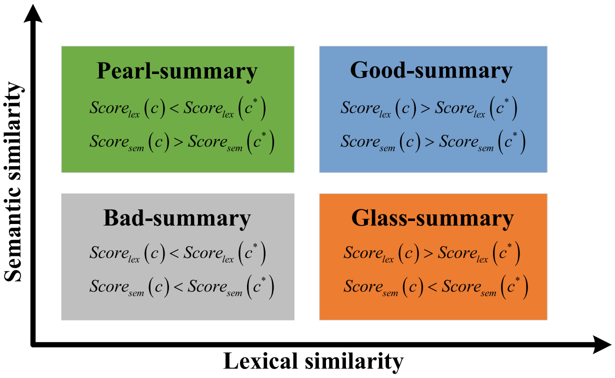

Pearl-summary
$Score_{lex}(c) < Score_{lex}(c^*)$
$Score_{sem}(c) > Score_{sem}(c^*)$
Good-summary
$Score_{lex}(c) > Score_{lex}(c^*)$
$Score_{sem}(c) > Score_{sem}(c^*)$
Bad-summary
$Score_{lex}(c) < Score_{lex}(c^*)$
$Score_{sem}(c) < Score_{sem}(c^*)$
Glass-summary
$Score_{lex}(c) > Score_{lex}(c^*)$
$Score_{sem}(c) < Score_{sem}(c^*)$
Semantic similarity
Lexical similarity

Fig. 3. The classification of candidate summaries based on lexical and semantic similarity. The X-axis represents the lexical similarity score and the Y-axis indicates the semantic similarity score.

3.3. Categories of candidate summary

For the first time, candidate summaries are divided into four categories based on the lexical similarity and semantic similarity, namely good-summary, pearl-summary, glass-summary, and bad-summary. Fig. 3 shows the classification of candidate summaries based on $Score_{lex}(c)$ and $Score_{sem}(c)$.

Good-summary. The good-summaries are those with higher lexical and semantic similarity, which are defined as follows. A candidate summary c is defined as a good-summary if there exists another candidate summary c^* that satisfies the inequality: $Score_{lex}(c) > Score_{lex}(c^*)$ and $Score_{sem}(c) > Score_{sem}(c^*)$. In short, the good-summary tends to have higher $Score_{lex}(c)$ and $Score_{sem}(c)$. Clearly, if a candidate summary is a good-summary, it is not difficult to evaluate it for metrics based on lexical or semantic similarity.

Pearl-summary. Inspired by Zhong et al. (2020), we define pearl-summaries to be those with a lower lexical similarity but a higher semantic similarity. Formally, it can be expressed mathematically as follows. A candidate summary c is defined as a pearl-summary if there exists another candidate summary c^* that satisfies the inequality: $Score_{lex}(c) < Score_{lex}(c^*)$ while $Score_{sem}(c) > Score_{sem}(c^*)$. Obviously, it is challenging to correctly evaluate a pearl-summary using the standard ROUGE. Such summaries are often underestimated and are therefore called pearly-like summaries.

Glass-summary. In contrast, summaries with lower semantic similarity and higher lexical similarity are classified as the glass-summary. Similarly, it can be expressed as follows. A candidate summary c is defined as a glass-summary if there exists another candidate summary c^* that satisfies the inequality: $Score_{lex}(c) > Score_{lex}(c^*)$ while $Score_{sem}(c) < Score_{sem}(c^*)$. Notably, the glass-summary is also hard to evaluate for metrics based on the lexical similarity. Hence, such summaries are often overestimated by the standard ROUGE and are therefore called glass-like summaries.

Bad-summary. It goes without saying that bad-summaries are those with lower lexical similarity $Score_{lex}(c)$ and semantic similarity $Score_{sem}(c)$. Similarly, a candidate summary c is defined as a bad-summary if there exists another candidate summary c^* that satisfies the inequality: $Score_{lex}(c) < Score_{lex}(c^*)$ and $Score_{sem}(c) < Score_{sem}(c^*)$. Clearly, most of the evaluation metric can effectively evaluate bad-summaries.

For the first time, we subdivide the candidate summaries into four groups with different characteristics of lexical and semantic similarity. Obviously, the evaluation performance of metrics varies greatly for different types of summaries. Both good-summary and bad-summary, which are consistent in lexical and semantic similarity, can be efficiently evaluated by ROUGE. However, pearl-summary and glass-summary with inconsistent lexical and semantic similarity tend to be mismeasured. In particular, ROUGE overestimates glass-summaries because of similar external expressions. Furthermore, ROUGE tends to underestimate pearl-summaries, which have the same semantics but different expressions. Therefore, the inaccurate evaluation of the pearl-summary and glass-summary is the main reason for the inconsistent correlation between ROUGE scores and human judgments.

4. Methodology

In this section, we introduce the general architecture of ROUGE-SEM, a pipeline framework that incorporates semantic similarity to improve ROUGE for better summarization evaluation. As shown in Fig. 4, it consists of three primary components: (1) a candidate summary classifier (Section 4.1), (2) a categorized summary rewriter (Section 4.2) and (3) a rewritten summary scorer (Section 4.3). These components constitute a pipeline approach that imitates human behavior, first classifying candidate summaries using semantic and lexical similarity, then rewriting summaries that are difficult to evaluate accurately, and finally rescoring summaries based on the classified and rewritten results.

4.1. Candidate summary classifier

From Fig. 4, the candidate summary classifier is composed of three modules, including the semantic similarity module, the lexical similarity module, and the classification module. The following sections go through each module in depth.

4.1.1. Semantic similarity module

As mentioned above, semantic similarity plays a crucial part in automatic summarization evaluation. Although many semantic embedding-based methods have been proposed, semantic similarity is still not measured effectively and accurately. Inspired by the siamese network structure (Mueller & Thyagarajan, 2016), we propose a novel Siamese-BERT network to measure the semantic similarity of two texts by contrastive learning. The overall structure of the semantic similarity module is shown in Fig. 5, which consists of an encoder layer, a contrastive learning layer, and a classification layer. Firstly, the semantic similarity module adopts the pre-trained BERT as the encoder layer to convert each token of input data into an embedding. Secondly, the contrastive learning loss between positive sample embedding and negative sample embedding is calculated in the contrastive learning layer. The semantic similarity module uses a softmax classifier to obtain positive and negative predictions, and then calculates the cross-entropy loss based on the predicted and trues labels. The final joint loss is a weighted sum of the contrastive learning loss and cross-entropy loss.

Word embedding. Each token of the input data, including candidate summaries, reference summaries, and source documents, should be converted into an embedding. Following Rani and Lobiyal (2021), we utilize pre-trained embeddings to represent each token of the input data to obtain semantic information. In particular, semantically contextual embeddings are derived from the input data by directly loading the bert-base-uncased model. Note that the bert-large-uncased model has better performance owing to its deeper architecture and greater hidden dimension.

BERT encoder. To measure the semantic similarity between candidate and reference summaries, the popular BERT model is adopted as a semantic encoder. In particular, BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee, & Toutanova, 2019) has been shown to have stronger semantic extraction capabilities than traditional neural networks due to the use of bidirectional transformers. It is worth mentioning that the basic transformer is composed of a

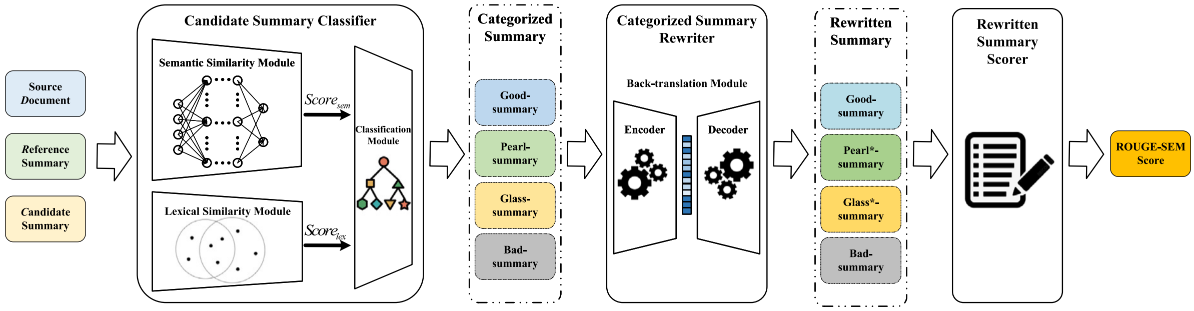


Fig. 4. An overview of the proposed ROUGE-SEM framework.

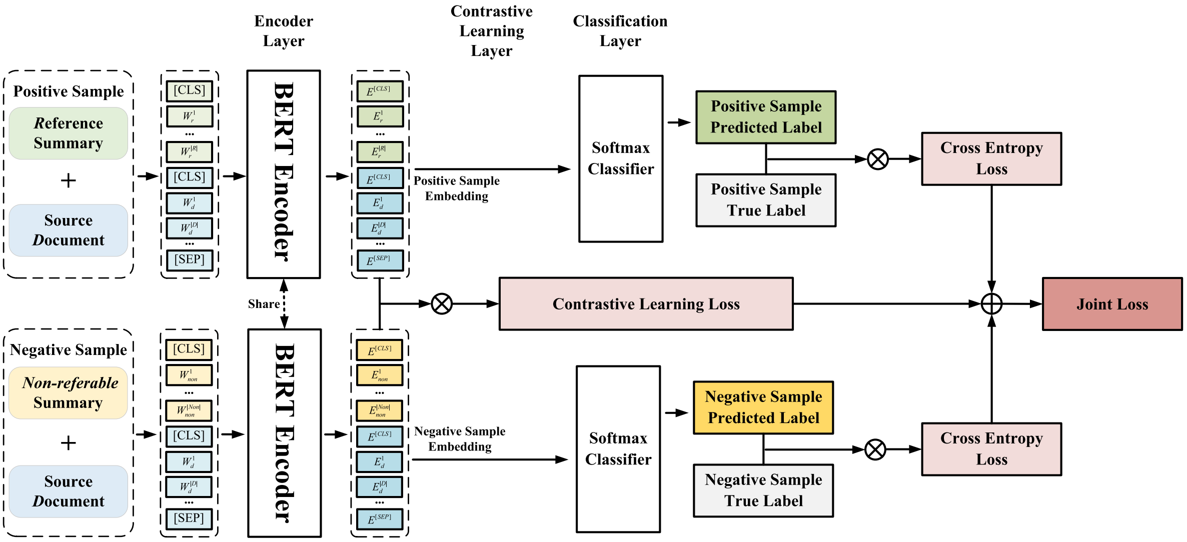


Fig. 5. The overall architecture of semantic similarity module.

multi-head self-attention module and a feed forward network. Furthermore, there are residual connections and normalization layers in both the multi-head self-attention module and the feed forward network. Due to limited space, a detailed introduction to the process of obtaining semantic embeddings from input data can be found in Liu, Yang, and Cai (2022).

