Text Summarization Model of Combining Global Gated Unit and Copy Mechanism

Shuxia Ren
School of Computer Science and Technology
Tianjin Polytechnic University
Tianjin, China
t_rsx@126.com

Kaijie Guo
School of Computer Science and Technology
Tianjin Polytechnic University
Tianjin, China
1831125493@stu.tjpu.edu.cn

Abstract—**Text summarization is a common task in NLP. Automatic text summarization aims to transform lengthy documents into shortened versions. Recently, the neural networks based on seq2seq with attention are good at generating summarization. However, the accuracy of the summarization too difficult are to guarantee. In addition, the Out-of-Vocabulary (OOV) problem is also an important factor affecting the quality of the generated summary. To solve these problems, we hybrid the advantages of the extractive and abstractive summarization systems to propose text summarization model of combining global gated unit and copy mechanism (GGUC). The experiment results demonstrate that the performance of the model is better than the other text summary system on LCSTS datasets.**

Keywords-text summarization; global gated unit; copy mechanism ; GGUC

I. Introduction

The rapid development of the Internet is accompanied by a large amount of text data produced every day. Therefore, users must spend a lot of time reading articles to get to know what they are interested in. The first paper on automatic summarization published by Luhn of IBM in 1958 opened the curtain of research in this field [1]. However, automatic summarization provides users with a quick way to understand the content of the source text. It has many application scenarios, such as automatic report generation and news headlines generation. Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information and without losing the overall meaning.

There are two main types of summarization. First, extractive summarization systems form summaries by selecting and copying text snippets from the original text [2]. For example, Mihalcea proposed the textrank method in 2004 [3]. Ko proposed an automatic summarization method that combines statistical information with contextual information [4]. However, it has some inherent problems, such as its inability to ensure the coherence and relevance of summaries. Second, abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text [5]. Compared to the former, the abstractive summarization can make more human-readable and grammatically correct summaries.

Recently, neural network models based on the sequence-to-sequence (seq2seq) model have become popular in summarization. In particularly, the attention mechanism based seq2seq models prevails in abstractive summarization task. For example, Rush employed an attentional feed-forward network for abstractive summarization on the Gigaword and DUC 2004 datasets [6]. Nallapati applied Large Vocabulary Trick (LVT) to solve the computational bottleneck of decoder in generating words on the CNN/Daily Mail and DUC 2002 datasets [7]. The above experiments are all aimed at English datasets. For Chinese corpus, Hu proposed a new Chinese dataset LCSTS and used GRU to build the seq2seq model for text summarization [8]. Wang proposed a reinforced topic-aware convolutional seq2seq model on LCSTS dataset [9].

However, the traditional attention based on seq2seq models has some challenges in abstractive summarization. In general, the summary is shorter than the original text in summarization task. So there is no obvious alignment between the original text and the summary. Meanwhile, although in each time step of the decoder, most of the existing models have used the attention mechanism to perform weighted summation operation on the input sequence. But in the encoding stage, the traditional encoder is calculating the vector representation of each word or the hidden layer state by considering only some words before (or after) the word, rather than the complete global information. These models often generate some repetitive and incoherent phrases. In order to solve this problem, See proposed the coverage mechanism to avoid this problem of traditional seq2seq models [10]. Lin proposed a model of global encoding for abstractive summarization [11].

Moreover, the Out-of-Vocabulary (OOV) problem has always been a stumbling block to the development of text summarization. Vinyals proposed the pointer networks could to copy words directly from the source text. So this method can improve the accuracy of summarization and process OOV problem [12]. Meanwhile, Gu explored a “copying mechanism” to effectively solve the problem [13].

In this paper, our main contributions are as follows: (a) we propose a text summarization model of combining global gated unit and copy mechanism (b) we simplified the traditional copy mechanism to solve the OOV problem.

978-1-7281-0945-9/19/$31.00©2019 IEEE

II. Our Model

In this paper, we propose a new model based on the seq2seq with attention. For the encoder, we choose the classic bi-directional LSTM structure and use Global Gated Unit (GGU) to filter the encoder information. Moreover, we introduce a new copy mechanism in the decoder. In this section, we introduce the details of the encoder, GGU and decoder respectively. Figure 1 shows the overall architecture of the GGUC model.

