

# SummScore: A Comprehensive Evaluation Metric for Summary Quality Based on Cross-Encoder

Wuhang Lin<sup>1\*</sup>, Shasha Li<sup>1\*</sup>, Chen Zhang<sup>1</sup>, Bin Ji<sup>1</sup>, Jie Yu<sup>1✉</sup>, Jun Ma<sup>1</sup>, and Zibo Yi<sup>2</sup>

<sup>1</sup> College of Computer, National University of Defense Technology, Changsha, China  
wuhang\_lin@163.com, shashali@nudt.edu.cn, chenzhang199705@163.com,

jibin@nudt.edu.cn, yj@nudt.edu.cn, majun@nudt.edu.cn

<sup>2</sup> Information Research Center of Military Science PLA Academy of Military Science  
100142 Beijing, China  
ziboyi@outlook.com

**Abstract.** Text summarization models are often trained to produce summaries that meet human quality requirements. However, the existing evaluation metrics for summary text are only rough proxies for summary quality, suffering from low correlation with human scoring and inhibition of summary diversity. To solve these problems, we propose SummScore, a comprehensive metric for summary quality evaluation based on Cross-Encoder. Firstly, by adopting the original-summary measurement mode and comparing the semantics of the original text, SummScore gets rid of the inhibition of summary diversity. With the help of the text-matching pre-training Cross-Encoder, SummScore can effectively capture the subtle differences between the semantics of summaries. Secondly, to improve the comprehensiveness and interpretability, SummScore consists of four fine-grained submodels, which measure Coherence, Consistency, Fluency, and Relevance separately. We use semi-supervised multi-rounds of training to improve the performance of our model on extremely limited annotated data. Extensive experiments show that SummScore significantly outperforms existing evaluation metrics in the above four dimensions in correlation with human scoring. We also provide the quality evaluation results of SummScore on 16 mainstream summarization models for later research.

**Keywords:** SummScore · Comprehensive metric · Summary quality evaluation.

## 1 Introduction

Automatic text summarization technology aims to compress a long document into a fluent short text, which is consistent with the key information of the original text and preserves the most salient information in the source document [6]. In recent years, automatic text summarization technologies have been significantly developed. However, the research on automatic summarization evaluation still fell behind [7]. Today, the mainstream evaluation metrics for automatic text

summarization, such as ROUGE, BLEU, and Meteor, simply calculate n-gram overlap between candidates and references [1,11,14]. Studies [12,19] have shown that they are only rough proxies for summary quality evaluation. Some concerns of these metrics are shown as follows.

Firstly, the existing evaluation metrics strongly rely on expert-generated summaries as references, which are difficult to obtain. What’s more, these metrics inhibit the diversity of summaries generated by the summarization model. Because the mainstream metrics only rely on the interaction between the reference summary. However, different summaries written by readers with different knowledge reserves and for different purposes are also correct. We cannot force different summaries to be evaluated simply by measuring the degree of alignment with a single reference summary. Such an evaluation metric will limit the diversity of summaries generated by the summarization model.

Secondly, some studies show that the mainstream evaluation metrics scoring do not correlate well with human scoring [3,19]. When humans evaluate the quality of summaries, they usually consider multiple fine-grained quality dimensions, such as rich information, non-redundancy, coherence, and well-structured. However, these metrics mainly focus on the similarity of literal and expressions, which cannot well evaluate semantic relevance and topic consistency. Moreover, they ignore the evaluation of language quality, such as logical consistency and language fluency. Many of the above-mentioned factors can affect the comprehensiveness and interpretability of the summary quality evaluation.

As illustrated in the examples in Figure. 1. Comparing the reference with the original text, when experts score the summary generated by model Bottom-Up, they find that the generated summary has *factual errors*(gray shaded fonts). The fact is that *Manuel Pellegrini (Manchester City)* wants to sign *Evangelos Patoulidis*. Therefore, except for Fluency, the experts give low scores for all quality dimensions. However, because of *the large overlap of n-grams*(blue fonts) between the summary and the reference, ROUGE scores high. For the BART model, because the generated summaries almost focus on *the important information* (orange fonts), and the text is of high quality and no redundant information. So, experts give it high marks. However, the wording is different from the reference summary, so ROUGE gives the summary a low rating. It can be seen from these two examples that ROUGE is a rough proxy that is unable to recognize semantic factual errors. Moreover, over-reliance on the literal matching of reference may lead to a suppression of the diversity of generated summaries. Therefore, a good summarization evaluation metric should be able to help identify: (i) semantically correct summaries with good word overlap with the original text or reference, and (ii) non-redundant and fluent summaries that contain enough correct facts, even if their wording is different from the reference.

To solve these problems, we propose SummScore, a comprehensive metric for summary quality evaluation based on Cross-Encoder. SummScore adopts the original-summary paired measurement mode. The summaries are scored by comparing the semantics of the original text, avoiding the suppression of the diversity of the summaries caused by the forced alignment of a single reference summary.

