Sentence Selection for Cloze Item Creation: A Standardized Task and Preliminary Results

Andrew M. Olney
University of Memphis
365 Innovation Drive, Suite 303
Memphis, Tennessee 38152
aolney@memphis.edu

ABSTRACT

Cloze items are commonly used for both assessing learning and as a learning activity. This paper investigates the selection of sentences for cloze item creation by comparing methods ranging from simple heuristics to deep learning summarization models. An evaluation using human-generated cloze items from three different science texts indicates that simple heuristics substantially outperform summarization models, including state-of-the-art deep learning models. These results suggest that sentence selection for cloze item generation should be considered a distinct task from summarization and that continued advances on this task will require large datasets of human-generated cloze items.

Keywords

cloze item, assessment, learning, extractive summarization

1. INTRODUCTION

Cloze items, also known as fill-in-the-blank questions, are common in educational practice, with applications both for assessing learning and for promoting learning [16]. Because cloze items may be created directly from text simply by deleting a word or phrase, automated methods for creating cloze items have been considered since their inception. Indeed, the work widely viewed as introducing the cloze item also proposed creating them by randomly deleting words or deleting every n^{th} word [24], and these methods became a common practice in the following decades [2]. For learning applications, however, such text-insensitive automated methods offer no control over content, and for assessment applications, research suggests that text-insensitive methods are better aligned with local properties of the text (e.g. grammar and vocabulary) than with non-local properties associated with text comprehension [2, 3, 4].

Advances in natural language processing (NLP) since 1990 have enabled text-sensitive approaches to cloze item creation for both learning and assessment applications. Research in this area has broadly organized around two different goals, creating cloze items for language learning (native or foreign language) and for text comprehension (i.e., learning from text). These two goals have led to different approaches for creating text-sensitive cloze items. Research on cloze items for language learning tends to be keyword-first [5, 8, 9, 22], meaning that sentences in the text are selected for cloze items depending on the presence of relevant keywords. These keywords are then deleted to make cloze items. Similar to text-insensitive methods, a keyword-first approach emphasizes local properties of the text and so aligns with common language-learning concerns like grammar and vocabulary, while allowing for more control over content. In contrast, research on cloze items for text comprehension tends to be sentence-first [1, 15, 19], meaning that important sentences in the text are selected first, followed by procedures for deleting words to make cloze items. A common approach to selecting important sentences for cloze items is to use extractive summarization techniques [1, 15]. Extractive summarization systems attempt to create a coherent summary of a text by filtering out unimportant sentences in a text (conversely selecting important sentences) [18] and so intuitively appear relevant for this task. Because sentence-first approaches focus on the non-local properties of the text, they are aligned with text comprehension concerns.

Research on automated cloze item creation has predominantly been theory-driven rather than data-driven, likely because large datasets of human-created cloze items have not been available until recently and only then for language-learning goals [26]. Given the absence of data with which to train and evaluate models, researchers have used rule-based and statistical techniques that are fundamentally heuristic, and they have evaluated their systems largely using rubric-based human evaluation of the cloze items created, rather than by comparing them to human-generated cloze items. One notable exception is Olney et al. [19], who compare their method with human-generated items and randomly generated items on learning outcomes. However, that work does not present a detailed comparison of automatic- and human-generated cloze items.

Research on automated cloze item creation could benefit from adopting common practices in other areas of NLP, such as common datasets, standard evaluation metrics, and the comparisons these allow with previous work. To this end, the present paper proposes sentence selection as a standardized task associated with cloze item creation. The sentence selection task is ideal for standard evaluation metrics because automated selections can be directly compared to human selections. The remainder of this paper compares multiple existing methods and their performance on the sentence selection task, including Olney et al. [19], a recent updated version of that model [20] with several variants, and three extractive summarizers.

