ATLANTIS PRESS

*Advances in Intelligent Systems Research, volume 147*

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

# Chinese Short Text Summary Generation Model Integrating Multi-Level Semantic Information

Guanqin Chen [a)]

*School of Computer, Guangdong University of Technology, Guangzhou 510006, China*

[a)] Corresponding author: 287212108@qq.com

**Abstract** Short text information is less, and short text comprehension abstract generation is currently a hot and difficult issue. We proposed an understanding-based short text summary generation model that combines multi-level semantic information. We improved the structure of encoder in the framework of encoder-decoder which consists of self-attention mechanism and selective network to focuses on the multi-level semantic information. Then our model fully exploits high-level global semantics and shallow semantic information of internal words of the text, and organically fuses the hidden state of the decoder and the original through two different attention mechanisms. The high-level and shallow semantic information adaptively provide the decoder with a syntactic semantic vector with abstract characteristics, so that the decoder can more accurately focus on the core content of the article. This paper selects the LCSTS data set for model training and testing. The experimental results show that compared with Seq2Seq, Seq2Seq with standard Attention, and Transformer model, the proposed model generates a Chinese short text summary with higher quality and performs better evaluation value in ROUGE.

**Key words:** Multi-level; Self-attention mechanism; Selective network; Global information; Seq2Seq; Joint semantic vector; Text abstract.

## INTRODUCTION

With the rapid development of Internet technology, online information is showing an explosive growth trend, and information overload is a serious problem. In the face of massive information, how to find useful information from it has become an urgent issue in the field of information acquisition and processing. Information extraction is a major area of research in natural language processing and refers to extracting important information from a large amount of text. Automatic summarization technology is the key technology for information extraction and compression. Since the emergence of automatic summarization technology in the 1950s, a wave of new automated abstract methods has emerged for every wave of new technologies. However, there is a certain gap between the results and artificial abstracts. According to the form of abstracts, automatic summarization can be divided into two major categories, which are Extractive and Abstractive [1]. Extractive summarization is based on the assumption that the core idea of an article can be summed up in one sentence and a few sentences of the article. Abstractive summarization is based on the understanding of the content of the article and a summary description is given. The description text does not have to be presented in the original text and it is closer to real intelligence.

At present, deep learning technology has been widely used in the field of natural language processing, including tasks such as machine translation, automatic question and answer, reading comprehension, automatic summarization, and creation [2]. The pure data-driven end-to-end automatic summarization generation method was originally borrowed from the neural network model of machine translation [3,4]. In the middle of 2015, K Lopvrev et al. proposed the use of LSTM (Long Short-Term Memory) [5] as a basic unit of RNN (Recurrent Neural Network) to construct an encoder-decoder-based decoder. The Encoder-Decoder model with the attention mechanism is used to generate news headlines [6]. Next, the two papers [7,8] published by Rush et al. from the Facebook Artificial Intelligence Research Institute from 2015 to 2016 to solve text summary generation tasks, based on the Encoder-

Decoder architecture, proposed different encoder approaches-based CNN (Convolutional Neural Network) and attention mechanisms, and decoder architecture based on the RNNLM (Recurrent Neural Network Language Model). Hu et al. [9] applied the RNN-based Encoder-Decoder architecture to the Chinese text digest task and constructed a Chinese text digest dataset LCSTS to facilitate the study of Chinese comprehension abstracts.

This paper mainly studies sentence-level Chinese short text comprehension abstract generation tasks and builds a summary generation model based on LCSTS data sets. The current generic abstract generation method is generally based on the Encoder-Decoder architecture, that is, a Seq2Seq(sequence-to-sequence) text representation learning model [10]. And then, the encoder and decoder structure and attention mechanism are studied and improved. At present, these models either use RNN or CNN encoder structures or hierarchically stacked RNN encoder structures to obtain long text sentence level and word level semantic information. The encoder based on RNN is essentially a Markov decision process, and it is unable to fully obtain the global representation of the original text. Moreover, the RNN model is a natural recursive structure and multi-layer RNN model results in slower performance in training; The CNN-based encoder is necessary to stack multiple layers and gradually expand the receptive field to obtain the global representation of the original text, and the CNN model is insensitive to position information. However, Chinese short texts have the characteristics of short text length and small number of sentences, and they require short and detailed abstracts. This requires the overall understanding of global semantics of short text as well as a comprehensive and accurate understanding of the semantics of the word in short text. Since every word in the original text has contextual semantics, combining the high-level global semantics and shallow semantics of the original word will help to obtain the abstract information in the original text. And focusing on keywords can also help generate summary. Therefore, this paper proposes a text summary generation model that integrates multi-level semantic information. The main points of improvement are three points: The first point is that by referring to the merits of the selective encoding model [19], an improved selective network is proposed to remove redundant information such as common words in the original text to provide more accurate input information for the follow-up. The second point is that by combining the advantages of the Transformer model [11] (Google Brain, 2017) and the RNN model which has natural advantage of catching location information we improve the encoder structure of the text summary model. Based on the RNN model, a multi-headed self-attention mechanism and feedforward network layer are added, enabling the model to grasp the internal relations of the original text, fully mining the global semantic representation and shallow semantic representation of each word and avoiding loss some important information of original text. The second point is that organic integration of hidden state of the decoder and high-level and shallow semantic information of encoder through the two attention mechanisms. It integrates the internal relations of the original global and local semantic information and the alignment relationship between the original and the abstract, which can obtain the multi-dimensional semantic representation of the original information, and combines them into a joint semantic vector, and adaptively provides appropriate original semantic information for the input of the decoder. Compared with the Seq2Seq, Seq2Seq with standard Attention, and the Transformer model based solely on the attention mechanism, the summary generated by the proposed model in this paper is more concise and consistent and performs better on the summary evaluation in ROUGE [12].

