JEFF - Just Another EFFicient Reading Comprehension Test Generation

Anonymous ACL submission

Abstract

We introduce a method for generating vocabulary questions on reading comprehension of a given English article. In our approach, the method involves selecting target words in the given English article, finding synonyms as answer keys, and generating seemingly reasonable words in context as distractors. At runtime, some target words in the inputted article will be identified as questions, and automatically generating one answer key and three distractors. We present a AQG (automatic question generation) system, JEFF, that applies the method to generate questions automatically. Evaluation on a set of questions generated by JEFF shows that the method is close to the human-designed ones.

1 Introduction

Many reading comprehension vocabulary tests (e.g. *The **atmosphere** on Mars was thicker and liquid water widespread. Question: The word **atmosphere** here is A. ice B. air C. soil D. crust*) are widely used in text books and standardized tests. For example, Vocabulary Workshop[1] used this kind of practice to familiarize English learners with the meaning of target words, while TOEFL [2] used to assess language learners' proficiency in English.

In this paper, we generate the vocabulary type test questions of reading comprehension.[3] This type of question is a four-choice multiple choice question. It have three main elements, which are target words, answer keys and distractors. Target words are the words in the article which is very difficult to test takers, so test takers don't know their meanings and only can conjecture them by contexts. Answer keys are the words have the same meaning as corresponding target words, which are correct answers of the questions. As for distractors, they should be the wrong options and the main object of distractors is attracting test takers to choose them but they can't be possible to be answer keys.

Vocabulary test items of reading comprehension such as Vocabulary Workshop and TOEFL are typically manually designed. However, it is almost impossible to generate suitable hand-crafting test items on demand for user-selected online reading materials. Such test items could be generated automatically and on the fly if we generate paraphrases of the target word (e.g., atmosphere) in the stem sentence and use these paraphrases as an answer key (e.g., air) and distractors (e.g., ice, soil, and crust).

Consider the sentence "The atmosphere on Mars was thicker and liquid water widespread." of a given article. The best target word for this sentence is probably not "water" for B2 level learners, but rather "atmosphere" or "widespread". A good answer key might be "air" with the similar meaning to "atmosphere". Proper distractors might be "ice", "soil" or "crust", because they are reasonable in phrasal context, but impossible in sentential context. Intuitively, by paraphrasing the target word in sentential and phrasal contexts, these answer key and distractors the can be retrieved.

We present a new system, JEFF, that automatically generates items of vocabulary tests on reading comprehension. An example JEFF generates a test item is shown in Figure 1. JEFF has found a target word (e.g., atmosphere) as questions and determined an answer key and three distrators. JEFF generates these options automatically by using a pre-trained language model with contexts of various width to predict replacements. We describe the JEFF question-generating process in more detail in

[1] It was published by William H. Sadlier. See https://www.sadlier.com/

[2] Test Of English as a Foreign Language, http://www.ets.org/toefl/

[3] According to The Official Guide to the TOEFL Test (https://www.ets.org/toefl/test-takers/ibt/prepare/guides-books/), the vocabulary type test questions of reading comprehension account for 25%-50% of TOEFL reading test, which is the highest proportion question type in the TOEFL reading test.

Running Water on Mars?

Photographic evidence suggests that liquid water once existed on the surface of Mars. Two types of flow features are seen: runoff channels and outflow channels. Geologists think that they are dried-up beds of long-gone rivers. Runoff channels on Mars speak of a time 4 billion years ago, when the **atmosphere** on Mars was thicker and liquid water widespread.

Question

The word **atmosphere** in the passage is closest in meaning to

A. ice
B. air
C. soil
D. crust

Figure 1: A sample question generated by JEFF.

Section 3.

At run-time, JEFF starts with an English article and the level[4] of the test taker submitted by the user, and selects proper target words, and then generates answer keys and distractors for each target word. In our prototype, JEFF returns the questions to the user directly (see Figure 1), alternatively the test items can be presented to a teacher for making exam papers by identifying ideal ones.

The rest of the article is organized as follows. We review the related work in the next section. Then, we present our method for automatically finding proper target words and corresponding answer keys and distractors(Section 3). As part of our evaluation, we use indicators, including discrimination and redability, which are usually used in the tests evaluation(Section 4). We obtained astonishing result(Section 5), which shows the machine-generating questions almost meet the standard of ideal tests.