Contrastive learning. To efficiently fine-tune the Siamese-BERT, the contrastive learning method is used to learn the latent semantic information of the input data. Intuitively, given a pair of source documents and reference summaries, their encoded representations should be as similar to each other as possible, while examples that do not belong to the same pair should be farther apart in the representation space. Similarly, the contrastive learning objective aims to map the representation of source documents $d \in D$ close to the representation of reference summaries $r \in R$, and away from the representation of non-referable summaries $non \in Non$. As shown in Fig. 5, we build positive samples using source documents and reference summaries, and negative samples using source documents and non-referable summaries.

Note that a margin-based ranking loss (Liu, Dou, Chen, Qin, & Heng, 2020) is adopted to update the weights, which can be formalized as:

$$\mathcal{L}_{CL} = \max\left(0, \Delta_m + f(D,R) - f(D,Non)\right) \tag{4}$$

where hyperparameter $\Delta_m = 1$ is a margin value, D represents source documents, R represents reference summaries, Non represents non-referable summaries, and $f(\cdot)$ denotes the distance in the BERT representation space.

Softmax classifier. As shown in Fig. 5, we use a softmax classifier to train the Siamese-BERT network by judging whether two sentence embeddings are semantically similar. Especially in the inference stage, the output of the softmax classifier is considered as the semantic similarity between candidate summaries and reference summaries.

Specifically, the classifier maps the input x to the corresponding class $y \in \{0, 1\}$, where 0 means semantically dissimilar and 1 indicates semantically similar. Following Reimers and Gurevych (2019), we take the word embeddings of two texts and their element-wise difference as input to the softmax classifier. Since source documents and reference summaries are semantically similar, while source documents and non-reference summaries are often semantically dissimilar, it is not difficult for us to obtain true labels for positive and negative samples. It is worth noting that the cross-entropy loss between the predicted and true labels is also used in the semantic similarity module, which can be formalized as:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n * \log\left(\delta\left(z_n\right)\right) + \left(1-y_n\right) * \log\left(1-\delta\left(z_n\right)\right)\right) \tag{5}$$

where y_n is the true label of nth sample, z_n represents probability of predicting nth sample as a positive sample, and δ denotes the sigmoid function.

Finally, the parameters of the Siamese-BERT network are trained to minimize both loss terms together as follows.

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\mathrm{CE}} + \lambda\mathcal{L}_{\mathrm{CL}} \tag{6}$$

where hyperparameter λ is a balancing coefficient.

4.1.2. *Lexical similarity module*

As the name suggests, the lexical similarity module is employed to calculate the lexical similarity between candidate summaries and reference summaries. In particular, the widely used ROUGE-1, ROUGE-2, and ROUGE-L metrics are adopted as the main components of the lexical similarity module. In literature, ROUGE-1, ROUGE-2, and ROUGE-L are often used to measure the quality of candidate summaries at different granularities, and they all contribute to the measure of lexical similarity. Specifically, ROUGE-1 represents the overlapping of unigrams between candidate summaries and reference summaries; ROUGE-2 refers to the overlapping of bigrams between two summaries; and ROUGE-L measures the longest common subsequence between two summaries. To obtain a more comprehensive lexical similarity score, we approximately consider the average of all ROUGE variants mentioned above while balancing precision and recall. Hence, the lexical similarity score $Score_{lex}$ can be calculated as:

$$Score_{lex}(c) = w_1 * ROUGE\text{-}1(c) + w_2 * ROUGE\text{-}2(c) \\ + w_3 * ROUGE\text{-}L(c) \quad (7)$$

herein, we set hyperparameters $w_1 = 0.3$, $w_2 = 0.3$, and $w_3 = 0.4$, which is similar to Dong, Shen, Crawford, van Hoof, and Cheung (2018).

4.1.3. Classification module

This part aims to classify candidate summaries based on the results of the semantic similarity module and lexical similarity module. As defined in Section 3, candidate summaries can be divided into four groups using semantic similarity score $Score_{sem}$ and lexical similarity score $Score_{lex}$. Therefore, the candidate summary classifier can be formalized as:

$$\text{Classifier}(c) = \begin{cases} \text{good-summary} & \text{if } Score_{sem} \geq \alpha \text{ and } Score_{lex} \geq \beta \\ \text{pearl-summary} & \text{if } Score_{sem} \geq \alpha \text{ and } Score_{lex} < \beta \\ \text{glass-summary} & \text{if } Score_{sem} < \alpha \text{ and } Score_{lex} \geq \beta \\ \text{bad-summary} & \text{if } Score_{sem} < \alpha \text{ and } Score_{lex} < \beta \end{cases} \quad (8)$$

where hyperparameters $\alpha = 0.6979$ and $\beta = 0.2839$. In general, a larger value of α should be considered to ensure semantic similarity. The effects of different parameters settings for α and β are discussed in Section 6.4

4.2. Categorized summary rewriter

With the performance of translation getting better and better in recent years, the back-translation technique has become a routine data augmentation method in the NLP community (Beddiar, Jahan, & Oussalah, 2021). Specifically, back-translation is the process of translating from one language into another and then back into the source language. Due to the encoder–decoder process, a rewritten sentence is different from the original sentence in many ways (Sugiyama & Yoshinaga, 2019), including synonym replacement, syntactic structure replacement, etc. Therefore, we can generate diverse expressions while preserving the semantics of the original sentence in this way. As shown in Fig. 6, the original English sentence is translated into German and then back into English to generate a new sentence. Compared to the original sentence, the new sentence retains most of semantic information of the original, but the lexical expression of the new sentence is different.

The back-translation module is adopted as the main component of the categorized summary rewriter to generate rewritten summaries with similar semantics but different expressions. To evaluate summaries more efficiently, we only rewrite pearl-summary and glass-summary which are inaccurately evaluated by ROUGE. In this way, pearl-summary can be theoretically rewritten into good-summary, and glass-summary can be rewritten into bad-summary. Hence, the categorized summary rewriter plays an important role in ROUGE-SEM, which converts the hard-to-evaluate pearl-summary and glass-summary into the easy-to-evaluate good-summary and bad-summary.

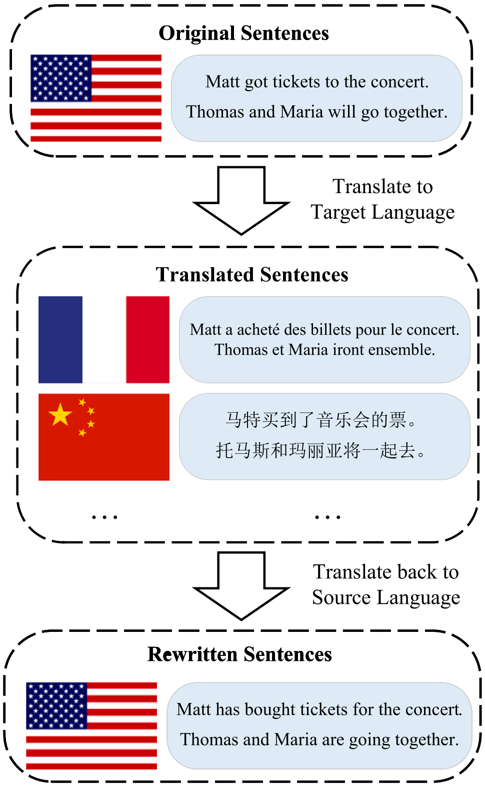

Original Sentences
Matt got tickets to the concert.
Thomas and Maria will go together.
Translate to
Target Language
Translated Sentences
Matt a acheté des billets pour le concert.
Thomas et Maria iront ensemble.
马特买到了音乐会的票。
托马斯和玛丽亚将一起去。
. . .
. . .
Translate back to
Source Language
Rewritten Sentences
Matt has bought tickets for the concert.
Thomas and Maria are going together.

Fig. 6. The process of back-translation.

The back-translation adopts the sequence-to-sequence model based on the encoder–decoder framework, and its basic model is also transformer. To reduce the time cost and computational overhead of evaluation process, the back-translation module is implemented by calling the reliable translation interface. In our experiments, German, Chinese, and French are selected as the intermediate languages of the back-translation module.

4.3. Rewritten summary scorer

The third part is used to calculate the evaluation scores based on the results of candidate summary classifier and categorized summary rewriter. Similar to the lexical similarity module, the standard ROUGE family is adopted as the main component of the rewritten summary scorer. Specifically, we output the maximum score of the rewritten pearl-summary and the minimum score of the rewritten glass-summary. To further improve the correlation between the proposed ROUGE-SEM and manual evaluation, the score of pearl-summary that can be accurately evaluated by humans but underestimated by ROUGE should be appropriately increased. Correspondingly, the score of glass-summary which is overestimated by ROUGE should be appropriately reduced. In the same way, we should appropriately increase the scores of good-summaries and reduce the scores of bad-summaries. Therefore, the evaluation results of ROUGE-SEM can be formalized as follows.

$$\text{ROUGE-SEM}(c) = \begin{cases} \gamma_1 * \text{ROUGE}(c, r) & \text{if } c \text{ is good-summary} \\ \gamma_2 * \max \text{ROUGE}(c^*, r) & \text{if } c \text{ is pearl-summary} \\ \gamma_3 * \min \text{ROUGE}(c^*, r) & \text{if } c \text{ is glass-summary} \\ \gamma_4 * \text{ROUGE}(c, r) & \text{if } c \text{ is bad-summary} \end{cases} \quad (9)$$

where different weights are assigned to different types of candidate summaries, the hyperparameters $\gamma_1, \gamma_2 \in [1, 1.5)$, $\gamma_3, \gamma_4 \in (0.5, 1]$, and c^* represents the rewritten summary of the candidate summary c.

Table 2
Hyperparameter settings of ROUGE-SEM along four quality dimensions on SummEval.

Dimensions	Hyperparameters settings					
	α	β	γ_1	γ_2	γ_3	γ_4
Coherence	0.6979	0.2839	1.0667	1.2000	0.5667	1.0000
Consistency	0.6979	0.2839	1.4000	1.1333	0.5000	0.8667
Fluency	0.6979	0.2839	1.1333	1.0333	0.6000	0.9333
Relevance	0.6979	0.2839	1.0000	1.1333	0.5667	1.0000

5. Experiments

In this section, extensive experiments are carried out to test ROUGE-SEM using two main benchmark datasets. To start off, these datasets, as well as implementation details and evaluation methods are provided. Following previous studies (Ng & Abrecht, 2015), the system-level correlations of Pearson (P), Spearman (S), and Kendall (K) are used as evaluation methods.

5.1. Datasets

The scarcity of human-annotated datasets has led to limited progress in the summarization evaluation task. To demonstrate the efficacy of the proposed ROUGE-SEM, thorough experiments are conducted on the SummEval (Fabbri et al., 2021) and DialSummEval datasets (Gao & Wan, 2022).

SummEval. It is an important benchmark dataset for the text summarization evaluation task proposed in 2021. It is the largest and most diverse collection of human judgments of generated summaries annotated by experts and crowdworkers on the CNN/DM dataset (Nallapati, Zhou, dos Santos, Gülçehre, & Xiang, 2016). More specifically, it contains 1,600 manually annotated summaries, each of which is evaluated on four quality dimensions of coherence, consistency, fluency, and relevance. To expand, coherence measures the overall quality of all sentences, consistency is the factual alignment between candidate summaries and source documents, fluency denotes the quality of individual sentences, and relevance indicates the selection of key content from original documents.

