ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

A Sentence Summarizer using Recurrent Neural Network and Attention-Based Encoder

Takashi Kuremoto*, Takuji Tsuruda, Shingo Mabu and Masanao Obayashi

Dept. Information Science and Engineering, Yamaguchi Univ., Tokiwadai 2-16-1, Ube, Yamaguchi, Japan

*Corresponding author

***Abstract*—For automatically summarizing sentences of nature languages, some cutting-age methods have been proposed since a decade ago. In this paper, an advanced model of abstractive sentence summarization is proposed by composing a recurrent neural network (RNN) and an attention-based encoder. The proposed model is an improvement version of Rush-Chopra-Weston's neural attention model, and main differences between the proposed model and the conventional one is that: 1) the novel model utilizes two RNNs instead of the feed-forward neural networks; 2) the length of summarized sentence (the output of these models) is variable (which is fixed in the conventional case). Experiments showed the effectiveness of the proposed sentence summarizer and these results suggest that it is possible to abstract long articles into shorten words.**

Keywords-abstractive summarization; recurrent neural network; auto-encoder;nature language understanding; artificial intelligence

I. INTRODUCTION

Natural language understanding is a dreamful challenge of artificial intelligence (AI). In the past decades, studies on machine translation [1]-[4], abstractive summarization [5] [6], and article generation have been proposed and gathered a lot of attention of researchers in AI filed and ordinary users.

In the present big data era, articles, books, contexts on the web sites written in the different natural languages are expected to be compressed, abstracted, or summarized. In [6], Rush, Chopra and Weston proposed a cutting-age neural attention model for abstractive sentence summarization utilizing 1) Bengio et al.'s feed-forward neural network language model (NNLM) [1] instead of conventional stochastic machine translation-inspired summarization methods; 2) Bahdanau el al.'s attention-based summarization (ABS) [5] as an encoder to compress the input words. However, both the output of ABS and the output of NNLM are fixed in the certain length of sentence predesigned. Moreover, Chopra, Auli and Rush proposed a recurrent attentive summarizer (RAS) recently [7], improving the former neural attention models by adopting convolutional network to encode input words, and a recurrent neural network for generation.

In this paper, we propose to adopt recurrent neural networks (RNNs) in to Rush-Chopra-Weston's neural attention model instead of NNLM, and remove the constraint of summarized sentences with fixed length. RNNs are utilized in the attention part and attention-based encoder in the proposed model. The training method of model utilizes minimizing negative log-likelihood (*NLL*) by mini-batch stochastic gradient descent. Experiment using articles from "*The Wall Street Journal*" (https://www.wsj.com/) showed the higher abstractive summarization ability of the proposed model comparing with conventional neural attention model.

II. MODEL

A. Definition of Problem

To define the problem of abstractive summarization, we can consider a sample (See Figure I) as follows:

Input sentence: "*Russian Defense Minister Ivanov called Sunday for the creation of a joint front for combating global terrorism*".

Output summarization: "*Russia calls for joint front against terrorism*".

Now let the input sentence $\mathbf{x}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_M)$ be composed by M words, $\mathbf{x}_i \in \{0, 1\}^V$, V is the size of a fixed vocabulary space *V*. The output summarization $\mathbf{y}$ $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_N)$, $\mathbf{y}_i \in \{0, 1\}^V$, $N<M$, is given by an abstractive system $s(\mathbf{x}, \mathbf{y})=\log p(\mathbf{y}|\mathbf{x}; \theta)$, where θ indicates parameters of the system.

To find $\mathbf{y}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_N)$, we can set a window of size *C* for $\mathbf{y}_j$, i.e., $\mathbf{y}_C = \mathbf{y}_{[j-C+1,\ldots,j]}$, and the conditional log-probability of the summarizing system can be given by following:

$$s(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \approx \sum_{j=0}^{N-1} \log p(\mathbf{y}_{j+1} \mid \mathbf{x}, \mathbf{y}_C; \boldsymbol{\theta}) \quad (1)$$

B. A Neural Summarizer

A neural summarizer is shown in Figure 1. Recurrent neural networks (RNNs) are used in the encoder which serves as an indicator, i.e., attention, and in the summary generation. The total model of this neural summarizer is an improved version of Rush-Chopra-Weston's neural attention model, where the improvement is the adoption of RNNs.

