Empowering Language Assessment and Education with Natural Language Processing: A Focus on Cloze Tests

TOKYO METROPOLITAN UNIVERSITY

Zizheng Zhang

Supervisor: Prof. Yasufumi Takama

Department of Computer Science
Graduate School of Systems Design
Tokyo Metropolitan University

This dissertation is submitted for the degree of
Doctor of Philosophy (Computer Science)

March 2024

Acknowledgements

I am profoundly grateful to my supervisor, Professor Mamoru Komachi, whose invaluable guidance was instrumental throughout my research and Ph.D. journey. Over four years, Professor Komachi meticulously nurtured my academic acumen and communication skills among various other competencies. Whenever I encountered research impediments, his insightful advice not only extricated me from these challenges but also fortified my mental resilience. His generosity in providing a robust platform allowed me the freedom to explore my research interests deeply. By initiating study groups, encouraging participation in academic events, and actively organizing such gatherings, he exemplified the epitome of a distinguished researcher and educator, an ideal I aspire to in my own professional path.

My sincere appreciation extends to Professor Yasufumi Takama, Professor Naoaki Okazaki, Professor Eri Shimokawara, and Professor Shohei Yokoyama. Their incisive questions and comments invariably sparked new perspectives and enhanced the efficiency of my research following each presentation I conducted.

I must express my gratitude to the Assistant Professors in our laboratory, Dr. Masato Mita and Dr. Teru Oka. Dr. Mita diligently reviewed my research progress weekly, offering invaluable feedback and references that safeguarded my time and efficiency. His expertise, particularly in refining my manuscripts, was pivotal in the smooth publication of my papers. Dr. Oka was instrumental in shaping my fellowship application, guiding me on presenting my research compellingly and accessibly, which significantly contributed to my representation before the fellowship committee.

I am equally thankful for the camaraderie and intellectual stimulation provided by my peers in the NLP laboratory at Tokyo Metropolitan University. Interactions with Zhousi Chen, Longtu Zhang, Masahiro Kaneko, Yuting Zhao, Hongfei Wang, Tosho Hirasawa, Siti Oryza Khairunnisa, Hwichan Kim, Taichi Aida, Aomi Koyama, Xiaomeng Pan, Naoya Ueda, Zhidong Ling, Yoshinari Nagai, Masamune Kobayashi, Wenjie Duan, and many others were sources of continuous inspiration. Our collaborative spirit in sharing innovative ideas and pivotal academic papers significantly enriched my research experience.

My deepest gratitude is reserved for my family and closest friends. While they may not have been directly involved in my academic pursuits, their unwavering support provided a nurturing environment that was my sanctuary during challenging times. Their strength and kindness illuminated my darker days, and it is my fervent hope to embody these qualities in my own life, becoming a beacon for others as they have been for me.

Lastly, I wish to extend my thanks to you, the reader, for showing interest in my work. Your engagement is the final, integral piece in this academic journey.

Abstract

Cloze questions are an integral tool in language assessment, where students, especially those learning English as a Second Language (ESL), encounter passages with omitted words or phrases and must select or fill in the most appropriate words to complete the text. Recognized for their efficiency in capturing a holistic understanding of a student's language proficiency, these questions are prominently featured in esteemed proficiency tests such as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL). Their ubiquitous presence underscores their significance in the realm of language education. However, beneath this prevalent use lies a multitude of complexities and challenges that drive continuous research and innovation in the domain.

The creation and evaluation of these cloze questions are far from trivial. While experts painstakingly design these questions to measure various language abilities, from grammatical knowledge to reading comprehension, their manual creation is both time-intensive and costly. This resource-intensive nature has led to the exploration of automated cloze generation methods. These automatic methods, while scalable, often fall short in replicating the quality and precision of expertly curated questions. Specifically, issues related to the reliability and validity of automatically generated questions have emerged as significant challenges. An unreliable cloze question might have multiple answers that fit the context almost equally well, making it a poor measure of a student's actual language proficiency. On the other hand, the validity of the question ensures it assesses the specific aspect of the language it was intended for, without any ambiguities.

Recognizing these challenges, we embarked on a mission to enhance the quality of cloze questions through the Cloze Quality Estimation (CQE) task. This innovative task is dedicated to evaluating the appropriateness of cloze tests for language assessment, with a special focus on evaluating distractors in the questions. Rooted in the time-honored

principles of test design, the CQE task is anchored around two pillars: reliability and validity. These foundational concepts ensure that the cloze questions are both accurate in their assessment and pertinent in measuring specific language phenomena. In the study CQE task, substantial advancements were achieved in evaluating the quality of cloze tests. We proposed the task considering both reliability and validity, and built a high-quality test set for the task. Through comprehensive experiments, the proposed option-aware baseline models demonstrated marked superiority over option-agnostic baselines. Specifically, the DNN-based option-aware approaches reached F1 scores of 19.5 and 58.5 on detecting whether a question is reliable and valid, respectively. Although DNN-based approaches outperformed rule-based option-aware baseline and option-agnostic baselines, it still shows the CQE task is a challenge.

However, our exploration did not stop at just the evaluation of cloze questions. Recognizing the profound impact of clear explanations on language learning, especially in self-paced learning contexts, we ventured into the realm of generating explanations tailored for cloze questions. These explanations, when crafted with clarity and precision, can serve as powerful learning tools, demystifying the intricacies of the English language and shedding light on the rationale behind correct and incorrect choices. It's an endeavor to empower learners, equipping them with insights that foster deeper comprehension and robust long-term knowledge retention. To this end, we introduced the `ClozEx` task. A groundbreaking initiative, the `ClozEx` task aims at generating explanations that are both fluent in their expression and valid in their content for English cloze questions. These explanations, we envisioned, should seamlessly blend readability with rich, contextual information, offering students a comprehensive understanding of the underlying language principles. Supporting this task, we curated an extensive dataset, boasting over 140k expert-assured pairs of cloze questions and their respective explanations. Each entry in this dataset stands as a testament to meticulous design and rigorous quality checks, ensuring their relevance and utility for the `ClozEx` task. Our research journey led us through a plethora of models, spanning from intricate encoder-decoder architectures to the behemoths of computational linguistics – Large Language Models (LLMs). Our findings were revelatory. The encoder-decoder models, particularly BART-large, emerged as the top performers. This model achieved a manual validity score of 4.43 out of 5 and a BLEU score of 27.33, highlighting its efficacy in generating fluent and valid English cloze explanations. In the realm of LLMs, the GPT3.5-turbo exhibited promising results in a zero-shot scenario, attaining the highest fluency score of 4.53. However, the performance of LLMs in the aspect of validity was not very satisfactory. It demonstrated the essentialness of training data in the task. Further, by analyzing the consistency between manual and automatic metrics, we found that reference-based metrics like BLEU could be used to measure supervised models trained on our dataset. This extensive analysis not only showcased the strengths and limitations of various models but also set a benchmark for future endeavors in the `ClozEx` domain.

In conclusion, this research offers a deep dive into the multifaceted world of cloze questions. By intertwining the realms of computational linguistics and language education, we endeavor to redefine language learning, making it a more insightful, tailored, and enriching experience for learners worldwide. Through a blend of innovative tasks, expansive datasets, and rigorous model evaluations, this work stands as a beacon, illuminating the path for future research and pedagogical advancements in the domain.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Language learning and assessment are deeply interconnected processes, with tools and methods evolving over time to measure and facilitate language proficiency. At the forefront of these tools, cloze tests have carved out a special niche [1]. Recognized for their efficiency and comprehensiveness, cloze tests are extensively employed in language proficiency evaluations, shedding light on diverse facets of language skills, from grammatical nuances to reading comprehension capabilities [2–5].

However, the rising reliance on such tests has underscored the pressing need for their meticulous crafting, validation, and the addition of supportive feedback mechanisms. High-quality cloze tests, especially those crafted by experts, can be resource-intensive and costly to produce. On the other hand, while automatically generated cloze tests offer a cost-effective alternative, they often suffer from quality inconsistencies [6].

Equally important is the provision of explanations for cloze questions. Clear explanations not only aid in clarifying the correct answer but also amplify the learning process, enabling learners to grasp underlying language principles and rectify misconceptions [7].

To address challenges inherent in cloze tests and their explanations, Natural Language Processing (NLP) presents a promising avenue. With its powerful computational capabilities and intricate language modeling techniques, NLP offers convenient technological solutions to refine, evaluate, and enhance cloze tests.

In response to the discernible absence of tasks and methodologies for estimating the quality of automatically generated cloze questions, our initial endeavor led to the conception of a novel task dedicated to the quality estimation of cloze tests, termed as CQE. Grounded in the test design principles delineated by the Association of Language Testers in Europe [8], we postulate that a cloze test of superior quality must inherently be both reliable and valid. Thus, the CQE task mandates a method to gauge the twin attributes of reliability and validity for a proffered cloze test. To ascertain the efficacy of the proposed CQE methodologies in truly estimating cloze test quality, we meticulously curated a benchmark dataset, christened as the Cloze Estimation dataset for Language Assessment (`CELA`). `CELA` is a diverse compendium, amalgamating expert-crafted cloze tests with those autonomously generated. These tests undergo rigorous quality appraisal by seasoned experts to ensure their veracity. Furthering our exploration, we proffer baseline methodologies tailored for the CQE task. We championed the design of option-conscious methods, emphasizing the nuanced evaluation of cloze questions by delving into the intricacies of their provided options. Subjecting these baselines to rigorous scrutiny using the `CELA` dataset and juxtaposing them against their option-indifferent counterparts revealed intriguing insights. It became palpable that the identification of unreliable questions posed intricate challenges, with the baseline methodologies exhibiting a discernible reticence in labeling a question as dubious in reliability. Notably, the architecture of our option-conscious methodologies bore significant fruit in appraising test validity. This was particularly pronounced in the Deep Neural Network (DNN)-oriented approaches, which not only eclipsed the performance of option-indifferent baselines but also signaled ample avenues for further enhancement.

To address the burgeoning need for automatically generating high-quality explanations accompanying cloze tests, we embarked on introducing an innovative task dedicated to crafting cogent explanations for given cloze tests. At its core, an adept explanation aimed at aiding the solution of a cloze query should seamlessly combine readability with a reservoir of pertinent background knowledge. This naturally predicates that any explanation generation should be suffused with fluency and rich informativeness.

In tandem with this, we proffered a voluminous dataset, housing in excess of 140k instances, which is ensured by experts, juxtaposing cloze questions with their congruent explanations. The genesis of this dataset involved a meticulous expansion of expertly conceived cloze questions and their associated explanations. Our methodology was marked by its ingenuity: we crafted a technique to distill patterns from a cloze question paired with its explanation. This distilled pattern subsequently served as the bedrock for autonomously generating new question-explanation dyads.

Delving deeper, our exploration sought to decode the intricate dynamics influencing the `ClozEx` task. To this end, we subjected a plethora of models, spanning encoder-decoder to decoder-exclusive architectures, to rigorous training regimens, enshrining them as our baselines. Our foray also extended to gauging the prowess of large language models (LLMs) within a zero-shot prediction paradigm. Herein, we harnessed the LLMs to conjure explanations for proffered cloze questions, eschewing any fine-tuning. The consequent evaluation of these baseline models unveiled that both encoder-decoder and decoder-only paradigms, post fine-tuning, exhibited commendable acumen in generating acceptable explanations. Concurrently, while LLMs showcased a predilection for crafting fluent explanations, they oftentimes fell short in imbuing them with ample informational content requisite for resolving the queries. Empowering LLMs solely with questions, even when coupled with rudimentary prompts, proved inadequate in consistently yielding top-tier explanations.

In this dissertation, our contributions are 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.
- We propose a new task toward generation of fluent and valid English cloze explanation (`ClozEx`) for ESL learning.
- We create a large-scale and expert-quality-assured dataset for `ClozEx` task, including more than 140k instances generated by a pattern-based method.
- We investigate model performance trained on our dataset. We also explore the ability of LLMs of generating appropriate explanations in zero-shot scenario.

- We examine the correlation between automatic evaluation metrics and manual evaluation in the context of the `ClozEx` task, providing insights into the reliability of these metrics for assessing the quality of generated explanations.

For clarity, this dissertation is structured as follows:

Chapter 2 delves into the background and related works, tracing the journey of cloze tests, underscoring their importance, and identifying existing gaps. Chapter 3 unravels the details of the Cloze Quality Estimation task, its inherent challenges, and our methodologies to surmount them. Chapter 4 pivots to the task of generating explanations for cloze questions, elucidating its importance and our innovative approach. Ensuing chapters encapsulate our experimental design, outcomes, discussions, and final thoughts, succinctly presenting our contributions and their broader implications in the world of language learning and assessment. Through this dissertation, we endeavor to harness the synergy of traditional language assessment techniques and advanced NLP tools, ensuring cloze tests remain an indispensable asset in the landscape of language proficiency evaluation in our increasingly digital age.

Chapter 2

Background and Related Work

2.1 Introduction to English Cloze Questions

The concept of cloze questions was pioneered in the early 1950s by Wilson Taylor [1]. Since its inception, this form of assessment has become an indispensable component in language assessment [9]. A typical cloze test presents learners with a passage with select words or phrases omitted. The challenge for the learner is to deduce the missing elements, either by choosing from given options or by producing the appropriate word or phrase, testing their grasp on syntax, semantics, and context [10, 11].

Educational establishments globally recognized the dual efficacy of cloze tests. Firstly, they serve as a robust evaluative tool, gauging a student's command over grammatical structures [2, 3] and their reading comprehension abilities [4, 5]. Secondly, as a pedagogical instrument, they aid in reinforcing language concepts and enhancing comprehension. Further testament to their significance is their inclusion in esteemed English proficiency examinations, such as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL), which employ cloze questions to measure the holistic language abilities of English as a Second Language (ESL) learners.

English cloze questions can be broadly categorized into two types based on the provision of options: open and closed.

Open Cloze Questions. In open cloze questions [12], test-takers are tasked with filling in the blanks without any provided options, relying solely on their knowledge and understanding of the language (as depicted in Table 2.1a). Such questions demand a higher level of linguistic proficiency, making them an effective tool for gauging deeper language comprehension and mastery [13–15]. However, they present a challenge when it comes to evaluation. Due to the potential for multiple correct answers, manual intervention, often by language experts or teachers, becomes necessary to judge the appropriateness of the response. This characteristic can be a limitation, especially in an era increasingly leaning towards automated scoring for language assessments.

Closed Cloze Questions (Multiple Choice Cloze Questions - MCCQs). Contrasting open cloze questions, MCCQs [16] present test-takers with options for each blank, making them choose the most suitable answer (refer to Table 2.1b). Their structured nature, with a definitive correct answer, makes them amenable to machine scoring. Yet, this format demands meticulous effort during the question design phase, particularly in crafting distractor options that are plausible yet incorrect.