DialSummEval. It is the only benchmark dataset for the dialogue summarization evaluation task proposed last year. It is a multifaceted dataset of human judgments consisting of human scores of 14 model outputs on the SAMSum (Gliwa, Mochol, Biesek, & Wawer, 2019) dataset. Similar to SummEval, it contains 1,400 human-annotated summaries, each of which is scored by three college students on four dimensions of coherence, consistency, fluency, and relevance.

5.2. Implementation details

In this section, we provide implementation details of the proposed pipeline approach, including the candidate summary classifier, the categorized summary rewriter, and the rewritten summary scorer.

Preprocessing. For the above datasets, we first remove samples without source documents, candidate summaries, or reference summaries. In addition, the maximum length of the input is set to 512 subwords, which is the same as the maximum sequence length that BERT can handle. It is worth noting that we keep the first 512 subwords of source documents instead of the last 512 subwords, which leads to better performance. Furthermore, we compute the average of the human-annotated scores for each candidate summary as the final manual score.

Candidate summary classifier. We design a semantic similarity module and a lexical similarity module to calculate the semantic similarity score and lexical similarity score of two summaries, respectively. Specifically, we implement the proposed semantic similarity module with the HuggingFace PyTorch (Wolf et al., 2019) library, which conveniently supports tokenization and preprocessing. With limited computational power, we adopt the bert-base-uncased as the text encoder to implement the Siamese-BERT. The basic implementation of BERT is made up of 12 transformer blocks, 768 hidden layers, 12 self-attention heads, and 110M parameters. Also, the dimension of word embeddings is set to 768. Following the common BERT fine-tuning procedure, we optimize the loss function using an Adam optimizer. Additionally, a dropout rate of 0.1 and an initial learning rate of 2E-5 are set. The number of fine-tuning epochs is set to 100, and the batch sizes for training and testing are both set to 8. For the lexical similarity module, we use the SacreROUGE (Deutsch & Roth, 2020) library to implement the standard ROUGE family. Further, we carry out a grid-search of hyperparameters α and β in the candidate summary classifier and adopt the hyperparameter settings in Table 2.

Categorized summary rewriter. We implement the back-translation module by calling the reliable translation interface to reduce the complexity of the evaluation process. Specifically, Google Translate is used to implement the back-translation module, and German, Chinese, and French are selected as intermediate languages.

Rewritten summary scorer. We adopt a similar implementation to the lexical similarity module of the candidate summary classifier, since the standard ROUGE family serves as the main component. In particular, we optimize hyperparameters $\gamma 1, \gamma 2, \gamma 3, \gamma 4$ using grid search strategy. The settings of selected hyperparameters are shown in Table 2.

Hardware. All the experimental results presented in this paper are the average of three runs. Each experiment is performed on a workstation with two Tesla P100 GPUs with 12 GB memory, and an Intel(R) Xeon(R) E5-2680 v4 CPU @ 3.60 GHz with 128 GB memory. The benchmark datasets and source codes for our experiments will be available at https://github.com/zhangming-19/ROUGE-SEM.

5.3. Evaluation methods

It is well known that an automatic evaluation metric is better if it correlates better with human judgments. On the basis of this knowledge, we use the correlation between evaluation metric scores and manual scores to measure the performance of automatic evaluation metrics. Thus, the performance of an automatic evaluation metric can be formalized as follows.

For a given evaluation metric M and a ground-truth metric G, we use M (such as ROUGE) to approximate G (such as manual evaluation). In other words, the performance of an automatic evaluation metric is measured by computing the correlation between automatic evaluation scores and human-annotated scores of candidate summaries. Let there be m source documents, and each document has n candidate summaries. In addition, we refer to M_i^j as the score of M on the jth candidate summary of the ith document. Similarly, G_i^j denotes the score of G on the jth candidate summary of the ith document. Let $Corr(\cdot)$ be a correlation measure. Then, the correlation between M and G can be calculated at the system level as follows.

The system level correlation is calculated as follows.

$$Corr_{M,G}^{sys} = Corr\left(\left\{\left(\frac{1}{m}\sum_{i=1}^{m} M_i^j, \frac{1}{m}\sum_{i=1}^{m} G_i^j\right)\right\}_{j=1}^{n}\right) \tag{10}$$

Here, $Corr_{M,G}^{sys}$ calculates the correlation between the scores for each system.

To sum up, we use the system level correlation $Corr^{sys}_{M,G}$ to measure the performance of automatic evaluation metrics following Fabbri et al. (2021). Typically, if $Corr^{sys}_{M_1,G} > Corr^{sys}_{M_2,G}$, we can deduce that metric M_1 is better than metric M_2. Furthermore, three popular correlation measures are adopted as $Corr(\cdot)$ following Ng and Abrecht (2015), including Pearson correlation, Spearman rank coefficient, and Kendall rank coefficient.

5.4. Comparison metrics

Based on the SummEval and DialSummEval datasets, several widely used and representative metrics are selected for systematic comparisons, including lexical overlap-based metrics and semantic embedding-based metrics.

Lexical overlap-based metrics. These metrics measure the lexical similarity between candidate summaries and reference summaries based on their lexical overlap. Following previous studies (Ermakova et al., 2019; Gao & Wan, 2022), the well-known BLEU and METEOR are selected as representatives of lexical overlap-based metrics, especially the well-established ROUGE family is considered as a strong baseline.

- BLEU (Papineni et al., 2002) is a precision-based metric proposed for automatic evaluation of machine translation, which measures the translation quality by calculating the n-gram overlap between candidate utterances and reference utterances. In particular, BLEU-1 measures the accuracy of word translation based on unigrams, while BLEU-2 measures the fluency of sentence translation based on bigrams.
- ROUGE (Lin, 2004) is a Recall-Oriented Understudy for Gisting Evaluation metric proposed for text summarization. The well-established ROUGE family includes many variants, commonly used ones such as ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. In particular, ROUGE-N is similar to BLEU-N, except that BLEU focuses on precision, while ROUGE focuses on recall. ROUGE-L is a well-known variant of ROUGE, which calculates the length of the longest common subsequence between candidate summaries and reference summaries. In addition, ROUGE-W and ROUGE-SU* are also selected as comparison metrics, which are improved variants of ROUGE. ROUGE-W uses the weighted longest common subsequence to improve ROUGE-L, so that consecutive correct sentences get higher scores. Besides, ROUGE-S/SU* are extensions of the simple n-gram co-occurrence counting mechanism. ROUGE-S is based on a skip bigram co-occurrence mechanism that allows arbitrary gaps between any pair of words. ROUGE-SU* extends ROUGE-S by counting unigrams.

$$ROUGE\left(w_1, w_2\right) = \begin{cases} 1, & \text{if } w_1 = w_2 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where w_1 and w_2 are the words of n-gram.
- METEOR (Banerjee & Lavie, 2005) is an F-measure based Metric for Evaluation of Translation with Explicit ORdering, which addresses some inherent flaws of BLEU. Specifically, WordNet synonyms, stemmed tokens, and paraphrases are used for exact token matching to create unambiguous alignments between two utterances.

Semantic embedding-based metrics. These metrics measure semantic similarity by calculating the similarity between vector representations of two summaries. Following Fabbri et al. (2021), we select GM, VE, ROUGE-WE, and S^3 as early representatives without pre-trained language models. As pre-trained language model technology has flourished, several strong metrics using the pre-trained contextualized embeddings are also used for comparison, including the widely used MoverScore, the noteworthy SMS, the popular BERTScore, and the recently proposed BARTScore. Two reference-free metrics BLANC and SUPERT proposed in recent years are adopted as unsupervised baselines for extrinsic and intrinsic evaluation, respectively.

- GM (Rus & Lintean, 2012) is a metric based on word-to-word similarity, which uses a greedy matching algorithm to measure the similarity of word embeddings.
- VE (Forgues et al., 2014) is a simple alternative to average sentence vectors, which uses extrema to obtain sentence-level vectors to measure the semantic similarity of two texts.
- ROUGE-WE (Ng & Abrecht, 2015)is an important variant of ROUGE, which uses word embeddings to compute the semantic similarity of the words used in summaries instead of measuring lexical overlaps. In particular, ROUGE-WE-1/2/3 are used in the following experiments.

$$ROUGE\text{-}WE\left(w_1, w_2\right) = \begin{cases} 0, & \text{if } v_1 \text{ or } v_2 \text{ are OOV} \\ v_1 \cdot v_2, & \text{otherwise} \end{cases} \tag{12}$$

where w_1 and w_2 are the words of n-gram, v_1 and v_2 are the corresponding vector representations, OOV means out-of-vocabulary.
- S^3 (Peyrard et al., 2017) uses an off-the-shelf implementation of support vector regression to combine the existing automatic evaluation metrics, including metrics using reference summaries (e.g., ROUGE-N, ROUGE-L, and ROUGE-WE), metrics using source documents (e.g., TFIDF), and metrics using candidate summaries only.
- MoverScore (Zhao et al., 2019) measures Word Mover's Distance between the contextualized embeddings of generated and reference texts to evaluate the quality of the generated text.
- SMS (Clark et al., 2019) is a metric based on Sentence Mover's Similarity, which calculates the semantic similarity of two texts in a continuous space using word and sentence embeddings. In particular, Sentence Mover's Similarity is an optimization of Word Mover's Distance, enabling access to higher-level representations of the text.
- BLANC (Vasilyev et al., 2020) is a reference-free metric that measures the performance gain from accessing reference summaries when pre-trained language models perform language understanding tasks. There are two variants of BLANC, where BLANC-help directly concatenates reference summaries to source documents at inference time, while BLANC-tune uses the reference summaries to fine-tune pre-trained language models. In our experiments, we implement BLANC-help as a representative for extrinsic evaluation.
- SUPERT (Gao et al., 2020) is an unsupervised evaluation metric for multi-document summarization, which measures the semantic similarity between candidate summaries and pseudo reference summaries by aligning contextualized embeddings at the token level. In the experiments, we take the same preprocessing steps in the paper to implement SUPERT as a representative for reference-free metrics.
- BERTScore (Zhang et al., 2020) uses contextualized embeddings trained with BERT to compute token similarity between candidate and reference sentences. For a given tokenized candidate sentence $\hat{x} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m\}$, the corresponding contextualized embeddings trained with BERT can be denoted as $\hat{\mathbf{x}} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_m\}$. Similarly, the tokenized reference sentence $x = \{x_1, x_2, \ldots, x_n\}$ is mapped to $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Hence, BERTScore can be calculated as follows.

$$\begin{aligned} BERTScore\text{-}r &= \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^{\top} \hat{\mathbf{x}}_j \\ BERTScore\text{-}p &= \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^{\top} \hat{\mathbf{x}}_j \\ BERTScore\text{-}f &= 2\frac{BERTScore\text{-}p * BERTScore\text{-}r}{BERTScore\text{-}p + BERTScore\text{-}r} \end{aligned} \tag{13}$$

where $BERTScore\text{-}r$, $BERTScore\text{-}p$, and $BERTScore\text{-}f1$ are the recall, precision, and f1 scores of BERTScore, respectively.

