A Task of Cloze Explanation Generation for ESL Learning

Zizheng Zhang ♠ Masato Mita ♠♣
Mamoru Komachi ♠♡

♠ Tokyo Metropolitan University. 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
♣ CyberAgent, Inc. 2-24-12 Shibuya Shibuya-ku, Tokyo 150-6121, Japan
♡ Hitotsubashi University. 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan
zhang-zizheng@ed.tmu.ac.jp, mita_masato@cyberagent.co.jp,
mamoru.komachi@r.hit-u.ac.jp

Abstract

Addressing the lack of dedicated tasks for generating language learner explanations of cloze questions, this paper introduces a new task focused on cloze explanation generation in language assessment, particularly for English as Second Language (ESL) learners. To support this task, we present a meticulously curated data creation method, which expands expert-designed high-quality cloze questions and explanations. This data aims to assess language proficiency and facilitates language learning by offering informative and accurate explanations.

1 Introduction

Cloze questions [1] are a fundamental component of language assessment (LA). They typically consist of a sentence or a passage with certain words or phrases omitted, and language learners are required to select or fill in the most appropriate word to complete the text. Cloze questions in language educational settings are usually used in language educational settings to evaluate language proficiency in terms of various aspects such as grammatical knowledge [2, 3] and reading comprehension skills [4, 5]. They are also widely employed in famous tests for English as a Second Language (ESL) learners, such as the International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL).

Explanations for cloze questions play a crucial role in language learning, particularly in self-study contexts. When learners encounter challenging cloze questions, having access to clear and concise explanations after answering the question can greatly aid their understanding of the correct answers [6]. Explanations provide learners with insights into the reasoning behind the correct and incorrect choices, helping them identify and rectify their own misconceptions. The provision of high-quality explanations can empower learners, fostering deeper comprehension and long-term knowledge retention.

Question:
The Westchester Philharmonic received a national _____ education program three years ago.
(A) awardable (B) award
(C) misawarded (D) well-awarded

Explanation:
The blank needs a word that fits as the object of "received," and it should follow the determiner "a" and the adjective "national." Therefore, we need a singular or mass noun. Option (A) "awardable" and option (D) "well-awarded" are adjectives. Option (B) "award" is a singular noun. Option (C) is a verb in past tense. Therefore, the correct answer is (B) "award."

Table 1: Examples of data for the ClozEx task. A system is expected to receive **Question** as input and produce **Explanation** as output. Words in green denote "hint words" that help reason the answer.

However, despite its usefulness, there has been almost no work on generating high-quality explanations for given cloze questions. One essential reason is that no dataset for such a task is available. Because of high costs in terms of time and human effort, employing experts to create such a dataset, to provide abundant data, is difficult, although it could guarantee the quality of the dataset. Furthermore, even if a dataset could be constructed, it would be challenging to automatically generate human-like cloze explanation.

To address these challenges, we propose the a natural language generation (NLG) task `ClozEx` of generating explanations for English cloze questions. Intuitively, a good explanation that helps answer a cloze question should be easy to read and provide sufficient background knowledge. Therefore, fluency and informativeness should be considered in explanation generation. We also provide a method to create expert-quality-assured (question, explanation) pairs automatically, an example of which is shown in Table 1.

The main contributions of this work are (1) we propose a new task toward generation of fluent and valid English cloze explanation (`ClozEx`) for ESL learning; (2) we designed a method for creating cloze questions and explanations, which is based on expert-designed data thus the quality is ensured.

2 Related Work

The effectiveness of cloze questions in language assessments has spurred research into their automatic generation. Unlike earlier studies using simplistic methods like fixed ratio word deletion and random distractor selection, modern research aims to enhance question validity. For instance, Sakaguchi et al. [7] utilized a corpus of English learner errors to create more challenging distractors, thus improving the assessment of language proficiency. Further studies [8, 9, 10, 11, 12] have explored features influencing cloze question validity, including part of speech, n-gram frequency, and word sense, leading to more targeted question generation. Despite these advances, the creation of explanatory content alongside questions remains largely unaddressed, an area where this research innovates by focusing on explanation generation.