## INTRODUCTION TO ENCODER-DECODER AND ATTENTION MECHANISM

The basic architecture of Encoder-Decoder is first introduced in this section, followed by a general paradigm based on attention mechanism.

### Introduction to Encoder-Decoder Architecture

In order to solve the problem of sequence-to-sequence text generation, scholars proposed an encoder-decoder architecture [3,10]. The encoder encodes the input sequence into a semantic vector representation. The structure can be a RNN, a CNN encoder, and other encoder models that can represent the input sequence as a semantic vector. The stage of the decoder can be seen as the inverse of the encoding stage. According to the specific task, the semantic vector is decoded to generate the output sequence. The decoder can also be a variety of kinds of sequence generation models. One difference is that the decoder cannot be a bidirectional sequence structure. Because the sequence can only be generated afterwards. At present, the Encoder-Decoder structure has been widely used in the field of texts such as automatic question and answer, machine translation, and automatic summarization. The model structure is shown in Fig.1, The encoder converts the input sequence $(x_1, x_2 ..., x_n)$ to a fixed dimension semantic

encoding vector. The decoder produces the output sequence $(y_1, y_2..., y_n)$ just according to the transformed semantic vector of the encoder.

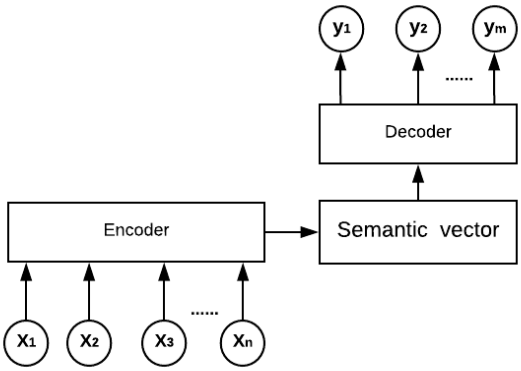


**FIGURE 1.** Encoder and Decoder architecture

The Encoder-Decoder model connects through a unique semantically encoded vector. The semantic encoding vector is an encoder that compresses the entire input sequence into a fixed-length semantic vector and cannot fully represent the entire sequence of information. Loss of encoded information does not allow the decoder to have enough semantic input information, leading to a reduction in the final decoding accuracy. In order to solve the incomprehensive problem of the output semantic vector information in the Encoder-Decoder model, scholars have improved the connection structure between encoder and decoder. In 2014, Bahdanau et al. introduced an attention mechanism [4] so that the decoder's input is no longer the single semantic vector output by the encoder, but the weighted sum of the hidden semantic vectors of the encoder input sequence.

## The General Paradigm of Attention Mechanism

The purpose of the attention mechanism is to do a weighted sum on each state vector of the original sequence to obtain different focused state vectors. Its general paradigm can be expressed as a feedforward neural network, as shown in Fig.2:

