生成 AI による多肢選択式語彙問題および錯乱肢の生成と訂正

内田 諭 [1] 畔元 里沙子 [2] 吉村 理一 [1] 伊藤 薫 [1]
[1] 九州大学 言語文化研究院 [2] 九州大学 地球社会統合科学府

{uchida,r-yoshimura,ito}@flc.kyushu-u.ac.jp risako.azemoto.773@s.kyushu-u.ac.jp

概要

本論では語彙学習を目的とした多肢選択式空所補充問題の作成のため生成 AI 等の技術を活用し，その性能を検証した．最初に英単語 300 語に対し，ChatGPT にプロンプトを与え問題文と錯乱肢を生成した．錯乱肢には人手で訂正の有無をアノテーションし，ChatGPT と Gemini を用いて自動訂正の可能性を検討した．結果として，問題生成については訂正が不要であった一方，錯乱肢生成では約 17.8%で訂正を要した．この結果を受けて，生成 AI の錯乱肢の訂正能力を検証したところ，F1 値が最大 0.291 となり，不十分な結果となった．提案手法として，予め準備した錯乱肢を BERT の確率により選別する方式を試した結果，訂正が必要な場合は 11.0%となり，生成 AI に対する優位性が明らかになった．

1 はじめに

LLM の台頭により自然言語処理技術は飛躍的な進歩を遂げたが，ChatGPT の登場はその技術の利用者層を格段に広げた．英語教育においてもその影響は大きく，翻訳，質問応答，文章生成，要約など様々なタスクをこなすことが可能なチャットボットは，教育者・学習者双方から注目されている．我々は英語教育における生成 AI の活用先の一つとして語彙問題生成に着目し，能力を検証した．

本論では，語彙問題生成の最高性能達成ではなく，実世界でのタスクにチャットボット応用した際に，どの程度の性能が得られるかを把握することを目的としている．具体的には，英語教員が汎用チャットボットを使用し，授業で使用する多肢選択式語彙問題を生成するという状況を想定している．そのため，アルゴリズムの改良や fine-tuning など複雑な操作が必要な方法ではなく，zero-shot や few-shot，プロンプト，単純なプログラミングなどにより実行できる手法でどの程度適切な問題が得られるかを明らかにすることがゴールとなる．

本論で生成したい問題は多肢選択式の空所補充問題であり，問題文，正解肢 1，錯乱肢 3 からなるフォーマットとした．以下に例題を示す．

> Masked-language model is a(n) _____ approach for improving natural language processing.
> **(a) successful** (b) devastated (c) tiny (d) optical

2 関連研究

Yoshimi ら [1] はコサイン類似度を用いた先行研究 [2] や BERT を用いた先行研究 [3] を踏まえ，英語多肢選択式問題の錯乱肢生成に取り組んでいる．当該研究では，日本の大学入試問題から収集した問題文に対し、問題タイプ（文法・機能語・文脈）を分類した上で，単語埋め込み表現による生成, ランキング，フィルタリングを行うことで錯乱肢を生成している。（生成 AI による検証は行っていない）．

主に読解に焦点を当てた多肢選択式問題生成の研究 [4] では，LLM とプロンプトエンジニアリングを用いることである程度質の高い問題が生成できることを明らかにすると同時に，ChatGPT が空所補充問題等，特定の形式の問題生成に適さない可能性を報告している．また，AI と教師の協働についても議論しており，人手介入の必要性も示唆している．

Wang ら [5] は ChatGPT3.5 turbo を使って穴埋め方式の多肢選択式問題の生成を行った．生成された 60 問のうち，設問文は 75%，錯乱肢は 66.85%で適切なものであったとし，LLM を利用することで旧来の手法から大幅に精度が上がったと報告している．しかし，最終的には人の目での確認も必要であることを示唆している．

以上のように，多肢選択式語彙問題の生成において，一定の質が担保できることは明らかであるが，同時にマニュアルでの確認も必要であることがわか

る．教材やテストのような教育に影響の大きいマテリアルでは教員が確認することが理想ではあるが，このプロセスもある程度自動化できることが望ましい．そこで本研究では，(1) 最新の LLM によって問題生成とその精度検証を行った後，(2)LLM によってどの程度訂正が可能であるかを複数の LLM やプロンプトによって検証する．