Table 3
The correlation coefficients (“Pearson correlation (P)”, “Spearman rank correlation (S)” and “Kendall rank correlation (K)”) of annotations computed on **SummEval** along four quality dimensions (“Coherence”, “Consistency”, “Fluency”, and “Relevance”) between automatic metrics and human judgments. As representatives of error metrics, the average (Ave.) and standard deviation (S.D) values of the evaluation performance of each metric are presented in the last two columns, respectively. The evaluation performance of the top five is bolded in each column.

Metrics	Pearson correlation (P)				Spearman rank correlation (S)				Kendall rank correlation (K)				Error metrics	
	Coherence	Consistency	Fluency	Relevance	Coherence	Consistency	Fluency	Relevance	Coherence	Consistency	Fluency	Relevance	Ave.↑	S.D ↓
BLEU-1	0.0299	0.6371**	0.6085**	0.3595	0.1534	−0.0728	0.1556	0.2026	0.0667	−0.0834	0.1137	0.1190	0.1908	0.2339
BLEU-2	0.0464	0.1462	0.0909	0.1112	−0.0481	0.0335	−0.1204	0.0519	−0.0151	0.0309	−0.0927	0.0379	0.0227	**0.0797**
ROUGE-1	0.1972	0.6696**	0.6164**	0.6285**	**0.5215***	0.4957*	0.6580**	**0.6581****	**0.3630***	0.3107	0.4926**	**0.5353****	0.5122	0.1523
ROUGE-2	**0.2687**	**0.6860****	**0.6311****	**0.6956****	**0.6410****	0.4550	**0.7141****	**0.8186****	**0.4519***	0.3258	**0.6138****	**0.6194****	**0.5768**	0.1657
ROUGE-L	0.1406	0.5209*	0.4668	0.4442	0.2000	0.3950	0.3901	0.3217	0.1111	0.2270	0.2804	0.2528	0.3126	0.1322
ROUGE-SU*	**0.2648**	0.6300**	0.6125**	0.6413**	0.4110	0.1048	0.5654*	0.5807*	0.2889	0.0682	0.4622*	0.3717*	0.4168	0.2013
ROUGE-W	0.0120	0.5512*	0.4791	0.4124	0.4160	0.5117*	**0.6704****	0.5635*	0.2593	0.3107	**0.5077****	**0.4758****	0.4308	0.1722
METEOR	0.1869	**0.7163****	**0.6381****	0.6175**	**0.4601***	**0.8545****	**0.7124****	0.6532**	0.3037	**0.6896****	**0.5380****	0.4610*	0.5693	0.1891
S^3	0.0919	0.5930*	0.4922*	0.4789	0.1706	**0.6880****	0.3864	0.3585	0.1259	0.5380**	0.2955	0.2825	0.3751	0.1897
ROUGE-WE-1	0.1227	0.6119**	0.5578*	0.5117*	0.4601*	0.4538	0.6161**	0.5562*	**0.3333**	0.2804	0.4471*	**0.4758****	0.4522	0.1448
ROUGE-WE-2	0.0657	0.5205*	0.4590	0.4327	0.3963	0.4044	0.5667*	0.4886*	0.2296	0.2652	0.4168*	0.4015*	0.3873	0.1388
ROUGE-WE-3	0.0950	0.4883*	0.4359	0.4405	0.3521	0.3231	0.4963*	0.4469	0.2296	0.2046	0.3865*	0.4015*	0.3584	**0.1243**
SMS	0.2584	0.6396**	0.6165**	0.3331	0.1387	**0.7571****	0.3111	0.1743	0.1387	**0.7571****	0.3111	0.1743	0.3842	0.2403
MoverScore	**0.3483***	0.5598*	0.5730*	0.6415**	0.4503	0.0555	0.5642*	0.5746*	0.3185	0.0227	0.4319*	0.3569*	0.4081	0.2014
BERTScore-p	0.2175	0.2507	0.2945	0.4277	0.2344	−0.2700	0.3432	0.4101	0.1111	−0.1440	0.2501	0.1933	0.1932	0.2086
BERTScore-r	0.1843	**0.7098****	**0.6333****	0.6038**	0.4258	**0.8693****	**0.6679****	0.6188**	0.2593	**0.7047****	**0.4926****	0.4164*	**0.5488**	0.1988
BERTScore-f1	0.2476	0.6216**	0.5969*	**0.6624****	0.3865	0.0752	0.5383*	0.6311**	0.2444	0.0834	0.4168*	0.4015*	0.4088	0.2093
SUPERT	0.0502	0.6665**	0.5814*	0.3189	−0.0307	**0.7213****	0.2259	0.1007	−0.0307	**0.7213****	0.2259	0.1007	0.3043	0.2924
BLANC	0.0643	0.6058**	0.5079*	0.3694	−0.0724	0.6770**	0.1370	0.1080	−0.0222	0.5077**	0.0834	0.1933	0.2633	0.2577
BARTScore	0.2245	0.6417**	0.5941*	**0.6681****	0.3658	0.5046*	0.5528*	**0.6557****	0.2681	**0.6041****	0.4513*	0.4261*	0.4964	0.1509
ROUGE-SEM-1	**0.2903**	**0.6750****	**0.6179****	**0.6902****	**0.6070****	0.6162**	0.5898*	**0.7632****	**0.4059***	0.4605*	0.4605*	**0.6371****	0.5678	0.1357
ROUGE-SEM-2	**0.3340**	**0.6792****	**0.6294****	**0.7181****	**0.6352****	0.5176*	**0.6872****	**0.7926****	**0.4520***	0.3397	**0.5209****	**0.6222****	**0.5773**	0.1461
ROUGE-SEM-L	0.1827	0.6157**	0.5488*	0.5726*	0.3691	0.3869	0.4207	0.4294	0.2583	0.2642	0.3095	0.3852	0.3953	**0.1331**

[*] Symbol represents significant for $p \leq 0.05$.

[**] Symbol represents significant for $p \leq 0.01$.

- BARTScore (Yuan et al., 2021) uses an encoder–decoder based pre-trained model (BART) to measure how similar the candidate summary is to the reference summary in a generative way. For a given sequence-to-sequence pre-trained model parameterized by θ, an input token sequence $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ is mapped to an output token sequence $\mathbf{y} = \{y_1, y_2, \dots, y_m\}$. Correspondingly, BARTScore can be defined as the weighted log probability of one output $\mathbf{y}$ given input $\mathbf{x}$.

$$BARTScore = \sum_{t=1}^{m} \omega_t \log p\left(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{x}, \theta\right) \tag{14}$$

where ω_t is the weights of different tokens.

For both datasets, we consider following strong and representative baselines, including all lexical overlap-based metrics (e.g., BLEU-1/2, ROUGE-1/2/L, ROUGE-W, ROUGE-SU* and METEOR) and most semantic embedding-based metrics (e.g., SMS, MoverScore, BERTScore-p/r/f, BLANC, and BARTScore). Specifically for SummEval dataset, we take ROUGE-WE-1/2/3, S^3, and SUPERT into consideration. Whereas for DialSummEval dataset, we take GM and VE into consideration.

6. Results and analysis

In this section, we demonstrate the strength and usefulness of our ROUGE-SEM framework by conducting extensive experiments and performing a comprehensive analysis. Comparisons of evaluation performance on the SummEval and DialSummEval benchmark datasets are examined in depth in the following sections. Furthermore, ablation study as well as hyperparameter tuning are discussed in detail.

6.1. Results on SummEval

To get a fuller picture of the effectiveness of our proposal, we not only use Kendall correlation coefficient to measure the evaluation performance calculated at the system level following (Fabbri et al., 2021), but also Pearson and Spearman correlation coefficients. Specifically, Table 3 shows the Pearson, Spearman and Kendall correlations of the scores produced by automatic metrics with human-annotated scores for coherence, consistency, fluency, and relevance respectively. Additionally, the average (Ave.) and standard deviation (S.D) values are calculated, the evaluation performance of the top five is bolded, and the symbol * represents the statistical significance with manual evaluation. From Table 3, we have observed the following interesting phenomena.

(1) Clearly, there are metrics that correlate strongly with human judgments in one dimension, but few metrics show good correlations across all dimensions. For example, we find that BERTScore-r correlates strongly with human judgments on coherence and fluency, but its performance on coherence and relevance does not correlate well.

(2) Among the lexical overlap-based metrics, the widely used ROUGE-2 shows better correlations with human judgments. Specifically, the average evaluation performance of ROUGE-2 is 0.75% higher than that of METEOR, while the standard deviation is lower. In addition, the BLEU metric, which is frequently used in machine translation tasks, performs poorly in text summarization tasks. Further, we find that METEOR performs better in the consistency and fluency dimensions, but worse in coherence and relevance. Compared to METEOR, the performance of ROUGE-SEM is improved in four dimensions. In particular, the comparison of ROUGE-SEM and ROUGE will be discussed in detail later.

(3) Among the semantic embedding-based metrics, the increasingly popular BERTScore shows better comprehensive performance in four dimensions. As shown in Table 3, the evaluation performance of BERTScore-r in four dimensions with three coefficients is often among the top five, which shows that it is a strong baseline. In addition, the average evaluation performance of BERTScore-r is 54.88%, which is 5.24% higher than BARTScore.

(4) We have also observed that most embedding-based metrics have higher correlations in terms of consistency, fluency and relevance, but lower correlations in the dimension of coherence. One possible reason for this is that the embedding-based metrics cannot measure the positional information of sentences well. Further, we observe that recently proposed embedding-based metrics generally outperform those without PLMs. One possible explanation is the pre-trained language models enable word embeddings to better capture the exact meaning of words.

(5) We have also observed that semantic embedding-based metrics tend to outperform lexical overlap-based metrics in the dimension of consistency. This may be due to the hard subsequence alignments of lexical overlap-based metrics, which do not measure the factual alignment between candidate summaries and source documents well.

(6) Compared to the proposed method, embedding-based metrics have shown competitive performance in terms of consistency, especially those proposed in recent years using pre-trained language models. However, the embedding-based metrics do not show advantages in

Table 4
A comparison between the ROUGE-SEM variants and their corresponding ROUGE variants based on **SummEval**. Three different variants of ROUGE-SEM with higher correlation (“Pearson”, “Spearman” and “Kendall”) compared to the corresponding variants of ROUGE in each column are bolded. As representatives of error metrics, the average (Ave.) and standard deviation (S.D) values of the evaluation performance of each metric are presented in the last two columns, respectively.

