A Topic Information Fusion and Semantic Relevance for Text Summarization

FUCHENG YOU, SHUAI ZHAO, AND JINGJING CHEN
Beijing Institute of Graphic Communication, Beijing 100000, China
Corresponding author: Fucheng You (youfucheng@bigc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773229 and in part by the Beijing Municipal Natural Science Foundation under Grant KZ201710015010.

ABSTRACT
With the continuous development of deep learning, pre-trained models have achieved sound effects in the field of natural language processing. However, text summarization research is far from what people want, especially in abstractive summarization. A high-quality summarization system needs to focus on the topic content of the document and the similarity between the summary and the source document. In this paper, we propose a topic information fusion and semantic relevance for text summarization based on Fine-tuning BERT(TIF-SR). Primarily, considering the critical role of topic information in summary generation, we extract topic keywords and fusion them with source documents as part of the input. Secondly, make the summary closer to the source document by calculating the semantic similarity between the generated summary and the source document, the quality of the abstract is improved. The experimental data indicate that the ROUGE index and readability have improved in this model, so these shreds of evidence suggest that the method proposed by our model is sufficient.

INDEX TERMS
Summarization generation, fine-tune BERT, topic information, semantic relevance, transformer, attention mechanism.

I. INTRODUCTION
Text summarization is to make the text easier to read and understand, especially for poor readers. There are two kinds of summary generation: extractive and abstractive. Extractive summarization models extract some important words or sentences from the original documents. Compared with abstractive methods, it is comparatively mature and straightforward, but its drawback is limited to the original text, and abstract coherence is weak. Abstractive summarization systems are not limited to the original text. Extractive summarization is easier, but abstractive summarization is more like the way humans process text. In this paper, we focus on abstractive summarization of single-document.

In some previous work, deep neural network approaches achieve favorable effects in various Natural Language Processing(NLP) tasks like image captioning [1], machine translation [2]–[4], speech recognition [5], and so on. In particular, the attention-based sequence-to-sequence(seq2seq) model with recurrent neural networks(RNNs) [6]–[8] is widely used in abstractive summarization. The model compresses the original document into dense vectors with the encoder, and the decoder generates summary using the compressed vectors. While promising, most of the current research on the summarization generation focuses on short text, because these methods fail to provide satisfactory performance when faced with long input sequence.

In this work, we focus on the key factors that affect the quality of the summary generation. The sequence-to-sequence framework consists of two parts: the encoder layer that processes the input and the decoder layer that generates the output. There have been lots of works on the design and alignment of the encoder and the decoder, like the choice of the neural network structure, the specific design of the attention mechanism [9]–[11], etc. In contrast, the topic information [12], [13], which directly affects the results of the summarization task, has received relatively much less attention.

At present, almost all attention mechanism models distribute their attention evenly on all the contents of the text, which has many advantages. Still, they ignore the influence of text topics on the generation of abstracts, as shown
in Table 1. Based on this problem, we will extract the subject keywords of the source text and fuse the subject information as input. Another common weakness of previous works is the sequence-to-sequence model tends to produce grammatical and coherent summary regardless of the semantic relevance to source texts. Although the generated summarization is similar to source texts literally, they have low semantic relevance. In this work, our goal also is to improve the semantic relevance [14] between source texts and generated summarization. In our model, we compress the source texts into vectors with the encoder and decode the vector into summarization with the decoder. The encoder layer produces the representation of original documents, and the decoder layer produces the representation of the generated summarization. A similarity evaluation component is introduced to measure the relevance of original texts and generated summarization. During training, it maximizes the similarity score to encourage high semantic relevance between source texts and generated summarization.

The contributions of this work are as follow:
1. We argue the BERT [15] model, with its pre-training on a huge dataset and the powerful architecture for learning complex features, can further boost the performance of abstractive summarization. We fine-tune the BERT model to generate text summary.
2. The attention mechanism distribute their attention evenly on all the contents of the text, which ignore the influence of text topics on the generation of abstracts. We will extract the subject keywords from the source text and fuse the subject information as input.
3. The sequence-to-sequence models tend to regardless of the semantic relevance to source texts. They have low semantic relevance. In this work, we will improve the semantic relevance between source texts and generated summarization.

### II. RELATED WORK

At present, the methods of summary generation are mainly divided into extractive and abstractive approaches. On the one hand, early extractive summarization systems mostly focused on statistical and linguistic features, such as word frequency and term frequency [16]. Classical approaches include graph-based methods [17], integer linear programming [18], LDA [19], Conditional random field [20], and classifier-based methods [21], [22]. On the other hand, due to the technology, early research on abstractive systems is relatively less. Most of the jobs just focused on sentence compression, using methods like syntactic tree pruning [23] and machine translation [24].

