Distractor Generation for Fill-in-the-Blank Exercises by Question Type

Nana Yoshimi¹, Tomoyuki Kajiwara¹, Satoru Uchida², Yuki Arase³, Takashi Ninomiya¹
¹Ehime University, ²Kyushu University, ³Osaka University
{yoshimi@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp
uchida@flc.kyushu-u.ac.jp, arase@ist.osaka-u.ac.jp

Abstract
This study addresses the automatic generation of distractors for English fill-in-the-blank exercises in the entrance examinations for Japanese universities. While previous studies applied the same method to all questions, actual entrance examinations have multiple question types that reflect the purpose of the questions. Therefore, we define three types of questions (grammar, function word, and context) and propose a method to generate distractors according to the characteristics of each question type. Experimental results on 500 actual questions show the effectiveness of the proposed method for both automatic and manual evaluation.

1 Introduction
Fill-in-the-blank questions, also known as cloze tests (Taylor, 1953), are one way to assess learners’ English proficiency and are widely used in examinations such as TOEIC¹ and in school education. As shown in Figure 1, the question format generally consists of a four-choice option with one correct answer and three distractors. These require substantial costs because they are manually created by question writers with extensive language teaching experience. This study automatically generates distractors to reduce workload.

Most of the previous studies on the automatic generation of cloze tests (Mitkov and Ha, 2003; Sumita et al., 2005; Zesch and Melamud, 2014; Jiang and Lee, 2017; Susanti et al., 2018; Panda et al., 2022) have generated words that are semantically similar to the correct word as distractors. These candidate words were ranked by similarity and filtered by word 3-gram. That is, if a 3-gram containing a candidate word appears in Wikipedia, that candidate is excluded. It filters out expressions that are actually used in a large-scale corpus to exclude appropriate examples from the distractor candidates.

Jiang and Lee (2017) utilized cosine similarity with word embeddings (Mikolov et al., 2013) to identify candidate words that are semantically similar to the correct word. These candidate words were ranked by similarity and filtered by word 3-gram. That is, if a 3-gram containing a candidate word appears in Wikipedia, that candidate is excluded. It filters out expressions that are actually used in a large-scale corpus to exclude appropriate examples from the distractor candidates.

In this study, we first manually classify English fill-in-the-blank questions in the entrance examinations for Japanese universities² by an expert. Next, we propose a method for automatic distractor generation according to the characteristics of each question type. Experimental results on 500 actual questions show the effectiveness of the proposed method for both automatic and manual evaluation.

2 Related Work
Previous studies have generated distractors in the following three steps: (1) candidate generation, (2) reranking, and (3) filtering.

Jiang and Lee (2017) utilized cosine similarity with word embeddings (Mikolov et al., 2013) to identify candidate words that are semantically similar to the correct word. These candidate words were ranked by similarity and filtered by word 3-gram. That is, if a 3-gram containing a candidate word appears in Wikipedia, that candidate is excluded. It filters out expressions that are actually used in a large-scale corpus to exclude appropriate examples from the distractor candidates.

Yeung et al. (2019) reranked the candidates generated from word embeddings by the mask-filling

Jeff didn’t accept the job offer because of the ____ salary.
(a) low (b) weak (c) cheap (d) inexpensive

It was certainly _crowded than I thought it would be.
(a) less (b) little (c) least (d) fewer ((a) is correct )

Figure 1: Example of English fill-in-the-blank question. (National Center Test for University Admissions, 2018)²

leads to bias in the characteristics of the generated distractors. Actual entrance examinations have multiple question types reflecting the purpose of the questions, such as grammatical knowledge and idiomatic expressions. Existing methods have difficulty in flexibly changing the characteristics of distractors for each question type.

In this study, we first manually classify English fill-in-the-blank questions in the entrance examinations for Japanese universities² by an expert. Next, we propose a method for automatic distractor generation according to the characteristics of each question type. Experimental results on 500 actual questions show the effectiveness of the proposed method for both automatic and manual evaluation.

²https://jcshop.jp/SHOP/18149/list.html

---

¹https://www.ets.org/toeic.html

---

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 276–281
July 10-12, 2023 ©2023 Association for Computational Linguistics
I hear that one of his three sisters __ four movies a week. sees seeing seen see grammar
My mother was surprised __ the news that I passed the test. at to for in function word
When you exercise, you should wear __ and loose clothing. comfortable delicate serious flat context

Table 1: Examples of question types. From top to bottom, the sources are (Toyo University, 2018), (Meijo University, 2017), (Nakamura Gakuen University, 2018).