The length of the context in cloze questions also plays a pivotal role, determining which language abilities are being assessed.

Short-term Cloze Questions. Often limited to a single sentence with a blank (as seen in Table 2.1), short-term cloze questions predominantly test a learner's grammatical knowledge and vocabulary skills. Their concise format ensures that the focus remains on specific linguistic constructs or words.

Long-term Cloze Questions. Offering a more extended context, long-term cloze questions span several paragraphs with multiple blanks interspersed (illustrated in Table 2.2, cited from [6]). They delve deeper, targeting vocabulary breadth and reading comprehension skills. Such questions often require the test-taker to engage in long-term reasoning, piecing together information from various parts of the text to arrive at the correct answers.

2.2 Cloze Questions Corpus

Language educators have traditionally crafted cloze tests with a methodological approach, drawing from their pedagogical experience to enhance reliability and validity. Such human-

Question:
She put the flowers in the _____ so they would get some sunlight.
Answer(s):
atrium, greenhouse, garden, sunroom, ...

(a) Example of open English cloze question.

Question:	
She put the flowers in the _____ so they would get some sunlight.	
Options:	**Answer:**
(A) garden (B) fridge (C) drawer (D) closet	(A) garden

(b) Example of closed English cloze question (MCCQ).

Table 2.1 Examples of short-term cloze questions.

created cloze tests, including collections like CLOTH [6], SCDE [17], and CEPOC [18], are highly regarded by experts for their efficacy in measuring English language proficiency. However, manual design by experts is not only expensive but also challenging to scale.

This has prompted a significant shift towards automated cloze generation methods, aiming to mitigate the high costs associated with expert involvement. Early attempts at automatic cloze question generation leaned on rudimentary strategies, such as fixed ratio word deletion and random distractor selection [2, 19]. Contemporary research, though, emphasizes the validity of the generated questions more than ever. A discriminative approach [20], for instance, generated distractor options based on an English learner writing correction corpus. This nuanced approach meant that words commonly misused in specific contexts became distractor options, making the cloze tests more proficient at discerning learner language proficiency.

Several recent works have investigated various features, such as part of speech (POS), n-gram frequency, and word sense, to enhance the validity of cloze questions [21–25]. The focus on distractor generation, which considerably influences test quality, led to the exploration of discriminative models, including conditional random fields and support vector machines, as well as large pre-trained language models (LMs). These LMs, due to their adeptness at capturing rich semantic nuances, facilitate the production of more plausible distractors, thus improving language ability measurements.

Passage:
Nancy had just got a job as a secretary in a company. Monday was the first day she went to work, so she was very __1__ and arrived early. She __2__ the door open and found nobody there. “I am the __3__ to arrive.” She thought and came to her desk. She was surprised to find a bunch of __4__ on it. They were fresh. She __5__ them and they were sweet. She looked around for a __6__ to put them in. “Somebody has sent me flowers the very first day!” she thought __7__ . “But who could it be?” she began to __8__ . The day passed quickly and Nancy did everything with __9__ interest. For the following days of the __10__ , the first thing Nancy did was to change water for the followers and then set about her work.
Then came another Monday. __11__ she came near her desk she was overjoyed to see a(n) __12__ bunch of flowers there. She quickly put them in the vase, __13__ the old ones. The same thing happened again the next Monday. Nancy began to think of ways to find out the __14__ . On Tuesday afternoon, she was sent to hand in a plan to the __15__ . She waited for his directives at his secretary’s __16__ . She happened to see on the desk a half-opened notebook, which __17__ : “In order to keep the secretaries in high spirits, the company has decided that every Monday morning a bunch of fresh flowers should be put on each secretarys desk.” Later, she was told that their general manager was a business management psychologist.

Questions:

1.	A. depressed	B. encouraged	**C. excited**	D. surprised
2.	A. turned	**B. pushed**	C. knocked	D. forced
3.	A. last	B. second	C. third	**D. first**
4.	A. keys	B. grapes	**C. flowers**	D. bananas
5.	**A. smelled**	B. ate	C. took	D. held
6.	**A. vase**	B. room	C. glass	D. bottle
7.	A. angrily	B. quietly	C. strangely	**D. happily**
8.	A. seek	**B. wonder**	C. work	D. ask
9.	A. low	B. little	**C. great**	D. general
10.	A. month	B. period	C. year	**D. week**
11.	A. Unless	**B. When**	C. Since	D. Before
12.	A. old	B. red	C. blue	**D. new**
13.	A. covering	B. demanding	**C. replacing**	D. forbidding
14.	**A. sender**	B. receiver	C. secretary	D. waiter
15.	A. assistant	B. colleague	C. employee	**D. manager**
16.	A. notebook	**B. desk**	C. office	D. house
17.	**A. said**	B. written	C. printed	D. signed

Table 2.2 Examples of long-term MCCQs. Correct answers are highlighted in bold.

Despite the advancements, generating high-quality distractors remains an intricate challenge. While recent research has proposed improved distractor generation methods, empirical comparisons against preceding works are hindered by the lack of uniform evaluation metrics.

Moreover, while the bulk of prior research in question generation sought to devise plausible questions from given texts, a notable gap was the overlooked importance of generating accompanying explanations. This is paramount for offering comprehensive assistance to language learners. Shifting the spotlight from solely generating questions to also crafting explanations paves the way for this research, marking a pivotal step in the evolution of language learning technologies.

2.3 Quality Estimation for Cloze Questions

The evaluation of cloze tests, traditionally, has relied on human judgment to assess the quality and the suitability of the questions. Crowdsourcing has emerged as a potent tool in this realm, offering a platform to engage a diverse set of individuals for quality assessment [26]. In typical setups, participants, or "workers", are tasked with filling blanks in open cloze-style sentences without any provided options. Their collective responses serve as a rich dataset, allowing for the computation of metrics such as Cloze Easiness [27]. Such metrics aim to gauge the suitability of a sentence in testing a learner's vocabulary depth and breadth.

Moreover, esteemed bodies like the Association of Language Testers in Europe have laid down guidelines for the development and evaluation of language tests [8]. These guidelines stress the importance of reliability and validity in test design, necessitating empirical evaluations where diverse examinees attempt the test. By studying metrics like accuracy and answer distribution, a deeper understanding of the test's effectiveness is achieved.

Nevertheless, while manual quality estimations offer depth and nuance, they come with their own set of challenges. Engaging human evaluators, especially experts in the field, is resource-intensive. The process is often time-consuming, and sourcing experts, particularly for niche or advanced topics, can pose significant hurdles. This motivates a demand for automated or semi-automated approaches that could streamline the evaluation while preserving the rigor and depth of manual evaluations.

For automatic evaluation of cloze tasks, researchers explored the applicability of information theory to predict task easiness, based on the context provided by surrounding words [28]. The theoretical foundation [29] emphasized the role of lexical transfer features in determining the predictability of a cloze task. The authors utilized a multi-stage filtering approach for cloze sentence generation, evaluating quality using metrics such as context restriction and Cloze Easiness. Crowdsourced evaluations from Amazon Mechanical Turk were juxtaposed with expert evaluations to gauge task quality. Notably, a significant correlation was observed between co-occurrence scores (how often words appear together) and Cloze Easiness, implying its utility as a cloze quality predictor. However, the anticipated correlation between reading levels of words and context restriction wasn't statistically substantiated.

This research is pivotal in the intersection of computational linguistics and educational assessment. By leveraging both traditional readability measures like Flesch-Kincaid and newer statistical models, the paper seeks to refine and improve the automated generation of cloze tasks. However, the word difficulty is only a limited part of the question quality, and only serve on vocabulary questions. Universal standard for automatic cloze quality estimation is still a challenge.

2.4 Benefits of Answer Explanations in Language Learning

Language, by its very nature, is riddled with nuances and subtleties that can often elude even the keenest of learners. As learners acquire a knowledge of a new language, they are frequently confronted with challenging cloze questions. In such scenarios, having access to clear and concise explanations can be helpful in language learning. These explanations elucidate the rationale behind both correct and incorrect choices, illuminating the subtleties and intricacies of the language, ensuring learners do not stray into the realm of misconceptions [30].

Explanations do not merely provide clarity; they delve deeper, offering a comprehensive understanding. Grasping the underlying rules and concepts helps learners understand not just the "what" but the "why" of a particular language structure. This deeper insight, in turn, paves the way for enhanced retention and application.

On a cognitive front, the benefits of explanations gain even more prominence. According to schema theory, our brains systematically organize knowledge into schemas – mental constructs pivotal for interpreting and assimilating new information [31]. As learners assimilate new linguistic elements, they leverage their existing schemas to correlate with new vocabulary or grammatical paradigms. Explanations serve a dual purpose in this scenario. They not only align new linguistic data into pre-established mental schemas but also fortify these schemas, ensuring a robust understanding.

Ultimately, the provision of high-quality explanations, underpinned by solid theoretical frameworks, can empower learners, nurturing profound comprehension and fostering long-term knowledge retention [7].

2.5 Explanation Generation Methods for Language Learning

The realm of language learning has seen various attempts at leveraging computational methods to facilitate feedback generation. Among these endeavors, the feedback comment generation (FCG) task [32] stands out as a notable advancement. This task aims at automating the generation of feedback comments such as hints or explanatory notes tailored for non-native learners of English embarking on writing exercises.

While the FCG task brings a unique perspective to the table and aids in grammar learning through real-time writing correction, it presents certain limitations. A central concern is the task's inherent structure; it is rooted in a bottom-up approach that derives grammatical knowledge from free English composition. This approach, while beneficial in some contexts, may not provide exhaustive coverage of all the grammar facets essential for learners. It is akin to learning grammar piecemeal rather than systematically.

In stark contrast, cloze questions emerge as a more holistic tool. Expertly crafted, these questions adhere to pedagogically sound guidelines, ensuring learners are exposed to a wide array of grammatical constructs they ought to internalize.

Another limitation of the FCG task is its narrow scope. It zeroes in on explaining the suitability of specific words within a composition, often neglecting to clarify why certain expressions, even if they seem plausible, should be sidestepped.

Moreover, the inherent nature of FCG's feedback, emerging from free compositions, presents scalability challenges. Generating high-quality, nuanced commentaries on a large scale without extensive manual oversight remains a formidable task. This underscores the need for more structured and comprehensive methods in the explanation generation domain for language learning.

Pivoting to the domain of explainable NLP, EXPECT [33], is a significant addition. The research introduced the EXPECT dataset, a unique resource annotated with evidence words and error categorization. This dataset serves as a foundation for training models to not only identify but also elucidate grammatical errors, bringing a layer of transparency to automated GEC systems. The work notably integrated syntactic knowledge into models, leading to enhanced performance in detecting and explaining errors. This was further validated through human evaluations, emphasizing the system's efficacy in assisting second-language learners in comprehending and addressing their errors.

However, transitioning from this GEC-centered research to cloze explanation generation isn't straightforward. At the heart of this challenge lies the differences between the two tasks. While GEC focuses on identifying and rectifying grammatical discrepancies, cloze questions aim at assessing vocabulary knowledge, contextual comprehension, or a blend of both. The explanations required for each task are distinctly nuanced. Additionally, the modeling techniques optimized for GEC, especially those incorporating syntactic embeddings, might not be directly translatable to cloze tasks. For cloze, there could be a greater emphasis on semantics and maintaining coherence in passages. Another challenge is the specificity of the EXPECT dataset, tailored for grammatical errors, which might not be readily adaptable for cloze-focused explanations.

Chapter 3

Cloze Question Quality Estimation

3.1 Motivation

Cloze tests have long been recognized as an effective tool in language assessment, measuring various aspects of language proficiency, such as grammatical knowledge and reading comprehension ability. Especially prominent in language proficiency tests, cloze questions often present passages with blanks, requiring examinees to select or fill in words or phrases that make the passage coherent.

The traditional method of designing cloze tests relies on experts who meticulously curate the questions to ensure their reliability and validity in assessing language proficiency. However, given the rising costs and demand, there has been a shift towards automatic cloze generation methods. Although these automatic methods offer the advantage of scalability, they suffer from issues related to quality. Automatically generated cloze tests often do not match the quality standards of manually created ones, leading to tests that might not be reliable or valid for assessing a learner's language proficiency.

To illustrate, invalid tests might not properly assess the specific aspect of language knowledge they are intended for, making it challenging for educators to pinpoint areas of improvement for learners, e.g., the question is too easy to measure language abilities (Question 3 in Table 3.1). On the other hand, unreliable cloze tests might have multiple options that fit a blank equally well, making it difficult for even knowledgeable examinees to select the "correct" answer (Question 4 in Table 3.1).

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:

1.	A. dark	B. bright	***C. dim***	D. colorful
2.	A. make	B. makes	C. making	***D. made***
3.	A. is stopped	***B. lighted***	C. burning	D. drop
4.	A. formed	***B. illuminated***	C. relieved	***D. showed***
...				

Table 3.1 Example of cloze test and qualities for each question. Correct answers are highlighted in bold.

Adding to the complexity, while various works have claimed advancements in distractor generation or cloze test creation, comparing their efficacy is a challenge. Most of these works utilize their own evaluation metrics, making cross-comparison problematic. The predominant evaluation method remains manual assessments. However, these manual methods are resource-intensive, time-consuming, and sometimes challenging to procure, especially when expertise is required.

The increasing reliance on automated systems, combined with the noted challenges in ensuring their quality and the limitations in evaluation methods, underscores the need for robust solutions in the domain of cloze questions. This dissertation sets out to tackle these challenges, aiming to bridge the gap between automatic generation and the quality standards of expert-designed cloze tests.

3.2 Overview of the Cloze Quality Estimation Task

Addressing the challenges surrounding the evaluation of cloze tests, particularly focusing on distractors, this dissertation introduces the Cloze Quality Estimation (CQE) task. It's an innovative approach to measure the appropriateness of cloze questions.

In the context of our work, a “high-quality” cloze question should be reliable and valid. The former means the question should denote the language ability of the test taker; the latter requires the question to indicate what language skills the test taker owns/lacks [8]. Therefore, the key aim of the CQE task is to gauge cloze tests with regards to their two corresponding primary dimensions: reliability and validity.

Subsequently, we designed and introduced a novel dataset, the Cloze Estimation dataset for Language Assessment (`CELA`). This dataset encompasses a diverse set of English cloze tests, comprising those crafted by linguistic experts and those formulated through algorithms. Native English speakers subsequently solved these tests, providing invaluable annotations on each question’s reliability and validity.

For the CQE task, we have formulated baseline methods. These methods, specifically option-aware ones, critically evaluate cloze questions by analyzing their answer choices. When these methods were put to the test against the CELA dataset, the results offered insightful revelations about the challenges in detecting unreliable questions.

Finally, we also test LLMs on the `CELA`. LLMs have shown remarkable performance across diverse tasks in zero-shot scenarios [34]. To explore the potential of LLMs in solving the CQE task, we tested the performance of LLMs with different prompts on `CELA`.