2 Related Work

Automatic question generation (AQG) has been an area of active research. Recently, the state-of-the-art in AQG research has been represented in Ch and Saha (2018) and Kurdi et al. (2020), which involves AGQ papers for educational purposes. In our work we address an aspect of multiple choice question generation, that was direct focus of some papers reviewed by Kurdi et al. (2020). We also consider levels of test takers, which tend to be different for test takers, and teachers might be interested in tailored levels of the test to the students.

More specifically, we focus on automatically generating multiple choice question, namely, deciding target words of a given article and generating answer keys and distractors. Generating distractors has long been an active topic of AQG research. An interesting approach presented by Goto et al. (2010) describes how to generate distractors by investigating existing human-designed tests. In general, this system generated distractors which is similar to existing tests based on the question patterns. For example, pattern "Interrogative" may generate "which", "what", "who", "when", or "where" as distrators. In contrast, we will show how to generate distractors not based on existing tests because some patterns like vocabulary will have many possibilities, and there will be few existing tests for most of vocabularies.

Most of the automatic vocabulary question generation systems are choosing distractors by existing database or dictionaries [Sumita et al. (2005)]. Similarly, Wita et al. (2018) describes a method for utilizing Japanese WordNet to generate distractors. This approach is not optimal because distractors generated by this method doesn't refer to the contexts of the target words and this method doesn't consider levels of vocabularies. However, the user (e.g., the teacher) is actually looking for distractors which almost make sense with the context. Generating context-related distractors is particularly important when making test takers hard to choose the answer key and help them force them to realize the given article.

Some of the traditional AQG systems have begun to generate distractors by large corpora. Brown et al. (2005) describes a system that generates distractors by pre-processing BNC[5]. The table used for generating distractors is built counting the frequencies of all words, because Coniam (1997) say distractors with similar frequency to the answer key are ideal. Additionally, the Liu et al. (2005) generated distractors based on the frequency and the collocation. However, this approach has not been

[4] In our system, we use Common European Framework of Reference(CEFR) level (https://www.eur.nl/en/education/language-training-centre/cefr-levels) as input.

[5] British National Corpus, which is a 100-million-word text corpus of samples of written and spoken English

shown to significantly make test takers confused if they don't understand the given article enough.

Recent work has been done on using word embedding with the goal of selecting distractors similar to the target word. For example, an embedding-based approach is described in Jiang and Lee (2017). Aldabe and Maritxalar (2010) show how to use a method almost identical to word embedding to create distrators in Basque. These methods considering meanings of words have a great breakthrough on generating distractors. However, it's hard to define how close a relationship is an appropriate distractor. If the meaning of the distractor is too similar to the target word, there will be more than two answer keys.

Recently, Sakaguchi et al. (2013) independently presented a method for producing distractors based on the context of the target word. This work took four words beside the target word as features. In contrast to our work, the length of the context this work reflecting on is too short.

In a study more closely related to our work, Susanti et al. (2015) generated vocabulary questions of reading comprehension but rather cloze tests. The main difference from cloze tests is that the cloze tests utilizing the target word as answer key. Hence, cloze tests are unnecessary to produce an extra answer key. In contrast, vocabulary questions of reading comprehension need to generate an answer key whose meaning is the same as the target word. However, Susanti et al. (2015) still used WordNet to find answer keys by looking up synonyms and distractors by words with the same hypernym. Again, this method didn't consider the context.

In contrast to the previous research in generating distractors, we present a system that generates distractors related to the large scope of the context. In addition, we have one more step to generate test items. We need to produce an answer key. We exploit the power of deep learning by paraphrasing the target word of the given article.

3 THE JEFF system

Manually generating test items (e.g., The test items for *The **atmosphere** on Mars was thicker and liquid water widespread.*) for the user-selected materials is almost impossible. Human-designed test items for one user is uneconomic. To generate test items for the user-selected materials, a promising approach is to automatically generate test items.

3.1 Selecting Target Words

In the first stage, we select a set of words in the given article. These words are functioned as target words which will be shown in the questions (e.g., In "The word atmosphere in the passage is closest in meaning to", "atmosphere" is the target word). One target word will generate one test item later. The input of this stage is an arbitrary article and the target word level[6] inputted by the user, and the output of this stage is a set of target words.

