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

Nana Yoshimi[1], Tomoyuki Kajiwara[1], Satoru Uchida[2], Yuki Arase[3], Takashi Ninomiya[1]
[1]Ehime University, [2]Kyushu University, [3]Osaka 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 TOEIC[1] 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 words as distractors. Other methods have been proposed, such as those based on co-occurrence with words in the carrier sentence (Liu et al., 2005; Hill and Simha, 2016), considering the whole context (Yeung et al., 2019), and considering the learner's error tendencies (Sakaguchi et al., 2013). However, these previous studies apply the same method to all questions, which 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.

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

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

In this study, we first manually classify English fill-in-the-blank questions in the entrance examinations for Japanese universities[2] 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

[1]https://www.ets.org/toeic.html

[2]https://jcshop.jp/SHOP/18149/list.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

Carrier sentence	Correct	Distractors			Type
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[2] are (Toyo University, 2018), (Meijo University, 2017), (Nakamura Gakuen University, 2018).

probability with BERT (Devlin et al., 2019). They also utilize BERT for filtering, eliminating candidates with too high and too low probabilities.

Panda et al. (2022) proposed candidate generation based on round-trip machine translation. That is, the carrier sentence was first translated into a pivot language and back-translated into English. Then, word alignment was used to obtain a candidate for the correct word and its corresponding word. These candidates were reranked using word embeddings and filtered by WordNet (Miller, 1995). Specifically, synonyms of the correct word in WordNet and words with a different part of speech from the correct word were excluded from the candidates.

These existing methods have been evaluated in different ways on different datasets, making it difficult to compare their performance. We have comprehensively evaluated them and propose further improvements on top of their combinations.

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.

Question type	Number of questions
Grammar	66 (13.2%)
Function word	195 (39.0%)
Context	239 (47.8%)

Table 2: Statistics of question types.

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[3] 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 θ_H to θ_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

[3]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.

Type	Method	Candidate	Reranking	Filtering	$k = 3$	$k = 5$	$k = 10$	$k = 20$
Grammar	Jiang-2017	fastText	fastText	Word 3-gram	24.7	21.6	**17.7**	**11.2**
	Yeung-2019	fastText	BERT	BERT (static)	1.5	1.9	3.0	3.4
	Panda-2022	Round-trip	fastText	WordNet	8.6	8.3	5.6	3.6
	Ours	fastText	fastText	POS+BERT (dynamic)	**27.8**	**25.0**	17.0	10.4
Function word	Jiang-2017	fastText	fastText	Word 3-gram	10.3	12.1	11.8	9.3
	Yeung-2019	fastText	BERT	BERT (static)	6.3	7.1	7.3	5.7
	Panda-2022	Round-trip	fastText	WordNet	15.9	16.7	13.1	7.8
	Ours	Round-trip	BERT	List of function words	**19.1**	**22.2**	**21.1**	**13.2**
Context	Jiang-2017	fastText	fastText	Word 3-gram	2.2	2.9	3.7	3.2
	Yeung-2019	fastText	BERT	BERT (static)	1.8	2.0	2.3	2.7
	Panda-2022	Round-trip	fastText	WordNet	**4.2**	5.1	4.6	3.2
	Ours	Round-trip	fastText	BERT (dynamic)	3.8	**5.3**	**5.8**	**4.4**

Table 3: Results of automatic evaluation of generated distractors by F1-score.

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 : discuss	discussed	discusses			
(Jiang and Lee, 2017)	debating	talking	discussion	commenting	mentioning	**discuss**	examining
(Yeung et al., 2019)	creating	talking	considering	promoting	deciding	initiating	exploring
(Panda et al., 2022)	talking	dealing	speaking	working	reporting	giving	wednesday
Proposed Method	discussion	**discuss**	**discussed**	discussions	**discusses**	about	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			
(Jiang and Lee, 2017)	concerning	regarding	relating	talking	what	telling	pertaining
(Yeung et al., 2019)	considering	up	the	seeing	than	just	discussing
(Panda et al., 2022)	the	any	and	afraid	affected	anxious	at
Proposed Method	by	after	**for**	at	**from**	**with**	of

Table 4: Examples of generated distractors. The example in the upper row is from (Ritsumeikan University, 2019),[2] and the example in the lower row is from (Morinomiya University of Medical Sciences, 2018).[2] 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.

Method	Pearson	Spearman	Kendall
(Jiang and Lee, 2017)	0.739	0.723	0.584
(Panda et al., 2022)	0.776	0.774	0.614
Proposed Method	**0.903**	**0.802**	**0.629**

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

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual Language Model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452.

Jennifer Hill and Rahul Simha. 2016. Automatic Generation of Context-Based Fill-in-the-Blank Exercises Using Co-occurrence Likelihoods and Google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30.

Shu Jiang and John 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.

Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

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.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations*.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Ruslan Mitkov and Le An Ha. 2003. Computer-Aided Generation of Multiple-Choice Tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, page 17–22.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation*, pages 314–319.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 2022. Automatic Generation of Distractors for Fill-in-the-Blank Exercises with Round-Trip Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401.

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*, pages 238–242.

Yangqiu Song and Dan Roth. 2015. Unsupervised Sparse Vector Densification for Short Text Similarity. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1275–1280.

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.

Yunik Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic Distractor Generation for Multiple-choice English Vocabulary Questions. *Research and Practice in Technology Enhanced Learning*, 13(15):1–16.

Wilson L Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism quarterly*, 30(42):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Chak Yan Yeung, John Lee, and Benjamin Tsou. 2019. Difficulty-aware Distractor Generation for Gap-Fill Items. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164.

Torsten Zesch and Oren Melamud. 2014. Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.