Pedagogically Aligned Objectives Create Reliable Automatic Cloze Tests

Brian Ondov, Dina Demner-Fushman
National Library of Medicine
Bethesda, MD, USA
{brian.ondov@,
demner@mail.}nih.gov

Kush Attal
NYU Grossman School of Medicine
New York, NY, USA
Kush.Attal@nyulangone.org

Abstract

The cloze training objective of Masked Language Models makes them a natural choice for generating plausible distractors for human cloze questions. However, distractors must also be both distinct and incorrect, neither of which is directly addressed by existing neural methods. Evaluation of recent models has also relied largely on automated metrics, which cannot demonstrate the reliability or validity of human comprehension tests. In this work, we first formulate the pedagogically motivated objectives of plausibility, incorrectness, and distinctiveness in terms of conditional distributions from language models. Second, we present an unsupervised, interpretable method that uses these objectives to jointly optimize sets of distractors. Third, we test the reliability and validity of the resulting cloze tests compared to other methods with human participants. We find our method has stronger correlation with teacher-created comprehension tests than the state-of-the-art neural method and is more internally consistent. Our implementation is freely available and can quickly create a multiple choice cloze test from any given passage.

1 Introduction

The cloze procedure, first introduced by Taylor (1953), is a widely used method for creating reading comprehension tests inspired by the Gestalt principle of "closure." Though many variations have been introduced and studied, the core concept is to mask words in prose and task the subject with providing the missing words. The fraction of words guessed correctly is used as a measure of comprehension of the document. A commonly used variant uses a multiple-choice response for each masked word, with the correct answer hidden among several "distractors." This simplifies scoring by removing ambiguity surrounding synonyms or misspellings. Choosing words to be distractors for each blank, however, requires either large pilots with free-text responses or experts in the domain of the text being tested. Good distractors must be incorrect, meaning they don't make sense in the blank, but still plausible enough to require comprehension of the passage to rule them out, requiring careful thought. Manual creation of distractors could also be subject to bias, especially if the test creator is testing two versions of a text and has an interest in one being seen as more comprehensible. Many have thus sought to automate distractor generation, using word co-occurence, lexical databases, and embeddings. More recently, pretrained Masked Language Models have been used, both with their original training objective and fine-tuned on reference distractors. The training objective of MLMs, which is essentially the cloze task, makes them natural choice for producing plausible words to act as distractors. However, pedagogical literature suggests that distractors should also be both incorrect and distinct (Haladyna et al., 2002; Moreno et al., 2015; Burton et al., 1990). To be incorrect, they must not make sense in the blank given the entire passage, which is not necessarily satisfied by choosing words that are slightly different from the answer. To be distinct, distractors should not overlap with each semantically, which requires optimization of *sets* of distractors, rather than ranking and choosing the top k.

Prior work largely assesses the quality of distractors using information retrieval metrics vs. reference distractors, and qualitative human judgments of each distractor. There are several issues with this approach. First, these metrics do not capture interactions between distractors, such as semantic overlap (which hurts distinctiveness). Second, there are only a few reference distractors for each blank, while many other possible good distractors (perhaps even better) exist. Third, these metrics rely on reference distractors for blank positions that were chosen by teachers and do not characterize how the methods would perform on arbitrary

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3961–3972

blanks, which would be necessary for creating new tests.

To address these issues, we first define the pedagogically motivated objectives of *plausibility*, *incorrectness* and *distinctiveness* in terms of embeddings and conditional probabilities from by MLMs. We then optimize for these objectives using simulating annealing. This makes our method both unsupervised (reference distractors are not required) and interpretable (the balance of the three objectives for a chosen set of distractors is known). We call the resulting method nCloze, for "neural cloze."

To address shortcomings with assessment, we measure how well automatically generated cloze tests from various methods actually perform their intended function, by giving tests to human readers and measuring their validity and reliability as psychometric instruments. As a reference instrument for measuring validity, we use a set of 18 teacher-created middle- and high-school-level reading comprehension passages from the CLOTH dataset (Xie et al., 2018). We test two recently reported neural distractor methods (Chiang et al., 2022; Wang et al., 2023), both with mechanical deletion, and nCloze with two deletion strategies. The best-performing version of nCloze improves correlation to the reference instrument by 17% and internal consistency by 18% vs. the next best existing method. To test domain applicability, we perform another experiment using the Newest Vital Sign (NVS) (Weiss et al., 2005) as a reference instrument for measuring health literacy. We find that nCloze tests on health-related text passages strongly correlate with performance NVS.

Our contributions are: (1) We define pedagogically motivated objectives, based on conditional distributions from masked language models, for ranking sets of distractors. (2) We present nCloze, an unsupervised, interpretable method for generating cloze tests based on these objectives. (3) We experimentally demonstrate the validity and reliability of nCloze tests, and neurally generated cloze tests generally, with human participants. (4) We provide open-source implementations of both our new method and the previously closed-source SotA method.

2 Background

Since the cloze procedure was first introduced, many variations have been proposed (Bickley et al., 1970). Two widely used variations are (1) "rational deletion" (as opposed to "mechanical deletion"), in which the words to blank are chosen based on their importance to the passage, and (2) multiple-choice (rather than free-text) responses, for which "distractors," or incorrect answer choices, must be chosen (Jonz, 1976). There has been interest in automating both of these tasks, with methods evolving along with Natural Language Processing techniques.