3　データセット

英語学習者にとって適切な難易度の単語となるよう，主に大学 1 年生の語彙学習を目的とした『九大英単』[6] に収録されている名詞，動詞，形容詞から，品詞ごとにそれぞれ 100 語の計 300 語を正解肢として使用した．これらは英語の授業で実施した単語の和訳を選択する形式の小テストで正解率が低いものを選んだ．正解率の低いものを使用した理由は，学習価値のある英単語を選定するためである．また，正解率の分布は高い方に偏っているため、低い方から選ぶことにより幅広い難易度の単語を選ぶことができる．

4　生成 AI を用いた手法

本節では，(1) ChatGPT を使った問題生成とその精度の検証，(2) ChatGPT および Gemini を使った訂正能力の検証の 2 つを行う．

(1) 問題生成に関しては，簡単な指示と出力すべき例を数例与える few shot learning を用いて（プロンプトは付録 A.1 参照），ChatGPT4o (執筆時点で最新のモデルである gpt-4o-2024-11-20) で問題の生成を行う．API (temperature=0，その他はデフォルト）を用いて JSON 形式で出力する．

AI による生成結果は，多くの場合で人手でチェックすることになる．特に教育応用を考えた場合，適切な問題文になっているかという確認は重要な手続きである．多肢選択式空所補充の問題の場合，正答が一意に決まるか，すぐに排除できる錯乱肢が含まれていないか，選択肢の品詞や時制，人称が統一されているか，選択肢の難易度のバランスは適切か，など確認すべき点は多岐に渡る．これらをすべて考慮して問題を確認するには多大な労力を要する．

(2) そこで，生成された問題および錯乱肢の適切性について，複数の生成 AI を用いて (A) 詳細な指示のないチェック（zero：付録 A.2 参照）と (B) 詳細な指示のあるチェック（note：付録 A.3 参照）の複数のパターンを検証し，もっとも効果の高いものを明らかにする．また，ベースラインとして BERT[7](Hugging Face の google-bert/bert-base-uncased モデルを利用）による確率による選択肢のチェックを行い，その結果と比較する．なお，ターゲットとなる単語が複数の要素に tokenize される場合は，それらの平均確率で代表することとする．

4.1　アノテーションの方法

ChatGPT4o を用いてターゲット語彙を対象とした 300 問を生成し，問題文，正解肢，錯乱肢（3 つ）をデータセットに含めた．これに対して，英語学または英語教育学が専門で日本の高等教育機関での教授経験を有する著者のうちの 2 名が独立に問題文，正解肢，錯乱肢の可否を判定した．なお，錯乱肢の訂正候補については一意に決めることが難しいため，今回は訂正が必要かどうかという点のみでアノテーションを行った．300 問中 294 問で判定が一致し（一致率 98 %），不一致のあった 6 件については合議の上，最終的なアノテーションを決定した．(公開先 https://github.com/discoursemarker/q-aacd/)

4.2　多肢選択問題の生成とその結果

問題文および正解肢については訂正が必要だと判断された項目はなかった．しかしながら，正解肢はすべて原形（入力と同じ形）で出力されており，多様な文脈での用例を生成するにはプロンプトを工夫する必要があることが示唆される．

錯乱肢について，900 件（300 問× 3）中 161 件（約 17.8%）が訂正が必要だと判断された．訂正が必要なものの中には問題文に含まれる冠詞によって文脈的に不適合なもの（例：an <affiliate> に対して competitor, rival が錯乱肢として生成される場合など）や錯乱肢も正答になりうると判断されたものなどが含まれている．この結果から，80 %以上のケースでそのまま使えるということになるが (Wang ら [5] の研究と比較すると LLM のモデルがアップデートされたことで精度が高くなっていることも伺える)，1 問ごとに生成されたにも関わらず選択肢に偏りが見られるという傾向が明らかになった．頻度が多いものは ignore(35 回), confusion(12 回), ordinary(10 回), boring(6 回), confusing(6 回), discussion(6 回), temporary(6 回), whisper(6 回) などであるが，全体で見たときにこのような偏りがあると，容易に不正解として排除できる可能性がある．

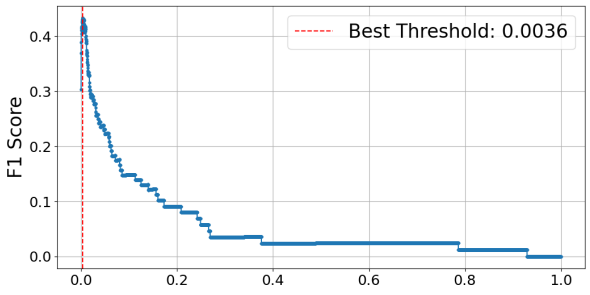

図 1　BERT の確率と F1 スコア

4.3　多肢選択問題のチェック

4.3.1　BERT

それぞれの錯乱肢について，BERT で文脈に生起する確率を計算し，アノテーション済みのデータと照合して Python の sklearn によって F1 Score が最大となる確率の閾値を計算した．その結果，確率 0.0036 を基準とした場合，F1 スコアは 0.434 となることがわかった（図 1 参照）．また，正解率・適合率・再現率は表 1 に示す．

