Cloze Quality Estimation for Language Assessment

Zizheng Zhang † and **Masato Mita** ‡† and **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
zhang-zizheng@ed.tmu.ac.jp, mita_masato@cyberagent.co.jp, komachi@tmu.ac.jp

Abstract

Cloze tests play an essential role in language assessment and help language learners improve their skills. In this paper, we propose a novel task called **C**loze **Q**uality **E**stimation (CQE) — a zero-shot task of evaluating whether a cloze test is of sufficient "high-quality" for language assessment based on two important factors: reliability and validity. We have taken the first step by creating a new dataset named CELA for the CQE task, which includes English cloze tests and corresponding evaluations about their quality annotated by native English speakers, which includes 2,597 and 1,730 instances in aspects of reliability and validity, respectively. We have tested baseline evaluation methods on the dataset, showing that our method could contribute to the CQE task, but the task is still challenging. [1]

1 Introduction

A cloze test (Taylor, 1953) is an efficient and comprehensive tool in language assessment, and thus it is widely used in language proficiency tests (**Passage** and **Questions** in Table 1), which measure multiple language abilities of examinees simultaneously, for example, grammatical knowledge (Rye, 1982; Alderson, 1979) and reading comprehension ability (Raymond, 1988; Klein-Braley, 1997). The most widespread format of the cloze test is the multiple-choice word-level cloze test, which consists of an incomplete passage with several blanks and a series of questions, where each question includes several options (four options in the usual setting), requiring examinees to fill the blanks by selecting words (or phrases) from options that make the passage coherent.

The significantly high price of creating cloze tests by experts has prompted automatic cloze generation methods (Goto et al., 2010; Sakaguchi et al., 2013; Hill and Simha, 2016; Panda et al., 2022). However, automatically generated cloze tests do not always contribute well to language assessment and suffer from low quality. In language assessment, a reliable and valid cloze test shows a high ability to measure examinees' language level (Bachman, 1985). Cloze question creation has two steps, word deletion and distractor generation, with the latter greatly affecting the ability to measure language level. As indicated by (Xie et al., 2018), some cloze tests, particularly automatically generated cloze tests, are created coarsely and cause two fatal issues in language assessment: (1) these tests do not guarantee that the answer is not ambiguous, which means there is a risk that multiple options fit almost equally well into the blank; (2) the test creation process considers less about which aspect of the language phenomenon is measured in the test; hence, such tests cannot measure the examinee's language level. These two issues make the cloze test unsuitable for measuring examinees' language level. The first issue makes the test unreliable (e.g., Question 4 in Table 1), which means even if an examinee has enough language knowledge to answer the test, the test might report a wrong score to indicate that the examinee lacks such knowledge. In other words, an unreliable cloze test cannot present the examinee's language level. The second issue makes the test invalid (e.g., Question 3 in Table 1). An invalid test cannot identify the aspect in which an examinee lags in terms of language knowledge, which indicates that educators cannot identify the knowledge that the examinee has not acquired.

In this paper, we tackle the issues of evaluating the appropriateness of cloze tests for language assessment focusing on distractors. By following the test design principle (ALTE, 2011), we define a zero-shot task to evaluate cloze tests for language assessment considering two aspects: reliability and validity, which is called **C**loze **Q**uality **E**stimation

[1] The CELA dataset and code of baselines are available at https://github.com/zz-zhang/cloze-quality-estimation.

Findings of the Association for Computational Linguistics: EACL 2023, pages 540–550
May 2-6, 2023 ©2023 Association for Computational Linguistics

Passage:

A policeman was walking along the street. In the doorway of a shop, a man was standing in the __*1*__ light, with an unlighted cigar in his mouth. The policeman slowed down and then walked up to the man. "I'm just waiting for a friend here," the man said "It's an appointment __*2*__ twenty years ago." The man struck a match and __*3*__ his cigar. The light __*4*__ a pale face with a little white scar near his right eye. "Twenty years ago tonight, when I said goodbye to Jimmy Wells, my best friend to start for the West to make my fortune ...

Questions:				**Evaluation:**
1. A. dark	B. bright	*C. dim*	D. colorful	(*reliable*, *valid_grammar*)
2. A. make	B. makes	C. making	*D. made*	(*reliable*, *valid_reading*)
3. A. is stopped	*B. lighted*	C. burning	D. drop	(*reliable*, *not_valid*)
4. A. formed	*B. illuminated*	C. relieved	*D. showed*	(*not_reliable*, *not_valid*)
...				...