For generation of the summary, we have:

$$p(\mathbf{y}_{j+1} \mid \mathbf{x}, \mathbf{y}_c; \boldsymbol{\theta}) \propto \exp(\mathbf{V}\mathbf{h}_t + \mathbf{W}\mathbf{p}^{\mathbf{T}}\bar{\mathbf{x}}) \tag{2}$$

$$\mathbf{h}_t = \tanh(\mathbf{U}\mathbf{y}_c + \mathbf{E}\mathbf{h}_{t-1}) \tag{3}$$

$$\mathbf{p} \propto \exp(\tilde{\mathbf{x}}\mathbf{P}\mathbf{h}_t^{enc}) \tag{4}$$

$$\mathbf{h}_t^{enc} = \tanh(\mathbf{G}\mathbf{y}_c + \mathbf{K}\mathbf{h}_{t-1}^{enc}) \tag{5}$$

$$\tilde{\mathbf{x}} = [\mathbf{F}\mathbf{x}_1 ..., \mathbf{F}\mathbf{x}_M] \tag{6}$$

$$\forall i, \bar{\mathbf{x}}_i = \sum_{i-Q}^{i+Q} \tilde{\mathbf{x}}_i / Q \tag{7}$$

where parameter Q is the size of a smoothing window.

And $\boldsymbol{\theta} \equiv (\mathbf{F}, \mathbf{G}, \mathbf{P}, \mathbf{U}, \mathbf{V}, \mathbf{W})$, and $\mathbf{F}$, $\mathbf{G}$, $\mathbf{U} \in R^{DxV}$ are matrixes for word embeddings, D is the size of the word embeddings, $\mathbf{P}$, $\mathbf{V}$, $\mathbf{W}$ are weight matrixes.

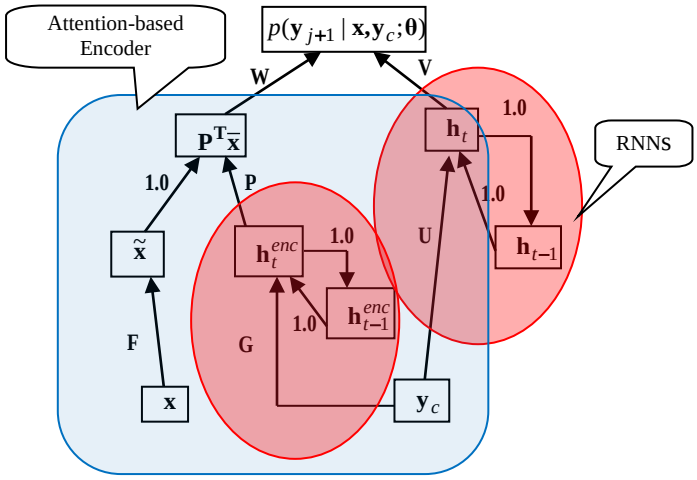

FIGURE I. A NEURAL SUMMARIZER USING RECURRENT NEURAL NETWORKS (RNNS) AND ATTENTION-BASED MODEL.

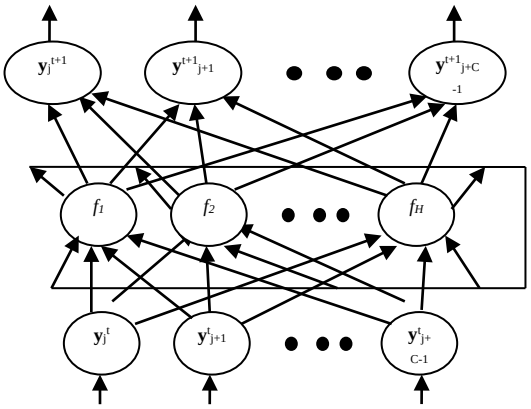

FIGURE II. A STRUCTURE OF RECURRENT NEURAL NETWORKS.

The originality of this study is the two RNNs are adopted into the attention part and the attention-based encoder part of the system as shown in Figure I. In Figure I, there are two recurrent neural networks (RNNs), which are used as summarizers of temporal output of the whole model. RNNs are attentive elements for a part of encoder, and a part of model. The structure of RNN is shown in Figure II. Different from the general recurrent neural networks, RNNs used as attentive summarizers in this model, the number of units in the input layer and the output layer are variable. In fact, the input to RNNs is $\mathbf{y_c}$, and the output of RNNs has the same size as the input, and they are dynamic according to Eq. (2).

