Chinese long text summarization using improved sequence-to-sequence lstm

Zanjie Yao¹, Aixiang Chen²* and Han Xie³

School of Statistics and Mathematics of Guangdong University of Finance & Economics, Guangzhou, Guangdong, 510000, China

*Corresponding author’s e-mail: cax413@163.com

Abstract. Text summarization is an important issue in natural language processing. The existing method has the problem of low accuracy when performing long text summarization. In this paper, we use the LSTM to construct the sequence-to-sequence model, and combine the attention mechanism to process automatic Chinese long text summarization. The experimental results indicate that our method can accurately extract key information from long text, generate high-quality summary.

1. Introduction
In the era of information explosion, with the development of network media, people can easily and quickly accept a large number of articles and obtain the required knowledge. Facing with a huge knowledge base, how to efficiently and quickly filter out important information from massive texts, improve learning efficiency, and save readers a lot of time is very important. When people are reading an article, in order to grasp the content of the article in the shortest time, they often need to quickly understand the central idea by reading the summaries. Traditional summaries are mainly generated manually, which requires professionals to have relevant professional knowledge without subjective explanations and comments, and to produce some high-quality summaries concisely and accurately. But it is often a waste of time and effort, and it is not realistic to rely on manual abstraction in massive texts. Therefore, the automatic summarization technology was born [1].

Text automatic summarization is mainly divided into two types: extractive text summarization and generative text summarization [2]. Extractive text summarization based on statistics or rules, extracting abstract sentences in text by calculating the weight of sentences or calculating the similarity among sentences. Common extractive methods such as TextRank algorithm [3-5] and TF-IDF [6] algorithm are the simplest and most widely used methods in the field of text summarization, and often do not need to train the model in advance. Its general idea is to rank the sentences from the original text and extract the sentences with the highest importance as the abstract of the article. The extractive text summarization is a summary of sentences extracted from the text. The connection between the sentence and the sentence is stiff, which is not conducive to the user's understanding and experience, and cannot meet the needs of people. Therefore, the generative text summarization also just appeared. Generative text summarization rely on the computer's own understanding of the text to summarize the articles, which have always been a more complex problem. It often needs to be trained first to let the algorithm remember the key information of the article. Therefore, the requirements on the data set and the computer configuration are much higher, and the training effect is often affected, especially in the case of some longer texts, which often have poor effect.
Deep learning has developed rapidly in recent years and has made good progress in text summarization. The Seq2Seq [7] model has become a common model in various fields of deep learning due to its advantages of accepting input and output of indefinite length [8-9], and it also provides another idea for the field of text summarization. The automatic summarization generation technology based on deep learning also came into being.

In 2015, Rush et al. [10] used RNN to construct a summary extraction model based on the Seq2seq model. They have done text summarization tasks on English data sets such as DUC-2014 corpus and Gigaword corpus, which is also the first attempt by people in the field of generative summarization. However, the domestic research tasks on Chinese text summarization have been in the state of less research. On the one hand, there are relatively few datasets for Chinese text summarization tasks. On the other hand, the deep learning model itself is not effective in processing Chinese text summarization.

In the research of Chinese short text summarization, most tasks use LCSTS corpus [11]. It is a commonly used data set for Chinese short text summarization. The corpus is from Sina Weibo, with short texts of 80 to 140 words and summaries of 10 to 20 words. In 2016, Gu et al. [12] introduced copy technology into neural-network-based Seq2Seq learning, and proposed a CopyNet model with codec structure. In the same year, Ayana [13] adopted the minimum risk training (MRT) strategy to directly optimize the impact of model parameters on evaluation indicators at the sentence level, thereby significantly improving the efficiency of summarization generation. In 2017, Li et al. [14] proposed a new text summarization framework based on deep recursive generative decoder (DRGN). In 2018, Lin [15] proposed a global coding model, which improved the effect on the Chinese short text corpus LCSTS. In 2019, Zhang Kejun et al. [16] improved the model's semantic understanding on the LCSTS corpus by improving the word vector generation technology. The research mentioned above is all based on short text summary research. Their ability to handle input text can only reach about 100 words.

However, the long text summarization task is very difficult and the existing methods for Chinese long text summarization have low accuracy. On the one hand, The kind of processing has long-distance dependence problems, and the super-long text task requires the encoder's ability to encode long texts and understand semantics. On the other hand, in order to ensure the comprehensiveness of the word list in the word vector, the size of the text to be trained cannot be too small, otherwise the effect of the trained neural network model will be reduced due to too many “UNK” (unknown) words in the text.

Actually, most of the time we need to deal with is super-long text. Therefore, in this work, we uses the Chinese news corpus provided by Sogou Lab to train the word vector model.