Table 1: Example of cloze test and qualities for each question in CELA. The input consists of a **passage** with multiple blanks and a series of **questions** (tuples of options). The expected output is **evaluation** tuples to indicate whether the questions are reliable and valid (and what language ability is tested if valid). Underlined options are the correct answers, which fit passage perfectly.

(CQE). In CQE, each question in a cloze test is asked to be estimated, whether it is reliable and valid. CQE provides a cloze as input (**Passage** and **Questions** in Table 1) and requires estimation of each question as output (**Evaluation** in Table 1). We introduce a new test set called the **C**loze **E**stimation dataset for **L**anguage **A**ssessment (CELA) which includes a variety of cloze tests and corresponding annotations. These cloze questions are specifically designed for junior high-school students in China. We prepared diverse English cloze tests including expert-designed and rule-generated tests and asked native English speakers to solve them and annotate the quality of each question in the two aspects of reliability and validity.

We also introduce baseline methods for the CQE task: we designed option-aware methods that evaluate cloze questions by analyzing their options. We tested the baseline methods with the CELA and compared them against option-agnostic baselines. We found that detection of unreliable questions is challenging and that all our baseline methods were wary to label a question as unreliable. The framework of our option-aware methods contributed to the validity evaluation, particularly when implemented by DNN-based approaches, which outperformed option-agnostic baselines significantly and showed potential for improvement.

The main contributions of this work are summarized as follows:

- We propose a new task of quality estimation of cloze tests (CQE) for language assessment. We design two sub-tasks: reliability evaluation and validity evaluation.

- We create a new CQE dataset (CELA) for English learners, including annotations for both expert-designed and automatically generated cloze tests.

- We propose the first CQE methods considering the options of cloze questions. We report the experimental results using rule-based and DNN-based approaches.

2 Related Work

Language educators are capable of creating cloze tests rationally; they select words to be blanked and design distractors by their experience in language education to improve reliability and validity. CLOTH (Xie et al., 2018), SCDE (Kong et al., 2020), and CEPOC (Felice et al., 2022) are collections of human-created cloze tests, which are highly evaluated by experts in terms of measuring the English ability of examinees. However, designing cloze tests by experts is costly and difficult to generalize.

Automatic cloze generation methods could decrease the cost of creating cloze tests. To avoid generating useless tests in language assessment, these methods focus on distractor generation, which affects the quality of tests significantly. Previous works have conducted trials designing good rules or machine learning models. Sakaguchi et al. (2013) described a method of generating distractors for assessing an English as second language (ESL) learner's ability to distinguish semantic nuances between vocabulary words. They could generate valid questions, but these questions are domain-limited, which is not easy to generalize to cloze tests for human language assessment. The Children's Book Test (CBT) (Hill et al., 2016) deletes named entities, common nouns, verbs, and prepositions and then designs distractors having the same part of speech (POS) as the deleted words to measure abilities in reading comprehension and vocabulary. Because of the naive distractor creation, the CBT method has the risk of generating unreliable questions. Coniam (1997); Goto et al. (2010); Correia et al. (2012); Hill and Simha (2016); Jiang et al. (2020) explore cloze test generation with various features, such as n-gram frequency and POS tag, and attempt to select deleted words and create distractors by using discriminative models including conditional random fields and support vector machine. Advanced distractor generation methods (Panda et al., 2022) that employ large pre-trained language models (LMs) can provide more valid distractors. These LMs produce better text representation that captures rich semantic information, and better text representation allows generation methods to produce more plausible distractors, which could measure language abilities better. However, designing good rules or models to improve the quality of cloze tests is not that easy. Furthermore, these works claimed they could generate better distractors, but it is difficult to perform comparisons to previous work. All the works performed human evaluations using their own metrics.

Crowdsourcing has been used to evaluate the quality of cloze tests and explore factors that affect quality (Skory and Eskenazi, 2010); workers were required to fill appropriate words in an open cloze-style sentence (a sentence with a blank, but without providing options). Answers from workers were used to calculate Cloze Easiness (Finn, 1977) to indicate whether the sentence is usable for the testing an examinee's vocabulary. The Association of Language Testers in Europe provides a manual for developing a language test (ALTE, 2011), which specifies that the statistics of a test's results reflect its reliability and validity. It asks various examinees to answer a test and analyzes the statistics of a question such as the accuracy and the answer distribution. However, these evaluation methods require human resources or experts, which is time-consuming and sometimes difficult to obtain.

