Automatic Question Generation System for English Reading Comprehension

Yin-Chun Fung[1](✉), Jason Chun-Wai Kwok[1], Lap-Kei Lee[1], Kwok Tai Chui[1], and Leong Hou U[2]

[1] School of Science and Technology, The Open University of Hong Kong, Ho Man Tin, Kowloon, Hong Kong SAR, China
{ycfung,g1203482}@study.ouhk.edu.hk, {lklee,jktchui}@ouhk.edu.hk
[2] SKL of Internet of Things for Smart City, Department of Computer and Information Science, University of Macau, Macao SAR, China
ryanlhu@um.edu.mo

Abstract. This paper presents a web-based automatic question generation (AQG) system to generate reading comprehension questions and multiple-choice (MC) questions on grammar from a given English text. Such system saves teachers' time on setting questions and facilitates students and their parents to prepare self-learning exercises. Our web-based system can automatically generate Wh-questions (i.e., what, who, when, where, why, and how) and MC grammar questions of selected sentences. Wh-questions can also be generated from user-specified answer phrases. The generation of Wh-questions exploits the pre-trained natural language understanding model, Text-To-Text Transfer Transformer (T5), and an adapted version of the SQuAD 2.0 machine reading comprehension dataset. The generation of MC questions involves identifying regular verbs in a text and using the verb's lexemes as the answer choices. Our system takes an average time of about 1 s to generate a Wh-question and it generates a MC question almost instantly. User evaluation indicated that our system is easy-to-use and satisfactory in usefulness, usability, and quality, revealing the effectiveness of our system for teachers and parents.

Keywords: Automatic question generation · English reading comprehension · Wh-questions · Multiple-choice questions · Natural language processing

1 Introduction

Language learning is compulsory in most schools. According to Cartwright (2002), reading is a cognitive demanding task. Reading comprehension can facilitate the development of cognitive skills and the learning of new vocabulary (Nagy 1988). Hence, reading comprehension is commonly adopted in the language learning process. Teachers give reading comprehension as assessment in schools. There are numbers of online platforms for students to practice reading comprehension nowadays. Some parents would like to set up exercises for their children to train their reading skills. There are

L.-K. Lee et al. (Eds.): ICTE 2020, CCIS 1302, pp. 136–146, 2020.
https://doi.org/10.1007/978-981-33-4594-2_12

a lot of articles on the Internet that can be used as reading comprehension material. However, setting up questions is a time-consuming and labour-intensive task (Mitkov et al. 2006). Recently, some researchers focused on automatic question generation (AQG) (see Kurdi et al. (2020) and the references therein). Yet, it is rare to find an existing system that is easy and free for non-technical users to generate reading comprehension questions in a fast and massive way. In addition, English has long been one of the most important language, supported by many research articles (Bury and Oka 2017; Qi 2016). Thus, we have decided to implement a web-based system to provide an easy-to-use platform for teachers and parents of junior students to generate English reading comprehension exercises.

1.1 Existing Approach on AQG

There are numbers of algorithms on automatic question generation (AQG). We can generally classify them into three categories: template-based, syntax-based, and semantics-based (Kurdi et al. 2020; Yao et al. 2012):

- **Template-based.** The first step of template-based approach is to define templates consisting of some fixed text, such as ‘What is X’ and ‘When did X begin’. The main purpose of this category of AQG algorithm is to find suitable keywords in a text to substitute ‘X’ into the template (Liu et al. 2018).
- **Syntax-based.** Syntax-based approach makes use of syntax structure of sentences in a text. Syntax rules and transformation are defined. If there are sentences that match the syntax rules, questions can be generated using transformation rules (Heilman and Smith 2009).
- **Semantics-based.** Semantic features are analysed in semantics-based approach. By recognizing the semantic meaning between phrases, questions are generated. (Flor and Riordan 2018).