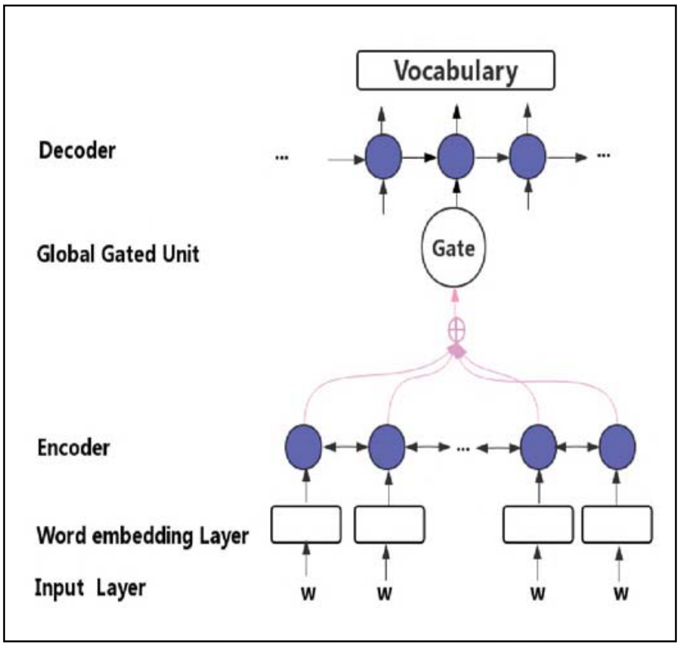


Figure 1. The overall diagram of GGUC

A. Encoder

First, we implement a bidirectional LSTM encoder consisting of a forward and a backward network. The input of the encoder is the word embedding of each word from source text and the hidden state at the previous time step. The forward hidden state $\overrightarrow{h_t}$ and the backward hidden state $\overleftarrow{h_t}$ are stitched together to represent the hidden layer state h_t of encoder. For each time step, we can get the output of the encoder, as shown in equation (1):

$$h_t=[\overrightarrow{h_t};\ \overleftarrow{h_t}] \tag{1}$$

B. Global Gated Unit

After obtaining the hidden-state of the encoder, we use the Global Gated Unit to filter the encoder information. The unit uses CNN network to extract key information to select the output information of encoder. This method can achieve the purpose of removing repetitive information. Here, we use the scaled dot-product attention proposed by Vaswani to calculate the relationship between information at each timestep and global information [14].

In general, scaled dot-product attention mechanism is essentially an addressing process. Given a task-related query vector q, attention value is calculated by attaching the attention distribution with key to Value. The mechanism simplifies the attention operation by calculating the following function, where the queries, keys, values and outputs both are vectors. In this paper, the query is the output of each timestep of the encoder, and the key and value are the global information obtained after self-matching. Then we use the softmax function to get the attention distribution, and it can be described by the following equations:

$$g=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

$$Q=W_Q h_t, K=W_K h_t, V=W_V h_t \tag{3}$$

$$h_t'=h_t*\sigma(g) \tag{4}$$

where W_Q, W_K, W_V are learnable parameter matrix, $\boldsymbol{\sigma}$ refers to the sigmoid function.

C. Decoder

Copy mechanism use attention to enable a seq2seq system to easily copy words and phrases from the input to the output. Allowing both copying and generating gives us a hybrid extractive/abstractive approach. Therefore, we have added a new copy mechanism to the decoder.

We use the hidden-state of the CGU output as the input of the decoder. There are two modes of generating and copying when generating words, so the model is a probabilistic model combining two modes. In addition, we use a switch to determine how the predicted words are generated at the timestep t is generated. When the switch is turned on, the predictive word v_t is derived from the corpus and is generated by the generation mode. On the contrary, the predicted word is derived from the vocabulary of the input text when the switch is turned off, as shown in equation (5):

$$p(t_{s=1})=\sigma(W_h h_t'+W_s s_t+W_c c_t) \tag{5}$$

where the $p(t_{s=1})$ is a value indicating whether the predicted word is derived from the generation mode at the time step t, the W_h, W_s, W_c are learnable parameter matrix, s_t is the hidden-state of the decoder and c_t is the sequence of the decoder output before time t. The generated words y_t is composed of two modes of generation and copying, as shown in equation (6):

$$p(y_t|X,y_{t-1})=p(t_s=1)p_g(y_t|X,y_{t-1})+ \\ (1-p(t_s=1))p_c(y_t|X,y_{t-1}) \tag{6}$$

Generate mode is essentially a normal decoder output. It can be obtained by the softmax function of the Decoder stage, as shown in equation (7):

$$p_g(y_t|X,y_{t-1})=\mathrm{softmax}(W_t[s_t,h_t']) \tag{7}$$

where X is the input sequence, y_{t-1} is the output at time t-1, s_t is the hidden- state of the decoder at time t.