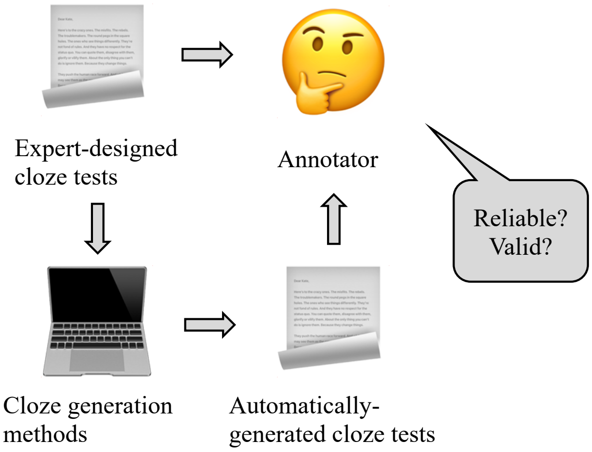

Expert-designed cloze tests
Annotator
Reliable? Valid?
Cloze generation methods
Automatically-generated cloze tests

Figure 1: Flow of creating CELA. Cloze generation methods use passages and blanks from expert-designed tests to eliminate the effect of the word deletion strategy.

3 Cloze Quality Estimation

Motivated by related work, we propose the task of evaluating the quality of the cloze test. We introduce the definition of CQE task and CELA, our new dataset designed specifically for the CQE task. Figure 1 shows the flow of creating CELA. We collected expert-designed and automatically generated cloze tests and asked native English speakers to annotate whether these tests are reliable and valid.

3.1 Task Definition

First, we define a CQE task as follows. In a CQE task, given an incomplete passage with blanks and a series of questions with tuples of options, a quality estimation model should predict the quality of the questions and return tuples to indicate whether these questions are reliable and valid for language assessment. Here, we formalize a CQE task as two classification sub-tasks: *reliability evaluation* and *validity evaluation*.

Reliability. In the reliability evaluation, given a cloze passage with multiple questions (tuples of options), we perform a binary classification of whether the questions are reliable (*reliable*) or not (*not_reliable*). In terms of reliability, if a cloze question has more than one option that fits the context perfectly, there is no guarantee that the question can report a stable test score even when taken by the same examinee. Thus, in the reliability evaluation, we define that if a cloze question has more than one correct answer, it is not reliable.

Validity. In the validity evaluation, we perform a three-class classification of either the question is valid for measuring grammatical knowledge (*valid_grammar*), valid for measuring reading comprehension ability (*valid_reading*), or invalid (*not_valid*). In terms of validity, a valid question should require examinees to use their language ability to distinguish the correct answer option from distractors. If a question measures more than one language ability, the question is considered to be too simple to measure the language ability of the examinee (Rankin, 1976). Thus, in the validity evaluation, we define that if a question requires the examinee's single language ability (grammar or reading comprehension) to distinguish the answer option, the question is valid, otherwise it is invalid.

3.2 CELA data preparation

We collected English cloze tests from Chinese senior-high-school examinations (Xie et al., 2018) called CLOTH, which is expert-designed. To explore whether automatically generated cloze questions are sufficient for language assessment, we also employed four automatically generated cloze tests using previous generation methods: Randomized, Hill, Jiang, and Panda. All generated tests were based on the same cloze passages from expert-designed tests, that is, these five settings share the passages, blanks, and correct answers but have respective distractors in questions.

Randomized is generated using a random sampling method. In this method, we built vocabulary from CLOTH and randomly selected words from the vocabulary as distractor options.

Hill is generated using the same method of the CBT dataset (Hill et al., 2016), which selects words that have the same POS tag with the answer from the vocabulary as distractors.

Jiang employs the method of Jiang et al. (2020), which also selects words from the vocabulary but considers more factors including POS tag, word frequency, and spelling similarity. Their method is designed for the Chinese cloze test, but we adapted it to the English test.

Panda employs the method of Panda et al. (2022), which uses round trip translation to paraphrase a passage and align the paraphrased passages with the original one. They use aligned words to the answer as distractor candidates and select a distractor from candidates considering the synonym and POS tag.

As a result, we collected and generated 150 cloze tests including 3,000 questions [2].

3.3 CELA annotation

We hired Amazon Mechanical Turkers to annotate the 3,000 questions. To ensure annotation quality, we required annotators to have approval rates over 98% and be native English speakers living in the United States. We also added attention checks to avoid bots and irresponsible annotators. Each question was annotated by three different annotators. Table 2 shows examples of our annotation task. As a reward, we paid each annotator $1.5 for a test, which included 20 questions and took 5 to 7 minutes for completion.