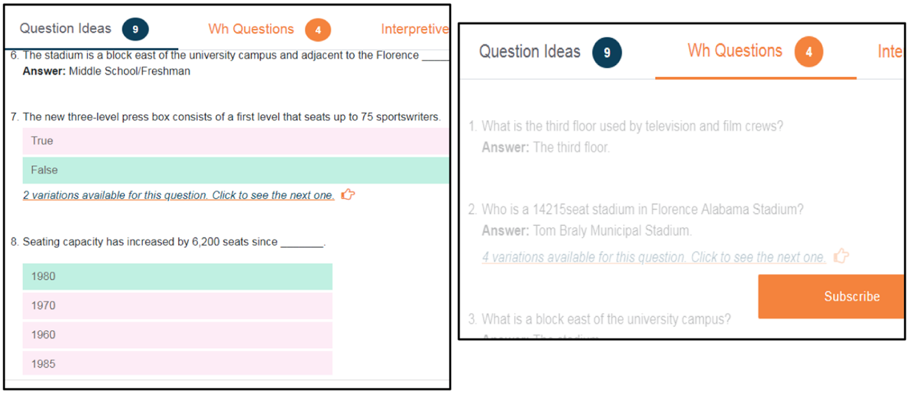
Recently, more studies are using sequence to sequence (seq2seq) encoder-decoder model with attention mechanism on AQG (Du et al. 2017; Zhou et al. 2018; Yuan et al. 2017; Zhao et al. 2018; Hosking and Riedel 2019). Seq2seq, introduced by Google (Sutskever et al. 2014), means that the input and output are both sequences. Encoder-decoder is an architecture that combines the encoder and decoder network. The encoder is a network that turns the input into a vector containing the information/features of the input. The decoder is the opposite of encoder, which is a network that turns the vector into an output item. They have usually been employed together as encoder-decoder architecture. Attention mechanism (Bahdanau et al. 2015) is often used with encoder-decoder to improve the performance. It allows the model to pay attention to the most important part of the text instead of all the text.

Transformer-based models (Vaswani et al. 2017) are also popular in AQG nowadays. It is an encoder-decoder architecture with a multi-head self-attention mechanism. Nearly all state-of-the-art models are based on the transformer. Bidirectional encoder representations from transformers (BERT) (Devlin et al. 2019) is the most popular model in natural language understanding. It consists of multi-layers of transformer encoders. GPT-2 (Radford et al. 2019) is the most powerful language model for text generation. It is based on the transformer decoder. These examples show how powerful and inspirational the transformer is.

Unlike the typical neural network model that requires to be trained from scratch, most of the transformer-based models are general-purpose language models that make use of transfer learning. These models are following a pre-training and fine-tuning process. In the pre-training stage, the model is trained with an extremely large corpus, so that the model is able to have a better representation of the language. This step usually requires a lot of computational resources. It can cost thousands of dollars per training. Therefore, researchers and organizations who proposed the models usually release the pre-trained models. Users can download the pre-trained models and perform fine-tuning for different downstream tasks with much less computational cost, while enjoying the performance gain from pre-training. It is worth noting that the fine-tuning process is equivalent to the tuning of hyperparameters. It is the training process that trains the model with the data for a specific task.

1.2 Existing System for AQG

Quillionz is an artificial intelligence (AI)-powered question generator. It provides both free and paid services. Free service provides only True-False questions, multiple choices questions and fill-in-the-blank questions, while Wh-questions and interpretive questions generation are paid service (Fig. 1). Users need to input a text between 300 and 3000 words and select the domain of the text. The system will initialize some keywords for the question generation process. Users can choose to include or exclude some of the keywords. It then requires users to review the content, which includes solving lengthy sentences, resolving pronouns, and modifying some subjective or incomplete sentences. This process aims to improve the quality of generated questions. Finally, questions will be generated.



6. The stadium is a block east of the university campus and adjacent to the Florence ______
Answer: Middle School/Freshman

7. The new three-level press box consists of a first level that seats up to 75 sportswriters.
True
False
2 variations available for this question. Click to see the next one.

8. Seating capacity has increased by 6,200 seats since _______.
1980
1970
1960
1985

1. What is the third floor used by television and film crews?
Answer: The third floor.

2. Who is a 14215seat stadium in Florence Alabama Stadium?
Answer: Tom Braly Municipal Stadium.
4 variations available for this question. Click to see the next one.

3. What is a block east of the university campus?

Fig. 1. Quillionz generates True-False, multiple choices and fill-in-the-blank questions for free (left). Wh- and interpretive questions generation are paid service (right).

The limitations of Quillionz are the minimum word count and the control of the content. The system rejects a text with less than 300 words which is not flexible for generating reading comprehension questions for junior students. If the text is not in the listed domain of the system, the generated questions may not be of good quality. Moreover, the system requires human effort to modify the text to fit the system in order to generate good quality questions. It is not user-friendly when the users are asked to "rewrite" the text but they do not think the text is that poorly written.

2 Our Web Application for AQG

Our AQG system consists of two major QG components: Wh-question generation and grammar question generation. When using the system, users need to input a text. The text will be processed in two ways to produce Wh-questions and multiple choice questions on grammar, respectively. Figure 2 shows the overall design of our system.