3.3 CQE Task Definition

Input: An incomplete passage containing several blanks and accompanied by a series of questions, each providing multiple option tuples (Table 3.1).

Output: An estimation of the quality of each question, delineated in terms of its reliability and validity [8].

Reliability Evaluation: A cloze question’s reliability is assessed through a binary classification system. In symbolic representation, a reliable question is designated as $\mathbb{REL}$, while an unreliable one is marked $\mathbb{NREL}$. A pivotal concern here is the presence of multiple plausible answers for a single question. Simply put, a question with more than one potentially correct answer is flagged as unreliable. For example, in Table 3.1, Questions 1 3 are reliable but Question 4 is not.

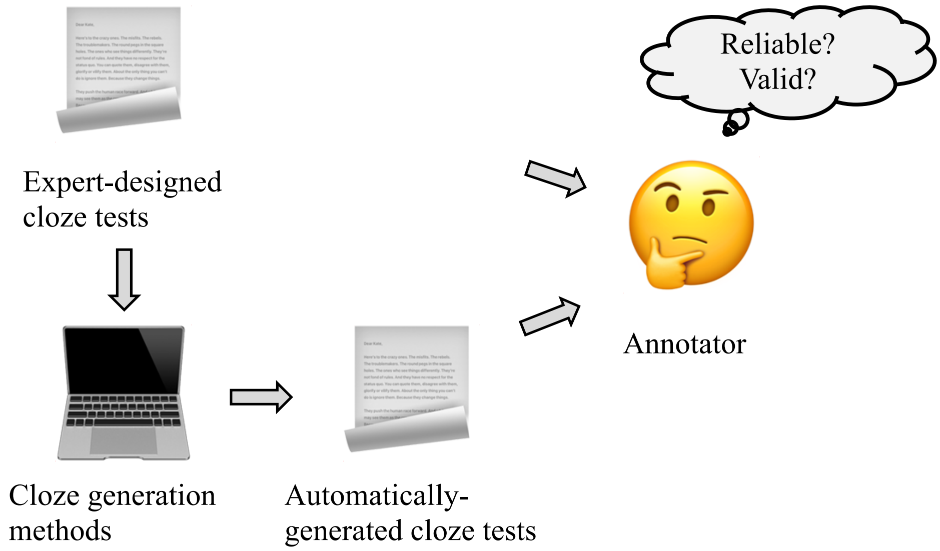


Fig. 3.1 Flow of creating CELA. Cloze generation methods use passages and blanks from expert-designed tests to eliminate the effect of the word deletion strategy.

Validity Evaluation: Validity is delineated through a tripartite classification system. In this schema, $\mathbb{VALG}$ represents questions valid for assessing grammatical knowledge, $\mathbb{VALR}$ signifies those valid for measuring reading comprehension abilities, and $\mathbb{NVAL}$ designates the questions as invalid. A valid question in this context necessitates the examinee to employ their linguistic expertise to differentiate the correct option from distractors. For example, in Table 3.1, Question 1 requires reading comprehension ability to answer and Question 2 asks grammatical knowledge.

Questions that potentially gauge multiple language proficiencies simultaneously are deemed invalid (Question 4 in Table 3.1). The question is considered to be too simple to measure the language ability of the examinee [35]. Therefore, questions that distinctly challenge either the examinee's grammatical knowledge or reading comprehension are marked valid; others are classified as invalid.

3.4 CELA Dataset

Figure 3.1 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.

3.4.1 CELA data preparation

We collected English cloze tests from Chinese senior-high-school examinations [6] 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 [36], which selects words that have the same POS tag with the answer from the vocabulary as distractors.

Jiang employs the method that selects words from the vocabulary but considers more factors including POS tag, word frequency, and spelling similarity [24]. Their method is designed for the Chinese cloze test, but we adapted it to the English test.

Panda uses round trip translation to paraphrase a passage and align the paraphrased passages with the original one [25]. They use aligned words to the answer as distractor candidates and select a distractor from candidates considering the synonym and POS tag.

Option examples (without passages) of each subset are shown in Table 3.2.

As a result, we collected and generated 150 cloze tests including 3,000 questions. The cloze tests are collected/generated in five ways, each accounting for one-fifth of the total.

3.4.2 CELA annotation

In the vast expanse of NLP, manual annotation emerges as a crucial, human-centric method of instilling interpretative layers upon raw data [37]. It involves the meticulous task of affixing metadata or labels, rendering the data intelligible, categorizable, and contextual within a predetermined framework. Predominantly, NLP leverages manual annotation for crafting labeled datasets, which subsequently serve as the empirical ground truth for both training machine learning models and evaluating their resultant proficiency.

Source: `CLOTH`				
1.	A. people	B. sound	C. fans	D. songs
2.	A. Saturday	B. day	C. time	D. concert
3.	A. cold	B. special	C. dark	D. successful
4.	A. young	B. famous	C. strong	D. black
5.	A. Classic	B. Country	C. Popular	D. Light
Source: `Randomized`				
1.	A. imagination	B. vain	C. novels	D. that's
2.	A. concerts	B. construction	C. finds	D. hydrant
3.	A. follow	B. well-tended	C. so	D. outlook
4.	A. morning	B. big	C. Today	D. when
5.	A. down-turned	B. Handbook	C. misdeed	D. No
Source: `Hill`				
1.	A. sacrificed	B. becomes	C. moved	D. surpassed
2.	A. youth	B. bombs	C. section	D. hand
3.	A. as	B. along	C. first	D. occurred
4.	A. Back	B. strangely	C. over,	D. hardly
5.	A. immediate	B. controlled	C. foolish	D. others
Source: `Jiang`				
1.	A. sudenly	B. shelly	C. golly	D. impatiently
2.	A. blamed	B. tilted	C. coordinated	D. resembled
3.	A. skill	B. improper	C. worry	D. preservation
4.	A. preservation	B. improper	C. min	D. skill
5.	A. bucephalus	B. neither	C. nor	D. and
Source: `Panda`				
1.	A. diminutive	B. small	C. walk	D. kid
2.	A. downstairs	B. across	C. around	D. over
3.	A. get	B. see	C. made	D. watch
4.	A. wonder	B. see	C. marveller	D. disbelief
5.	A. off	B. around	C. by	D. over

Table 3.2 Example of options in each subset in `CELA`.

The indispensability of labeled datasets in the realm of supervised machine learning in NLP cannot be overstated [38]. Such labels, standing as the "ground truth", pave the way for models to discern underlying patterns in data and, in parallel, provide a yardstick for scholarly performance evaluations. Given the intricate and context-sensitive nature of many linguistic phenomena, human intervention often becomes paramount to ensure precise annotations.

Manual annotations act as a bridge, facilitating the dialogue between lofty linguistic theories and pragmatic computational models. Such data, curated in line with linguistic tenets, not only aids in scrutinizing and honing linguistic theories but also steers the trajectory of algorithmic innovations in NLP.

Linguistic ambiguity remains a defining characteristic of natural languages [39]. While computational models grapple with linguistic subtleties, humans excel at teasing out nuanced meanings anchored in contextual cues. Manual annotation capitalizes on this inherent human prowess, illuminating the ambiguities, and thereby rendering data more digestible for models.

In the iterative process of NLP algorithmic or model development, the availability of standardized, quality-assured datasets emerges as a sine qua non. Such datasets, borne out of rigorous manual annotations, set the gold standard, enabling methodological comparisons and evaluations.

Annotation in NLP exhibits a rich tapestry of types: sequence classification (like sentiment analysis [40]), sequence labeling (such as POS Tagging [41]), and text generation tasks (e.g., document summarization [42]), among others.

However, the ever-present ambiguity in natural language propels the annotation process into a domain where individual experience and perception hold sway, introducing a measure of subjectivity. Hence, the pursuit of impeccable annotation quality becomes imperative for the fruition of robust NLP systems. To this end, several methodologies have been instituted to ascertain annotation quality.

Among these, inter-annotator agreement (IAA) stands out as the predominant metric [43], quantifying the concordance across annotations from multiple annotators. Renowned measures in this context include Cohen's Kappa [44], Fleiss' Kappa [45], and the more straightforward percentage agreement. Additionally, the institution of exhaustive and unambiguous guidelines [46, 47] ensures that annotators navigate the task with a shared understanding, thereby mitigating potential variances in annotations. Also, before embarking on large-scale annotation, a small subset of the data is annotated as a trial. This helps in identifying potential challenges, ambiguities, or misunderstandings in the guidelines.

A nuanced appreciation of the intricacies of manual annotation ensures that researchers and practitioners are better equipped to sculpt high-caliber datasets, which invariably underpin the successes in NLP applications.

In `CELA`, 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 3.3 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 [45]. 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 $\mathbb{VALR}$ instead of $\mathbb{VALG}$, 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 3.4. 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.

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 do not 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 do not 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 3.3 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 3.3 for examples."

3.4.3 CELA analysis

We observed that the five types of cloze tests have various qualities. Figure 3.2 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

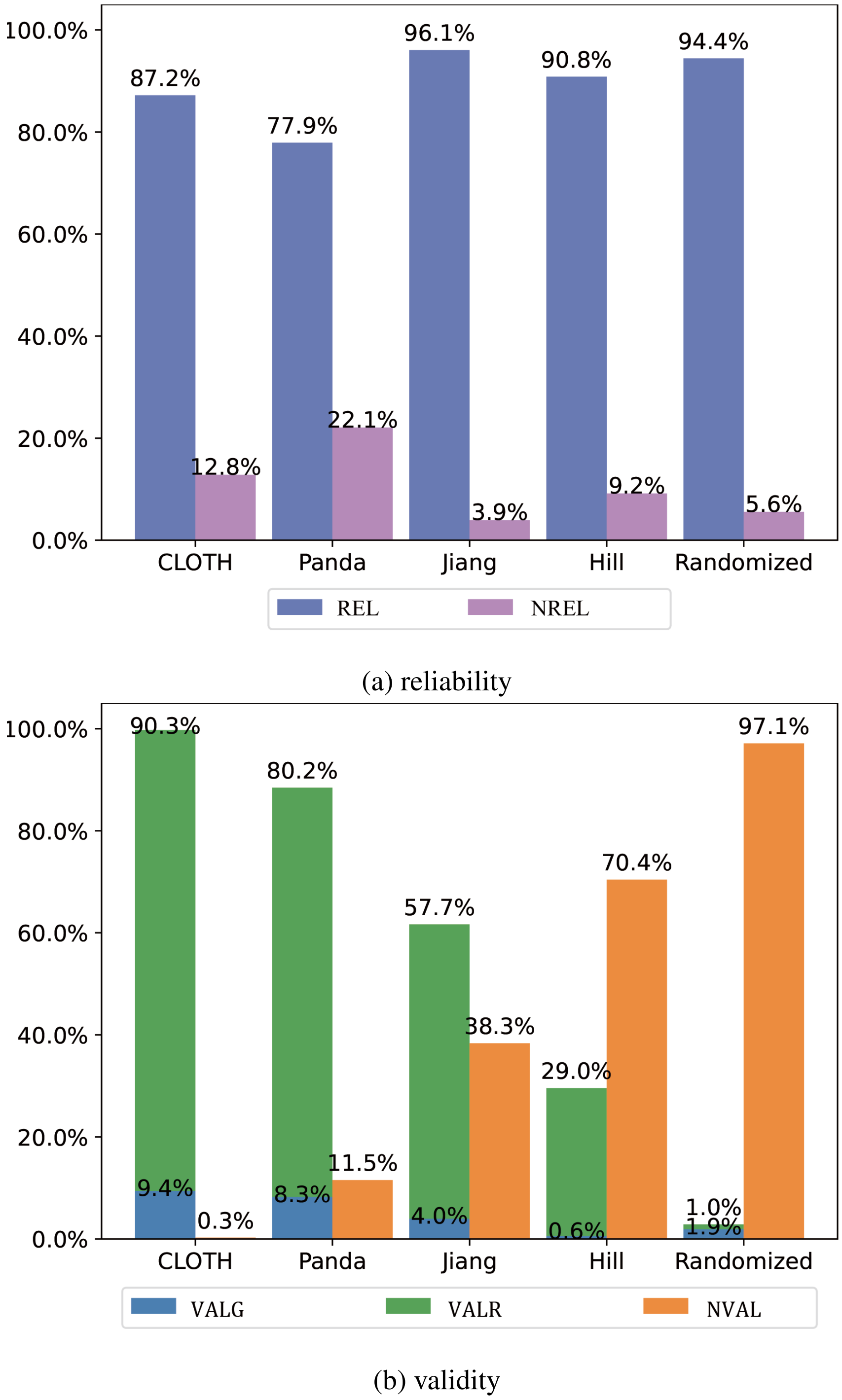


Fig. 3.2 Quality statistics of cloze tests in CELA. The left and right buckets represent the ratio of high-quality and low-quality questions, respectively.

Type	#
Reliability questions	2,597
$\mathbb{REL}$	2,324
$\mathbb{NREL}$	273
Validity questions	1,730
$\mathbb{VALG}$	86
$\mathbb{VALR}$	921
$\mathbb{NVAL}$	723

Table 3.4 Statistics of the processed data. Because reliability is easier to annotate, it has higher agreement, and more annotations are retained than validity.

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 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.

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.

3.5 Methodology

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

3.5.1 Option-aware Method

Intuition We designed an option-aware CQE method considering how options in the question affect reliability and validity. We followed the definition in section 3.3 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, $\mathit{BreakGrammar}(\cdot)$ and $\mathit{BreakSemantics}(\cdot)$, to judge whether an option breaks grammatical or semantic coherence.

The detailed description of the overall framework is shown in Algorithm 1. To realize these two functions, we designed two different approaches: a rule-based approach and a DNN-based approach. 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
$r \in \{\mathbb{REL}, \mathbb{NREL}\}$, $v \in \{\mathbb{VALG}, \mathbb{VALR}, \mathbb{NVAL}\}$

`// 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**
    $r \leftarrow \mathbb{REL}$;
    if $types.count(grammar) = n-1$ **then** $v \leftarrow \mathbb{VALG}$;
    else if $types.count(reading) = n-1$ **then** $v \leftarrow \mathbb{VALR}$;
    else $v \leftarrow \mathbb{NVAL}$;
else
    $r \leftarrow \mathbb{NREL}$;
    $v \leftarrow \mathbb{NVAL}$;
end
return (r, v);

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 [1].

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.

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.

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

Prompt For Reliability	Prompt For Validity
Evaluate given English cloze questions in aspect of reliability. In reliability, if a question only includes one correct answer, it is reliable. Otherwise, it is not reliable. Return the result in the format of "question index: *reliable/not_reliable*". The cloze questions are: [CLOZE TEST]	Evaluate given English cloze questions in aspect of validity. In validity 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. For valid questions, please indicate which language ability is measured (grammar or reading comprehension). Reture the result in the format of "question index: *valid_grammar/valid_reading/not_valid*". The cloze questions are: [CLOZE TEST]

Table 3.5 Zero-shot prompts for LLMs.

