Generating Multiple Choice Questions with a Multi-Angle Question Answering Model

Andrew M. Olney
University of Memphis
aolney@memphis edu

ABSTRACT

Multi-angle question answering models have recently been proposed that promise to perform related tasks like question generation. However, performance on related tasks has not been thoroughly studied. We investigate a leading model called Macaw on the task of multiple choice question generation and evaluate its performance on three angles that systematically reduce the complexity of the task. Our results indicate that despite the promise of generalization, Macaw performs poorly on untrained angles. Even on a trained angle, Macaw fails to generate four distinct multiple-choice options on 17% of inputs. We propose augmenting multiple-choice options by paraphrasing angle input and show this increases overall success to 97.5%. A human evaluation comparing the augmented multiple-choice questions with textbook questions on the same topic reveals that Macaw questions broadly score highly but below human questions.

Keywords

multiple choice, question generation, multi-angle, paraphrase, Macaw

1. INTRODUCTION

Multiple choice questions are commonly used for assessment of learning but are expensive to produce, with significant costs associated with the development of distractor options in particular [4, 7]. As a result, researchers have attempted a variety of approaches to generating multiple-choice questions over the past two decades. A review of this literature [18] identified four core subtasks for generating multiple-choice items from text: sentence selection, answer selection from selected sentences, corresponding question generation, and distractor generation. Because training data has historically been scarce, researchers have applied related natural language processing (NLP) techniques to address these core subtasks (e.g. summarization for sentence selection), typically in a sequential pipeline architecture.

In recent years, researchers have leveraged large-scale training data and deep neural network architectures that were previously unavailable [6, 16, 22], but with a corresponding focus on distractor generation for reading comprehension questions supported by the training data. The Macaw model was very recently proposed [21] and has the potential to expand the generality of the multiple-choice question generation beyond this recent deep learning work.

Macaw expands upon the multi-task T5 model [17] by representing various question answering/generation tasks as angles rather than separate tasks as in T5. Each angle consists of slots like A (answer), Q (question), M (multiple choice options), and C (context) as well as a mapping from input to output slots. For example, $QM \rightarrow A$ represents an angle for answering a multiple-choice question, and $AC \rightarrow QM$ represents an angle for generating a multiple-choice question from an answer and context. The angle representation allows Macaw to perform data augmentation by combining multiple data sets in a common format with as much overlap as possible. For example, a QA dataset can be augmented by a multiple-choice dataset by dropping M from $QM \rightarrow A$. Likewise, angles support reverse mappings and a complementary pooling of data like $A \rightarrow Q$. By training on a large number of datasets and a large number of angles, Macaw promises more general and robust performance for a variety of question-related tasks, including zero-shot.

While the Macaw approach is intriguing, evaluations so far have focused on question answering rather than question generation [21]. Therefore it is unclear how well generation angles for multiple-choice questions perform in practice. The present study explores the suitability of Macaw for generating multiple-choice questions for academic text. Our primary research questions are (1) how does Macaw perform across three angles that systematically reduce the complexity of the task, $C \rightarrow QMA$, $AC \rightarrow QM$, and $QAC \rightarrow M$ (2) what post-processing might improve angle performance, and (3) how does the best Macaw angle's output compare to textbook questions in a human evaluation study.

2. ANGLE INPUT ABLATION STUDY

Three angles were selected for evaluation. The first of these, $C \rightarrow QMA$, is an end-to-end angle that takes a context sentence and returns an item based on it. Considering the four core subtasks, it presumes sentence selection and performs the remaining subtasks jointly. The second, $AC \rightarrow QM$, further presumes an answer has been identified and performs

the remaining subtasks jointly. Finally, $QAC \rightarrow M$ presumes all subtasks but distractor generation. These angles were chosen to determine the effectiveness of Macaw on the four core subtasks and whether simplifying the problem by providing more input improves performance. The primary outcome measure of interest is failure to generate the output slots defined by the angle. In the case of M, we additionally count the distinct options generated (defined by exact string match) and define 4 distinct options as the only success case. As each angle presumes at least one subtask result as an input slot, Macaw-external systems were used to provide these inputs. All angles were evaluated using `macaw-answer-11`, the largest Macaw model that is trained without explanations, which we cannot easily supply as an input slot.

