問題タイプを考慮した英単語穴埋め問題の不正解選択肢の自動生成

吉見 菜那 [1]　梶原 智之 [1]　内田 諭 [2]　荒瀬 由紀 [3]　二宮 崇 [1]
[1] 愛媛大学　[2] 九州大学　[3] 大阪大学
yoshimi@ai.cs.ehime-u.ac.jp　{kajiwara,ninomiya}@cs.ehime-u.ac.jp
uchida@flc.kyushu-u.ac.jp　arase@ist.osaka-u.ac.jp

概要

本研究では、日本の大学入試の英単語穴埋め問題を題材として、不正解選択肢の自動生成に取り組む。先行研究では全ての問題に対して同一の手法を適用していたが、本研究では 3 種類の問題タイプ（文法・機能語・文脈）を定義し、それぞれの特徴を考慮して不正解選択肢を自動生成する手法を提案する。日本の大学入試で実際に出題された 500 問を用いた評価実験の結果、提案手法は特に機能語の問題において既存手法を大きく上回る性能を達成した。

1　はじめに

英単語穴埋め問題 [1] は、学習者の英語習熟度を評価する方法のひとつであり、TOEIC[1)]や実用英語技能検定[2)]などの試験から学校教育の現場まで、広く使用されている。問題形式は図 1 のとおり、ある 1 単語を空欄とする問題文に対して、空欄に当てはまる正解選択肢が 1 つ、当てはまらない不正解選択肢が 3 つの 4 択形式で構成されるのが一般的である。これらの選択肢は、教員などの豊富な言語教育経験を持つ作問者によって人手で作成されるため、作問コストが大きい。本研究では、作問者の負担軽減のために、不正解選択肢の候補を自動生成する。

英単語穴埋め問題における不正解選択肢の自動生成に関する先行研究では、正解単語と意味的に類似する単語を生成する手法 [2–7] が多く提案されている。また、問題文中の単語との共起情報を用いる手法 [8, 9]、文脈全体を考慮する手法 [10]、学習者の誤り傾向を考慮する手法 [11] なども提案されてきた。しかし、これらの先行研究は全ての問題に対して同一の手法を適用するため、生成された不正解選択肢の特徴に偏りがある。実際の問題には、作問者の意図が反映されており、文法知識を問う問題や慣用表現を問う問題など複数の問題タイプが存在する。そのため、既存手法には不正解選択肢の特徴を問題タイプごとに柔軟に変更できないという課題がある。

Jeff didn't accept the job offer because of the ___ salary.
(a) low　(b) weak　(c) cheap　(d) inexpensive
(a)が正解選択肢

図 1　英単語穴埋め問題（大学入試センター試験, 2018）[4)]

本研究では、日本の大学入試における英単語穴埋め問題を対象に、人手で問題タイプを分類する。その上で、問題タイプごとの選択肢の特徴を考慮して、それぞれに適した不正解選択肢の自動生成手法を提案する。作問者によって用意された実際の不正解選択肢と自動生成した不正解選択肢の一致率を評価したところ、提案手法は多くの設定において既存手法 [5, 7, 10] を上回る性能を達成した。

2　関連研究

先行研究 [5, 7, 10] では、英単語穴埋め問題の不正解選択肢の自動生成を、候補生成・リランキング・フィルタリングの 3 つのステップで行っている。本節では、この 3 ステップで各手法の概要をまとめる。

単語分散表現に基づく手法 [5] は、word2vec [12] の余弦類似度によって正解単語との意味的類似度が高い単語を候補生成し、候補単語を類似度によってランキングし、フィルタリングとして単語 3-gram を用いている。単語 3-gram によるフィルタリングでは、候補単語を空欄に当てはめた際に、候補単語とその前後の単語からなる単語 3-gram が Wikipedia[3)]上に出現する場合、その単語を候補から除外する。

