inventions

MDPI

Communication

An Automatic Question Generator for Chinese Comprehension

Yin-Chun Fung *, Lap-Kei Lee and Kwok Tai Chui

School of Science and Technology, Hong Kong Metropolitan University, Hong Kong SAR, China
* Correspondence: ycfung@study.hkmu.edu.hk

Abstract: Question generation (QG) is a natural language processing (NLP) problem that aims to generate natural questions from a given sentence or paragraph. QG has many applications, especially in education. For example, QG can complement teachers' efforts in creating assessment materials by automatically generating many related questions. QG can also be used to generate frequently asked question (FAQ) sets for business. Question answering (QA) can benefit from QG, where the training dataset of QA can be enriched using QG to improve the learning and performance of QA algorithms. However, most of the existing works and tools in QG are designed for English text. This paper presents the design of a web-based question generator for Chinese comprehension. The generator provides a user-friendly web interface for users to generate a set of wh-questions (i.e., what, who, when, where, why, and how) based on a Chinese text conditioned on a corresponding set of answer phrases. The web interface allows users to easily refine the answer phrases that are automatically generated by the web generator. The underlying question generation is based on the transformer approach, which was trained on a dataset combined from three publicly available Chinese reading comprehension datasets, namely, DRUD, CMRC2017, and CMRC2018. Linguistic features such as parts of speech (POS) and named-entity recognition (NER) are extracted from the text, which together with the original text and the answer phrases, are then fed into a machine learning algorithm based on a pre-trained mT5 model. The generated questions with answers are displayed in a user-friendly format, supplemented with the source sentences in the text used for generating each question. We expect the design of this web tool to provide insight into how Chinese question generation can be made easily accessible to users with low computer literacy.

Keywords: natural language processing; question generation; deep learning

check for updates

Citation: Fung, Y.-C.; Lee, L.-K.; Chui, K.T. An Automatic Question Generator for Chinese Comprehension. *Inventions* **2023**, *8*, 31. https://doi.org/10.3390/inventions8010031

Academic Editor: Leo Chen

Received: 14 December 2022
Revised: 18 January 2023
Accepted: 19 January 2023
Published: 28 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Listening, writing, reading, and speaking are the four major components in every human language [1,2]. For language learning in the modern education system, students learn and are evaluated through these four components. Among these components, reading is a cognitively demanding task [3]. Through reading a text, one can develop cognitive skills and can learn new vocabulary [4]. Hence, reading comprehension is commonly used in the learning of any language, as well as in language assessments for university/school admission and job recruitment. Hence, there is a strong demand for reading comprehension exercises.

The Internet provides rich resources of texts for reading comprehension such as Wikipedia and the news. However, it is time-consuming and labor-intensive to set up comprehension questions manually [5]. Hence, automatic question generation (AQG) has become a hot topic in scientific research [6]. There is a high volume of research on AQG, but most of these works focus on various aspects of the generated question quality. It is not easy to find an AQG system that allows users to generate reading comprehension questions quickly and in quantity, especially for non-English languages.

This paper presents the design of a web-based system to provide an easy-to-use platform to generate Chinese reading comprehension exercises. Given a Chinese text passage, the system automatically identifies a set of possible answer phrases from the passage,

which can also be easily refined through a user-friendly interface. Based on the passage and the answer phrases, a large number of wh-questions (on what, who, when, where, why, and how) can be generated instantly. The generated questions with answers are displayed in a user-friendly format, supplemented with the source sentences in the text used for generating each question.

The underlying question generation is based on the transformer approach, which was trained on a dataset combined from three publicly available Chinese reading comprehension datasets, namely, CRMC2017 [7], CRMC2018 [8], and DRCD [9]. The system extracts linguistic features from the text such as parts of speech (POS) and named-entity recognition (NER), which together with the original text and the answer phrases, are fed into a machine learning algorithm based on a pre-trained mT5 model [10]. We expect the design of this web tool to provide insight into how Chinese question generation can be made easily accessible to users with low computer literacy.

The organization of the paper is as follow: Section 2 reviews the existing AQG approaches. Section 3 provides the details of our web-based AQG system. Section 4 provides the evaluation results of our AQG system. Finally, Section 5 discusses the evaluation results and concludes our work with possible directions for future work.

2. Related Works

2.1. AQG Methods

Natural language processing (NLP) is a well-studied and important research topic with many applications [11–14]. AQG is one of the NLP research problems and was studied as early as 1976 [4], which is an era when machine learning was not yet discovered. At that time, AQG was carried out in a rule-based manner. Now, most of the existing works in AQG can be classified as the following approaches with a typical example.

2.1.1. Template-Based Approach

In 2010, Kalady et al. [15] defined a way to generate definitional questions, which are questions that ask the definition of a selected term, which is called an "up-key". The "up-key" can be found by statistical methods such as the occurrence of terms. By recognizing the named entity of the "up-key", a question word can be chosen. The "up-key" and question word are then substituted into the template "<Question word> is <up-key>?" to form a definitional question.