<b>Original Text:</b> Manchester City are keen to sign Anderlecht teenager Evangelos Patoulidis. The 14-year-old playmaker is regarded as one of the best talents to emerge from Anderlecht's youth set-up and has also attracted attention from Arsenal and Barcelona. The Belgian starlet rejected a move to Barcelona's La Masia academy when he was 12 as his family wanted him to continue his studies. He has continued to impress and City have held discussions with Anderlecht chairman Roger Vanden Stock in the hope of agreeing a compensation package. Manuel Pellegrini is looked to build for the future by snapping up hot property Evangelos Patoulidis.				
<b>Reference:</b> evangelos patoulidis also attracted interest from barcelona and arsenal. anderlecht rejected a move to barcelona when he was 12. city in talks with anderlecht chief roger vanden stock to complete a deal.				
<b>Bottom-Up:</b> evangelos patoulidis has been linked with arsenal and barcelona. the belgian starlet rejected a move to barcelona. anderlecht chairman roger vanden is keen to sign manuel pellegrini.				
Human	Coherence: 2.33,	Consistency: 1.67,	Fluency: 5.00,	Relevance: 2.67
ROUGE	ROUGE_1: 0.508,	ROUGE_2: 0.211,	ROUGE_L: 0.407	
<b>BART:</b> manuel pellegrini is keen to sign anderlecht youngster evangelos patoulidis. The 14-Year-old playmaker is regarded as one of the best talents to emerge from the belgian club 's youth set-up. arsenal and barcelona are also interested in the youngster.				
Human	Coherence: 5.00,	Consistency: 5.00,	Fluency: 5.00,	Relevance: 5.00
ROUGE	ROUGE_1: 0.324,	ROUGE_2: 0.028,	ROUGE_L: 0.162	

**Fig. 1.** A typical example showing ROUGE’s problems.

With the help of the text-matching pre-training Cross-Encoder, SummScore can effectively capture the subtle differences between the semantics of summaries. To improve the comprehensiveness and interpretability, SummScore consists of four fine-grained submodels, which measure Coherence, Consistency, Fluency, and Relevance separately.

We conduct our experiment in SummEval[7] dataset and measure the quality of our SummScore by calculating the Pearson correlation and Spearman correlation coefficient between SummScore scores and human annotation scores. We use semi-supervised multi-rounds of training to improve the performance of our model on extremely limited annotated data. Extensive experiments show that SummScore significantly outperforms existing evaluation metrics. In addition, we evaluate 16 mainstream summarization models with SummScore and publish the results for later research. Our contributions are summarized as follows:

- We propose SummScore, a novel evaluation metric for summary quality, which uses original text instead of the hard-to-obtain expert-generated gold summary as the reference to evaluate the quality of the generated summary.
- We trained four submodels of SummScore based on the Cross-Encoder framework to automatically evaluate the four fine-grained qualities of Relevance, Consistency, Fluency, and Coherence respectively. Experiments show SummScore has strong human relevance on all the four fine-grained dimensions.
- We evaluate 16 mainstream summarization models with SummScore and publish the results for later research.

## 2 Related Work

In this section, we will first introduce the common metrics on summarization and their main problems. Next, we will introduce the context-dependent metrics and the trained metrics in the evaluation of related natural language generation tasks. By borrowing the principles and advantages of the context-dependent metrics and the trained metrics, we design SummScore for summary quality evaluation.

**Common Metrics in Summarization** The early common summary metrics are mainly represented by ROUGE [11], BLEU [14] and METEOR [1]. All of them obtain the summary quality score by calculating the token n-gram overlap between the summary and the reference. However, these lexical-based overlap metrics cannot capture the changes in semantics and grammar. Therefore, BERTscore [20] and MoverScore [21] use BERT to extract contextual embeddings and use embeddings matching to complete the similarity calculation between summary and references. However, these metrics, which rely on the alignment of single-reference abstracts, bring about suppression of abstract diversity.

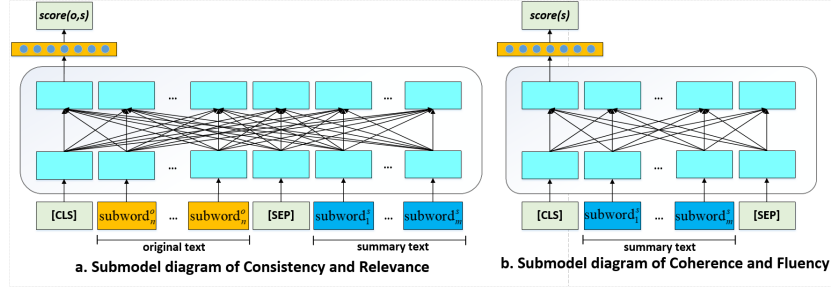
**Context-dependent Metrics** To get rid of the constraints of reference summaries, ROUGE-C [9] improves ROUGE, which compares summaries with the original texts instead of reference summaries. ROUGE-C proves that using original text instead of reference can yield positive benefits, especially when the reference summary is not available. SUPERT [8] is an unsupervised reference-less summarization evaluation metric. SUPERT enables the quality assessment of the generated summaries with the help of pseudo-reference summaries created by salient sentences from the original text. Our model is also a context-dependent metric. Experiments show that our method not only gets rid of the comfort of reference summary but also supports the diversity of summary text generation.

**Trained Metrics** There are training-based evaluation metrics in related natural language generation tasks. For machine translation, BLEND [13] and BEER [18] train a scoring model by combining a variety of existing untrained metrics, such as BLEU, METEOR, and ROUGE. As the pre-training models show promising performance, BERT for MTE [17] and BLEURT [16] are proposed for the machine translation system. By performing BERT fine-tuning training on a small amount of labeled data, they compute the similarity of the candidate and reference sentences. The difference is that BLEURT innovatively designs a set of pre-training signals and pre-trains BERT. We propose a trained-based summary evaluation metric SummScore, which consists of four submodels, corresponding to four quality dimensions. We believe that a single model may not be able to take into account the evaluation of various quality dimensions of the summary texts. At the same time, the independent scoring of multiple dimensions also helps to improve the interpretability of the summary quality score.