3.5.2 LLM Method

We employed GPT3.5-turbo [2] to explore the potential of LLMs on CELA. We designed zero-shot prompts for reliability and validity, respectively. The prompts we used are shown in Table 3.5. To invesgate whether LLMs could perfrom better with more information, we also designed few-shot prompts. Specifically, we add annotation examples shown in Table 3.3 in zero-shot prompts.

3.6 Experimental Setup for CQE Task

We conduct experiments to determine whether option-aware CQE methods (Section 3.5) can be a good baseline to estimate the quality of cloze tests, by comparing them with option-agnostic and LLM baseline methods (Section 3.6.2, Section 3.5.2).

3.6.1 Configurations

To implement an option-aware baseline with a rule-based approach, we built an English synonym dictionary [3]. 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.

[2] https://platform.openai.com/docs/guides/gpt
[3] collected from https://www.thesaurus.com/

The word lemmatization was implemented by employing the NLTK library [4] and using the lemmatizer based on WordNet [48]. 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 [49], a grammatical error corrector, that provided trained parameters and achieved a considerable performance on both CoNLL-2014 and BEA-2019 shared task [50, 51]. We used GECToR which was implemented by RoBERTa [52]. We fed original and corrected sentences into the ERRor ANnotation Toolkit (ERRANT) [5] 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.

3.6.2 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 $\mathbb{REL}$ and $\mathbb{VALR}$ in the sub-task of reliability and validity classification, respectively.

3.6.3 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

[4] https://www.nltk.org/
[5] https://github.com/chrisjbryant/errant

evaluation, we used F_1, precision, and recall score. In a binary classification context, these scores are defined as:

$$Precision = \frac{TP}{TP+FP} \tag{3.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.2}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \tag{3.3}$$

where TP, FP, and FN represent true positive, false positive, and false negative in prediction, respectively. Because unreliable cloze tests are harmful to language assessment, we must focus on how well CQE models can recognize unreliable tests; thus we set $\mathbb{NREL}$ 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 $\mathbb{VALR}$, $\mathbb{VALG}$, and $\mathbb{NVAL}$.

3.7 Result and Discussion

The performance of the baselines on the `CELA` dataset is presented in Table 3.6. For option-agnostic baselines, because of imbalanced data distribution, the majority prediction baseline was not able to detect the $\mathbb{NREL}$ 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 $\mathbb{REL}$ 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.

Methods	Reliability			Validity				
	F_1	prec.	recall	mic. F_1	mac. F_1	$r.F_1$	$g.F_1$	$n.F_1$
Option-agnostic								
- Random	17.5	10.6	49.8	32.3	27.9	39.9	8.4	35.5
- Majority pred.	0.0	0.0	0.0	53.2	23.2	69.5	0.0	0.0
Option-aware								
- Rule-based	2.9	1.5	<u>66.7</u>	42.4	41.3	30.3	37.1	56.6
- DNN-based ($\bar{O}$)	**19.5**	**98.5**	10.8	54.8	43.1	**72.3**	<u>47.0</u>	10.0
- DNN-based (O)	<u>19.3</u>	<u>97.8</u>	10.7	<u>58.5</u>	<u>48.3</u>	<u>71.9</u>	**53.8**	19.1
LLMs								
- Zero-shot	4.4	2.5	18.2	46.7	35.0	17.7	21.1	<u>66.1</u>
- Few-shot	14.2	7.7	**90.9**	**60.8**	**49.7**	56.0	21.1	**72.0**

Table 3.6 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 $\mathbb{VALR}$, $\mathbb{VALG}$, and $\mathbb{NVAL}$ questions, respectively. **Bold** and <u>underline</u> 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.

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 $\mathbb{NVAL}$ 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.

LLMs did not perform well in the case of zero-shot prompting, which means LLMs are not able to understand the task well when providing only instructions in prompts. Once providing several examples, LLMs could perform better, especially in *recall* of reliability and $r.F_1$ of validity. As a new task, CQE requires more detailed instructions in prompts to help the model understand.

3.8 Implications for Language Learning

The introduction and comprehensive exploration of the Cloze Quality Estimation (CQE) task marks an advancement towards a data-driven approach to language assessment. By placing a spotlight on the two pivotal components of reliability and validity in cloze questions, this research aligns more intimately with the foundational elements of efficient language education and assessment. Here are the primary implications of this study for language learning:

Enhanced Question Quality: By emphasizing the quality of cloze questions, this research brings forth the necessity of precise and meaningful language assessment tools. Through the CQE task, both educators and test creators have a framework that ensures the designed language tests are reflective of a learner's true linguistic capabilities.

Bridging Computational Linguistics and Education: This research exemplifies the harmonization of computational linguistics with language education. Automated CQE methods, with their specific attention to options, introduce an innovative dimension to the field. This synergy can facilitate the creation of effective test materials, ensuring students are evaluated through premium quality instruments.

Resource Encouragement: The establishment of the `CELA` dataset for the CQE task is not just a technical accomplishment but also a new direction for future academic pursuits. This dataset's specificity and design can serve as a foundational stone for subsequent studies, potentially catalyzing more advancements at the intersection of computational linguistics and language education.

In summary, the ramifications of this research are not limited to the technological domain. By enhancing the methodology behind cloze test evaluations, a ripple effect is initiated, ultimately leading to a qualitative improvement in language education. Through refined assessment tools, learners can be ensured of accurate and constructive feedback, aiding their linguistic journey.

3.9 Limitations

This research, while proposing approach to evaluating the quality of cloze tests, has several limitations:

Coverage of Cloze Test: The CQE task, as defined in this study, focuses on evaluating cloze tests generated by distractor generation methods, using expert-designed blanks. However, another pivotal aspect of cloze test creation, the word deletion methods determining which word should be blanked, has been overlooked. The influence of these methods on the overall quality of cloze tests remains an area yet to be explored. Furthermore, the composition of the `CELA` corpus is rooted in an expert-designed dataset intended for senior high school students in China, predominantly aligning with CEFR levels C2 to B1 [53]. Consequently, while `CELA` serves as a robust measure for questions within this specific difficulty spectrum, its efficacy in reliably evaluating questions of varying or extreme difficulty levels remains unverified. This limitation is significant when considering the principles of the Item Response Theory (IRT) [54]. IRT posits that the discriminatory power of a question, defined as its ability to distinguish between test-takers of differing proficiency levels, is an essential facet of question quality. The current scope of `CELA`, primarily tailored to a narrow proficiency band, may not adequately address this critical aspect, underscoring the need for a more diversified corpus encompassing a broader range of difficulty levels to fully encapsulate the complexities of cloze test assessment.

Scalability of Annotation: The meticulous annotation of question quality by experts, while ensuring high accuracy, poses challenges in terms of scalability. Constructing a large-scale dataset becomes a complex endeavor due to this expert-dependent annotation approach. A potential remedy to this limitation could be a strategic shift in the target data. Replacing native English speakers with non-native speakers, especially for data that does not necessitate high-level English proficiency (like CEFR-A level), might ease this constraint.

Language Specificity: The design of the CQE task and its accompanying `CELA` dataset is tailored specifically for the English language. Although the foundational principles of test design, such as reliability and validity, are universally applicable across languages, the intricate specifics might diverge based on linguistic characteristics. For instance, assessment in hieroglyph-based languages could necessitate glyph identification, altering the criteria for reliability and validity. Transitioning the CQE task to accommodate a different language would require access to a public cloze question dataset in that language, proficient cloze question generation techniques, and domain experts for quality evaluation. The vision for the future is to craft an automated adaptation mechanism to seamlessly extend the task and dataset to a plethora of languages.

Chapter 4

Explanation Generation for Cloze Questions

4.1 Motivation

As introduced in Chapter 3, cloze questions hold significant value in language assessment and are extensively used in renowned tests such as IELTS and TOEFL.

Providing explanations for answers to cloze questions can significantly enhance the language learning experience, especially for those engaged in self-study. Examples of explanations for cloze questions are shown in Table 4.1. When learners come across challenging cloze questions, access to clear and concise explanations can be immensely beneficial. Such explanations furnish the learner with insights into the reasoning behind the correct and incorrect choices, facilitating the identification and rectification of misconceptions. The provision of these well-constructed explanations can be instrumental in promoting deeper understanding and ensuring long-term knowledge retention.

Despite the evident importance of explanations, there is a noticeable void when it comes to generating high-quality explanations tailored for cloze questions. Existing research like the feedback comment generation (FCG) task proposed by Nagata [32] offers feedback on free English composition, but this approach is not without its limitations. For instance, the FCG task primarily focuses on explaining the appropriateness of specific words within a sentence and often misses out on highlighting why certain plausible expressions should be

Question 1: As expected, the infectious period had a clear _____ relationship with mean offspring number. (A) positive (B) positively (C) positives (D) positivity
Explanation 1: For this cloze question, you need to choose an adjective that modifies the word "relationship." Option (A) "positive" is an adjective, while options (B), (C), and (D) are not adjectives. Therefore, the correct answer is option (A), "positive."
Question 2: The couple were _____ Saturday on disorderly conduct charges by officers investigating a family dispute at their home in New Canaan, Conn. (A) arresting (B) arrested (C) arrest (D) arrests
Explanation 2: The blank in the cloze question requires a past participle verb, as indicated by the passive auxiliary "were" and the passive nominal subject "couple." The options given are: (A) arresting (gerund or present participle), (B) arrested (past participle) (C) arrest (base form), and (D) arrests (3rd person singular present). Based on this information, the correct answer is option (B), "arrested."
Question 3: The work was done _____ the Rehabilitation Institute of Chicago under an \$8-million grant from the Army. (A) at (B) down (C) round (D) of
Explanation 3: Based on the context of the sentence, option (A) "at" is the appropriate choice for the cloze question. "At" indicates a specific location or arrival at a particular place or position, such as "he is at the store."

Table 4.1 Examples of different types of cloze questions and corresponding explanations.

avoided. Furthermore, its reliance on free composition restricts the scalability of producing quality commentaries. This contrasts with the potential of cloze questions which, when designed by experts, ensures comprehensive coverage of grammatical constructs.

Therefore, a task and methods in Natural Language Generation (NLG) that focuses on cloze explanation is being expected. NLG is a specialized subfield of NLP that emphasizes the generation of coherent and fluent textual output from non-textual data or structured textual information. Contrasting with Natural Language Understanding (NLU), which deals with interpreting and understanding human language, NLG revolves around the construction and production of human-like language.

NLG has various applications. It is harnessed for automatically crafting written reports from data, such as financial summaries or medical reports [55, 56]. In the realm of journalism, NLG aids in generating news articles from structured data about events [57]. Moreover, it plays a significant role in narrative generation in machine translation where text is translated from one language to another [58].

Several challenges persist in the domain of NLG. Ensuring that the generated text maintains coherence and cohesion is crucial [59]. It should be logically consistent and flow in a manner natural to human readers. Another challenge is handling ambiguity to avoid producing vague or unclear statements. Diversifying outputs is essential, ensuring the system avoids repetitive patterns and generates varied responses. Keeping the content relevant to the user's intent or the given context is also vital. An ethical dimension also exists, as there is the responsibility of not generating misleading or biased content. Additionally, obtaining high-quality training data for NLG is a significant challenge. Creating manual annotations for NLG often requires more resources than for tasks like sequence classification, making the process expensive and time-consuming.

Historically, NLG began with rule-based and template-based systems [60]. These early systems operated on handcrafted rules and predefined templates, employing explicit language rules and slot-filling mechanisms. As machine learning evolved, statistical models such as Hidden Markov Models [61] and n-gram models [62] started gaining traction, particularly in machine translation. The advent of deep learning brought further innovations. Recurrent Neural Networks (RNNs) [63] and their advanced versions, like LSTM and GRU [64, 65], became popular for tasks that involved sequences. Later, Transformer-based architectures [66] became the mainstay. Models like OpenAI's GPT [67] and Google's T5 [68] set new benchmarks.

NLG, with its amalgamation of linguistic, statistical, and deep learning techniques, continues to bridge the divide between structured data and human-like language, promising innovations that span various industries.

4.2 Overview of the Cloze Explanation Generation Task

Building on the aforementioned importance of cloze questions in language learning and the necessity for detailed explanations, our work introduces a structured approach to bridge this gap. Recognizing the challenges in generating high-quality explanations for cloze questions, particularly in a self-study context, we formulated the `ClozEx` task. The primary goal of this task is the generation of concise and coherent explanations for English cloze questions, with a focus on achieving both fluency and informativeness.

A proficient explanation, besides offering a rationale for the answer, must also be accessible and impart relevant linguistic insights. The creation of our dataset, featuring over 140k pairs of questions and expert-reviewed explanations, substantiates this requirement. A snapshot of this dataset can be glimpsed in Table 4.1.

Diving deeper into the task, our investigative endeavors involved trialing multiple models, ranging from encoder-decoder frameworks to decoder-only structures. An intriguing aspect of our exploration was evaluating the capabilities of LLMs in a zero-shot scenario. This involved leveraging LLMs to spontaneously generate explanations without any prior fine-tuning on the specific task. Initial assessments revealed a dichotomy: while the fine-tuned encoder-decoder and decoder-only models showcased promising results, LLMs, though adept at crafting fluent text, occasionally fell short on the informativeness criterion. Furthermore, we observed that mere naive prompting of LLMs was inadequate for consistently producing high-caliber explanations.

This comprehensive undertaking not only positions `ClozEx` as a pivotal task in language learning but also underscores the complexities and nuances of generating cogent and insightful explanations for cloze questions.

4.3 ClozEx Task Definition

The ClozEx task, short for Cloze Explanation, is devised to address a pivotal gap in cloze question comprehension. For any given cloze question q (as illustrated by "Questions" in Table 4.1), methods tailored to handle the ClozEx task primarily act upon this input.

Delving into the structure, a cloze question embodies a sentence with a blank, which we represent as *sent*. Accompanying this are a set of options, denoted as $OPT = [opt_1, opt_2, ..., opt_n]$. Most commonly, n equates to four distinct options. The overarching goal is then to generate an explanatory text, *exp*, that sheds light on the correct choice for the given question. Such explanations can be found under "Explanations" in Table 4.1.

For an explanation to be deemed effective, it must adhere to two foundational criteria:

- Fluency [69]: The crafted explanation should exude coherence, ensuring it is easily comprehensible. A convoluted or challenging-to-decipher explanation would defeat its purpose, as it would nnt aid language learning effectively.
- Validity [70]: Beyond just being readable, the explanation must impart requisite knowledge. This involves disseminating pertinent linguistic insights that can guide the reader to the correct answer.

