Scientific Research Publishing

Journal of Computer and Communications, 2021, 9, 161-168
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

The Automatic Question Generation System for CET

Xinya Zhang[1,2], Xiaodong Yan[1,2*], Zhou Yao[2]

[1]School of Information Engineering, Minzu University of China, Beijing, China
[2]Center of Minority Languages, National Language Resource Monitoring & Research, Beijing, China
Email: *yanxd3244@sina.com

How to cite this paper: Zhang, X.Y., Yan, X.D. and Yao, Z. (2021) The Automatic Question Generation System for CET. *Journal of Computer and Communications*, **9**, 161-168.
https://doi.org/10.4236/jcc.2021.99013

Received: April 20, 2021
Accepted: September 27, 2021
Published: September 30, 2021

Abstract

In this paper, we apply the abstractive text summarization technology to automatic generation system of reading comprehension, which is part of College English Test (CET) in China. At present, there is a growing demand of English reading examination questions, yet the manual examination question generating is time-consuming and labor-intensive. To relieve the pressure on question generating task, we put the related automatic technology into application, which aims to assist teachers in question generating, meanwhile, to provide more CET exercises for students. We combine seq2seq model and attention mechanism to generate the abstractive text summarization. The abstract generated by this method is easy to understand and in line with the question generating of long reading comprehension, the experiment showed good results of question generating.

Keywords

Text Summarization, seq2seq Model, Attention Mechanism, College English Test

1. Introduction

College English Test (CET-4 and CET-6) is a national English examination in charge of the Ministry of Education to measure the comprehensive English application ability of Chinese college students. The examination paper consists of writing, listening, reading and translation [1]. Among these 4 parts, the reading accounts for 35% of the total score [2]. Since 2012, CET-4 and CET-6 have been reformed from "one question with multiple papers" to "multiple questions with multiple papers". There are about 6 - 8 sets of papers in each test, and they are made and generated by related government authorities. After December 2013,

the paper structure and test question type of CET-4 are the same as CET-6, thus creating a great deal of English reading comprehension test questions. In this paper, we mainly focus on the automatic question-generating system on the long reading comprehension in CET. The long reading comprehension is a kind of matching reading question type, it provides an article and a topic. The topic is a brief summary of the given paragraph; the examinees are asked to match the topic with the related paragraph according to the key information of the brief summary. As the pattern of question generation is similar as the text summarization, we try to apply the abstractive text summarization technology to the automatic generation of the topic belonging to the long reading comprehension.

The text summarization technology can be roughly divided into 2 methods in terms of algorithm design principles: extractive text summarization method and abstractive text summarization method [3]. Based the value in the text, the extractive text summarization method extracts the sentences provided with more value in the text, to form the summary. The structure of the summary generated with this method is relatively complete, yet the central idea of the abstract is vague and hard to understand. The abstractive text summarization method uses the neural network model to encode the text, then decode the text vector, which is like the summary generated after re-understanding. Compared to the extractive text summarization method, the summary generated with the abstractive text summarization method is more in line with the author's original intention, easy to understand, and applicable to more fields [4].

As far as we know, there is no study or research on automatic generation system for English Reading based on text summarization technology for now. Our work applies the abstractive text summarization method to the automatic generation of the topic belonging to the long reading comprehension. Our work can largely raise the efficiency of reading questions.

2. Design of Automatic Generation System of English Reading Questions

2.1. Corpus Source and Preprocessing

The object of this article is mainly English reading comprehension in CET-4 and CET-6. After analyzing the reading questions in previous years, it is found that it mainly comes from The New York Times, The Economist, The Atlantic, Time, Newsweek, The Guardian and other newspapers. This article will crawl information from the websites of the above newspapers as a corpus. This article uses java crawler architecture. Crawl article content according to the front-end HTML/CSS code of different. Finally, it is stored in the database using UTF-8 encoding website pages [5].

Preprocessing the text is very important to the task directly affect downstream processing. His process mainly includes word segmentation, clause, morphological restoration, removal of stop words, etc. [6].

But English has natural spaces as separators. So English doesn't need word

segmentation. Usually use the defined punctuation between sentences for sentence processing. So will create a punctuation file. Which includes punctuation marks, such as ".,?,!" and so on.

Lemmatization is based on the lexical characteristics of words. It restores words in different tenses to words in the general present tense, usually only need to record the common tense conversion rules [7].

Invalid words, short words or punctuation in English sentences need to be removed, such as "a", "to" when doing feature processing. When doing feature processing, use the parameter stop words to introduce an array as a stop word list. Use the stop word list to remove the stop words in the feature word list.

2.2. Research on Generative Text Summarization Technology

This article uses a model that combines the seq2seq model with the attention mechanism. This model is very good at solving problems such as repetition and incompatibility in the generation of words and sentences. And the quality of abstract generation has been improved [8].