The method for selecting target words is based on the level. That significantly influences the difficulty of test items. Hence, we consider the level of test takers to select target words. We utilize the Oxford 3000 and the Oxford 5000[7] to pick the ideal target words. The level of the selected target words should be higher than the level of test takers. Then, they can't directly realize the meaning of the target words, but rather they will conjecture them by the context. Additionally, based on the analysis of the The Official Guide to the TOEFL Test[8], we only found noun, verb, adjective and adverb words as the target words. Consequently, we follow this rule and only take words of these four classifications as the target words. We combine the two conditions (i.e., the level and the POS) to select the target words meeting conditions.

3.2 Paraphrasing Target Words

In the second stage, we find some words as the replacement of the target word to paraphrase the given article. These words also or almost make sense in their context. Then, they will be taken as answer keys or distractors (Section 3.3).

For this stage, we use a robust language model (i.e., Devlin et al. (2018)) considering the context to paraphrase the article. The input of this stage is a set of target words, which is selected from the last stage, and we use this language model with the context of different widths to get the replacement of these target words. These replacements will be the output of this stage.

We address this problem by using a simple heuristic that worked well in our experiments. In

[6]Users can choose A1, A2, B1, B2, C1, C2, or FL. They can choose one or more of them.

[7]They were presented by oxford dictionary (https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000). They classified most vocabularies into six levels, and we add one to put words which is not in these six levels.

[8]https://www.ets.org/toefl/test-takers/ibt/prepare/guides-books/

the question type of vocabulary test of reading comprehension, some test takers may choose the distractor as a result of the context of partial width. For example, if the test taker only considers the sentence "The **atmosphere** on Mars was thicker", the distractor "crust" may be a possible answer here. By this idea, we input the context of various widths into the language model. We set widths from 1 to 49 words and use 3 words as a gap.

3.3 Generating Answer Keys and Distractors

In the third stage, we use some rules to select answer keys and distractors from the replacements from section 3.2. Thus, the output of this stage are an answer key and 3 distractors for each target word. If the number of replacements passing the rules meets the setting of multiple choice questions (i.e., an answer key and 3 distractors), it will be a test item in section 3.4.

The first rule is following the option level, which is another input from users. The replacements matching the option level are kept. Next, POS of replacements different from that of target words are eliminated. The third rule is used to distinguish answer keys and distractors. If the replacement is in the synonym dictionary[9] of its target word, then this may be an answer key. Otherwise, if the replacement isn't in the synonym dictionary of its target word, this may be a distractor.

Unfortunately, some replacements, which are selected from the third rule and considered as distractors, are still reasonable in the given article and that will make the possible answer more than one. Therefore, we use an pre-trained natural language inference model[10] to dismiss them. If the score of ENTAILMENT of the replacement is higher than 0.97, this replacement will be removed. Similarly, for potential answer key, if the score of ENTAILMENT of this is lower than 0.98, this will be knocked out.

3.4 Generating Test Items

In the final stage, we process the distractors and answer keys found in the former stage and then generate last test items.

The target words with more than 3 distractors and more than 1 answer key will be kept. The tense or the singular and plural of options of these target words will be adjusted to match corresponding target words. Afterwards, we randomly array the positions of options. Then, join the target word to the stem (e.g., "The word atmosphere in the passage is closest in meaning to"). Finally, we complete generating test items for a given article with given levels.

4 Experiments

JEFF was designed to generate questions for given articles. As such, JEFF will be evaluated by comparing with human-designed test items.

We collected the articles from *VOA Learning English*[11]. For consistency, we only select the articles which are marked as advanced level and presented in 2021. We got 304 articles in the end. Then, we set target word level as B2, C1, C2, and FL and set option level as A2 to FL to implement on all collected articles. Hence, we had 304 articles with questions and kept 85 articles with more than 3 test items.

Next, we randomly reviewed 50 and selected the articles with more than three excellent test items. There were 22 articles selected.

On the other hand, we gathered the human-designed questions from The Official Guide to the TOEFL Test. There were 8 articles in Reading and 7 of them with more than 3 test items about vocabulary. Equally, we kept these 7.

Finally, we have three kinds of questions. One is machine-generated, another is machine-generated but selected by us and the other is human-designed. We randomly selected 2 articles with 3 randomly-selected test items for each kind and randomly arrayed the six articles to make an exam paper. By this way, we made 5 exam papers, in which two kinds of machine-generated test items were unrepeatable. Therefore, there were 30 test items for both machine-generated kinds and 21 test items for human-designed kind.

The two kinds of experimental setting simulate the two kinds of applications of JEFF. Directly machine-generated one simulates that learners read arbitrary online materials and want to test themselves. The kind which is machine-generated and selected by us simulates that teachers use JEFF to accelerate making exam papers.