### 3 Our Methodology

#### 3.1 Problem Definition

Our SummScore model is based on the **Cross-Encoder** [5] model in the field of information retrieval. In QA(Question answering) retrieval, when sorting the candidate answers, the higher the similarity score between the answer and the question, the more accurate the answer is considered. The specific process can be realized by stitching the subword sequences of question text and answer text with



**Fig. 2.** Structural diagram of SummScore’s submodels.

[SEP] and inputting them into the Cross-Encoder model for training. Similarly, a summary can be regarded as a semantically similar text obtained after the original text is compressed. A heuristic idea is that the more similar the summary is to the original text, the higher the quality of the summary. The similarity here includes semantic similarity, content consistency, etc. Inspired by QA retrieval, we also regard the scoring of summary quality as a process of text similarity calculation between the original text and summary text.

As shown in Figure. 2, we formally define the summary quality evaluation problem as follows. Given the subword sequence  $O$  of the source document, where  $O = \{o_1, \dots, o_n\}$ . Suppose that the subword sequence of the generated summary is  $S$ , where  $S = \{s_1, \dots, s_m\}$ . The goal is to implement a function  $score(O, S)$  and predict a score  $y$  to represent the similarity between document  $O$  and summary  $S$ . Given the training data with human annotation scores on summary quality, our goal is to train the function  $score(O, S)$  so that it can regress to the human annotation score  $y'$ .

### 3.2 Structure of Model

The structure of SummScore’s submodels is based on the Cross-Encoder. The Cross-Encoder [15] believes that the spliced sentence pair is a reasonable input mode, which is suitable for NSP(Next Sentence Prediction) [4] pre-training task and natural language inference task. Our SummScore is designed based on the principle of semantic similarity computation, and the used Cross-Encoder is pre-trained on related tasks. Hugging Face SentenceTransformers provides researchers with Cross-Encoder<sup>3</sup> after training on the semantic similarity benchmark dataset STS [2]. By inputting sentence pairs, the Cross-Encoder will predict a score between 0 and 5 representing the semantic similarity of the two texts.

Subsequently, we use the pre-trained Cross-Encoder to perform fine-tuning on fine-grained quality human-annotated data. Specifically, we add a regression model based on MLP(Multilayer Perceptron) to Cross-Encoder to evaluate the score. The format of the input is a patchwork of sentence pairs. The first token of each sentence pair is always a special mark [CLS], and the sentences are

<sup>3</sup> <https://www.sbert.net/examples/training/cross-encoder/README.html>

separated by [SEP]. Finally, the final hidden state corresponding to the first special [CLS] token is taken as the sentence feature of the overall input. Feed the [CLS] embedding  $V_{[CLS]}$  into the MLP to get the predicted score  $y$ :

$$V_{[CLS]} = \text{Cross-Encoder}([CLS], O, [SEP], S) \quad (1)$$

$$y = WV_{[CLS]} + b \quad (2)$$

where  $W$  and  $b$  are learnable parameters. The learning goal of the whole model is to fit the gold label  $y'$  with  $y$ . Our squared regression loss is:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \|y - y'\|^2 \quad (3)$$

where  $N$  is the size of the sample.

### 3.3 Training Method of Submodels

Proxy metrics such as ROUGE and BERTscore usually return only a single value for the summary quality. It is difficult for people to clearly know how good or bad the current summary is from this score value. For example, does this summary capture the topic of the original text? How fluent is this summary? What are the main problems in this summary? That is, proxy metrics such as ROUGE and BERTscore are not well interpretable. Due to the poor interpretability of metrics scoring, it is also difficult for the summarization model to further improve the quality of the generated summary and the performance of the model.

The counterpart to machine scoring is human evaluation. It is a common fact that the human evaluation will first divide the quality of summaries into multiple fine-grained quality dimensions, and then score on the specific dimensions. A popular division is to divide the quality of the summary into four fine-grained quality dimensions (**Coherence**, **Consistency**, **Fluency**, and **Relevance**) [7,10]. Specifically, **Coherence**: the summary should be a coherent set of information about a topic, and whether the organizational structure between sentences is logical. **Consistency**: the summary should contain only the facts and themes of the original text. Both should be presented consistently and without hallucinatory facts. **Fluency**: the quality of the language. Whether there are grammatical errors that affect reading. **Relevance**: the summary should only contain important information from the source document, penalizing the summary that contains redundant information. Following this principle, our SummScore is also composed of four scoring submodels, and each corresponds to one of the above quality dimensions. Therefore, SummScore has the human-like ability to comprehensively evaluate the quality of summaries across multiple quality dimensions.