1) seq2seq model

The seq2seq model can be applied to tasks such as part-of-speech tagging, text summarization, translation, etc. The main research of this article is a text Summarization, the correspondence between the input sequence and the output sequence is not obvious. So the model in this article is built by two RNN (recurrent neural networks). An RNN is an encoder, it is responsible for encoding the input sequence X, convert the input sequence $\mathrm{X} = < x_1, x_2, x_3, \ldots, x_n >$ into a fixed-length semantic vector C, this process is a non-linear change. [6] Another RNN is decoder, It is responsible for generating the y-i to be generated at time I, it according to the semantic vector C and the generated information y_1, y_2, y_3, …, y_{i-1}. Finally generated sequence $\mathrm{Y} = < y_1, y_2, y_3, \ldots, y_m >$ (**Figure 1**).

$$\mathrm{C} = \mathrm{F}\left(x_1, x_2, x_3, \ldots, x_n\right) \tag{1}$$

$$y_i = G\left(\mathrm{C}, y_1, y_2, y_3, \ldots, y_{i-1}\right) \tag{2}$$

The end-to-end process of this model combines semantic understanding and text generation. But in the coding stage, regardless of the length of the input sequence, the length of the output semantic vector C is fixed. This leads to data loss when generating sequences. Therefore, we introduce an attention mechanism in the seq2seq model to solve the problem [9].

2) Attention Mechanism

The idea of the Attention mechanism is similar to manual abstract writing. First read through the text to understand the main idea, and then review the key content in the text to complete an abstract [10] (**Figure 2**).

In the summary generation process, the Attention mechanism provides a weight reference for the words in the sequence. There are three calculation steps (**Figure 3**).

First step, define an information element in the output sequence of the decoder as "Query", the sequence information in the encoder input sequence as a

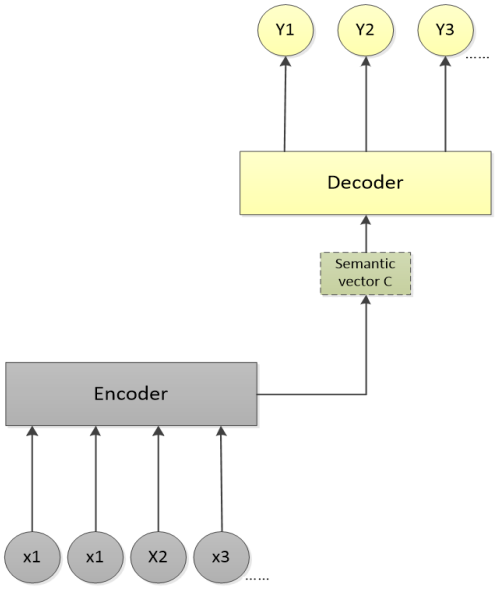

Figure 1. seq2seq model.

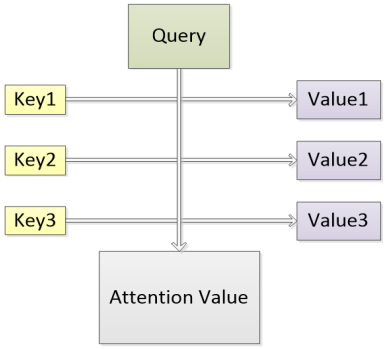

Figure 2. Attention frame.

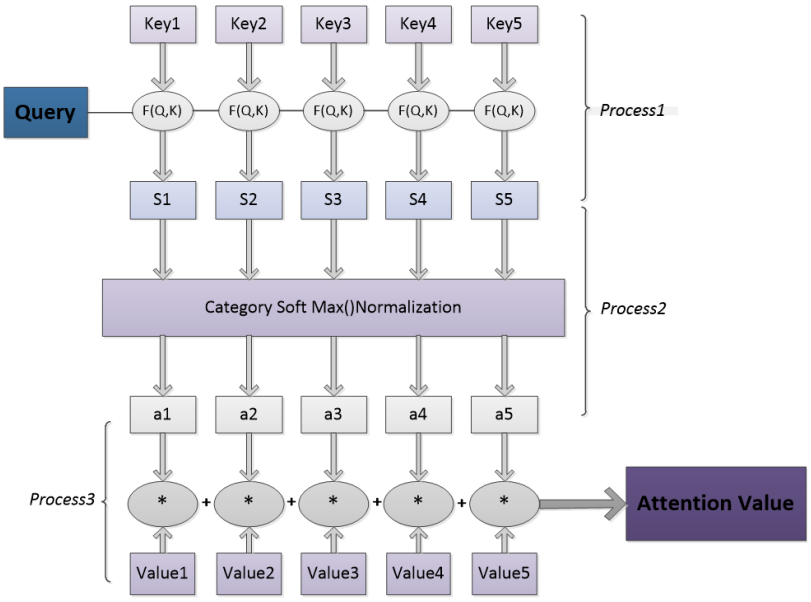

Figure 3. Attention calculation process.