We performed inter-annotator analysis on the annotations using Fleiss' kappa score (Fleiss, 1971). Kappa scores were 0.67 and 0.45 for reliability (binary) and validity (3-class), respectively. Moderate kappa scores indicate that the annotation task was well-defined and the annotation result was trustable. Furthermore, to improve the annotation quality, we discarded all disagreed annotations. The majority of annotations that were rejected on the grounds of reliability pertained to long-term reasoning questions. These questions necessitated the integration of information from multiple sentences, and without taking into account this information, the distractors appeared to be equally plausible. This led to a divergence of opinions among some annotators and ultimately resulted in the determination that these questions were unreliable. The reasons for rejection in terms of validity were more varied. One pattern that emerged was the use of prepositions, where some annotators classified questions regarding preposition usage as *valid_reading* instead of *valid_grammar*, despite our explicit instructions on this matter. We posit that this may have been due to the fact that certain questions involving prepositions necessitate contextual information in order to deduce the correct answer (e.g., prepositions of location), causing some annotators to consider them as reading comprehension questions.

The processed data statistics are shown in Table 3. Because most blanks in CLOTH are content words and corresponding questions are designed to measure reading comprehension ability, there are few questions that measure grammatical knowledge.

[2] The cloze tests are collected/generated in five ways, each accounting for one-fifth of the total.

Passage	. . . He wished to find a good job. One day, he went to a company to ____ for a job.
Example 1	
Question	A. apply B. vote C. prepare D. wait
Explanation	In this question, only option A fits the passage perfectly, so please select “One” in Number of answer; options B, C, and D don’t fit the passage logically, and you will eliminate them by the knowledge (ability) of reasoning, so please select “Reading” option in Measured ability.
Example 2	
Question	A. apply B. applied C. look D. has applied
Explanation	In this question, both option A and C fit the passage perfectly, so please select “More than one” in Number of answer; except correct answers (option A and C), options B and D don’t fit the passage grammatically, and you will eliminate them by the knowledge (ability) of grammar, so please select “Grammar” option in Measured ability.
Example 3	
Question	A. apply B. vote C. applying D. waiting
Explanation	In this question, only option A fits the passage perfectly, so please select “One” in Number of answer; option B doesn’t fit the passage logically, option C doesn’t fit the passage grammatically, and you will eliminate them by the both of knowledge (abilities). So please select the “None” option in Measured ability. Also, since option D fits the passage neither logically nor grammatically, you will eliminate it by any of knowledge (abilities). So you can select the “None” option in Measured ability only considering option D.

Table 2: Example of annotation. We used following instructions: “Please select an option in the Number of answers list to indicate whether there is more than one option that fits the passage perfectly; please select what kind of language ability the question measures in the Measured ability list. You can refer to Table 2 for examples.”

Type	#
Reliability questions	2,597
reliable	2,324
not_reliable	273
Validity questions	1,730
valid_grammar	86
valid_reading	921
not_valid	723

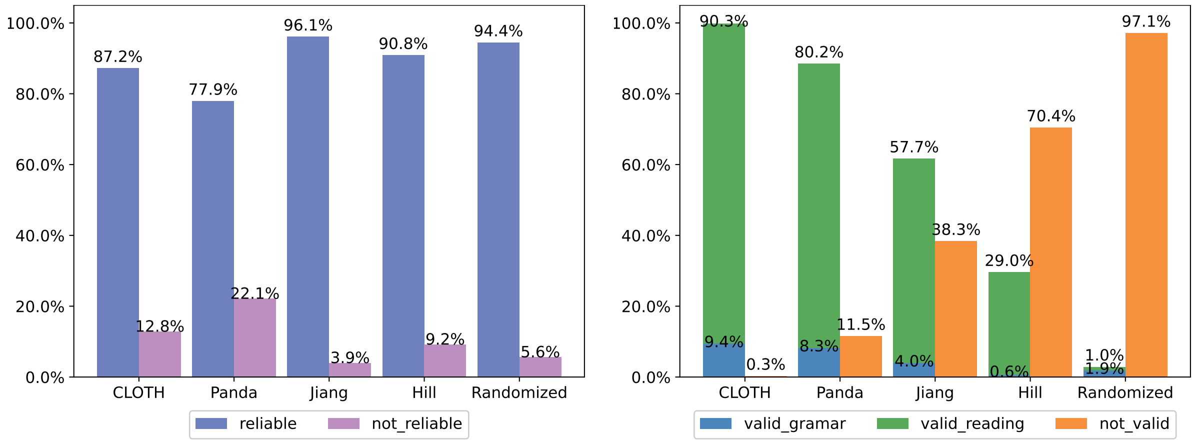
Table 3: Statistics of the processed data. Because reliability is easier to annotate, it has higher agreement, and more annotations are retained than validity.