Early methods to choose distractors (Brown et al., 2005; Mitkov et al., 2006), and some more recent (Sun and Wang, 2023), relied on WordNet (Miller, 1995) to find words having the same part of speech as the answer and semantic similarity. For the goal of testing language proficiency, morphological modifications and orthographic similarity have been used (Pino and Eskenazi, 2009; Goto et al., 2010). Once word embeddings were introduced, e.g. word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014), many distractor generation methods incorporated them (Guo et al., 2016; Kumar et al., 2015; Jiang and Lee, 2017; Hill and Simha, 2016; Ren and Zhu, 2021). Frequency of n-grams has also been used (Hill and Simha, 2016; Mostow and Jang, 2012).

Transformer-based pretrained language models, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and GPT (Radford et al., 2018), trained to fill blanks or predict the next word, offered new opportunities for creating cloze tests. One use of such language models in this domain has been to guide rational deletion (i.e., choosing which words to blank). Keim and Littman (2022) use language models to estimate conditional probabilities for differing contexts, hypothesizing that good blanks will have options that are likely in a local context but not a broader context. However, since they use GPT-2 (Radford et al., 2019), only previous, and not subsequent, context can be included. They also find that inclusion of further domain-specific context (namely, the Wikipedia entry for "psychology") is required; our method needs only a provided passage and a pretrained language model. Matsumori et al. (2023) use language model distributions to ensure blanked words have unique answers. Yang et al. (2021) perform additional supervised training with an MLM to identify the optimal word to blank in a passage, based on how important the word is to the overall passage. All three of these rational deletion methods assume free-text response and thus do not address distractor generation.

Another application of language models to cloze tests is scoring of potential distractors. Yeung et al. (2019), for example, use BERT to rerank pools of words selected by similarity to the answer via non-contextual embeddings. Gao et al. (2020) use conditional probabilities from a pretrained MLM as features for supervised classification of whether words could be distractors. Similarly, Andersson and Picazo-Sanchez (2023) use BERT with a restricted softmax to assess distractors, but use morphological perturbations specific to Japanese generate them. Zhang et al. (2023) use a transformer-based grammatical error correction model to predict whether a given potential distractor would require grammatical knowledge or reading comprehension too rule out.

Finally, neural language models have been used to create sets of distractors for given blanks, which is the main focus of our work. Panda et al. (2022) use transformer-based round-trip Neural Machine Translation to generate distractors appropriate for specific language pairs for second-language learners. Chiang et al. (2022) fine-tune encoder-only MLMs to gap-fill with likely distractors rather than likely words, training on the CLOTH set. The top k can be chosen as a set, but they are not optimized jointly, which is important for nonredundant distractors and thus effective multiple choice questions. Similarly, Murugan and Ramakrishnan (2022) fine-tune BERT to create distractors for agglutinative languages. Wang et al. (2023) choose distractor sets jointly by training encoder-decoder language models to produce sets of distractors. However, they find their model often produces duplicate distractors, making it impractical despite its state-of-the-art performance via automatic metrics. This underscores the need for both more advanced methods and better evaluation tools.

3 Methods

We will first describe the process of generating a set of distractors for a single blank, which can be generalized to any number of blanks in the passage. Given a passage of text $X = \{x_1, ..., x_l\}$ of length l and the index b of the word to blank, x_b, our task is to generate a set of distractors D of size n. Our method has two stages: (1) select a pool of words W that are highly ranked for both plausibility and incorrectness, based the probabilities of words appearing in the blank, then (2) choose an optimal set of distractors D from W to balance the latter objectives with distinctiveness, based on the similarity of potential distractor sets. The following sections will define the three objectives, discuss the optimization of D from W, and provide a rational deletion strategy based on the objectives to be optimized.

3.1 Plausibility and Incorrectness

Since we are interested in testing comprehension of passages by fluent English speakers (as opposed to testing the proficiency of English as a second language), we follow Hill and Simha (2016) and Keim and Littman (2022) in seeking distractors that require context from the passage to rule out, rather than ungrammatical or irrelevant distractors. We thus compare the probabilities of words appearing in the blank given either just the surrounding sentence as context or the entire passage as context. A high probability in the *sentence* context ensures plausibility, while a low probability in the *passage* context ensures incorrectness. In Table 1 it can be seen that likely words (as estimated by the MLM) given the entire example passage are related to contagious infections. Likely words given only the sentence as context, however, are related to freight. These words would be desirable as distractors, since they are both syntactically correct and semantically sensible in the local context, but can be easily ruled out by a reader who comprehends the passage and knows it is people that are arriving. High on the sentence-based list, however, is "contaminated" which would not be a good distractor because it actually makes sense in this passage. This is a case that would be difficult for rule-based systems to handle because it is not a direct synonym of the answer ("exposed") but completes passage in a subtly different but accurate way.