"key". The "value" usually represents the sequence information with the same key (it is the semantic code corresponding to each word in the input sentence). That is to say, the input sequence is composed of "key-value" data pairs.

The second step is to calculate the attention weight coefficient. First calculate the similarity of key and value. Then normalize the calculated similarity. The final result is the normalized coefficient, which is the weight coefficient a_i.

$$a_i = \text{softmax}\left(\text{f}_{att}\left(s_{i-1}, h_j\right)\right) \tag{3}$$

The third step is the weighted sum of value, according to the weight coefficient. The final result is the weight of the value.

$$c_i = \sum_{j-1}^{m} a_{ij} h_j \tag{4}$$

After the above steps, different "Query" will calculate the corresponding "Value" weight, and the attention model can be obtained.

2.3. Generative Text Summarization Model Based on seq2seq Model and Attention Mechanism

It can be seen from the above, the seq2seq model converts the input sequence into a fixed-length semantic vector C, and then decodes the semantic vector C to generate a sequence Y [11]. But the length of the semantic vector C is fixed, so data loss will occur when the sequence is generated. This article adds an attention mechanism to the seq2seq model. In the encoding stage, the input sequence can not only be transformed into a fixed-length semantic vector C, the input sequence X_i will correspond to a different semantic vector C_i. In this way, the data loss will be reduced when generating the sequence [10] (**Figure 4**).

The abstract generated in this way is more similar to the abstract generated after manual re-understanding, which is more in line with the author's original intention and high intelligibility [12]. At the same time, it conforms to the characteristics of English long reading.

3. The Application of Automatic Generation System of Reading Questions

The data used in this paper are selected from the official websites of The Guardian, The New York Times, The Economist, The Atlantic, News Week, etc. The data contain articles on domains of economics, technologies, and humanities. We take the data crawling from The Guardian as the example to show the training result of the model (**Table 1**).

In this experiment, 8 paragraphs are summarized with the model. We checked the results and found that 3 paragraphs are effective and available, the precision is 37.5% (**Table 2**).

The simple summarization generated by the model can be used as the topic of the paragraph. A complete long reading comprehension topic can be generated in the process of numbering the topics, rearranging the sequences, adding the interference options.

Table 1. The result of the summary generation.

	Paragraph	Result
Example 1	The first way to solve a problem is by identifying what the problem is. Identify the circumstances and emotions that lead you to stress-eat. These are your emotional eating triggers, and once you recognize them, you can take steps to avoid them or at least be prepared for them.	Identify the stress problem and prepare.*
Example 2	It might sound odd, but this is the first step to combat stress eating. So, keep a diary and make note of what you eat, how much you eat, when you eat, what kind of emotions do you have while eating, and how hungry you are. You can show your doctor the diary so as to help you measure and screen what you should or should not eat.	It show the combat stress eating.*
Example 3	Often, exercise helps you reduce stress. If you are physically fit, you're more resistant to the effects of stress according to Psycom. Exercise causes chemical changes in the brain that reduce stress but, unfortunately, stress itself can prevent some people from taking steps, like exercising, that could make a difference in their mental and physical health.	You exercise physical health.*
Example 4	If you think it becomes easier for you to eat junk food when you are filled with emotions, then why not remove all of that food, you can always go back to grocery shopping when your emotions are in check. Some people shop when they are emotionally down, some take to alcohol, but yours is eating; to avoid eating junk excessively, it is important to do away with them.	Go back to the grocery store and go shopping.#
Example 5	It is possible that you eat unhealthy because of boredom, so there is a need to address the root cause of the problem. Engage in other things that can take your mind off food such as reading, music, or hanging out. When boredom strikes, instead of eating, try these things and see how it works.	You eat unhealthy, you reading, music, or hanging out.*
Example 6	Discuss your feelings and your unhealthy responses to stress with close friends and family who can give you the support you need to get through this situation. If you often feel guilt, shame, or regret over your eating habits, you may want to speak with a professional counselor. There is nothing too much for you to discuss with your counsellor, they will hold your hands and help you through the situation.	You often feel guilt, shame, or regret over your unhealthy responses to stress.#
Example 7	Stress is a common trigger for emotional eaters because so many everyday life circumstances cause the stress and anxiety that leads to overeating. Some stressors come from within, like the stress you put on yourself to be perfect or the anxiety you feel when you want to do some important things for yourself. Other stressors come from outside of yourself, such as the demands of your job, medical issues, family obligations, and social pressure from friends.	The stress comes from inside and outside of yourself.#
Example 8	People cope differently with stress, while some take to shopping, others over-eat and do what is called stress-eating. Stress-eating can, however, be controlled and some steps can be taken to reduce this problem.	While some take to shopping steps can be taken to take to shopping.*

Note: # means the summary sentence is available, # means the summary sentence is not available.