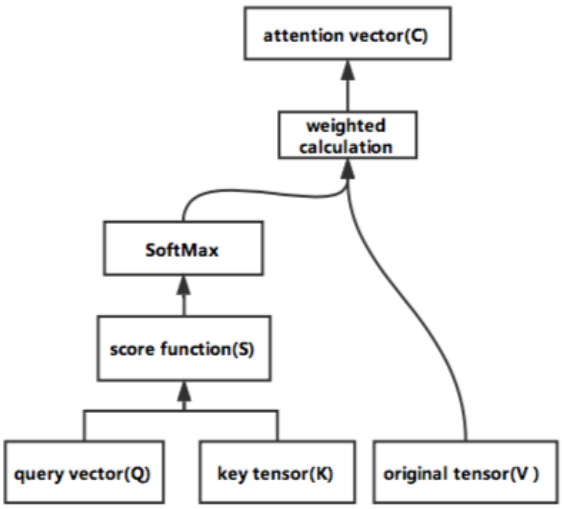


**FIGURE 2.** General mechanism of Attention Mechanism

In Figure 2, the original state tensor $V$ is a state matrix that requires weighted summation, which can be composed of the semantic vectors of each word of the source text; the state tensor $K$ is another representation of the linear transformation of the state tensor $V$ , which can make $V$ and the query vector $Q$ be in the same semantic space so that the similarity score of function S can be easily calculated. The query state vector $Q$ is a reference vector for computing the similarity scores of the original texts, and it may be a hidden state vector of the decoder or a hidden state matrix of the encoder. The similarity score function S has different forms [13]. It can be represented by a small forward feed neural network. It is mainly used to calculate the weight of each row vector of the original state matrix $V$ . The higher the score, the greater the weight. Finally, the output score of the S function is normalized by the Softmax activation function, and the original state vector is weighted and summed to obtain an appropriate attention vector $C$ .

$$C = \text{Attention}(Q,\ K,\ V) \tag{1}$$

## OUR PROPOSED MODEL

At the 2017 Nips conference, Google Brain proposed a purely attention-based transformation model [11] and used this model to achieve optimal translation performance on translation tasks. This paper implements the code of the model, migrates the model to the text summary generation task for the first time, and retrains the model on the LSCTS [7] data set to obtain the performance of the transformer model on the short text summary. It can be seen for details in the experimental section. However, the abstract generation of comprehension requires the comprehension of the original text and then compression and restatement. It is more complex than the translation task and it is impossible to generate high-quality abstracts through parallel alignment of two language. The transformer model completely uses the attention mechanism and has certain defects in the expression of the position sequence information. It is not applicable to the understanding abstract generation task. Therefore, this dissertation proposes a Chinese short text summary generation model that combines multi-level semantic information by learning the part structure of the encoder of the transformer model and the RNN model structure incorporating multiple attention mechanisms. Firstly, the proposed model introduces the encoder structure of selective network and multilayer self-attention mechanism on the basis of RNN model and obtains the high-level global semantics and shallow semantic information of the original text more fully. Then, under two different attention mechanisms, the model organically fuses the original multi-level semantic information and the decoder's hidden state information, generates context vectors with multi-channel information, and fuses them into a joint semantic vector with abstract characteristics. Finally, the decoder of the RNN language model generates a digest based on the joint semantic vector decoding. The above points ensure that the summary content generated by the model is more comprehensive and concise. The concrete structure of the text summary generation model that integrates multi-level semantic information in this paper is shown in Fig.3:

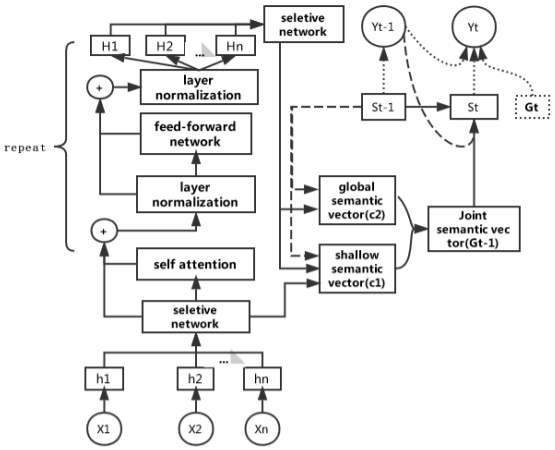


**FIGURE 3.** Text summary generation model with multi-level semantic information

In Fig.3, the model is divided into two parts: encoder and decoder. The encoder part is a cascade of RNN model, selective network and four identical layers. Each floor has two branches. The first branch is a multi-headed self-attention mechanism layer. The second branch is a simple full-connection feed-forward network. A mapped residual connection [14] and layer normalization [15] have been added outside the two branches. The decoder part is a decoder structure of an RNN language model. In the connection part of the encoder and the decoder, global semantic vectors and shallow semantic vectors are respectively obtained through two different attention mechanisms and are cascade-spliced into a joint semantic vector as the input of the decoder. This model is mainly to improve the structure of the encoder and design the relevant Attention mechanism. The main structure of the model is shown in the following three points:

selective network removes redundant information such as common words in the original text to provide more accurate input information for the follow-up. The specific formulas are as follows:

$$h = \left[h_1, h_2, ..., h_n\right] = \mathrm{RNN}(x_1, x_2, ..., x_n) \tag{2}$$

$$m = V_h \tanh(W_h h) \tag{3}$$

$$e_i = \frac{\exp(m_i)}{\sum_i^n \exp(m_i)} \tag{4}$$

$$h_{att} = \sum_i^n e_i h_i \tag{5}$$

$$a_i = \mathrm{Sigmoid}(U_a h_{att} + W_s h_i + b) \tag{6}$$

$$h_i^{'} = a_i h_i \tag{7}$$

First, a consistent representation of the hidden semantics $(h_1, h_2 ..., h_n)$ of the input sequence $(x_1, x_2 ..., x_n)$ is obtained by using the RNN model of equation (2). Then, through attention mechanism described by the formula (3), (4) and (5), $h_{att}$ which represents the global meaning of hidden semantics $(h_1, h_2 ..., h_n)$ is obtained. Finally, formula (6) is used to combine the global meaning $h_{att}$ local state $h_i$ ,and the weight $a_i$ is got by the sigmoid function. The new state $h_i^{'}$ is the output of selective network, which is described in formula (7).