Of particular note is the task's specificity towards grammatical questions. While cloze questions span both grammatical and reading comprehension domains, ClozEx focuses exclusively on the former. The rationale behind this is the innate challenge associated with generating explanations for grammatical questions. For reading comprehension questions, explanations often gravitate towards defining individual words. If a learner comprehends the meanings of all the words in the question, they can answer it effortlessly. In contrast, grammatical questions require an understanding of external constructs, such as specific grammatical rules or conventions. This added layer of complexity mandates more nuanced explanations, presenting a challenging yet rewarding endeavor for the ClozEx task.

4.4 ClozEx Dataset

4.4.1 Data Preparation

Experts in English education can be hired to write explanations for cloze questions to provide very high-quality data. However, because of the consumption of time and human effort, datasets created in such a way are scale-limited. To mitigate the considerable cost associated with manual explanation generation, we need to explore an automated method for creating both the questions and explanations in our dataset.

Experts design cloze questions in a top-down manner, starting with a specific grammatical item. Subsequently, they designed various questions based on the grammatical item [2]. Such grammatical items could be regarded as a pattern of a specific group of cloze questions. A pattern can also be used to create new cloze questions with explanations. Thus, we designed a pattern-based method for automatic cloze question and explanation generation. This method extracts patterns from expert-designed cloze questions and explanations to ensure the quality. Then these patterns are used to generate new questions and explanations.

The data creation process is outlined in Figure 4.1. This method involves the extraction of patterns from expert-designed cloze questions and their corresponding explanations. These patterns serve as the foundation for generating new questions and explanations based on a publicly available corpus. During the question creation phase, sentences from a news corpus that align with a given pattern are selected. Distractor options are then generated based on which aspect of language is measured. For the explanation generation process, templates tailored to the question type are designed. These templates are populated with question and pattern information to yield initial explanations. Finally, we employ LLMs to paraphrase the template-based explanations, enhancing their fluency and diversifying their expression. To avoid redundancy, or an excessive amount of irrelevant information, in the generated explanation, we set a maximum length for the explanation (128 words).

4.4.2 Creation Methods by Question Types

We begin by focusing on three specific types of cloze questions: affix, verb-tense, and preposition. These question types have been selected based on their prominence in lan-

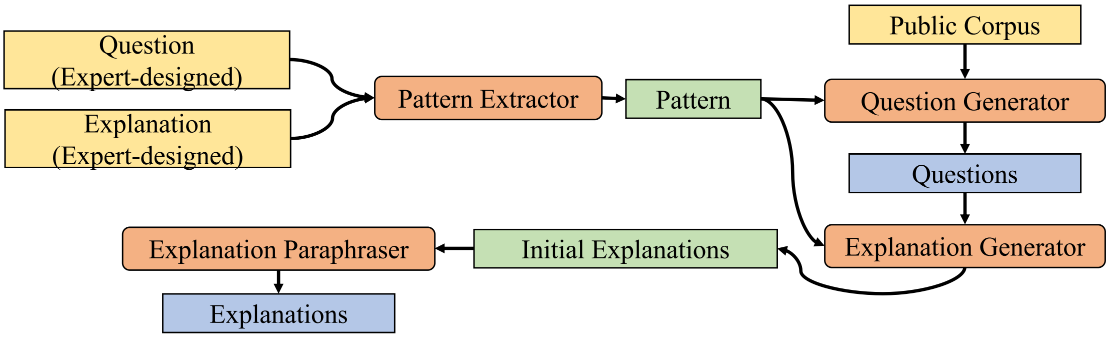


Fig. 4.1 Pipeline of data creation method. Yellow rectangles symbolize input to the pipeline, whereas blue rectangles represent output. Modules are depicted in orange, and their corresponding intermediate results are highlighted in green.

guage assessment [71–73], particularly in the context of the Test of English for International Communication (TOEIC). Affix questions require ESL learners to differentiate POS of options by analyzing prefixes or suffixes. Verb-tense questions prompt learners to identify the appropriate tense of the sentence and options. Preposition questions necessitate learners to comprehend the meaning of a sentence and consider the potential senses of the options. Questions 1 to 3 in Table 4.1 represent affix, verb-tense, and preposition questions, respectively.

The comprehension of affix and verb tense questions often relies on a narrower context within the sentence, allowing learners to answer without necessarily reading the entire sentence. By contrast, preposition questions require a comprehensive understanding of the sentence and an awareness of the various senses associated with prepositions. Therefore, affix/tense and preposition questions necessitate different focal points for extracting patterns and generating informative explanations.

Affix/Tense Questions 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 4.2). 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 4.2, 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 in formal writing and leads to fewer grammatical errors. If a parsing tree, $\tilde{tree}_i$, includes the extracted pattern $pattern_j$, we consider the corresponding sentence, $\tilde{sent}_i$, 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 4.2).

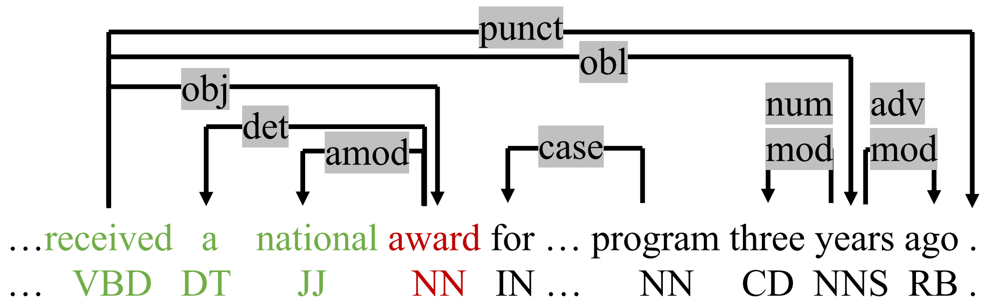
To select distractors for the new question, we built candidate dictionaries for affix and verb-tense questions, respectively. Distractor options are selected from the corresponding dictionary. For example, if an affix question has the answer option "contractor", the distractor candidates could be in ["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 4.2). To improve fluency and diversity, we employ LLM to paraphrase the template-based explanation.

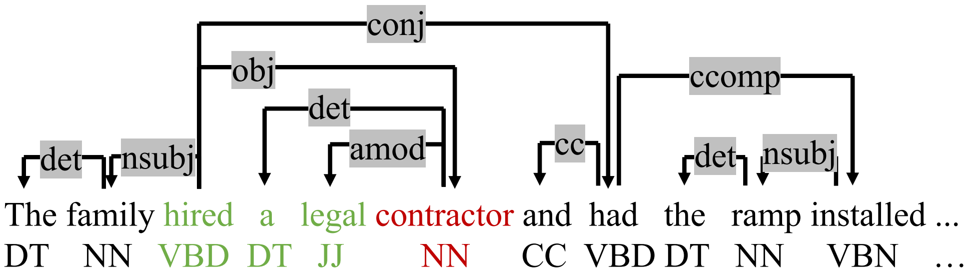
Prep. Questions Preposition questions require a comprehensive understanding of sentence meaning and the specific senses associated with the preposition options. Consequently, the pattern for a preposition question should incorporate the answer option along with its corresponding sense within the given sentence. To achieve this, we employed a preposition

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 4.1).



(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 4.1).

Fig. 4.2 Examples of process of generating a new question with its explanation.

sense disambiguation (PSD) model to determine the sense of the answer option within a particular sentence, denoted as $sent^{ans}$.

Subsequently, we consider the answer option together with its identified sense as the pattern, denoted as $pattern$. We then apply PSD to sentences extracted from a publicly available news corpus. If a sentence, denoted as $\tilde{sent}_i$, contains the pattern $pattern_j$, it is considered a viable candidate for generating a new preposition question.

When selecting distractor options for preposition questions, a straightforward approach would involve randomly choosing prepositions from a pool of available options. However, this method may yield simple questions that are easy to answer. Such simplistic questions fail to effectively gauge the language proficiency of ESL learners or aid in language learning [8]. As highlighted by previous research [74], prepositions sharing the same semantic relation often appear in similar contexts. By utilizing prepositions with similar semantic roles as distractor options, we can enhance the difficulty level of preposition questions. To facilitate this, we construct a dictionary to cluster prepositions based on their semantic roles, which aids in the selection of appropriate distractor options.

Finally, similar to the approach described in Section 3.2.1, we design a template to generate initial explanations, which are then refined by employing an LLM to enhance their fluency and diversity.

4.4.3 Implementation of Dataset Creation

Patterns for affix/tense questions were extracted from a published TOEIC practice book [75]. A total of 231 patterns were extracted from 432 affix questions, while 99 patterns were extracted from 219 tense questions. For preposition questions, we focused on 34 prepositions used in the PSD dataset [76] as question patterns.

To generate new questions and explanations, we selected the ag_news [77], cc_news [78], and multi_news [79] corpora from the public news corpus.

In the process of creating new preposition questions, we employed BERT-PSD [1] to identify the pattern present in each given sentence. Although BERT-PSD is a state-of-the-art model in the PSD task, it achieved an accuracy of only 90.84%, leading to potential noise in the dataset. To address this, we set a threshold of 0.8 for the model's prediction confidence.

[1]https://github.com/dirkneuhaeuser/preposition-sense-disambiguation

Affix/tense	The word in the blank should be [RELATION TO PARENT OF ANSWER] of "[PARENT OF ANSWER]". "[CHILD i OF ANSWER]" is [RELATION TO CHILD OF ANSWER] of the blank. Thus, a [POS OF ANSWER] is required. [OPTION i] is a [POS OF OPTION i]. Therefore, the correct answer is [ANSWER].
Prep.	According the meaning of this sentence the option [ANSWER] is suitable, which means "[SENSE OF ANSWER]".

Table 4.2 Templates used for generate initial explanations.

If the model predicted the pattern of a sentence with a confidence equal to or higher than 0.8, we retained the sentence along with its pattern for producing new questions and explanations. Otherwise, the sentence was discarded. With this threshold, the prediction accuracy improved to 97.78%.

For creating distractor options in affix questions, we prepared a distractor candidate dictionary in advance. We collected words from an English dictionary website[2] that share the same root but have different prefixes or suffixes. A similar process was followed for tense questions, where the distractor candidate dictionary focused specifically on verbs and their various tense forms. In the case of preposition questions, the distractor candidate dictionary was created based on preposition semantic relations [74]. Prepositions that share the same semantic relations are considered as distractor options for each other.

To avoid ambiguous questions that have multiple correct answers, we utilized a GPT2-based LM scorer [3]. If a distractor option obtained a higher LM score than the answer, as determined by the scorer, the option was discarded. The templates used for generating initial explanations for questions are shown in Table 4.2.

These initial explanations were further paraphrased using GPT3.5-turbo. The prompt for paraphrasing (the parameter of OpenAI API) is shown in Table 4.3.

[2]https://www.vocabulary.com/
[3]https://github.com/simonepri/lm-scorer

Role	Content
system	You are an English teacher.
user	Paraphrase the following explanation of a cloze question within 128 words: [exp].

Table 4.3 Prompt for paraphrasing initial explanations.

4.4.4 Dataset Analysis

To validate the quality and suitability of our created dataset for training models in the ClozEx task, we conducted a thorough manual quality assessment. As outlined in Section 4.3, the evaluation focused on two aspects: fluency and validity.

For the fluency assessment, we enlisted the expertise of two native English speakers from Tokyo Metropolitan University. These experts independently evaluated 100 randomly selected instances from our dataset 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 from Tokyo Metropolitan University, because these learners possess a strong understanding of textbook grammar [80]. All ESL learners in annotation hold public English test certificates to indicate they have a CEFR A2 level or higher. 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.

For the detail of manual estimation, human evaluators were tasked with rating the quality of generated explanations from each method in terms of fluency and validity using a 1-5 scale. The following criteria were provided to guide their ratings:

For fluency, the ratings were as follows:

- 1=Bad: The explanation was unreadable.
- 2=Unacceptable: The explanation was disfluent.
- 3=Borderline: The explanation fell between unacceptable and acceptable fluency.

	IAA		Estimation Score		
	Pearson	p-value	Average	Median	Variance
Fluency	0.82	<0.001	4.29	4.20	0.52
Validity	0.77	<0.001	4.51	4.50	0.45

Table 4.4 Inter-annotator agreement and manual estimation result.

- 4=Acceptable: The explanation was clear and understandable, but with room for improvement.
- 5=Good: The explanation was fluent and easy to understand.

For validity, the ratings were as follows:

- 1=Bad: The explanation included factual errors or was unrelated to the question.
- 2=Unacceptable: The explanation was related to the question but provided knowledge that did not contribute to answering it.
- 3=Borderline: The explanation fell between unacceptable and acceptable validity.
- 4=Acceptable: The explanation provided some necessary knowledge for answering the question, but there were still some missing elements.
- 5=Good: The explanation provided sufficient language knowledge to answer the question.

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 4.4. 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 dataset and support its publication as a reliable resource for the `ClozEx` task.

For a comprehensive understanding of our dataset, Table 4.5 presents a statistical analysis, providing relevant insights into its characteristics.

	#(Q, E)	Q average length	E average length
Train	102,930	28.99	58.53
Dev.	22,056	29.00	58.69
Test	22,057	28.95	58.47

Table 4.5 Statistics of our dataset. **#(Q, E)** represents number of (question, explanation) pairs. Average length of questions and explanations denote the number of tokens.

4.5 Experiment

To address the `ClozEx` task, we conducted an investigation into baseline models under various scenarios and architectures. To evaluate the performance of these baseline models, we conducted thorough assessments using development and test data from our dataset, encompassing both manual and automatic evaluation metrics.

4.5.1 Baseline Models

As a generation task, we employed encoder-decoder and decoder-only models for fine-tuning. In the case of the encoder-decoder models, we performed fine-tuning on BART [81] and T5 [82] architectures. For fine-tuning, we tailored cloze questions as input for the encoder-decoder models in the format of "$\{sent\}$[OPT]$\{opt_1\}$[OPT]...$\{opt_4\}$," where "[OPT]" is a special token that is used for concatenation among sentence and options. The output of the encoder-decoder models is the corresponding explanation. We explored different model sizes, including base and large, to assess their performance in the `ClozEx` task.

On the other hand, in the case of the decoder-only models for fine-tuning, we selected GPT2 and GPT2-medium [83]. For decoder-only models, the input is a question with an explanation that is connected with a prompt. We then fine-tuned models with such input instances.