Formally, let x_s denote the first word of the sentence containing the blanked word x_b, and x_e denote the last word of the sentence containing x_b. We will define the plausibility $\phi(D)$ as the sum of the log probabilities of each distractor w appearing in the blank given only the sentence containing the blank as context, as estimated by the MLM:

$$\phi(D) = \sum_{w}^{D} \ln P(w|x_s, ..., x_{b-1}, x_{b+1}, ..., x_e) \quad (1)$$

We will define the incorrectness $\zeta(w)$ as the probability of the word appearing in the blank given the entire passage as context:

Sentence			Passage			Sentence ÷ Passage		
Rank	**Word**	**Log prob.**	**Rank**	**Word**	**Log prob.**	**Rank**	**Word**	**Log ratio**
1	damaged	-2.4819	1	sick	-0.9100	1	damaged	9.8501
2	loaded	-2.9512	2	ill	-1.1443	2	used	8.9756
3	used	-3.1726	3	contagious	-1.8857	3	sold	8.7024
4	**contaminated**	-3.8820	4	infected	-3.1546	4	inspected	8.0821
...	...	...	...	...	...	...	...	...
9	dangerous	-4.4017	9	**contaminated**	-6.0685	9	broken	7.6788
...	...	...	...	...	...	...	...	...
83	normal	-6.9792	83	threatened	-10.4127	83	**contaminated**	2.1865
...	...	...	...	...	...	...	...	...
125	susceptible	-9.3653	125	abandoned	-12.4955	125	vaccinated	-3.6264
126	positive	-9.6393	126	defective	-12.6861	126	ill	-4.5058
127	hospitalized	-10.1444	127	broken	-12.7448	127	infectious	-4.5791
128	immune	-10.2835	128	sold	-13.2481	128	contagious	-6.6855

Passage: Mr. Frieden had this to say: "We won't be able to check travelers for fever when they leave or when they arrive. **We won't be able, as we do presently, to take a detailed history to see if they were _____ when they arrive.** When they arrive, we wouldn't be able to impose quarantine as we now can if they have high-risk contact." **Answer:** "exposed"

Table 1: The highest and lowest ranking words from an initial pool of 128 for a blank, ordered by either the probability of appearing in the blank given just the sentence (left), the probability of appearing in the blank given the entire passage (center), or the ratio of these probabilities (right). Bottom, the passage in question, with the sentence containing the blank in bold. The rankings of one undesirable potential distractor "contaminated" illustrate why it is important to consider not only the probabilities in either context, but their ratio. This example is an excerpt from a CLOTH passage, using one of its blank locations. For the sake of illustration we omit further sentences before and after the three shown here; in experiments the entire original passages are given to the masked language model.

$$\zeta(D) = \sum_{w}^{D} \ln P(w|x_1, ..., x_{b-1}, x_{b+1}, ..., x_l) \tag{2}$$

3.2 Distictiveness

If two or more distractors in a multiple choice question are synonymous, they could signal that they are incorrect (since a reader could assume the correct answer would not have synonyms as distractors). This also reduces the effective number of distractors, thus increasing the chance of guessing correctly, which is undesirable. It would also be ideal not to have hypernymy, hyponymy, and co-hyponymy among distractors. To minimize these types of relationships among distractors, we add pairwise semantic dissimilarity to our objectives. Let $\overrightarrow{e(w)}$ be the MLM embedding of word w in the entire passage context. We use cosine similarity of embeddings to estimate semantic similarity $sim(w, v)$ of words w and v (Eq. 3), defining the dissimilarity $dis(w, v)$ as the 1 minus their similarity (Eq. 4), and the scaled dissimilarity $dis'(w, v)$ as the dissimilarity normalized across all pairs of words in W, such that these pairwise scaled dissimilarities also range from 0 to 1 (Eq. 7):

$$sim(w, v) = \frac{\overrightarrow{e(w)} \cdot \overrightarrow{e(v)}}{||\overrightarrow{e(w)}|| \cdot ||\overrightarrow{e(v)}||} \tag{3}$$

$$dis(w, v) = 1 - sim(w, v) \tag{4}$$

$$dis_{min} = \min_{w,v \in W} dis(w, v) \tag{5}$$

$$dis_{max} = \max_{w,v \in W} dis(w, v) \tag{6}$$

$$dis'(w, v) = \frac{dis(w, v) - dis_{min}}{dis_{max} - dis_{min}} \tag{7}$$

Finally, we define distinctiveness δ as the harmonic mean of the normalized dissimilarity of all the distractors and the answer x_b to each other:

$$\delta(D) = \frac{n(n-1)}{2 \sum_{w,v}^{D \cup \{x_b\}} \frac{1}{dis'(w,v)}} \tag{8}$$

The harmonic mean in Eq. 8 ensures that the contributions of dissimilarity relationships to the objective are balanced.