Definition question generation generates questions on an article basis. It is useful for generating questions for expository essays, or postreading reflection questions to lead readers to think. However, the answer to a generated question may not appear in the source article. Consider a news report on an earthquake. One possible generated question could be "What is an earthquake?" The answer is usually outside the scope of the news article and cannot be found.

2.1.2. Syntax-Based and Semantic-Based Approach

Understanding the text is one of the key ways to make good questions. A parse tree represents a text in a tree structure with a syntactic structure. By investigating the structure, scientists can decide on rules to generate questions. Heilman [16] generated factoid questions by changing and replacing the structure of the parse tree with the use of tree regular expression manipulation.

By further exploring the parse trees, semantics inside the text can be found and higher-level questions such as reasoning questions can be generated. However, the quality is greatly depending on the tree parser and the writing style of the text.

2.1.3. Sequence-to-Sequence (seq2seq) Approach

Question generation is recognized as a text-generation task, where the input and output text are sequences of words. The process is defined as a sequence-by-sequence task by Google in 2014 [17]. In the process, the input text, together with its features, is decom-

posed into a vector of numbers in an encoder network. The output from a machine learner, which is also a vector of numbers, will be decoded into text by a corresponding decoder.

2.1.4. Transformer Approach

In most NLP tasks, there are some common subtasks such as extracting language features such as the part of speech and semantic meaning of words. A transformer-based model, which is usually a general-purpose language model, will undergo a pre-training phase that aims to learn the better representation of languages with an extremely high number of corpus. Then, the pre-trained model is fine-tuned with a task-specific dataset to solve the target problem.

To save the long training process and high-computational power, developers reuse the pre-trained model for different tasks. This is a transfer-learning process. Examples of popular pre-trained models are the bidirectional encoder representations from transformers (BERT) [18] and text-to-text transfer transformers (T5) [19].

3. Materials and Methods

3.1. Dataset Preparation

The dataset was built from three datasets for question answering (QA) tasks, namely the CRMC2017 [7], CRMC2018 [8], and DRCD [9].

3.1.1. CRMC2017 Dataset

This dataset was used in a machine reading comprehension competition in 2017. The dataset consists of 5000 entries of pre-tokenized passages, questions, and one corresponding answer in simplified Chinese. Most passages are narrative stories, and a few are explanatory texts. All questions are wh-questions with a majority of "who", "where", "when" and "which". Answers can be extracted from the passages. The level of the passages and questions is suitable for primary students.

3.1.2. CRMC2017 Dataset

Used in a machine reading comprehension competition in 2018, this dataset contains 14,363 question–answer pairs, where some pairs share a common passage from Wikipedia in simplified Chinese. On top of the wh-questions, "why" and "how" questions are also included. Each question in the dataset comes with three possible answers. The three answers are extracted from the passage, with minor differences in their length. In our data preparation, we select the first answer. This dataset can be categorized into secondary levels.

3.1.3. DRCD Dataset

This is the first machine reading comprehension in traditional Chinese. It shares a similar structure as the CRMC2017 dataset. A total of 40,970 question–answer pairs are included. We translated the passages, questions, and answers into simplified Chinese in our dataset preprocessing.

3.2. Encoder and Decoder

We adopted the Google mT5 model in the encoding and decoding processes. Figure 1 shows the flow of the machine learning process. We compressed the passage and answer in a form of "context: <passage>, answer: <answer>". This form was encoded with the mT5 pre-trained model. Together with the attention vector extracted, they were fed into the machine learner for training.

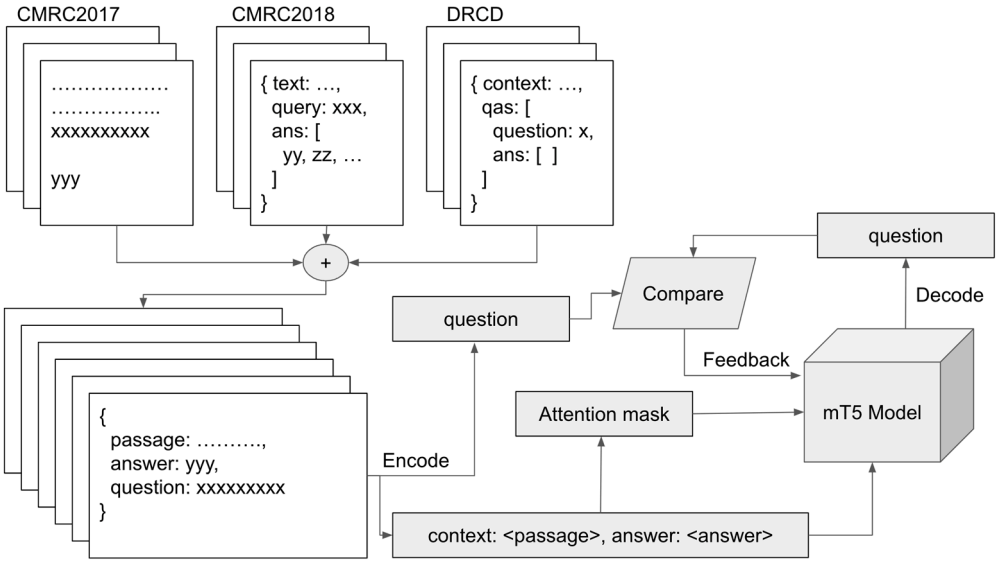

Figure 1. The flow of the machine learner: dataset preparation, encoder, decoder, and transformer.