4.3.2　生成 AI

次に生成 AI を用いた訂正の精度について検証する．ここでは問題生成にも利用した ChatGPT4o と，別の LLM である Gemini を用いて，詳細指示なし (zero）および詳細指示あり（note）の 2 つのパターンを検証する．さらに，Gemini の推論モデル (gemini-2.0-flash-thinking-exp-1219) の性能も合わせて検証する．結果は表 1 の通りである．

この結果から以下のような傾向が読み取れる．(1) まず，問題生成に用いたものと同じ LLM を利用した場合，極端に再現率が低くなる．実際，訂正が行われたのはそれぞれ 7 件（zero）と 4 件 (note）に留まり，自己訂正が十分に機能しないことが示唆される．(2) 一方，別の LLM を用いた検証では再現率が比較的高くなる傾向にあり，特に詳細な指示をした場合 (note) はその傾向が顕著である．Gemini_note では 900 件中 754 件 (約 83.8 %) が訂正されているが，これは過剰な訂正になっている．(3) 推論モデルを用いてもそれほど精度（F1 スコア）は高くならない．(4)BERT を用いたモデルがもっとも F1 スコアが高く，バランスがよい．以上の点から，生成 AI を用いて錯乱肢を訂正するということには困難が伴うことが示唆される．

表 1　各モデルの評価指標

	正解率	適合率	再現率	F1
BERT	0.782	0.405	0.466	**0.434**
ChatGPT_zero	0.818	0.286	0.012	0.024
ChatGPT_note	**0.821**	**0.500**	0.012	0.024
Gemini_zero	0.719	0.201	0.193	0.197
Gemini_note	0.279	0.176	**0.826**	0.291
Gemini_thinking	0.773	0.264	0.149	0.190

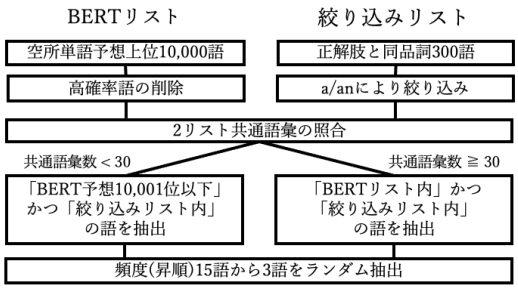

図 2　絞り込みリストを用いた手法の概略

5　リストを用いた錯乱肢生成

生成 AI を用いた手法では，錯乱肢の頻度の偏りや，空所直前の *a/an* により排除できてしまう錯乱肢の生成などの問題が生じた．これらはプロンプトによって容易には解決できなかったため，本論では BERT と候補リストを用いた錯乱肢生成を考案し，その結果を人手により評価した．本節では，その方法と結果について示す．

5.1　錯乱肢の生成方法

手法の概略を図 2 に示す．最初に前節で生成した問題文からランダムに 100 問を選び，その空所を BERT に fill-mask タスクとして解かせ空所に入る単語を予想させた．予想結果は高確率なものから上位 10,000 語出力し，単語と生起確率の組を得た．さらに，正解選択肢となりうる語を排除するため，生起確率が 4.3.1 で得た確率 0.0036 よりも十分小さい確率 1.0×10^{-6} を超える語彙を削除した．この手続きによって得た単語の集合を **BERT リスト**と呼ぶことにする．

次に，**絞り込み用リスト**の作成について示す．英語学習者は，学習者の習熟度に対し容易すぎる単語や難しすぎる単語を不正解として排除することが容易であると想定できる．このため，ある程度錯乱肢に使用する単語を絞り込んでおくことが有用であ

ると考えられる．絞り込み用リストはこのような絞り込みを行うためのリストである．本論では，『九大英単』に含まれる名詞 490 語, 動詞 386 語, 形容詞 503 語のうち，正解選択肢と同じ品詞を絞り込み用リストとして使用した．生成 AI を用いた手法では空所直前の *a/an* による問題が生じたため，問題文の空所直前に *a* が出現する場合は a, e, i, o, u で始まる単語を，*an* が出現する場合はそれ以外の単語を絞り込み用リストから削除した．