In recent years, with the development of deep learning techniques and it’s widely investigated both on extractive and abstractive summarization tasks. In terms of neural networks extractive summarization systems, the sequence-to-sequence model gains impressive performance by applying neural networks. The improvement on automatic metrics like ROUGE [25] has reached satisfactory results, such as Dong [26], Narayan [27], and Zhang [28].

In the meantime, the sequence-to-sequence neural network models have provided a feasible new method for abstractive summarization systems. Rush [7] were the first to apply neural networks in abstractive text summarization and used Attention Mechanism to automatic summarization. Experimental data show that the attention mechanism has achieved good results; their model includes encoder and decoder, in which the encoder layer used Convolutional Neural Network (CNN), and the decoder layer used neural language model. Chopra [6] used Recurrent Neural Network (RNN) to replace the decoder, and Nallapati [8] improved the performance by applying RNN encoder-decoder model. Hou [29] used the subject information as part of the encoder input to improve the quality of the summary generated by the model. Ma [14] used the RNN model based on the attention mechanism to calculate the similarity between the generated summary and the source document and achieved excellent results. However, feature extractors with RNN or CNN as the core often have some problems that cannot be ignored. For instance, parallel computing power, unable to capture long-distance features. Therefore, Ming et al. proposed using the Transformer [30] as feature extractor to generate abstracts and achieved excellent results. The NEUSUM [31] adopts an end-to-end extractable text summary, which solves the problems of sentence scoring and sentence selection splitting in the previous model. Chen [32] combined the reinforcement learning method to train the extractive model and the abstractive model separately. Sentence-level strategy gradient method is used to connect the two networks. Recently, language model pre-training has been shown to be effective for improving many natural language processing tasks. Such as BERT [15], which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. Besides, there are some improved methods based on attention mechanism, optimization method, and text information embedding. Inspired by BERT’s pre-training and fine-tuning, Song [33] proposed the masked mass model to generate text in the encoder-decoder model. Zhang [34] used BERT to

### TABLE 1. An example of summarization.

| Source: 华谊兄弟在昨日收盘后发布公告称，公司拟以自有资产3978亿元收购浙江永乐影视股份有限公司若干股东持有的永乐影视51%的股权，对于该项收购，华谊兄弟董事长胡明昨日表示：“和永乐影视的合并是对华谊兄弟电视剧业务的一个加强。”  
Huayi Brothers issued an announcement after yesterday’s closing that it intends to acquire 51% of the shares of Yongle Film and Television held by several shareholders of Zhejiang Yongle Film and Television Co., Ltd. with its own funds of 397.8 million yuan. Regarding this acquisition, Hu Ming, the Secretary of Huayi Brothers said yesterday, “The merger with Yongle Film and Television is an enhancement of Huayi Brothers’ TV drama business.”  
Reference: 华谊兄弟收购永乐影视51%股权
Huayi Brothers intends to acquire 51% equity of Yongle Film and television.
Seq2Seq: 华谊兄弟合并永乐影视股权。
Huayi Brothers merges the equity of Yongle Film and television.  
In the meantime, the sequence-to-sequence neural network model tends to regardless of the semantic relevance to source texts. Although the generated summarization is similar to source texts, they have low semantic relevance. In this work, our goal also is to improve the semantic relevance [14] between source texts and generated summarization. In our model, we compress the source texts into vectors with the encoder and decode the vector into summarization with the decoder. The encoder layer produces the representation of original documents, and the decoder layer produces the representation of the generated summarization. A similarity evaluation component is introduced to measure the relevance of original texts and generated summarization. During training, it maximizes the similarity score to encourage high semantic relevance between source texts and generated summarization.

The contributions of this work are as follow:
1. We argue the BERT [15] model, with its pre-training on a huge dataset and the powerful architecture for learning complex features, can further boost the performance of abstractive summarization. We fine-tune the BERT model to generate text summary.
2. The attention mechanism distribute their attention evenly on all the contents of the text, which ignore the influence of text topics on the generation of abstracts. We will extract the subject keywords from the source text and fuse the subject information as input.
3. The sequence-to-sequence models tend to regardless of the semantic relevance to source texts. They have low semantic relevance. In this work, we will improve the semantic relevance between source texts and generated summarization.

### II. RELATED WORK

At present, the methods of summary generation are mainly divided into extractive and abstractive approaches. On the one hand, early extractive summarization systems mostly focused on statistical and linguistic features, such as word frequency and term frequency [16]. Classical approaches include graph-based methods [17], integer linear programming [18], LDA [19], Conditional random field [20], and classifier-based methods [21], [22]. On the other hand, due to the technology, early research on abstractive systems is relatively less. Most of the jobs just focused on sentence compression, using methods like syntactic tree pruning [23] and machine translation [24].