2. Models

2.1. Encoder-decoder framework
The essence of text summarization technology is to input a piece of text and then map it through a function to get a piece of text to form a summary. This is essentially a sequence-to-sequence model, called a Sequence-to-sequence structure, or Seq2seq structure for short. The structure was first proposed by Sutskever [7].

The Encoder-decoder framework is shown in Figure 1. Define the sequence representation of the input text as \( \{x_1, x_2, \ldots, x_n\} \) (\( n \) denotes an input length). When the sequence is input, the encoder encodes the sequence as an intermediate vector \( C \), as an intermediate semantic encoding \( C \), and then inputs \( C \) to the decoder part. The decoder then uses the existing historical information and semantic coding \( C \) to make a prediction value \( y_i \) at the current time \( i \). The predicted values at all times as a summary of the text. Define the sequence representation of the predicted output as \( \{y_1, y_2, \ldots, y_m\} \) \((m < n)\), and define the reference summary at time \( i \) as \( \hat{y}_i \). Our goal is to make \( y_i \) and \( \hat{y}_i \) the same.
2.2. LSTM neural unit

The traditional fully-connected BP network cannot perform sequence modeling tasks because the nodes between the hidden layer and the hidden layer are connectionless. The recurrent neural network RNN, due to its structural design, links the past information with the current moment, making it very suitable for dealing with many problems of sequence model tasks, and has gained tremendous success in the fields of natural language processing and speech recognition. However, due to the problem of gradient disappearance or gradient explosion in the deep RNN model, it is difficult to transfer the error during back propagation. In 1997, S Hochreiter et al. [17] proposed LSTM to overcome this problem.

The basic model framework of this paper is shown in Figure 2. The encoder and decoder select LSTM neural unit as the basic unit.

2.3. Attention mechanism

The biggest advantage of Encoder-decoder is that the input sequence length and output sequence length are variable. However, there is a problem that no matter how short or long the input sequence is, its processing method is uniformly encoded as an intermediate semantic vector $C$. In the case of a long input length, the intermediate semantic vector $C$ cannot encode too much information, and it is easy to cause the loss of semantic information.

Therefore, this paper uses the attention mechanism, the most commonly used by natural language processing. It was first used by Bahdanau [19] in the task of machine translation. When translating a word, the attention mechanism embodies an alignment mechanism between the input word and the translated word. In the text summarization generation, there is also an alignment mechanism for each word between the target output word $y_i$ and the input $x_i$. There are usually many words in the input sequence, but not every word contains the key to determine the prediction result, so it is required to focus on one of the input words in the input text.

After the attention mechanism is introduced into the encoder-decoder framework, when each word is generated, the semantic vector $C$ that represents the original sequence information with the original value unchanged will be replaced and become a semantic vector that changes according to the currently generated word.
The Seq2seq model with the attention mechanism is shown in Figure 3. When using a text summarization generation model based on the Seq2Seq with attention mechanism for text summarization generation, the specific process is as follows:

(1) Vectorize the text first and enter it into the model;
(2) Get distributed representation of articles using LSTM;
(3) Use the attention mechanism to get more accurate expressions;
(4) Input the distributed expression of the article into the LSTM unit to predict the distributed expression of the summary;
(5) Transform the distributed representation of the summary into text form to get the summary.

![Figure 3. Seq2seq model with the attention mechanism.](image)

Among them, the calculation formula of $C_i$ is like equation (1) and equation (2). When predicting the output $y_i$ at time $i$, the Attention structure will match each input $x_i$ to $x_n$ with the output at the current time $i$, and then automatically calculate the distribution value $a_\beta$ of each attention probability distribution. The calculation formula of $a_\beta$ is as in equation (2). $S_j$ represents the activation value of the hidden neuron in the input part.

$$C_i = \sum_{j=1}^{n} a_\beta S_j$$  \hspace{1cm} (1)

$$a_\beta = \frac{\exp(e_\beta)}{\sum_{i=1}^{n} \exp(e_\beta)}$$  \hspace{1cm} (2)

### 3. Objective function

The objective function of this article is a negative logarithm Likelihood function $J(\theta)$. Given the input text sequence \{ $x_i, x_2, \ldots, x_n$ \}. To maximize the probability of the output summary, that is to make $J(\theta)$ the smallest. $\theta$ represents the parameters of the model, $N$ represents the training set. The formula is shown in equation (3)

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{m} \log P(y^{(i)} \mid x, y^{(1)}, \ldots, y^{(i-1)})$$  \hspace{1cm} (3)

### 4. Experiments and results

#### 4.1. Data description

The data in this article uses the Sogou News Corpus, which is a total of 100263 data from June to July 2012. Each data consists of a news and a corresponding summary. All of these data contain the contents of various fields, including domestic and foreign entertainment, culture, sports, Internet, business, public welfare and other news, including a wide range.
4.2. Data preprocessing
To preprocess the data, the first step is to delete text less than 100 in length. We get 83309 samples. The data set is divided into 8: 1: 1, corresponding to the training set, test set, and validation set. Then perform sentence and word segmentation operations on all data samples, count the word frequencies of all training sets, and rank them from high to low. The resulting vocabulary contains 135487 words.