Because LLMs have shown remarkable performance across diverse tasks in zero-shot scenarios [84], to explore the potential of LLMs in solving the `ClozEx` task without the need for additional training data, we employed LLMs of different sizes and structures to generate explanations without fine-tuning. We employed GPT2-large, GPT2-XL, GPT3.5-turbo [4], and

[4]https://platform.openai.com/docs/models

Model	Prompt
GPT2 & LLaMa	$\{sent\}$ Options: (A) $\{opt_1\}$ (B) $\{opt_2\}$ (C) $\{opt_3\}$ (D) $\{opt_4\}$ Explanation:$\{exp\}$
GPT3.5-turbo	Generate an explanation of the following cloze question: $\{sent\}$ Options:(A) $\{opt_1\}$ (B) $\{opt_2\}$ (C) $\{opt_3\}$ (D) $\{opt_4\}$'

Table 4.6 Prompt for generating explanations using LLMs. The parameter "role" of GPT3.5-turbo has the same value in Table 4.3.

LLaMa-7B [85] to generate explanations in the zero-shot scenario. The prompts used for the LLMs can be found in Table 4.6.

4.5.2 Evaluation Metrics

We engaged human annotators to estimate the fluency and validity of the generated explanation, following the same estimation process as described in Section 4.4.4. We randomly selected 100 samples of generated explanations from each model to be estimated. All instances were estimated without reference explanations, ensuring a reference-free evaluation.

To complement the manual annotation, which can be time-consuming and less generalizable, we also employed automatic metrics to assess the generated explanations. For reference-based metrics, we used BLEU-4 [86] from the Huggingface Evaluate library [5] to measure the similarity between the generated explanations and the reference labels.

$$\mathrm{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.1}$$

where p_n is the precision for n-grams. w_n are the weights for each precision (usually $w_n = \frac{1}{N}$ for N-gram BLEU, N equals to four in our experiment). BP is the brevity penalty, defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \quad (4.4) \\ \exp\left(1 - \frac{r}{c}\right) & \text{otherwise} \quad (4.5) \end{cases}$$

Here, c is the length of the candidate translation, and r is the effective reference length.

[5] https://huggingface.co/docs/evaluate/index

Evaluator	Prompt
GPT-Fluency	Estimate whether the given text is fluent. Here is the score definition: 1=Bad: The explanation was unreadable. 2=Unacceptable: The explanation was disfluent. 3=Borderline: The explanation fell between unacceptable and acceptable fluency. 4=Acceptable: The explanation was clear and understandable, but with room for improvement. 5=Good: The explanation was fluent and easy to understand. The input is: $\{exp\}$
GPT-Validity	Estimate whether the given explanation could explain the cloze question well. Here is the score definition: 1=Bad: The explanation included factual errors or was unrelated to the question. 2=Unacceptable: The explanation was related to the question but provided knowledge that did not contribute to answering it. 3=Borderline: The explanation fell between unacceptable and acceptable validity. 4=Acceptable: The explanation provided some necessary knowledge for answering the question, but there were still some missing elements. 5=Good: The explanation provided sufficient language knowledge to answer the question. The input question is: $\{q\}$ The explanation is: $\{exp\}$

Table 4.7 Prompt for GPT3.5-based evaluator. The parameter "role" of GPT3.5-turbo has the same value in Table 4.3.

According to ChatGPT evaluator [87], LLMs such as GPT3.5-turbo can evaluate the quality of generated text and exhibit a moderate correlation with human annotators. Therefore, we utilized GPT3.5-turbo as a reference-free metric to evaluate the fluency and validity of the generated explanations (named GPT-Fluency and GPT-Validity, respectively). The reliability of GPT evaluators will be discussed in Section 4.6.2. Samples used for the GPT evaluator are the same as manual estimation. All metrics except BLEU are based on the Likert 5-point scale.

Prompts for the GPT evaluator can be found in Table 4.7.

	Manual		Automatic		
	Fluency	Validity	BLEU	GPT-Fluency	GPT-Validity
BART-base	4.13	4.38	25.64 / 25.53	4.88	3.75
BART-large	4.11	**4.43**	**27.33 / 27.01**	4.84	2.90
T5-base	2.03	1.52	7.62 / 7.59	2.53	1.32
T5-large	3.99	4.26	22.70 / 22.68	**4.95**	2.31
GPT2	3.87	2.78	15.40 / 15.41	4.03	1.77
GPT2-medium	3.91	1.85	16.85 / 16.84	4.16	2.03
LLM-GPT2-large	3.97	1.73	0.50 / 0.51	3.94	1.58
LLM-GPT2-XL	3.97	1.70	0.60 / 0.60	4.00	1.58
LLM-GPT3.5-turbo	**4.53**	2.70	1.39 / 1.34	4.93	**4.87**
LLM-LLaMa-7B	3.81	1.78	1.06 / 1.08	3.81	1.44

Table 4.8 Performance of baseline models. BLEU scores are based on dev. and test sets, respectively. In evaluation metrics, GPT-Fluency and GPT-Validity indicate fluency and validity estimation, respectively, using GPT3.5-turbo. Prefix LLM- denotes LLM-generated explanations. Except for BLEU, all scores are ranged in [1, 5].

4.6 Result and Discussion

4.6.1 Result

The evaluation results are presented in Table 4.8, and Table 4.9 provides examples of generated explanations from each model. With regard to the manual metrics, the encoder-decoder models generally exhibited the ability to generate fluent and valid explanations, except for T5-base. BART-large achieved the highest level of validity performance. By contrast, the decoder-only models based on GPT-2 produced acceptably fluent texts but did not effectively explain the questions. Across all the fine-tuned models, the size of the model did not have a substantial impact on performance, except for T5 base and large, where it hindered the generation of more valid explanations. Despite the large amount of training data, texts generated by LLMs demonstrated acceptable fluency, particularly with LLM-GPT3.5-turbo yielding the most fluent texts. Surprisingly, the LLMs in our experiment received low evaluations in terms of producing valid explanations. This highlights the ongoing challenge of using LLMs to generate cloze question explanations for LA, without mentioning the generation of a dataset specifically tailored for the `ClozEx` task. A detailed discussion regarding the performance of LLMs is included in Section 4.6.2.

Model: BART

Question:

The repayment, which settles a suit brought under an international legal tribunal, is _____ from the tens of billions of dollars in frozen foreign accounts that Iran can now access after the end of nuclear sanctions.

(A) separates (B) separability (C) separateness (D) separate

Explanation (BART-base):

The blank in the cloze question is filled by the copula "is". This means that an adjective is needed to complete the sentence. The options given include a verb in the third person singular present tense, a singular or mass noun, an adjective, and another adjective. The correct answer is the adjective "separate".

Explanation (BART-large):

The blank in the cloze question is filled by the copula "is". This means that an adjective is needed to complete the sentence. The options given are: (A) separates, which is a verb in the 3rd person singular present tense; (B) separability, a singular or mass noun; (C) separateness, also a noun in the same category; and (D) separate, an adjective. Based on this information, the correct answer is option (D), which is "separate".

Model: T5

Question:

South Korean shares slid on Friday, with foreign investors _____ their biggest daily volume in nearly 20 months, hurt after aggressive easing from the Bank of Japan sent the yen reeling, as well as by the tension over North Korea.

(A) selling (B) undersells (C) resell (D) sell

Explanation (T5-base):

The stock market slid on Friday, with investors ________________________

Explanation (T5-large):

The correct answer for the cloze question is (A) selling, which is a verb in the form of a gerund or present participle. This is indicated by the fact that the word "selling" can be used as a verb in this context.

Model: GPT2

Question:

The father of the former Meghan Markle says he wishes he could have walked her _____ the aisle during her wedding to Prince Harry.

(A) with (B) down (C) as (D) for

Explanation (GPT2):

Based on the definition provided, option (b) “down” is the appropriate choice for the blank in this sentence. This word indicates movement towards a lower position or level, as in the example “he descended the hill.” This word can also be used to describe the action of

Explanation (GPT2-medium):

The correct answer for the cloze question is option (b) down, which refers to the action of moving towards a lower point or level, such as when someone goes down a hill. This is based on the sentence’s meaning. For example, “he

Models: LLMs

Question:

The repayment, which settles a suit brought under an international legal tribunal, is _____ from the tens of billions of dollars in frozen foreign accounts that Iran can now access after the end of nuclear sanctions.

(A) separates (B) separability (C) separateness (D) separate

Explanation (GPT2-large):

The first option is to separate the accounts from the rest of the assets. This would allow the U.S. to keep the money in the accounts, but not the assets. The second option is to separate the accounts from the rest of the assets. This would allow the U

Explanation (GPT2-XL):

Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option. Separation is the most common option.

Explanation (LLaMa-7B):

The repayment, which settles a suit brought under an international legal tribunal, is from the tens of billions of dollars in frozen foreign accounts that Iran can now access after the end of nuclear sanctions. Question: The repayment, which settles a suit brought under an international legal tribunal, is from the tens of billions of dollars in frozen foreign accounts that Iran can now access after the end of nuclear sanctions
Explanation (GPT3.5-turbo): In this context, the word "separate" (option D) is the correct choice to complete the sentence. The sentence is discussing a repayment that is distinct or independent from the tens of billions of dollars in frozen foreign accounts. The word "separate" means to set apart or make a distinction between two things, which fits the context of the sentence. Options A, B, and C do not convey the intended meaning in this context.

Table 4.9 Examples of generated explanations.

With regard to the automatic metrics, the BLEU score exhibited a strong correlation with manual fluency and validity scores when evaluating models fine-tuned with our training data. However, because the LLMs did not learn the distribution from our training data, the generated text varied from the reference. Because a good explanation for a cloze question is not unique, reference-based metrics should focus on evaluating models trained with our data. In this regard, BART-large achieved the best performance once again.

The GPT evaluator demonstrated stability in terms of fluency. These GPT-Fluency scores showed a positive correlation with manual fluency scores. However, in terms of validity, the GPT evaluator was less consistent, assigning varying scores to models that received similar validity scores from human annotators (such as BART-base, BART-large, and T5-large). Notably, LLM-GPT3.5-turbo was highlighted, because the GPT evaluator exhibited more leniency toward it than human annotators.

Finally, although these automatic scores showed some correlation with human evaluation, they were calculated under the macro average. To determine the reliability of these automatic metrics in the `ClozEx` task, we will discuss the micro-averaged Pearson correlation coefficient between manual and automatic scores in Section 4.6.2.

4.6.2 Discussion

Do LLMs Explain Cloze Questions Well? Given the remarkable performance of LLMs across various tasks without fine-tuning [88], there is a reasonable expectation that they would excel in generating high-quality explanations for cloze questions. However, our experimental 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 **LLM-GPT3.5-turbo Question 1** Table 4.10, 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 **LLM-GPT3.5-turbo Question 2** in Table 4.10, 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.

The outcomes of our research underscore the inherent challenges associated with generating cloze explanations in a zero-shot scenario using LLMs. Our findings revealed that the explanations generated by LLMs, based on the current set of prompts, did not meet the expectations of our annotators. However, it is crucial to note that this does not necessarily imply the unsuitability of LLMs for the `ClozEx` task in its entirety. In fact, there are specific techniques, such as in-context few-shot learning and chain-of-thought prompting, that are known to significantly enhance the capabilities of LLMs [89]. These methods, which effectively leverage the underlying power of LLMs, hold promise for more efficacious application in cloze explanation generation. We anticipate that with the integration of suitable tricks and meticulously designed prompts, LLMs could demonstrate marked improvements in their performance, offering more precise and informative explanations for cloze questions. This potential advancement signals a promising direction for future exploration in harnessing the power of LLMs for more nuanced and effective language comprehension tasks.

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.

Table 4.10 Negative examples of explanations generated by LLM-GPT3.5-turbo.

	Manual Fluency	Manual Validity
BLEU	0.04 / 0.17	0.08 / 0.11
- w/o LLMs	0.39 / 0.43	0.44 / 0.47
GPT-Fluency	0.57 / 0.61	—
GPT-Validity	—	−0.03 / 0.05

Table 4.11 Pearson correlation coefficient between manual and automatic evaluation scores. The automatic scores yielded two correlated coefficients because each instance is assessed by two annotators.

Are Automatic Metrics Reliable in `ClozEx`? The evaluation of automatic metrics, specifically BLEU and GPT-Fluency scores, aligns with the trends observed in manual evaluation scores (Section 4.6.1). To ascertain the reliability of these metrics in reflecting the quality of generated explanations, we computed the micro-averaged Pearson correlation coefficient between manual and automatic evaluation scores.

As shown in Table 4.11, the BLEU score is largely independent of the manual fluency score. However, when excluding explanations generated by LLMs, the BLEU score exhibits a moderate correlation with the manual fluency score. The validity correlation reported a similar tendency. As a reference-based metric, BLEU demonstrates limitations in recognizing explanations with different styles from our dataset, implying that a low BLEU score does not necessarily indicate a poor explanation. However, due to the high quality of our dataset, an explanation with a high BLEU score can generally be considered good.

As a reference-free metric, GPT-Fluency exhibits a strong correlation with manual fluency scores, even when considering LLM explanations. Unlike the correlation observed between GPT-Fluency and Manual Fluency, GPT-Validity fails to effectively reflect the manual validity score. Furthermore, for explanations generated by LLM-GPT3.5-turbo, as mentioned previously, GPT-Validity tends to assign higher scores. In light of these findings, when a reference-free evaluation is conducted, it is acceptable to employ LLMs such as GPT3.5-turbo to assess fluency in the `ClozEx` task. However, using LLMs to evaluate validity is not recommended.

4.7 Implications for Language Learning

The introduction of the `ClozEx` task has brought to the fore the importance of generating insightful and coherent explanations for English cloze questions. This endeavor is paramount, especially when considering the context of ESL learning. The task does not just echo the intrinsic significance of cloze questions in language education, as articulated in Chapter 3, but further underscores the profound impact of pertinent explanations in bolstering the learners' comprehension.

A major milestone achieved in this regard is the collation of an expansive dataset, amassing over 140k instances of cloze questions coupled with their respective explanations. The meticulous pattern-based method employed for dataset creation draws patterns directly from expert-designed cloze questions and explanations, ensuring the curated questions and explanations resonate with high standards of quality. It is worth noting that this dataset's caliber has been affirmed by experts, underlining its robustness and aptness for the `ClozEx` task.

Diving deeper into the mechanics of the task, various models were rigorously fine-tuned to cater to explanation generation needs. The gamut ranged from encoder-decoder frameworks to decoder-only architectures. Alongside these, the prowess of LLMs was harnessed, particularly in a zero-shot scenario where they were tasked with spontaneously generating explanations. While the encoder-decoder models demonstrated commendable prowess in churning out high-grade explanations, the LLMs showcased an intriguing behavior. Their knack for crafting fluent texts was evident; however, their ability to consistently generate valid explanations wavered. This underscores the challenges inherent to LLMs, particularly when they are thrust into producing valid explanations without the benefit of prior fine-tuning.

Yet another critical aspect delved into was the exploration of the correlation between manual and automated evaluation metrics. It is heartening to observe that automated metrics, although not flawless, do exhibit a promising degree of reliability when juxtaposed against manual assessments for the `ClozEx` task.