Moreover, we implement its attention for the decoder at each time. It can be described by the following equations:

$$e^t=V^t[s_t^T h_1',\ldots,s_t^T h_N'] \tag{8}$$

$$a^t=\mathrm{softmax}(e^t) \tag{9}$$

The predicted word probability of the copy mode is obtained by equation (10):

$$p_c(y_t|X,y_{t-1})=\text{argmax}(a_t) \qquad (10)$$

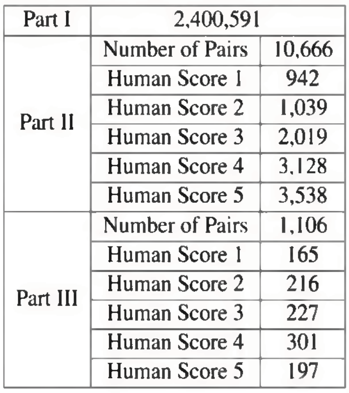
III. EXPERIMENT

In this section, we introduce the dataset, experiment settings and baseline models.

A. Dataset

Hu created a dataset of Weibo summaries posted by media organizations [8]. Short texts are about 100 characters, summaries are about 20 characters. We use the large-scale Chinese short text summarization (LCSTS) dataset. The dataset consists of three parts shown as Table II.

TABLE I. LCSTS DATA STATISTICS



Part I	2,400,591	
Part II	Number of Pairs	10,666
	Human Score 1	942
	Human Score 2	1,039
	Human Score 3	2,019
	Human Score 4	3,128
	Human Score 5	3,538
Part III	Number of Pairs	1,106
	Human Score 1	165
	Human Score 2	216
	Human Score 3	227
	Human Score 4	301
	Human Score 5	197

B. Exeperiment Settings

TABLE II. HYPERPARAMETERS SETTINGS

characters embedding dimension	300
vocabulary size	5000
optimizer	Adam
LSTM hidden unit dimension	512
Batch size	64

We evaluate our model with the ROUGE metric. ROUGE is an automatic summarization evaluation method proposed by Lin [15]. The main idea is that multiple experts generate manual abstracts to form a standard summary set. The quality of the summarization is evaluated by comparing the number of overlaps between the automatically and manually generated summaries. ROUGE scores are reported separately for each n-gram. The most commonly reported F1 scores for ROUGE-1, ROUGE-2 and ROUGE-L.

For the experiment, we use PyTorch deep learning framework to implement our models on an NVIDIA GPU. In order to alleviate the risk of large vocabulary, we split the Chinese sentences into characters. The hyperparameters settings are shown in Table II:

C. Baseline Models

In order to evaluate the performance of our proposed model (Global Gated Unit and Copy Mechanism, referred to as (GGUC) in the text summarization task. We compare the GGUC with the experimental results of the following models on the LCSTS dataset.

RNN: In these models, Hu use the sequence-to-sequence with GRU encoder architecture [8].

RNN-context: The difference between the RNN is that the RNN-context uses the attention mechanism and the other does not [8].

CopyNet: Gu explore the copy mechanism in seq2seq with attention models [13]. This method is still an encoder-decoder model, but the decoder part adds a copy mode to the traditional Seq2Seq model.

CGU： Lin introduce a convolution gated unit that filters information from source text [11]. The unit is used to perform global coding to improve the representation of source information.

DRGD: Li propose a new framework to learn the latent structure information implied in the target summary to improve the quality of the summarization [16].

D. Experiment Result

First, we follow the existing models and adopt the ROUGE metric for evaluation. Table 3 shows the results of our model and several baselines. The GGUC model we proposed is very competitive in terms of ROUGE and achieves better performance than almost all the existing baseline systems.

TABLE III. ROUGE SCORES ON LCSTS

Methods	Rouge Scores		
	R-1	***R-2***	***R-L***
RNN	21.5	8.9	18.6
RNN-context	29.9	17.4	27.2
CopyNet	34.4	21.6	31.3
CGU	39.4	26.9	36.5
DRGD	37.0	24.2	34.2
GGUC(Our impl.)	**39.8**	**28.2**	**37.6**

Compared to other baseline models, the GGUC model has increased by at least 0.4, 1.3 and 1.1 percentage points on Rouge-1, Rouge-2 and Rouge-L respectively. It indicating that it is very useful to use global gated unit with copy mechanism in text summarization task.