Nagata [13] introduced the Feedback Comment Generation (FCG) task, which automatically generates feedback for English writing exercises for non-native learners. While FCG aids in grammar learning, it's limited in offering systematic grammar knowledge. It relies on free composition, which doesn't cover all grammar aspects comprehensively. Additionally, FCG mainly explains word appropriateness in sentences, not why some expressions should be avoided. Also, producing high-quality feedback from free compositions is labor-intensive. In contrast, this research's `ClozEx` task uses a top-down approach to construct cloze questions based on grammar item "patterns." This method enables the automated generation of numerous high-quality explanations.

3 `ClozEx` Task and Data Creation

3.1 Task Definition

Methods devised to address the `ClozEx` task are expected to operate on a cloze question q as input. A cloze question comprises a sentence with a blank, denoted as $sent$, and a set of options $OPT = [opt_1, opt_2, ..., opt_n]$ (typically, n equals to 4). The objective of the methods is to generate an explanation text exp as output for the given question. The generated explanation should satisfy two criteria: (1) fluency [14], meaning that the explanation should be coherent and easily comprehensible, because an explanation that is difficult to read would not effectively aid language learning; (2) validity [15], indicating that the explanation should provide sufficient information, such as relevant language knowledge, to facilitate answering the question accurately.

3.2 Data Preparation

To create high-quality data for cloze questions, employing English education experts for writing explanations is ideal but time-consuming and not scalable. To overcome these limitations, we propose an automated method for generating both questions and explanations.

Experts typically construct cloze questions in a top-down approach, focusing on a specific grammatical item. These items serve as a pattern for a group of questions. Our method, based on this pattern approach, automates the creation of new cloze questions and their explanations by extracting and utilizing patterns from expert-designed materials, ensuring quality retention.

Our data creation process, depicted in Figure 1, starts with extracting patterns from expert-crafted questions and explanations. Using these patterns, we generate new questions and explanations from a public corpus. In the question generation phase, relevant sentences from a news corpus are matched with the identified patterns. Distractors are created based on the targeted language aspect. For generating explanations, we design templates specific to the question type, filled with relevant details from the question and pattern. These explanations are then refined and diversified using Language Models (LLMs) to enhance fluency, with a set limit of 128 words to avoid redundancy.

Initially, our focus is on two common cloze question types in language assessment, particularly in tests like

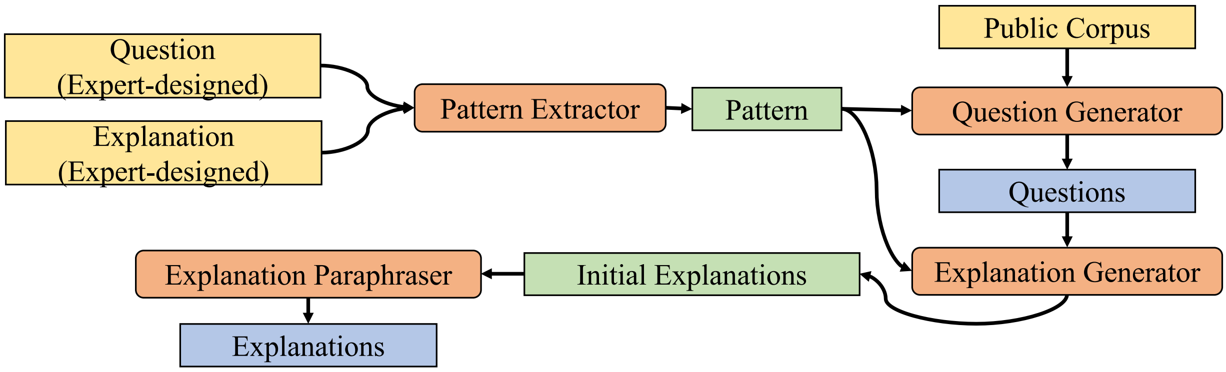


Figure 1: Pipeline of data creation method. Yellow rectangles symbolize input to the pipeline, whereas blue rectangles represent output. Modules are depicted in orange, and their corresponding intermediate results are highlighted in green.