<table>
<thead>
<tr>
<th>Carrier sentence</th>
<th>Correct</th>
<th>Distractors</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>I hear that one of his three sisters __ four movies a week.</td>
<td>sees</td>
<td>seeing</td>
<td>grammar</td>
</tr>
<tr>
<td>My mother was surprised __ the news that I passed the test.</td>
<td>at</td>
<td>to</td>
<td>function word</td>
</tr>
<tr>
<td>When you exercise, you should wear __ and loose clothing.</td>
<td>comfortable</td>
<td>delicate</td>
<td>flat context</td>
</tr>
</tbody>
</table>

Table 2: Statistics of question types.

<table>
<thead>
<tr>
<th>Question type</th>
<th>Number of questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>66 (13.2%)</td>
</tr>
<tr>
<td>Function word</td>
<td>195 (39.0%)</td>
</tr>
<tr>
<td>Context</td>
<td>239 (47.8%)</td>
</tr>
</tbody>
</table>

3 Definition of Question Types

An experienced English teacher specializing in English education has categorized the question types for English fill-in-the-blank questions. The analysis covers 500 randomly selected questions from the entrance examinations for Japanese universities in the five-year period from 2017 to 2021. As shown in Table 1, the following three question types were defined:

- **Grammar**: Questions that mainly use the conjugated form of the same word as choices.
- **Function word**: Questions that are choices from a prescribed list of function words.
- **Context**: Questions with choices determined by context or idiomatic expressions.

Table 2 shows the number of occurrences for each question type. Approximately half of the questions were on context, 40% were on function word, and 10% were on grammar. In the next section, we propose how to generate distractors according to the characteristics of each question type.

4 Generating Distractors

Following previous studies (Jiang and Lee, 2017; Yeung et al., 2019; Panda et al., 2022), we also generate distractors through three steps. For candidate generation and reranking, we selected combinations of the existing methods described in Section 2 that maximize performance on the validation dataset for each question type. For filtering, we propose methods according to the characteristics of each question type, which are described below.

4.1 Filtering for Questions on Grammar

For questions on grammar, the conjugated forms of the correct word should be obtained as candidates. Therefore, we apply POS filtering. That is, we exclude candidates that have the same part of speech or the same conjugation as the correct word. Furthermore, to avoid unreliable distractors that could be the correct answer, we exclude candidates with a high mask-filling probability by BERT (Devlin et al., 2019). Unlike Yeung et al. (2019), called BERT (static), which used two fixed thresholds to select the top $\theta_H$ to $\theta_L$, our filter, called BERT (dynamic), dynamically changes the thresholds. Specifically, we exclude candidates that have a higher probability than the correct word. The example of the first sentence in Table 1 shows that “thinks” is eliminated as a candidate for the same

For the validation dataset, 500 questions were randomly selected in addition to the evaluation dataset annotated in Section 3. These questions were automatically annotated with question types by BERT (Devlin et al., 2019). The accuracy of BERT was 84.8% in the 10-fold cross-validation.
part of speech, and “watches” is eliminated as a high probability candidate.

4.2 Filtering for Questions on Function Word
For questions on function words, only function words such as prepositions and conjunctions are basically used as choices. Therefore, we utilize the list of function words\(^4\) for entrance examinations for Japanese universities to exclude candidates not included in this list. The example of the second sentence in Table 1 shows that “time” and “taken” are eliminated.

4.3 Filtering for Questions on Context
Since the questions on context are designed to test knowledge of collocations or idioms, candidates should be obtained for words that often co-occur with surrounding words in the carrier sentence. However, as with questions on grammar, to avoid unreliable distractors, candidates with a high mask-filling probability by BERT are excluded. The example of the third sentence in Table 1 shows that “comfy” and “cosy” are eliminated.

5 Experiments
We evaluate the method of distractor generation on the 500 questions constructed in Section 3.

5.1 Setting
Implementation Details For candidate generation, we implemented methods based on word embeddings (Jiang and Lee, 2017) and round-trip machine translation (Panda et al., 2022). We utilized fastText (Bojanowski et al., 2017) as word embeddings and Transformer (Vaswani et al., 2017), trained on English-German language pairs\(^5\) (Ng et al., 2019; Ott et al., 2019) according to the previous study (Panda et al., 2022), as machine translators. For word alignment, we used Hungarian matching (Kuhn, 1955) based on word embeddings (Song and Roth, 2015).

For reranking, we implemented methods based on word embeddings (Jiang and Lee, 2017) and BERT (Yeung et al., 2019). We utilized BERT-base-uncased (Devlin et al., 2019) via HuggingFace Transformers (Wolf et al., 2020). Note that the candidate words are restricted to the intersection of the vocabulary of fastText and BERT.

For filtering, NLTK (Bird and Loper, 2004) was used for pos tagging. We used 166 function words.\(^4\)

Comparative Methods We compared the proposed method with three existing methods described in Section 2: methods based on word embeddings (Jiang and Lee, 2017), masked language models (Yeung et al., 2019), and round-trip machine translations (Panda et al., 2022). For word 3-gram filtering, we used preprocessed English Wikipedia (Guo et al., 2020). For BERT (static) filtering, we used thresholds of \(\theta_H = 11\) and \(\theta_L = 39\) following Yeung et al. (2019).