1）https://www.ets.org/toeic.html
2）https://www.eiken.or.jp/eiken/
3）https://en.wikipedia.org/

表 1　本研究で構築した英単語穴埋め問題データセットの例

問題文	正解	不正解			タイプ	出典
I hear that one of his three sisters __ four movies a week.	sees	seeing	seen	see	文法	(東洋大学, 2018)[4]
My mother was surprised __ the news that I passed the test.	at	to	for	in	機能語	(名城大学, 2017)[4]
When you exercise, you should wear __ and loose clothing.	comfortable	delicate	serious	flat	文脈	(中村学園大学, 2018)[4]

表 2　問題タイプごとの出現頻度

問題タイプ	問題数
文法	66 (13.2%)
機能語	195 (39.0%)
文脈	239 (47.8%)

マスク言語モデルに基づく手法 [10] は、単語分散表現に基づく手法 [5] を改良したもので、単語分散表現によって候補生成し、BERT [13] の単語穴埋め確率によって文脈を考慮して候補をリランキングする。フィルタリングにも BERT の単語穴埋め確率を用いており、確率の上位 θ_H 位以上は正解となり得る信頼性の低い不正解選択肢であり θ_L 位以下は妥当性の低い不正解選択肢であるとして、確率の上位 θ_H 位から θ_L 位までの候補単語を抽出している。

折り返し翻訳に基づく手法 [7] は、問題文をドイツ語などのピボット言語に機械翻訳し、それをさらに英語へと折り返し翻訳した際に、単語アライメントによって正解単語と対応する候補単語を獲得する。リランキングには、単語分散表現の余弦類似度 [5] や BERT の単語穴埋め確率 [10] など、任意の既存手法を用いる。フィルタリングでは、WordNet [14] を用いて正解単語の同義語を除外し、さらに正解単語と異なる品詞の単語を除外する。

3　問題タイプの定義

2017 年から 2021 年までの 5 年間に出題された日本の大学入試問題[4] の中から無作為に 500 問を抽出し、英語教育を専門とする大学教員が問題タイプを分類した。表 1 に例示するように、英単語穴埋め問題を以下の 3 種類の問題タイプに分類した。

- 文法：主に同じ語の活用形を選択肢とする問題
- 機能語：規定の機能語リストからの選択肢である問題
- 文脈：文脈または慣用表現によって選択肢が決まる問題

表 2 に、各問題タイプの出題頻度の内訳を示す。

4)　https://jcshop.jp/SHOP/18149/list.html

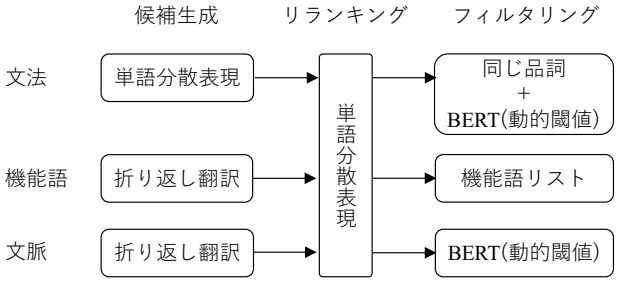

図 2　提案手法の概要

文脈の問題が約半数を占め、機能語の問題が約 4 割、文法の問題が 1 割強の割合であった。次節では、これらの問題タイプごとの特徴を考慮した不正解選択肢の生成手法を検討する。

4　不正解選択肢の生成

本研究でも先行研究 [5, 7, 10] と同様に、候補生成・リランキング・フィルタリングの 3 つのステップによって不正解選択肢を生成する。提案手法の概要を図 2 に示す。候補生成およびリランキングにおいては、2 節で説明した既存手法のうち、検証セット[5] における性能が最高となる手法の組み合わせを問題タイプごとに選択した。フィルタリングにおいては、以下で説明する各問題タイプの特徴を考慮した新しい手法を提案する。