The model structures of the four submodels are consistent, but the mode of data input of the submodels of Coherence and Fluency is different. As shown in Figure. 2, among them, the scoring submodels for Fluency and Coherence no longer use the training mode of sentence pair. Because Fluency evaluates

**Algorithm 1** The semi-supervised multi-round training

**Input:** Initial submodel  $M_0$ , annotated dataset  $D_L(D_L^{train} \cup D_L^{val})$ , unannotated data subset  $D_U = \{D_1, \dots, D_k\}$ , epoch size for fine-tuning  $ep$

**Output:** The best submodel  $M_{best}$

```

1: /*Part1: the first round of supervised training with  $D_L^{train}$ .*/
2: Let  $M_0^{best} = M_0$ 
3: for each  $i \in \{0, 1, \dots, ep - 1\}$  do
4:   Train  $M_0$  on  $D_L^{train}$  an epoch agin and obtain  $M_{i+1}$ 
5:   if  $f(M_{i+1}, D_L^{val}) > f(M_{best}, D_L^{val})$  then
6:      $M_0^{best} = M_{i+1}$ 
7:   end if
8: end for

9: /*Part2: multiple rounds of semi-supervised training.*/
10: Let  $D = D_L^{train}$ ,  $M^{best} = M_0^{best}$ 
11: for each  $t \in \{1, 2, \dots, k\}$  do
12:   Annotate  $D_t$  with  $M_{t-1}^{best}$  and obtain pseudo-annotated data  $D_t^{pseudo}$ 
13:    $D = D \cup D_t^{pseudo}$ 
14:   Let  $M_t^{best} = M_0$ 
15:   for each  $i \in \{0, 1, \dots, ep - 1\}$  do
16:     Train  $M_i$  on  $D_L^{train}$  an epoch agin and obtain  $M_{i+1}$ 
17:     if  $f(M_{i+1}, D_L^{val}) > f(M_t^{best}, D_L^{val})$  then
18:        $M_t^{best} = M_{i+1}$ 
19:     end if
20:   end for
21:   if  $f(M_t^{best}, D_L^{val}) > f(M^{best}, D_L^{val})$  then
22:      $M^{best} = M_t^{best}$ 
23:   end if
24: end for

```

the linguistic quality of the summary itself. When experts annotate Fluency’s scores, they can do it without referring to the original text. For the Coherence dimension, experts only focus on whether the summary text itself has a clear theme and rigorous sentence logic. In contrast, when experts score the quality dimensions of Consistency and Relevance, it is necessary to repeatedly compare the generated summary with the original text. Therefore, for the submodels of Coherence and Fluency dimension, we remove the original text information and change the formula (1) to the following form:

$$V_{[CLS]} = Cross - Encoder([CLS], S, [SEP]) \quad (4)$$

Because the annotation data resources are very limited, we adopt a semi-supervised multi-round training method to maximize the correlation between SummScore and human ratings. The input of the algorithm includes the pre-trained Cross-Encoder  $M_0$ , which is used as the initial state of the SummScore’s submodel. We have a small-scale manually annotated supervised dataset  $D_L$ . We divide  $D_L$  into the training set  $D_L^{train}$  and validation set  $D_L^{val}$ . In addition, we have a large amount of unsupervised data  $D_U$  generated by several main-stream summarization models.  $D_U$  is randomly divided into sub-datasets of the

same size  $\{D_1, \dots, D_k\}$ . Moreover, we also have a scoring function  $f(\cdot)$  to judge whether the submodel is good or bad, which is achieved by comparing the correlation between the scores predicted by the submodel and the manually annotated scores on  $D_L^{val}$ .  $f(\cdot)$  can be chosen from  $\max(Pearson)$ ,  $\max(Spearma)$  and  $\max(Pearson * Spearma)$ . Our goal is to obtain the globally optimal submodel  $M_{best}$  with limited annotation data.

Our training is mainly divided into two parts, as shown in lines 1-8 and 9-24 of the Algorithm. 1 respectively. In the first part of the algorithm, we first train the submodel on the small-scale supervised data  $D_L^{train}$ . In the beginning, we assume that the best submodel  $M_0^{best}$  in the first round of training is  $M_0$  (line 2). Then, we perform fine-tuning training for  $ep$  times (line 3). After the  $i$ -th fine-tuning, the submodel is trained from  $M_i$  to  $M_{i+1}$  (line 4). After each fine-tuning, we compare the quality of  $M_0^{best}$  and  $M_{i+1}$ , and save the best model as  $M_0^{best}$  (line 5-7). After the first round of supervised training, we get the best model of the first round  $M_0^{best}$ .

In the second part of the algorithm, we will carry out multiple rounds of semi-supervised training to improve the performance of the submodel using unlabeled data. We assume that the initial global optimal model is  $M_0^{best}$ , and the current training available dataset  $D$  is  $D_L^{train}$  (line 10). Because the unsupervised dataset  $D_U$  is divided into  $k$  blocks, the algorithm will perform  $k$  rounds of semi-supervised training (Line 11). At the beginning of the  $t$ -th round of semi-supervised training, we will use the optimal model of the previous round  $M_{t-1}^{best}$  to label the sub-data  $D_t$ , and get the pseudo-labeled dataset  $D_t^{pseudo}$  (line 12). Then, the newly obtained pseudo-label data  $D_t^{pseudo}$  is extended to the available dataset  $D$  for the next round of semi-supervised training (line 13). After that, like the steps of Part1, start with the initial Cross-Encoder  $M_0$  and fine-tune the submodel  $ep$  epochs on data  $D$  (line 14-20). After the end of each epoch fine-tuning, the optimal submodel  $M_t^{best}$  of round  $t$  is retained (line 18). After each  $t$ -th round of semi-supervised training, we will also compare  $M^{best}$  and  $M_t^{best}$ , and keep the best model as the global optimal model  $M^{best}$  (line 22). After all  $t$  rounds of semi-supervised training, we finally obtain the globally optimal submodel  $M^{best}$  of SummScore for each fine-grained quality dimension.