In recent years, with the development of deep learning techniques and it’s widely investigated both on extractive and abstractive summarization tasks. In terms of neural networks extractive summarization systems, the sequence-to-sequence model gains impressive performance by applying neural networks. The improvement on automatic metrics like ROUGE [25] has reached satisfactory results, such as Dong [26], Narayan [27], and Zhang [28].

In the meantime, the sequence-to-sequence neural network models have provided a feasible new method for abstractive summarization systems. Rush [7] were the first to apply neural networks in abstractive text summarization and used Attention Mechanism to automatic summarization. Experimental data show that the attention mechanism has achieved good results; their model includes encoder and decoder, in which the encoder layer used Convolutional Neural Network (CNN), and the decoder layer used neural language model. Chopra [6] used Recurrent Neural Network (RNN) to replace the decoder, and Nallapati [8] improved the performance by applying RNN encoder-decoder model. Hou [29] used the subject information as part of the encoder input to improve the quality of the summary generated by the model. Ma [14] used the RNN model based on the attention mechanism to calculate the similarity between the generated summary and the source document and achieved excellent results. However, feature extractors with RNN or CNN as the core often have some problems that cannot be ignored. For instance, parallel computing power, unable to capture long-distance features. Therefore, Ming et al. proposed using the Transformer [30] as feature extractor to generate abstracts and achieved excellent results. The NEUSUM [31] adopts an end-to-end extractable text summary, which solves the problems of sentence scoring and sentence selection splitting in the previous model. Chen [32] combined the reinforcement learning method to train the extractive model and the abstractive model separately. Sentence-level strategy gradient method is used to connect the two networks. Recently, language model pre-training has been shown to be effective for improving many natural language processing tasks. Such as BERT [15], which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. Besides, there are some improved methods based on attention mechanism, optimization method, and text information embedding. Inspired by BERT’s pre-training and fine-tuning, Song [33] proposed the masked mass model to generate text in the encoder-decoder model. Zhang [34] used BERT to
extract sentence features on the encoder side and Transformer and Bert on the decoder side respectively to make predictions. Wang [35] used BERT to encode each sentence, tried three training strategies of Mask, Replace and Switch, and proved its effectiveness. The HIBERT [36] uses an unsupervised method to pre-train on a large corpus to better adapt to downstream tasks.

Compared to previous work, we propose a topic information fusion and semantic relevance for text summarization based on Fine-tuning BERT(TIF-SR). Our method focuses on the topic information and the semantic relevance using the BERT model. Previous models limited their attention to source documents and ignored important subject information in the original text. The sequence-to-sequence models regardless of the semantic relevance to source texts. We will extract the topic keywords as part of the input and improve the semantic relevance between source texts and generated summaries. Unlike previous models, which only fuse subject information or calculate semantic similarity, our model consists of three components: the topic information is used as part of the encoder input, and the input is compressed into semantic vectors using BERT. The decoder uses the Transformer to produces semantic vectors of the generated summaries. Finally, the similarity function evaluates the relevance between the semantic vectors of source texts and generated summaries.

III. MODEL

A. MODEL INPUT

The topic of a document is a concise summary of the main content of a document, which is usually composed of multiple topic keywords. The topic keywords are the smallest unit to represent the content of the document, reflecting the main content of the document. In this model, we first preprocess the document, extract 10 subject keywords of the article using the TextRank algorithm, and use these keywords as part of the model input.

The core of the TextRank [37] algorithm is to divide the text into several constituent units (words, sentences) and build a graph model. It uses the voting mechanism to sort the scores of important components in the text and then takes the words with the highest scores as the topic keywords.

As illustrated in Figure 1, we insert a [CLS] token before the first sentence and a [SEP] token after each sentence. The [CLS] is used as a symbol to aggregate features from one sentence or a pair of sentences. In particular, we add topic token embedding as part of the input. To make the generated summary fully reflect the subject information, in the first place, our model extracts the subject keyword information from the source document and takes it as a part of the input. The input content of our model is similar to that of BERT model, but the subject information is added. The input information mainly includes input document, token embedding, topic token embedding, interval segment embedding, and position embedding.

B. MODEL ARCHITECTURE

Our model conforms to the typical encoder and decoder technical framework: after preprocessing, the original document is input from the encoder side, and the decoder side generates statements as the result of the summary. The main idea behind this architecture is to use the transfer learning from pretrained BERT, a masked language model. We have replaced the encoder part with BERT and the decoder is used transformer [26]. As illustrated in Figure 2.

