# An Automatic Question Generation System for High School English Education

Tianqi Wang[1] Teruhiko Takagi[1] Masanori Takagi[2] Atsushi Tamura[3]
[1]Classi Corp. [2]The University of Electro-Communications.
[3]Iwate Prefectural University.
{ tianqi.wang, teruhiko.takagi }@classi.jp
takagi-m@uec.ac.jp, atsushi_t@iwate-pu.ac.jp,

## Abstract

The multiple-choice question is a highly prevalent assessment tool in educational contexts, yet manual question creation can be a costly endeavor. In this paper, we present a system designed to generate multiple-choice questions for English education. Our approach involves breaking down the question generation process into distinct steps and applying various techniques to each step. We explore the enhancements in question generation efficiency and also address the limitations of this system. The system's performance is evaluated based on feedback from human expert question creators.

## 1 Introduction

The multiple-choice question (MCQ) serves as a pivotal tool in assessing students' comprehension of various constructs [1]. As illustrated in Figure 1(1-3), an MCQ primarily comprises three components [2]. The **(1)stem** presents the question's text. In the case of fill-in-the-blank questions, students are required to select the correct choice to complete the stem sentence. In our English education setting, stems are provided in both English and Japanese. The correct choice is denoted as the **(2)key**, while the other choices are referred to as **(3)distractors**. Beyond these components, we introduce **(4)commentary** as an integral element of MCQs. Commentary offers concise instructions to elucidate the answer, particularly beneficial when students make errors.

Due to the brief time required to answer an MCQ, educational environments demand a substantial quantity of these questions. However, crafting effective MCQs presents challenges for human creators. Guidelines for creating

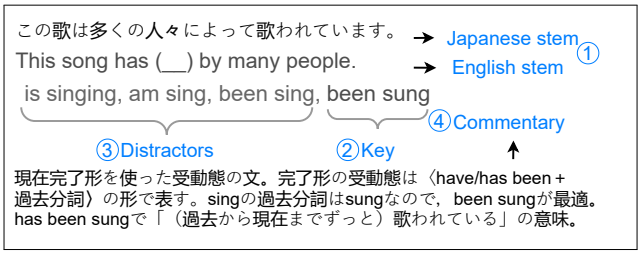


**Figure 1** Question components

MCQs and distractors encompass numerous rules and criteria [3, 4, 5]. Automatic Question Generation (AQG) emerges as a viable solution, yet several factors render this task complex. Firstly, question creators aim to generate questions aligned with specific course objectives, necessitating knowledge verification. Secondly, beyond content, fine-tuning difficulty levels within the format is crucial. Lastly, automatic generation of commentary proves necessary but notably challenging

In this paper, we present an automatic question generation system designed for English education context. Our objective does not entail creating a fully automated question generation approach. Instead, we acknowledge the necessity of manual revision to ensure question quality, particularly since the generated questions will be released to students for educational purposes. Our aim is to have human experts refine the automatically generated content, thereby enabling the creation of a larger volume of questions at a reduced cost.

## 2 Related Work

Ontology-based approaches utilizing structured knowledge resources are commonly employed in factual question generation [6, 7, 2]. Typically, given the course objective and input texts, linked data or an Instance Tree is generated. The answer choice is selected from the nodes, and

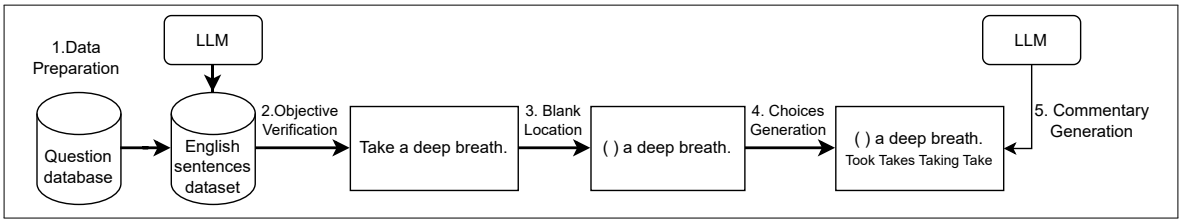


**Figure 2** The proposed overall workflow.

the question stem and distractors are generated based on neighboring nodes to the answer choice. Conversely, some research [8] focuses on transforming input sentences into question stems utilizing linguistic features such as POS, lexical patterns, and NER.