## 4 Experiments Settings

We conduct experiments on SummEval [7] dataset, which contains 1600 manually annotated summaries. Each summary is evaluated on the four fine-grained quality dimensions according to criteria [10] and is scored by 5 independent crowdsourcing workers and 3 independent experts. Annotation scores span from 1 (worst) to 5 (best). We calculate the average of the annotation scores of the 3 experts as the final supervision score for each data and randomly divide the data into a training set  $D_L^{train}$  of 1000 pieces of data and a test set  $D_L^{test}$  of 600 pieces of data. At the beginning of each round of semi-supervised training, we randomly sample 100 pieces of data from the training set  $D_L^{train}$  as the validation set  $D_L^{val}$  for model selection and pass it to  $f(\cdot)$  for model selection.

In addition to the above small-scale annotated data, we also use a large amount of unannotated data consisting of original texts and machine-generated summaries. These unannotated data will be randomly divided into  $k$  equal-sized parts in the experiment. Specifically, these divided data are mainly used in the semi-supervised training of the model to further help SummScore improve performance.

Our SummScore is based on Cross-Encoder<sup>4</sup> of Hugging Face SentenceTransformers. We expect that the scoring process of SummScore will be as fast as possible without taking up too much video memory of the machine. Therefore, we abandon the pre-training model with large-scale parameters, such as *RoBERTa<sub>LARGE</sub>* (24 layers), and only select the public *DistilRoBERTa<sub>BASE</sub>* (6 layers), and *RoBERTa<sub>BASE</sub>* (12 layers) for fine-tuning. So the GeForce GTX 1060 can meet all the experimental needs of SummScore. We set the epoch size for fine-tuning to be 6 and the batch size to be 4. When the amount of newly expanded pseudo-annotated data reaches 10,000 (about ten times the annotated data), the model can obtain satisfactory performance. At the beginning of each new round of semi-supervised training, SummScore will perform linner warmup training with 1/10 of the single round steps. We use Adam as our optimizer with a learning rate of 2e-5 and a weight decay of 0.01. Consistent with previous research works [17,20], we use Pearson and Spearman correlation coefficients to judge the correlation between the scoring metrics and manual scoring.

## 5 Experiments

### 5.1 Comparative Experiments

For the convenience of comparison, we conduct our comparative experiments in groups. First, we compare our model with several well-known training-free metrics. These metrics include BLEU [14], TF-IDF, ROUGE, BERTscore, and SUPERT. These metrics have their innovative principles and advantages, which have a profound impact on the development of the corresponding field. In particular, ROUGE and BERTscore are very popular and well-received in summary quality evaluation. Our SummScore is based on pre-training fine-tuning. Therefore, we also select two representative metrics based on the pre-trained model fine-tuning: BLEURT and BERT for MTE. For a fair comparison, we maintain the experimental design consistent with SummScore and conduct fine-tuning on the same data. Similarly, BLEURT and BERT for MTE also adopt multi-round semi-supervised training to eliminate the influence of training methods.

Table.1 shows our experimental results. Scores represent the Pearson correlation and Spearman correlation of each metric with respect to human annotations. It can be seen that compared with the training-free metrics, SummScore far exceeds them. Moreover, except for SUPERT, these metrics have a low correlation with human annotations in all fine-grained dimensions. However, we find that they (e.g. BLEU, ROUGE, and SUPERT) tend to be more relevant to human

<sup>4</sup> <https://www.sbert.net/examples/training/cross-encoder/README.html>

**Table 1.** The results of the correlation experiment of the evaluation metrics on the test set of SummEval

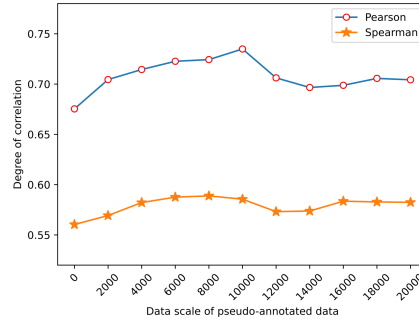
Quality Dimension		Coherence		Consistency		Fluency		Relevance	
Metric		Pearson Spearma		Pearson Spearma		Pearson Spearma		Pearson Spearma	
Training-free	BLEU-1	0.0278	0.0272	0.2023	0.1552	0.1367	0.0696	0.2459	0.1992
	BLEU-2	0.0419	0.0384	0.1531	0.1456	0.1206	0.0810	0.2104	0.2002
	BLEU-3	0.0588	0.0668	0.1129	0.1367	0.1092	0.0841	0.1901	0.2240
	BLEU-4	0.0513	0.0764	0.0896	0.1271	0.1053	0.0898	0.1567	0.2193
	TF-IDF	0.0667	0.0689	0.0772	0.0930	0.0797	0.0727	0.0893	0.0472
	ROUGE-1	0.1757	0.1593	0.2242	0.1767	0.1651	0.1055	0.3264	0.2955
	ROUGE-2	0.1229	0.1237	0.1515	0.1489	0.1297	0.1062	0.2502	0.2509
	ROUGE-3	0.1219	0.1339	0.1184	0.1319	0.1182	0.1106	0.2160	0.2472
	ROUGE-L	0.1569	0.1457	0.1671	0.1486	0.1578	0.1499	0.2415	0.2284
	BERTscore-r	0.1414	0.1297	0.2456	0.1921	0.2142	0.1553	0.3474	0.2960
	BERTscore-p	0.1746	0.1428	0.1059	0.0821	0.2414	0.1683	0.1930	0.1513
	BERTscore-f	0.1792	0.1534	0.2043	0.1750	0.2614	0.1811	0.3135	0.2694
	SUPERT	0.2853	0.2599	0.3230	0.2931	0.3062	0.2280	0.3703	0.3256
Ours Trained	BLEURT	0.4631	0.4410	0.3206	0.2233	0.4639	0.2193	0.5621	0.5286
	BERT for MTE	0.5532	0.5324	0.3721	0.3058	0.4601	0.2645	0.5638	0.5315
	BERT for MTE <sub>DistilRobertaBase</sub>	0.6080	0.6036	0.4630	0.3512	0.4787	0.3509	0.5813	0.5500
	SummScore <sub>DistilRobertaBase</sub>	<b>0.6704</b>	<b>0.6684</b>	<b>0.4839</b>	<b>0.4080</b>	<b>0.7071</b>	<b>0.5586</b>	<b>0.6018</b>	<b>0.5538</b>
Ours Trained		<b>0.7061</b>	<b>0.7116</b>	<b>0.4852</b>	<b>0.4497</b>	<b>0.7348</b>	<b>0.5855</b>	<b>0.6746</b>	<b>0.6391</b>