2.1 Data

Sentences were selected from a college-level textbook on anatomy and physiology [19]. Candidate sentences were selected by a cloze item generation system which was designed to identify the most important sentences in a text for study [15]. Sentence selection was performed using coreference chains, which are sequences of nominal phrases across the text that refer to a single entity. In principle, important sentences should contain multiple such chains, and in so doing define relationships between important ideas in the text. Sentences were selected using the heuristic that they contain at least three chains of length at least two, and sentences were otherwise ranked according to the summed length of chains they contain. Previous studies have shown that this heuristic selects sentences more like a human teacher [13] than typical NLP summarization techniques and can select sentences more efficacious for study than a human expert [14]. In the present study, the top 15% of such sentences were selected, and then stratified sampling of sentences containing non-adjunct semantic arguments was used to select 5 sentences each from all 24 chapters for a total of 120 sentences.

2.2 Input Slot Sources

Two external systems were used to fill the input slots of the evaluated angles. The C slot was simply filled by one of the 120 sentences described above. AC slots were provided by the same cloze item generation system, i.e. with C provided by the same sentence, where the A slot corresponds to deleted spans of text in each sentence to make cloze items (i.e., flashcards) for study. In a departure from that previous work, which uses text spans defined by coreference chains, syntactic arguments, and semantic arguments, in the present study we only used non-adjunct semantic arguments in order to maintain parity with QA slots, as will be explained shortly. Therefore, inputs to AC slots consisted of non-adjunct semantic arguments for slot A and the same aforementioned selected sentences for slot C. An example C is "Mannitol is used in some patients to increase urinary excretion of toxins," and a corresponding A is "urinary excretion."

Inputs to QA slots were provided by a simple question generation system that was designed to provide tutorial dialogue for the larger cloze item practice system [15]. The simplicity of the system stems from its use of semantic arguments as WH targets for question generation without WH movement (e.g. *what*) , which would require careful handling of syntactic transformations. Continuing the mannitol example, the system would generate "Mannitol is used in some patients to increase what" as the input to a Q slot rather than the more sophisticated "What is mannitol used in some patients to increase?" It is important to note that the corresponding A slot input, "urinary excretion of toxins" is not identical to A slot input paired with C, "urinary excretion." This difference again stems from the simplicity of the question generation system and its design to avoid syntactic transformations. In order to maintain comparability between AC and QA inputs, questions and answers were selected such that, for all the questions generated from a sentence, we selected the question whose answer contained the answer from AC and was the shortest of such possible answers. If no such question and answer existed, one was chosen at random.

Angle	Distinct Options			
	1	2	3	4
$C \rightarrow QMA$			1	
$AC \rightarrow QM$		5	15	100
$QAC \rightarrow M$	6	16	16	80

Table 1: Distinct multiple-choice output options per angle for otherwise successful outputs.

2.3 Results

Results are displayed in Table 1 for cases where all output slots defined by the angle were output. The worst performing angle was $C \rightarrow QMA$, which failed to generate an output answer 119 times, and the one time it succeeded to generate an answer only managed to generate 3 distinct multiple-choice options. Angle $AC \rightarrow QM$ generated all outputs and successfully generated 4 distinct options 100 times (83%). Finally $QAC \rightarrow M$ failed to generate multiple-choice options twice and successfully generated 4 distinct options 80 times (67%). We note that the exact string match metric cannot distinguish paraphrases; more rigorous output quality metrics are considered in Section 4.

The two most surprising results are the failure of $C \rightarrow QMA$ and the null effect of simplifying the task, which did not lead to an increase in performance. The most likely explanation for the failure of $C \rightarrow QMA$ is that it is the only angle of the three that was not used in training, i.e. is zero-shot. Its failure casts some doubt on the generalization of Macaw to untrained angles, though more such angles would need to be tested to establish poor generalization results with confidence. As the other two angles are trained, it is counterintuitive that the simpler angle $QAC \rightarrow M$ performed so much worse than $AC \rightarrow QM$. Indeed, the $QAC \rightarrow M$ angle was trained on two more datasets (RACE and MCTest) than $AC \rightarrow QM$, so the performance difference could not be due to less training data. A possible explanation is that the form of the questions and answers input to $QAC \rightarrow M$ did not sufficiently match those used in training and that Macaw is sensitive to these differences. Further evaluation of the $QAC \rightarrow M$ angle using questions drawn from a range of textbooks with different question styles would more firmly establish this result.