In contrast to factual questions explored in prior research, the determination of key choices and distractors in language learning questions cannot rely solely on the ontology graph. Therefore, our research emphasizes the utilization of linguistic features to validate course objectives.

## 3 Proposal

Our automatic question generation process is divided into several key steps, as illustrated in Figure 2, depicting the overall workflow.

**Dataset Preparation** We created an English text dataset from manually created fill-in-the-blank MCQs as candidates for question stems. This involved populating the answer blanks and utilizing paraphrasing with GPT-3.5 model[1)] to prevent the generation of duplicate questions. All sentences underwent translation into Japanese using the pretrained FuguMT model [2)]. Additionally, creators are able to input additional sentences as candidates for the question stem.

**Objective Verification** We developed linguistic features to validate specific course objectives, leveraging POS tags and dependency trees as depicted in Figure 3. These features are extracted using dependency parser and the en_core_web_sm model provided by Spacy[3)].

For instance, in the case of *passive voice* course objective, we identify spans that adhere to the following criteria: 1) the root of the span in the dependency tree is a verb $v$ tagged as *VBN*, 2) $v$ possesses a left-child node $c$, and 3) the POS tag of $c$ is *VBZ*, with the lemma text of $c$ being *be*. Additional constraints such as tense matching can be incorporated into these rules as necessary.

1) https://platform.openai.com/docs/models/gpt-3-5
2) https://staka.jp/wordpress/
3) https://spacy.io/

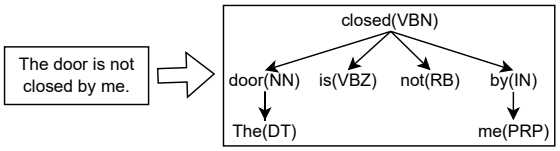


**Figure 3** An example of dependency tree.

**Blank Location** For a given course objective, we mask some specific words in the extracted sentences to generate the English stem. These masked words become part of the choices as the key to the question. This approach to locate blank words is similar to the objective verification method, but instead of masking the entire matched span, we only mask the most relevant words. For instance, the word *have* and its subsequent *verb in past participle form* are matched for course objective focused on *past perfect tense*, but only the *verb* is masked as the blank within the generated stem.

**Distractor Generation** When learning a particular English grammar item, students are expected to achieve two objectives: 1) grasp the meaning associated with the grammar item and 2) understand how to apply the grammar rule (e.g., transforming a verb to its past participle form). To address the first objective, we employ a pretrained masked language model, specifically the BERT model[4)] [9]. Given an English sentence containing a blank to fill in, we treat this blank as a masked word and predict it using the BERT model. The output from BERT consists of a list of words along with their associated scores. While traditionally, words with the highest scores might be chosen as correct options, we opt to select the top-N words after the 20th highest score as distractors. This approach ensures that selected distractors are highly unlikely to be correct answers but still maintain some degree of semantic relevance, thus that a basic understanding is required to distinguish the key choice from the distractors. Notably, only words sharing the same POS tag as the key are considered in this phase. For the second skill, we generate distractors by transform-

4) https://huggingface.co/bert-base-uncased

**Table 1** Correctness on dimensions.

| Objective | Word level (Q) | Word level (C) | Grammar | Context | Blank Location | Distractors | Commentary |
|---|---|---|---|---|---|---|---|
| 0.95 | 0.81 | 0.90 | 0.86 | 0.62 | 0.95 | 0.52 | 0.0 |

<GENERATED_QUESTION>
正答を教えてください。
この設問文の文型と意味を説明してください。最後に、誤答選択肢の単語の意味と正答にならない理由を教えてください。
<GENERATED_QUESTION>
*Please select the correct answer*
*and give explain the sentence pattern and meaning, and explain why it is correct and why the other choices are not.*

**Figure 4** Prompt for commentary generation. The English version is for reference.

ing the masked key into various forms.

**Commentary Generation** We include commentaries to complement the generated questions, and the GPT-3.5 model provided by OpenAI is utilized in our research. In an effort to establish a baseline for future endeavors, we do not do much prompt engineering in this phase, allowing us to gain a comprehensive understanding of the model's performance and limitations. Figure 4 illustrates the prompt utilized for commentary generation.

**Structural Similarity** In language learning, there's often a requirement to generate question stems that mirror the structure of a provided sentence. For instance, a user might request question stems focused on the *passive voice* while excluding instances with a *by phrase*. To address this, we calculate the structural similarity between sentences in the dataset and an input sentence, utilizing their dependency trees. Subsequently, we output only the top-N most similar sentences for question generation. This process enables the creation of question stems closely aligned with the desired sentence structure.