judgments than Coherence and Fluency on Relevance and Consistency quality dimensions. The reason is that these metrics all need to compare the literal n-gram or semantic information of the reference summary when scoring. Because the reference summary is a compressed text that captures the central idea of the original text. Therefore, these metrics can achieve the purpose of preliminarily measuring the Relevance and Consistency of the original text topic semantics of the generated summary. Unfortunately, they are designed without considering the quality requirements of Coherence and Fluency. So these metrics tend to work poorly in the Coherence and Fluency dimensions.

Compared with the trained metric group, our model also outperforms all of them. However, we find that these metrics also perform well after multiple rounds of semi-supervised training on data. To eliminate the influence of the pre-trained language model, we also replace the pre-trained model of BERT for MTE with the same *DistilRobertaBase* trained on the STS dataset. We find that the performance of BERT for MTE model is more competitive. This proves that the idea of SummScore’s quality evaluation design, which is inspired by the similarity matching principle of information retrieval, is reasonable.

## 5.2 Ablation Results

We believe that multi-round semi-supervised training is an important factor for improving SummScore. Because this training method brings about the rapid expansion of pseudo-annotated data and alleviates the problem of the small amount of data. In order to explore the influence of the amount of pseudo-annotated data, we conduct corresponding ablation experiments. Only the ablation experiments of *SummScoreRobertaBase* on the Fluency quality dimension are introduced here, and other quality dimensions have the same conclusion. In the ablation experiments, we expand the pseudo-annotation data with a span of 2000 pieces per round. The results of the ablation experiment are shown in Figure. 3. It can be



**Fig. 3.** Ablation experiments on the impact of pseudo-annotated data volume on the Fluency dimension of *SummScore<sub>RobertaBase</sub>*.

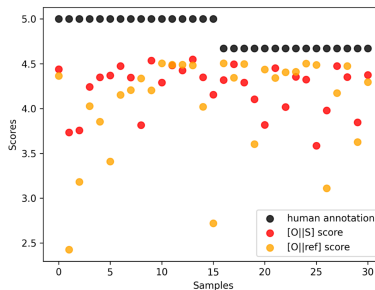
clearly found that the performance of the model is significantly improved with the increase of pseudo-annotation data in the early stage. This indicates that the scale of data volume is an important factor to limit the model performance during this period. When the expanded pseudo-annotation data reaches about 10,000 (10 times the annotation data), the correlation of Fluency reaches its peak. This shows that at this time, the model has maximized the benefits from the increase in data volume. Subsequently, even with more data, the performance of the model does not improve and even begins to degrade. The ablation experiments show that reasonable multi-round semi-supervised training can effectively improve the performance of the SummScore in the case of scarce annotated data. This also provides a new training idea for later researchers to alleviate the limitation of small data volume in similar experimental scenes.

To explore the difference between original-summary pairing  $[O||S]$  and common summary-reference  $[S||ref]$  (adopted by BLEURT, BERT for MTE, and other models), we also conduct relevant analysis experiments. Table. 2 shows the correlation of the two input methods on the *SummScore<sub>RobertaBase</sub>* model with human evaluation in the dimensions of Consistency and Relevance, respectively. We can find that  $[O||S]$  can achieve better results than  $[S||ref]$ . Originally, we are also worried that the longer original information in  $[O||S]$  may be more difficult for SummScore to learn than the short reference summary in  $[S||ref]$ . Analyzing the reasons for the better results of  $[O||S]$ , we believe that the reason is that the form of  $[O||S]$  may be more consistent with the scoring process of humans in the dimensions of Consistency and Relevance. Because, normally, humans start to write a summary after reading the original text. In real life, few golden summaries can be repeatedly referred to write the new summaries. In the scoring process, experts often score only after reading the original text. The input mode  $[O||S]$  is also consistent with the expert scoring process.

Further experiments, we find that the original-summary mode  $[O||S]$  can support the diversity of textual representations of summaries. The lower the ROUGE score, the more different the expression of the summary and the reference. However, those semantically correct summaries, which are expressed differently from the reference summaries, are also qualified summaries. Qualified summary met-

**Table 2.** Ablation experiments on the influence of input modes  $[S||ref]$  and  $[O||S]$  on Consistency and Relevance.