3.2.1 Optimization

We define our energy function $E(D, W, X)$ as a sum of plausibility (Eq. 1), incorrectness (Eq. 2) and distinctiveness (Eq. 8), with the latter two balanced in relation to the former by the hyperparameters α and β:

$$E(D, W, X) = \phi(D) + \alpha\zeta(D) + \beta\delta(D) \tag{9}$$

Given an initial pool of the top k words $W = \{w_1, ..., w_k\}$ according to plausibility and incorrectness, we wish to find the set of distractors D^* of size n that maximizes this function:

$$D^* = \underset{D \subset W, |D|=n}{\arg\max} \; E(D, W, X) \tag{10}$$

Since the objective function is non-convex and has discrete inputs (the k possible distractors), we optimize using simulated annealing, in which new points are created by randomly replacing a distractor in D with another one from the pool W. We decrease the temperature linearly from 1 to 0 over the course of 1,000 iterations. Figure 1 shows an example of a solution found by the optimization for a given passage and blank. In this case, though other animal words are the most semantically similar to the blanked word, they have been avoided in favor of types of people, due to the distinctiveness and incorrectness objectives. This both prevents the distractors from being unintentionally accurate (e.g. "chimpanzees") and creates distractors that would make more sense a reader who is using only local context cues (which here suggest human subjects) rather than comprehending the passage. Further, the distinctiveness component of the objective avoids including, for example, both "teens" and "teenagers." These both have high ratios of sentence probability to passage probability but would be redundant and thus could signal their incorrectness to a test-taker.

3.3 Rational Deletion

Not all words in a passage have the same amount of plausible, but incorrect alternatives. For example, conjunctions often do not have many syntactically correct alternatives. We thus experiment with "rational deletion," or choosing which words to blank. Specifically, we use a deletion algorithm designed to give us the best possibilities for our distractor optimization. We score each word by the first two terms of Eq. 9, which we term *contextuality*. We then greedily choose the highest-scoring positions until the desired number of blanks is reached, with the constraint that a blank cannot be within m words of a previously chosen blank, with $m = 7$ for experiments.

3.4 Implementation and Performance

For experiments, we implemented our proposed nCloze method in Python 3, using RoBERTa-large (Liu et al., 2019) within the Hugging Face framework (Wolf et al., 2019) for transformer models. Performance scales approximately linearly with k, the number of potential distractors in the pool. With $k = 32$ our method can generate about 90 multiple choice cloze questions per minute using an Nvidia Tesla P100 GPU.

4 Experiment 1

Recent work has framed distractor generation as an information retrieval problem (Chiang et al., 2022; Ren and Zhu, 2021; Murugan and Ramakrishnan, 2022). Under this paradigm, models are trained in a supervised manner and information retrieval metrics are used to evaluate them. However, ranking reference distractors highly does not necessarily translate into creating effective sets of distractors, because (1) the reference distractors may be mixed with ineffective distractors near the top of the list, (2) sets chosen from the top k may contain overlapping distractors, and (3) arbitrary blanks in unseen text may have different properties than those of the reference set. Further, as our method is unsupervised and optimizes sets of distractors, there is no clear way to evaluate it using these types of metrics. We thus instead perform human experiments to characterize validity and reliability, which are standard measurements used to evaluate psychometric instruments.

- **Validity** is whether a test measures the phenomenon it is intended to measure. In this case, the phenomenon is reading comprehension. To measure this, we look at Pearson correlation of nCloze scores with teacher-created cloze tests.

- **Reliability** is whether repeated testing gives similar results. When participants are only sampled once, internal consistency is used. Since scoring for each question (that is, each blank) is binary (either the correct answer was chosen or a distractor was chosen) we use Spearman-Brown split-half correlation to measure internal consistency.

4.1 Hyperparameters

We begin with the assumption that better hyperparameters will result in the reference distractors scoring higher by our objectives. We set hyperparameters in the order they are used in the optimization. First we set α to maximize the Mean Reciprical Rank (MRR) of the CLOTH validation set

Passage: Most animals, including snakes and fish, yawn, but it is only contagious in humans and chimps and, according to a recent study, dogs. **The researchers, from the University of London's Birbeck College, put 29 _____ in a room with a yawning man and found that 21, or 72%, also started to yawn.** They said the skill may allow the pet to build stronger bonds with their owners. **Answer:** "dogs"