4.1　文法問題のためのフィルタリング

文法問題では、基本的に同じ語の活用形を選択肢とするため、正解単語の活用形を不正解選択肢の候補として取得したい。そこで本研究では、正解単語と同じ品詞および活用形を持つ候補を除外する。

また、信頼性の低い（正解となり得る）不正解選択肢を避けるために、空欄に当てはめた際に自然な文を構成する候補は除外する必要がある。そこで本研究では、BERT [13] の単語穴埋め確率を用いて、正解単語よりも高確率の候補単語を除外する。表 1 の 1 文目の例では、同じ品詞の候補として “thinks”、高確率の候補として “watches” などが削除される。

5)　3 節でラベル付けした評価用の 500 問とは別に 500 問を無作為抽出し、BERT [13] で問題タイプを自動分類した。なお、自動分類の正解率は 10 分割交差検証で 84.8% であった。

表 3　問題タイプごとの不正解選択肢の自動生成の F 値

問題タイプ	手法	候補生成	リランキング	フィルタリング	$k=3$	$k=5$	$k=10$	$k=20$
文法	[5]	fastText	fastText	単語 3-gram	24.7	21.6	**17.7**	**11.2**
	[10]	fastText	BERT	BERT（静的閾値）	1.5	1.9	3.0	3.4
	[7]	折返翻訳	fastText	同義語+異なる品詞	8.6	8.3	5.6	3.6
	Ours	fastText	fastText	同じ品詞+BERT（動的閾値）	**27.8**	**25.0**	17.0	10.4
機能語	[5]	fastText	fastText	単語 3-gram	10.3	12.1	11.8	9.3
	[10]	fastText	BERT	BERT（静的閾値）	6.3	7.1	7.3	5.7
	[7]	折返翻訳	fastText	同義語+異なる品詞	15.9	16.7	13.1	7.8
	Ours	折返翻訳	fastText	機能語リスト	**21.4**	**22.7**	**20.0**	**13.1**
文脈	[5]	fastText	fastText	単語 3-gram	2.2	2.9	3.7	3.2
	[10]	fastText	BERT	BERT（静的閾値）	1.8	2.0	2.3	2.7
	[7]	折返翻訳	fastText	同義語+異なる品詞	**4.2**	5.1	4.6	3.2
	Ours	折返翻訳	fastText	BERT（動的閾値）	3.8	**5.3**	**5.8**	**4.4**

4.2　機能語問題のためのフィルタリング

機能語問題では、基本的に前置詞や接続詞などの機能語のみを選択肢として用いる。そこで本研究では、日本の大学受験用の機能語リスト[6]を用意し、ここに含まれない候補単語を除外する。表 1 の 2 文目の例では、“time” や “taken” などが削除される。

4.3　文脈問題のためのフィルタリング

文脈問題は、コロケーションやイディオムの知識を問う問題であるため、問題文中の周辺単語とよく共起する単語を候補として使用したい。ただし、文法問題と同様に、信頼性の低い不正解選択肢を避けるために、BERT の単語穴埋め確率において正解単語よりも高確率の候補単語は除外する。表 1 の 3 文目の例では、“comfy” や “cosy” が削除される。

5　評価実験

3 節で収集した 500 件の英単語穴埋め問題を対象に、不正解選択肢を自動生成する評価実験を行う。

5.1　実験設定

モデル　候補生成には、2 節で紹介した単語分散表現に基づく手法 [5] および折り返し翻訳に基づく手法 [7] を用いた。ここで、単語分散表現には fastText[7] [15] を使用した。また、折り返し翻訳には先行研究 [7] の設定に従って英独および独英の Transformer[8] [16, 17] を使用し、単語アライメントの計算には単語分散表現および Hungarian 法 [18, 19] を用いて候補単語を収集した。