Metrics	Coherence			Consistency			Fluency			Relevance			Error metrics	
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Ave. ↑	S.D ↓
ROUGE-1	0.1972	0.5215*	0.3630*	0.6696**	0.4957*	0.3107	0.6164**	0.6580**	0.4926**	0.6285**	0.6581**	0.5353**	0.5122	0.1523
ROUGE-2	0.2687	0.6410**	0.4519*	0.6860**	0.4550	0.3258	0.6311**	0.7141**	0.6138**	0.6956**	0.8186**	0.6194**	0.5768	0.1657
ROUGE-L	0.1406	0.2000	0.1111	0.5209*	0.3950	0.2270	0.4668	0.3901	0.2804	0.4442	0.3217	0.2528	0.3126	0.1322
ROUGE-SEM-1	**0.2903**	**0.6070****	**0.4059***	**0.6750****	**0.6162****	**0.4605***	**0.6179****	0.5898*	0.4605*	**0.6902****	**0.7632****	**0.6371****	**0.5678**	**0.1357**
ROUGE-SEM-2	**0.3340**	0.6352**	**0.4520***	0.6792**	**0.5176***	**0.3397**	0.6294**	0.6872**	0.5209**	**0.7181****	0.7926**	0.6222**	**0.5773**	**0.1461**
ROUGE-SEM-L	**0.1827**	**0.3691**	**0.2583**	**0.6157****	0.3869	**0.2642**	**0.5488***	**0.4207**	**0.3095**	**0.5726***	**0.4294**	**0.3852**	**0.3953**	0.1331

* Symbol represents significant for $p \leq 0.05$.

** Symbol represents significant for $p \leq 0.01$.

other dimensions, and most of them perform worse than ROUGE-SEM in the dimensions of relevance, coherence, and fluency. In particular, we observe that ROUGE-SEM-1 and ROUGE-SEM-2 have shown better evaluation performance across all dimensions under different coefficients in most cases. Specifically, the average and standard deviation of the evaluation performance of ROUGE-SEM-2 are 57.73% and 14.61%, respectively. Compared to BERTScore-r, ROUGE-SEM-2 exhibits an average improvement of 18.33%, 16.46% and 1.46% with respect to coherence, relevance and fluency, respectively. These results confirm that the proposed ROUGE-SEM is more effective in the dimensions of coherence, relevance, and fluency than embedding-based metrics for automatic summarization evaluation. Although the performance of the proposed method has declined in the dimension of consistency, its comprehensive evaluation performance is better due to improvement of other dimensions.

For a more succinct comparison of ROUGE-SEM and ROUGE, we reconstruct Table 4 from Table 3. As shown in Table 4, we compare three ROUGE-SEM variants (ROUGE-SEM-1, ROUGE-SEM-2, and ROUGE-SEM-L) to their corresponding ROUGE variants (ROUGE-1, ROUGE-2, and ROUGE-L) on SummEval dataset. Note that the variants of ROUGE-SEM with a higher correlation than the corresponding variants of ROUGE are in bold. From this table, we have the following findings.

In particular, we observe that ROUGE-SEM-1 is superior to ROUGE-1 most of the time. Specifically, ROUGE-SEM-1 outperforms ROUGE-1 by 5.56% and 1.66% in the average and standard deviation of the evaluation performance. Compared to ROUGE-1, ROUGE-SEM-1 exhibits an average improvement of 7.38%, 9.19% and 8.95% with respect to coherence, consistency, and relevance, respectively. However, the performance of ROUGE-SEM-1 degrades in the fluency dimension, including a 6.82% decrease under Spearman rank correlation and a 3.21% drop under Kendall rank correlation. These results confirm that the proposed ROUGE-SEM is more effective in the dimensions of coherence, consistency, and relevance for automatic summarization evaluation.

In addition, we also find that ROUGE-SEM-2 performs better than ROUGE-2 when evaluating coherence and consistency. Compared to ROUGE-2, the performance of ROUGE-SEM-2 is slightly improved, including a 1.99% increase in coherence and a 2.32% increase in consistency. Although not as dramatic, these results suggest that the proposed ROUGE-SEM can better evaluate the coherence and consistency of the summarization. With acceptable declines in fluency and relevance assessment, ROUGE-SEM-2 can still be considered a competitive metric for automatic summarization evaluation. In summary, the average evaluation performance of ROUGE-SEM-2 is similar to ROUGE-2, but the standard deviation is lower.

Lastly, ROUGE-SEM-L is found to correlate better than ROUGE-L in all quality dimensions, regardless of the correlation coefficients used. Compared to ROUGE-L, the average evaluation performance of ROUGE-SEM-L has an absolute improvement of 8.27%. Specifically, the average performance of ROUGE-SEM-L in four quality dimensions is greatly improved by 11.95%, 4.13%, 4.72% and 12.28%, respectively. This lends further support to our proposal of using ROUGE combined with semantics.

In most cases, the variants of ROUGE-SEM correlate better than the corresponding variants of ROUGE when measured with Pearson, Spearman and Kendall rank correlation on four quality dimensions. This demonstrates the effectiveness of improving ROUGE with semantic information for better summarization evaluation.

6.2. Results on DialSummEval

Following the experimental setup of Section 6.1, we use Pearson, Spearman, and Kendall rank correlation coefficients to measure the performance of automatic evaluation metrics on DialSummEval. As shown in Table 5, we calculate three correlations between automatic evaluation scores and manual scores in four quality dimensions on DialSummEval. Note that each column in bold represents the evaluation performance of the top five, the last two columns show the average (Ave.) and standard deviation (S.D) values of the evaluation performance, and the symbol * indicates statistical significance. Compared with the findings in the previous section, some similar but not identical conclusions can be drawn from Table 5.

(1) Specifically, metrics that perform well on certain dimensions tend to perform poorly on others. For example, we find that BERTScore-p correlates strongly with human judgments in terms of coherence and fluency, but performs worse in terms of consistency and relevance. In contrast, BERTScore-r performs well on consistency and relevance, but poorly on measuring coherence and fluency. One possible reason for this is that there is a correlation between the quality dimensions, which corroborates with the findings of Gao and Wan (2022).

(2) Among the lexical overlap-based metrics, METEOR and ROUGE-1 have shown better performance in four dimensions, since their average evaluation performance is approximately the same. Similar to the results in SummEval, BLEU performs worse than other metrics on several dimensions. Different from the results on SummEval, METEOR is observed to have a good correlation with consistency, fluency, and relevance when measured with Spearman and Kendall rank correlation. In particular, METEOR has better comprehensive evaluation performance due to the lower standard deviation than ROUGE-1. Compared to METEOR, the performance of ROUGE-SEM is improved in all correlation coefficients. In addition, we will also discuss the comparison of ROUGE-SEM and ROUGE in detail later.

(3) Among semantic embedding-based metrics, SMS has shown better comprehensive performance across all dimensions. As shown in Table 5, the average evaluation performance of SMS is 47.49%, which is higher than recently proposed BARTScore. Additionally, embedding-based metrics are observed to have higher correlations with human judgments when measured with a certain kind of correlation coefficient. We find that BARTScore performs better with Pearson and Spearman rank correlation, but worse with Kendall rank correlation, which is different from the findings of Gao and Wan (2022).

(4) Similar to the results in SummEval, semantic embedding-based metrics tend to perform better than lexical overlap-based metrics in terms of consistency and relevance. We have also observed that recently proposed embedding-based metrics generally outperform those earlier representatives in the dimension of consistency and relevance. These

Table 5
The correlation coefficients (“Pearson correlation (P)”, “Spearman rank correlation (S)” and “Kendall rank correlation (K)”) of annotations computed on **DialSummEval** along four quality dimensions (“Coherence”, “Consistency”, “Fluency” and “Relevance”) between automatic metrics and human judgments. As representatives of error metrics, the average (Ave.) and standard deviation (S.D) values of the evaluation performance of each metric are presented in the last two columns, respectively. The evaluation performance of the top five is bolded in each column.

Metrics	Pearson correlation (P)				Spearman rank correlation (S)				Kendall rank correlation (K)				Error metrics	
	Coherence	Consistency	Fluency	Relevance	Coherence	Consistency	Fluency	Relevance	Coherence	Consistency	Fluency	Relevance	Ave. ↑	S.D ↓
BLEU-1	0.3515	0.3399	0.3035	0.3602	0.3256	0.1385	0.6484*	0.1804	0.2652	0.1429	0.4505*	0.1989	0.3088	0.1440
BLEU-2	0.3143	0.3477	0.2514	0.3704	0.3872	0.1912	0.6791**	0.2442	0.3094	0.1868	0.4945*	0.2431	0.3349	0.1399
ROUGE-1	**0.5889***	0.4203	**0.5756***	0.4040	**0.5589***	0.2484	**0.7846****	0.2992	**0.4199***	0.2967	**0.5604****	**0.4199***	**0.4647**	0.1538
ROUGE-2	0.4710	0.4085	0.4324	0.4096	0.5391*	0.2088	0.7802**	0.2794	**0.4199***	0.2527	**0.5604****	0.3757	0.4281	0.1538
ROUGE-L	**0.5733***	0.3865	**0.5380***	0.3725	0.5215	0.2000	0.7538**	0.2640	0.3978*	0.2308	0.5385**	0.3536	0.4275	0.1622
ROUGE-SU*	0.4709	0.3800	0.4249	0.3785	0.5149	0.2352	0.7802**	0.2904	0.3757	0.2967	**0.5604****	0.3978*	0.4255	0.1454
ROUGE-W	0.5256	0.3094	0.4475	0.2990	**0.5963***	0.1473	0.7099**	0.1760	**0.4862***	0.1868	0.4945*	0.2210	0.3833	0.1848
METEOR	0.3741	0.4203	0.3325	0.4346	0.4840	**0.4242**	**0.7934****	**0.4928**	0.3315	**0.3846**	**0.6044****	**0.4862***	0.4636	**0.1292**
GM	0.4336	0.3545	0.4307	0.3556	0.4444	0.2396	0.7626**	0.2838	0.3094	0.2967	**0.5604****	0.3978*	0.4058	0.1424
VE	0.4653	0.3526	0.4284	0.3479	0.5193	0.2747	0.7714**	0.3014	0.3978*	**0.3626**	0.5385**	**0.4199***	0.4317	0.1337
SMS	0.3293	0.3778	0.2684	0.3962	**0.5809***	**0.4374**	**0.8154****	**0.4664**	0.3978*	**0.4505***	**0.6703****	**0.5083***	**0.4749**	0.1519
MoverScore	0.4976	0.3868	0.4637	0.3822	0.4840	0.2484	0.7802	0.2948	0.3757	**0.3407**	**0.5604****	0.3978*	0.4344	0.1399
BERTScore-p	**0.5671***	0.1108	0.4976	0.0767	**0.6557***	0.1297	0.6484*	0.1144	**0.5304****	0.1429	0.4066	0.1326	0.3344	0.2356
BERTScore-r	0.4332	**0.4509**	0.4184	**0.4634**	0.3916	**0.3011**	0.7363**	0.3630	0.3315	0.2527	**0.5604****	0.3094	0.4177	**0.1311**
BERTScore-f1	0.5296	0.2818	0.4831	0.2689	0.5127	0.1429	0.7363**	0.1782	**0.4420***	0.1868	0.5385**	0.1989	0.3750	0.1896
BLANC	−0.3697	**0.5017***	−0.1791	**0.5607***	−0.0176	**0.6923****	0.3143	**0.8053****	−0.0442	**0.6044****	0.2088	**0.7072****	0.3153	0.3907
BARTScore	0.4951	**0.5453****	**0.5136**	**0.5561***	0.3916	**0.3011**	0.7363**	**0.3630**	0.3315	0.2527	**0.5604****	0.3094	0.4463	0.1434
ROUGE-SEM-1	**0.6140***	**0.4882***	**0.6484***	**0.4700**	**0.6051***	**0.3495**	**0.7978****	**0.4092**	**0.5083***	**0.3407**	**0.6044****	**0.4641***	**0.5250**	**0.1338**
ROUGE-SEM-2	0.4621	0.4151	0.4343	0.4170	0.5017	0.2176	0.7714**	0.2882	**0.4199***	0.2747	**0.5824****	0.3978*	0.4319	0.1467
ROUGE-SEM-L	**0.6018***	**0.4589**	**0.6153***	**0.4424**	0.5017	0.2352	**0.7890****	0.2794	0.3978*	0.2967	**0.5604****	0.3757	0.4629	0.1610

* Symbol represents significant for $p \leq 0.05$.