First, we input the constructed word embedding to the encoder layer, which is the Bert layer (BERT-Base, Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters). The Bert layer makes use of the encoder part of the Transformer and stacks multiple Bidirectional
Transformers as the language model. In a random mask, 15% of words in each sentence are predicted by their context, and next sense prediction is used to understand sentence meaning and relationship better. The encoder layer compresses the input sequence into a semantic vector.

After obtaining the semantic vectors from BERT, we build the Transformer layer after the Bert layer to capture the document features and generate the corresponding summary. One of the advantages of using Transformer networks is that training is much faster than LSTM [38] based models as we eliminate sequential behavior in Transformer models. It is built by many multi-head self attentions blocks. Transformer has high parallel computing speed and long-distance feature acquisition ability by use self-attention, which is another useful feature extractor apart from RNN and CNN.

The traditional attention mechanism is to give a set of vector set values, and a vector query. The attention mechanism is a mechanism for calculating the weighted summation of values based on the query. Because of the seq2seq model, the encoder compresses the input information into a fixed-length semantic vector, but as the length of the input statement increases, the fixed-length semantic vector cannot fully express all the information of the text, which affects the decoder effect conversely the final output sequence. But self-attention encourages the model to learn long-term dependencies and does not create much computational complexity. The attention is applied to a set of queries simultaneously and relationship better. The encoder layer compresses the input sequence into a semantic vector.

In addition to the attention layer, Transformer adds fully connected feed-forward network to each layer of encoder and decoder, which is applied to each position separately and identically.

\[ FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \]  

Transformer uses sine and cosine functions of different frequencies to represent the positional encodings and adds the positional encodings to the input layer by adding:

\[ PE(pos, 2i) = \sin(pos/10000^{2i/d_{model}}) \]  
\[ PE(pos, 2i + 1) = \cos(pos/10000^{2i/d_{model}}) \]

Transformer-based models generate more correct and coherent sentences grammatically. The decoder focus on summarization tasks from the BERT outputs:

\[ \tilde{h}_l = LN(h_{l-1} + MHAtt(h_{l-1})) \]  
\[ h_l = LN(h_l + FNN(\tilde{h}_l)) \]

where \( h_0 = PosEmb(T) \) and \( T \) are the sentence vectors output by BERT; LN is the layer normalization operation; MHAtt is the multi-head attention operation; the superscript \( l \) indicates the depth of the stacked layer. The final output layer use a softmax function to get the predicted score:

\[ p(y_t|x_t) = \text{softmax}(W_o h_t + b_o) \]

**C. SEMANTIC RELEVANCE**

In this section, our purpose is to compute the semantic relevance of source documents and generated summary given the source semantic vector \( V_1 \) and the generated summarization semantic vector \( V_2 \).

As illustrated in Figure 3. We focus on the task of semantic similarity calculation, and there are many ways to do this task. For example, Inner Product, Dice Coefficient, Jaccard Coefficient. Because the source document and the generated summary are in the same language, it is assumed that
their semantic vectors are distributed in the same space. Cosine similarity [39] works better in calculating the distance between two vectors in the same space. The critical problem of using semantic relevance metrics is how to get the semantic vector $V_1$ and $V_2$. In our model, the semantic vector $V_1$ is represented by the output of BERT layer, which generates the semantic vector by compressing and encoding the source document. Similarly, the semantic vector generated by the encoder side is input into the transformer layer as the initial state, and Transformer layer generates the semantic vector $V_2$ by decompressing. We use cosine similarity to calculate semantic relevance. The closer the cosine value is to 1, the more similar the two vectors are:

$$
cos(V_1, V_2) = \frac{V_1 \cdot V_2}{||V_1|| ||V_2||} \tag{10}
$$

D. TRAINING
Set the input document to $x$ and given the model parameter $\theta$. The model generates summary $y$ and semantic vector $V_1$ and $V_2$. The objective is to minimize the loss function:

$$
L = -p(y|x; \theta) - \lambda cos(V_1, V_2) \tag{11}
$$

where $p(y|x; \theta)$ is the conditional probability of summaries given source documents and is computed by our model. $cos(V_1, V_2)$ is cosine similarity of semantic vectors $V_1$ and $V_2$. This term tries to maximize the semantic relevance between source texts and generated summarization.

IV. ANALYSIS OF EXPERIMENTAL RESULTS
In this section, we present the evaluation of our model and show its performance on Chinese corpus. Besides, we perform a case study to explain the semantic relevance between generated summary and source text.