3.4 CELA analysis

We observed that the five types of cloze tests have various qualities. Figure 2 shows the quality statistic in CELA according to generation methods.

In reliability, Jiang is the most reliable and only includes 3.9% of unreliable questions, and Panda has 22.1%, which is the most unreliable. Surprisingly, CLOTH and Panda, which are expert-designed and generated by an advanced generation method, respectively, are not as reliable as the others. We conjecture that these two types of tests tend to produce more plausible distractors that break only little coherence of the context. Plausible distractors are good at measuring learners’ language ability but have a higher risk of making the question unreliable. In particular, the Panda system utilizes round-trip translation and alignment to generate distractor candidates, which limits the scope of possible candidates and tends to produce more credible options compared to those generated by other systems. Furthermore, the Panda system does not impose strict limitations on eliminating distractors that are also suitable for the blank, which increases the likelihood of generating unreliable questions. On the other hand, the Randomized system selects distractors from the vocabulary without any constraints, which reduces the chance of selecting distractors that are also appropriate for the blank.

In validity, meeting our conjecture, there are fewer invalid questions in CLOTH and Panda, which means these two test types are better at measuring language ability than others. For automatic distractor generation methods, Panda has the strictest



(a) reliability

(b) validity

Figure 2: Quality statistics of cloze tests in CELA. The left and right buckets represent the ratio of high-quality and low-quality questions, respectively.

restrictions on distractor selection and produces the fewest invalid questions. Jiang has more filters for eliminating distractor candidates than Hill and could generate more valid questions. Randomized does not have any restrictions and is difficult to produce valid questions for language assessment.

We provide an example of the CELA dataset in Table 1. In the CELA dataset, each instance includes an incomplete passage with blanks and corresponding sets of options as input (questions). In a question, at least one option can be filled into the corresponding blank to make the passage coherent both grammatically and semantically. The label for each question is a tuple that denotes whether the question is reliable and valid, and if the question is valid the tuple also indicates which aspect of language ability the question measures.

4 Option-aware CQE Method

We propose two methods to tackle the CQE task, which analyze all options of cloze questions as baseline methods for the CQE task.

4.1 Intuition

We designed an option-aware CQE method considering how options in the question affect reliability and validity. We followed the definition in Subsection 3.1 and considered that the reliability and validity of a question is decided by its options. Thus, to tackle two sub-tasks in CQE, we need to inspect each option in terms of (1) whether it can be regarded as the sole answer to the question, and (2) what language ability it measures.

For the former (reliability), we consider that if an option breaks neither grammatical nor semantic coherence of the context, it fits the context perfectly and can be regarded as an answer option. For the latter (validity), if a distractor option only breaks grammatical (or semantic) coherence, examinees will use grammatical knowledge (or reading comprehension ability) to eliminate it, and in these cases, we say the distractor option is a grammatical (or reading) option; if a distractor option breaks both coherence, because it is too simple to measure one's ability, we say it is a purposeless option.

For example, given a context: *I remember sitting in that dark hall listening to Mr. Zigler ____ everyone's spirits up to the ceiling.* and options: *[raise, rise, educate, disappointed]*, the option *raise* does not break neither grammatical nor semantic coherence, so it is an answer option; the option *rise* breaks the grammatical coherence because the blank requires a transitive verb, so it is a grammatical option; the option *educate* obeys the grammatical rule but does not fit context semantically, so it is a reading option; the option *disappointed* is a purposeless option because it breaks both grammatical and semantic coherence of context.

Based on this intuition, we implement two functions, $BreakGrammar(\cdot)$ and $BreakSemantics(\cdot)$, to judge whether an option breaks grammatical or semantic coherence. See Appendix A for a detailed description of the overall framework. To realize these two functions, we designed two different approaches: a rule-based approach and a DNN-based approach.

4.2 Rule-based approach

The rule-based approach is straightforward. It compares options with the answer to a question. The answer to a question can fit the context perfectly and does not break either grammatical or semantic coherence. Thus, we consider that if an option has the same grammatical/semantical feature as the answer, it does not break corresponding coherence either. In this case, functions $BreakGrammar(\cdot)$ and $BreakSemantics(\cdot)$ require one more parameter $answer$.