3.3. Application Design

The question generation system is presented in a web-based application. Users are required to input a passage in the first step (Figure 2). The passage can be in simplified Chinese or traditional Chinese. Then, the system will annotate some of the possible answer phrases to generate questions (Figure 3). Users can modify the default choices to fit their needs. Users can click on the red cross button to remove the pre-annotated phrase. User can highlight the answer phrase and a green plus button will appear (Figure 4). The user can add their answer phrase with the button (Figure 5). When the user finishes their annotation and proceeds, the passage with a list of selected answer phrases will be sent to the server side (Figure 6). The fine-tuned model will produce the question set, which will be sent to the client side. The client-side script will display the questions to the user (Figure 7). The user can remove any unwanted questions with the "remove" button. They can hide the answer by toggling the checkbox and copying the content to the clipboard by clicking the "copy" button.

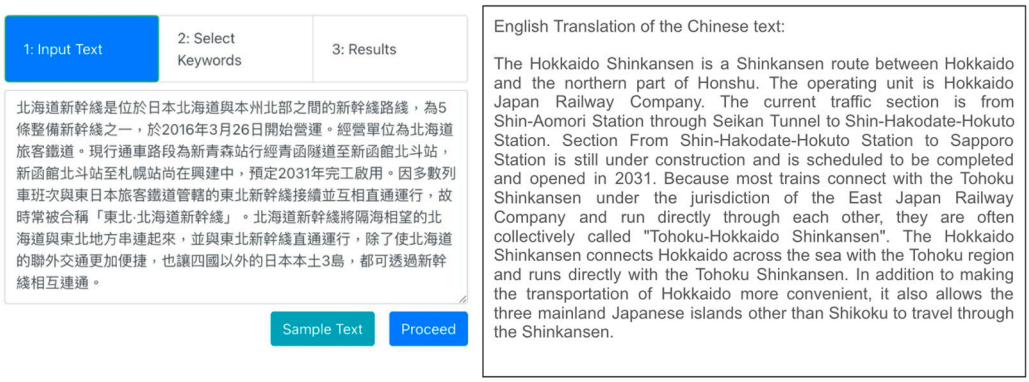

Figure 2. Step 1: user needs to input a passage.

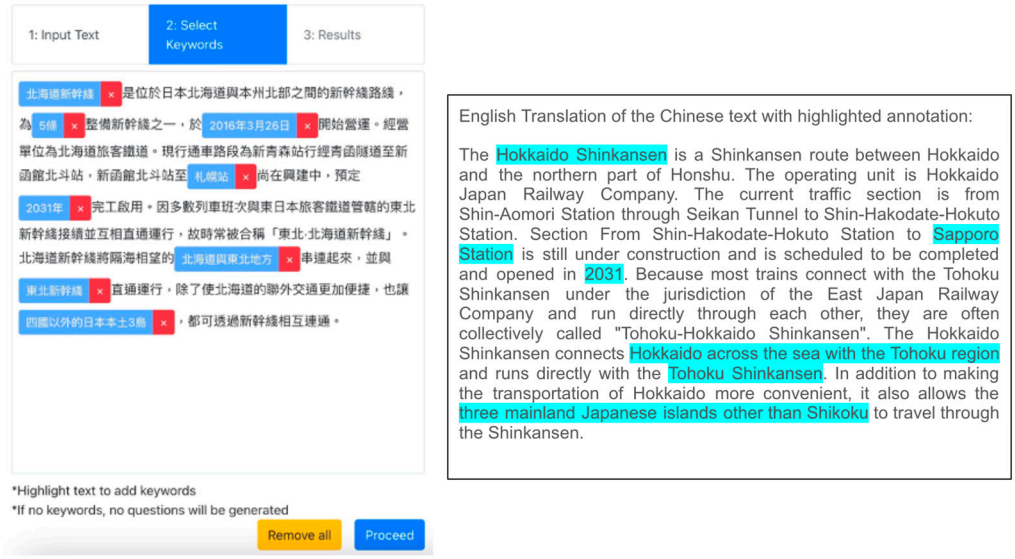

Figure 3. Step 2: possible answer phrases are pre-annotated.

Figure 4. Step 2: the user can edit the annotation by removing or adding new phrases.

Figure 5. Step 2: the pre-selected phrase is replaced by the user annotation.

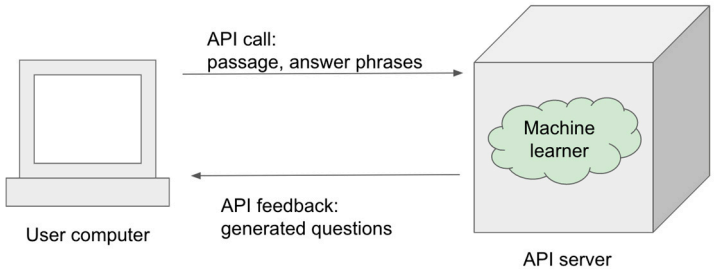

Figure 6. API communication between the user side and the API server.