TOEIC: affix and verb-tense questions. These types were chosen for their significance in language proficiency evaluation. Affix questions challenge ESL learners to discern different parts of speech through prefixes or suffixes, while verb-tense questions require identifying the correct tense in a given sentence.

3.3 Pattern Extraction

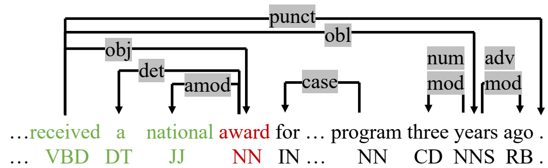
Affix/tense questions necessitate ESL learners to identify and analyze a specific context referred to as "hint words," which serve to modify or be modified by the word in the blank to answer the question accurately. To capture the patterns inherent in these questions, we focus on the relationship between the hint words and the answer option.

To extract the pattern from each expert-designed question, we begin by inserting the answer option into the sentence, resulting in a completed sentence denoted as $sent^{ans}$. Next, we extract the hint words from the expert-designed explanation, and we mark their corresponding positions in $sent^{ans}$ (see (a) in Figure 2). Subsequently, we employ dependency parsing on $sent^{ans}$ to generate its dependency tree. Given that the hint words and the answer option play crucial roles in the question, we extract a sub-tree from the dependency tree that encompasses all the hint words and the answer node. This sub-tree serves as the pattern for the question and is denoted as $pattern$ (see (b) in Figure 2, the pattern could be summarized as "A noun works as an object that is modified by an article and adjective.").

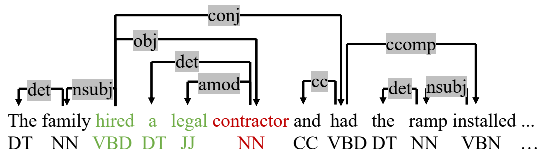
After obtaining the pattern for a specific question, we utilize it to generate new questions. We parse all sentences, denoted as $[\tilde{sent}_1, ..., \tilde{sent}_m]$, from publicly available news corpus to acquire their respective parsing trees, denoted as $[\tilde{tree}_1, ..., \tilde{tree}_m]$. We use a news corpus because news is

The Westchester Philharmonic received a national award for its education program three years ago.

(a) Example of $sent^{ans}$; red word represents the answer option, and green ones denote hint words extracted from expert-designed explanation.



(b) Partial dependency parsing tree of $sent^{ans}$ in (a). Only nodes of colored words are extracted as $pattern$ (Pattern in Figure 1).



(c) Partial $\tilde{sent}_i$ and its dependency parsing tree. Because $\tilde{tree}_i$ consists of $pattern$ (marked in colored text), $\tilde{sent}_i$ could be used to generate a question.

Question:
The family hired a legal _____ and had the ramp installed at the front of their home at the Woodlands at Copperstone in Brentwood.
(A)contractual (B) contractor (C) contracted (D) contractable
Initial Explanation:
The word in the blank should be the object of "hired".
"a" is the determiner of the blank. "legal" is the the adjective modifier of the blank.
Thus, a Noun, singular or mass is required.
(A) contractual is a Adjective. (B) contractor is a Noun, singular or mass. (C) contracted is a Verb, past tense. (D) contractable is a Adjective.
Therefore, the correct answer is (B) contractor.

(d) Example of generated question and corresponding initial explanation (Initial Explanations in Figure 1).

Figure 2: Examples of process of generating a new question with its explanation.

in formal writing and leads to fewer grammatical errors. If a parsing tree, $t\tilde{ree}_i$, includes the extracted pattern $pattern_j$, we consider the corresponding sentence, $s\tilde{ent}_i$, as a suitable candidate for generating a new question that belongs to $pattern_j$. It is important to note that our focus lies in capturing the modification relationship between the hint words and the answer option (e.g., dependency relations), and their grammatical classes within the sentence (e.g., POS), rather than the specific words used in the question generation process (see (c) in Figure 2).