In essence, the `ClozEx` task has not only unveiled the significance of valid and fluent explanations in the realm of ESL learning but has also shed light on the multifaceted challenges and prospects tied to generating such explanations. The insights gleaned from this task promise to shape future endeavors in ESL education, making learning more intuitive, informed, and impactful.

4.8 Limitations

The exploration into the `ClozEx` task, designed to augment ESL learning, has shed light on numerous insights and novel methodologies. However, like any scientific endeavor, our study is not without its limitations.

Firstly, while the primary objective of the `ClozEx` task was to support ESL learning, the effectiveness of this endeavor in truly enhancing language proficiency remains unevaluated. Though our dataset, vetted through expert estimation, offers promising results, it provides only a surrogate indicator of the utility of the generated explanations for ESL learning. Direct evidence of efficacy remains elusive. To address this, subsequent experiments should be orchestrated, potentially involving pre- and post-exposure evaluations of ESL learners' proficiency after interacting with our dataset. Such an approach would yield a clearer picture of the tangible impacts of our materials on language acquisition.

The breadth of question types within our dataset constitutes another limitation. Our current dataset, built upon pattern extraction methodologies, has its genesis in three specific types of questions. Yet, the expansive domain of language assessment presents a diverse array of question types, encompassing areas such as word meaning identification and the nuanced use of linguistic components like pronouns and conjunctions. Given that our dataset's creation method was narrowly focused on certain question types, it may not adequately represent the full spectrum of language assessment. Future endeavors would benefit from refining pattern extraction techniques to encompass a more diverse set of questions, thereby offering a more comprehensive dataset.

Another constraint stems from the evaluative metrics utilized. While we identified a positive correlation between BLEU and GPT-Fluency scores against manual evaluations, these metrics bring their own set of challenges. BLEU, anchored in its reference-based nature, often falters when multiple plausible explanations exist. On the other hand, while reference-free metrics such as GPT-Fluency present a promising alternative, their consistent reliability is yet to be firmly established. Furthermore, the closed-source nature of models like GPT3.5-turbo, which underpin our GPT evaluators, poses potential impediments for future research endeavors, potentially restricting replicability and further advancements.

Lastly, our reliance on LLMs unveiled their limitations, particularly in the context of generating explanations for cloze questions. Despite their evident prowess in paraphrasing during dataset creation, they fell short when generating explanations without auxiliary information. This highlights an intrinsic limitation of LLMs, pointing towards their occasional inability to weave precise and contextually relevant explanations. Tackling this would necessitate innovative approaches, potentially involving the infusion of external grammatical knowledge, to enhance LLMs' capability in delivering clear, informative, and relevant explanations for cloze questions.

Chapter 5

Conclusion

5.1 Conclusion

In this dissertation, two significant avenues in the realm of cloze test question analysis and explanation generation were delved into, each contributing invaluable insights to the overarching goal of enhancing English language learning, especially for ESL learners.

5.1.1 Cloze Quality Estimation.

The meticulous exploration and investigation into the quality estimation of cloze tests have made several foundational contributions to the domain:

Introduction of a Novel Task: One of the standout contributions of this research was the proposition of a novel task aimed at assessing the quality of cloze tests. The introduction of this task filled a noticeable gap in the realm of language assessment, particularly for ESL learning environments. By setting the stage for an entire domain of study, this task laid the foundation for subsequent explorations, methodologies, and innovations.

Development of a Manual Annotated Dataset: A significant step in any machine learning or computational linguistics task is the creation of a reliable dataset. Recognizing this necessity, the research provided a meticulously curated, manual annotated dataset. This dataset not only served as a platform for testing and evaluating the proposed models but also offers an invaluable resource for future research endeavors.

Establishment of Baseline Models and Performance Insights: The research did not stop at just proposing the task and providing a dataset. It further delved into the introduction of baseline models, leveraging both rule-based and deep neural network approaches. These models were tailored to the unique requirements of the quality estimation task. Moreover, the rigorous evaluation of these models offered clear insights into their strengths and limitations. The results from these evaluations illuminated potential pathways for further refinement, highlighted the challenges faced, and provided a clear understanding of where current methodologies stand in the context of the task.

In summary, the investigation into cloze test quality estimation, through its pioneering task proposition, dataset creation, and baseline model evaluations, has significantly advanced the field. It has not only set the stage for further research but also provided the necessary tools and insights to guide subsequent explorations.

5.1.2 Cloze Explanation Generation.

Our rigorous exploration into cloze explanation generation has led to multiple significant milestones in the realms of linguistics and education. Central to our efforts was the initiation of a groundbreaking task focused on the generation of fluent and valid English cloze explanations for ESL learning. This novel undertaking aimed to fill a pivotal void in the sphere of language assessment and learning, setting the foundation for innovative methodologies and targeted research in this area.

Complementing this was the development of a comprehensive, expert-quality-assured dataset tailored for the `ClozEx` task. This dataset, encompassing over 140k instances, was meticulously curated using a pattern-based method. Its creation, coupled with expert validation, signifies its importance as a premier resource, facilitating the benchmarking and training of models for this distinct task.

Our investigation did not restrict itself to task formulation. We undertook an exhaustive examination of model performances when fine-tuned on our dataset. Furthermore, our exploration extended to the capabilities of LLMs, shedding light on their potential to churn out appropriate explanations in a zero-shot scenario. This multifaceted evaluation provided a holistic view of the strengths and challenges associated with various modeling approaches.

Rounding off our contributions, we delved into the dynamics of evaluation metrics. A critical correlation analysis between automatic and manual evaluations in the context of the `ClozEx` task unveiled the degree of trustworthiness of these metrics. This introspection aimed to discern the alignment of automatic metrics with human judgements, ensuring the robustness and reliability of the explanations generated.

In totality, our journey in cloze explanation generation has not only erected foundational pillars for this domain but has also offered rich insights, a dependable dataset, and a thorough model evaluation, setting a robust trajectory for subsequent research endeavors.

5.1.3 Real-World Impact

The advancements made in this dissertation, encompassing both CQE and `ClozEx`, have far-reaching implications that extend beyond theoretical academic contributions, making a substantial impact on the practical aspects of language learning and assessment. By automating and refining the processes involved in creating and evaluating cloze tests, this research directly addresses several pressing challenges in contemporary education and language studies.

For educators and assessment designers, the innovative models and methodologies introduced significantly alleviate the logistical and financial burdens traditionally associated with developing language assessment tools. This streamlining of resources is particularly transformative for institutions with limited access to linguistic expertise or financial constraints, democratizing the availability of high-quality educational content. The CQE allows the on-the-fly high-quality cloze question creation, which means educators or automatic question generation methods could obtain real-time feedback on the question quality and thus modify them promptly.

For learners, particularly those engaged in ESL education, the enhancements in assessment quality and explanation generation contribute to a more enriched learning experience. They receive more reliable, understandable, and detailed feedback on their performance, allowing for more targeted personal development. This nuanced educational support was largely unfeasible in previous frameworks due to the prohibitive costs of individualized, expert-led assessment.

Additionally, by establishing more objective and consistent assessment criteria, this research combats the subjectivity and bias that can often infiltrate language evaluations. This move toward greater objectivity not only fosters a more equitable academic environment but also ensures that language proficiency assessments are more reflective of a learner's true capabilities, thereby facilitating fairer educational and professional opportunities.

In the grander scheme, the synergistic effect of the CQE and `ClozEx` research encapsulates a move toward smarter, more efficient, and just educational practices. It signals a shift in educational paradigms, where advanced artificial intelligence (AI) and linguistic research break down barriers to quality education, making learning more accessible, personalized, and effective. This dissertation, therefore, marks a significant stride forward in harnessing technology's power to transcend traditional educational limitations, providing scalable solutions that could reshape the future landscape of language education and beyond.

5.2 Future Work

Building upon the foundational research in both the realms of Cloze Test Quality Estimation and Cloze Explanation Generation, several intriguing avenues lie ahead for deeper investigation and advancements. Each stride forward promises to deepen our understanding, enhance methodologies, and most importantly, contribute meaningfully to the domain of language assessment.

5.2.1 Cloze Quality Estimation.

In the sphere of Cloze Test Quality Estimation, our exploration unveiled several opportunities for further refinement. The continuous enhancement and diversification of our dataset to encompass an even broader range of cloze test question types and linguistic intricacies can provide a more comprehensive training ground for models [90]. As some deep learning models have already shown promise in initial experiments, the quest for cutting-edge architectures that cater specifically to the nuances of the CQE task remains. Ensuring the robustness and reliability of models, especially in detecting unreliable cloze tests, is paramount [91]. This challenge might be met through the harnessing of advanced NLP techniques and algorithms.

The potential of transfer learning and domain adaptation [92], especially considering the IRT, could be a cornerstone for future endeavors. Integrating IRT can significantly enhance this approach by considering three critical aspects: (1) difficulty: IRT emphasizes that the difficulty of each item in a test is crucial. This means analyzing how challenging each cloze question is for the test population, which can provide insights into the appropriateness of the test for different proficiency levels. (2) discrimination: This refers to an item's ability to differentiate between test-takers of different proficiency levels. High discrimination items are effective in distinguishing between more and less skilled individuals. (3) guessability: This concept deals with the probability of correctly guessing an answer. In cloze tests, it's important to assess how likely it is for a test-taker to guess the correct answer without actually knowing it, as this can affect the test's validity.

Envisioning real-time application, the development of tools providing instantaneous feedback on cloze test quality to educators and test-makers could revolutionize language assessment material creation. Moreover, broadening the horizons to encompass CQE for languages other than English could set the stage for globally relevant research [93]. Collaborative efforts, bringing together computational experts, educators, and linguists, could further enrich the research landscape, ensuring a holistic approach to language assessment [94].

5.2.2 Cloze Explanation Generation.

Transitioning to Cloze Explanation Generation, our introduction of a new task specifically tailored for the generation of fluent and valid English cloze explanations for ESL learning marks just the beginning. The creation of our large-scale and expert-quality-assured dataset, encompassing more than 140k instances, provides a robust platform for further experiments. While our investigation into model performance trained on this dataset yielded significant insights, further explorations into advanced architectures and techniques could refine explanation generation. Moreover, while LLMs displayed remarkable capabilities in many domains, their proficiency in the cloze explanation generation context, especially in a zero-shot scenario, remains a fertile ground for research [95, 96]. Our examination of the correlation between various evaluation metrics emphasized the need for more nuanced and reliable measures [97]. Delving deeper into the synergy between manual and automatic evaluation methods could bolster the assessment paradigm for generated explanations.

Additionally, taking into account the diverse language proficiency levels of learners, the development of an adaptive, on-the-fly explanation generation system would significantly enhance the utility of `ClozEx`. Such a system, capable of tailoring explanations to suit individual learners' language abilities, would not only improve the inclusivity and reach of `ClozEx` but also ensure that each learner receives optimally beneficial and comprehensible feedback. This customization is pivotal in catering to a broad spectrum of learners, each with their unique linguistic needs and learning trajectories.

5.2.3 Advancing NLP in Language Assessment and Education

Beyond the immediate next steps for enhancing CQE and `ClozEx`, our research's implications prompt a more expansive vision for the future of NLP-powered language education. This broader perspective considers not just incremental improvements in existing methodologies, but also how these advancements could revolutionize learning paradigms.

The integration of advanced NLP techniques has the potential to dramatically reshape educational ecosystems. Personalized learning experiences, bolstered by AI, could become the norm, with systems capable of adapting content, feedback, and assessments to individual learners' needs. This approach could mitigate current one-size-fits-all models, accommodating diverse learning paces and styles, and ultimately, democratizing access to quality education across various socio-economic contexts.

The future beckons a surge in interdisciplinary research, combining cognitive science, pedagogy, and computational linguistics to develop holistic educational tools [98]. These tools would transcend traditional language education confines, facilitating not just grammar or vocabulary acquisition, but also critical thinking, creativity, and cultural understanding. By harnessing insights from diverse fields, we can foster more nuanced, context-aware applications that resonate with complex human learning processes.

As we advance, ethical considerations will take center stage in AI-powered education [99, 100]. Future work must address bias in machine learning models to ensure fair and equitable language education. This involves creating diverse and inclusive datasets and continually scrutinizing AI methodologies' implications on all demographic groups. Ensuring that AI-driven tools uphold the highest ethical standards is imperative for their sustainable integration into educational frameworks.

Looking ahead, there is immense scope for global collaboration in developing universally accessible language learning resources. Consortia of researchers, educators, and policymakers worldwide could unify to share knowledge, tools, and best practices, driving innovation and quality in language education. This international synergy might lead to multilingual and multicultural educational platforms, broadening learners' horizons and fostering global citizenship.

Lastly, the evolution of NLP-powered education must continually assess its impact on learning outcomes. This means establishing robust feedback mechanisms, where student performance data refine educational AI tools iteratively. Harnessing real-time analytics and longitudinal studies will be vital in understanding these technologies' long-term effects and ensuring they substantively contribute to education quality improvement globally.

Ethical Considerations

In conducting this research, we placed a significant emphasis on the ethical considerations related to the recruitment and treatment of human annotators. The following section outlines the ethical principles adhered to during the experiments:

- Informed Consent: Prior to participation, all annotators were informed about the nature and purpose of the experiment. They were made aware of what was expected of them, the time required, and the compensation they would receive. Annotators were also assured that they could withdraw from the task at any time without any negative repercussions.
- Fair Compensation: Annotators in the CQE experiment were paid $1.5 USD for each batch, which took approximately 5-7 minutes to annotate. This ensures that their pay rate was considerably above the local minimum hourly wage. In the `ClozEx`, annotators were compensated at $13.5 USD per batch, which took about an hour to complete. This rate is nearly twice the local minimum wage of $7.5 USD, reflecting a commitment to fairly compensating participants for their time and expertise.
- Privacy and Confidentiality: All personal information, including English test certificates for non-native speakers, was treated with strict confidentiality. The data was stored securely, and any identifying information was separated from the annotated results to ensure anonymity in the dataset.
- Institutional Approval: Importantly, before embarking on this research, the experiments were reviewed and approved by the ethical committee at Tokyo Metropolitan University. This external review process ensured that our research methodologies and practices aligned with established ethical standards.

- Feedback and Queries: An open channel of communication was maintained with all participants. They were encouraged to ask questions, provide feedback, or express any concerns they might have throughout the process.

In conclusion, this research was committed to upholding the highest ethical standards in its dealings with human participants. The well-being, privacy, and fair treatment of annotators were of paramount importance throughout the study. Our adherence to these principles ensures the validity and reliability of the results while respecting the dignity and rights of every individual involved.

List of Publications

Journal Paper

Zizheng Zhang, Masato Mita and Mamoru Komachi. **Cloze Quality Estimation for Language Assessment**. Journal of Natural Language Processing, Vol. 31, No. 2. Forthcoming.