(ii) The encoder part builds a high-level global semantic representation within the source word sequence by adding a multi-headed self-attention mechanism and a feedforward network based on the RNN model.

(1) The multi-head self-attention mechanism is a repeated self-attention mechanism and is a way of obtaining the internal relations of the original text. The specific formulas are as follows:

$$h^{'} = \left[h_1^{'}, h_2^{'}, ..., h_n^{'}\right] \tag{8}$$

$$C_h = \mathrm{Attention}(h^{'}, h^{'}, h^{'}) \tag{9}$$

$$Q = \mathrm{Relu}(h^{'}, W_q) \tag{10}$$

$$K = \mathrm{Relu}(h^{'}, W_k) \tag{11}$$

$$V = \mathrm{Relu}(h^{'}, W_v) \tag{12}$$

$$C_h = \text{Softmax}(QK)\bullet V \tag{13}$$

First, the selective hidden semantics $(h_1^{'}, h_2^{'}..., h_n^{'})$ is concated to $h^{'}$ in equation (8). The input of the attention mechanism comes from the same kind of information, and its paradigm is shown in formula (9). The specific calculation methods for the attention mechanism of this paper are shown in formulas (10) (11) (12) and (13). Equations (10), (11), and (12) respectively indicate that the input matrix $h$ is convert to a query matrix $Q$, a state matrix $K$, and semantic matrix $V$ through the non-linear mappings. The non-linear mapping is advantageous to transform the matrix $Q$, $K$ and $V$ to the same subspace. Then, formula (13) represents the similarity operation of the inner product of the query matrix $Q$ and the state matrix $K$ and the normalization of the weighted coefficients by the SoftMax function. Finally, the weighted coefficient matrix and the source semantic matrix $V$ are multiplied together to obtain the weighted global semantic matrix $C_h$. The self-attention mechanism can obtain the relative semantic relationship between each word and all words in the source text. That is, the semantics of each word can be linearly represented by the semantic vectors of all the original words. The multi-headed self-attention mechanism repeats multiple self-attention mechanisms through different parameter templates, obtains semantic information of a variety of different characteristics, and then forms a global semantic matrix through splicing.

(2) Feedforward network is a non-linear full-connection network. Its main function is to further convert the information output by the attention mechanism layer to provide the input for the next self-attention mechanism layer and two attention-mechanism layers at the decoder. The feedforward network is defined as shown in Equation (14).

$$\text{F}(X) = \text{Relu}(XW_1 + b_1)W_2 + b_2 \tag{14}$$

In order to effectively obtain high-level global semantic information and training effects of the model, the model repeatedly operates the self-attention mechanism layer and the feedforward network layer four times and adds some skill operations between the self-attention mechanism layer and the feedforward network layer. Including the mapped residual connection [14] and layer normalization [15].

(iii) As shown in Fig. 3, the model firstly obtains the shallow semantic vector $C_1$ and high-level global semantic vector $C_2$ of the original text through two attention mechanisms, and then fused into a joint semantic vector $G_t$ to adaptively provide the input of the digest characteristics to the decoder.

(1) The shallow semantic vector $C_1$ is obtained by weighted summation of the hidden semantic vectors of the RNN model at encoder. The calculation of the attention weight coefficient introduces high-level global semantic information. The specific calculation formula is as follows:

$$H_i^{'} = \text{selective}([H_1, H_2, ..., H_n]) \tag{15}$$

$$e_{t_i}^1 = \text{S}^1(h_i^{'}, h_{att}, H_i^{'}, s_{t-1}) \tag{16}$$

In formula (15), the vector $H_i$ emphasizes the high-level semantic encoding of the internal relationship of each word in the original text, $H_i^{'}$ is obtained by the selective network introduced in the first point (i). In formula (16), the vector $h_i^{'}$ emphasizes the shallow semantic encoding of words. The vector $h_{att}$ is obtained by weighted average of all shallow semantic vectors $h_i$ according to formula (5). The combination of the three vectors $h_i$, $h_{att}$, and $H_i^{'}$ is designed as concatenation, defined as formula (17) and (18).