** Symbol represents significant for $p \leq 0.01$.

Table 6
A comparison between the ROUGE-SEM variants and their corresponding ROUGE variants based on **DialSummEval**. Three different variants of ROUGE-SEM with higher correlation (“Pearson”, “Spearman” and “Kendall”) compared to the corresponding variants of ROUGE in each column are bolded. As representatives of error metrics, the average (Ave.) and standard deviation (S.D) values of the evaluation performance of each metric are presented in the last two columns, respectively.

Metrics	Coherence			Consistency			Fluency			Relevance			Error metrics	
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Ave. ↑	S.D ↓
ROUGE-1	0.5889*	0.5589*	0.4199*	0.4203	0.2484	0.2967	0.5756*	0.7846**	0.5604**	0.4040	0.2992	0.4199*	0.4647	0.1538
ROUGE-2	0.4710	0.5391*	0.4199*	0.4085	0.2088	0.2527	0.4324	0.7802**	0.5604**	0.4096	0.2794	0.3757	0.4281	0.1538
ROUGE-L	0.5733*	0.5215	0.3978*	0.3865	0.2000	0.2308	0.5380*	0.7538**	0.5385**	0.3725	0.2640	0.3536	0.4275	0.1622
ROUGE-SEM-1	**0.6140***	**0.6051***	**0.5083***	**0.4882***	**0.3495**	**0.3407**	**0.6484***	**0.7978***	**0.6044****	**0.4700**	**0.4092**	**0.4641***	**0.5250**	**0.1338**
ROUGE-SEM-2	0.4621	0.5017	**0.4199***	**0.4151**	**0.2176**	**0.2747**	**0.4343**	0.7714*	**0.5824****	**0.4170**	**0.2882**	**0.3978***	**0.4319**	**0.1467**
ROUGE-SEM-L	**0.6018***	0.5017	**0.3978***	**0.4589**	**0.2352**	**0.2967**	**0.6153***	**0.7890***	**0.5604****	**0.4424**	**0.2794**	**0.3757**	**0.4629**	**0.1610**

* Symbol represents significant for $p \leq 0.05$.

** Symbol represents significant for $p \leq 0.01$.

results further corroborate that pre-trained language models enable word embeddings to better capture the full meaning of words.

(5) Compared to embedding-based metrics, we observe that the proposed method has better comprehensive performance across all dimensions with different coefficients in most cases. Compared with SMS, ROUGE-SEM-1 has an absolute improvement of 5.01% in average and 1.81% in standard deviation, which reiterates the effectiveness of using semantic information to improve ROUGE. Furthermore, ROUGE-SEM-1 achieves a dramatic performance improvement over BARTScore, including a 7.87% improvement in average and a 0.96% increase in standard deviation. These results demonstrate that the proposed ROUGE-SEM is more effective than strong embedding-based metrics for automatic summarization evaluation. In general, ROUGE-SEM performs better than embedding-based metrics on DialSummEval in most cases.

As mentioned above, we reconstruct Table 6 from Table 5 for a more detailed comparison of ROUGE-SEM and ROUGE. In detail, we compare ROUGE-SEM-n (ROUGE-SEM-1, ROUGE-SEM-2, and ROUGE-SEM-L) with ROUGE-n (ROUGE-1, ROUGE-2, and ROUGE-L) using three correlation coefficients along four quality dimensions. It is notable that ROUGE-SEM-n performs better than the corresponding ROUGE-n is bolded. From Table 6, we can easily draw the following conclusions.

To start with, we observe that ROUGE-SEM-1 correlates better than ROUGE-1 in four quality dimensions, regardless of the correlation coefficients used. Compared to ROUGE-1, ROUGE-SEM-1 exhibits an improvement of 6.03% and 2.00% regarding average and standard deviation, respectively. Specifically, ROUGE-SEM-1 achieves an average performance improvement of 5.21%, 7.10%, 4.33% and 7.34% in terms of coherence, consistency, fluency and relevance, respectively. Similar to the results of SummEval, these results further demonstrate the comprehensive capabilities of ROUGE-SEM in terms of automatic summarization evaluation.

Moreover, ROUGE-SEM-2 is observed to outperform ROUGE-2 in terms of consistency, fluency, and relevance, but performs poorly in coherence. Compared to ROUGE-2, the performance of ROUGE-SEM-2 is slightly improved, including a 0.38% increase in average and a 2.00% increase in standard deviation. On the one hand, ROUGE-SEM-2 has improved average performance compared to ROUGE-2, with a 1.25% increase in consistency, a 0.50% improvement in fluency, and a 1.28% increase in relevance. On the other hand, the performance of ROUGE-SEM-2 is slightly decreased in coherence, including a 0.89% decrease under Pearson correlation and a 3.74% drop under Spearman rank correlation. Similar to the results on SummEval, these results still demonstrate that ROUGE-SEM-2 is an effective improvement of ROUGE for better evaluating the consistency, fluency, and relevance of candidate summaries.

Lastly, we also observe that ROUGE-SEM-L performs better than ROUGE-L in most cases. Except for a slight decrease of 1.98% in coherence measured by Spearman rank correlation, the other cases have different degrees of improvement. Compared to ROUGE-L, the average performance of ROUGE-SEM-L in terms of consistency, fluency, and relevance is improved by 5.78%, 4.48% and 3.58%, respectively. In particular, there is an improvement of 3.54% between ROUGE-SEM-L and ROUGE-L in terms of the average evaluation performance.

Table 7
Ablation study of ROUGE-SEM on the SummEval dataset.

Component	Coherence			Consistency			Fluency			Relevance			Error metrics	
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Ave. ↑	S.D ↓
w/o Classifier	0.2819	0.3433	0.2583	0.5654*	0.1429	0.0830	0.5352*	0.3233	0.2491	0.6466**	0.5669*	0.4148*	0.3676	0.1795
w/o Rewriter	0.1724	0.5518*	0.3911*	0.6666**	0.6174**	0.4454*	0.5961*	0.5947*	0.4454*	0.5885*	0.6933**	0.4741*	0.5197	0.1442
with BERTScore	0.2889	0.4243	0.3173	0.6069**	0.1713	0.0981	0.5760*	0.3862	0.2944	0.6555**	0.6491**	0.4889*	0.4131	0.1862
ROUGE-SEM	**0.2903**	**0.6750****	**0.6179****	**0.6902****	0.6070**	**0.6162****	0.5898*	**0.7632****	0.4059*	0.4605*	0.4605*	**0.6371****	**0.5678**	**0.1357**

* Symbol represents significant for $p \leq 0.05$.

** Symbol represents significant for $p \leq 0.01$.

In general, ROUGE-SEM performs better than ROUGE on DialSummEval in most cases. The above results further confirm the validity of using semantics to improve ROUGE for better summarization evaluation.

6.3. Ablation study

To assess the relative contribution of different components, we compare the proposed ROUGE-SEM with three ablated variants in an ablation analysis. In this analysis, we first remove the removable components of ROUGE-SEM, including the candidate summary classifier ("w/o Classifier") and the categorized summary rewriter ("w/o Rewriter"). In another experiment ("with BERTScore"), we replace the semantic similarity module with BERTScore to obtain the semantic similarity scores. It is important to note that the f1 score of BERTScore is used in this experiment. Due to the limited space, the ablation results of the unigram-overlap ROUGE-SEM on the SummEval dataset are reported in Table 7.

From Table 7, we can easily find that ROUGE-SEM shows better comprehensive performance in four dimensions. Specifically, the average and standard deviation of the evaluation performance of ROUGE-SEM are 56.78% and 13.57%, respectively. Furthermore, we find that "w/o Rewriter" performs better than "w/o Classifier" in most cases, indicating that the candidate summary classifier plays a more important role in ROUGE-SEM. The candidate summary classifier has a positive impact on the performance of summarization evaluation, and the categorized summary rewriter can further improve the performance. Compared to the performance of "with BERTScore", ROUGE-SEM exhibits a significant improvement when measured with Spearman and Kendall rank correlation. This result shows that the proposed semantic similarity module can better measure the semantic similarity between candidate summaries and reference summaries than BERTScore.

The above ablation results demonstrate that the removed or replaced components of ROUGE-SEM are essential for better summarization evaluation. In other words, the candidate summary classifier, the categorized summary rewriter, and the semantic similarity module together contribute to the performance of automatic summarization evaluation, validating the effectiveness of the individual components of ROUGE-SEM.

6.4. Effects of different hyperparameters settings

As mentioned in Section 6.1, different settings of hyperparameters α and β have a significant impact on the performance of ROUGE-SEM. Specifically, the hyperparameter α is used to judge whether the candidate summary is semantically similar to the reference summary, and the hyperparameter β determines whether the candidate summary is lexically similar to the reference summary. To study the effect of hyperparameters on ROUGE-SEM, we use the control variable method to carry out experiments with different settings. The experimental results of the unigram-overlap ROUGE-SEM on the SummEval dataset measured by Pearson correlation are reported in Table 8.

From Table 8, we can find the following observations. It is consistent with our intuition that the proportions of good-summary and pearl-summary decrease as α increases, and the proportions of good-summary and glass-summary decrease as β increases. Furthermore, the proposed ROUGE-SEM exhibits competitive performance when the average of semantic similarity scores and lexical similarity scores are chosen as hyperparameters α and β, respectively. Even when α is slightly larger than the average of semantic similarity score and β is unchanged, ROUGE-SEM shows a stronger correlation with human judgments, indicating that the proposed method still has room for performance improvement. Based on the above findings, the suggested hyperparameter α is the average of semantic similarity scores or slightly greater than the average, and the suggested hyperparameter β to be the average of lexical similarity scores.

6.5. Limitations

No metric is perfect, and without exception, our proposed method has some limitations that should be carefully considered.