To select distractors for the new question, we built candidate dictionaries for affix and verb-tense questions, respectively. Distractor options are selected from the corresponding dictionary. For example, if an affix question has the answer option “contractor”, the distractor candidates could be in [“contractual”, “contraction”, “contracted”, “contractable”]. Similarly, distractor options for verb-tense questions are also selected from another pre-defined dictionary.

Finally, we design templates for specific types of questions to present all the necessary information for answering the question, including *pattern* and options (see (d) in Figure 2). To improve fluency and diversity, we employ LLM to paraphrase the template-based explanation.

3.4 Data Analysis

To validate the quality and suitability of our created data for training models in the `ClozEx` task, we conducted a thorough manual quality assessment. As outlined in Section 3.1, the evaluation focused on two aspects: fluency and validity.

For the fluency assessment, we enlisted the expertise of two native English speakers from the university with which the authors are associated. These experts independently evaluated 100 randomly selected instances from our data using a 5-point Likert scale (1 denotes the worst and 5 denotes the best), solely considering the fluency of the generated explanations and disregarding their validity. To evaluate the validity aspect, we recruited four advanced ESL learners [1)] from the university with which the authors are associated, because these learners possess a strong understanding of textbook grammar [16]. Similarly, these annotators used a 5-point Likert scale to assess the validity of 100 instances. To ensure the independence between fluency and validity, we selected fluent instances in advance for the validity estimation. The validity assessment aimed to determine whether the explanations provided the necessary information to answer the corresponding question.

	IAA		Estimation Score		
	Pear.	p-value	Avg.	Med.	Var.
Fluency	0.82	<0.001	4.29	4.00	0.52
Validity	0.77	<0.001	4.51	4.50	0.45

Table 2: Inter-annotator agreement and manual estimation result. **Pear.** denotes Pearson’s correlation coefficient. **Avg.**, **Med.**, and **Var.** indicate the average, median, and variance of scores, respectively.

To ensure robustness, each instance underwent double annotation for both fluency and validity. We performed the Pearson correlation test to assess the inter-annotator agreement between the different annotators. Result of inter-annotator agreement and manual estimation are shown in Table 2. The high correlation coefficients indicate a strong agreement among the annotators, underscoring the reliability of our manual estimation. The scores for both fluency and validity exhibited high median values and low variance. These findings confirm the high quality of our data and support its publication as a reliable resource for the `ClozEx` task.

Further, to investigate whether LLMs could provide receivable explanations without any training, we also asked annotators to estimate explanations generated by various LLMs. The result demonstrate that though LLMs are good at generating fluent text, these text generally do not explain cloze questions well. Please refer to Appendix A for details.

4 Conclusion

This paper introduced a novel task, `ClozEx`, aimed at generating fluent and valid explanations for English cloze questions to support ESL learning. We designed a comprehensive data creation method, which aims to expand expert-designed cloze questions and explanations. Generated questions and explanations, which are based on expert-designed ones, ensure the quality in the cases of LA and ESL learning. Expert evaluations also confirmed the high quality and suitability of our data for the `ClozEx` task.

1) They hold public English test certificates to indicate they have a CEFR A2 level or higher.

Acknowledgment

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2139.

References

[1] Wilson L Taylor. "Cloze procedure": A new tool for measuring readability. **Journalism quarterly**, Vol. 30, No. 4, pp. 415–433, 1953.

[2] James Rye. **Cloze procedure and the teaching of reading**. London, 1982.

[3] J Charles Alderson. The cloze procedure and proficiency in English as a foreign language. **TESOL quarterly**, pp. 219–227, 1979.

[4] Patricia M Raymond. Close procedure in the teaching of reading. **TESL Canada journal**, pp. 91–97, 1988.

[5] Christine Klein-Braley. C-tests in the context of reduced redundancy testing: An appraisal. **Language testing**, Vol. 14, No. 1, pp. 47–84, 1997.

[6] Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. Why does explaining help learning? insight from an explanation impairment effect. In **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 32, 2010.

