Using large language models to develop readability formulas for educational settings

Scott Crossley[1][0000-0002-5148-0273] Joon Suh Choi[2][0000-0002-7732-0366] Yanisa Scherber[2][0000-0003-1857-0776] and Mathis Lucka[3]

[1] Vanderbilt University
[2] Georgia State University
[3] deepset

scott.crossley@vanderbilt.edu, jchoi92@gsu.edu
yscherber1@gsu.edu, mathis.lucka@deepset.ai

Abstract. Readability formulas can be used to better match readers and texts. Current state-of-the-art readability formulas rely on large language models like transformer models (e.g., BERT) that model language semantics. However, the size and runtimes make them impractical in educational settings. This study examines the effectiveness of new readability formulas developed on the CommonLit Ease of Readability (CLEAR) corpus using more efficient sentence-embedding models including doc2vec, Universal Sentence Encoder, and Sentence BERT. This study compares sentence-embedding models to traditional readability formulas, newer NLP-informed linguistic feature formulas, and newer BERT-based models. The results indicate that sentence-embedding readability formulas perform well and are practical for use in various educational settings. The study also introduces an open-source NLP website to readily assess the readability of texts along with an application programming interface (API) that can be integrated into online educational learning systems to better match texts to readers.

Keywords: Text Readability, Large Language Models, Natural Language Processing.

1 Introduction

Instructional texts aligned with students' reading levels are beneficial for gradually improving reading skills [1]. Providing students with texts that are matched to their reading abilities can ensure a stronger understanding of material and increased knowledge. Readability formulas are often used to match texts to readers. However, traditional readability formulas are based on weak proxies of language features related to text comprehension and do not measure many text features that are essential to determining text difficulties such as text cohesion and semantics. Additionally, many traditional readability formulas are biased because they were normed based on readers from specific demographics or on small text samples from specific domains.

To address these concerns, the CommonLit Readability Prize (CRP) was hosted on Kaggle (https://www.kaggle.com/c/commonlitreadabilityprize) in the summer of 2021. The CRP tasked data scientists on the Kaggle platform to build open-source readability formulas to predict the reading ease of ~5,000 reading excerpts normed for grade 3-12 classrooms. The reading ease scores scalar values unique to each individual text and were the result of teacher judgments of text difficulty using pairwise comparisons. Over 70,000 readability formulas were developed. The winning formulas were strongly predictive of text readability for the excerpts in the CommonLit Ease of Readability (CLEAR) Corpus [2][3], explaining around 90% of the variance. All winning formulas were ensemble models based on large language models (LLMs) which use contextualized token embeddings approaches to model language phenomena (e.g., question answering, sentence classification, and sentence-pair regression). However, while the winning models were highly predictive, they were not efficient. For example, the second-place model was 105 gigabytes in size and reported runtimes of 39 minutes to process seven texts. The inefficiency of the winning models makes it unlikely that they will be readily adapted by teachers, administrators, materials developers, and learning engineers regardless of their predictive power.

The goal of this study is to develop and assess the efficiency of new readability formulas derived from the CLEAR corpus that are based on contextual token embedding approaches similar to those found in the LLMs that won the CRP but are more efficient, allowing for quicker run times and less data storage. To do this, we develop three efficient readability formulas based on contextual token embedding approaches. These approaches include Doc2Vec [4], Universal Sentence Encodings (USE) [5], and SentenceBert (SBERT) [6]. We compare the performance of these readability formulas to traditional readability formulas (e.g., Flesch-Kincaid Grade Level [7], Flesch Reading Ease [8]), more advanced linguistic features formulas (The Crowdsourced Algorithm of Reading Comprehension [9], The Coh-Metrix Second Language Readability Index [10], and the model reported in Crossley et al., 2022 [3]), and the winning formulas from the CRP competition.

2 Method

2.1 Traditional Readability Formulas

We derived traditional readability scores using the Automatic Readability Tool for English (ARTE) [11]. ARTE is a freely available tool that automatically calculates a number of existing readability formulas on batches of text files. The formulas selected were not trained or normed on the CLEAR corpus. The formulas included Flesch - Kincaid Grade Level [7], Flesch Reading Ease [8], Automated Readability Index ARI), The Simple Measure of Gobbledygook (SMOG) [12], and the New Dale-Chall [13].

2.2 NLP-informed Readability Formulas

We calculated two NLP-informed readability formulas using ARTE that were not trained or normed on the CLEAR corpus: The Crowdsourced Algorithm of Reading

Comprehension (CAREC) [9] and The Coh-Metrix Second Language Readability Index (CML2RI) [10]. We also report on an NLP-informed readability formula that was developed specifically for the CLEAR corpus and was reported in Crossley, 2022 [3].