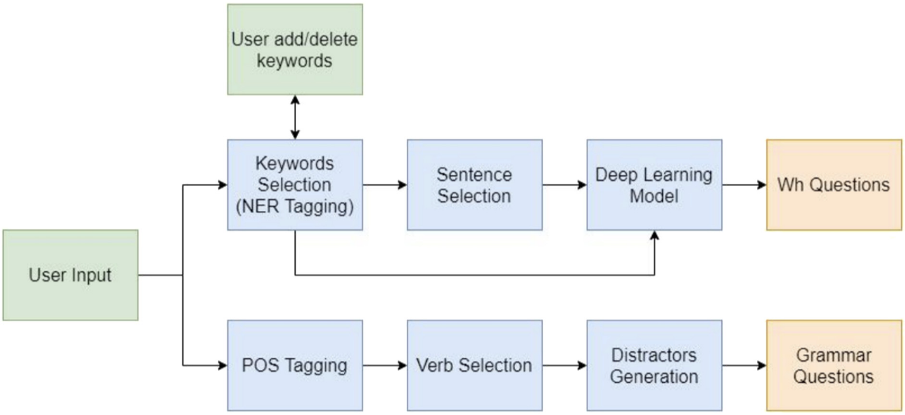


Fig. 2. The overall design of our web-based AQG system.

2.1 Grammar Question Generator

Our grammar questions are multiple choice questions on tenses. The first step is to identify which are the verbs in the text. We used a part-of-speech (POS) tagger (De Smedt and Daelemans 2012) to identify the part of speech of all individual words in the text. We choose all the verbs and produce its other lexemes as the distractors (i.e., wrong answers) of the question. If the number of lexemes is less than three for a particular verb, this verb will be ignored. There is no pre-training in this part of the generation process.

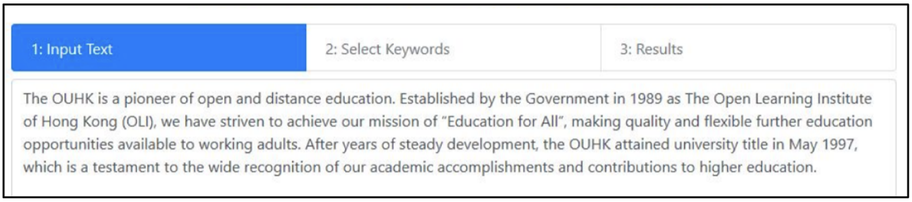
2.2 Wh-Question Generator

We have used a deep learning approach to generate Wh-questions. A pre-trained English model, Text-To-Text Transfer Transformer (T5), was adopted (Raffel et al. 2019) as the base model. Considering the processing power and the responding time of the system, we use “t5-small”, the smallest model of the T5 family, in our work. We fine-tune the T5 model using the benchmark SQuAD 2.0 dataset (Rajpurkar et al. 2018) to make it suitable for generating questions. Since the SQuAD 2.0 dataset is for question answering, we simply treat the text, question-answer pairs of the dataset as the inputs and outputs for fine-tuning our model to better fit the question generation task: the input is a source sentence with an answer phrase, whereas the output is a Wh-question.

When using our system, the input text will go through a named-entity tagging process to find the possible answer phrases. Then, the selected phrases, called keywords, will be passed into our fine-tuned model together with the source sentence. A Wh-question will then be produced as the output.

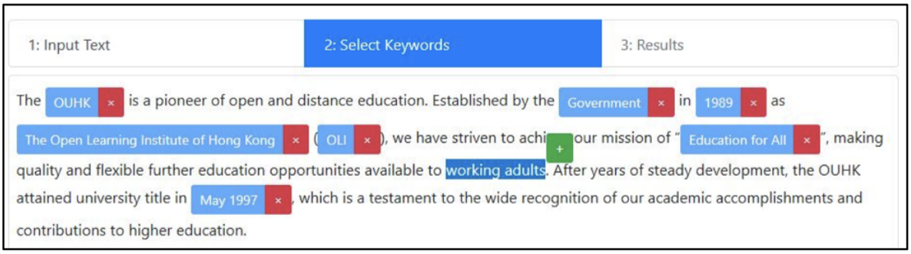
2.3 Wrapping the Two Generators

We used a web application to wrap the whole process. To generate questions, users will be required to input a text (Fig. 3). The system will select and display the possible answer phrases to the users. Users can choose new keywords and un-select any keyword according to their needs (Fig. 4).