A. EVALUATION STANDARD
ROUGE (recall-Oriented Understudy for Gisting Evaluation) [25] is a similarity measurement method based on the recall rate. Similar to BLEU, which is a commonly used evaluation standard in automatic summarization. At present, The ROUGE evaluation criteria include ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L. ROUGE-1 represents the amount of information contained in automatic abstract, ROUGE-2, ROUGE-3, ROUGE-4 represents the fluency of automatic abstract, and ROUGE-L is the degree of information covered by abstract to the original text. The ROUGE-N calculation method as shown in Formula (12), as shown at the bottom of this page.

Among them, Re is an artificial standard abstract, the gram is an N-element word, and count represents the maximum number of N-element words that the model generates the same abstract as the standard abstract. The basic idea is that the experts manually write a summary to form a standard summary set, and then compare the abstract produced by the model with the standard artificial summary. The quality of the summary generated by the model is further evaluated by comparing the number of basic units that overlap the two.

B. EXPERIMENTAL CORPUS
There are two corpora used in this study, LCSTS [40], and NLPCC2017 [29]. LCSTS is constructed by Hu, Chen. The dataset consists of more than 2.4 million text-summary pairs, constructed from a famous Chinese social media website called Sina Weibo. It is split into three parts, with 2,400,591 pairs in PART I, 10,666 pairs in PART II and 1,106 pairs in PART III. All the text-summary pairs in PART II and PART III are manually annotated with relevant scores ranged from 1 to 5, and we only reserve pairs with scores no less than 3. Following the previous work, we use PART I as the training set, PART II as the development set, and PART III as the test set. In addition, NLPCC2017 is a public data set provided by Headline, which mainly contains 50,000 news-summary pairs, and the length of each article ranges from 10-10000 Chinese characters.. In this experiment, 49,500 texts are selected as the training set and verification set, and the remaining 500 texts are taken as the test set.

C. MODEL COMPARISON
In order to evaluate the performance of the proposed model in summary generation, six typical models were selected for comparisons, Where RNN and RNN-context are the benchmark models in the data.

RNN [40]: With RNN as the encoder, the input of the decoder is the last hidden state of the encoder, and no context is used during the decoder.

RNN-context [40]: Based on the previous model, the context is used during the decoder, and the input to the decoder is a combination of all hidden states of the encoder.

Bi-MulRnn [41]: The model encoder uses a bi-directional recurrent neural network. The decoder uses a multi-layer recurrent neural network and the attention mechanism.

SRB [42]: Introducing the semantic correlation neural model, the encoder compresses the original text into the semantic vector, and the decoder generates the summary and the semantic vector of the summary. Then, the similarity function is used to evaluate the correlation between the semantic vector of the original text and the generating summary.

Seq2seq [43]: The encoder is a bidirectional LSTM structure containing information in both directions of the context; the decoder is a unidirectional LSTM structure that reads the

$$
ROUGE - N = \frac{\sum_{S \in \text{ReferencesSummaries}} \sum_{\text{gram}_n \in \text{SCountmatch}(\text{gram}_n)}}{\sum_{S \in \text{ReferencesSummaries}} \sum_{\text{gram}_n \in \text{SCount}(\text{gram}_n)}} \tag{12}
$$
input and generates a summary to map the target vocabulary to the high latitude space.

CopyNet [44]: Copy mechanism is introduced. The Encoder is bidirectional RNN; decoder is hybrid generation mode and copy mode. Survival mode selects words from preset vocabulary and copy mode selects words from the input sequence.

CGU [45]: A framework of global coding is introduced to perform global coding to improve the representation effect of source information.

MEAD [46]: This method considers the centroid, position, common sub-sequence and keywords of a sentence to score its features.

Submodular [47]: This model uses submodule functions to select important sentences by calculating their diminishing returns and generate summary.

NLP-ONE [48]: This method includes the input sequence attention mechanism, the output sequence attention mechanism, and the fusion of two attention mechanisms for the generation of the summary.

AS-TKF [29] This model used the topic information as part of the encoder input to improve the quality of the summary generate.

V. ANALYSIS
In this section, we show the results of our experiments and analyze the performance of our model. Also, we provide an example to show that the generated summary is closer to the topic and has higher semantic similarity with the source document.

A. RESULTS ON LCSTS DATASET
In the experiments on the LCSTS dataset, we compare our TIF-SR with the above baseline systems, including RNN, RNN-context, Bi-MuRnn, SRB, Seq2seq, CopyNet, and CGU. Our model achieves the advantages of ROUGE score over the baselines, and the advantages of ROUGE score on the LCSTS are significant. Table2 shows the results of our TIF-SR and baseline systems on the LCSTS. Compared with the CGU model, the advantage of our model is that ROUGE-1 is increased by 2.7, ROUGE-2 is increased by 1.7, and ROUGE-L is increased by 0.9. The experimental results are as Table 2 and Figure 4:

| model      | ROUGE-1 | ROUGE-2 | ROUGE-L |
|------------|---------|---------|---------|
| RNN        | 21.5    | 8.9     | 18.6    |
| RNN-context| 29.9    | 17.4    | 27.2    |
| Bi-MuRnn   | 27.6    | 17.1    | 25.9    |
| SRB        | 33.3    | 20.0    | 30.1    |
| Seq2Seq    | 33.8    | 23.1    | 32.5    |
| CopyNet    | 34.4    | 21.6    | 31.3    |
| CGU        | 39.4    | 26.9    | 36.5    |
| TIF-SR     | 42.1    | 28.6    | 37.4    |