3. PARAPHRASE POST-PROCESSING

As the results in Section 2.3 indicated $AC \rightarrow QM$ is the best evaluated angle for multiple-choice generation, only failing to generate distinct options in 17% of cases, we further investigated post-processing options to improve these results. One possible approach would be to use an external system to generate more options, e.g. [6, 16, 22]. However, it is challenging to generate options that are sufficiently distinct from each other and the correct answer, and an external system not designed to be conditioned on an existing set of distractors and correct answer will not be able to apply these constraints. Therefore we elected to use the simpler strategy of paraphrasing the context sentence C and leverage the inherent instability of deep neural networks to generate different outputs given slightly different inputs.

To paraphrase C, we used a T5 model created by [12] that was trained on the same textbook using sentences as input and paraphrases as output. The sentence-paraphrase training data was created using Google Translate to translate the sentences into Czech and Russian and then back into English, a process called back translation. Using this model, we repeatedly paraphrased C, used the paraphrase in $AC \rightarrow QM$, collected the M, and used exact string match to determine if we had four distinct options across all the options previously generated. Once four distinct options were found, the process terminated for that item; otherwise, the process terminated when the paraphrases were exhausted (maximum of 10 generated using top-k and top-p sampling). We considered but did not pursue more aggressive paraphrasing, such as paraphrasing the last paraphrase, due to concerns about drift from the source sentence.

The above paraphrase approach applied to the 20 cases with less than four distinct options was successful on 17 cases. Thus the $AC \rightarrow QM$ angle, with paraphrase postprocessing, was successful in generating full multiple-choice questions on 97.5% of evaluated cases.

4. HUMAN EVALUATION STUDY

A human evaluation study was conducted to compare the best performing version of Macaw, the $AC \rightarrow QM$ angle with paraphrase post-processing described in Section 3, to textbook questions on the same topic.

4.1 Method

4.1.1 Design

The evaluation study used a within-subjects design with two conditions, the best performing version of Macaw and textbook questions on the same topic. Conditions were presented in alternating order to prevent carryover effects between conditions and make fatigue effects equivalent across conditions. Evaluator judgments were analyzed using mixed-effects beta regression with random intercepts for judge and question using the `glmmTMB` R package [3]. These random intercepts for judge and question account for natural variability, e.g. a judge consistently producing higher or lower ratings than other judges. Beta regression is appropriate for continuous bounded outcome variables, unlike linear regression, which isn't suitable for bounded outcomes, and logistic regression, which can be used for proportions, but only when the proportion is a ratio of two counts [9]. Because beta regression is defined on the open interval (0,1), we use a standard transformation to squeeze our closed interval outcome variables to the open interval [20]. We conducted statistical tests at $\alpha = .05$ to address our research questions.

4.1.2 Participants

Raters ($N = 5$) were recruited through the Amazon Mechanical Turk (AMT) marketplace in June of 2022 using the CloudResearch platform [10]. Raters were required to be native English speakers, or have learned English before the age of 7, reside in the U.S., Canada, New Zealand, United Kingdom, or Australia, have completed at least an Associate Degree, and be employed as a nurse or physician. The educational and occupational constraints we designed to ensure raters were experts in the evaluation subject domain: they had passed anatomy and physiology in their studies and used this knowledge on a daily basis. Demographic constraints are enforced by CloudResearch based on rater responses to previous demographic surveys. Raters were further required to have completed at least 100 previous AMT tasks with at least a 95% approval rating. Raters were paid $12 regardless of reliability, based on an estimated 100 minutes to complete the task. In addition, raters were paid up to $50 in bonuses for passing quality checks: a $5 bonus for passing each check, and an additional $20 bonus for passing a comprehensive check.

4.1.3 Materials

The 120 questions generated using the $AC \rightarrow QM$ angle with paraphrase post-processing described in Section 3 formed the evaluation set for that condition (including the three questions with incomplete distractions options). Unfortunately, the textbook from which these questions were derived had no corresponding multiple-choice questions to use for the textbook condition. Therefore, a separate textbook on anatomy and physiology [1] from OpenStax was used as the source of textbook questions. The questions were web scraped from the OpenStax website[1] and manually checked an aligned with the answer key accessible by registering as an instructor. The 120 questions for this condition were obtained by sampling the first 4-5 questions from each of the textbook's 28 chapters, such that both conditions were approximately matched for topics.