The minimum number of words set to 3, the number of vocabulary words that appear less than 3 is deleted and replaced with a tag <UNK>. And remove special characters, punctuation, remove stop words and so on, all the rest as the default vocabulary. Additionally added <EOS> and <GO> tag. As shown in Figure 1, the beginning of each text is replaced with <GO> and the end is replaced with <EOS>. Therefore, the actual generated vocabulary has 50,000 words. The line number of the vocabulary will be used as an index. When a certain word is input, the word will read the index and be converted into a One-hot vector as input.

4.3. Experimental parameter settings
This experiment was trained on Ubuntu 16.04 system. The whole training process took one week. When training the Seq2Seq with Attention model, the parameter settings are as follows: Encoder and decoder are two layers. Dropout is set to 0.1. Select the Adam algorithm as the optimization function. The initial learning rate is set to 0.1, and the learning rate decline frequency is 0.5.

4.4. Rouge evaluation indicators
The ROUGE [20] method is an evaluation index proposed by Lin and Hovy in 2004. It compares the reference summary with the prediction summary to obtain a corresponding score for evaluating the prediction summary. This method has become one of the common standards for evaluating the text summarization extraction technology.

\( N \) in ROUGE-\( N \) refers to the model of \( N \)-words. Generally, \( N \in [1, 4] \), and its formula is as shown in equation (4).

\[
\text{ROUGE-}N = \sum \sum \frac{\text{Count}_{\text{match}}(\text{gram}_n)}{\text{Count}(\text{gram}_n)}
\]

(4)

Among them, the numerator part is the number of n-grams appearing at the same time as the reference summary and the prediction summary (The number of repeated units), and the denominator part is the number of n-grams of the reference summary. The Rouge score evaluates the similarity between the prediction summary and the reference summary. The higher the similarity, the better the prediction quality. [21]. Since the evaluation index is for English, for Chinese text summarization, before using the evaluation index, first perform word segmentation on the Chinese summary, and then compare the similarity between the prediction summary and the reference abstract.

4.5. Experimental results
The Seq2Seq with Attention module is trained, and the set iteration is 90 epochs. The curve of the loss function of the model training is shown in Figure 4.

![Loss Function graph.](image)
50 samples were randomly selected in the test set for testing, and the prediction summary output from the three algorithms was compared with the reference summary. According to Rouge's formula, the Rouge evaluation value was obtained. The obtained results are shown in Table 1.

| Model     | Rouge-1 | Rouge-2 | Rouge-L |
|-----------|---------|---------|---------|
| LSTM      | 0.8325  | 0.6896  | 0.8283  |
| Textrank  | 0.1969  | 0.0946  | 0.1610  |
| TF-idf    | 0.1640  | 0.0761  | 0.1383  |

From the experimental results in Table 1, it can be seen that the experimental evaluation value of the model based on Seq2Seq lstm in the text summarization task of super-long text is better than the other two extractive text summarization models. Compared with the Textrank algorithm, the Rouge-1 value is increased by about 0.64, the Rouge-2 value is increased by about 0.59, and the Rouge-L value is increased by about 0.66. Compared with the TF-IDF algorithm, the Rouge-1 value is increased by about 0.67, the Rouge-2 value is increased by about 0.61, and the Rouge-L value is increased by about 0.69. According to the data in Table 1, it is proved that the generative text summarization model is better than the extractive text summarization model after a long training. Therefore, the Seq2Seq lstm model designed in this paper has a better understanding of the semantics of super-long article content, and the article summary generated by this model is more accurate and more in line with the results of article summary.

4.6. Qualitative Analysis

Next, we will show one of these test cases. The following table shows the results of three algorithms for abstracting the same article. The total length of the article is 1000 words, which is a very long text.

| Reference summary                  | One Foundation Rescue Alliance holds flood rescue exercise to teach Self-help knowledge to Orphans |
|------------------------------------|--------------------------------------------------------------------------------------------------|
| LSTM                               | One Foundation prepares for flood relief and conducts public welfare rescue exercises              |
| Textrank                           | In order to do a good job in flood prevention and preparedness, and to launch rescue operations as soon as the flood arrives, the One Foundation Rescue Alliance and local relief organizations have developed a series of nationwide flood rescue exercises. The drill locations include Hubei, Guangxi, Yunnan, Sichuan and other places |
| TF-IDF                             | The rescue team strictly complied with the "Standard Practice Plan for Flood Rescue and Drilling Operations" previously established, and established a level 5 and 3 order command system to conduct practical relief drills, such as temporary communication relay erection, mass transfer, mass rescue, and charge. Boat driving, diving operations and the use of new types of emergency rescue equipment, through this exercise training, team members can master the disaster relief skills more proficiently. |

By comparing the three kinds of prediction summaries with reference summary, we can find that the first two algorithms, the Seq2Seq lstm model and Textrank are all good. Both can extract the subject words "One Foundation" and "Rescue Exercises", which mainly indicate the central theme Semantic features. But in comparison, the previous Seq2Seq lstm model generalizes better, and the generated summary is concise and highly generalized. The latter's Textrank algorithm has a more detailed language.
The third algorithm, TF-IDF, performs poorly, and the extracted sentences do not match the center of the article.