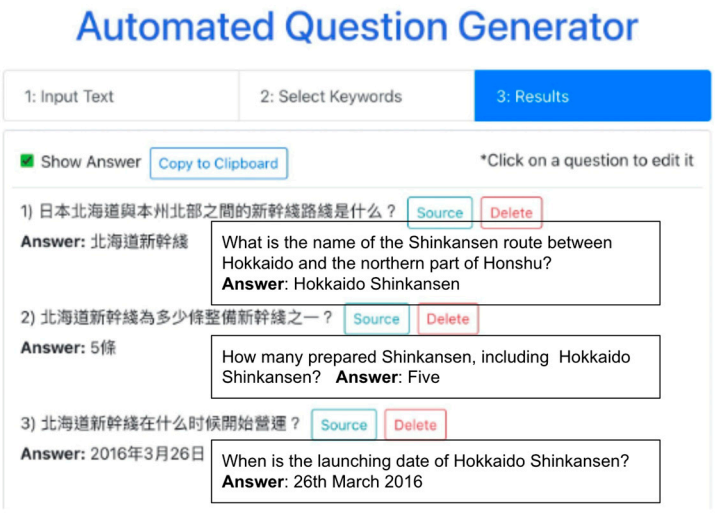

Figure 7. Step 3: questions are generated and displayed on the user screen.

4. Preliminary Results

We performed three preliminary tests on our application.

4.1. Response Time

We performed a preliminary test on the time required to generate questions using our web application. We selected four texts in traditional Chinese and one text in simplified Chinese from different categories. The text and the English version are included in Appendix A. The number of words and the number of questions from each text are shown in Table 1.

Table 1. Result of response-time test.

	No. of Words	No. of Questions	Time Used (sec)	Time Used per Questions (sec)
Text 1	305	13	6.73	0.518
Text 2	296	7	4.14	0.591
Text 3	371	11	4.96	0.451
Text 4	243	10	5.02	0.502
Text 5	223	29	13.17	0.454
Mean average time for one question (seconds)				0.486

In our experiment, we annotated the answer phrases instead of choosing the pre-selected ones. This was to ensure there were enough questions generated for comparison. The pre-trained model was located on a server with two GPUs (RTX 3090). From the result, it was found that the time to generate questions was not dependent on the length of the text. It depended on the number of answer phrases marked. On average, it took half a second to generate a question.

4.2. User Survey

We invited 20 computer science students from Hong Kong Metropolitan University to test the system and complete a survey. There were six questions on the usefulness, usability, and look-and-feel of the system in the survey. The five-point Likert scale (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree) was used. Table 2 shows the results.

Questions 1 and 4 were about the quality of the generated questions and the expectations for the system. The responses showed that more than half of the students disagreed with the quality of the generated questions. Hence, they concluded that the system worked out their expectations. There is great room for improvement. In the next subsection, we investigate the quality of the generated questions.

Table 2. Results of the user survey.

	Question	Percentage (%)				
		1	**2**	**3**	**4**	**5**
1	I am satisfied with the quality of questions.	35.00	30.00	20.00	10.00	5.00
2	I think the system is easy to use.	0.00	10.00	15.00	45.00	30.00
3	I think the system is easy to learn.	10.00	5.00	30.00	20.00	35.00
4	The system works the way I expected.	30.00	35.00	15.00	15.00	5.00
5	I like the user interface of the system.	5.00	5.00	40.00	30.00	20.00
6	Overall, I am satisfied with the system.	15.00	35.00	30.00	15.00	5.00

Questions 2, 3, and 5 asked about the usability of our system. The interviewees agreed that our system was easy to use. The interface provided sufficient information to first-time users. They could easily generate questions without seeking aid from the developers. The interviewees feedbacked that the ability to go back to previous stages was a friendly design feature for them to annotate new keywords in the passage. The source button was also helpful for them to reference what they had annotated without switching backward.

Despite the usability of the system, interviewees were not satisfied due to the poorly generated questions.

4.3. Quality of Generated Question

We investigated the varieties of questions that could be generated by the system using the text in Appendix A. Table 3 shows some of the acceptable outputs of our system. The system could generate "What", "Who", "When", "Where", "Which", "Why" and "How" -type questions when we annotated the answer phrases.

Yet, defective questions were also generated. Table 4 shows a list of the defects and their descriptions. In general, the machine learner faced a major difficulty in the knowledge of unseen phrases. The Chinese language does not have a similar concept of letters and words in English. Before encoding the text, it is necessary to perform tokenization to separate phrases into standalone tuples. It becomes a blind guess when the tokenizer has not encountered the phrase. Even if the tokenization is correct, the part of speech and named entity will be other guesses in the pipeline architecture. As Chinese sentences are usually long, the attention mechanism is important to force the focus on the important part of the text. However, it is still relying on the tokenization accuracy.

Table 3. Results of question-type test.