続いて，BERT リストと絞り込み用リストを照合しつつ，最終的な錯乱肢を選定した．まず，BERT リストと絞り込み用リストの双方に含まれる単語が十分多く含まれるかどうかを調べ，十分な場合は 2 つのリストの共通部分を，不十分な場合は絞り込み用リストのうち，BERT の予想した生起確率上位 10,000 語に含まれない語を抽出した．閾値は 30 語とした．さらに，錯乱肢の頻度を制御するため，それまでに使用した錯乱肢の頻度が低い順に残った単語を並び替え，頻度の低い 15 語を残した．そのうち 3 語をランダム抽出し，最終的な錯乱肢とした．なお，錯乱肢の頻度を制御した結果，すべての錯乱肢の頻度が 2 以下となったため，選択肢の偏りは解消されたとみなすことができる．

5.2　人手による評価

4.1 でアノテーションを行った 2 名が，本節の手法で生成した錯乱肢 300 (100 問 × 3) に対しても独立に訂正の要否を判定した．判定が異なる点については合議の上統一した．その結果，300 件中 33 件 (11.0%) が要訂正と判断された．訂正が必要な原因の多くは錯乱肢が正確肢の品詞と異なるためであり，33 件中 30 件 (90.9%) を占めている．生成 AI で問題文を作成する際に入力した単語には複数の品詞として解釈できる可能性を持つものがあり，使用者の意図とは異なる品詞に合わせて問題文が生成されていた．この点は生成 AI による錯乱肢生成では障害にならなかったが，予め錯乱肢の品詞を限定した当該手法では性能を下げる原因となった．LLM が品詞を正しく理解できるかという点については検討が必要な課題である（cf. [8]）．

4.2 に示した ChatGPT による錯乱肢生成では 17.8% に訂正が必要と判断されたが，当該手法では訂正が必要な選択肢は 11.0% に押さえられた．また，問題文における正解肢の品詞をあらかじめ出題意図に沿ったものにすることで，訂正が必要な割合はさらに抑えることができると考えられる．一方で，今回定量的に評価していない性質として，絞り込み用リストを用いた手法では正解肢と錯乱肢の意味が遠く，正解肢の意味を知らずとも消去法で正解を導きやすい問題となりやすいのに対し，ChatGPT による正解肢と錯乱肢の意味が近く，良い意味で難易度の高い問題となりやすいことがアノテーション時に観察されている．典型例を表 2 に示す．

表 2　各手法による難易度の典型例

問題文	After being sick for a week, he felt too <feeble> to go back to work.
ChatGPT	strong / energetic / sturdy
絞り込みリスト	local / minus / naval

注：上段の <> は空所の箇所および正解肢を表す．下 2 段は各手法で生成した錯乱肢．

5.3　2 手法についての総評

ChatGPT による錯乱肢生成では正解になりうる（正解が一意に決まらない原因となる）選択肢の生成や，冠詞による文脈不適合，語彙の偏りといった問題が生じた．絞り込みリストと BERT を用いた手法では明示的なプログラミングで冠詞や語彙の偏りを排除し，正解になりうる選択肢の生成も大幅に抑えることができた．人手訂正が少なく済むという点では絞り込みリストによる手法は有用だと思われるが，難易度や良問か否かといった観点では評価できていない面も残る．錯乱肢の生成は，文脈を考慮しつついかに正解肢から適切な類似度の語彙を選ぶかという問題に帰着できることが示唆される．

6　おわりに

実験の結果，生成 AI が生成した問題文には訂正の必要がなく十分な性能を得られていると考えられる一方，錯乱肢には訂正が必要なものが（実用性という観点からは）多く見られた．また，生成 AI は錯乱肢の訂正性能が十分でないことが明らかになった．特に，問題生成と同じモデルを使用した場合は F1 値が極端に低い結果となり，BERT を用いたモデルが最高性能を示した．絞り込みリストと BERT を用いた錯乱肢生成手法は，訂正の必要性という観点では生成 AI よりも錯乱肢生成において高い性能を発揮しており，問題文生成の際に品詞を制御できれば高い実用性があると言える．しかし，問題の難易度という観点では生成 AI の方が有用な可能性もあり，今後は多角的に性能を検証する必要がある．

謝辞

本研究は九州大学人社系学際融合プログラムおよび JSPS 科研費 JP20H00095，JP23K21949 の助成を受けたものです.