It can be concluded from this that the extractive text summarization is very dependent on the presence of the central sentence in the article. If the article does not have a central sentence, a specific algorithm is needed to understand the meaning, and the trained Seq2Seq lstm model has a higher ability to understand the article.

4.7. Conclusion
In this paper, by researching generative text summarization, we designed a Seq2Seq lstm model for the problem of summary generation of Chinese long texts. LSTMs are used in both the encoder and decoder to make full use of contextual information to understand semantic features, and are trained on the Sogou news dataset. The trained model is compared with two extractive text summarization algorithms, Textrank and TF-IDF, and analyzed. We finally concluded: In the super-long text summary generation, the trained Seq2Seq lstm model has a higher Rouge index value, indicating that it has a higher degree of similarity to the real summary, and the prediction summary is more accurate and realistic.

Acknowledgments
This article is supported by the National Social Science Project of China: Research on Deep Learning Technology from a Statistical Perspective (19FTJB003).

References
[1] Luhn, H.P. (1958) The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2):159-165.
[2] Gambhir,M., Gupta,V. (2017)Recent automatic text summarization techniques: A survey. Artificial Intelligence Review, 47(1):1-66.
[3] Mihalcea,R., Tarau,P. (2004)TextRank: Bringing Order into Texts. Emnlp, 404-411.
[4] Li,N. N.,Liu,P. Y.,Liu,W. F. (2019) TextRank-based Automatic Digest Optimization Algorithm. Application Research of Computers, 36(04):91-96
[5] L. Page, S. Brin, R. Motwani and T. Winograd.(1998) The PageRank Citation Ranking: Bring Order to the Web. Technical report, Stanford.
[6] Zhang,Y. F.,Peng,S. M.,Lv,J.(2006) Improvement and Application of TFIDF Method Based on Text Classification. Computer Engineering. 32(19):76-78.
[7] Sutskever,I., Vinyals,O., Le,Q. V. (2014)Sequence to Sequence Learning with Neural Networks. Advances in neural information processing systems.
[8] Bahdanau,D., Cho,K., Bengio,Y. (2014)Neural Machine Translation by Jointly Learning to Align and Translate. Computer Science.
[9] Chorowski,J.K., Bahdanau,D., Serdyuk,D., et al. (2015)Attention-based models for speech recognition. In: NIPS.pp.577-585.
[10] Rush,A.M., Chopra,S., Weston,J.(2015) A neural attention model for abstractive sentence summarization. In:Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 379–389.
[11] Hu, B.T.,Chen,Q.C.,Zhu,F.Z.(2018).LCSTS:A large scale Chinese short text summarization dataset.https://arxiv.org/abs/1506.05865.
[12] Gu,J., Lu,Z. , Li,H., et al. (2016)Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In: ACL 2016.
[13] Ayana, Shen,S., Zhao,Y. , et al. (2016)Neural Headline Generation with Sentence-wise Optimization. In: arXiv:1604.01904.
[14] Li,P., Lam,W., Bing,L., et al. (2017)Deep Recurrent Generative Decoder for Abstractive Text Summarization. In: EMNLP 2017.
[15] Lin, J., Sun, X., Ma, S., et al. (2018) Global Encoding for Abstractive Summarization. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 163-169.

[16] Zhang, K.J., Li, W.N., Qian, R., et al. (2019) Automatic text summarization scheme based on deep learning. Computer Application, 39(02):7-11.

[17] Hochreiter, Sepp, Schmidhuber, Jürgen. (1997) Long Short-Term Memory. Neural Computation, 9(8):1735-1780.

[18] Chen, A.X. (2019) Deep Learning. Tsinghua university press, Beijing.

[19] Cho, K., Van Merrienboer, B., Bahdanau, D. (2014) On the properties of neural machine translation: Encoder-decoder approaches. In: Proceedings of the SSST @EMNLP, pp. 103-111.

[20] Lin, C.Y. (2004) ROUGE: A Package for Automatic Evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), pp. 74-81.

[21] Lin, J.Y., Sun, X., Ma, S., et al. (2018) Global Encoding for Abstractive Summarization. In: The 56th Annual Meeting of the Association for Computational Linguistics 2018. Australia, pp. 163–169.