ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 147

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

Chinese Short Text Summary Generation Model Combining Global and Local Information

Guanqin Chen [a)]

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

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

Abstract. Short text comprehension summary generation is currently a hot issue. In this paper, we improve the attention mechanism under the framework of encoder-decoder and proposes a comprehensible short text abstract generation model that integrates the global and local semantic information. The model consists of a dual encoder and a decoder. The dual encoder structure can combine the global and local semantic information and fully obtain the abstract features of the original text. And the improved mechanism can adaptively combine all information of short text to provide the input with summary characteristics for the decoder, so that the decoder can more accurately focus on the core content of the source text. In this paper, LCSTS dataset is used to train and test the model. The experimental results show that compared with the Seq2Seq and Seq2Seq with standard attention models, the proposed method can produce high-quality summary which consists of less repetitive words and performs better evaluation value in ROUGE.

Key words: Dual encoder; Attention mechanism; Global information; Local information; Text summary; Seq2Seq.

INTRODUCTION

Automatic text summarization is the use of computers to automatically generate text summaries. It is a traditional and cutting-edge research field. According to the form of abstracts, automatic summarization can be divided into two major categories, which are Extractive and Abstractive [1,2]. Extract summarization is based on the hypothesis that the core idea of an article can be summed up in one sentence and a few sentences of the article. On the contrast, Abstractive summarization is based on the understanding of the content of the article. The description text does not have to be presented in the original text and is closer to real intelligence. Compared with extractive abstracts, the comprehension abstract generation process is closer to human thinking and can more accurately reflect the meaning of the text. However, the comprehension summary is more difficult, involving the understanding of natural language and the re-creation of text. The summary generated by the current study still cannot effectively solve the difficulties such as incomplete understanding and semantic logic of text.

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. The pure data-driven end-to-end automatic summarization generation method was originally borrowed from the neural network model of machine translation [3, 4]. K Lopvrev et al. built an abstract generation model based on the encoder-decoder framework in 2015 by using RNN(Recurrent Neural Network) with unit of LSTM(Long Short-Term Memory)[5] and used an attention mechanism to generate news headlines[6].Secondly, the two papers[7, 8] published by Rush et al. from the Facebook Artificial Intelligence Research Institute from 2015 to 2016 to solve the text abstract generation task, 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 RNN-based Encoder-Decoder architecture to Chinese text digest tasks 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 methods are all based on the Encoder-Decoder architecture, namely the Seq2Seq (Sequence to Sequence) text representation learning model [10]. Then the encoder and decoder structure and attention mechanism are studied and improved. Currently, these models either use a single encoder structure or use a hierarchically superimposed encoder structure to obtain long text sentence level and word level semantic information. However, Chinese short texts have the characteristics of short text length and small number of sentences and require that the generated abstracts are short and concise. When multiple independent encoders work together, they can make full use of the global and local information of the original text and grasp the abstract characteristics of short texts. The process in which we humans write abstracts is to read the article first, grasp the overall meaning of the article, and then write a summary based on the original text and its own understanding. Then the text summary generation model needs to simulate the process by first having an encoder that obtains the global semantic information of the article, and then combining the global semantic information with the local semantic information of the original word, the decoder side generates a summary of the article verbatim from the forward and backward. Therefore, the text summary generation model studied in this paper has two major improvements. The first point is that at the encoder side, two independent encoder structures are used. The global encoder focuses on obtaining the high-level semantic vector representation of the full text. The local encoder focuses on obtaining the semantic vector representation of each word of the original sequence. The two encoders fully express the semantic information of the original text and avoid excessive loss of important information in the original text. The second point is that on the decoder side, the attention mechanism fuses the semantic information of the two encoders and the hidden state of the decoder. Further, it is the internal relationship between the global and local information of the original text and the alignment between the original text and the abstract, which can make decoder obtain more comprehensive and focused input information in the original text. The seq2seq model that incorporates global and local semantic information is also an attempt to simulate the process of reading passage and writing abstract by human readers. By testing on the LSCTS dataset, compared with the classical Seq2Seq with Attention model, the summary generated by the improved model in this paper is more concise and consistent, and performs better on the summary evaluation in ROUGE [14].

ENCODER-DECODER AND BASIC PRINCIPLES OF 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 decoder can be seen as the inverse stage of the encoding. 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 fixed dimension semantic encoding vector.