	Pearson Spearma	
<b>Consistency</b> $_{[S  ref]}$	0.4291	0.3290
<b>Consistency</b> $_{[O  S]}$	<b>0.4852</b>	<b>0.4497</b>
<b>Relevance</b> $_{[S  ref]}$	0.6519	0.6172
<b>Relevance</b> $_{[O  S]}$	<b>0.6746</b>	<b>0.6391</b>

**Fig. 4.** Ablation experiments exploring the effect of input modes  $[S||ref]$  and  $[O||S]$  on the diversity of summaries.

rics should be able to identify summaries that are diverse in expression but of acceptable quality. From the SummEval annotation data, we extract summary data with a low ROUGE score but a high human score. We plot the scatter plots of human scores and SummScore scores for these two input modes, respectively. Only the experiments in the Relevance dimension are listed here, and the results are shown in Figure. 4. We can find that  $[O||S]$  is closer to the distribution of human scoring. However, for the summaries with high human scores,  $[S||ref]$  is more likely to give low scores. Therefore, this can be illustrated that  $[O||S]$  can recognize summaries with different literal expressions but qualified quality. This also shows that the  $[O||S]$  mode can help to improve the diversity of summary generation.

### 5.3 Case Analysis

In Figure. 5, we show a typical example for case analysis. By reading the original text and the reference summary, we know that the original text is about *River Plate are keen to sign Manchester United striker Radamel Falcao* (orange fonts), and then some information about Radamel Falcao is introduced. However, we can find that the summary under test completely fails to capture the central idea of the original text. Therefore, in the Relevance dimension, both SummScore and experts give a low score of less than 2 points, but the baseline BERT for MTE scores a qualified score of 3 points. Further analysis, we find that there are *hallucination errors* (blue shaded text) in the summary under test. Radamel Falcao has good goalscoring form in *Colombia* rather than *Manchester United*. So both SummScore and experts give low marks for Consistency. Analyzing the

<b>Original Text:</b> River Plate are keen to sign Manchester United striker Radamel Falcao but admit a deal is complicated. The Colombia forward spent eight years with the Argentine side before leaving for Porto in 2009 and River Plate are open to Falcao returning. During an interview with Esto es River program, vice president Matias Patanian said: 'We dream of Falcao Garcia. The doors are open.' River Plate are keen to sign former forward Radamel Falcao who has struggled on loan at Manchester United. River Plate vice president Matias Patanian admits the club 'dream of Falcao' and that 'the doors are open'. The 29-year-old has struggled during a season-long loan spell at Old Trafford this term - scoring just four Premier League goals - and it remains to be seen whether United will exercise the option to keep the frontman or whether he will return to parent club Monaco. However, Falcao has been in good goalscoring form for his country this week, finding the net three times in two games to equal Colombia's all-time goalscoring record with 24 goals. Joining River Plate at the age of 15 in 2001 before making his first-team debut four years later, Falcao went on to score 34 goals in 90 appearances for the Primera Division club. Falcao scored 34 goals in 90 appearances for the Argentine club during his four seasons in the first team.				
<b>Reference:</b> river plate admit they 'dream' of manchester united striker radamel falcao. the colombia international spent eight years with the argentine club. falcao has managed just four goals in 19 premier league appearances.				
<b>Summary:</b> radamel falcao has been in good goalscoring form for manchester united. the 29-year-old has struggled in a season-long loan at manchester united. the frontman has been scoring just four premier league goals.				
	<b>Coherence</b>	<b>Consistency</b>	<b>Fluency</b>	<b>Relevance</b>
<b>Human:</b>	1.33	2.33	5.00	1.33
<b>SummScore:</b>	1.773	2.531	4.979	1.894
<b>BERT for MTE:</b>	2.031	4.896	4.928	3.052
<b>ROUGE:</b>	ROUGE_L:0.436	ROUGE_2:0.169	ROUGE_L:0.399	

**Fig. 5.** A classic example of performance comparison between SummScore and other metrics.

structure between sentences, we find that the semantics of the summary to be tested is lack logic. For one moment, the summary tells us Radamel Falcao has good goalscoring form and another point that he struggles at Manchester United. Due to the lack of logic between sentences, it is difficult to read. So both SummScore and experts score low on the Coherence dimension, but BERT for MTE scores a high score close to 5. In terms of fine-grained quality dimension, it can be said that SummScore has better scoring ability than BERT for MTE, and the scoring effect is closer to human scoring. Because of the good n-gram overlap between the summary and the reference, ROUGE-1 gave this incomplete summary a high score of 0.436. As you can see, ROUGE is indeed a rough proxy indicator without explanatory power. ROUGE cannot tell us the specific quality of the summary, such as whether factual errors and grammatical errors exist.