B. RESULTS ON NLPCC2017 DATASET
In the experiments on the NLPCC2017 dataset, we compare our TIF-SR with the above baseline systems, including MEAD, SUbmodular, NLP-ONE and AS-TKF. By comparing the experimental results, the effect of the fusion topic information and semantic similarity calculation model proposed in this paper over the baselines, and the best effect is achieved. The results show that the fusion of topic information and the calculation of semantic similarity play a decisive role in the generation of high-quality summary and can better guide the generation of summary. The experimental results are as Table 3 and Figure 5:

C. DISCUSSION
We present the summary generated by our TIF-SR and compare it with the reference summary and the seq2seq model. As illustrated in Table 4. The original text introduced the express company’s violent sorting of courier services, but only one express company admitted to apologize, and the other express companies did not speak. The seq2seq model only states that the express company apologizes, but ignores
the text theme to emphasize the silence of other express companies. Our model can filter the trivial details that are irrelevant to the core meaning of the source text and focuses on the information that contributes most to the main idea. In comparison, the summary generated by our model have better fluency, sentence coherence, and more abundant information.

As our model incorporates topic keyword information, it can better grasp the subject information of the original document, thereby leading to the generation of summary closer to the topic. And we calculate the semantic similarity to improve the similarity between the generated summary and the original document, making the generated summary closer to the original document.

VI. CONCLUSION

Seq2Seq models are widely used in text summarization generation research. However, traditional models often have problems such as ignoring topic information and generating low similarity in summaries. In this paper, we propose a topic information fusion and semantic relevance for text summarization on Fine-tuning BERT. First of all, we take into consideration the impact of topic information on the quality of the summary, and the topic information is given higher weight by extracting topic keywords as part of the input. Secondly, in order to obtain the summary with higher semantic similarity, we calculate the semantic similarity between the generated summary and the original document, and maximize the similarity score through training. The experimental data show that our model has acquired good results on the LCSTS dataset, and the generated abstract is more readable and coherent.