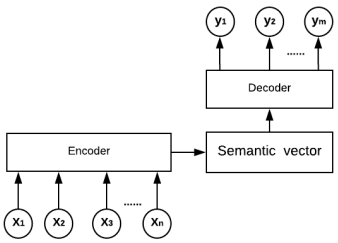

FIGURE 1. Encoder and Decoder architecture

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.

The Basic Principle of Classical Attention Mechanism

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. Due to the encoder, the RNN model is classical. The Encoder-Decoder model and attention mechanism based on the recurrent neural network will be described in detail below.

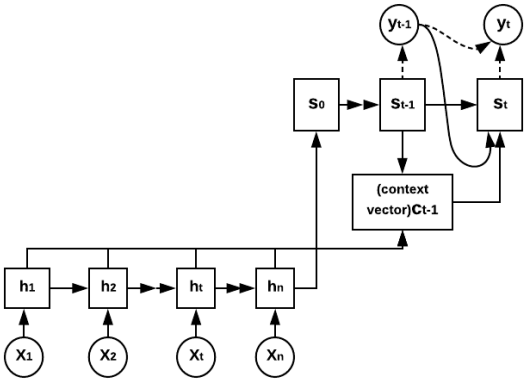

FIGURE 2. RNN-based Encoder-Decoder Attention Mechanism model

As shown in Figure 2, the RNN encoder converts the input sequence $(x_1, x_2..., x_n)$ to a high-dimensional hidden representation $(h_1, h_2..., h_n)$ and uses the output hidden state at the last moment as the initial hidden state of the RNN decoder model. The RNN decoder model combines the context vector to decode the output sequence $(y_1, y_2..., y_n)$. The final conditional probability formula defined by the decoder is as follows:

$$\mathrm{p}(y_t \mid y_{t-1}, \ldots, y_1, x) = \mathrm{g}(y_{t-1}, s_t, c_t) \tag{1}$$

Formula (1) shows that under the conditions given by y_{t-1}、 s_t and c_t, the probability of which word appears at time t is the greatest. y_{t-1} is the output label of the RNN decoder at time step t-1, s_t represents the hidden state of the decoder at time step t, c_t represents the context vector obtained by the attention mechanism at time step t, and the g function represents the conditional probability likelihood function. The calculation of s_t is as follows (2).

$$s_t = \mathrm{f}(y_{t-1}, s_{t-1}, c_{t-1}) \tag{2}$$

Equation (2) indicates that the RNN decoder obtains the output state at time step t through the output word at time step t-1, the hidden state at time t-1, and the context vector at time step t-1 obtained from the attention mechanism. The f function is a nonlinear activation function. The context vector in formula (2) is the core of the

entire attention mechanism, which is obtained by weighted summation of all the hidden states of the RNN encoder, indicating which word information in the source sequence needs to be emphasized by the decoder at time step t. The calculation is defined by the following formulas (3), (4) and (5).

$$c_{t\text{-}1} = \sum_{i=1}^{n} a_{t_i} h_i \tag{3}$$

$$a_{t_i} = \frac{\exp(e_{t_i})}{\sum_{i=1}^{n} \exp(e_{t_i})} \tag{4}$$

$$e_{t_i} = \mathrm{S}(h_i, s_{t-1}) \tag{5}$$

Formula (3) represents the weighted sum of the hidden state vectors of the encoder. The coefficient a_{t_i} is obtained by the softmax function of equation (4), which means that the contribution to each hidden state of the encoder is normalized. e_{t_i} in Equation (4) is calculated by Equation (5), which depends on the hidden state of decoder at time step t-1 and the i-th hidden state of the encoder. The similarity score calculation function S can be a direct inner product, a bilinear multiplication [11], and the like.

The attention mechanism was originally used for the automatic alignment of words between the source sentence and the target translation sentence in the machine translation task [3], that is, the local alignment of the hidden semantic information of the decoder and the hidden semantic information of the encoder. However, the comprehensible summary generation task is not only to achieve the word alignment of the source and the summary, but also to use the global semantic information to generate the comprehension abstract. In this paper, through the dual encoding and the attention mechanism of the global and local semantic information, the multi-encoder summary generation model makes the calculation of context vector weights more reliable and can adaptively provide the input information of the summary characteristics for each time step in the decoder.