Conference Papers

Zizheng Zhang, Masato Mita and Mamoru Komachi. **ClozEx: A Task toward Generation of English Cloze Explanation**. Findings of the 2023 Conference on Empirical Methods in Natural Language Processing.

Zizheng Zhang, Masato Mita and Mamoru Komachi. **Cloze Quality Estimation for Language Assessment**. Findings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.

Other Publications

Zizheng Zhang, Tosho Hirasawa, Wei Houjing, Masahiro Kaneko, Mamoru Komachi. **Translation of New Named Entities from English to Chinese**. Proceedings of the 7th Workshop on Asian Translation.

Zizheng Zhang, Sshigemi Ishida, Shigeaki Tagashira, Akira Fukuda. **Danger-pose detection system using commodity Wi-Fi for bathroom monitoring**. Sensors 19 (4).

References

[1] Taylor WL. "Cloze procedure": A new tool for measuring readability. Journalism quarterly. 1953;30(4):415–433.

[2] Rye J. Cloze procedure and the teaching of reading. London; 1982.

[3] Alderson JC. The cloze procedure and proficiency in English as a foreign language. TESOL quarterly. 1979; p. 219–227.

[4] Raymond PM. Close procedure in the teaching of reading. TESL Canada journal. 1988; p. 91–97.

[5] Klein-Braley C. C-Tests in the context of reduced redundancy testing: An appraisal. Language testing. 1997;14(1):47–84.

[6] Xie Q, Lai G, Dai Z, Hovy E. Large-scale Cloze Test Dataset Created by Teachers. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 2344–2356. Available from: https://aclanthology.org/D18-1257.

[7] Williams JJ, Lombrozo T, Rehder B. 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.

[8] ALTE. Manual for Language Test Development and Examining: For Use with the CEFR. Language Policy division, Council of Europe; 2011. Available from: https://books.google.co.jp/books?id=Ot4ozQEACAAJ.

[9] Oller JW. Language tests at school: A pragmatic approach. (No Title). 1979;.

[10] Brown JD, et al. Cloze item difficulty. JALT journal. 1989;11(1):46–67.

[11] Brown JD, Yamashiro AD, Ogane E. Tailoring cloze: Three ways to improve cloze tests. 1999;.

[12] Pino J, Eskenazi M. Measuring Hint Level in Open Cloze Questions. In: FLAIRS Conference. Citeseer; 2009.

[13] Das B, Majumder M. Factual open cloze question generation for assessment of learner's knowledge. International Journal of Educational Technology in Higher Education. 2017;14:1–12.

[14] Felice M, Taslimipoor S, Buttery P. Constructing Open Cloze Tests Using Generation and Discrimination Capabilities of Transformers. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 1263–1273. Available from: https://aclanthology.org/2022.findings-acl.100.

[15] Matsumori S, Okuoka K, Shibata R, Inoue M, Fukuchi Y, Imai M. Mask and Cloze: Automatic Open Cloze Question Generation Using a Masked Language Model. IEEE Access. 2023;11:9835–9850.

[16] Bachman LF, Palmer AS. Language testing in practice: Designing and developing useful language tests. vol. 1. Oxford University Press; 1996.

[17] Kong X, Gangal V, Hovy E. SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 5668–5683. Available from: https://aclanthology.org/2020.acl-main.502.

[18] Felice M, Taslimipoor S, Øistein E Andersen, Buttery P. CEPOC: The Cambridge Exams Publishing Open Cloze dataset. In: Proceedings of the 2022 International Conference on Language Resources and Evaluation. European Language Resources Association; 2022.

[19] Bachman LF. Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. TESOL Quarterly. 1985;19(3):535–556.

[20] Sakaguchi K, Arase Y, Komachi M. 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). Sofia, Bulgaria: Association for Computational Linguistics; 2013. p. 238–242. Available from: https://aclanthology.org/P13-2043.

[21] Goto T, Kojiri T, Watanabe T, Iwata T, Yamada T. Automatic generation system of multiple-choice cloze questions and its evaluation. Knowledge Management & E-Learning: An International Journal. 2010;2(3):210–224.

[22] Correia R, Baptista J, Eskenazi M, Mamede N. Automatic generation of cloze question stems. In: International Conference on Computational Processing of the Portuguese Language. Springer; 2012. p. 168–178.

[23] Hill J, Simha R. 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; 2016. p. 23–30.

[24] Jiang Z, Xu FF, Araki J, Neubig G. How Can We Know What Language Models Know? Transactions of the Association for Computational Linguistics. 2020;8:423–438. doi:10.1162/tacl$_{a0}$0324.

[25] Panda S, Palma Gomez F, Flor M, Rozovskaya A. 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. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 391–401. Available from: https://aclanthology.org/2022.acl-srw.31.

[26] Skory A, Eskenazi M. 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. Los Angeles, California: Association for

Computational Linguistics; 2010. p. 49–56. Available from: https://aclanthology.org/W10-1007.

[27] Finn PJ. Word Frequency, Information Theory, and Cloze Performance: A Transfer Feature Theory of Processing in Reading. Reading Research Quarterly. 1977;13(4):508–537.

[28] Skory A, Eskenazi M. 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; 2010. p. 49–56.

[29] Finn PJ. Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. Reading Research Quarterly. 1977; p. 508–537.

[30] Hattie J, Timperley H. The power of feedback. Review of educational research. 2007;77(1):81–112.

[31] Anderson RC, Pearson PD. A schema-theoretic view of basic processes in reading comprehension. Handbook of reading research. 1984;1:255–291.

[32] Nagata R. 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). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3206–3215. Available from: https://aclanthology.org/D19-1316.

[33] Fei Y, Cui L, Yang S, Lam W, Lan Z, Shi S. Enhancing Grammatical Error Correction Systems with Explanations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 7489–7501. Available from: https://aclanthology.org/2023.acl-long.413.

[34] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners; 2022.

[35] Rankin E. Sequence strategies for teaching reading comprehension with the cloze procedure. In: Reading: Theory, Research and Practice, 26th Yearbook of the National Reading Conference. Atlanta, GA: National Reading Conference; 1976. p. 92–98.

[36] Hill F, Bordes A, Chopra S, Weston J. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In: Bengio Y, LeCun Y, editors. International Conference on Learning Representations (ICLR); 2016.Available from: http://dblp.uni-trier.de/db/conf/iclr/iclr2016.html#HillBCW15.

[37] Pustejovsky J, Stubbs A. Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc."; 2012.

[38] Liu Q, Wu Y. In: Seel NM, editor. Supervised Learning. Boston, MA: Springer US; 2012. p. 3243–3245. Available from: https://doi.org/10.1007/978-1-4419-1428-6_451.

[39] Piantadosi ST, Tily H, Gibson E. The communicative function of ambiguity in language. Cognition. 2012;122(3):280–291.

[40] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal. 2014;5(4):1093–1113.

[41] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Conference on empirical methods in natural language processing; 1996.

[42] Kupiec J, Pedersen J, Chen F. A trainable document summarizer. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval; 1995. p. 68–73.

[43] Artstein R. Inter-annotator agreement. Handbook of linguistic annotation. 2017; p. 297–313.

[44] Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37–46.

[45] Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971;76(5):378.

[46] Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project. 1990;.

[47] Saurí R, Littman J, Knippen B, Gaizauskas R, Setzer A, Pustejovsky J. TimeML annotation guidelines. Version. 2006;1(1):31.

[48] Miller GA. WordNet: An electronic lexical database. MIT press; 1998.

[49] Omelianchuk K, Atrasevych V, Chernodub A, Skurzhanskyi O. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Seattle, WA, USA â†' Online: Association for Computational Linguistics; 2020. p. 163–170. Available from: https://www.aclweb.org/anthology/2020.bea-1.16.

[50] Ng HT, Wu SM, Briscoe T, Hadiwinoto C, Susanto RH, Bryant C. The CoNLL-2014 Shared Task on Grammatical Error Correction. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 1–14. Available from: https://aclanthology.org/W14-1701.

[51] Bryant C, Felice M, Andersen ØE, Briscoe T. The BEA-2019 Shared Task on Grammatical Error Correction. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Florence, Italy: Association for Computational Linguistics; 2019. p. 52–75. Available from: https://aclanthology.org/W19-4406.

[52] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al.. RoBERTa: A Robustly Optimized BERT Pretraining Approach; 2019.

[53] Council of Europe. Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press; 2001.

[54] Baker FB. The basics of item response theory. ERIC; 2001.

[55] Sakaji H, Murono R, Sakai H, Bennett J, Izumi K. Discovery of rare causal knowledge from financial statement summaries. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE; 2017. p. 1–7.

[56] Iroju OG, Olaleke JO. A systematic review of natural language processing in healthcare. International Journal of Information Technology and Computer Science. 2015;8:44–50.

[57] Leppänen L, Munezero M, Sirén-Heikel S, Granroth-Wilding M, Toivonen H. Finding and expressing news from structured data. In: Proceedings of the 21st International Academic Mindtrek Conference; 2017. p. 174–183.

[58] Stahlberg F. Neural machine translation: A review. Journal of Artificial Intelligence Research. 2020;69:343–418.

[59] Pishdad L, Fancellu F, Zhang R, Fazly A. How coherent are neural models of coherence? In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics; 2020. p. 6126–6138. Available from: https://aclanthology.org/2020.coling-main.539.

[60] Reiter E, Dale R. Building applied natural language generation systems. Natural Language Engineering. 1997;3(1):57–87.

[61] Eddy SR. Hidden markov models. Current opinion in structural biology. 1996;6(3):361–365.

[62] Broder AZ, Glassman SC, Manasse MS, Zweig G. Syntactic clustering of the web. Computer networks and ISDN systems. 1997;29(8-13):1157–1166.

[63] Medsker LR, Jain L. Recurrent neural networks. Design and Applications. 2001;5(64-67):2.

[64] Graves A, Graves A. Long short-term memory. Supervised sequence labelling with recurrent neural networks. 2012; p. 37–45.

[65] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014;.

[66] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[67] Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. 2018;.

[68] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research. 2020;21(1):5485–5551.

[69] Cotter J. Understanding the relationship between reading fluency and reading comprehension: Fluency strategies as a focus for instruction. 2012;.

[70] Cross CB. Explanation and the Theory of Questions. Erkenntnis (1975-). 1991;34(2):237–260.

[71] Mochizuki M, Aizawa K. An affix acquisition order for EFL learners: an exploratory study. System. 2000;28(2):291–304. doi:https://doi.org/10.1016/S0346-251X(00)00013-0.

[72] Collins L. L1 differences and L2 similarities: teaching verb tenses in English. ELT Journal. 2007;61(4):295–303. doi:10.1093/elt/ccm048.

[73] Chodorow M, Gamon M, Tetreault J. The utility of article and preposition error correction systems for English language learners: Feedback and assessment. Language Testing. 2010;27(3):419–436.

[74] Srikumar V, Roth D. Modeling Semantic Relations Expressed by Prepositions. Transactions of the Association for Computational Linguistics. 2013;1:231–242. doi:10.1162/tacl$_{a0}$0223.

[75] Kato T. TOEIC L&R tesuto bunpo mondai deru 1000mon (TOEIC L&R Test Grammar Problems 1000 Questions). ASK Publishing; 2017.

[76] Gong H, Mu J, Bhat S, Viswanath P. Preposition Sense Disambiguation and Representation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural

Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 1510–1521. Available from: https://aclanthology.org/D18-1180.

[77] Zhang X, Zhao J, LeCun Y. Character-level Convolutional Networks for Text Classification. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc.; 2015.Available from: https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

[78] Hamborg F, Meuschke N, Breitinger C, Gipp B. news-please: A Generic News Crawler and Extractor. In: Proceedings of the 15th International Symposium of Information Science; 2017. p. 218–223.

[79] Fabbri A, Li I, She T, Li S, Radev D. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 1074–1084. Available from: https://aclanthology.org/P19-1102.

[80] Glisan EW, Drescher V. Textbook grammar: Does it reflect native speaker speech? The Modern Language Journal. 1993;77(1):23–33.

[81] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 7871–7880. Available from: https://aclanthology.org/2020.acl-main.703.

[82] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research. 2020;21(140):1–67.

[83] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. 2019;.

[84] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems. vol. 35. Curran Associates, Inc.; 2022. p. 22199–22213. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[85] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al.. LLaMA: Open and Efficient Foundation Language Models; 2023.

[86] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics; 2002. p. 311–318. Available from: https://aclanthology.org/P02-1040.

[87] Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, et al.. Is ChatGPT a Good NLG Evaluator? A Preliminary Study; 2023.

[88] Liu Z, Yu X, Zhang L, Wu Z, Cao C, Dai H, et al.. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4; 2023.

[89] Hessel J, Marasovic A, Hwang JD, Lee L, Da J, Zellers R, et al. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 688–714. Available from: https://aclanthology.org/2023.acl-long.41.

[90] Khurana D, Koli A, Khatter K, Singh S. Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications. 2023;82(3):3713–3744.

[91] Wang X, Wang H, Yang D. Measure and Improve Robustness in NLP Models: A Survey. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Seattle, United States: Association for Computational Linguistics; 2022. p. 4569–4586. Available from: https://aclanthology.org/2022.naacl-main.339.

[92] Gnehm AS, Bühlmann E, Clematide S. Evaluation of Transfer Learning and Domain Adaptation for Analyzing German-Speaking Job Advertisements. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2022. p. 3892–3901. Available from: https://aclanthology.org/2022.lrec-1.414.

[93] Wang X, Ruder S, Neubig G. Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. p. 863–877. Available from: https://aclanthology.org/2022.acl-long.61.

[94] Settles B, T LaFlair G, Hagiwara M. Machine learning–driven language assessment. Transactions of the Association for computational Linguistics. 2020;8:247–263.

[95] Huang J, Gu SS, Hou L, Wu Y, Wang X, Yu H, et al. Large language models can self-improve. arXiv preprint arXiv:221011610. 2022;.

[96] Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus. 2023;15(5).

[97] Kocmi T, Federmann C, Grundkiewicz R, Junczys-Dowmunt M, Matsushita H, Menezes A. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In: Proceedings of the Sixth Conference on Machine Translation. Online: Association for Computational Linguistics; 2021. p. 478–494. Available from: https://aclanthology.org/2021.wmt-1.57.

[98] Abu-Ghuwaleh M, Saffaf R. Integrating AI and NLP with Project-Based Learning in STREAM Education. 2023;.

[99] Valentini F, Rosati G, Fernandez Slezak D, Altszyler E. The Undesirable Dependence on Frequency of Gender Bias Metrics Based on Word Embeddings. In: Findings of the

Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 5086–5092. Available from: https://aclanthology.org/2022.findings-emnlp.373.

[100] Kumar S, Balachandran V, Njoo L, Anastasopoulos A, Tsvetkov Y. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics; 2023. p. 3299–3321. Available from: https://aclanthology.org/2023.eacl-main.241.