Type		
What	Content	違例駕駛記分制度是為了改善道路安全而設的。 The DOPs system is a measure designed to promote safety on the road.
	Answer phrase	為了改善道路安全 to promote safety on the road
	Generated question	違例駕駛記分制度的目的是什麼? What is the purpose of the DOPs system?
Who	Content	外交部发言人汪文斌17 日主持例行记者会, Foreign Ministry Spokesperson Wang Wenbin hosted a regular press conference on the 17th
	Answer phrase	汪文斌 Wang Wenbin
	Generated question	哪一个人主持例行记者会? Who hosted a regular press conference?
When	Content	「安心出行」自2020 年11 月啟用至今, Since the launch of "Leave Home Safe" in November 2020,
	Answer phrase	2020 年 The year 2020
	Generated question	「安心出行」什麼時候啓用至今? When did "Leave Home Safe" launch?

Table 3. *Cont.*

Type		
Where	Content	都大物理治療中心, 將於 2022 年 12 月 1 日在賽馬會校園盛大開幕。 The grand opening of the HKMU Physiotherapy Centre Limited in the Jockey Club Campus will be on 1 December 2022.
	Answer phrase	賽馬會校園 Jockey Club Campus
	Generated question	都大物理治療中心在哪兒舉行開幕? Where will be the opening ceremony of HKMU Physiotherapy Centre Limited?
Which	Content	醫務衛生局局長盧寵茂表示, Lo, Director of the Medical and Health Bureau, said
	Answer phrase	醫務衛生局 Medical and Health Bureau
	Generated question	盧寵茂是哪一個單位的負責人? Which unit is Lo in charge of?
Why	Content	疫苗通行證仍是重要措施, 包括可鼓勵市民加速接種疫苗, 因此市民要進入一些處所仍要出示疫苗通行證 The vaccine pass is still an important measure, including encouraging citizens to speed up vaccination, so citizens still have to show the vaccine pass to enter some premises
	Answer phrase	可鼓勵市民加速接種疫苗 encouraging citizens to speed up vaccination.
	Generated question	爲何市民要進入一些處所仍要出示疫苗通行證? Why are citizens required to show their vaccine pass?
How	Content	北海道新幹綫無法與東海道新幹綫、山陽新幹綫、九州新幹綫直通運行，需在東京站轉乘。 The Hokkaido Shinkansen cannot run directly with the Tokaido Shinkansen, Sanyo Shinkansen, and Kyushu Shinkansen, and needs to transfer at Tokyo Station.
	Answer phrase	需在東京站轉乘 Transfer at Tokyo Station.
	Generated question	北海道新幹線如何轉乘東海道新幹線、山陽新幹線? How to transfer Hokkaido Shinkansen to Tokaido Shinkansen and Sanyo Shinkansen?

Table 4. Result of some generated questions.

Defect		
Wrong question word	Content	都大物理治療中心, 將於 2022 年 12 月 1 日在賽馬會校園盛大開幕。 The grand opening of the HKMU Physiotherapy Centre Limited in the Jockey Club Campus will be on 1 December 2022.
	Answer phrase	都大物理治療中心 HKMU Physiotherapy Centre Limited
	Generated question	哪一個實驗室將於2022 年12 月1 在賽馬會校園盛大開幕? Which laboratory will be opened on 1st December 2022?
	Explanation	The machine learner misunderstood that the HKMU Physiotherapy Centre Limited is a laboratory.

Table 4. *Cont.*

Defect		
Incorrect answer	Content	現行通車路段為新青森站行經青函隧道至新函館北斗站, 新函館北斗站至札幌站尚在興建中, 預定2031 年完工啟用。 The current traffic section is from Shin-Aomori Station through Seikan Tunnel to Shin-Hakodate-Hokuto Station. The section from Shin-Hakodate-Hokuto Station to Sapporo Station is still under construction and is scheduled to be completed and opened in 2031.
	Answer phrase	新函館北斗站至札幌站 From Shin-Hakodate-Hokuto Station to Sapporo Station
	Generated question	現行通車路段爲新青森站行經青函隧道至哪站? Which station the current traffic section is from Shin-Aomori Station through Seikan Tunnel to?
	Explanation	The source is too long. The machine learner is not able to realize that the former part and the later part of the section are under different construction stages.
Vague	Content	駕駛人如在兩年內被記滿 15 分或以上, 法庭可以取消他持有或領取駕駛執照的資格一段時期。 If 15 or more DOPs have been incurred within 2 years, the driver can be disqualified by a court from holding or obtaining a driving license for a certain period.
	Answer phrase	15 分或以上 15 or more DOPs
	Generated Question	駕駛人若在兩年內被記滿多少? How many DOPs have been incurred within 2 years?
	Explanation	The implication part of the content is not included in the question, leading it to be vague.
Does not make sense	Content	醫務衛生局局長盧寵茂表示, 當局決定撤銷市民掃瞄「安心出行」的相關法例及行政安排 Lo, director of the Medical and Health Bureau, said that the authorities have decided to revoke the relevant laws and administrative arrangements for citizens to scan "Leave Home Safe".
	Answer phrase	撤銷市民掃瞄「安心出行」的相關法例及行政安排 revoke the relevant laws and administrative arrangements for citizens to scan "Leave Home Safe"
	Generated Question	盧寵茂爲何要撤銷市民掃瞄「安心出行」的相關法例及行政安排? Why Lo has decided to revoke the relevant laws and administrative arrangements for citizens to scan "Leave Home Safe"
	Explanation	The machine learner copied the answer once as the question.
Ungrammatical	Content	北海道新幹綫除了使北海道的聯外交通更加便捷, 也讓四國以外的日本本土3 島, 都可透過新幹綫相互連通。 In addition to making the transportation of Hokkaido more convenient, it also allows the three mainland Japanese islands other than Shikoku to travel through the Shinkansen.
	Answer phrase	四國以外的日本本土3 島 the three mainland Japanese islands other than Shikoku
	Generated Question	北海道新幹線除了使北海道的聯外交通更加便捷還有什麼? What else does the Hokkaido Shinkansen do besides making transportation in Hokkaido more convenient?
	Explanation	The machine learner does not know what the three mainland Japanese islands are, other than Shikoku.