Measure	Scale
The question contains correct information	0-100
The question is grammatical and fluent	0-100
The given correct answer is correct	0-100
The given correct answer is present in the answer options	0-100
Number of answer options that give a correct answer	0-4
Number of answer options that are distinct (no duplicates)	1-4
Quality of the question, given answer, and answer options combined	0-100

Table 2: Ratings used in human evaluation study

[1] `https://openstax.org/details/books/anatomy-and-physiology`

Survey	Question informative		Question fluent		Answer correct		Answer in options		Correct options		Distinct options		Combined quality	
	α	n	α	n	α	n	α	n	α	n	α	n	α	n
1	...	1	.78	2	.87	3	.94	2	...	1	1.00	3	.94	2
2	.57	2	.94	3	...	1	...	0	...	0	...	1	.96	3
3	.94	3	.98	3	.88	3	.94	3	...	0	.98	3	.96	3

Table 3: Inter-rater reliability per survey for included raters.

Three surveys were created with Qualtrics, an online survey tool, using 40 questions from each condition in alternating order. Each question, correct answer, and answer options were formatted vertically in that order on a single survey page using the direct assessment methodology [8, 5]. These three elements each had two associated ratings, followed by an overall quality rating, for a total of seven ratings per question, as shown in Table 2. All ratings were in horizontal slider format and arranged in descending order. The 0-100 sliders had no numeric indicators and were initialized at the midpoint. The remaining sliders had numeric indicators and snapped to integer positions. Each survey had instructions at the beginning to explain the task and the seven ratings.

Following the direct assessment methodology, degraded items were created to evaluate the internal reliability of each rater [5, 8, 2]. Degraded items were created by copying the question, answer, and options on an existing survey page and then applying the following transformations. Questions were degraded by deleting a span of words [8], where the length of the span was determined by the equation $span_{length} = 0.21696 * word_{count} + 0.78698$ [11]. Degraded answers were created by replacing the answer with one of the other answer options selected at random. Degraded answer options were created by randomly selecting an answer option and then duplicating it while removing another option at random. Thus each survey of 100 pages contained 80 distinct pages and 20 degraded versions of distinct pages.

We refer to a distinct page and its degraded version as a control pair. A sample size of 20 control pairs is sufficient to detect a large (.8 SD) effect using a Wilcoxon signed-ranks test for matched pairs at $\alpha = .05$ and .95 power on a one-tailed test. Thus if we do not detect a large effect between ratings of distinct pages and their degraded versions, we infer the rater is not reliable. The degraded pages were in randomly assigned positions in each survey and were evenly distanced from their matched distinct pages, modulo 50. This ensured that pages in control pairs had 50 other items between them, making it less likely that raters would remember their rating on a previous item. Because of the complexity of the survey design and survey length, a Qualtrics export file was reverse engineered and the survey items were programmatically generated and imported into Qualtrics.

4.1.4 Procedure

The three surveys were released in two waves. The first wave included a single survey which served as a pilot to ensure that intra-rater reliability using the control pairs was achievable. The other two surveys were released in the second wave. Each survey was terminated once it has been completed by three participants, based on the finding of generally high reliability in the first wave. Raters were allowed to participate in more than one survey if they passed the comprehensive quality check.

In all waves, raters accessed the surveys through AMT and completed the surveys using Qualtrics. Because the study is a system evaluation and not human subjects research, informed consent was not obtained. Raters saw the instructions for the survey twice, once as a preview on AMT before undertaking the survey, and again once they clicked on the survey link. On each following page, raters read the question, the correct answer, and the answer options, and then completed the ratings described in Table 2. Raters were paid upon completion of the survey and received bonuses based on the quality checks passed, i.e. based on their intra-rater reliability for each rating, with the final rating in Table 2 serving as the comprehensive check.

4.2 Results and Discussion

Median completion time across surveys was 178 minutes, giving approximately 107 seconds to read the question, answer, and options and make 7 judgments. Control checks were considered to be passed if $p < .05$ on the aforementioned Wilcoxon signed-ranks test. While one rater failed to pass the comprehensive check, all raters passed 4-6 checks out of seven. Notably one rating, "Number of answer options that give a correct answer," was failed by all but one rater. Low intra-rater reliability on this item may be explained by the randomness of the degradation strategy, which does not guarantee the removal of the given correct answer and would only remove the correct answer with 25% probability.