$$\text{S}^1(h_i^{'}, h_{att}, H_i^{'}, s_{t-1}) = V_a^T \text{ Tanh}(U_h \left[ h_i^{'}, h_{att}, H_i^{'} \right] + U_s s_{t-1}) \tag{17}$$

$$C_{t-1}^{1}=\sum_{i=1}^{n}e_{t_i}^{1}h_i^{'} \tag{18}$$

In equation (17), $U_h$, $U_s$, and $V_a^T$ are the matrix and vector template parameters that need to be optimally solved. The specific calculation process of formula (17) is as follows: firstly, the shallow word semantic vectors $h_i^{'}$, global shallow semantic vectors $h_{att}$, and high-level semantic vectors $H_i^{'}$ are connected to a vector and it is multiplied with template parameters $U_h$ to be converted into a representation vector that combines the full-text and current word information; The hidden state vector $s_{t-1}$ of decoder and template parameter $U_s$ are multiplied to convert into the vector which represents the current state information of decoder. Then, the two do the additive operation of the corresponding elements and convert it into the fusion state vector which combines the information of the high-level and shallow semantic information of the $i$ word of the encoder global semantic information and the state information of the decoder at the time t-1 through the nonlinear activation function tanh. Finally, the fused state vector and the template parameter $V_a$ do an inner product operation to get the real value. The larger the value, the greater the contribution to the shallow semantic vector $C_{t-1}^{1}$. The final calculation of $C_{t-1}^{1}$ is as above formula (18).The attention mechanism here is mainly to integrate the original global semantic information, high-level and shallow word semantic information, so that the calculation of $C^1$ is more reliable and can pay more attention to the original abstract information. Keeping the proper shallow semantic information of the source text is conducive to decoding the core words of the original text.

(2) The high-level global semantic vector $C^2$ is obtained by weighted summation of high-level semantic vectors $H_i^{'}$ at the encoder side, and the definition formulas for the attention weight coefficient are as shown in formulas (19) and (20).

$$e_{t_i}^{2}=\mathrm{S}^2(H_i^{'},s_{t-1})=H_i^{'}Ms_{t-1} \tag{19}$$

$$C_{t-1}^{2}=\sum_{i=1}^{n}e_{t_i}^{2}H_i^{'} \tag{20}$$

In formula (19), $M$ is a matrix template parameter, and a similarity score value is obtained by merging the implicit state vector $s_{t-1}$ in decoder at the time t-1 and the high-level global semantic vector $H_i^{'}$ by a bilinear operation. The higher the current high-level semantic vector $H_i^{'}$ contributes to the global semantic vector $C_{t-1}^{2}$. The final calculation of $C_{t-1}^{2}$ is as above formula (20). Here, the weight coefficients calculated by the bilinear similarity operations are mainly used to adaptively retain appropriate high-level global semantic information and provide global semantic reference information for the decoder-generated decoding abstract. Here, the weight coefficient of $C_{t-1}^{2}$ is calculated mainly through the bilinear similarity operation, adaptively retains the appropriate high-level global semantic information, and provides global semantic reference information for the decoder.

The shallow semantic vector $C_{t-1}^{1}$ and the high-level global semantic vector $C_{t-1}^{2}$ are converted into a joint semantic vector $G_{t\text{-}1}$ through a layer of non-linear transformation. The calculation of $G_t$ is similar to $G_{t\text{-}1}$. And $G_t$ represents the compressed textual information and acts as an input vector to the decoder at time t. The decoder finally normalizes the probability of each predicted word at time t through the fully connected layer and the SoftMax activation function, defined as equation (21). And our goal is to maximize the output sequence of summary probability given the input sequence of short text. The objective function is log-likelihood function, defined as equation (22).

$$p(y_t|x) = \text{Softmax}(W_c G_t + W_o s_t + b_o) \tag{21}$$

$$J(\vartheta) = \frac{1}{|D|} \sum_{(x,y)\in D} \log p(y \mid x) \tag{22}$$

From equation (21) and (22), it can be seen that the joint semantic vector $G_t$ and the hidden state $s_t$ of the decoder are subjected to a fully connected neural network and SoftMax activation function to make the probability prediction of the output digest word. Therefore, the shallow semantic vector $C^1$ and the high-level global semantic vector $C^2$ obtained by introducing the improved encoder structure and dual attention mechanism in this model are significant for decoding and generating digests to maximize the objective function.

## EXTRACT HIGH-QUALITY DATA FROM LCSTS DATA SET

The LCSTS data set was proposed by Hu et al. [7] and is a large-scale Chinese short text summary data set taken from Sina Weibo. The data set includes political, economic, military, film, games, and people's livelihoods. More than 2 million real Chinese short texts and abstracts given by each text author. The data set is divided into three parts. The first part is the main part of the data set and contains 2,400,591 pairs of short text abstracts. This part of the data is used to train the model for generating abstracts. The second part consists of 10,666 pairs of manually annotated short texts. Each sample is scored 1-5. The score is used to judge the relevance of the short text to the abstract. 1 represents the least relevant and 5 represents the most relevant. This part of the data is randomly sampled from the first part of the data to analyze the distribution of the first part of the data. Among them, the sample texts labeled with 3, 4, and 5 scores are more relevant to the abstract. The third part included 1106 pairs of short text abstracts, and three people rated them.