5. Discussion

An easy-to-use, web-based system embedded with an automatic question generator for Chinese reading comprehension was built and received a preliminary evaluation. Users could generate wh-questions by inputting Chinese passages into the system. The analysis revealed that the question generator achieved a satisfactory response time of about half a

second per question. However, the survey conducted concluded that the questions generated with the current system were not satisfactory. There is room for improvement in future work from two perspectives.

First, a more accurate machine learner should be trained. In this work, the model was fine-tuned from the mT5 pre-trained model. Although the pre-trained model contains rich language features, the absence of knowledge of the unseen phrases is critical when the machine learner is supposed to understand the phrases and make good questions about them, considering the quality of the generated questions. Existing works could be an alternative underlying AQG component to the web application.

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 question types, such as parts of speech and open-ended questions, should 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, more question types could be included in the generation process. Wh-questions are not the only question type seen in the education system. Common types such as multiple-choice questions and reflective questions are possible.

In future work, a systematic evaluation will be performed. Prevailing automatic evaluation metrics for NLP, such as BLEU, and ROUGE-L, can be applied to evaluate the accuracy of the machine learner. More users from different knowledge backgrounds, such as secondary school teachers, students, and users with limited computer skills, could be invited to test the application. A statistical power and significance test on the results could be achieved.

Looking to the future, we hope that AQG can be greatly helpful to the education industry. Teaching materials can be generated with less effort. An interactive system embedded with AQG could hopefully motivate students' intention to read. We look forward to the further development of AQG.

Author Contributions: Conceptualization, Y.-C.F. and L.-K.L.; methodology, Y.-C.F.; software, Y.-C.F.; validation, Y.-C.F., L.-K.L. and K.T.C.; formal analysis, Y.-C.F. and L.-K.L.; investigation, Y.-C.F.; resources, L.-K.L.; data curation, L.-K.L. and K.T.C.; writing—original draft preparation, Y.-C.F.; writing—review and editing, L.-K.L. and K.T.C.; visualization, Y.-C.F.; supervision, L.-K.L.; project administration, L.-K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Test Data

Appendix A.1.1. Text 1

Original text: 北海道新幹綫是位於日本北海道與本州北部之間的新幹綫路綫, 為5 條整備新幹綫之一，於2016 年3 月26 日開始營運。經營單位為北海道旅客鐵道。現行通車路段為新青森站行經青函隧道至新函館北斗站，新函館北斗站至札幌站尚在興建中, 預定2031年完工啟用。因多數列車班次與東日本旅客鐵道管轄的東北新幹綫接續並互相直通運行, 故時常被合稱「東北· 北海道新幹綫」。北海道新幹綫將隔海相望的北海道與東北地方串連起來, 並與東北新幹綫直通運行, 除了使北海道的聯外交通更加便捷, 也讓四國以外的日本本土3 島, 都可透過新幹綫相互連通, 但無法與東海道新幹綫、山陽新幹綫、九州新幹綫直通運行, 需 在東京站轉乘，九州新幹綫則需在新大阪站或博多站轉乘。

Translated text: The Hokkaido Shinkansen is a Shinkansen route between Hokkaido and the northern part of Honshu. The operating unit is the Hokkaido Japan Railway

Company. The current traffic section is from Shin-Aomori Station through the Seikan Tunnel to Shin-Hakodate-Hokuto Station. The section from Shin-Hakodate-Hokuto Station to Sapporo Station is still under construction and is scheduled to be completed and opened in 2031. Because most trains connect with the Tohoku Shinkansen under the jurisdiction of the East Japan Railway Company and run directly alongside each other, they are often collectively called "Tohoku-Hokkaido Shinkansen". The Hokkaido Shinkansen connects Hokkaido across the sea with the Tohoku region and runs directly with the Tohoku Shinkansen. In addition to making transportation to Hokkaido more convenient, it also allows the three mainland Japanese islands other than Shikoku to travel through the Shinkansen. However, they cannot run directly with the Tokaido Shinkansen, Sanyo Shinkansen, and Kyushu Shinkansen. They need to transfer to Tokyo Station. For the Kyushu Shinkansen, they need to transfer to Shin-Osaka Station or Hakata Station.