Given an answer option $answer$ and an option opt, we fill $answer$ and opt into context and obtain POS tags for them. If opt has the same POS tag as $answer$, we consider that it does not break grammatical coherence, otherwise it breaks grammatical coherence. For the implementation, we employed POS tagger in the Stanza library [3].

Similarly, we use a synonym dictionary to judge if the option breaks grammatical coherence. If opt is a synonym of $answer$, opt does not break the semantic coherence, otherwise it breaks semantic coherence.

4.3 DNN-based approach

We also designed a DNN-based approach to implement these two functions. By using pretrained DNN models, we can plug in both grammatical and semantic knowledge into the CQE model. Unlike the rule-based approach, the DNN-based approach does not use $answer$ but opt information for CQE.

We employ an English grammatical error corrector that can detect both grammatical and semantic errors and output the error types. We fill each option into context as input of the corrector and check the output. If the output indicates that there is no grammatical/semantic error, we regard that the option does not break grammatical/semantic coherence; otherwise, we think it breaks such coherence. We need to distinguish grammatical and semantic errors which affect the output of $BreakGrammar(\cdot)$ or $BreakSemantics(\cdot)$. We design such a filter based on error types. To recognize the error type, we use the output tag of an error annotation toolkit.

5 Experiment

We conduct experiments to determine whether option-aware CQE methods (§4) can be a good baseline to estimate the quality of cloze tests, by comparing them with option-agnostic baseline methods (§5.2).

5.1 Configurations

To implement an option-aware baseline with a rule-based approach, we built an English synonym dictionary [4]. Considering that the word inflection or tense do not affect the meaning, we lemmatized both $answer$ and opt into their basic form to judge if they were synonyms. The word lemmatization was implemented by employing the NLTK library [5] and using the lemmatizer based on WordNet (Miller, 1998). We also employed POS tagger in the Stanza library to assign POS tags to $answer$ and opt.

As for a DNN-based approach, we employed GECToR (Omelianchuk et al., 2020), a grammatical error corrector, that provided trained parameters and achieved a considerable performance on both CoNLL-2014 and BEA-2019 shared task (Ng et al., 2014; Bryant et al., 2019). We used GECToR which was implemented by RoBERTa (Liu et al., 2019). We fed original and corrected sentences into the ERRor ANnotation Toolkit (ERRANT) [6] to obtain ERRANT tags. If the detected error's ERRANT tag is one of **ADJ**, **ADV**, **NOUN**, and **VERB**, we considered the error to be a semantic one and not a grammatical one. Furthermore, we observed that the tag **OTHER** might contain both grammatical and semantic errors; therefore, we set two configurations for errors with tag **OTHER** as either grammatical or semantic errors.

5.2 Option-agnostic baselines

We employed the following *random* baseline and *majority prediction* baseline to show how well option-agnostic methods could perform on the CELA. Option-agnostic baselines can also be regarded as weak baselines.

Random baseline The random baseline predicts random class in reliability and validity classification. We chose the output class from the uniform distribution.

Majority prediction baseline The majority prediction baseline predicts the majority class in each classification sub-task. According to our CELA dataset, it always predicts *reliable* and

[3] https://github.com/stanfordnlp/stanza

[4] collected from https://www.thesaurus.com/

[5] https://www.nltk.org/

[6] https://github.com/chrisjbryant/errant

Methods	Reliability			Validity				
	F_1	prec.	recall	micro F_1	macro F_1	$r.F_1$	$g.F_1$	$n.F_1$
Option-agnostic (weak)								
- Random	17.45	10.56	49.82	32.31	27.90	39.87	8.37	35.45
- Majority prediction	0.00	0.00	0.00	53.24	23.16	69.48	0.00	0.00
Option-aware (strong)								
- rule-based	2.87	1.47	**66.67**	42.35	41.31	30.28	37.05	**56.61**
- DNN-based ($\bar{O}$)	**19.50**	**98.53**	10.82	54.79	43.11	**72.33**	47.02	9.96
- DNN-based (O)	19.31	97.80	10.71	**58.54**	**48.25**	71.90	**53.79**	19.05

Table 4: Performance of CQE baseline methods on the CELA dataset. $r.F_1$, $g.F_1$, and $n.F_1$ represent binary F_1 score for *valid_reading*, *valid_grammar*, and *not_valid* questions, respectively. **Bold** and underline indicate the best and second-best result, respectively. $\bar{O}$ and O indicate we regard errors from GECToR with tag **OTHER** as grammatical and semantic errors, respectively.

valid_reading in the sub-task of reliability and validity classification, respectively.