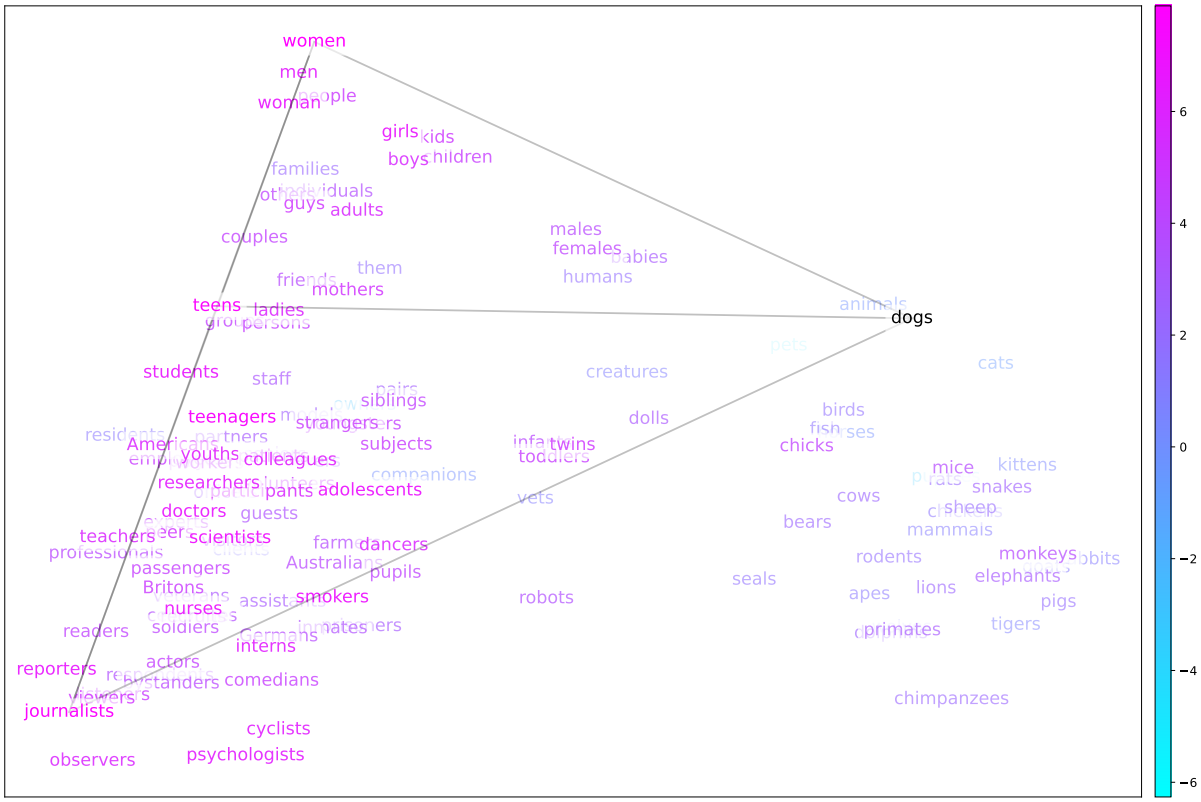

Figure 1: Interpretable, joint optimization for plausibility, incorrectness, and distinctiveness of a set of distractors. Above, a passage with a blank for which distractors are being generated, with the sentence containing the blank in bold. Below, the pool of possible distractors ($k = 128$) embedded in the RoBERTa model and projected to the 2 axes of highest variance from PCA. Color and opacity of each word corresponds to the log of the ratio of sentence probability to passage probability, which captures plausibility and incorrectness. The correct answer ("dogs") is show in black. Distractors chosen from the pool by simulated annealing ("women," "teens," "journalists") are connected by lines, which represent pairwise relationships that are included in the energy function. The distance of the chosen distractors in this projection illustrates how the distinctiveness objective avoids including pairs of high-scoring but redundant distractors (e.g. "teens" and "teenagers").

distractors, trying $\{0.1, 0.3, 1, 3, 10\}$. These values resulted in MRRs, respectively, of $\{0.0018, 0.0022, 0.0019, 0.0019, 0.0020\}$. We thus set $\alpha = 0.3$. We then optimize distractor sets for blanks in the CLOTH validation for $\beta = \{0.1, 0.3, 1, 3, 10\}$ and compute F1 scores for retrieving the reference distractors. This leads to $\beta = 0.3$ at an F1 of 0.002. Note that our method functions very differently to supervised methods that train on reference distractors, and thus we do not expect it to generate many of the same ones, hence the low scores. However, this lets us set hyperparameters in an unbiased and automatic fashion. The potential distractor pool size k mainly serves to balance efficiency and quality. We set $k = 32$ for experiments. The number of distractors per question is set to 3, as in CLOTH and the systems we will compare with (Chiang et al., 2022; Wang et al., 2023).

4.2 Data

We randomly choose 6 high school and 12 middle school passages from the CLOTH test set. As this dataset originates from printed tests that were digitized via Optical Character Recognition (OCR), it contains artifacts that are likely to be digitization errors. In order to be as faithful as possible to the teacher-created tests, the chosen passages were thus proofread by a native English speaker to correct erroneous punctuation and word splitting or merging, which are known weak points for OCR (Mei et al., 2018).

4.3 Systems

We compare two existing systems and two versions of our system. Note that systems are not given access to the choice of blanks from the CLOTH set, as these were teacher-chosen and would not be available when creating new tests. All systems except `nCloze-r` thus use mechanical deletion, in which blanks are evenly spaced throughout the text to reach the same number of blanks as the original CLOTH test. The systems we tested are:

- `CDGP`: The method of (Chiang et al., 2022). First, a pretrained BERT model is fine-tuned to predict CLOTH distractors given a masked passage. Then, candidate distractors predicted by the model are ranked according to a scoring formula that includes (1) model-predicted probability, (2) whether the distractor is the same part of speech as the answer, (3) cosine similarity of the distractor to the answer using word-level embeddings, and (4) cosine similarity of the entire sentence with the distractor vs. with the answer, by averaging word-level embeddings of each word in the sentence. Code and trained models from the paper were downloaded from `https://github.com/AndyChiangSH/CDGP/tree/main`. Mechanical deletion is used.

- `T5-multi`: The best performing method from (Wang et al., 2023), according to the majority of metrics they reported. This method was multitask training that included Distractor Finding and Cloze Test Answering in addition to text-to-text distractor generation. As code was not provided either in the paper or on request, we reimplemented the method. We consider our implementation faithful since it achieves similar, and in fact slightly better, results on the test set both for F1@3 (21.85 vs. the original 19.82) and NDCG@3 (37.89 vs. the original 36.26).

- `nCloze-r`: Our distractor generation method, with rational deletion based on finding blanks to produce high contextuality scores.