The OUHK is a pioneer of open and distance education. Established by the Government in 1989 as The Open Learning Institute of Hong Kong (OLI), we have striven to achieve our mission of "Education for All", making quality and flexible further education opportunities available to working adults. After years of steady development, the OUHK attained university title in May 1997, which is a testament to the wide recognition of our academic accomplishments and contributions to higher education.

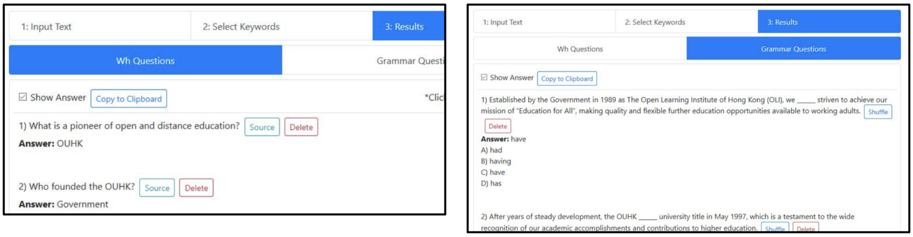
Fig. 3. First step of using our system: input a text.



The OUHK × is a pioneer of open and distance education. Established by the Government × in 1989 × as The Open Learning Institute of Hong Kong × (OLI ×), we have striven to achi our mission of "Education for All ×", making quality and flexible further education opportunities available to working adults. After years of steady development, the OUHK attained university title in May 1997 ×, which is a testament to the wide recognition of our academic accomplishments and contributions to higher education.

Fig. 4. Second step of using our system: select keywords.

When all settings are ready, questions will be generated automatically (Fig. 5), as introduced in Sects. 2.1 and 2.2. Generated Wh-questions are shown on the left in Fig. 5. Users can click on the “Source” button to see the original sentence in case they want to make sure the questions are set correctly, or would like to examine the answer (Fig. 6). The right of Fig. 5 shows the generated multiple choice questions on grammar. Users can shuffle the choices of the questions by clicking the “shuffle” button.



1) What is a pioneer of open and distance education?

Answer: OUHK

2) Who founded the OUHK?

Answer: Government

1) Established by the Government in 1989 as The Open Learning Institute of Hong Kong (OLI), we _____ striven to achieve our mission of "Education for All", making quality and flexible further education opportunities available to working adults.

Answer: have

A) had

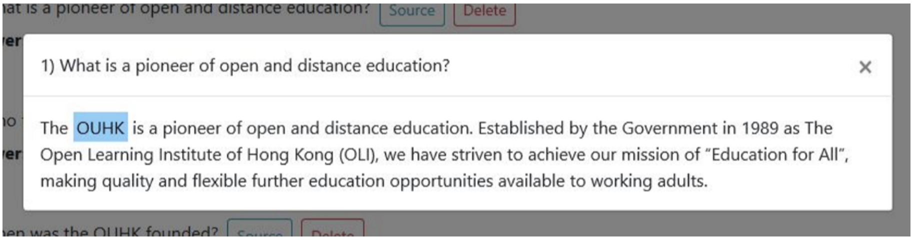
B) having

C) have

D) has

2) After years of steady development, the OUHK _____ university title in May 1997, which is a testament to the wide recognition of our academic accomplishments and contributions to higher education.

Fig. 5. Wh-questions (left) and Grammar multiple choice questions (right) are generated.



1) What is a pioneer of open and distance education?

The OUHK is a pioneer of open and distance education. Established by the Government in 1989 as The Open Learning Institute of Hong Kong (OLI), we have striven to achieve our mission of "Education for All", making quality and flexible further education opportunities available to working adults.

Fig. 6. Users can view the source of the generated question.

In both question pages (Fig. 5), a user would remove some questions directly using the "Delete" button next to those questions. All answers will be hidden by un-selecting the "Show Answer" option. The "Copy to Clipboard" button allows users to copy the questions to other file for further processing. Users can also print the web page directly.

3 Evaluation

3.1 Response Time

We conducted a response time test to evaluate the time needed to generate a question using GPU and CPU. There are 5 randomly chosen texts, the number of words and number of questions is shown in Table 1 below. We collected the time used to generate questions in these 5 texts, and the average time needed per question, and tried running the model on CPU (i7-4770) and GPU (RTX 2060).

The result shows that the average time cost per question using CPU is 1.3 s, and that of GPU is 1.07 s. It shows that GPU outperforms CPU by 18% in AQG.