Because there are duplicate text and abstract samples in the LCSCT data set, and there is low match between text and abstract in some samples of the LCSCT data set. The LCSCT data set needs to be filtered to extract high quality text summary data. First, the text preprocessing of the LCSTS data set includes the elimination of the text summarization pair, the removal of summary whose segment length is less than 3 and which does not consist of Chinese characters, replacement of numbers, English, special characters, URLs, dates, and more in text and summary. And then we build the category matching model for text and abstract by the Dual-LSTM model [16] improved by ourselves to get the good high-quality text summarization Paris. Finally, a total of about 1.3 million high-quality text summaries were obtained for the modeling and evaluation of automated text summaries. We segment high-quality textual abstractions into training, validation, and test sets, where the training set is from the first part of the LCSTS data. The validation set is a portion of the LCSTS data set that is randomly selected from the first and second parts that are greater than or equal to 3. The short text summary pairs are composed of 1000 short text summary pairs with a score above 3 points randomly selected from the remaining second and third parts of the LCSTS data set.

## ABSTRACT GENERATION EXPERIMENT SETUP AND RESULT ANALYSIS

### Implementation Details

In the experiments of the abstract generation model in this paper, for the encoder part, the underlying RNN model uses a four layers structure of 400 GRU units. The selective network uses the matrix parameters with 800 dimensions to get the attention vector. The self-attention mechanism layer uses six identical self-attention mechanism layer cascaded structures, which is similar to the multi-convolution kernel form in CNN. The final output of the feedforward network layer is a high-level semantic vector of 800 dimensions. For the decoder part, the RNN decoding model uses the four layers unidirectional RNN structure of 400 GRU units. To alleviate the risk of word segmentation mistakes [20], we use Chinese character sequences [17] based on character splitting as both source inputs and target outputs. The vocabulary size of model is set to 4000. The number of 4,000 characters in the whole corpora accounts for about 99% of the character vocabulary frequency. In order to speed up the convergence in model training, a pre-trained character vector with 300 dimensions is used in the character embedding layer. Due to the large loss in the initial training period, all the parameters of the word embedding layer are set to be untrainable.

After a few iterations of the initial training, the character embedding layer parameters are fine-tuned. And it is to use a self-adaptive strategy to adjust the learning rate in model training. when training the model, the cross-entropy solution of the loss function uses a label smoothing [18] strategy, which is beneficial to improve the generalization ability of the model.

## Analysis of Experimental Results

The document abstract evaluation method is roughly divided into two categories: (1) Intrinsic Methods. On the premise of providing a reference abstract, the quality of the system summary is evaluated based on the reference abstract. In general, the more the system summary and the reference summary match, the higher the quality. (2) Extrinsic Methods. This method does not provide a reference abstract. Instead of the original document, the document abstract is used to execute a certain document-related application, such as document retrieval, document clustering, document classification, etc. The summary that can improve the application performance is considered to be a high-quality summary. This article uses an internal evaluation method ROUGE [12] that is commonly used for automatic textual summarization. ROUGE is based on n-gram co-occurrence information in the abstract to evaluate the quality of abstracts. It is an evaluation method based on the recall rate of n-grams. This article uses the ROUGE-1, ROUGE-2, and ROUGE-L to evaluate summary performance of the models.

Table 1 shows the ROUGE evaluation values of the summary generation of the four models implemented by ourselves on the test set, where En-De represents the basic Encoder-Decoder model based on RNN, and En-De-ATT represents the Encoder-Decoder model with Bahdanau's attention mechanism [4]. En-De-MLS represents our proposed model that fuses multi-level semantic information, and Transformer [11] represents an Encoder-Decoder model based entirely on the attention mechanism. Beam search means the model uses the beam search algorithm when predicting the generation of a summary, and the beam size is set to 10. Greed means only obtains the best result at every time step of the decoder. From the experimental results, it can be seen that the evaluation performance in ROUGE of the classical attention mechanism model is better than that of the En-De model. The evaluation performance of Transformer model and En-De-ATT model generation summary is similar, but the training time efficiency in transformer model is better than the En-De-ATT model. The evaluation performance in ROUGE of our abstract model combined with multi-level semantic information is better than all other baseline model. Beam search algorithm in decoding always gets the better result than the greedy algorithm. The Transformer model completely uses the self-attention mechanism and it is good to model the internal relations of the text. The En-De-ATT model performs better on the position information of the sequence. Our model combines the advantages of the En-De-ATT model and the Transformer model structure, catching the characteristics of the text summary task through the attention mechanism both in encoding and decoding stages. Therefore, the quality of the summary generated by our proposed model is better, and the summary assessment has a higher ROUGE value.