5.3 Meta-evaluation metrics

To demonstrate the efficiency of CQE methods in estimating the quality of cloze tests, we provide baseline meta-evaluation metrics for the CQE task. Specifically, in the reliability evaluation, we used F_1, precision, and recall score. Because unreliable cloze tests are harmful to language assessment, we must focus on how well CQE models can recognize unreliable tests; thus we set *not_reliable* as the positive label. For the validity evaluation, we used the micro-averaged and macro-averaged F_1 score. To indicate how well models perform in each class, we also split the overall F_1 score into three parts: F_1 for *valid_reading*, *valid_grammar*, and *not_valid*.

5.4 Result

The performance of the baselines on the CELA dataset is presented in Table 4. For option-agnostic baselines, because of imbalanced data distribution, the majority prediction baseline was not able to detect the *not_reliable* questions. Both baselines of random and majority prediction did not perform well on reliability compared with validity. Moreover, unreliable question detection is important to language assessment. In future work, improving the performance on reliability should be considered preferentially.

The option-aware baseline implemented by the rule-based approach performed worse than random baselines on some metrics. Although it achieved a moderate recall value, the precision was nearly zero, which denotes it tends to assign *reliable* to all questions. On the validity performance, it outperformed option-agnostic baselines on some metrics, but it is still insufficient for evaluating the quality of cloze tests. One reason is that rules using the POS tag and synonym list are so naïve that they only consider partial cases of the option type. For example, given a context *This music made everyone want to ____. It was an early form of jazz.* and options *[dance, sing, laugh, ...]*, though options *sing* and *laugh* are not the synonyms of the answer *dance*, they also fit the context semantically and should have not been classified into the reading option.

In most cases, the option-aware method using the DNN-based approach outperformed option-agnostic baselines. The DNN models utilized in this paper were straightforward and rudimentary, and there is potential for further improvement to make them more suitable for widespread use. Regarding reliability, errors with **OTHER** as grammatical or semantic errors have little effect on the performance. In terms of validity, when we regard **OTHER** errors as semantic errors, the micro F_1 value increased because the model could predict more *not_valid* questions correctly, which accounted for a significant proportion in CELA.

Except for hyperparameters, the mis-prediction caused by the underlying DNN models also leads to errors. GECToR did not perform well on long-term reasoning; thus, it was not able to detect some semantical errors. For example, given a context *I was ____ of flying, ... In order to get rid of my fear I decided to try a helicopter ride*, when filling word *proud* into the blank, we expect GECToR to correct the sentence with some words similar to *afraid*, but GECToR did not report any error.

6 Conclusion and Future Work

We proposed a novel task for evaluating cloze questions for human language assessment (CQE), which involved two important factors that affect the quality of cloze questions, reliability and validity, and also provided the CELA dataset. In addition, we explored automated CQE methods that can estimate the quality of cloze tests, by designing option-aware methods. Our experimental results on the CELA dataset showed that imbalanced data bring challenges.

In future work, we would like to investigate more factors that affect the quality of cloze questions and expand the CELA dataset. For example, given the context *The sun also* ____. and two different sets of options *[raises, rises, lifts, elevates]* as well as *[raises, lives, runs, lefts]*, although these two sets of options both measure reading comprehension ability, answering the question with the former set of options is more difficult than the latter and requires a higher language level. Improving the performance of DNN models and optimizing implementation of $BreakGrammar(\cdot)$ and $BreakSemantics(\cdot)$ can improve the evaluation performance. For example, designing more detailed filters in the grammatical error corrector to filter out potential semantic errors or using fine-designed data to train different language models for grammatical and semantic error detection separately may boost the performance of DNN models.

We hope that the task and our resource will encourage further exploration from both computational linguistics and language education.

Limitations

The first limitation of this study is the coverage. In this study, the CQE task is defined to evaluate cloze tests that are generated by distractor generation methods. Cloze tests in this study are all based on expert-designed blanks. However, word deletion methods, which decide which word to be blanked, affect the quality of cloze tests, too. Investigation of how blanks influence the quality of cloze tests is necessary.