- `nCloze-m`: Our distractor generation method, with mechanical deletion.

4.4 Participants

We recruited 200 unique participants using Amazon Mechanical Turk (MTurk), allowing only participants in North America with at least 95% approval ratings on at least 5,000 prior tasks.

4.5 Procedure

Each experimental condition compares one system to a teacher-created cloze test from the CLOTH set as a control. Design is thus within-subjects vs. the control and between-subjects for experimental conditions. Ordering of the system vs. the control is randomized. Specific passage pairings for each participant were also randomized, though participants received either all middle- or high-school level tests. Since the middle school passages have roughly half the blanks as the high school ones, middle school passages were paired, so each condition had either one high school passage or two middle school passages. Both the experimental and control passages each had one question replaced by an attention check with nonsense distractors.

4.6 Results

We compute the Pearson correlation coefficient for accuracy (number of questions correct / total number of questions) of each system versus accuracy on CLOTH, with the null hypothesis that there is no correlation. We discard participants who got an attention check wrong on either the experimental or control condition or scored chance level (25%) or below on the control. As seen in Table 3, `nCloze-m`, achieves the highest correlation with the reference instrument and higher internal consistency than the previously reported state-of-the-art methods, and is much more internally consistent. At 0.6347, `nCloze-m` (our method with mechanical deletion) shows a borderline moderate/strong relationship (Akoglu, 2018). The most internally consistent system, however, was `nCloze-r` (our method with rational deletion).

4.7 Importance of Objectives

To measure the importance of each objective experimentally, we compute the Discrimination Index (DI) (Oosterhof, 2001) of a set of questions answered by participants. The DI is a commonly used measure of how well an individual question distinguishes high-scoring from low-scoring test-takers. We would expect that questions with high DIs have distractors with high values for pedagogi-

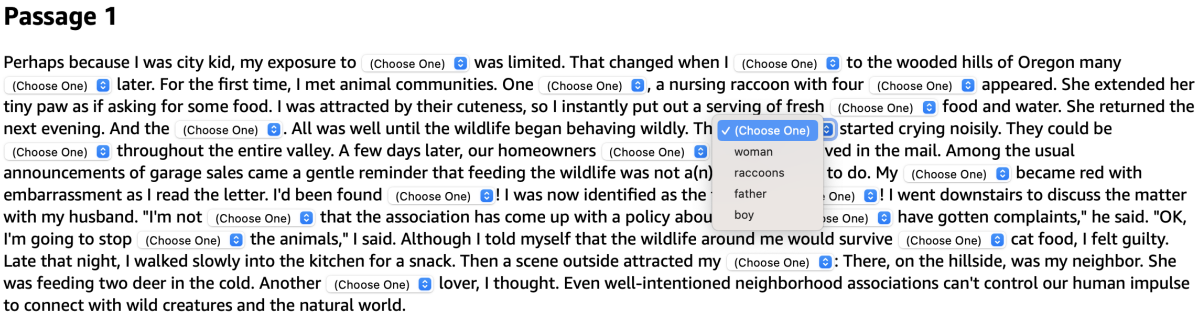

Figure 2: An example multiple-choice Cloze task within Amazon Mechanical Turk. Multiple choice options are provided in dropdowns inline with the text.

System	Pearson	Spearman-Brown
CDGP	0.5361***	0.4021*
T5-multi	0.5426**	0.5671***
nCloze-m	**0.6347*** **	0.6708***
nCloze-r	0.5228***	**0.7592*** **

Table 2: Experimental results. Pearson's correlations are with CLOTH controls, indicating validity. Spearman-Brown is split-half correlation, indicating internal consistency, and thus reliability. *p<0.05, **p<0.01,***p<0.001

cal objectives, according to our definitions. Since each question of the experimental conditions was only given to a handful of participants, we instead compute DIs for responses to the control conditions (the teach-created CLOTH questions), which were the same for all participants and thus have many more responses. We then compute each objective for each question, followed by the correlation of DI with each objective.

Incorrectness (ζ) and distinctiveness (δ) have weak, but statistically significant, relationships with DI. Interestingly, however, plausibility (ϕ) has a negative correlation with DI.

Objective	Pearson vs. DI
ϕ (plausibility)	-0.08*
ζ (incorrectness)	0.09*
δ (distinctiveness)	0.15***

Table 3: Pearson correlation of each objective with the Discrimination Index (DI), as computed on CLOTH questions. *p<0.05,***p<0.001

5 Experiment 2

As a domain experiment, we use a model trained to identify medically related text (Gupta et al., 2023) to choose 5 biomedical passages and generate corresponding nCloze distractors for each one for a total of 5 CLOTH/nCloze pairs, with an average length of 299 words per passage. We measure correlation of each with Newest Vital Sign (NVS) (Weiss et al., 2005), a commonly used health literacy test that provides a nutrition label and ask 6 questions that require reading, understanding, and reasoning about the label.

5.1 Participants

For each combination of text passage, Cloze version, and order between Cloze task and NVS task, we recruited at least 6 participants, with a total of at least 120 participants: 5 (passages) $\times$ 2 (conditions) $\times$ 2 (orderings) $\times$ 6 (participants).