参考文献

[1] Nana Yoshimi, Tomoyuki Kajiwara, Satoru Uchida, Yuki Arase, and Takashi Ninomiya. Distractor generation for fill-in-the-blank exercises by question type. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)**, pp. 276–281, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] Shu Jiang and John Lee. Distractor generation for Chinese fill-in-the-blank items. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis, editors, **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 143–148, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Chak Yan Yeung, John Lee, and Benjamin Tsou. Difficulty-aware distractor generation for gap-fill items. In Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay, editors, **Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association**, pp. 159–164, Sydney, Australia, 4–6 December 2019. Australasian Language Technology Association.

[4] Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. **Education and Information Technologies**, Vol. 29, pp. 1–33, 10 2023.

[5] Qiao Wang, Ralph Rose, Naho Orita, and Ayaka Sugawara. Automated generation of multiple-choice cloze questions for assessing English vocabulary using GPT-turbo 3.5. In Mika Hämäläinen, Emily Öhman, Flammie Pirinen, Khalid Alnajjar, So Miyagawa, Yuri Bizzoni, Niko Partanen, and Jack Rueter, editors, **Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages**, pp. 52–61, Tokyo, Japan, December 2023. Association for Computational Linguistics.

[6] 九州大学英語表現ハンドブック編集委員会 (編). 九大英単ー大学生のための英語表現ハンドブック. 研究社, 2014.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Satoru Uchida. Using early llms for corpus linguistics: Examining chatgpt's potential and limitations. **Applied Corpus Linguistics**, Vol. 4, No. 1, p. 100089, 2024.

A 付録：プロンプト集

A.1 問題生成

Instruction

Create a fill-in-the-blank question for CEFR A2-B1 level English learners, where learners must choose the correct answer from four options, including one correct answer (target word) and three distractors (incorrect options). Avoid creating options that can be eliminated solely based on grammar, such as articles (a/an) or verb tenses.

The target word

{target}

Output Format

Provide the question in the following JSON format:

{{"question": "[sentence_with_blank]", "correct_answer": "[target_word]", "distractors": ["[distractor_1]", "[distractor_2]", "[distractor_3]"]}}

- "sentence_with_blank": A sentence where learners must fill in the blank with the target word. The blank should be marked as <blank>. - "target_word": The correct answer. - "distractors": List here the items that cannot be considered correct answers based on meaning or grammar.

Examples

[{{"question": "She received an award for her <blank> performance in the competition, which impressed all the judges.", "correct_answer": "outstanding", "distractors": ["forgettable","distracting","smooth"]}},
{{"question": "The artist was known for his meticulous <blank> of famous paintings, making them almost indistinguishable from the originals.", "correct_answer": "reproduction", "distractors": ["destruction","misinterpretation","translation"]}},
{{"question": "The company decided to <blank> a team to investigate the issue.", "correct_answer": "dispatch", "distractors": ["delay","terminate","acquire"]}}]

A.2 問題チェック：zero

Instruction

Below is a fill-in-the-blank multiple-choice question in the JSON format shown in "#Input Format." Please check the appropriateness of each item, and if there are any issues with the English quiz, provide a corrected version in the same JSON format. Do not add any extra comments before or after the json part (only json part should be generated).

#Input Format

{{"question": "[sentence_with_blank]", "correct_answer": "[target_word]", "distractors": ["[distractor_1]", "[distractor_2]", "[distractor_3]"]}}

#Target

{target}

A.3 問題チェック：note

Instruction

Below is a fill-in-the-blank multiple-choice question in the JSON format shown in "#Input Format." Please check the appropriateness of each item, and if there are any issues with the English quiz, provide a corrected version in the same JSON format with revisions. Be sure to follow the guidelines provided in the "#Notes" section. Do not add any additional comments before and after the JSON format.

#Input Format

{{"question": "[sentence_with_blank]", "correct_answer": "[target_word]", "distractors": ["[distractor_1]", "[distractor_2]", "[distractor_3]"]}}

#Target

{target}

#Notes

(1) The correct answer must be unambiguous and determined by the context. Pay special attention when using synonyms or antonyms as options.
(2) Articles should not provide hints that eliminate options. For example, in the case of "an <blank>", all options must begin with a vowel sound. In the case of "a <blank>", no options should begin with a vowel.
(3) All options must be of the same part of speech.
(4) If verbs are used as options, their tense must be consistent.
(5) The same word should not appear as multiple options.
(6) Please make changes to the options only when absolutely necessary. If changes are made, use words that are approximately at the A2 to B2 level according to the CEFR, and avoid overly difficult terms.
(7) The "correct_answer" must not be altered.
(8) The question text ([sentence_with_blank]) should only be modified when absolutely necessary.