Semantic similarity. One of the limitations is the use of meaningful word embeddings. The proposed classification of candidate summaries relies on semantic similarity and lexical similarity, but it is challenging to accurately measure semantic similarity. We adopt a Siamese-BERT network with contrastive learning as the semantic similarity module to address this issue. However, this can be affected by many potential factors, including the choice of text encoders, the quality of pre-trained language models, and the representativeness of word embeddings. Specifically, BERT encoder may be biased towards certain types of language and may not account for the diversity of expressions and writing styles. We propose to choose other task-specific text encoders (Gaci, Benatallah, Casati, & Benabdeslem, 2022) instead of BERT encoder to mitigate the bias of text encoders. In addition, word embeddings may be biased towards certain contexts and may not capture the full meaning of words. We propose to use the recently proposed bias-reduced word embeddings (An, Liu, & Zhang, 2022) to measure semantic similarity, which not only improves accuracy of semantic similarity measurement, but also further improves evaluation performance of the proposed metric. Especially for texts in different domains or languages, their nuances and contextual meanings may not be accurately captured. To mitigate this, we propose to replace the standard bert-base-uncased with domain- or language-specific pre-trained language models on HuggingFace, which may play a key role. To further improve the semantic similarity, it is necessary to use domain- or language-specific text summarization datasets to train the semantic similarity module according to this paper.

Evaluation time. Another limitation is the use of pre-trained language models. To improve the evaluation performance of text summarization, sophisticated calculations are performed on the similarity between candidate and standard summaries, which leads to increased computational complexity. More precisely, the increase in computational complexity mainly comes from the semantic similarity module due to the use of pre-trained language models. During the real evaluation process, since more attention is paid to the evaluation time, the computation time is adopted to measure the computational complexity of the evaluation process. Like most semantic embedding-based methods using PLMs, our evaluation process requires more time to

Table 8
Summarization evaluation performance of ROUGE-SEM on the SummEval dataset using different settings of hyperparameters α and β. Bold shows the best performance.

α	β	Good-summary	Pearl-summary	Glass-summary	Bad-summary	Coherence	Consistency	Fluency	Relevance
0.4979	0.2839	49.52%	47.17%	0.17%	3.14%	0.2246	0.6072**	0.5558*	0.6037*
0.5979		48.00%	39.17%	1.70%	11.13%	0.1981	0.5893*	0.5389*	0.5879*
0.6979		36.47%	17.88%	13.23%	32.42%	0.3029	0.6811**	0.6203**	0.6909**
0.7979		9.88%	2.29%	39.82%	48.01%	**0.4433**	**0.6925****	**0.6576****	**0.7381****
0.8979		0.00%	0.00%	49.70%	50.30%	0.3982	0.1560	0.4738	0.6163*
0.6979	0.0839	54.35%	0.00%	45.58%	0.07%	0.2079	0.6565**	0.5924*	0.6273**
	0.1839	54.00%	0.35%	39.88%	5.77%	0.2777	0.6563**	0.5997*	0.6591**
	0.2839	36.47%	17.88%	13.23%	32.42%	**0.3029**	**0.6811****	**0.6203****	**0.6909****
	0.3839	4.82%	49.52%	0.88%	44.78%	0.2265	0.6374**	0.5703*	0.6512**
	0.4839	0.11%	54.23%	0.00%	45.66%	0.2608	0.6091**	0.5649*	0.6710**

* Symbol represents significant for $p \leq 0.05$.

** Symbol represents significant for $p \leq 0.01$.

train the semantic similarity module, while the inference time for automatic summarization evaluation is acceptable. Due to the particularity of measuring semantic similarity, there is no need to retrain for texts in the same source (e.g., domain and language). Therefore, we choose inference time as evaluation time overhead instead of training time when measuring computational complexity. Compared with ROUGE, the evaluation time of our method is relatively long, but the performance improvement makes up for this.

7. Conclusion and future work

In this paper, we propose a new evaluation metric, ROUGE-SEM, that enhances the popular ROUGE by combining semantic information. To achieve this, the candidate summary classifier, categorized summary rewriter, and rewritten summary scorer as the main components constitute the pipeline framework in a manner consistent with human behavior. Specifically, the candidate summary classifier employs a semantic similarity module to calculate semantic similarity and a lexical similarity module to calculate lexical similarity between candidate summaries and reference summaries. Then, candidate summaries are classified into four groups according to the differences of semantic similarity and lexical similarity, including good-summary, pearl-summary, glass-summary and bad-summary. For pearl-summary and glass-summary which are incorrectly evaluated by ROUGE, the categorized summary rewriter employs the back-translation technique to alleviate lexical bias through more diverse synonymous expressions. Finally, the rewritten summary scorer outputs more accurate evaluation scores based on the results of the candidate summary classifier and the categorized summary rewriter. The experimental results show that the performance of ROUGE-SEM is comparable to existing strong baselines and widely used metrics, measured with three coefficients. In particular, variants of ROUGE-SEM consistently outperform corresponding variants of ROUGE.

In future work, we will employ some task-specific pre-trained language models as semantic encoders for more accurate semantic similarity. We will consider replacing the back-translation module with various text-generation models for offline evaluation. In addition, we will adopt more efficient parameter optimization strategies for parameter tuning. Finally, we will apply the proposed metric to evaluate existing baselines and state-of-the-art summarizers. We hope that this work can have a positive impact on future research on text summarization systems.

CRediT authorship contribution statement

Ming Zhang: Conceptualization, Methodology, Software, Writing – original draft. **Chengzhang Li:** Software, Data curation, Investigation. **Meilin Wan:** Validation, Writing – review & editing. **Xuejun Zhang:** Visualization, Writing – review & editing. **Qingwei Zhao:** Funding acquisition, Project administration, Supervision.

Table A.1
Abbreviations of phrases and models.

Abbreviation	Full name
ATS	Automatic Text Summarization
Ave.	Average value
NLP	Natural Language Processing
PLMs	Pre-trained Language Models
QA	Question Answering
BERT	Bidirectional Encoder Representations from Transformers
BLEU	BiLingual Evaluation Understudy
ENMS	Enhance existing N-gram based evaluation Metrics with Semantics
FactCC	Factual Consistency Checking
FEQA	Faithfulness Evaluation with Question Answering
GM	Greedy Matching
METEOR	Metric for Evaluation of Translation with Explicit ORdering
QAFactEval	QA-based Factual consistency Evaluation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
ROUGE-SEM	ROUGE with SEMantics
ROUGE-WE	ROUGE with Word Embedding
SDC	Semantic Distribution Correlation
SPEED	Sentence Pair EmbEDdings score
SLN	Semantic Link Network
SEM-nCG	SEMantic-aware nCG(normalized cumulative gain)-based evaluation metric
SMS	Sentence Mover Similarity
S.D	Standard Deviation value
SUPERT	SUmmarization evaluation with Pseudo references and bERT
S^3	Supervised Summarization Scorer
VE	Vector Extrema

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and constructive guidance that improved this paper. This research has been supported by the Youth Innovation Promotion Association of the Chinese Academy of Sciences, China (E1291902), Jun Zhou (2021025).

Appendix A. Abbreviations

For your convenience, phrases and model abbreviations commonly used in this paper are listed in Table A.1.

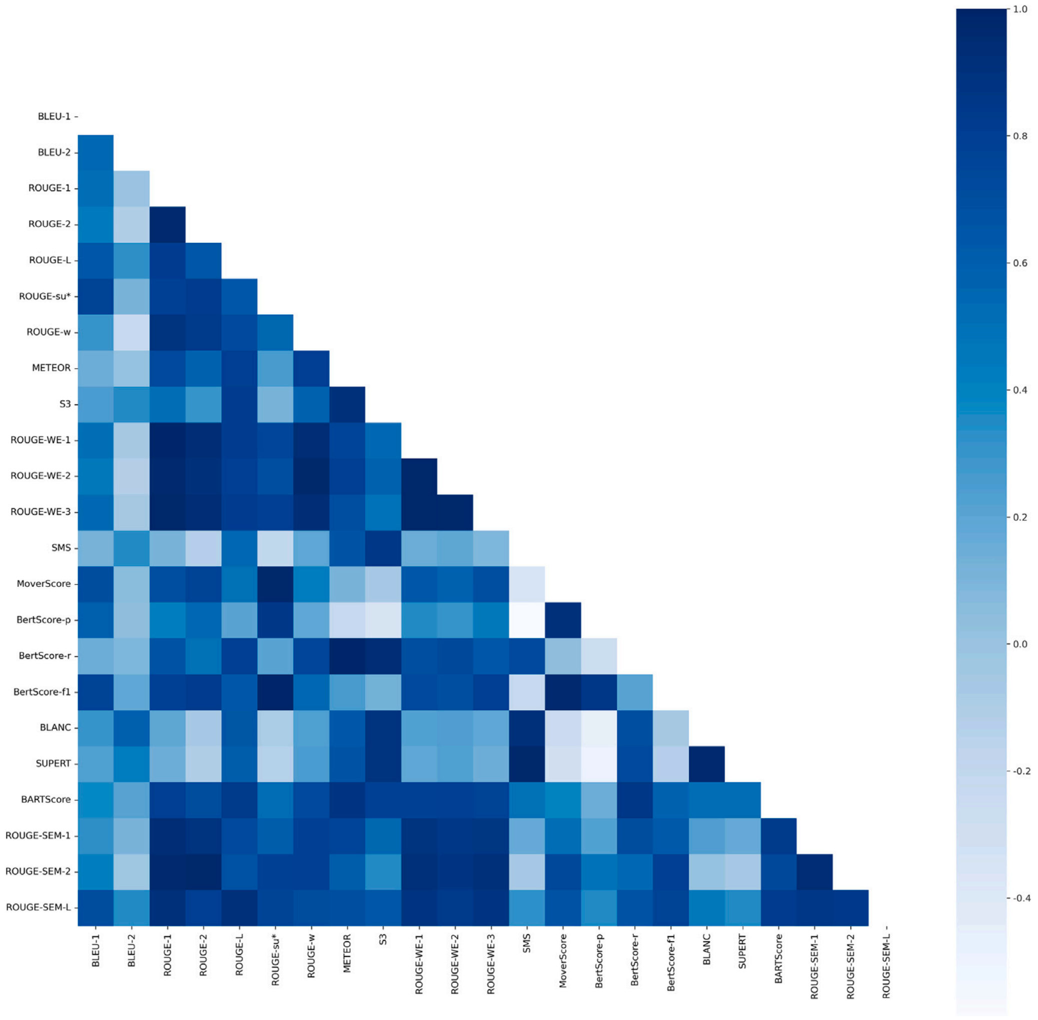


Fig. B.1. The statistical correlations between different automatic evaluation metrics.

Appendix B. Statistical correlations between different metrics.

Fig. B.1 shows the system-level Pearson correlation between different automatic evaluation metrics.

References

Akter, M., Bansal, N., & Santu, S. K. K. (2022). Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In *Findings of the association for computational linguistics* (pp. 1547–1560). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.findings-acl.122.

Akula, R., & Garibay, I. (2022). Sentence pair embeddings based evaluation metric for abstractive and extractive summarization. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6009–6017). European Language Resources Association, URL: https://aclanthology.org/2022.lrec-1.646.