[7] 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 (Volume 2: Short Papers)**, pp. 238–242, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[8] Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. Automatic generation system of multiple-choice cloze questions and its evaluation. **Knowledge Management & E-Learning: An International Journal**, Vol. 2, No. 3, pp. 210–224, 2010.

[9] Rui Correia, Jorge Baptista, Maxine Eskenazi, and Nuno Mamede. Automatic generation of cloze question stems. In **International Conference on Computational Processing of the Portuguese Language**, pp. 168–178. Springer, 2012.

[10] 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.

[11] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 423–438, 2020.

[12] 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, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] Ryo Nagata. Toward a task of feedback comment generation for writing learning. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3206–3215, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Jennifer Cotter. Understanding the relationship between reading fluency and reading comprehension: Fluency strategies as a focus for instruction, 2012.

[15] Charles B. Cross. Explanation and the theory of questions. **Erkenntnis (1975-)**, Vol. 34, No. 2, pp. 237–260, 1991.

[16] Eileen W Glisan and Victor Drescher. Textbook grammar: Does it reflect native speaker speech? **The Modern Language Journal**, Vol. 77, No. 1, pp. 23–33, 1993.

[17] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. DeID-GPT: Zero-shot medical text de-identification by GPT-4. **arXiv preprint arXiv:2303.11032**, 2023.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

A LLMs for Cloze Explanation Generation

Given the remarkable performance of LLMs across various tasks without fine-tuning [17], there is a reasonable expectation that they would excel in generating high-quality explanations for cloze questions.

We conducted manual estimation for explanations that generated by various LLMs in a zero-shot scenario, including GPT2-large, GPT2-XL, GPT3.5-turbo [2], and LLaMa-7B [18]. The estimation process followed Section 3.4.

Our experimental (Table 3) findings indicate that no LLM achieved an acceptable validity score in manual evaluation. Upon analyzing the explanations generated by GPT3.5-turbo, we identified two critical shortcomings of LLMs in effectively explaining cloze questions.

Firstly, LLMs exhibit a tendency to generate factual errors, thereby failing to ensure the accuracy of the generated texts. This deficiency is exemplified in **GPT3.5-turbo Question 1** Appendix B, where an evident error is observed in the verb tense following the word "did not," a discrepancy that can have detrimental consequences in the context of LA.

Secondly, LLMs have the propensity to produce explanations that lack meaningful and informative content, failing to provide the necessary knowledge required for comprehending the reasons behind the answer options. As illustrated by **GPT3.5-turbo Question 2** in Appendix B, such explanations leave ESL learners unaware of why the given answer option is necessary, while also failing to elucidate the distinctions among the options resulting from affixes. Furthermore, these explanations may even present incorrect answers and flawed analyses, further diminishing their utility.

	Fluency	Validity
GPT2-large	3.97	1.73
GPT2-XL	3.97	1.70
GPT3.5-turbo	**4.53**	**2.70**
LLaMa-7B	3.81	1.78

Table 3: Manual estimation result of LLMs. All scores are ranged in [1, 5].

B Examples from GPT3.5-turbo

Question 1: Isn't it just that it didn't ____ to sell the sets it had produced?
(A) manage (B) manages (C) managing (D) managed

Explanation 1: The correct answer is (D) managed. The sentence is in the past tense and "did not" is the negative form of the auxiliary verb "did". Therefore, the next verb has to be in the past tense as well. The verb "managed" is the correct form in the past tense, meaning "succeeded in doing something despite difficulties or obstacles". Example: "The company didn't manage to sell all of its products before the end of the year."

Question 2: The innards of the machine were complex but ____.
(A) manageable (B) manageably (C) unmanageably (D) mismanage

Explanation 2: The correct option is (C) unmanageably. The word "innards" means the internal parts of a machine. The sentence suggests that the parts of the machine were complex and difficult to manage. The word "manageably" (option B) means something that is able to be managed easily which contradicts the sentence. The word "mismanage" (option D) means to handle or operate wrongly or inefficiently, which does not fit in the sentence. Therefore, the correct answer is (C) unmanageably, meaning that the innards of the machine were difficult to manage.

2) https://platform.openai.com/docs/models