Automatic Evaluation To evaluate whether the generated distractors are matched with the actual entrance examinations, an automatic evaluation is performed. We generated 100 words of candidates for each method and compared the top

---

\(^4\)https://ja.wikibooks.org/wiki/大学受験英語_英単語/機能語・機能型単語一覧

\(^5\)As a pivot language, we also tried Japanese, the native language of the examinees, but German performed better.
Carrier sentence: There are three people __ school events.
Question type: Grammar
Correct answer: discussing
Distractors: discussing
discussing
discuss
discuss
discuss
discuss
discuss
discusses
discussing
examining
discussing
talking
discussing
considering
promoting
deciding
initiating
exploring
talking
discussing
considering
promoting
deciding
initiating
exploring
(Panda et al., 2022)
talking
dealing
speaking
working
reporting
giving
Wednesday
conversation

Carrier sentence: They are a little worried __ their daughter’s trip to the Amazon.
Question type: Function word
Correct answer: about
Distractors: for
with
from
concerning
regarding
relating
talking
what
telling
pertaining
considering
up
the
seeing
than
just
discussing
the
any
and
afraid
affected
anxious
at
by
after
for
at
from
with
of

Table 4: Examples of generated distractors. The example in the upper row is from (Ritsumeikan University, 2019), and the example in the lower row is from (Morinomiya University of Medical Sciences, 2018). Candidates matching the gold distractors are highlighted in bold.

$k \in \{3, 5, 10, 20\}$ words, after reranking and filtering, to the three gold distractors. Note that if there are fewer than $k$ candidates, the remainder were randomly selected from the vocabulary. We employed the F1-score as the evaluation metric.

**Manual Evaluation** To assess the correlation of examinee performance between the generated questions and the actual entrance examinations, a manual evaluation is performed. First, distractors are generated for each of the 60 randomly selected questions in each of the proposed and two comparative methods (Jiang and Lee, 2017; Panda et al., 2022). Next, ten university students, who are native Japanese speakers, took 100 English fill-in-the-blank questions from the actual entrance examinations, as well as these 180 generated questions. Note that these questions are sampled evenly by question type, with no duplication. Finally, we calculated the correlation of accuracy between the generated and actual questions.

**5.2 Results**

**Automatic Evaluation** Table 3 shows the results of the automatic evaluation. The top three rows show the performance of the comparison method and the bottom row shows the performance of the proposed method for each question type. The proposed method achieved the best performance in 9 out of 12 settings and the second best performance in the remaining 3 settings. This implies the effectiveness of filtering according to the characteristics of question types. The improvement in performance was particularly noticeable for questions on function words, with greater improvement as the number of candidates $k$ increased.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Jiang and Lee, 2017)</td>
<td>0.739</td>
<td>0.723</td>
<td>0.584</td>
</tr>
<tr>
<td>(Panda et al., 2022)</td>
<td>0.776</td>
<td>0.774</td>
<td>0.614</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>0.903</td>
<td>0.802</td>
<td>0.629</td>
</tr>
</tbody>
</table>

Table 5: Correlation of accuracy between actual entrance examinations and generated questions.

**Manual Evaluation** Table 5 shows the results of the manual evaluation. The proposed method has the highest correlation with the performance of the actual entrance examinations for all correlation coefficients. This means that the proposed method is most effective in identifying the English proficiency of examinees.

**Output Examples** Table 4 shows examples of generated distractors. In questions on grammar, existing methods without consideration of question types generate candidates that are semantically close to the correct word, but the proposed method correctly generates conjugated forms of the correct word. In questions on function words, the existing methods include candidates other than function words, but the proposed method generates only function words, correctly ranking the gold distractors higher. In questions on context, as shown in Table 3, the proposed method is not much different from the existing method until the top five, but may be followed by good candidates even after that.

**6 Conclusion**

To reduce the cost of creating English fill-in-the-blank questions in entrance examinations for Japanese universities, this study addressed automatic distractor generation. First, we identified
three question types and constructed a fill-in-the-blank corpus annotated by an expert with those question types. Next, we proposed methods to generate distractors that take into account the characteristics of each question type, focusing on candidate filtering. Experimental results based on automatic and manual evaluations demonstrate the effectiveness of the proposed method. Specifically, our method is able to generate candidates that match the gold distractors better than existing methods and has the highest correlation with the examinees’ English proficiency as assessed in actual entrance examinations. For future work, we plan to expand the corpus size by estimating question types, to generate distractors by supervised learning.

Acknowledgements

We thank anonymous reviewers for valuable comments and suggestions. This work was supported by JSPS KAKENHI Grant Number JP21H03564 and JP22H00677.

References