An, H., Liu, X., & Zhang, D. (2022). Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the association for computational linguistics* (pp. 1139–1152). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.findings-acl.90.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Ann Arbor, Michigan: Association for Computational Linguistics, URL: https://aclanthology.org/W05-0909.

Barbella, M., Risi, M., & Tortora, G. (2021). A comparison of methods for the evaluation of text summarization techniques. In *Proceedings of the 10th international conference on data science, technology and applications* (pp. 200–207). SCITEPRESS, http://dx.doi.org/10.5220/0010523002000207.

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Soc. Networks Media*, *24*, Article 100153. http://dx.doi.org/10.1016/j.osnem.2021.100153.

Bhandari, M., Gour, P. N., Ashfaq, A., Liu, P., & Neubig, G. (2020). Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 9347–9359). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.751.

Cao, M., & Zhuge, H. (2022). Automatic evaluation of summary on fidelity, conciseness and coherence for text summarization based on semantic link network. *Expert Systems with Applications*, *206*, Article 117777. http://dx.doi.org/10.1016/j.eswa.2022.117777.

Clark, E., Celikyilmaz, A., & Smith, N. A. (2019). Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2748–2760). Florence, Italy: Association for Computational Linguistics, URL: https://aclanthology.org/P19-1264.

Cohen, N., Kalinsky, O., Ziser, Y., & Moschitti, A. (2021). WikiSum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol. 2* (pp. 212–219). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-short.28.

Crystal, M., Baron, A., Godfrey, K., Micciulla, L., Tenney, Y. J., & Weischedel, R. M. (2005). A methodology for extrinsically evaluating information extraction performance. In *HLT/EMNLP 2005, human language technology conference and conference on empirical methods in natural language processing, proceedings of the conference* (pp. 652–659). The Association for Computational Linguistics, URL: https://aclanthology.org/H05-1082/.

Deutsch, D., Dror, R., & Roth, D. (2021). A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, *9*, 1132–1146. http://dx.doi.org/10.1162/tacl_a_00417.

Deutsch, D., & Roth, D. (2020). SacreROUGE: An open-source library for using and developing summarization evaluation metrics. CoRR abs/2007.05374, URL: https://arxiv.org/abs/2007.05374.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 1* (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1423.

Dong, Y., Shen, Y., Crawford, E., van Hoof, H., & Cheung, J. C. K. (2018). BanditSum: Extractive summarization as a contextual bandit. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3739–3748). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d18-1409.

Durmus, E., He, H., & Diab, M. T. (2020). FEQA: a question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5055–5070). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.454.

Egan, N., Vasilyev, O. V., & Bohannon, J. (2022). Play the Shannon game with language models: A human-free approach to summary evaluation. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence* (pp. 10599–10607). AAAI Press, URL: https://ojs.aaai.org/index.php/AAAI/article/view/21304.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, Article 113679. http://dx.doi.org/10.1016/j.eswa.2020.113679.

Ermakova, L., Cossu, J., & Mothe, J. (2019). A survey on evaluation of summarization methods. *Information Processing and Management*, *56*(5), 1794–1814. http://dx.doi.org/10.1016/j.ipm.2019.04.001.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, *9*, 391–409.

Fabbri, A. R., Li, I., She, T., Li, S., & Radev, D. R. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th conference of the association for computational linguistics, vol. 1: Long Papers* (pp. 1074–1084). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-1102.

Fabbri, A. R., Wu, C., Liu, W., & Xiong, C. (2022). QaFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 2587–2601). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.naacl-main.187.

Forgues, G., Pineau, J., Larchevêque, J.-M., & Tremblay, R. (2014). Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop, vol. 2* (p. 168).

Gaci, Y., Benatallah, B., Casati, F., & Benabdeslem, K. (2022). Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9582–9602). Association for Computational Linguistics, URL: https://aclanthology.org/2022.emnlp-main.651.

Ganesan, K. (2018). ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. CoRR abs/1803.01937, URL: http://arxiv.org/abs/1803.01937.

Gao, M., & Wan, X. (2022). DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 5693–5709). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.naacl-main.418.

Gao, Y., Zhao, W., & Eger, S. (2020). SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1347–1354). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.124.

Garg, M., & Kumar, M. (2022). KEST: A graph-based keyphrase extraction technique for tweets summarization using Markov decision process. *Expert Systems with Applications*, *209*, Article 118110.

Ghadimi, A., & Beigy, H. (2022). Hybrid multi-document summarization using pre-trained language models. *Expert Systems with Applications*, *192*, Article 116292.

Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd workshop on new frontiers in summarization* (pp. 70–79). Hong Kong, China: Association for Computational Linguistics, URL: https://aclanthology.org/D19-5409.

Haveliwala, T. H. (2003). Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, *15*(4), 784–796. http://dx.doi.org/10.1109/TKDE.2003.1208999.

He, J., Jiang, W., Chen, G., Le, Y., & Ding, X. (2022). Enhancing N-gram based metrics with semantics for better evaluation of abstractive text summarization. *Journal of Computer Science and Technology*, *37*(5), 1118–1133. http://dx.doi.org/10.1007/s11390-022-2125-6.

Ke, P., Zhou, H., Lin, Y., Li, P., Zhou, J., Zhu, X., et al. (2022). CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2306–2319). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.164.

Koto, F., Baldwin, T., & Lau, J. H. (2022). FFCI: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, *73*, http://dx.doi.org/10.1613/jair.1.13167.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 9332–9346). Online: Association for Computational Linguistics, URL: https://aclanthology.org/2020.emnlp-main.750.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From word embeddings to document distances. In *JMLR workshop and conference proceedings*: *vol. 37*, *Proceedings of the 32nd international conference on machine learning* (pp. 957–966). JMLR.org, URL: http://proceedings.mlr.press/v37/kusnerb15.html.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics, URL: https://aclanthology.org/W04-1013.

Lin, W., Li, S., Zhang, C., Ji, B., Yu, J., Ma, J., et al. (2022). SummScore: A comprehensive evaluation metric for summary quality based on cross-encoder. In *Lecture notes in computer science*: *vol. 13422*, *Web and big data - 6th international joint conference, APWeb-WAIM 2022, Nanjing, China, November 25-27, 2022, proceedings, Part II* (pp. 69–84). Springer.

Liu, L., Dou, Q., Chen, H., Qin, J., & Heng, P. (2020). Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE Transactions on Medical Imaging*, *39*(3), 718–728. http://dx.doi.org/10.1109/TMI.2019.2934577.

Liu, Y., Jia, Q., & Zhu, K. Q. (2022). Reference-free summarization evaluation via semantic correlation and compression ratio. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 2109–2115). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.naacl-main.153.

Liu, S., Yang, L., & Cai, X. (2022). SEASum: Syntax-enriched abstractive summarization. *Expert Systems with Applications*, *199*, Article 116819.

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, *37*(1), 1–41. http://dx.doi.org/10.1007/s10462-011-9216-z.

Mani, I. (2001). Summarization evaluation: An overview. In *Proceedings of the third second workshop meeting on evaluation of Chinese & Japanese text retrieval and text summarization*. National Institute of Informatics (NII), URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41. http://dx.doi.org/10.1145/219717.219748.

Mohd, M., Jan, R., & Shah, M. B. (2020). Text document summarization using word embedding. *Expert Systems with Applications*, *143*, http://dx.doi.org/10.1016/j.eswa.2019.112958.

Mueller, J., & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the thirtieth AAAI conference on artificial intelligence* (pp. 2786–2792). AAAI Press, URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12195.

Nallapati, R., Zhou, B., dos Santos, C. N., Gülçehre, Ç., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL conference on computational natural language learning* (pp. 280–290). ACL, http://dx.doi.org/10.18653/v1/k16-1028.

Ng, J., & Abrecht, V. (2015). Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1925–1930). The Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d15-1222.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). USA: Association for Computational Linguistics.

Peyrard, M., Botschen, T., & Gurevych, I. (2017). Learning to score system summaries for better content selection evaluation. In *Proceedings of the workshop on new frontiers in summarization* (pp. 74–84). Copenhagen, Denmark: Association for Computational Linguistics, URL: https://aclanthology.org/W17-4510.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). *Improving language understanding by generative pre-training*. OpenAI.

Rani, R., & Lobiyal, D. K. (2021). A weighted word embedding based approach for extractive text summarization. *Expert Systems with Applications*, *186*, Article 115867. http://dx.doi.org/10.1016/j.eswa.2021.115867.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3980–3990). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1410.

Rus, V., & Lintean, M. (2012). An optimal assessment of natural language student input using word-to-word similarity metrics. In *Intelligent tutoring systems: 11th international conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11* (pp. 675–676). Springer.

Schluter, N. (2017). The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics, vol. 2: short papers* (pp. 41–45). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/e17-2007.

Scialom, T., Dray, P., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., et al. (2021). QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6594–6604). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.529.

Scialom, T., Lamprier, S., Piwowarski, B., & Staiano, J. (2019). Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3246–3256). Hong Kong, China: Association for Computational Linguistics, URL: https://aclanthology.org/D19-1320.

ShafieiBavani, E., Ebrahimi, M., Wong, R. K., & Chen, F. (2018). A graph-theoretic summary evaluation for rouge. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 762–767). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d18-1085.

Shapira, O., Pasunuru, R., Ronen, H., Bansal, M., Amsterdamer, Y., & Dagan, I. (2021). Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 657–677). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.54.

Sugiyama, A., & Yoshinaga, N. (2019). Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the fourth workshop on discourse in machine translation* (pp. 35–44). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-6504.

Vasilyev, O., Dharnidharka, V., & Bohannon, J. (2020). Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 11–20). Online: Association for Computational Linguistics, URL: https://aclanthology.org/2020.eval4nlp-1.2.

Wang, L. L., Otmakhova, Y., DeYoung, J., Truong, T. H., Kuehl, B., Bransom, E., et al. (2023). Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 9871–9889). Association for Computational Linguistics, URL: https://aclanthology.org/2023.acl-long.549.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2019). HuggingFace's transformers: State-of-the-art natural language processing. CoRR abs/1910.03771, URL: http://arxiv.org/abs/1910.03771.

Xiao, L., He, H., & Jin, Y. (2022). FusionSum: Abstractive summarization with sentence fusion and cooperative reinforcement learning. *Knowledge-Based Systems*, *243*, Article 108483. http://dx.doi.org/10.1016/j.knosys.2022.108483.

Xie, Q., Bishop, J., Tiwari, P., & Ananiadou, S. (2022). Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, *252*, Article 109460. http://dx.doi.org/10.1016/j.knosys.2022.109460.

Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating generated text as text generation. In *Advances in neural information processing systems 34: annual conference on neural information processing systems 2021* (pp. 27263–27277). URL: https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *8th international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=SkeHuCVFDr.

Zhao, B., & Lui, Y. M. (2022). Towards a reliable text summarization evaluation metric using predictive models. *International Journal of Pattern Recognition and Artificial Intelligence*, *36*(10), 2251011:1–2251011:35. http://dx.doi.org/10.1142/S0218001422510119.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 563–578). Hong Kong, China: Association for Computational Linguistics, URL: https://aclanthology.org/D19-1053.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive summarization as text matching. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6197–6208). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.552.