2. SENTENCE SELECTION MODELS

2.1 Olney et al. (2017)

Olney et al. [19] used a coreference resolution system [12] for selecting sentences. A coreference chain is a sequence of repeated mentions of the same entity across a text. A common example of a coreference chain is between a noun and corresponding pronouns (e.g., "Jill" and "her"), but mentions can be less obviously connected (e.g., "Queen of England" and "Elizabeth"). Intuitively, a long chain represents an entity that is important to the discourse, and a sentence containing multiple such chains is important because it involves multiple such entities. Olney et al. operationalized this intuition with the heuristic that important sentences should contain at least three coreference chains (i.e., should contain mentions in these chains) and that the chains themselves should have a length of at least two mentions. These sentences were then filtered using criteria from a discourse parser [23], specifically nuclearity of elementary discourse units [11]. Under the theory implemented by the parser, clauses that carry little or no meaning are called satellites and are contrasted with nuclei that carry substantial meaning. Thus, selected sentences were deselected if they consisted of only satellite discourse units. This two-step heuristic was developed by inspecting a single text on the circulatory system and selecting criteria such that the number of selected sentences exactly matched the number of human-selected cloze sentences; the sentences themselves were not observed in the development of the heuristic. In later unpublished work, the above method was extended by ranking the sentences on the above criteria as well as the summed length of all coreference chains in a sentence. This extension makes it straightforward to return the top n sentences that meet the original two-step heuristic criteria while also relaxing these criteria when more sentences are requested.

2.2 Pavlik et al. (2020)

Pavlik et al. [20] describe a reimplementation of Olney et al. [19]. The reimplementation differs in several respects, including using a new coreference system based on deep learning [7] and doing away with the discourse parser constraint of nuclearity. It preserves the first step of the heuristic, prioritizing sentences having at least three coreferences chains of at least length two, and similarly ranks sentences using that criteria as well as the summed length of all coreference chains in a sentence. No comparison with Olney et al. [19] was reported.

2.3 MEAD summarizer

The MEAD summarizer [21] is a widely-used, publicly available summarizer applicable to multiple documents and multiple languages. Although MEAD has an orientation to extractive summarization of multiple documents on the same topic (e.g., a news story), it can also be used to summarize a single document. MEAD uses a variety of features to select sentences for summarization, including sentence length, position in the document, cosine with other sentences, keyword match, and LexPageRank, a measure of sentence centrality with respect to words in the document. By default, MEAD uses a linear combination of these features to identify important sentences and can be used to return the specified top n such sentences, skipping sentences that are too similar to already included sentences.

Table 1: Text characteristics

Text	FK Grade	Words	Sents	Selected
Circulatory	6.2	987	73	21
Nitrogen cycle	8.2	976	94	26
Photosynthesis	8.2	977	75	24

2.4 SMRZR summarizer

The SMRZR summarizer focuses on summarizing lectures using deep learning, is open source, and is freely available at `https://smrzr.io/` [13]. The summarizer uses BERT [6] to project each sentence in the document to an $sxwxe$ matrix, where s is the number of requested summary sentences, w is the words, and e is the embedding dimension. This matrix is then reduced to an sxe matrix by averaging over words, and each of the s sentence vectors in this reduced matrix is submitted to K-means clustering using $k = n$, the number of requested sentences. The sentences returned by the summarizer are those closest to the centroid of each of the clusters. SMRZR was not trained on a corpus but rather used a pre-trained BERT model. The layer from which the $sxwxe$ matrix is extracted was manually selected based on experimentation with a small set of test cases.

2.5 BERTSumExt summarizer

The BERTSumExt summarizer is a document-level BERT encoder that stacks inter-sentence Transformer [25] layers on top of BERT and is open source and freely available [10]. In this BERT variant, input sentences are separated by `[cls]` tokens to learn sentence representations encoded in corresponding token vectors at the output layer. These sentence representation vectors are then input to inter-sentence Transformer layers with position embeddings to capture sentence position, and these lead to a sigmoid classifier output layer that indicates the importance of the sentence. The top n such sentences can be returned to create an extractive summary. Unlike SMRZR and MEAD, BERTSumExt is directly trained on news corpora. BERTSumExt was state of the art on extractive summarization for the CNN/Daily Mail dataset [14] and was only recently surpassed by a system with less than a 1 point improvement in recall [27].