Note that the time for the NER tagging and generating grammar questions is very short (much less than one second) and are therefore omitted from the response time test. To sum up, our web-based AQG system provides satisfactory response time on the question generation.

3.2 Quality of Generated Questions

We evaluated the capability of the system in generating different types of questions, including "What", "Who", "When", "Where", "Why", "How", and "How many" (Table 2).

Table 1. Result of response time test.

	No. of words	No. of questions	Time used (second)		Avg. time cost per question (second)	
			CPU	GPU	CPU	GPU
Text 1	312	26	35.1	28.7	1.35	1.10
Text 2	180	25	32.0	26.5	1.28	1.06
Text 3	340	18	23.1	19.0	1.28	1.06
Text 4	160	19	25.8	20.5	1.36	1.08
Text 5	117	23	28.7	24.6	1.25	1.07
Mean average time cost per question (second)					1.30	1.07

The results indicated that the system can generate different types of Wh-questions if the answers are selected properly. Our auto-selected answers are based on NER tagging, which can select the answers for generating “What”, “Who”, “When”, “Where”, and “How many” type of questions. Yet it is not suitable for automatically generating the “Why” and “How” type of questions. The generation of these types of questions highly relies on user involvement and thus the user may need to manually select the answers to ensure the quality of the questions.

3.3 User Survey

To evaluate the performance of our web-based AQG system, we invited 15 participants to test the system and complete a survey. The survey has 8 questions, where Questions 1 to 7 used a 5-point Likert scale (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree) and are on the usefulness, usability, and look-and-feel of the system, and Question 8 is an open-ended question. Table 3 shows the result on Questions 1–7.

The first two questions are on the usefulness of our system. The result shows that most of the users are satisfied with the quality of the questions. However, we can see the satisfaction of grammar questions is higher than Wh-questions, which implies a room for improvement in the Wh-questions generation component.

Questions 3 and 4 are on the usability of our system. It reflects if the design of the system and user interface is logical. The result shows that a majority (over 80%) of users gave the highest rating 5 on this part.

Questions 5 and 6 are on the look-and-feel of the system, which is mostly related to the user interface. The result shows most of the users are satisfied with the user interface, but there is still a little room for improvement since there are still 33% of the users who did not select the highest score. We expect adding more instructions and functions to the user interface could be useful.

As indicated in the result of Question 7, all users were satisfied with our system.

Table 2. Result of question type test.

Type		
What	Content	Amazon is the most valuable brand in the world.
	Selected Keyword	Amazon
	Generated Question	What is the most valuable brand in the world?
Who	Content	Donald Trump is the current president of the United States
	Selected Keyword	Donald Trump
	Generated Question	Who is the current president of the United States?
When	Content	Donald Trump was born on June 14, 1946
	Selected Keyword	June 14, 1946
	Generated Question	When was Donald Trump born?
Where	Content	Donald Trump was born at the Jamaica Hospital in the borough of Queens, New York City
	Selected Keyword	Jamaica Hospital
	Generated Question	Where was Donald Trump born?
Why	Content	The increased solubility of O2 at lower temperatures (see Physical properties) has important implications for ocean life, as polar oceans support a much higher density of life due to their higher oxygen content
	Selected Keyword	higher oxygen content
	Generated Question	Why do polar oceans support a much higher density of life?
How	Content	The principal Treaties that form the European Union began with common rules for coal and steel, and then atomic energy, but more complete and formal institutions were established through the Treaty of Rome 1957 and the Maastricht Treaty 1992 (now: TFEU)
	Selected Keyword	with common rules for coal and steel
	Generated Question	How did the Treaties that form the European Union begin?
How many	Content	Only a few common complex biomolecules, such as squalene and the carotenes, contain no oxygen
	Selected Keyword	A few
	Generated Question	How many complex biomolecules contain no oxygen?

Question 8 is an open-ended question for users to comment on the system. It aims at collecting users' feedback to improve the limitations of the system. There are three main limitations: (i) The first one concerns that the auto-selected keywords are not good enough. The keywords are selected based on NER tagging. It is possible that the whole input text has no or very little named-entity. We also noticed that it is a challenging task to generate the "why" question if the answer is a named-entity. Therefore, it is one of

Table 3. Result of user survey on Questions 1 to 7.