REFERENCES

[1] H. Schwenk, “Continuous space translation models for phrase-based statistical machine translation,” in Proc. COLING, Mumbai, India, vol. 12, 2012, pp. 1071–1080.
[2] O. Vinyals and T. Show, “A neural image caption generator,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Nov. 2015, pp. 3156–3164.
[3] I. Sutskever, “Sequence to sequence learning with neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 12, 2014, pp. 3104–3112.
[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv:1409.0473. [Online]. Available: http://arxiv.org/abs/1409.0473.
[5] A. Graves, “Speech recognition with deep recurrent neural networks,” in Proc. 38th IEEE Int. Conf. Acoust., Speech, Signal Process., Vancouver, BC, Canada, Oct. 2013, pp. 6645–6649.
[6] S. Chopra, “Abstractive sentence summarization with attentive recurrent neural networks,” in Proc. Conf. North Amer. Assoc. Comput. Linguistics, 2016, pp. 93–98.
[7] A. A. Rush, “Neural attention model for abstractive sentence summarization,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 379–389.
[8] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, “Abstractive text summarization using Sequence-to-sequence RNNs and beyond,” in Proc. 20th SIGMOD Conf. Comput. Natural Lang. Learn., 2016, pp. 280–290.
[9] Z. Cao, “Joint learning of focusing and summarization with neural attention,” in Proc. COLING, 2016, pp. 547–556.
[10] B. Sankaran, H. Mi, Y. Al-Onaizan, and A. Ittycheriah, “Temporal attention model for neural machine translation,” 2016, arXiv:1608.02927. [Online]. Available: http://arxiv.org/abs/1608.02927
[11] W. Zeng, W. Luo, S. Fidler, and R. Urtasun, “Efficient summarization with read-again and copy mechanism,” 2016, arXiv:1611.03382. [Online]. Available: http://arxiv.org/abs/1611.03382.
[12] M. Ailem, B. Zhang, and F. Sha, “Topic augmented generator for abstractive summarization,” 2019, arXiv:1908.07026. [Online]. Available: http://arxiv.org/abs/1908.07026
[13] L. Wang, J. Yao, Y. Tao, L. Zhang, W. Liu, and Q. Du, “A reinforced topic-aware convolutional Sequence-to-Sequence model for abstractive text summarization,” 2018, arXiv:1805.03616. [Online]. Available: http://arxiv.org/abs/1805.03616
[14] S. Ma, X. Sun, J. Xu, H. Wang, W. Li, and Q. Su, “Improving semantic relevance for Sequence-to-Sequence learning of chinese social media text summarization,” 2017, arXiv:1706.02459. [Online]. Available: http://arxiv.org/abs/1706.02459
[15] J. Devlin, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL, 2019, pp. 4171–4186.
[16] A. Nenkova, L. Vanderwende, and K. McKeown, “A compositional context-sensitive multi-document summarizer: Exploring the factors that influence summarization,” in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2006, pp. 573–580.
[17] R. Mihalcea, “Language independent extractive summarization,” in Proc. ACL Interact. Poster Demonstration Sessions, 2005, pp. 49–52.
[18] D. Gillick and B. Favre, “A scalable global model for summarization,” in Proc. Workshop Integer Linear Program. Natural Language Process., 2009, pp. 10–18.
[19] I. Titov, “A joint model of text and aspect ratings for sentiment summarization,” in Proc. Conf. Assoc. Comput. Linguistics, 2008, 308–316.
[20] H. Nishikawa, “Learning to generate coherent summary with discriminative hidden Semi-Markov model,” in Proc. Assoc. Comput. Linguistics, 2014, pp. 1648–1659.
[21] D. Shen, “Document summarization using conditional random fields,” in Proc. IJCAI, 2007, pp. 2862–2867.
[22] J. M. Conroy and D. P. O’Leary, “Text summarization via hidden Markov models,” in Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2001, pp. 406–407.
[23] K. Knight, “Summarization beyond sentence extraction: A probabilistic approach to sentence compression,” Artif. Intell., vol. 139, pp. 91–107, Jul. 2002.
[24] M. Banko, V. O. Mittal, and M. J. Witbrock, “Headline generation based on statistical translation,” in Proc. 38th Annu. Meet. Assoc. Comput. Linguistics, 2000, pp. 318–325.
[25] C.-Y. Lin and E. Hovy, “Automatic evaluation of summaries using N-gram co-occurrence statistics,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol. (NAACL), 2003, pp. 71–78.
[26] Y. Dong, Y. Shen, E. Crawford, H. van Hoof, and J. Chi Kit Cheung, “BanditSum: Extractive summarization as a contextual bandit,” 2018, arXiv:1809.09672. [Online]. Available: http://arxiv.org/abs/1809.09672
[27] S. Narayan, S. B. Cohen, and M. Lapata, “Ranking sentences for extractive summarization with reinforcement learning,” 2018, arXiv:1802.08636. [Online]. Available: http://arxiv.org/abs/1802.08636

TABLE 4. An example of summarization.

Source: 快递公司发布“暴力分拣”丑闻，据中央电视台24日播出的节目，申通等快递公司的分拣工人还在包裹堆上随意行走，25日，网上又传出了圆通快递“暴力分拣”的视频。截至目前，申通快递已经承认错误，其他快递公司却保持沉默。
The courier company reappeared the scandal of “violent sorting.” According to the picture broadcast by CCTV on the 24th, the sorting workers from courier companies such as Shentong are still walking freely on the parcel stack; on the 25th, a video of “violent sorting” by YTO EXPRESS was also sent out on the Internet. As of yesterday, STO EXPRESS has acknowledged the mistake, while other express companies have remained silent.

The courier company reappeared the scandal of “violent sorting.” According to the picture broadcast by CCTV on the 24th, the sorting workers from courier companies such as Shentong are still walking freely on the parcel stack; on the 25th, a video of “violent sorting” by YTO EXPRESS was also sent out on the Internet. As of yesterday, STO EXPRESS has acknowledged the mistake, while other express companies have remained silent.
X. Zhang, M. Lapata, F. Wei, and M. Zhou, “Neural latent extractive document summarization,” 2018, arXiv:1808.07187. [Online]. Available: http://arxiv.org/abs/1808.07187

L. Hou, “Automatic chinese abstractive summarization with topical keywords fusion,” Acta Automatica Sinica, vol. 45, no. 3, pp. 530–539, Mar. 2019.

A. Vaswani, “Attention is all you need,” 2017, arXiv:1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762

Q. Zhou, “Neural document summarization by jointly learning to score and select sentences,” in Proc. Assoc. Comput. Linguistics, 2018, pp. 654–663.