Appendix A.1.2. Text 2

Original text: 醫務衞生局局長盧寵茂表示, 當局決定撤銷市民掃瞄「安心出行」的相關法例及行政安排, 場所負責人亦不用再張貼二維碼, 有關安排明天開展。他說,「安心出行」自2020 年11 月啟用至今, 發揮相重要的抗疫角色, 便利市民記錄行程, 亦可幫助個案追蹤。他又說, 疫苗通行證仍是重要措施, 包括可鼓勵市民加速接種疫苗, 因此市民要進入一些處所仍要出示疫苗通行證, 但強調與「安心出行」不是絕對掛勾, 市民在「安心出行」程式右下角有疫苗通行證, 同時亦可以用「智方便」或「醫健通」出示疫苗通, 入境旅客抵港時有臨時紙本疫苗通, 如果長者等人士無手機程式, 亦可出示紙本, 出示手機截圖亦可以, 指市民有不同方法出示疫苗通行證進入指定處所。

Translated text: Lo, director of the Medical and Health Bureau, said that the authorities have decided to revoke the relevant laws and administrative arrangements for citizens to scan "Leave Home Safe", and the person in charge of the venue no longer needs to post the QR code. The relevant arrangements will start tomorrow. He said that since its launch in November 2020, "Leave Home Safe" has played an important role in fighting the epidemic. It is convenient for citizens to record their itineraries and can also help with case tracking. He also said that the vaccine pass is still an important measure, including encouraging citizens to get vaccinated quickly. Therefore, citizens still have to show the vaccine pass when entering some places, but he emphasized that it is not linked to "Leave Home Safe". There is a vaccine pass in the lower right corner. At the same time, you can use "iAM Smart" or "eHealth" to show the vaccine pass. Inbound passengers will have a temporary paper vaccine pass when they arrive in Hong Kong. If the elderly and other people do not have a mobile phone program, they can also show a paper copy. It is also acceptable to show a screenshot of the mobile phone, which means that citizens have different ways to show the vaccine pass to enter the designated premises.

Appendix A.1.3. Text 3

Original text: 各位同學, 我們很高興地宣布大學的全資附屬公司, 都大物理治療中心, 將於 2022 年 12 月 1 日在賽馬會校園盛大開幕。都大物理治療中心致力於提供各類優質物理治療服務, 包括肌肉骨骼及運動物理治療、神經康復、女性健康和淋巴水腫管理等。中心還致力於為本校修讀物理治療的學生提供早期學習環境和臨床培訓。都大物理治療中心將於 2022 年 12 月 5 日正式運作。其主要服務對象是大學成員(包括理事會成員、員工、學生、退休人員及校友) 和公眾。大學全日制學生每次於中心接受45 分鐘理療諮詢治療, 均可享 30% 折扣優惠。中心將於2023 年推出多項活動, 例如網上研討會、健康講座及健康教育意識訓練班。此外, 專業資訊文章和資源也將適時經中心網站分享。都大物理治療中心的成立將為護理及健康學院和大學樹立新的里程碑。請與我一起祝賀該中心的成立, 並給予大力支持。

The counterpart in English: Dear students, we are excited to announce the grand opening of the HKMU Physiotherapy Centre Limited (HKMUPC), which is a wholly owned subsidiary of Hong Kong Metropolitan University (HKMU), at the Jockey Club Campus on 1 December 2022. HKMUPC is dedicated to delivering a variety of quality physiotherapy services, namely, musculoskeletal and sports physiotherapy, neurological rehabilitation, women's health, and lymphedema management, etc. We are also committed to providing

early learning exposure and clinical training to our physiotherapy students. HKMUPC will be officially operating on 5 December 2022. Our main clientele are the HKMU Community (including Council members, staff, students, retirees, and alumni) and the general public. The University's full-time students can enjoy a discount rate of 30% for a 45 min physiotherapy consultation treatment session. Several activities such as online webinars, health seminars, and exercise classes for health educational awareness will be launched in the coming year, 2023. Furthermore, professional informative articles and resources will also be shared on the HKMUPC website in due course. The establishment of the HKMUPC will set a new milestone for both the School of Nursing and Health Studies and the University. Please join me in congratulating the opening of HKMUPC and lend it your generous support.

Appendix A.1.4. Text 4

Original text: 違例駕駛記分制度是為了改善道路安全而設的。其主要目的是阻嚇經常違反交通規例的人士和提高駕駛水平, 從而減少交通意外的發生。除了原有懲罰外, 某些交通違例事項還會有違例駕駛記分, 觸犯這些表列違例事項將會被記相應的分數。駕駛人如在兩年內被記滿 15 分或以上, 法庭可以取消他持有或領取駕駛執照的資格一段時期。駕駛人如觸犯任何表列違例事項, 而就該事項被法庭定罪; 或而負上繳付定額罰款的法律責任, 即須在違例該日被記分。有關違例駕駛記分制度、表列違例事項及其相應分數的詳情, 請參閱「更多參考資料」。

The counterpart in English: The DOPs system is a measure designed to promote safety on the road. The main purposes are to deter habitual traffic offenders and to improve the standards of driving to reduce the accident toll. Certain traffic offenses will carry driving-offense points, in addition to other penalties. Committing any of these scheduled offenses will result in the recording of the corresponding points. If 15 or more DOPs have been incurred within 2 years, the driver may be disqualified by a court from holding or obtaining a driving license for a certain period. Any scheduled offense committed as from that date attracts DOPs if the driver is convicted by a court, or becomes liable to a fixed penalty for that offense. For more information on the DOPs system, scheduled offenses, and their corresponding DOPs, see "Further Reference Materials".