3. EVALUATION

3.1 Procedure

Evaluation data were obtained by asking expert judges to create cloze items for three texts on science topics, including the circulatory system, the nitrogen cycle, and photosynthesis. The text and cloze items for the circulatory system were taken from Olney et al. [19]. The other texts were created by a graduate student blind to the purpose of the study to match the length and difficulty of the circulatory system text. As shown in Table 1, texts matched closely in number of words but somewhat less so in terms of difficulty, with both nitrogen cycle and photosynthesis texts being approximately two Flesch-Kincade grades level units higher in difficulty than the circulatory system text.

Cloze items for the circulatory text were created by a graduate student who operationalized the task as selecting sen-

Table 2: Recall of Sentence Selection

Model	Circ. Sys.	Nit. Cyc.	Photosyn.	M
Olney et al.	**.57**	.19	.33	.37
Pavlik et al.	**.57**	.35	**.46**	**.46**
MEAD	.29	**.42**	.33	.35
SMRZR	.33	.19	.38	.30
BERTSumExt	.10	.27	.38	.25
Random	.29	.28	.32	.29
Two chains	.48	.27	.38	.37
# chains	.52	.27	.38	.39
No restriction	.29	.35	.42	.35

tences conveying the main ideas. Cloze items for the other two texts were created by a high school biology teacher who was blind to the purpose of the study. Both human judges selected similar numbers of sentences across texts.

Each of the three texts was input into the models described in Section 2 along with the parameter n, the number of sentences selected by a human judge for that text. The primary evaluation metric was the number of sentences returned that were selected by human judges (i.e. overlap), divided by n. This metric is equivalent to recall for extractive summarization, which some have argued is more appropriate than precision given the variability in human sentence selection [17].

Additionally, we evaluated several variants of the Pavlik et al. model that varied according to the primary heuristic of having at least three coreferences chains of at least length two. The variants included having at least two coreferences chains of at least length two, replacing this restriction by ranking by the total number of chains in the sentence, and removing this restriction entirely. Each variant ranks the sentences, post-constraint, by the summed length of all coreference chains in a sentence, just as the original.

3.2 Results

Results are presented in Table 2, which shows the best model recall score per text in bold font, with the final column showing the average recall across texts. The initial rows of Table 2 correspond to the models in Section 2, followed by a random baseline (i.e., random selection of n sentences), followed by the variants of the Pavlik et al. model.

The best performing model is Pavlik et al. [20], which has the best average score as well as the top score (or tied) for every text with the exception of the nitrogen cycle, for which MEAD achieves the highest score. The increased performance of Pavlik et al. model relative to the original Olney et al. [19] suggests that the discourse parser constraint of nuclearity is not contributing heavily to performance and that these contributions are easily overwhelmed by using a higher-performing coreference resolution system. However, it is notable that although the systems achieve the same score on the circulatory system, they do not make identical predictions: 25% of the correct predictions differ between the two models.

It is remarkable both how badly the summarization models perform on this task as well as how their performance seems to improve as their simplicity increases. The most sophisticated model, BERTSumExt, which is near state of the art on extractive summarization, performs below chance on 2/3 of the texts as well as below chance on average. SMRZR, another deep learning model, is similarly below chance on 1/3 of the texts and only 1% above chance on average. MEAD, the simplest and oldest model, is approximately at chance on 2/3 texts, though its average score is elevated by its top performance on the nitrogen cycle text. Overall, these results suggest that the intuition that summarization models are suitable for the sentence selection task of cloze item creation is incorrect. Indeed it appears that models trained on newswire text, like BERTSumExt, may be particularly poorly suited for this task.

Finally, the variant results indicate that the current heuristics used by Pavlik et al. are not overfitted to the original circulatory system text. No variant achieves a higher score on any single text or overall. However, the variant results suggest that heuristics involving the number of chains in a sentence are particularly significant for improving the score of the circulatory system text.