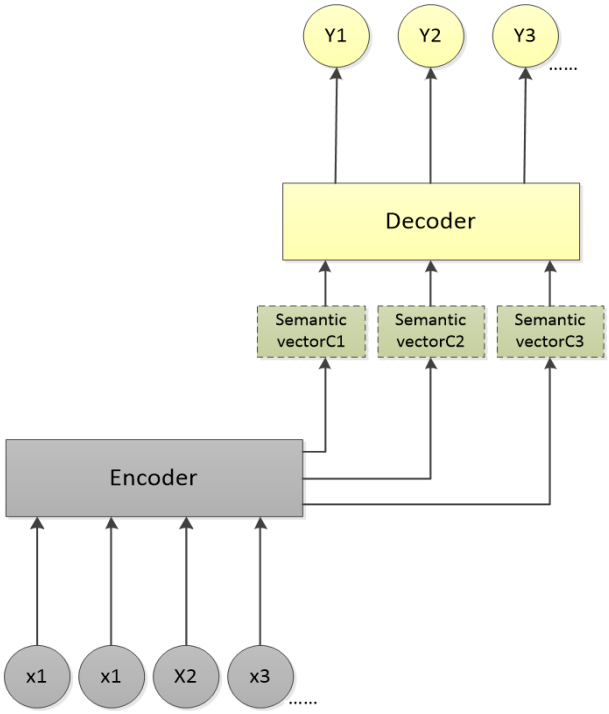

Figure 4. Seq2seq model architecture with Attention mechanism.

Table 2. Availability of generated results.

Sample size	P
200	75/200 = 37.5%

4. Conclusion

In this paper, we applied the technology of abstractive text summarization technology to the automatic generation of the topic belonging to the English reading comprehension. To achieve the goal, the seq2seq model and the attention mechanism are united to make the summary generated by the automatic system be more fit for the summary generated by human beings. Besides, we reduced the data loss rate in the process of the summary generation. As illustrated above, a large number of reading questions are needed by both official language examination board and private education organization. We put the related automatic technology into application that could largely relieve the pressure on question generating task. Also, our work could provide a great deal of CET exercises for students. Our future work is more about the other types of CET reading comprehension, such as the careful reading, whose question generating type is different from long reading question generation. We are going to try to use deep learning to generate the careful reading question generation.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Xi, C. (2017) Research on Reading Texts of College English Band 4 and Band 6 Based on Self-built Corpus. *Overseas English*, No. 20, 50-51.

[2] Lei, G. (2012) Text Analysis of Fast Reading in College English Band 4 and Band 6: Empirical Research Based on Self-Built Small Corpus. *Journal of Shanxi Radio & TV University*, **17**, 69-71.

[3] Qi, Z., Ling, Z. and Ya, Z. (2008) A Survey of Chinese Word Segmentation Algorithms. *Information Research*, No. 11, 53-56.

[4] Wei, L., Xiao, Y. and Xiao, X. (2020) An Improved TextRank for Tibetan Summarization. *Journal of Chinese Information Processing*, **34**, 36-43.

[5] Wei, L. (2020) Research on Tibetan News Abstract Generation Based on Unified Model. M.S. Thesis, Minzu University of China, Beijing.

[6] Qing, L. (2020) Research and Implementation of Text Summarization Technology Based on Machine Learning. M.S. Thesis, University of Electronic and Science and Technology of China, Chengdu.

[7] Tong, R. (2018) Study of Deep Learning Based Text Summarization. M.S. Thesis, Xidian University, Xi'an.

[8] Nallpati, R., Zhou, B., Santos, C., *et al.* (2016) Abstractive Text Summarization Using Sequence-to-Sequence Rnns and beyond. *Proceedings of the* 20*th SIGNLL Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Berlin, 280-290. https://doi.org/10.18653/v1/K16-1028

[9] Sutskever, I., Vinyals, O. and Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, MIT Press, MA, 3104-3112.

[10] Yu, L. (2019) Research and Implementation of Text Summarization Technology Based on Attention Mechanism. M.S. Thesis, Huazhong University of Science and Technology, Wuhan.

[11] Hong, G. (2018) Research on Abstractive Text Summarization Based on Deep Learning. M.S. Thesis, Harbin Institute of Technology, Harbin.

[12] Xiao, X. (2021) Research on the Generation of Korean Multi-Document Summaries. M.S. Thesis, Minzu University of China, Beijing.