C. Learning Rule

The modification of the parameters of the model described in the above is given by a gradient decent mini-batch learning algorithm using negative log-likelihood (*NLL*) function as follows.

$$NLL(\theta) = -\sum_{l=1}^{L} \sum_{j=0}^{N-1} \log p(\mathbf{y}^l{}_{j+1} \mid \mathbf{x}^l, \mathbf{y}_c; \boldsymbol{\theta}) \tag{8}$$

Where L is the size of training samples.

Input	[1]:<s> schizophrenia patients whose medication could n't stop the imaginary voices in their heads gained some relief after researchers repeatedly sent a magnetic field into a small area of their brains .<eos> [2]: <s> a yale school of medicine study is expanding upon what scientists know about the link between schizophrenia and nicotine addiction. <eos>
Output (Expected)	[1]: <s> Magnetic treatment may ease or lessen occurrence of schizophrenic voices. <eos> [2]: <s> Researchers examining evidence of link between schizophrenia and nicotine addiction. <eos>

FIGURE III. EXAMPLES OF ABSTRACTIVE SUMMARIZATION OF ENGLISH SENTENCES.

Epoch	Output (Proposed Model)
30:[1]	['<s>', 'This', 'treatment', 'may', '<s>', 'Researchers', 'lessen', '<s>', 'of', 'schizophrenic', 'voices.', '<eos>']
30:[2]	['<s>', 'Researchers', 'examining', 'evidence', 'of', 'link', '<eos>']
• • •	• • •
540:[1]	['<s>', 'Schizophrenia', 'treatment', 'may', 'ease', 'or', 'lessen', 'occurrence', 'of', 'schizophrenic', 'voices.', '<eos>']
540:[2]	['<s>', 'Neuroscientist', 'examining', 'evidence', 'of', 'link', 'between', 'schizophrenia', 'and', 'nicotine', 'addiction.', '<eos>']
• • •	• • •
5000:[1]	['<s>', 'Magnetic', 'treatment', 'may', 'ease', 'or', 'lessen', 'occurrence', 'of', 'schizophrenic', 'voices.', '<eos>']
5000:[2]	['<s>', 'Neuroscientist', 'examining', 'evidence', 'of', 'link', 'between', 'schizophrenia', 'and', 'nicotine', 'addiction.', '<eos>']
Output (Expected)	[1]: <s> Magnetic treatment may ease or lessen occurrence of schizophrenic voices. <eos> [2]: <s> Researchers examining evidence of link between schizophrenia and nicotine addiction. <eos>

FIGURE IV. EXPERIMENT RESULT OF THE PROPOSED METHOD

Epoch	Output (Rush-Chopra-Weston's Model)
30:[1]	['<s>', 'Schizophrenia', 'treatment', 'may', 'ease', 'or', 'lessen', 'occurrence', 'of', 'schizophrenic', 'voices.', '<eos>']
30:[2]	['<s>', 'Neuroscientist', 'examining', 'evidence', 'of', 'link', 'between', 'schizophrenia', 'and', 'nicotine', 'addiction.',
• • •	'<eos>']
540:[1]	['<s>', 'Schizophrenia', 'treatment', 'may', 'ease', 'or', 'lessen', 'occurrence', 'of', 'schizophrenic', 'voices.', '<eos>']
540:[2]	['<s>', 'Researchers', 'examining', 'evidence', 'of', 'link', 'between', 'schizophrenia', 'and', 'nicotine', 'addiction.',
• • •	'<eos>']
	• • •
5000:[1]	['<s>', 'This', 'treatment', 'may', 'ease', 'or', 'lessen', 'occurrence', 'of', 'schizophrenic', 'voices.', '<eos>']
5000:[2]	['<s>', 'Family', 'examining', 'evidence', 'of', 'link', 'between', 'schizophrenia', 'and', 'nicotine', 'addiction.', '<eos>']
Output (Expected)	[1]: <s> Magnetic treatment may ease or lessen occurrence of schizophrenic voices. <eos> [2]: <s> Researchers examining evidence of link between schizophrenia and nicotine addiction. <eos>

FIGURE V. EXPERIMENT RESULT OF THE CONVENTIONAL METHOD [6].

$$\theta^{new} = \theta^{old} + a\frac{\partial}{\partial\theta}NLL(\theta) \qquad (9)$$

Where $0 < \alpha < 1$ is the learning rate.