リランキングには、単語分散表現に基づく手法 [5] およびマスク言語モデルに基づく手法 [10] を用いた。後者のリランキングには、HuggingFace Transformers [20] 経由で BERT-base-uncased[9] [13] を使用した。なお、候補単語は fastText および BERT の語彙に共通して出現する単語のみに制限している。

フィルタリングでは、NLTK[10] [21] を用いて品詞タグ付けした。機能語は、166 単語[6]を収集した。

比較手法　本研究では、2 節で説明した単語分散表現に基づく手法 [5]、マスク言語モデルに基づく手法 [10]、折り返し翻訳に基づく手法 [7] の 3 手法を提案手法と比較した。比較手法のフィルタリングにおいて、Wikipedia は前処理済み[11] [22] の英語テキストを使用した。また、先行研究 [10] の設定に従い、閾値には $\theta_H = 11$ および $\theta_L = 39$ を使用した。

評価　各手法において 100 種類の単語を候補生成し、リランキングおよびフィルタリングした上位 $k \in \{3, 5, 10, 20\}$ 単語を、実際の不正解選択肢 3 件と比較する。評価指標には、再現率および適合率を求め、その調和平均（F 値）を用いる。ただし、候補が k 単語に満たない場合は、語彙の中から無作為に選択した。

6）https://ja.wikibooks.org/wiki/大学受験英語_英単語/機能語・機能型単語一覧
7）https://fasttext.cc/docs/en/english-vectors.html
8）https://github.com/facebookresearch/fairseq/blob/main/examples/wmt19/README.md
9）https://huggingface.co/bert-base-uncased
10）https://www.nltk.org/
11）https://www.tensorflow.org/datasets/catalog/wiki40b

表 4　不正解選択肢の出力例

問題文： There are three people __ school events.									（立命館大学, 2019）[4]
問題タイプ： 文法		正解選択肢： discussing			不正解選択肢： discuss		discussed	discusses	
比較手法 [5]	debating	talking	discussion	commenting	mentioning	**discuss**	examining	arguing	describing
比較手法 [10]	creating	talking	considering	promoting	deciding	initiating	exploring	reviewing	meeting
比較手法 [7]	talking	dealing	speaking	working	reporting	giving	wednesday	undertaking	tyne
提案手法	discussion	**discuss**	**discussed**	discussions	**discusses**	about	conversation	debate	talked
問題文： They are a little worried __ their daughter's trip to the Amazon.									（森ノ宮医療大学, 2018）[4]
問題タイプ： 機能語		正解選択肢： about			不正解選択肢： for		with	from	
比較手法 [5]	concerning	regarding	relating	talking	what	telling	pertaining	concerned	exactly
比較手法 [10]	considering	up	the	seeing	than	just	discussing	going	out
比較手法 [7]	the	any	and	afraid	affected	anxious	at	after	as
提案手法	what	that	**for**	some	**with**	of	the	much	something
問題文： Would you __ for more coffee?									（産業能率大学, 2018）[4]
問題タイプ： 文脈		正解選択肢： care			不正解選択肢： want		like	drink	
比較手法 [5]	cared	caring	health	cares	healthcare	treatment	patient	quality	education
比較手法 [10]	caring	good	bother	rest	trust	nurse	comfort	life	service
比較手法 [7]	need	**want**	desire	mean	little	fancy	think	prefer	have
提案手法	need	**want**	desire	mean	wish	to	looking	**like**	little

5.2　実験結果

表 3 に実験結果を示す。各問題タイプにおいて、上から 3 行は比較手法、最下行は提案手法の性能を示している。提案手法は、12 件中の 9 設定において最高性能を達成し、残りの 3 設定においても 2 番目に高い性能を示したため、問題タイプの特徴を考慮したフィルタリングの有効性を確認できた。特に、機能語問題において性能の向上が顕著であり、候補数 k の増加に伴ってより大きな改善が見られた。