**Word-Level Control** As the questions are tailored for students across various proficiency levels, it's crucial to avoid the inclusion of overly hard words. We utilize the CEFR-J Wordlist version 1.6 [10], which categorizes 7,801 words into four proficiency levels (A1, A2, B1, and B2), with A1 being the simplest and B2 considered the most challenging. While this classification serves as a guideline, human judgment remains essential in determining the appropriateness of a word for students. The user assigns expected word levels for both stems ($l_s$) and key choices ($l_k$). If a stem contains a word surpassing the designated $l_s$ level, we highlight the word along with its level, leaving the final decision to human creators. Moreover, if the key choice's level exceeds $l_k$, we opt to remove the generated question from the dataset. This methodology ensures that questions adhere to specified word level criteria, thereby catering to diverse student proficiencies.

# 4 Results and Discussion

We conducted an evaluation comprising 21 questions designed for the *passive voice* course objective. We initially generated a pool of 120 questions, from which an editor curated 64 questions for subsequent revision. Then an expert question creator meticulously selected and revised 21 questions for release. To assess the quality of these 21 generated questions, we employed two evaluation methods: manual scoring and normalized editor distance.

## 4.1 Evaluation approaches

To comprehensively assess the quality of questions across various views, our expert question creator provided binary scores (0 for negative, 1 for positive) on specific dimensions for each question. These dimensions encompassed the following: **Objective**: Alignment with the target course objective, **Word Level (Stem)**: Appropriate word level in the question stem, **Word Level (Choices)**: Appropriate word level in the choices, **Grammar**: Accuracy of grammar in the question stem, **Context**: Naturalness of the question stem, **Blank**: Correct words placed as blanks, **Distractors**: Generation of suitable distractors, **Commentary**: Appropriateness of the provided commentary. The average scores across the 21 questions are detailed in Table 1. Notably, due to our more stringent policy on controlling word levels within choices, the **Word Level (Choices)** dimension received a higher score compared to **Word Level (Stem)**. It is apparent that the proposed method performed inadequately on dimensions related to stem context, distractors, and commentary. This discrepancy can be attributed to two main reasons. Firstly, a positive score was awarded only when no revision was needed for a particular dimension, making it challenging to attain high scores for dimensions relevant to longer content. Secondly, the subjective nature of the scores provided by the expert creator may have inclined towards revisions that aimed to enhance comprehensibility for students.

As previously mentioned, we view revisions by the expert creator as an integral part of the refinement process. We we consider the revised questions as the ground truth and measure the edit distance, specifically using the Levenshtein distance on character level, between the generated contents and the revised questions. A smaller edit distance signifies a lesser need for revision, reflecting a higher efficiency in the question generation process

Due to the sensitivity of the edit distance to text length, we normalize the edit distance as $d = \frac{e(g,t)}{\max(l_g,l_t)}$, where $e(\cdot,\cdot)$ represents the Levenshtein distance function, $g$ and $t$ denote the generated content and the ground truth, and $l_g$ and $l_t$ signify the lengths of the generated content and the ground truth, respectively.

We compute the distance for question stems (both in English and Japanese), choices, and commentary, presenting the results in Table 2. The table indicates that 1) minimal revision was generally required across most cases, and 2) commentaries frequently underwent revisions.

**Table 2** Normalized edit distances

| | stems(en) | stems(ja) | choices | commentaries |
|---|---|---|---|---|
| Mean | 0.244 | 0.205 | 0.064 | 0.861 |
| Median | 0.125 | 0.000 | 0.000 | 0.858 |

## 4.2 Instance analysis

We discuss the quality of generated questions with some instance in this section.

**Revision on question stems** 7 of 21 stems are revised, and three prevalent patterns are revealed. Firstly, certain revisions aimed to enhance contextual naturalness, such as expanding context for clarity (*e.g., She was surprised by the unexpected gift. → She was surprised by the unexpected gift from her friends.*). Secondly, simplification of sentences was observed ( *e.g., The new software has been downloaded by users from various countries. → This software has been used by many young people.* ). Lastly, some revisions aimed to avoid similar contexts, potentially improving diversity ( *e.g., My parents, as well as my brother, have been invited to the party. → I have been invited to give a speech at the university.*). It's noteworthy that while the last pattern involved rewriting the whole sentence, the overall sentence structure and blank word remained consistent. This highlights the usefulness of the generated contents as guiding hints for the expert editor, even in cases requiring substantial rewriting.