4. DISCUSSION

We have proposed sentence selection as a standardized task associated with automated cloze item creation. Unlike previous work that has used rubrics to evaluate cloze items, sentence selection allows automated selections to be directly compared to human selections using standard evaluation metrics like recall. Because our results show that simple heuristics outperform extractive summarization models, including a state of the art deep learning model, we argue that sentence selection for cloze item generation should be considered a distinct task from extractive summarization, particularly extractive summarization in the context of newswire text, where it has historically focused. Previous researchers have raised concerns with the type of direct evaluation we propose, based in part on the variability of sentences human judges will select for extraction [17]. We believe that these concerns are more valid for newswire text as opposed to academic text, which by definition is designed for learning. While experts may not agree on what parts of a current news story are most important in a summary, we suspect that experts on photosynthesis generally agree on key ideas, and thus key sentences in a text. However, we have not presented evidence confirming this suspicion in this paper, nor are we aware of research that has investigated this question. This suggests a new direction in automated cloze item creation: the creation of large datasets of cloze items on diverse texts, where each text has been annotated by a large enough sample of human judges that we can estimate human agreement reliably enough to calculate whether an automated method agrees as much (or more) with humans as humans do with each other. Without common datasets, standard evaluation metrics, and the comparisons these allow with previous work, we fear that researchers will continue to create novel systems and evaluate them in isolation, which will ultimately contribute little to progress on automated cloze item creation.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants 1918751 and 1934745 by the Institute of Education Sciences under Grant R305A190448. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Institute of Education Sciences.

6. REFERENCES

[1] M. Agarwal and P. Mannem. Automatic gap-fill question generation from text books. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64, Portland, Oregon, June 2011. Association for Computational Linguistics.

[2] J. C. Alderson. The cloze procedure and proficiency in english as a foreign language. *TESOL Quarterly*, 13(2):219–227, 1979.

[3] L. F. Bachman. The trait structure of cloze test scores. *TESOL Quarterly*, 16(1):61–70, 1982.

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

[5] D. Coniam. From text to test, automatically - an evaluation of a computer cloze-test generator. *Hong Kong Journal of Applied Linguistics*, 3(1):41–60, 1998.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] A. Kurtasov. A system for generating cloze test items from Russian-language text. In *Proceedings of the Student Research Workshop associated with RANLP 2013*, pages 107–112, Hissar, Bulgaria, Sept. 2013.

[9] C.-L. Liu, C.-H. Wang, Z.-M. Gao, and S.-M. Huang. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 1–8, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[10] Y. Liu and M. Lapata. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[11] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[12] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[13] D. Miller. Leveraging BERT for extractive text summarization on lectures. *CoRR*, abs/1906.04165, 2019.

[14] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics, 2016.

[15] A. Narendra, M. Agarwal, and R. Shah. Automatic cloze-questions generation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 511–515, Hissar, Bulgaria, Sept. 2013.

[16] National Institute of Child Health and Human Development. *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* NIH Publication No. 00-4769. U.S. Government Printing Office, Washington, DC, 2000.

[17] A. Nenkova. Summarization evaluation for text and speech: issues and approaches. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*. ISCA, 2006.

[18] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233, 2011.

[19] 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.

[20] 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.

[21] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD - a platform for multidocument multilingual text summarization. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[22] A. Skory and M. Eskenazi. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages

49–56, Los Angeles, California, June 2010. Association for Computational Linguistics.

[23] M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escarcega. Two practical rhetorical structure theory parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, Denver, Colorado, June 2015. Association for Computational Linguistics.

[24] W. L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Proceedings of the Thirty-first Annual Conference on Neural Information Processing Systems*, pages 5998–6008, 2017.

[26] Q. Xie, G. Lai, Z. Dai, and E. Hovy. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[27] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online, July 2020. Association for Computational Linguistics.