**TABLE 1.** Text summary ROUGE value evaluation results

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| En-De (beam search) | 0.225 | 0.093 | 1.98 |
| En-De-ATT(greedy) | 0.298 | 0.182 | 0.276 |
| En-De-ATT (beam search) | 0.302 | 0.186 | 0.281 |
| Transformer (beam search) | 0.298 | 0.173 | 0.275 |
| En-De-MLS(greedy) | 0.312 | 0.94 | 0.288 |
| En-De-MLS (beam search) | 0.314 | 0.197 | 0.290 |

Table 2 is several short text and summary samples from test sets and the summaries generated by En-De-ATT, Transfomer, and our model with beam search algorithm. It can be seen from Table 2 that although the summary generation model of this paper is a pure data-driven end-to-end model, the generated summary is straightforward, and it captures the key content of short texts, and gets the time, object, location, and events and other key information. At the same time, the abstract generated by our model is not only a simple copy of the original text utterance, but also a new sentence, including the use of recombination of original words and the generation of new words to reconstruct the abstract sentence. In case 2, 3, and 4, compared with the En-De-ATT and Transformer models, the summary generated by our model has a stronger expressive force to grasp the central content of the short text. In case 4, the summary generated by the En-De-ATT model appears duplicate words, but the model in this article can generate key words in summary. Statistics show that the number of summary generated by our model

which contains new character not in corresponding short text is more than the En-De-ATT model, but less than the Transformer model. The strong ability to create new characters is one of the reasons that summary evaluation value of ROUGE in Transformer model is not high. What is more, the number of the summary generated by our model which consists of duplicated words is less than that of En-De-ATT and Transformer model. And the number of summary generated by our model consists of UNK (unknown words) is also less than that of other models.

According to Case 3 in Table 2, given the En-De-ATT model and the proposed model in this paper, the attention mechanism weights can be visually analyzed. The visualized result of the weight of our model in this article is to add the weights of the two attentional weights to obtain the final weight. The final attention weight coefficient of our model shown in Fig.5 is obtained by correspondingly adding the weight coefficients of two different attention mechanisms. The visualization results of the attention weights of the En-De-ATT model and our model are shown in Fig. 4 and Fig. 5, respectively, where the abscissa represents the short text. Due to size constraints, there is only part of the original text in the following figure, and the ordinate represents the summary. The darker the color in the figure, the greater the weight of the corresponding position of character.

In Fig. 4 and Fig. 5, the weights of the characters in the summary correspond to the same characters in the original text is the largest, and the weights of other characters are very small. The models pay attention to the corresponding characters information in the original short text when generating the abstract. It is shown that our proposed model can obtain the central content of the short text. Selective network of our model can remain the central information of short text and two mechanisms provides input that contains the joint semantic information of the source short text for the decoder.

Finally, considering the space and time complexity, we simply analyze the space-time overhead of the model. Compared with the classical Seq2Seq+Attention model, our model contains more a part of parameters of the self-attention mechanism layer, selective network, the feedforward network layer and the multi-attention mechanism, so the parameter amount are about twice as much, but within the acceptable range. In terms of time efficiency, from the number of single-layer parallel computations, because the RNN model needs to be recursively steps, the time complexity is $O(n)$. However, the time complexity of the selective network, self-attention mechanism layer and feedforward network layer is $O(1)$. From the analysis of the computational complexity of a single layer, the complexity of the RNN model is $O(nd^2)$, where $d$ is the number of hidden units. The complexity of self-attention mechanism layer is $O(n^2 d)$ and of selective network is $O(nd^2)$, which can be run parallel. The dual attention mechanism of our model can also run in parallel, so the training and testing time of our model would not be much slower than that of the Seq2Seq with Attention model. However, the quality of the summary generated by our model is better.

## SUMMARY AND OUTLOOK

Under the framework of encoder-decoder, we propose a new encoder structure. Through combining the RNN model with the selective network and the encoder structure of the self-attention mechanism, we can focus on original multi-level semantic information including the high-level and shallow semantics of the original text. Then through two different attention mechanisms to perform their duties, the model can adaptively merge the original joint semantic information and part of the summary status information of decoder at the current moment to generate the next character of summary. Experiments show that compared with the Seq2Seq with standard Attention model and the Transformer model based entirely on the attention mechanism, the performance of the summary generated by our model in ROUGE [12] is better, and the more concise and consistent summary can be generated, which means the generated summary by our model consists of less duplicated characters than that of other models in this paper. For the summary generation model of this paper, there are many research directions that can be explored in the future. The first is that for the structure of the decoder, the self-attention mechanism layer can be added so that the decoder can learn more internal relationship information. Secondly, for the decoding process of the decoder, a multi-stage deliberate decoding process can be explored.

## ACKNOWLEDGMENTS

Author brief introduction: Chen Guanqin (1992-), male, master graduate, the main research direction includes deep learning and natural language processing.