表 4 に、問題タイプごとの不正解選択肢の出力例を示す。まず、文法問題においては、比較手法は正解単語と意味的に類似する単語を候補として出力する傾向が見られた。一方で、提案手法は実際の不正解選択肢である “discuss” や “discussed” に加えて、“discussion” や “discussions” などの正解単語の活用形を出力しており、文法問題の特徴を反映した候補生成ができた。次に、機能語問題においては、比較手法は機能語以外の単語も出力しており、実際の不正解選択肢は出力できていない。一方で、提案手法は問題の特徴を考慮したフィルタリングによって機能語のみを候補として生成し、実際の不正解選択肢のうち “for” および “with” を上位 5 件に含めることに成功している。また、文脈問題においては、表 3 に示したとおり、上位 5 件程度までは比較手法 [7] と提案手法が同様の候補を出力しているが、提案手法はそれ以降にも良い候補を出力できる場合が多い。

It was certainly __ crowded than I thought it would be.
(a) less　(b) little　(c) least　(d) fewer
(a)が正解選択肢

図 3　自動生成が難しい文法問題の例（会津大学, 2018）[4]

文法問題において、一部の設定では提案手法が有効ではなかった。この要因には、文法問題と分類された問題の中に文脈問題との区別が難しい問題が含まれる点が考えられる。図 3 の例のような、正解単語の活用形を不正解選択肢としない文法問題（66 問中の 16 問）は、提案手法では対処できなかった。

6　おわりに

本研究では、日本の大学入試における英単語穴埋め問題の作問コストを削減するために、問題タイプごとの特徴を考慮した不正解選択肢の自動生成手法を提案した。まず、英単語穴埋め問題には文法問題・機能語問題・文脈問題の 3 タイプが存在することを明らかにし、問題タイプのラベル付きコーパスを構築した。そして、候補のフィルタリングを中心に、各問題タイプの特徴を考慮した不正解選択肢の自動生成手法を提案した。実験の結果、提案手法は多くの設定において比較手法の性能を上回り、不正解選択肢の生成において問題タイプを考慮することの有効性を確認した。今後の課題として、問題タイプを自動分類してコーパスの規模を拡大し、教師あり学習による不正解選択肢の生成を検討したい。

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP21H03564, JP22H00677）の助成を受けたものです。

参考文献

[1] Wilson L Taylor. “Cloze Procedure”: A New Tool for Measuring Readability. **Journalism quarterly**, Vol. 30, No. 42, pp. 415–433, 1953.

[2] Ruslan Mitkov and Le An Ha. Computer-Aided Generation of Multiple-Choice Tests. In **Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2**, p. 17–22, 2003.

[3] Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 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**, pp. 61–68, 2005.

[4] Torsten Zesch and Oren Melamud. 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**, pp. 143–148, 2014.

[5] Shu Jiang and John Lee. Distractor Generation for Chinese Fill-in-the-blank Items. In **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 143–148, 2017.

[6] Yunik Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. Automatic Distractor Generation for Multiple-choice English Vocabulary Questions. **Research and Practice in Technology Enhanced Learning**, Vol. 13, No. 15, pp. 1–16, 2018.

[7] Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 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**, pp. 391–401, 2022.

[8] Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In **Proceedings of the Second Workshop on Building Educational Applications Using NLP**, 2005.

[9] Jennifer Hill and Rahul Simha. 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**, pp. 23–30, 2016.

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

[11] Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 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**, pp. 238–242, 2013.

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

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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**, pp. 4171–4186, 2019.

[14] George A. Miller. WordNet: A Lexical Database for English. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

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

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

[17] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 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)**, pp. 48–53, 2019.

[18] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. **Naval Research Logistics Quarterly**, Vol. 2, 1-2, pp. 83–97, 1955.

[19] Yangqiu Song and Dan Roth. 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**, pp. 1275–1280. Association for Computational Linguistics, 2015.

[20] 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. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.

[21] Steven Bird and Edward Loper. NLTK: The Natural Language Toolkit. In **Proc. of ACL**, 2004.

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