OUR PROPOSED MODEL

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, which is similar to translation task. Therefore, this paper proposes an attention mechanism generating model that fuses global and local semantic information. Firstly, our model introduces the dual encoder structure, containing the global encoder and local encoder. Dual encoders perform their own tasks to extract the global and local semantic information of the original text. Then, under the improved attention mechanism, the model organically integrates the original global and local semantic information as well as the state information of the decoder to generate a context vector with abstract characteristics. Finally, the model adaptively combines global and local context semantic information to generate semantic vectors for multi-channel information fusion, providing the decoder with appropriate information input. The above points ensure that the summary content generated by the model is more comprehensive and concise. The concrete structure of the attention mechanism model for integrating global and local semantic information in this paper is shown in Figure 3:

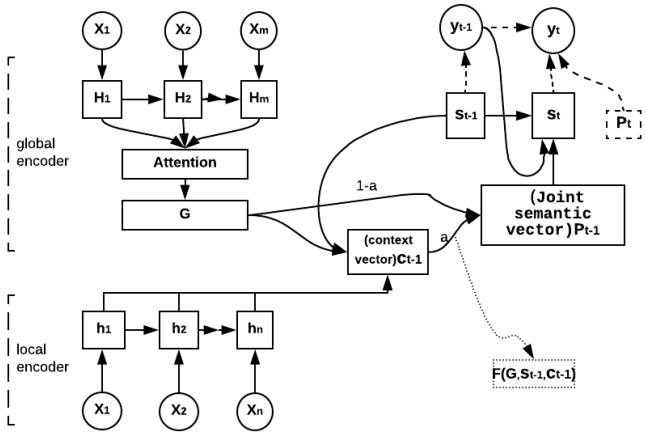

FIGURE 3. Attention mechanism model for integrating global and local semantic information

In Fig. 3, the global encoder focuses on the global semantic vector representation of the text, transforms the input sequence $(x_1, x_2 ..., x_m)$ into a high-level semantic representation vector G through a combination of RNN with attention model; the local encoder focuses on the consistent representation of the local semantics of the original text, The RNN model converts the input sequence $(x_1, x_2 ..., x_n)$ into a hidden representation $(h_1, h_2 ..., h_n)$. Dual-encoder models perform their duties, and the global and local semantic vectors of dual encoders are achieved through an improved Attention mechanism. The double-encoder models perform their duties. Through the improved Attention mechanism, the global and local semantic vectors of the dual encoder are combined to obtain joint semantic vector, which provides the decoder with appropriate and rich high-level semantic input information. Compared with the classical Seq2Seq with Attention, the improved model of this paper is mainly to introduce the dual encoder structure and design the related attention mechanism. The main internal structural changes of the model are shown in the following two points:

(i) We introduce the global RNN encoder which can fully obtain the global meaning of the short text. And its specific calculation process is shown in following formula (6), (7), (8) and (9). First, a consistent representation of the hidden semantics states $(H_1, H_2 ..., H_m)$ of the input sequence $(x_1, x_2 ..., x_m)$ is obtained by using the global RNN model of equation (6). And then, the import score s_i^H is obtained by the similarity function of formula (7) in which V_H and W_H are the parameters that should be to be optimized. At last the importance scores are normalized to the global vector G which represents the global meaning of hidden semantics states $(H_1, H_2 ..., H_m)$ by weighted sum:

$$H = [H_1, H_2, ..., H_m] = \text{RNN}(x_1, x_2, ..., x_m) \tag{6}$$

$$s_i^H = V_H \tanh(W_H H) \tag{7}$$

$$e_i^H = \frac{\exp(s_i^H)}{\sum_i^m \exp(s_i^H)} \tag{8}$$

$$G = \sum_i^m e_i^H H_i \tag{9}$$

In formula (10), the vector G emphasizes the global semantic encoding of the original text, and the vector h_i emphasizes the consistent semantic encoding of input sequence word. The combination of semantic vector h_i and G are designed in the form of cascade concatenation. The specify calculation of C_{t-1} is defined as follows:

$$S_{t_i}^h = \mathrm{S}(h_i, G, s_{t-1}) = v_e^T \tanh(U_h[h_i, G] + U_s s_{t-1}) \tag{10}$$