We invited 10 English experts to evaluate these exam papers. One expert randomly reviewed one

[9] https://www.lexico.com/synonyms, which is presented by Oxford.

[10] The model output three scores for ENTAILMENT, NEUTRAL, and CONTRADICTION. Total of them are 1. See https://huggingface.co/roberta-large-mnli.

[11] https://www.voanews.com/

exam paper, so each paper was reviewed twice. The background of experts are 2 professors, 7 English teachers and 1 native speaker. The length of teaching years of 9 teachers is 8.5 years on average. They were asked to score test items by the instruction "Based on the quality of test items, score each of them (1-5 points)".

5 Results

In this section, we report the results of the experimental evaluation using the methodology described in the previous section. 74 questions are scored by 10 English experts at least 2 times.

The average score of human-designed test items is 4.06. The test items JEFF directly generated got 3.74, which is close to the human-designed ones. As for those generated by JEFF and selected by us, they got 4.11 which is even a little higher than human-designed ones.

Figure 2 shows the frequencies of three kinds of questions for each point. As we can see, the kind which is machine-generated and selected by us performs the best on 5 points, and the machine-generated kind is comparable to the human-designed kind. However, in the low point area (i.e., 1 and 2 points), there is only one review result of human-designed questions, which means almost all human-designed questions are almost above the level, while there are 9 and 14 ones for questions which is machine-generated and selected by us and machine-generated respectively. We think since we are not experts and we didn't totally select the acceptable questions for the kind which is machine-generated and selected by us, this kind doesn't perform as good as the human-designed kind in low score area.

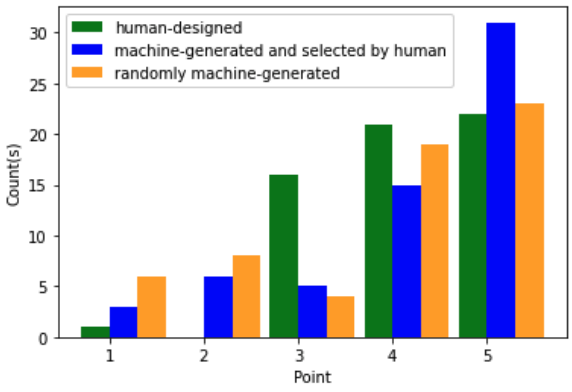

Figure 2: Frequency for each point

It is interesting to note that JEFF can almost replace humans in the generating question task. If JEFF has a little assistant by human, it is even better than only human-designed questions.

6 Future Work and summary

Many avenues exist for future research and improvement of our system. For example, joining a mechanism for filtering out the questions may be marked as 1 or 2 points. Additionally, an interesting direction to explore is selecting target words or generating options with more than one word (e.g., moving forward). Yet another direction of research would be to consider the attributes of options. For example, in four choice "A. explorers", "B. traders", "C. immigrants" and "D. ships", the first three are identities of people and different from the option D. The test takers may easily eliminate this option.

In summary, we have introduced a method for automatically generating questions of vocabulary tests on reading comprehension with controllable level that is suitable for using them on user-selected materials for learners and making exam papers with the assistance of teachers for them. The method involves selecting target words by inputted level, paraphrasing the given articles, and generating answer keys and distractors. We have implemented and thoroughly evaluated the method as applied to practical tests. In evaluations of three kinds of test source, we have shown that the method is close to the human-designed one and even outperforms that with our help.

References

Itziar Aldabe and Montse Maritxalar. 2010. Automatic distractor generation for domain specific texts. In *International Conference on Natural Language Processing*, pages 27–38. Springer.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826.

Dhawaleswar Rao Ch and Sujan Kumar Saha. 2018. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25.

David Coniam. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *Calico Journal*, pages 15–33.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal*, 2(3):210–224.

Shu Jiang and John SY Lee. 2017. Distractor generation for chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Chao-Lin Liu, Chun-Hung Wang, Zhao Ming Gao, and Shang-Ming Huang. 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 1–8.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative approach to fill-in-the-blank quiz generation for language learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–242.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005. Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 61–68.

Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. 2015. Automatic generation of english vocabulary tests. In *CSEDU (1)*, pages 77–87.

Ratsameetip Wita, Sahussarin Oly, Sununta Choomok, Thanabhorn Treeratsakulchai, and Surarat Wita. 2018. A semantic graph-based japanese vocabulary learning game. In *International Conference on Web-Based Learning*, pages 140–145. Springer.