Question		Percentage (%)				
		1	2	3	4	5
1	I am satisfied with the quality of Wh-questions	0	0	13.3	40	46.7
2	I am satisfied with the quality of grammar questions	0	0	0	13.3	86.7
3	I think the system is easy to use	0	0	0	13.3	86.7
4	I think the system is easy to learn	0	0	0	20	80
5	The system works the way I expected	0	0	0	33.3	66.7
6	I like the user interface of the system	0	0	0	33.3	66.7
7	Overall, I am satisfied with the system	0	0	0	46.7	53.3

the key issues to be solved in the future; (ii) the second one concerns the long loading time. The major reason for the long loading time is that the model we used is large. Powerful hardware or optimizing the program could help with this problem; and (iii) the last one concerns the quality of the generated questions. The quality of the questions mostly relies on the model. There is a trade-off of performance and response time.

4 Conclusion and Future Work

An easy-to-use and web-based English reading comprehension question generation system has been built. It can generate Wh-questions and multiple choice grammar questions on tenses. Analysis revealed that the question generation takes has a satisfactory response time (about one second per question). Survey was conducted and participants were satisfied with the current system. Though, there is room for improvement in future work. Future research directions are suggested in three perspectives.

First of all, considering the quality of generated questions. The auto-selected keywords need to include those other than named entities such that the "why" type questions can be generated without human involvement. Moreover, more questions types, like part of speech and open-ended questions, can be included in the future. If the difficulty of the questions can be increased, the system will be useful to senior students and their teachers and parents.

Second, the response time should be reduced. When processing a long text, the response time is not impressive. Rather than using a pre-trained model, we can train a model just for AQG from scratch. In future, we may also use other new models that are much powerful but with smaller size than “t5-small”.

Third, we can mimic similar approaches for different languages such as Chinese. Reading comprehension is also playing an important role in other languages. If different languages can be supported, more people can enjoy the convenience of AQG.

Looking to the future, we believe that AQG on reading comprehension can be greatly helpful to teachers, parents, and students. Teachers can save their time on preparing teaching materials. Parents can prepare more exercise for their children and the children as students can learn more from the exercises. We look forward to the well development on AQG.

References

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, January 2015

Bury, J., Oka, T.: Undergraduate students’ perceptions of the importance of English in the tourism and hospitality industry. J. Teach. Travel Tour. **17**(3), 173–188 (2017)

Cartwright, K.B.: Cognitive development and reading: the relation of reading-specific multiple classification skill to reading comprehension in elementary school children. J. Educ. Psychol. **94**(1), 56 (2002)

De Smedt, T., Daelemans, W.: Pattern for Python. J. Mach. Learn. Res. **13**, 2031–2035 (2012)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1), January 2019

Du, X., Shao, J., Cardie, C.: Learning to ask: neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1342–1352, July 2017

Flor, M., Riordan, B.: A semantic role-based approach to open-domain automatic question generation. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 254–263, June 2018

Heilman, M., Smith, N.A.: Question generation via overgenerating transformations and ranking (No. CMU-LTI-09–013). Carnegie-Mellon Univ Pittsburgh pa language technologies insT (2009)

Hosking, T., Riedel, S.: Evaluating rewards for question generation models. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2278–2283, June 2019

Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2020)

Qi, G.Y.: The importance of English in primary school education in China: perceptions of students. Multiling. Educ. **6**(1), 1–18 (2016). https://doi.org/10.1186/s13616-016-0026-0

Liu, T., Wei, B., Chang, B., Sui, Z.: Large-scale simple question generation by template-based seq2seq learning. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Yu. (eds.) NLPCC 2017. LNCS (LNAI), vol. 10619, pp. 75–87. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73618-1_7

Mitkov, R., Le An, H., Karamanis, N.: A computer-aided environment for generating multiple-choice test items. Nat. Lang. Eng. **12**(2), 177 (2006)

Nagy, W.E.: Teaching vocabulary to improve reading comprehension. National Council of Teachers of English, Urbana, IL.; ERIC Clearinghouse on Reading and Communication Skills, Urbana, IL.; International Reading Association, Newark, DE (1988)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)

Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer (2019). arXiv preprint arXiv:1910.10683

Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 784–789, July 2018

Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)

Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

Yao, X., Bouma, G., Zhang, Y.: Semantics-based question generation and implementation. Dialogue Discourse **3**(2), 11–42 (2012)

Yuan, X., et al.:Machine comprehension by text-to-text neural question generation. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 15–25, August 2017

Zhao, Y., Ni, X., Ding, Y., Ke, Q.: Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3901–3910, October 2018

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M.: Neural question generation from text: a preliminary study. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Yu. (eds.) NLPCC 2017. LNCS (LNAI), vol. 10619, pp. 662–671. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73618-1_56