2.3 CRP Readability Formulas

We replicated the top four readability formulas reported in the CRP. All formulas were based on transformer models, which depend on contextualized token embeddings that capture semantic information about a language. Transformer models use neural networks with multiple hidden layers that include millions of parameters that interact in complex ways, making it difficult to fully explain or interpret what the model is doing at each pass [**Error! Reference source not found.**]. Transformer models take i nto consideration the order in which words appear (e.g., *peace* and *war* would be represented differently than *war* and *peace*) and have attention mechanisms which allow input weights to be based on importance in a task [15]. These self-attention mechanisms allow the models to dynamically select which words are important for its calculations from a wider context window (the full length of the input). This approach allows transformer models to distinguish differences between the uses of the word *bank* in the sentences *The man robbed the bank* and *The man sat on the river bank*.

Because of the cost associated with developing transformer models, it is common to use pre-trained language models like BERT [15] and fine-tune them for different tasks. Fine-tuning leverages knowledge about diverse language-related tasks to influence the pre-trained weights of the model by providing labeled training data that are specific to the downstream task (i.e., the final target task). In the case of the CLEAR corpus, this would be the prediction of the ease of readability scores. All model winners from the CRP competition were fine-tuned on the CLEAR corpus.

2.4 Efficient Word Embedding Readability Formulas

We derived three sentence-embedding readability formulas based on Doc2Vec, Universal Sentence Encodings, and SentenceBert (SBERT) models. The three models map sentences and paragraphs (rather than individual words) to a multi-dimensional vector space, which can be used to train and predict readability scores. These models are more efficient based on size and run-time. It was not possible to fine-tune the Doc2vec and Universal Sentence Encoding models on the CLEAR corpus. However, the SBERT model was fine-tuned on the CLEAR corpus.

To develop readability formulas for the sentence embeddings models above, we did the following:

- For the SBERT model only, fine-tune the core BERT model using the CLEAR corpus.
- Calculate the sentence embeddings for each excerpt in the training and test set.

- For each excerpt in the test set, calculate the cosine similarity of each except in the train set to determine which excerpt from the train set is the most similar (i.e., has a cosine similarity score closest to one).
- Assign the known reading ease score for the most similar training excerpt to the test excerpt.
- Repeat for each excerpt in the test set until all items have been assigned a predicted reading ease score.

2.5 Statistical Analysis

Correlation analyses were conducted to compare how the scores from the traditional readability formulas, the NLP-informed readability formulas, the winning CRP models, and the light-weight sentence-embeddings models correlated with the reading ease scores from the CLEAR corpus. Only the test set for the CLEAR corpus as defined in the CRP was used.

3 Results

3.1 Traditional and NLP-informed Readability Formulas

Correlations for traditional readability formulas reported medium to strong effects with reading ease scores ($r \sim .5$) for the test set. The strongest correlation was reported for the New Dale Chall Readability Formula while the weakest correlation was reported for ARI. The variance explained by the formulas varied from 23% to 31%. Correlations for NLP-informed models reported strong effects with reading ease scores (r between .546 to .711) for the test set. The strongest correlation was found for the Crossley et al. model (2022) that was trained specifically on the CLEAR corpus. The weakest correlation was reported for CML2RI formula. The variance explained by the formulas varied from 29% to 51% (see Table 1).

Table 1
Correlations between ease of readability score and traditional/NLP-informed readability formulas

Variable	2	3	4	5	6	7	8	9
1. BT Ease	0.540	-0.517	-0.484	-0.551	-0.556	-0.588	0.546	0.711
2. FRE	1	-0.912	-0.860	-0.936	-0.837	-0.725	0.703	0.742
3. FKGL		1	0.987	0.835	0.682	0.587	-0.691	-0.677
4. ARI			1	0.778	0.631	0.537	-0.688	-0.635
5. SMOG				1	0.808	0.708	-0.659	-0.742
6. DC					1	0.738	-0.667	-0.738
7. CAREC						1	-0.583	-0.723
8. CML2RI							1	0.791
9. Crossley 2022								1

*FRE = Flesch Reading Ease, FKGL = Flesch Kincaid Grade Level, ARI = Automated Readability Index, SMOG = Simple Measure of Gobbledygook, DC = New Dale Chall, *CAREC = The Crowdsourced Algorithm of Reading Comprehension, CML2RI = The Coh-Metrix Second Language Readability Index, Crossley 2022 = Crossley et al. (2022) Model

3.2 Transformer-Based Readability Formulas.

Correlations for the CRP models reported large effects with reading ease scores (r = .90). The strongest correlation was reported for the third-place model (r = .903) while the weakest correlation was reported for the fourth-place model (.901). The variance explained by all the models was ~.81. Correlations for the sentence embedding models reported large effects with reading ease scores. The strongest correlation was reported for SBERT (r = .84) while both Doc2Vec and the Universal Sentence Encoder reported correlations of .51. The variance explained by the SBERT model was .71.