Cronbach's alpha was calculated for raters passing control checks in each survey. Three alphas for ratings on surveys 1 and 2 were negative, suggesting that some raters, while internally consistent, were performing the task in a different way from other raters. Since there was a common rater on these items, that rater was dropped for these items only. After recalculating alpha, one of the aforementioned alphas was still negative, so ratings from another rater in common were dropped for those items. The removal of ratings for low intra-rater reliability and the above removals for negative inter-rater reliability meant that alpha could not be calculated for all ratings, because a single rater or no rater was used on a particular combination of survey and scale. The final intra-rater reliabilities are shown in Table 3 in the same order as Table 2 but using abbreviated labels. Final alphas were otherwise high overall ($\alpha > .75$) with the exception of "The question contains correct information" on

Rating	Macaw		OpenStax		p
	M	SD	M	SD	
Question informative	89.34	22.03	96.59	7.62	.020
Question fluent	94.63	14.65	97.45	9.07	.039
Answer correct	81.30	36.11	93.04	22.50	.004
Answer in options	89.52	26.76	95.10	16.78	.264
Correct options	3.30	1.54	3.70	1.07	.169
Distinct options	3.86	.50	3.97	.27	.017
Combined quality	85.26	24.89	94.36	14.81	.000

Table 4: Results of mixed-effects beta regressions comparing the best performing Macaw version and OpenStax on each rating category.

survey 2, which was moderate, $\alpha = .57$. Ratings shown in Table 3 were used in all further analyses, including those with $n = 1$.

To answer our research question of how our best version of Macaw compares to textbook questions on the same topic, we ran separate mixed-effects beta regressions with random intercepts for rater and question, using the source of the question as the fixed effect (Macaw or OpenStax). The combined regression results are shown in Table 4.

Significant differences in favor of OpenStax were found on all but two measures, `Answer in options` and `Correct options`. However, these metrics should be discounted. `Answer in options` is relatively easy to meet, as it only requires Macaw to copy the given correct answer to one of the options. As previously discussed, the control check for `Correct options` was failed by all but one rater, and the mean for OpenStax, 3.70, indicates that rater did not understand the task, as we can be relatively certain there is only one correct answer in the options in a published textbook.

The largest differences between condition means are for `Question informative` and `Answer correct`, and together these explain the difference in `Combined quality`. Since Macaw used the $AC \rightarrow QM$ angle and the answer A was copied verbatim onto each survey page, one possible explanation for these differences in ratings is not that the answer is incorrect but rather there is a mismatch between the question Macaw generated and the provided answer. Examining low-scoring items to better understand these differences, however, suggests they are very subtle for a non-domain expert. For example, the question in Figure 1 received an average score of 50 on `Question informative` but seems quite reasonable.

Thus another possibility is that the raters are responding to other aspects of the questions, perhaps stylistic differences of even differences in difficulty. A possible avenue for future work would be to conduct a think-aloud study where raters explained the reasoning behind their ratings. Additionally, question pairs could be created by manually converting OpenStax questions into sentences as C and then using $AC \rightarrow QM$ to generate matched Macaw questions. This would reduce variability in question difficulty and create additional opportunities to study the differences between textbook questions and questions generated by Macaw.

Answer	Using partial pressures
Context	Using partial pressures simplifies the calculation of the concentration of a dissolved gas: the partial pressure of a gas dissolved in a liquid equals the partial pressure of that gas in the air with which the liquid has reached equilibrium.
Question	Which of the following simplifies the calculation of the concentration of a dissolved gas?
Multiple choice options	(A) Using moles; (B) Using atoms; (C) Using atomic masses; (D) Using partial pressures

Figure 1: Item generated by Macaw angle $AC \rightarrow QM$ that received an average score of 50 on `Question informative`

5. CONCLUSIONS

We have examined the performance of the recently proposed Macaw model for generating multiple-choice questions. Our main findings were that systematically reducing the complexity of the task by providing more input slots to Macaw angles, i.e. $C \rightarrow QMA$, $AC \rightarrow QM$, and $QAC \rightarrow M$, did not improve performance, and that $AC \rightarrow QM$ has the best evaluated performance, successfully generating complete questions in 83% of cases. We proposed paraphrase post-processing that provided missing answer options to the majority of incomplete cases, raising the success rate to 97.5%. A human subjects evaluation of this enhanced Macaw model revealed that natural textbook questions on the same topic were generally rated more highly than Macaw questions, though Macaw questions scored fairly high in absolute terms. Because our approach can be applied to any textbook, this work has potentially broad implications for scaling up multiple-choice question generation. Possible applications include using the generated questions as-is or creating draft questions for manual review and correction by a domain expert, potentially reducing authoring effort.