$$e_{t_i}^h = \frac{\exp(S_{t_i}^h)}{\sum_{i=1}^{n} \exp(S_{t_i}^h)} \tag{11}$$

$$C_{t-1} = \sum_{i=1}^{n} e_{t_i}^h h_i \tag{12}$$

In formula (10), U_h, U_s and v_e^T are vector parameters that need to be optimized. First, the full text semantic vector G and h_i are cascaded and multiplies with template parameters U_h to convert into the representation vector of the current state information; the decoder hidden state s_{t-1} multiplies with the template parameters U_s to convert into the representation vector of the current state information in decoder. Then, the two do the additive operation of the corresponding elements and pass through the tanh nonlinear activation function to convert into a fusion state vector combining the word semantic information of the i step at the encoder and the state information of decoder at time step t-1. Finally, the merged state vector and template parameters do an inner product operation, which is graphically mapped to a real value by a similarity operation. The larger the value, the greater the contribution to the input context vector. The similarity scores are converted into probabilities by Softmax function by equation (11). The final calculation formula of context vector C_{t-1} is weighted sum of hidden state vector of encoder, which is shown as equation (12) above. Here, the dual-encoder structure is used to introduce the original and global dual-channel feature information, so that the calculation can pay more attention to the original feature of the abstract.

(ii) The global semantic vector introduced at each time step in the decoder should be focused instead of simply repeating. Therefore, a global semantic vector G and context vector C_{t-1} should be input adaptively for the decoder, and the specific formula is as follows:

$$a = \mathrm{Sigmoid}(V_a^T \tanh(W_G G + W_c C_t + W_s s_{t-1})) \tag{13}$$

$$P_{t-1} = \left[(1-a) * G, a * C_{t-1}\right] \tag{14}$$

In equation (13), V_a^T, W_G, W_c and W_s are the template parameters that should be optimized. The weight coefficient a is obtained through the perceptron defined in equation (13) whose input are global vector G,context vector C_{t-1} and the state vector s_{t-1} of the decoder at time step t-1. And in equation (14) the joint semantic vector P_t is obtained through the concatenation between global vector multiplied by the weight coefficient a and context vector multiplied by the weight coefficient $(1-a)$. It means how much proportion of global semantic and local context information should be left as the input of decoder at time step t. The weight coefficient a controls the contribution between global vector and local context vector to the joint semantic vector which should be input for the decoder. If the weight coefficient a is equal to one, it means the joint semantic vector P_{t-1} is equal to C_{t-1},

which means the input for the decoder only from the context vector. The purpose of this is to avoid the decoder introducing too much redundant information and generate a summary of repeated strings.

Because the word of summary often appears in source text, and the copy network [15,16] is used to solve this problem and copy network also can produce the unknown word that is not in vocabularies. According to this, we also improved probability prediction of our model when predicting the summary word at every time step. In equation (15), the decoder uses the SoftMax activation to normalize the probability of each predicted word at time t through the fully connected layer whose inputs are s_t, y_{t-1}, and P_t whose calculation is similar to the P_{t-1} and sums it with the other probability P_a which is considered to produces the better summary word of source text instead of the wild symbols. P_a is defined as equation (16) in which when the output y_t is in the original text and belong to the P set which only consists of wild symbols we defined in data prepressing stage, P_a is probability of $e^h_{t_i}$ defined in equation (11) and otherwise P_a is zero. And we jointly our encoder and decoder by maximizing the log-likelihood to decode the correct word at each time step. Therefore, we optimize the negative log-likelihood loss function, defined as equation (17), where D denotes a set of parallel text-summary pairs and ϑ is the model parameter.

$$p(y_t|x) = \mathrm{Softmax}(W_c\mathrm{P}_t + W_o s_t + y_{t-1} + b_o) + P_a(y_t \mid x) \tag{15}$$

$$\begin{aligned} P_a(y_t \mid x) &= \max(e^h_{t_i}) \quad \text{if } y_t = x_i \in \mathrm{P} \quad i \in (1,2,\ldots n) \\ P_a(y_t \mid x) &= 0 \quad\quad\quad\quad\quad\quad \text{if } y_t \notin \mathrm{P} \text{ or } y_t \neq x_i \end{aligned} \tag{16}$$

$$J(\vartheta) = -\frac{1}{|D|}\sum_{(x,y)\in D} \log p(y \mid x) \tag{17}$$

DATA SET INTRODUCTION AND DATA PREPROCESSING

LCSTS Data Set Introduction

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.