Automatic generated

**正答選択肢**: been sung
**設問文の文型**: 現在完了形の受動態
**設問文の意味**: その歌は過去から現在まで、多くの才能あるアーティストによって歌われています。
**誤答選択肢の意味と正答にならない理由**:
- is sing: この文型は現在進行形の被動態を作るものであり、現在完了形の受動態ではありません。
- am singing: この文型は現在進行形の被動態を作るものであり、現在完了形の受動態ではありません。
- being sing: この文型は現在進行形の被動態を作るものであり、現在完了形の受動態ではありません。
**正答の意味と正誤理由**: been sungは「歌われている」という過去からの継続的なアクションを表す現在完了形の受動態です。設問文の意味に合致しています。

Revised

現在完了形を使った受動態の文。
完了形の受動態は〈have/has been + 過去分詞〉の形で表す。singの過去分詞はsungなので，been sungが最適。has been sungで「（過去から現在までずっと）歌われている」の意味。

**Figure 5** An example of revision on commentary.

**Revision on commentaries** All commentaries underwent rewriting in the revision process across all 21 instances. Figure 5 provides an example of a generated commentary, comprising five components: 1) The correct answer. 2) Grammar or sentence pattern. 3) The meaning of the question stem. 4) Explanation of distractors. 5) Explanation of the key choice. The rewritten commentaries significantly condensed compared to the automatically generated system. The revised versions retained only the $2^{nd}$ and $5^{th}$ components from the original structure. It's essential to note that the correctness of the generated content remains reliable. Despite the reduction in length, these commentaries continue to serve as valuable hints for human experts.

# 5 Conclusion and Future work

In this research, we introduced an automatic question generation system tailored for English study. The majority of automatically generated content required minimal to no revision, enhancing the efficiency of question creation. Even instances necessitating slight revisions were aided by the generated content, serving as hints for human experts.

Two primary avenues for future exploration emerge as we move forward. Firstly, the current system utilizes rule-based methods for blank location, presenting limitations in both performance and application. To address this, our future endeavors will focus on training machine learning models to replace these rule-based methods. Additionally, we aim to extend the application of this system beyond English to encompass other subjects in future studies.

## References


[1] MJ Wise. The effective use of negative stems and “all of the above” in multiple-choice tests in college courses. **J Educ Teach Soc Stud**, Vol. 2, No. 4, p. 47, 2020.

[2] Archana Praveen Kumar, Ashalatha Nayak, Manjula Shenoy K, Chaitanya, and Kaustav Ghosh. A Novel Framework for the Generation of Multiple Choice Question Stems Using Semantic and Machine-Learning Techniques. **International Journal of Artificial Intelligence in Education**, March 2023.

[3] Kosaku Nagasaka. Multiple-choice questions in mathematics: Automatic generation, revisited. In **The 25th Asian technology conference in mathematics, virtual format, Radford University, Virginia, USA and Suan Sunandha Rajabhat University, Thailand. https://atcm. mathandtech. org/EP2020/invited/21785. pdf**, 2020.

[4] An Evidence-Based Approach to Distractor Generation in Multiple-Choice Language Tests. **Journal of Higher Education Theory and Practice**, Vol. 21, No. 10, September 2021.

[5] 坪田彩乃, 石井秀宗. 多枝選択式問題作成ガイドラインの実証的検討. 日本テスト学会誌, Vol. 16, No. 1, pp. 1–12, 2020.

[6] Ruslan Mitkov, et al. Computer-aided generation of multiple-choice tests. In **Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing**, pp. 17–22, 2003.

[7] Fumika Okuhara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. Generation of Multiple Choice Questions Including Panoramic Information using Linked Data:. In **Proceedings of the 11th International Conference on Agents and Artificial Intelligence**, pp. 110–120, Prague, Czech Republic, 2019. SCITEPRESS - Science and Technology Publications.

[8] Miroslav Blšták and Viera Rozinajová. Automatic question generation based on sentence structure analysis using machine learning approach. **Natural Language Engineering**, Vol. 28, No. 4, pp. 487–517, July 2022.

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

[10] Yukio Tono. The cefr-j and its impact on english language teaching in japan.”. In **JACET international convention selected papers**, 第 4 巻, pp. 31–52, 2016.