**Table 3.** Evaluation results of mainstream models on the SummScore indicator on the CNN/DailyMail.

Metrics	ROUGE-1	ROUGE-2	ROUGE-L	Coherence	Consistency	Fluency	Relevance
<b>Extractive Models</b>							
LEAD-3	0.3994	0.1746	0.3606	<b>3.9146</b>	<b>4.9766</b>	<b>4.9430</b>	<b>4.4264</b>
NEUSUM	<b>0.4130</b>	<b>0.1893</b>	0.3768	3.1327	<b>4.9712</b>	<b>4.9031</b>	4.1876
BanditSum	<b>0.4137</b>	0.1868	0.3759	3.2399	<b>4.9738</b>	<b>4.9139</b>	4.2007
RNES	0.4088	<b>0.1878</b>	0.3719	<b>3.7673</b>	<b>4.9771</b>	<b>4.9041</b>	<b>4.4521</b>
<b>Abstractive Models</b>							
Pointer Generator	0.3921	0.1723	0.3599	3.3892	4.9654	<b>4.9401</b>	4.1721
Fast-abs-rl	0.4057	0.1774	<b>0.3806</b>	2.2031	4.9255	4.6389	3.9024
Bottom-Up	0.4124	0.1870	<b>0.3815</b>	2.8551	4.9113	4.7716	3.8890
Improve-abs	0.3985	0.1720	0.3730	2.1961	4.8243	4.5633	3.6758
Unified-ext-abs	0.4038	0.1790	0.3675	3.4100	<b>4.9736</b>	4.8955	4.2684
ROUGESa	0.4016	0.1797	0.3679	3.2674	4.9700	4.8688	4.1793
Multi-task (Ent + QG)	0.3952	0.1758	0.3625	3.3573	4.9633	4.8870	4.1208
Closed book decoder	0.3976	0.1760	0.3636	3.3825	4.9688	4.8908	4.1866
T5	<b>0.4479</b>	<b>0.2205</b>	<b>0.4172</b>	3.6991	4.9126	4.8703	<b>4.3365</b>
GPT-2 (supervised)	0.3981	0.1758	0.3674	<b>3.7410</b>	3.9252	3.8176	3.6069
BART	<b>0.4416</b>	<b>0.2128</b>	<b>0.4100</b>	<b>4.2064</b>	4.9707	4.8545	<b>4.5683</b>
Pegasus	<b>0.4408</b>	<b>0.2147</b>	<b>0.4103</b>	<b>3.7148</b>	4.9176	4.8522	<b>4.3421</b>

#### 5.4 Mainstream Summarization Models Evaluation

Finally, we use SummScore to evaluate the performance of 16 mainstream summarization models on the CNN/DailyMail dataset, and the scoring results are shown in Table. 3. Please refer to the work SummEval [7] for a detailed introduction to these mainstream summarization models. We bold the top 5 scores of each quality dimension for further experimental analysis. We find that ROUGE favors the abstractive models, but SummScore seems to prefer the extractive models. In particular, the LEAD-3 model has achieved high SummScore scores on all four fine-grained qualities. The reason is that the first three sentences of the news are the most important part of the full text and the LEAD-3 is very suitable for news data. For Fluency and Consistency, extractive models tend to achieve higher scores. This is reasonable, because the summary of the abstractive model is generated according to the probability distribution of words, and the problems of fragment repetition and syntax errors can not be avoided. The generated summary may also have illusory facts that are inconsistent with the facts of the original text. However, the extractive model produces summaries by splicing sentences extracted from the original text. Because the sentences are written manually, this avoids grammatical errors and repetition. Moreover, the sentences are derived from the original, so there is no illusory fact. For Coherence and Relevance, there is a strong correlation between the two scores. Moreover, the ROUGE score is also correlated with the score of SummScore in these two quality dimensions. Almost models with high ROUGE scores also have high scores of Coherence and Relevance and vice versa.

## 6 Conclusion

In this paper, we propose SummScore based on the semantic matching principle of information retrieval, which is a trained scoring metric for summary quality evaluation based on Cross-Encoder. SummScore has good interpretability. It consists of four submodels, which measures the quality of the summary comprehensively from four fine-grained qualities of Coherence, Consistency, Fluency, and Relevance. We use semi-supervised multi-round training to improve model performance on limited annotated data. Extensive experiments show that SummScore significantly outperforms existing metrics in terms of human relevancy and helps improve the diversity of generated summaries. Finally, we also use SummScore to evaluate 16 mainstream summarization models and publish the results for later research.

## References

1. Banerjee, S., Lavie, A.: 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 (2005)

2. Cer, D., Diab, M., Agirre, E., Specia, L.: Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation (2017)
3. Chen, W., Li, P., King, I.: A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy (2021)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.N.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2018)
5. Ding, Y., Liu, J., Liu, K., Ren, R., Wang, H.: Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering (2020)
6. Durmus, E., He, H., Diab, M.: Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization (2020)
7. Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R., Radev, D.: Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics **9**, 391–409 (2021)
8. Gao, Y., Zhao, W., Eger, S.: Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. arXiv preprint arXiv:2005.03724 (2020)
9. He, T., Chen, J., Ma, L., Gui, Z., Li, F., Shao, W., Wang, Q.: Rouge-c: A fully automated evaluation method for multi-document summarization. In: 2008 IEEE International Conference on Granular Computing. pp. 269–274. IEEE (2008)
10. Kryściński, W., Keskar, N.S., Mccann, B., Xiong, C., Socher, R.: Neural text summarization: A critical evaluation (2019)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
12. Lloret, E., Plaza, L., Aker, A.: The challenging task of summary evaluation: an overview. Language Resources and Evaluation **52**(1), 101–148 (2018)
13. Ma, Q., Graham, Y., Wang, S., Liu, Q.: Blend: a novel combined mt metric based on direct assessment—casict-dcu submission to wmt17 metrics task. In: Proceedings of the second conference on machine translation. pp. 598–603 (2017)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
15. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)
16. Sellam, T., Das, D., Parikh, A.P.: Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696 (2020)
17. Shimanaka, H., Kajiwar, T., Komachi, M.: Machine translation evaluation with bert regressor. arXiv preprint arXiv:1907.12679 (2019)
18. Stanojević, M., Sima'an, K.: Fitting sentence level translation evaluation with many dense features. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 202–206 (2014)
19. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Christiano, P.: Learning to summarize from human feedback (2020)
20. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
21. Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M., Eger, S.: Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv preprint arXiv:1909.02622 (2019)