Extract High-Quality Text Summary Data

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 with the wild symbols including NUM, UNK, NUMPERCENT, NER, TIME and so on. Then build a model that automatically extends large number of high quality textual summaries, as follows:

(1) Construct a positive and negative sample of whether the short text and the digest pair match: In the second part of the LCSCT data set, the text summarization pair whose score labeled by human is less than 3 is marked as a negative sample, and the text summary pair whose score is greater than or equal to 3 labeled by human is marked as a positive sample. And by using the mean word vector method to find the cosine similarity between the short text and the abstract and setting the threshold, the method is extended to obtain some positive and negative samples. Using these positive and negative samples as the basic corpus, construct a two-category matching model for text and abstract.

(2) The structure of the two-category model is shown in Fig. 4. It is based on the Dual-LSTM model framework and consists of two BiRNN (Bidirectional Recurrent Neural Network) that do not share LSTM unit parameters. They are BiRNN model of short text and BiRNN model of summary. First, the short text BiRNN model convert the short text sequence to obtain a semantic vector representation of the short text and multiplied with the template parameter W to be converted into a new text semantic vector T. Then, the summary RNN model convert the summary sequence to obtain the semantic vector representation of the summary and multiplied by the corresponding element of the new text semantic vector to obtain the similarity fusion vector of the text summary pair. Finally, after a fully connected layer and cross-entropy loss, a two-category model of matching degree between text and abstract is obtained. This model is based on the simple improvement of the Dual-LSTM model [12] in the Q & A retrieval field, which is better than the direct application of the text digest to match degree, both accuracy and recall.

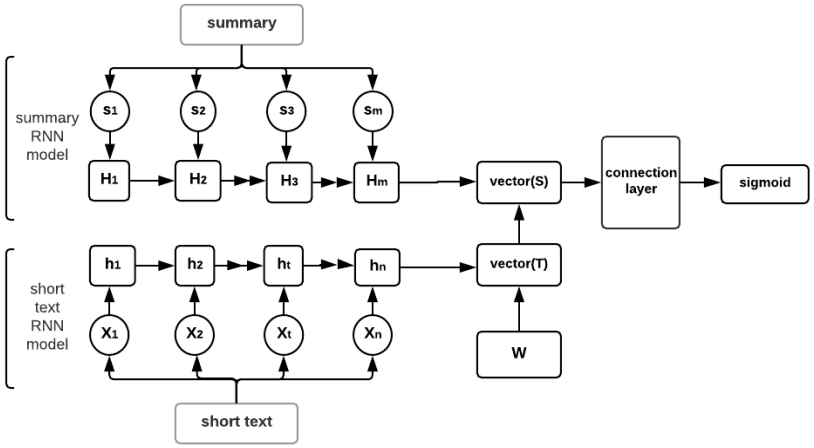

FIGURE 4. Text summary pair matching model

(3) First, the two-category matching model for text and abstract is used to predict the score of the first part of the LSCTS data set and a positive sample with a score of more than 0.95 and a negative sample of less than 0.1 are taken out and added to the training set again. Then, using this model to perform secondary modeling on the new training set and predicting the score of the unlabeled text summary pair, the sample with the predicted score greater than 0.9 is taken out as a sample with a better matching degree between the text and the abstract. 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 experiment of this summary generation model, the global encoder uses the RNN with attention model, in which the RNN uses a 4-layer BiRNN structure of 300 GRU units. the local encoder uses a 4-layer BiRNN structure of 300 GRU units; the decoder uses a 4-layer RNN structure of 300 GRU units. The input corpora of the model use two processing methods. One is a word sequence based on jieba segmentation, and the other is a character sequence based on character splitting [13]. The dictionary size of these two corpus processing methods is also different. The dictionary size for a word sequence is set to 40000, for the character sequence is set to 4000. For word sequence corpora, 40,000 dictionaries accounted for about 94% of word frequency in all corpora; for character sequence corpora, 4,000 dictionaries accounted for about 99% of character frequency in all corpora. In order to speed up the convergence of model training, pre-trained 300-dimensional character vectors and word vectors are used at the first embedding layer. Due to the large loss in the initial training period, all the parameters of the first embedding layer are set to be untrainable, and after a few iterations of the training, the word embedding layer parameters are fine-tuned. At the same time, in order to concentrate much on the global semantic representation in the global encoder, a Seq2Seq model with global encoder is pre-trained, and the parameter values of the encoder part are taken as parameter initialization values of the global encoder of this model, which can speed up the convergence speed and effect of the summary model. Finally, when the model predicts to generate a summary, the beam search algorithm is used to predict summary word and the beam size is set to 10.