Y.-C. Chen and M. Bansal, “Fast abstractive summarization with reinforcement-selected sentence rewriting,” 2018, arXiv:1805.11080. [Online]. Available: http://arxiv.org/abs/1805.11080

K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “MASS: Masked sequence to sequence pre-training for language generation,” 2019, arXiv:1905.02450. [Online]. Available: http://arxiv.org/abs/1905.02450

H. Zhang, J. Xu, and J. Wang, “Pretraining-based natural language generation for text summarization,” 2019, arXiv:1902.09243. [Online]. Available: http://arxiv.org/abs/1902.09243

H. Wang, X. Wang, W. Xiong, M. Yu, X. Guo, S. Chang, and W. Yang Wang, “Self-supervised learning for contextualized extractive summarization,” 2019, arXiv:1906.04466. [Online]. Available: http://arxiv.org/abs/1906.04466

R. Mihalcea, “TextRank: Bringing order into texts,” in Proc. Conf. Empirical Methods Natural Lang. Process. 2004, pp. 404–411.

S. Ma and X. Sun, “A semantic relevance based neural network for text summarization and text simplification,” 2017, arXiv:1710.02318. [Online]. Available: http://arxiv.org/abs/1710.02318

N. Kalchbrenner, I. Danihelka, and A. Graves, “Grid long short-term memory,” 2015, arXiv:1507.01526. [Online]. Available: http://arxiv.org/abs/1507.01526

B. Hu, “LCSTS: A large scale Chinese short text summarization dataset,” in Proc. Comput., 2015, pp. 2667–2671.

K. Lopyrev, “Generating news headlines with recurrent neural networks,” 2015, arXiv:1512.01712. [Online]. Available: https://arxiv.org/abs/1512.01712

S. Ma, X. Sun, W. Li, S. Li, W. Li, and X. Ren, “Query and output: Generating words by Querying distributed word representations for paraphrase generation,” 2018, arXiv:1803.01465. [Online]. Available: http://arxiv.org/abs/1803.01465

I. Sutskever, “Sequence to sequence learning with neural networks,” 2014, arXiv:1409.3215. [Online]. Available: https://arxiv.org/abs/1409.3215

J. Gu, “Incorporating copying mechanism in sequence-to-sequence learning,” in Proc. Conf. Assoc. Comput. Linguistics, vol. 3, 2016, pp. 1531–1640.

J. Lin, “Global encoding for abstractive summarization,” in Proc. Conf. Assoc. Comput. Linguistics, vol. 5, 2018, pp. 163–169.

D. Radov, “MEAD-a platform for multimodal multilingual text summarization,” in Proc. ELRA, 2013, pp. 699–702.

H. Lin, “Multi-document summarization via budgeted maximization of submodular functions,” Proc. Conf. Assoc. Comput. Linguistics, 2010, pp. 912–920.

L. Hou, “Abstractive document summarization via neural model with joint attention,” in Proc. Natural Lang. Process. Chin. Comput., 2018, pp. 329–338.

---

FUCHENG YOU is currently a Tutor, a Doctor, and a Professor of computer science with the School of Information Engineering, Beijing Institute of Graphic Communication. He is also the Director of Digital Image Processing Research Office with the Beijing Key Laboratory of High-End Printing Equipment Signal and Information Processing and the Director of Research and Recruitment Office with the School of Information Engineering. His main courses include database principle, digital image processing, image processing and analysis, and so on. He presided over many key projects with the Beijing Natural Science Foundation, the scientific research projects with the Beijing Municipal Education Commission, and the school-level key projects. More than 80 academic articles have been published in national and foreign publications, of which SCI/EI has searched more than 50. He holds over 12 national invention patents, five utility model patents, and two books. His main research interests include natural language processing, digital image processing, machine vision application, and so on. He received the Yachang Education Award, in 2016.

SHUAI ZHAO is currently pursuing the degree in computer science with the School of Information Engineering, Beijing Institute of Graphic Communication. He has participated in many critical projects with the Beijing Natural Science Foundation, the scientific research projects with the Beijing Municipal Education Commission, and the school-level vital projects. He has published three academic articles in national and foreign publications, including three SCI/EI searches. He holds over two national invention patents, one utility model patent, and one publication and compilation. His research interests include natural language processing, deep learning, and so on. He received the First-Class Scholarship, in 2019.

JINGJING CHEN is currently pursuing the degree in computer science with the School of Information Engineering, Beijing Institute of Graphic Communication. She has participated in many critical projects with the Beijing Natural Science Foundation, the scientific research projects with the Beijing Municipal Education Commission, and the school-level vital projects. She has published one academic article in EI searches. She holds one national patent for utility model and participated in publishing and compiling one. Her research interests include natural language processing, deep learning, and so on. She received the First-Class Scholarship, in 2019.

---

* * *