The second limitation of this study is the scalability of the annotation. In this study, the annotation of question quality is done by experts, which makes creating a large-scale dataset not that easy. This limitation could be mitigated by alternative choice of target data, i.e., there is room to replace native speakers with non-native speakers by selecting target data that do not require English knowledge at high-level proficiency (e.g., CEFR-A).

Finally, the CQE task and corresponding corpus are designed specifically for the English language. However, we are interested in exploring the possibility of adapting the task and dataset to other languages. The principles of test design, such as reliability and validity, apply to other languages as well, but the specific details may vary based on the language. For example, questions for a hieroglyph-based language may require learners to identify glyphs, which must be taken into consideration when defining reliability and validity. Adapting the CQE task to a new target language and creating a corresponding dataset requires a publicly available cloze question dataset or effective cloze question generation techniques in the target language, as well as experts in the language to evaluate question quality. In the future, we hope to develop an automatic adaptation method to transfer our task and dataset to multiple languages.

Acknowledgements

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

References

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

ALTE. 2011. *Manual for Language Test Development and Examining: For Use with the CEFR*. Language Policy division, Council of Europe.

Lyle F Bachman. 1985. Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, 19(3):535–556.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

David Coniam. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *Calico Journal*, pages 15–33.

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

Mariano Felice, Shiva Taslimipoor, Øistein E. Andersen, and Paula Buttery. 2022. CEPOC: The Cambridge exams publishing open cloze dataset. In *Proceedings of the 2022 International Conference on Language Resources and Evaluation*. European Language Resources Association.

Patrick J. Finn. 1977. Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, 13(4):508–537.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal*, 2(3):210–224.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations (ICLR)*.

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.

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

Christine Klein-Braley. 1997. C-tests in the context of reduced redundancy testing: An appraisal. *Language testing*, 14(1):47–84.

Xiang Kong, Varun Gangal, and Eduard Hovy. 2020. SCDE: Sentence cloze dataset with high quality distractors from examinations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5668–5683, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Earl Rankin. 1976. Sequence strategies for teaching reading comprehension with the cloze procedure. In *Reading: Theory, Research and Practice, 26th Yearbook of the National Reading Conference*, pages 92–98, Atlanta, GA. National Reading Conference.

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

James Rye. 1982. *Cloze procedure and the teaching of reading*. London.

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 (Volume 2: Short Papers)*, pages 238–242, Sofia, Bulgaria. Association for Computational Linguistics.

Adam Skory and Maxine Eskenazi. 2010. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–56, Los Angeles, California. Association for Computational Linguistics.

Wilson L Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

A Algorithm for option-aware models

Algorithm 1 shows the framework of option-aware methods considering how to classify questions in aspects of reliability and validity by using types of options.

Please note that Algorithm 1 takes only content word as options. If the option is a functional word, we only assign *answer* or *grammar* as its type because questions including functional words as options only measure grammatical knowledge.

Algorithm 1: Framework of option-aware baseline

```
Input: context $c$; a set of options in a question $O = \{opt_1, ..., opt_n\}$;
function to judge if option breaks grammatical coherence
$BreakGrammar(\cdot) \in \{true, false\}$;
function to judge if option breaks semantic coherence
$BreakSemantics(\cdot) \in \{true, false\}$
Output: reliability and validity tuple of the input question $(r, v)$, where
$r \in \{reliable, not\_reliable\}$,
$v \in \{valid\_grammar, valid\_reading, not\_valid\}$
// Assign type to each option
$types$ = [] ;
for $i \leftarrow 1$ to $n$ do
    if $BreakGrammar(c, opt_i) \wedge BreakSemantics(c, opt_i)$ then $types[i] \leftarrow purposeless$;
    if $\neg BreakGrammar(c, opt_i) \wedge BreakSemantics(c, opt_i)$ then $types[i] \leftarrow reading$;
    if $BreakGrammar(c, opt_i \wedge \neg BreakSemantics(c, opt_i)$ then $types[i] \leftarrow grammar$;
    if $\neg BreakGrammar(c, opt_i) \wedge \neg BreakSemantics(c, opt_i)$ then $types[i] \leftarrow answer$;
end
// Classify question in terms of reliability and validity by using option types
if $types.count(answer) = 1$ then
    $r \leftarrow reliable$ ;
    if $types.count(grammar) = n - 1$ then $v \leftarrow valid\_grammar$;
    else if $types.count(reading) = n - 1$ then $v \leftarrow valid\_reading$;
    else $v \leftarrow not\_valid$;
else
    $r \leftarrow not\_reliable$ ;
    $v \leftarrow not\_valid$ ;
end
return $(r, v)$;
```