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 as a high-quality summary. This article uses an internal evaluation method ROUGE [14] 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 paper uses the ROUGE-1, ROUGE-2, and ROUGE-L to evaluate summary performance of the models.

Table 1 gives the evaluation values in ROUGE of the four models implemented by ourselves on the test set, where En-De represents a basic Encoder-Decoder model based on BiRNN, and En-De-ATT represents an En-De model that adds Bahdanau's attention mechanism, En-De-GAL represents the attention model integrating the global and local semantic information and En-De-GAL++ represents the En-De-GAL model that adds other probability P_a to predict the summary word, in which we can copy the original word from the source text according to the position of word with max predicted probability when the predicted word is in wild symbols. Word represents that experiments were performed on the corpora of the jieba segmentation, and Char represents that experiments were performed on character-separated corpora. From the experimental results, it is seen that the performance of the summary evaluation of the En-De-ATT model is better than that of the En-De model, and En-De-GAL model is superior to the En-De-ATT model. The result shows that the attention mechanism of merging global and local semantic information is helpful to further enhance the effect of generating the abstract, so that the model can capture the global semantics of original short text combine the local semantics to generate better quality summaries. And because En-De-GAL++ model can copy the words about wild symbols from source text, En-De-GAL++ model obtains the best ROUGE evaluation value in all models. Furthermore, the result also shows that the character-level model generates a better summary than that of the word-level model. The reason has two folds. The first is that output dictionary size of the character level model is small, which makes the number of classification in the Softmax function layer of the neural network be less, the parameter amount be less, the training be faster and convergence be easier than that of word-level model. Second, although the output dictionary size in the character-level model is small, the percentage of character frequency in the total corpora is up to 99%, and unknown character in generating summary is less likely to appear. Based on the char-level model without word segmentation, it can further reduce the

errors caused by word segmentation errors. However, the input based on the character level also has drawbacks. Since there is no word segmentation, it is inevitable to lose the opportunity to further optimize the model by integrating more part of speech and syntactic structure information as input features.

TABLE 1. Text summary ROUGE value evaluation results

Model Name	ROUGE-1	ROUGE-2	ROUGE-L
En-De(Word)	0.181	0.086	0.162
En-De(Char)	0.225	0.093	0.198
En-De-ATT(Word)	0.269	0.148	0.256
En-De-ATT(Char)	0.302	0.186	0.281
En-De-GAL(Word)	0.276	0.162	0.261
En-De-GAL(Char)	0.314	0.198	0.293
En-De-GAL++(Word)	0.287	0.174	0.273
En-De-GAL++(Char)	0.316	0.203	0.298

From the case of Table 2, we can see that the summary generation model of this paper is a pure data-driven end-to-end model, but it can generate a summary of the succession, and the generated summary captures the key content of the short text and gives information such as time, person, place, and event. At the same time, the abstract generated by this 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, the model generates abstracts that contain new words. In case 3 and 4, the En-De-GAL++ model can generate a better summary that directly replaces the wild symbols with the proper word in original text than that of En-De-GAL model. It can be seen from case 4 that due to the huge number of words, the UNK (unknown word) appears in the summary generated by En-De-GAL model based on the word segmentation, resulting in a poor model-generating summary. However, in case 4 the summary generated by En-De-GAL++ model can copy the unknown word from source text. Admittedly, sometimes En-De-GAL++ model may copy the error word from source text, which resulting the confuse meaning of short text. What is more, the summary generated by character-level model is better than that of word-level model. The word-level model also has its advantages. Because it is a word unit that better reflects the semantic information of Chinese, in some test cases, a higher quality summary can be generated compared to the char-level model.