Project fund: National Natural Science Foundation of China (61472089, 61502108); NSFC-Guangdong Joint Fund (U1501254); Guangdong Natural Science Foundation (2014A030308008, 2014A030306004).

**TABLE 2.** Summary of the model generated case

| |
|---|
| Case 1:<br>Short text: Apple today released its fiscal first fiscal quarter results. According to the report, Apple's net revenue for the first fiscal quarter was US$74.599 billion, a 30% increase from the US$57.594 billion of the same period last year, a record high; net profit was US$18.024 billion, an increase of 38% from US$13.072 billion in the same period of last year. It also set a new record high.<br>Reference summary: Apple announced its first fiscal quarter earnings: net profit rose 38% year-on-year.<br>En-De-ATT: Apple's first-quarter net profit of NUM billion US dollars up NUMPERCENT year-on-year<br>Transformer: Apple's first quarter net profit NUM billion US dollars hits a record high<br>En-De-MLS: Apple reported first-quarter net profit increased NUMPERCENT year-on-year.<br>Case 2:<br>Short text: One of the characteristics and purposes of the Shanghai Free Trade Zone is that it can lead the reform of the Chinese financial market. The financial elements in the products in the region are implied, and they must achieve the level of liberalization of the elements. For companies, all involved in the same renminbi business, they can choose the most efficient one of offshore and onshore financial derivatives and reduce costs through financial innovation.<br>Reference summary: What financial innovations can Shanghai Free Trade Zone have<br>En-De-ATT: Shanghai Free Trade Zone features and the same RMB business<br>Transformer: NUM steps of financial reform in Shanghai Free Trade Zone.<br>En-De-MLS: Shanghai Free Trade Zone Financial Innovation.<br>Case 3:<br>Short text: The A-share company Qunxing Toys issued an announcement yesterday, announcing that it intends to acquire 100% of the shares of the "National Hero" developer and game company Xingchuang Interchange by issuing shares and paying cash. The total transaction price is 1.44 billion yuan. Qunxing Toys believes that toys and games, as entertainment and entertainment carriers, have more consistency, interoperability, and complementarity.<br>Reference summary: Qunxing Toys plans to acquire 1.44 billion to acquire the National Hero Developer<br>En-De-ATT: Qunxing Toys plans to acquire NUM billion shares of Xingchuang<br>Transformer: A shares announce A shares of NUM billion yuan to acquire Hero Internet Corporation<br>En-De_MLS: Qunxing Toys plan to acquire the National Hero Developer Game<br>Case 4:<br>Short text: "We are suffering from Internet finance's madness," Xiao Wang said in a jocular manner. "It's not so much about Internet finance as it is about 'big money'." The Internet has brought a very good vision to the fund. Everyone is very "hot" now, but we must realize that it is not "If you invest money today, you will be able to invest money," and you should have a long-term strategic vision.<br>Reference summary: Fund ecommerce suffers from internet financial madness<br>En-De-ATT: Internet Finance Internet Finance<br>Transformer: Wang Xiaochuan Internet finance is a hot investment<br>En-De-MLS: Internet financial madness. |

## REFERENCES

1. Sanderson M. Book Reviews: Advances in Automatic Text Summarization [J]. Information Retrieval, 2000, 4(1):82-83.
2. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing[J]. Computer Science, 2015.
3. Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [J]. Computer Science, 2014.
4. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, 2015.
5. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8): 1735–1780, 1997.
6. Lopyrev K. Generating News Headlines with Recurrent Neural Networks [J]. Computer Science, 2015.
7. Rush A M, Chopra S, Weston J. A Neural Attention Model for Abstractive Sentence Summarization [J]. Computer Science, 2015.

8. Chopra S, Auli M, Rush A M. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks[C]. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016:93-98.
9. Hu B, Chen Q, Zhu F. LCSTS: A Large-Scale Chinese Short Text Summarization Dataset [J]. Computer Science, 2015.
10. Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks [J]. 2014, 4:3104-3112.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems. 2017: 6000-6010.
12. Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C]. The Workshop on Text Summarization Branches Out. 2004:10.
13. Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[J]. Computer Science, 2015.
14. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. 2015:770-778.
15. Ba J L, Kiros J R, Hinton G E. Layer normalization [J]. arXiv preprint arXiv:1607.06450, 2016.
16. Lowe R, Pow N, Serban I, et al. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems [J]. Computer Science, 2015.
17. Zhang H, Li J, Ji Y, et al. Understanding Subtitles by Character-Level Sequence-to-Sequence Learning [J]. IEEE Transactions on Industrial Informatics, 2017, 13(2):616-624.
18. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:2818-2826.
19. Zhou Q, Yang N, Wei F, et al. Selective Encoding for Abstractive Sentence Summarization[J]. 2017.
20. Jingjing Xu and Xu Sun. 2016. Dependency-based gated recursiveneural networkfor chinese word segmentation. In Meeting of the Association for Computational Linguistics. pages 567–572.