5.2 Procedure

Design is again within-subjects vs. the control (NVS) and between-subjects for experimental conditions (nCloze vs. CLOTH). Ordering of the experimental condition vs. the control is randomized. As the NVS test was originally administered by interview, we implement a version similar to that of Mansfield et al. (2018), using the same distractors, and using dropdowns for multiple choice response.

5.3 Results

We compute the Pearson correlation coefficient for accuracy (number of questions correct / total number of questions) of each cloze condition versus accuracy on the NVS task, with the null hypothesis that there is no correlation. As seen in Figure 3, performance on both cloze versions significantly correlates with NVS task, though original CLOTH tests had stronger correlation (r=0.78, p=8.33e-16) than nCloze versions (r=0.69, p=1.05e-10).

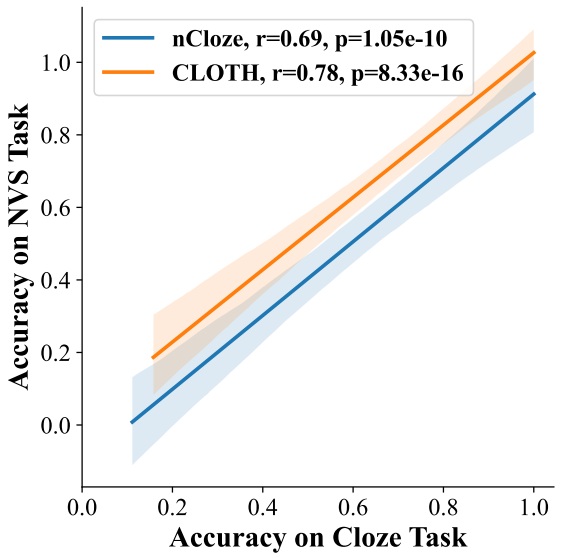

Figure 3: Validity of nCloze in the health domain, as measured by correlation with Newest Vital Sign (NVS), a health literacy test that does not use cloze-style questions. Lines represent regression lines, and shaded areas represent 95% confidence intervals, estimated with bootstrapping. Pearson correlation coefficients (r) and p-values are given in the legend.

6 Discussion and Future Work

The experiments we have run show initial evidence of validity and reliability of our method and support the idea of using pedagogical objectives. The large discrepancies between information retrieval metrics (MRR and F1 scores of our distractors from hyperparameter tuning) and testing instrument metrics (validity and reliability) show that there are many other possible distractors, and that some can in fact be better than reference distractors. However, they also show there is much work to be done in this space, despite intense interest and progress. Interestingly, rational deletion did not help our method, in fact placing it last in terms of validity, despite having the highest reliability. This is surprising considering the rational deletion method was chosen to optimize the same objectives as the distractor generation method. One possibility is that this leads to too many similar words being blanked, essentially testing similar concepts over and over. However, further exploration is warranted. Another line of investigation is constructing multi-token distractors using a beam search. This could allow common phrases or out-of-vocabulary terms, which may be especially useful in technical domains with jargon.

7 Conclusion

We have presented a distractor generation method based on pedagogical objectives that exhibits higher correlation with teacher-created distractors compared to the state-of-the-art neural method, and is more internally consistent. In addition to combining existing ideas in the field with new ideas and methods, we demonstrated the effectiveness of neurally generated distractors experimentally with human readers, which represents a large advance for the field that we hope will inspire further work. We formulate the task of distractor generation to align with the MLM pretraining objective, i.e., estimating conditional probabilities over a vocabulary given surrounding contexts. Consequently, our method requires no further training or layers, using only fast, gradient-free methods to optimize further desiderata for sets of distractors, namely plausibility, incorrectness and distinctiveness. The use of pretrained models directly also allows models to be easily swapped into our system for domain-specific applications. We implement our method in an open-source tool that allows researchers to easily generate distractors from arbitrary sources of text. Despite remaining unknowns, it is our hope that the work presented here is a step toward a reliable, automated method for creating reading comprehension tests for a wide variety of domains and applications. Our implementation is available at https://github.com/ondovb/nCloze.

Acknowledgements

We thank Dr. Elizabeth Mansfield for providing answer frequency data for use in selecting multiple choice options for NVS tests. This work was funding by the Intramural Research Program of the National Institutes of Health.

References

Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.

Tim Andersson and Pablo Picazo-Sanchez. 2023. Closing the gap: Automated distractor generation in japanese language testing. *Education Sciences*, 13(12):1203.

AC Bickley, Billie J Ellington, and Rachel T Bickley. 1970. The cloze procedure: A conspectus. *Journal of Reading Behavior*, 2(3):232–249.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826.

Steven J Burton, Richard R Sudweeks, Paul F Merrill, and Bud Wood. 1990. *How to prepare better multiple-choice test items: Guidelines for university faculty*. Ph.D. thesis, Brigham Young University. Department of Instructional Science.

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. 2022. Cdgp: Automatic cloze distractor generation based on pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5835–5840.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lingyu Gao, Kevin Gimpel, and Arnar Jensson. 2020. Distractor analysis and selection for multiple-choice cloze questions for second-language learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 102–114.

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

Qi Guo, Chinmay Kulkarni, Aniket Kittur, Jeffrey P Bigham, and Emma Brunskill. 2016. Questimator: generating knowledge assessments for arbitrary topics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3726–3732.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30.