TABLE 2. Summary of the model generated case

Case 1 :
Short text: SRC's spokesperson Zhang Xiaojun stated that Shanghai-Hong Kong Stock Connect conducte d a market-wide test on September 13th. This test focuses on system stress test scenarios and failover scen arios to verify the processing performance of all parties. As for the specific opening time of Shanghai-Ho ng Kong Stock Connect, Zhang Xiaojun said that all preparations are in order, but the specific opening ti me has not yet been finalized.
Reference summar: SRC: The specific opening time of Shanghai-Hong Kong Stock Connect has not been determined
En-De-GAL(word): SRC: The specific opening time of Shanghai-Hong Kong Stock Connect has not been determined
En-De-GAL(Char): SRC: The opening time of Shanghai-Hong Kong Stock Connect has not been determi ned
Case 2:
Short text: In June 2014, China's Manufacturing Purchasing Managers Index (PMI) was 51.0%, up 0.2 per centage points from the previous month and rebounded for four consecutive months, indicating that the m anufacturing industry continued its steady growth. However, it is worth noting that the impetus for the inc rease in PMI is not balanced, and the import index and practitioner index are still below the critical point.
Reference summary: China's manufacturing PMI was 51% in June
En-De-GAL(word): China's manufacturing PMI was NUMPERCENT in NUM year NUM month
En-De-GAL(Char): China's manufacturing PMI was NUMPERCENT in NUM month
En-De-GAL++(word): China Manufacturing PMI was 51.0% in June 2014
En-De-GAL++(Char): China's manufacturing PMI was 51.0% in June

From the cases of Table 3, there are some differences between our proposed model and En-De-ATT model. In case 5 and 6, the summary generated by our model has a stronger expression of the central content of the short text. In Cases 5 and 6, the abstract generated by the En-De-ATT model appears duplicate words, but our proposed model can generate abstracts that are straightforward. According to statistics, the number of the summary generated by our model which consists of duplicated words is less than that of En-De-ATT model. And the number of summary generated by our model which consists of UNK is also less than that of En-De-ATT model.

TABLE 3. Comparison of generated cases

Case 4: Short text: The Peugeot 308R high-performance hybrid model debuted at the Shanghai Auto Show was built on the basis of the Peugeot Quartz concept car. Compared with the regular 308 models, the power is more sturdy, the performance is superi or, the design exaggeration, and the two-color body configuration scheme makes it highly recognizable. Referen cesummary: Peugeot Shanghai Auto Show debuts 308R Hybrid En-De-ATT(char): Shanghai Auto Show Shanghai Auto Show debut En-De-GAL(char): Shanghai Motor Show debuts Peugeot NER high-performance hybrid model En-De-GAL++(char): Shanghai Auto Show debuts standard 308R high-performance hybrid model Case 5: Short text: The China Securities Regulatory Commission (SFC) spokesman said on the 17th that the China Securities Regul atory Commission has publicly solicited opinions from the public on the "Regulations on the Trial Implementation of Case I nvestigations Entrusted by the China Securities Regulatory Commission to the Shanghai Shenzhen Stock Exchange." After t

SUMMARY AND OUTLOOK

Based on the encoder-decoder model with attention mechanism, we propose a dual encoder summary generation model and improves the attention mechanism. Our proposed model fuses the multi-channel semantic information of short texts and makes the model able to adaptively combine the information of global and local contexts and some of the summary status information at every time step to generate the next summary word. Experimental results show that performance of the ROUGE [14] in our model is better than that of the classical attention mechanism model, and the summary generated by our model is even more straightforward, and the probability of appearing repeated words in generated summary is lower. For our proposed model, there are many research directions that can be explored in the future. Firstly, for the structure of the encoder and the decoder, a sequence-to-sequence convolution structure can be used instead. Secondly, how to shorten the abstract generation time of the attention mechanism model that fuses global and local semantic information. Thirdly, how to use more Chinese words and syntactic features to optimize the model and combine the advantages of character and word to generate higher-quality summary.

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).

REFERENCES

1. Sanderson M. Book Reviews: Advances in Automatic Text Summarization [J]. Information Retrieval, 2000, 4(1):82-83.
2. Bhatia N, Jaiswal A. Automatic text summarization and it's methods - a review [C]. Cloud System and Big Data Engineering. IEEE, 2016:65-72.
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 Network s[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. Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation [J]. Computer Science, 2015.
12. 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.
13. 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.
14. Flick C. ROUGE: A Package for Automatic Evaluation of summaries [C]. The Workshop on Text Summarization Branches Out. 2004:10.
15. Gu J, Lu Z, Li H, et al. Incorporating Copying Mechanism in Sequence-to-Sequence Learning [J]. 2016:1631-1640.
16. See A, Liu P J, Manning C D. Get to The Point: Summarization with Pointer-Generator Networks [J]. 2017.