Appendix A.1.5. Text 5

Original text: 外交部发言人汪文斌17 日主持例行记者会, 共同社记者提问说, 据美国媒体报道, 美国国务卿布林肯将于2 月5 日访问中国, 与中国外交部部长秦刚举行会晤。汪文斌表示, 中方欢迎布林肯国务卿访华。中美双方正就有关具体安排保持沟通。中方始终按照习近平主席提出的相互尊重、和平共处、合作共赢三原则看待和发展中美关系, 也希望美方树立正确的对华认知, 坚持对话而非对抗, 双赢而非零和, 同中方相向而行, 不折不扣落实两国元首达成的重要共识, 推动中美关系重回健康稳定发展轨道。

Translated text: Foreign Ministry Spokesperson Wang Wenbin presided over a regular press conference on the 17th. A reporter from Kyodo News said that according to US media reports, US Secretary of State Blinken will visit China on February 5 and meet with Chinese Foreign Minister Qin Gang. Wang Wenbin said that China welcomes Secretary of State Blinken's visit to China. China and the United States are maintaining communication on relevant specific arrangements. China always views and develops Sino–U.S. relations following the three principles of mutual respect, peaceful coexistence, and win–win cooperation, as proposed by President Xi Jinping. It also hopes that the U.S. will establish a correct understanding of China, and insist on dialogue rather than confrontation, and win–win rather than zero-sum, meeting each other halfway, fully implementing the important consensus reached by the two heads of state, and pushing China–US relations back on the track of healthy and stable development.

References

1. Sadiku, L.M. The importance of four skills reading, speaking, writing, listening in a lesson hour. *Eur. J. Lang. Lit.* **2015**, *1*, 29–31. [CrossRef]
2. Walton, E. The Language of Inclusive Education: Exploring Speaking, Listening, Reading and Writing. Routledge: London, UK, 2015.
3. 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.* **2002**, *94*, 56. [CrossRef]
4. Wolfe, J.H. Automatic question generation from text-an aid to independent study. In Proceedings of the Acm Sigcse-Sigcue Technical Symposium on Computer Science and Education, Anaheim, CA, USA, 12–13 February 1976; pp. 104–112.
5. Mitkov, R.; Le An, H.; Karamanis, N. A computer-aided environment for generating multiple-choice test items. *Nat. Lang. Eng.* **2006**, *12*, 177–194. [CrossRef]
6. 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.* **2020**, *30*, 121–204. [CrossRef]
7. Cui, Y.; Liu, T.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. Dataset for the First Evaluation on Chinese Machine Reading Comprehension. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
8. Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. *arXiv* **2018**, arXiv:1810.07366.
9. Shao, C.C.; Liu, T.; Lai, Y.; Tseng, Y.; Tsai, S. Drcd: A chinese machine reading comprehension dataset. *arXiv* **2018**, arXiv:1806.00920.
10. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computer Linguistics: Human Language Technologies, 6–11 June 2021; pp. 483–498.
11. Ranaldi, L.; Fallucchi, F.; Zanzotto, F.M. Dis-Cover AI Minds to Preserve Human Knowledge. *Future Internet* **2021**, *14*, 10. [CrossRef]
12. Lin, H.; Zhang, P.; Ling, J.; Yang, Z.; Lee, L.K.; Liu, W. PS-Mixer: A Polar-Vector and Strength-Vector Mixer Model for Multimodal Sentiment Analysis. *Inf. Process. Manag.* **2023**, *60*, 103229. [CrossRef]
13. Lee, L.K.; Chui, K.T.; Wang, J.; Fung, Y.C.; Tan, Z. An Improved Cross-Domain Sentiment Analysis Based on a Semi-Supervised Convolutional Neural Network. In *Data Mining Approaches for Big Data and Sentiment Analysis in Social Media*; IGI Global: Hershey, PA, USA, 2022; pp. 155–170.
14. Fung, Y.C.; Lee, L.K.; Chui, K.T.; Cheung, G.H.K.; Tang, C.H.; Wong, S.M. Sentiment Analysis and Summarization of Facebook Posts on News Media. In *Data Mining Approaches for Big Data and Sentiment Analysis in Social Media*; IGI Global: Hershey, PA, USA, 2022; pp. 142–154.
15. Kalady, S.; Elikkottil, A.; Das, R. Natural language question generation using syntax and keywords. In *Proceedings of QG2010: The Third Workshop on Question Generation*; Questiongeneration. Org: Pittsburgh, PA, USA, 2010; pp. 5–14.
16. Heilman, M.; Smith, N.A. *Question Generation Via Overgenerating Transformations And Ranking*; Carnegie-Mellon University Pittsburgh pa Language Technologies InsT: Pittsburgh, PA, USA, 2009.
17. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
18. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
19. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.