Shu Jiang and John SY Lee. 2017. Distractor generation for chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.

Jon Jonz. 1976. Improving on the basic egg: the m-c cloze. *Language learning*, 26(2):255–265.

Greg Keim and Michael Littman. 2022. Selecting context clozes for lightweight reading compliance. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 167–172.

Girish Kumar, Rafael E Banchs, and Luis Fernando D'Haro. 2015. Revup: Automatic gap-fill question generation from educational texts. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Elizabeth D Mansfield, Rana Wahba, Doris E Gillis, Barry D Weiss, and Mary L'Abbé. 2018. Canadian adaptation of the newest vital sign©, a health literacy assessment tool. *Public health nutrition*, 21(11):2038–2045.

Shoya Matsumori, Kohei Okuoka, Ryoichi Shibata, Minami Inoue, Yosuke Fukuchi, and Michita Imai. 2023. Mask and cloze: Automatic open cloze question generation using a masked language model. *IEEE Access*, 11:9835–9850.

Jie Mei, Aminul Islam, Abidalrahman Moh'd, Yajing Wu, and Evangelos Milios. 2018. Statistical learning for ocr error correction. *Information Processing & Management*, 54(6):874–887.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ruslan Mitkov, Ha Le An, and Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. *Natural language engineering*, 12(2):177–194.

Rafael Moreno, Rafael J Martínez, and José Muñiz. 2015. Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, 27(4):388–394.

Jack Mostow and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 136–146.

Shanthi Murugan and Balasundaram Sadhu Ramakrishnan. 2022. Automatic morpheme-based distractors generation for fill-in-the-blank questions using

listwise learning-to-rank method for agglutinative language. *Engineering Science and Technology, an International Journal*, 26:100993.

Albert Oosterhof. 2001. *Classroom applications of educational measurement*. ERIC.

Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 2022. Automatic generation of distractors for fill-in-the-blank exercises with round-trip neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Juan Pino and Maxine Eskenazi. 2009. Semi-automatic generation of cloze question distractors effect of students' l1. In *International Workshop on Speech and Language Technology in Education*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Siyu Ren and Kenny Q Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4339–4347.

Yicheng Sun and Jie Wang. 2023. Constructing cloze questions generatively. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023. Distractor generation based on text2text language models with pseudo kullback-leibler divergence regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491.

Barry D Weiss, Mary Z Mays, William Martz, Kelley Merriam Castro, Darren A DeWalt, Michael P Pignone, Joy Mockbee, and Frank A Hale. 2005. Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

Albert CM Yang, Irene YL Chen, Brendan Flanagan, and Hiroaki Ogata. 2021. Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society*, 24(3):147–158.

Chak Yan Yeung, John SY Lee, and Benjamin K Tsou. 2019. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164.

Zizheng Zhang, Masato Mita, and Mamoru Komachi. 2023. Cloze quality estimation for language assessment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 540–550.

A Limitations

This work provides initial evidence that the proposed method creates reliable and valid comprehension tests. However, comprehension tests, like all psychometric constructs, require extensive characterization along both of these axes in order to be trusted and deployed. The original cloze procedure and its manually-crafted multiple-choice variant have the benefit of having been widely studied as educational tools. Further modifications to the cloze procedure thus have a high bar to clear in order to justify saving cost and effort. Additional evidence is warranted before deploying such a system in a setting with real-world consequences.

Another limitation of this work is the relatively narrow scope of the language, domain, and register. The passages we tested are all intended to test middle and high school English comprehension. They cover a wide range topics but do not require deep subject matter knowledge. It is not clear how valid the method would remain when passages contain technical language or jargon, or when they subject matter knowledge is required to comprehend the passage.

Finally, in its current form, the method does not construct out-of-vocabulary words from word pieces, which may preclude potential distractors for

technical topics. It is also not clear how important multi-token distractors would be for languages that compound more frequently than English, such as Dutch or Chinese. Further investigation of validity and reliability, and potentially further development of the subword extension algorithm, would be important for these languages and domains.

B Ethical Considerations

In this work we propose algorithms to be used to assess text with human readers. If implemented as envisioned, this could have real-world impact on either how materials are presented or how readers are rated, potentially influencing downstream decisions. We thus caution that this research is still experimental and more validation of the method is needed before widespread deployment drives decision-making.

Additionally, Masked Language Models are known to reflect biases present in their training data. This could cause some distractors to perpetuate stereotypes or make certain questions more or less difficult based on the alignment of the training corpus with the test taker's background. Further research on this method could investigate debiasing methods and analyze external factors associated with test performance.

Finally, as our validity and reliability experiments required human participants, we ensured they were ethically treated according to the Common Rule. The studies were thus designed to be non-invasive, to collect no personal information, to have no risk of harm, and not to target vulnerable populations. All participant data was stored on secured servers and participant identities were hidden from researchers throughout using anonymized identifiers. As the study activity involved common educational practices and utilized the above protections, it did not meet the standard for requiring full review by an Institutional Review Board. To ensure that we were not taking advantage of those with lesser means, we piloted tasks among colleagues, friends, and family to gauge completion times and used these estimates to adjust compensation to target above US federal minimum wage rate.