Table 2
Correlations between ease of readability score and transformer-based readability formulas

Variable	2	3	4	5	6	7	8
1. BT Ease	0.902	0.902	0.903	0.901	0.512	0.513	0.843
2. First place	1	0.994	0.991	0.991	0.578	0.571	0.931
3. Second place		1	0.993	0.994	0.582	0.568	0.932
4. Third place			1	0.992	0.572	0.563	0.927
5. Fourth place				1	0.578	0.566	0.931
6. Doc2Vec					1	0.456	0.553
7. USE						1	0.542
8. Sbert							1

*USE = Universal Sentence Encoder Model, Sbert = SentenceBert Model

4 Discussion and Conclusion

Readability formulas are an important component for assigning appropriate texts to readers to help enhance the development of reading skills. Previous research has indicated that NLP-informed formulas, which measure more advanced features of texts than traditional readability formulas, are strong predictors of readability. However, as NLP techniques have advanced, concurrent advances in readability formulas had not, at least until the release of the CommonLit Readability Prize (CRP) competition. The BERT models from the CRP that measured semantic similarity far outperformed traditional readability formulas that are based on weak proxies of word decoding and syntactic parsing. As well, BERT-based readability formulas outperform readability formulas derived from more advanced NLP features. However, the BERT models required large amounts of storage and long run-times, making them generally impractical in most educational settings.

The more computationally efficient readability formula derived from SBert performed on par with the less efficient BERT models from the CRP competition. Lower performance was reported for the Doc2Vec and Sentence Encoder models. The SBERT model also outperformed traditional and NLP-informed readability formulas. Importantly, the SBERT model has a storage space that is 200 time smaller than the first-place CRP model and its runtime is a 100 time faster. Thus, the SBERT model's size and runtime make it feasible to scale in learning technologies that include assessments of text readability as well as for providing feedback to teachers, matching

texts to students, or assessing student products. To assist with the integration of the SBERT model, we have developed an application programming interface (API) for the Automatic Readability Tool for English (ARTE) that is freely available (please see https://nlp.gsu.edu/APIdoc) so that learning engineers can integrate the SBERT readability formula into learning technologies to better match texts to readers. In addition, we have developed ARTE into an intuitive and easy to use web tool that will automatically read in texts and provide readability results based on the SBERT formula so that researchers and teachers interested in reliably assessing text readability can do so on their devices (https://nlp.gsu.edu).

References

1. Mesmer, H.A.E.: Tools for matching readers to texts: Research-based practices. Guilford Press (2008).
2. Crossley, S.A., Heintz, A., Choi, J.S., Batchelor, J., Karimi, M., Malatinszky, A.: The CommonLit Ease of Readability (CLEAR) Corpus. In: Proceedings of the 14th International Conference on Educational Data Mining, pp. 755-760 (2021).
3. Crossley, S.A., Heintz, A., Choi, J.S., Batchelor, J., Karimi, M., Malatinszky, A.: A large-scaled corpus for assessing text readability. Behavior Research Methods, 1-17 (2022).
4. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International conference on machine learning, pp. 1188-1196 (2014).
5. Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes M., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzwell, R.: Universal Sentence Encoder. arXiv preprint arXiv:1803.11175 (2018).
6. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019).
7. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch (1975).
8. Flesch, R.: A new readability yardstick. Journal of applied psychology, 32(3), pp. 221-233 (1948).
9. Crossley, S. A., Skalicky, S., Dascalu, M.: Moving beyond classic readability formulas: New methods and new models. Journal of Research in Reading, 42(3-4), 541-561 (2019).
10. Crossley, S. A., Greenfield, J., McNamara, D.S.: Assessing text readability using cognitively based indices. Tesol Quarterly, 42(3), pp. 475-493 (2008).
11. Choi, J.S., & Crossley, S.A.: Advances in Readability Research: A New Readability Web App for English. In: Proceedings of the 22nd International Conference on Advanced Learning Technologies, pp. 1-5 (2022).
12. McLaughlin, G.H.: SMOG grading-a new readability formula. Journal of reading, 12(8), pp. 639-646 (1969).
13. Chall, J.S., Dale, E.: Readability revisited: The new Dale-Chall readability formula. Brookline Books (1995).
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp. 206-215 (2019).
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019).