Our study has several limitations. First, we were unable to calculate inter-rater reliability for all combinations of ratings and surveys. Thus it is possible that some ratings used in analysis were not reliable. Second, our evaluation was conducted using a single topic, anatomy and physiology, and these results may not generalize well to other domains. Finally, we sampled the first 4-5 multiple choice questions from each OpenStax chapter, and these questions may have been easier and therefore more highly rated.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the Institute of Education Sciences under Grant R305A190448 and by the National Science Foundation under Grants 1918751 and 1934745.

7. REFERENCES

[1] J. G. Betts, P. Desaix, E. Johnson, J. E. Johnson, O. Korol, D. Kruse, B. Poe, J. A. Wise, M. Womble, and K. A. Young. *Anatomy and Physiology.* OpenStax, 2017.

[2] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, P. Koehn, and C. Monz. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics.

[3] M. E. Brooks, K. Kristensen, K. J. Van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Machler, and B. M. Bolker. glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400, 2017.

[4] S. M. Downing. Selected-response item formats in test development. In T. M. Haladyna and S. M. Downing, editors, *Handbook of Test Development*, pages 287–301. Routledge, New Jersey, 2006.

[5] C. Federmann, O. Elachqar, and C. Quirk. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-Generated Text*, pages 17–26, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[6] Y. Gao, L. Bing, P. Li, I. King, and M. R. Lyu. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, 2019.

[7] M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6):1082–1116, 2017.

[8] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics.

[9] R. Kieschnick and B. D. McCullough. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling*, 3(3):193–213, 2003.

[10] L. Litman, J. Robinson, and T. Abberbock. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2):433–442, 2017.

[11] A. M. Olney. Generating response-specific elaborated feedback using long-form neural question answering. In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, L@S '21, page 27–36, New York, NY, USA, 2021. Association for Computing Machinery.

[12] A. M. Olney. Paraphrasing academic text: A study of back-translating anatomy and physiology with transformers. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, editors, *Proceedings of the 22nd International Conference on Artificial Intelligence in Education*, pages 279–284, Cham, 2021. Springer International Publishing.

[13] A. M. Olney. Sentence selection for cloze item creation: A standardized task and preliminary results. In T. W. Price and S. San Pedro, editors, *Joint Proceedings of the Workshops at the 14th International Conference on Educational Data Mining*, volume 3051 of *CEUR Workshop Proceedings*, pages LDI–6. CEUR-WS.org, 2021.

[14] A. M. Olney, P. J. Pavlik Jr., and J. K. Maass. Improving reading comprehension with automatically generated cloze item practice. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, editors, *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 262–273. Springer, 2017.

[15] P. I. Pavlik Jr., A. M. Olney, A. Banker, L. Eglington, and J. Yarbro. The mobile fact and concept textbook system (MoFaCTS). In S. Sosnovsky, P. Brusilovsky, R. Baraniuk, and A. Lan, editors, *Proceedings of the Second International Workshop on Intelligent Textbooks 2020 co-located with 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, pages 35–49, 2020.

[16] Z. Qiu, X. Wu, and W. Fan. Automatic distractor generation for multiple choice questions in standard tests. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2096–2106, Barcelona, Spain, Dec. 2020. International Committee on Computational Linguistics.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[18] D. C. Rao and S. K. Saha. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25, 2020.

[19] D. Shier, J. Butler, and R. Lewis. *Hole's Human Anatomy & Physiology*. McGraw-Hill Education, 15th edition, 2019.

[20] M. Smithson and J. Verkuilen. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, 2006.

[21] O. Tafjord and P. Clark. General-purpose question-answering with Macaw, 2021.

[22] X. Zhou, S. Luo, and Y. Wu. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 9725–9732. AAAI Press, 2020.