IV. RELATED WORK

Pre-neural summarization systems were mostly extractive. Litvak used the Hypertext-Induced Topic Search (HITS) algorithm to extract keywords [17]. Goyal uses the Bernoulli model to generate the subject terms and builds a graph model with them as nodes [18]. Neto combines TextTiling and TF-IDF algorithms to extract text summarization [19]. In 2014, the seq2seq model was first used in machine translation task and achieved great success [20].

From 2015 to the present, most researchers use neural networks to solve text summarization. Rush publish the first seq2seq summarization paper [6]. Chung systematically compares the two seq2seq components of LSTM and GRU [21]. Takase proposed attention-based AMR seq2seq model to improve headline generation benchmarks [22]. But the seq2seq models are bad at copying over details (like rare words) correctly. Fortunately, the copy mechanism is very suitable for solving this problem [13, 23, 24]. Gulcehre solved the unknown words problem in NLP tasks by using pointers on the input sequence [25].

V. CONCLUTION

In this work we presented a text summarization model with hybrid global gated unit and copy mechanism. The model uses the classic seq2seq with attention architecture. Experiments on LCSTS corpus show that GGUC has a significant improvement in ROUGE evaluation compared with the current baseline models.

In the future, the accuracy of the text summarization task needs to be further improved. In addition, our model only tests short text dataset. Therefore, studying the text summary method of long texts will be an important direction for our future efforts.

ACKNOWLEDGMENT

We would like to appreciate Qingcai Chen for providing the datasets and for the support.

REFERENCES

[1] Jones, Karen Spärck. "Automatic summarising: The state of the art." Information Processing & Management 43.6 (2007): 1449-1481.

[2] Paulus, Romain, Caiming Xiong, and Richard Socher. "A deep reinforced model for abstractive summarization." arXiv preprint arXiv:1705.04304 (2017).

[3] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. (2004).

[4] Ko, Youngjoong, and Jungyun Seo. "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization." Pattern Recognition Letters 29.9 (2008): 1366-1371.

[5] Yao, Jin-ge, Xiaojun Wan, and Jianguo Xiao. "Recent advances in document summarization." Knowledge and Information Systems 53.2 (2017): 297-336.

[6] Rush, A. M., Harvard, S. E. A. S., Chopra, S., and Weston, J. (2017). A Neural Attention Model for Sentence Summarization. In ACLWeb. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

[7] Nallapati, Ramesh, Bing Xiang, and Bowen Zhou. "Sequence-to-sequence rnns for text summarization." (2016).

[8] Hu, Baotian, Qingcai Chen, and Fangze Zhu. "Lcsts: A large scale chinese short text summarization dataset." arXiv preprint arXiv:1506.05865 (2015).

[9] Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., and Du, Q. "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization." arXiv preprint arXiv:1805.03616 (2018).

[10] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368 (2017).

[11] Lin, J., Sun, X., Ma, S., and Su, Q. "Global encoding for abstractive summarization." arXiv preprint arXiv:1805.03989 (2018).

[12] Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." Advances in Neural Information Processing Systems. (2015).

[13] Gu, J., Lu, Z., Li, H., and Li, V. O. "Incorporating copying mechanism in sequence-to-sequence learning." arXiv preprint arXiv:1603.06393 (2016).

[14] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. (2017).

[15] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. (2004).

[16] Li, P., Lam, W., Bing, L., and Wang, Z. "Deep recurrent generative decoder for abstractive text summarization." arXiv preprint arXiv:1708.00625 (2017).

[17] Litvak, Marina, and Mark Last. "Graph-based keyword extraction for single-document summarization." Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization. Association for Computational Linguistics, (2008).

[18] Goyal, Pawan, Laxmidhar Behera, and Thomas Martin McGinnity. "A context-based word indexing model for document summarization." IEEE Transactions on Knowledge and Data Engineering 25.8 (2012): 1693-1705.

[19] Neto, J. L., Santos, A. D., Kaestner, C. A., and Freitas, A. A. "Generating text summaries through the relative importance of topics." Advances in Artificial Intelligence. Springer, Berlin, Heidelberg, (2000). 300-309.

[20] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[21] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[22] Takase, S., Suzuki, J., Okazaki, N., Hirao, T., and Nagata, M. "Neural headline generation on abstract meaning representation." Proceedings of the 2016 conference on empirical methods in natural language processing. (2016).

[23] Miao, Yishu, and Phil Blunsom. "Language as a latent variable: Discrete generative models for sentence compression." arXiv preprint arXiv:1609.07317 (2016).

[24] Nallapati, R., Zhou, B., Gulcehre, C., and Xiang. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

[25] Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., and Bengio, Y. "Pointing the unknown words." arXiv preprint arXiv:1603.08148 (2016).