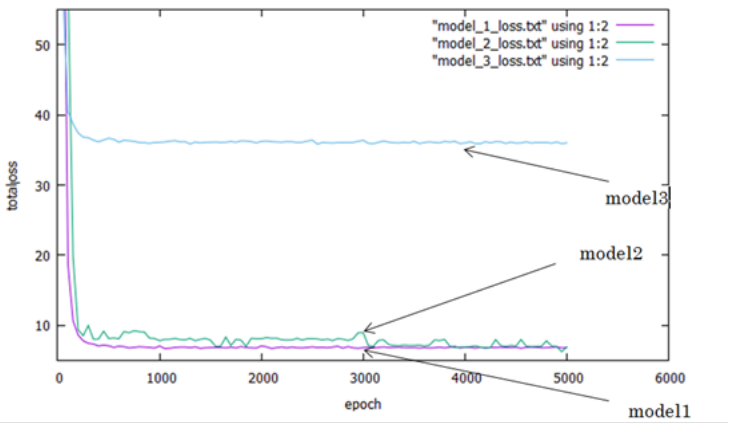

FIGURE VI. COMPARISON OF DIFFERENT MODELS: MODEL 1, 2, AND 3 INDICATE THE PROPOSED MODEL, THE CONVENTIONAL MODEL [6], AND THE PROPOSED MODEL WITH TEACHING SIGNALS LACKING OF THE FIRST WORD OF SENTENCES, RESPECTIVELY.

III. EXPERIMENT

Two summarization samples of English sentences are shown in Figure III. Articles on the site of "*The Wall Street Journal*" (https://www.wsj.com/) were used in our summarization experiment. The first sentences of articles were input of the model of summarizers, and the teach signals were the titles of articles. Starting symbol <s> and ending symbol <eos> were added in to both input sentences and output ones. There were 50 sets sentences were used in the experiment, and arbitrary extracted 4 sets from 50 samples were used as test samples.

A part of summarization experiment results are shown in Figure IV (the results of the proposed model) and Figure V (the results of the conventional model). In Figure IV, the output summarization of the proposed model after 5,000 times training, is completely as same as the teacher signal "Magnetic treatment may ease or lessen occurrence of schizophrenic voices" in a case of the first sample, and in the second case, only the subject of the sentence is different, i.e., the output of the model used "Neuroscientist", whereas the teacher' subject was "Researchers". The conventional model output the same result of the first sentence after training, however, the word of subject "Family" was not suitable for the second sentence contextually. The comparison of the learning performance of the different models is shown in Figure VI. In Figure 5, model 1, 2, 3 indicate the proposed model, the proposed model with teacher signals lacking of the first word of sentence. Hamming distance, for example, the distance between word "10111" and "11001" equals to 3, was used to evaluate the performance of different models (total loss). It can be confirmed that the proposed model had a prior learning convergence comparing to the conventional one [6], and it worked even using teach data with lacked elements.

IV. CONCLUSION AND FUTURE WORKS

To create the abstractive summarization of sentences in natural languages, an improved neural attention-based summarizer using recurrent neural networks was proposed in

this paper. The proposed model adopted the ideas of neural language translation model [1] [2], attention-based encoder [3], and attention-based summarization (ABS) [6]. Comparison of the experiment results showed the effectiveness of the proposed method. Meanwhile, the future work of this study is to compare it theoretically and experimentally with a novel abstractive summarizer with recurrent neural networks proposed by Chopra, Auli and Rush recently [7].

ACKNOWLEDGMENT

This work is supported by JSPS KAKENHI Grant No. 26330254 and No.25330287.

REFERENCES

[1] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin. 2003. “A Neural Probabilistic Language Model”, Journal of Machine Learning, Vol. 3, pp. 1137-1155.

[2] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk. and Y. Bengio. 2014. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation ”,Conference on Empirical Methods in Natural Language Processing, pp.1724-1734.

[3] D. Bahadanau, K. Cho, Y. Bengio. 2014. “Neural Machine Translation by Jointly Learning to Align and Translate”, CoRR, abs/1409.0473

[4] Sutskever, I., Vinyal, O., and Le, Q. V. 2014. “Sequence to Sequence Learning with Neural Networks”, In Advances in Neural Information Processing System, pp. 3104-3112.

[5] T. Luong, H. Pham, and C. D. Manning. 2015. “Effective Approaches to Attention-Based Neural Machine Translation.” Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421.

[6] A. M. Rush, S. Chopra, and J. Weston. 2015. “A Neural Attention Model for Abstractive Sentence Summarization”, In Proceedings of Empirical Methods in Natural Language Processing 2015, pp. 379-389.

[7] S. Chopra, M. Auli, A. M. Rush. 2016. “Abstractive Sentence Summarization with Attentive Recurrent Neural Networks”, In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016), pp. 93-98.