Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Novel Interpretation of Classical Readability Metrics: Revisiting the Language Model Underpinning the Flesch-Kincaid Index

Yo Ehara[a*]
[a]*Faculty of Education, Tokyo Gakugei University, Japan*
*ehara@u-gakugei.ac.jp

Abstract: In the realm of natural language processing (NLP), the quantification of text readability remains crucial, with pivotal applications in education. While the Flesch-Kincaid GradLevel (FKGL) has been a foundational metric for English text readability, recent advancements, particularly with models like Bidirectional Encoder Representations from Transformers (BERT) , have heralded a new age of language model-based assessments. Contrary to popular belief about the FKGL's legacy nature, our research elucidates that FKGL encapsulates language model complexities. We introduce a novel interpretation that views FKGL as a linear blend of perplexities from specific unigram models. Leveraging the OneStopEnglish dataset, we enhanced FKGL by incorporating perplexity values from state-of-the-art language models for sentence boundaries. Our results highlight that integrating BERT's capabilities significantly bolsters FKGL's performance. The implications are vast, suggesting potential expansion to multi-lingual FKGL applications and providing theoretical backing for FKGL-based research in languages like Japanese.

Keywords: Readability Assessment, Language Models, Perplexity

1. Introduction

The quantification of text readability stands as a fundamental task in natural language processing (NLP), having pronounced importance in various application domains, notably in education. The Flesch-Kincaid GradLevel (FKGL) [1] has historically been the benchmark metric, widely employed for the automatic assessment of English text readability. In contrast, recent advancements in deep learning, spearheaded by models like Bidirectional Encoder Representations from Transformers (BERT) [2], have ushered in an era where large-scale language models are increasingly anticipated to play significant roles in automated readability evaluation. Numerous studies, as delineated in [3], have put forth methodologies predicated on language models, reportedly outperforming FKGL in precision. A common conjecture attributed to this heightened accuracy is the inherent legacy nature of FKGL, given its non-reliance on any contemporary language models.

Challenging this narrative, our study underscores that FKGL, in fact, encapsulates complexities intrinsic to language models. More specifically, we posit that the FKGL can be aptly represented as a linear combination of the perplexity derived from a specific unigram model. This association between FKGL and perplexity is, to the best of our awareness, unprecedented, and its potential ramifications deem it worthy of comprehensive documentation.

Perplexity, as expounded in [4], acts as a gauge for the complexity of test data within the purview of language models. It is mathematically defined as the inverse of the probability of a word manifesting within the test data, computed on a per-word basis. Intuitively, perplexity provides insight into the average number of choices a language model, predicated on its training, can delineate from contextually when making predictions on the test data. For

illustrative purposes, if the occurrence probability of a word is set at 1/3, it implies the model's capability to narrow its choices down to one out of a potential three words.

2. Methods

The Flesch-Kincaid GradLevel (FKGL) is traditionally defined as:

$$FKGL=0.39\times\text{Average Word Count per Sentence}+11.8\times\text{Average Syllables per Word}-15.59$$

where the average word count per sentence is calculated as the ratio of total words to the number of sentences [1].

Consider a text where sentence boundaries are marked with the token “[SS]”. If the text is partitioned such that the last “[SS]” demarcates the test data, with everything preceding it as training data, the output probability of “[SS]” for the unigram model trained on this partitioned data is given by the number of “[SS]” occurrences in the training data divided by the total word count, which equivalently is the ratio of the number of sentences to the total word count. The perplexity for the single-word test data “[SS]” on this unigram model (which is the inverse of its probability) is thus given by Word Count/Sentence Count, aligning with the average word count per sentence.

To elucidate, consider the text segment: “[SS] This is a pen. [SS] That is a cat. [SS]”. Using the final “[SS]” as test data and the preceding as training, the unigram model yields p("[SS]") = 12/2 . The perplexity of the test data (“[SS]”) is 1/p("[SS]") = 12/2 = 6, matching the average word count.

Given that perplexity serves as a quantitative measure of test data complexity using a language model, the average word count in FKGL can be interpreted as representing text difficulty using a unigram model by quantifying the complexity of sentence boundaries. This understanding opens avenues for various applications.

We propose an enhanced FKGL by integrating the perplexity values derived from state-of-the-art language models for sentence boundaries and examining the potential improvement in performance. For evaluation, we utilized the OneStopEnglish dataset comprising 567 texts, each manually rated on a three-tier readability scale tailored for English learners. (More details can be found at: https://paperswithcode.com/dataset/onestopenglish)

Rather than using “[SS]”, the BERT model was employed to mask the next word following each text's end, and the probability of the BERT “[SEP]” token appearing was determined. We engaged the pretrained model 'bert-large-cased-whole-word-masking' for this endeavor.

3. Results

To evaluate the efficacy of the proposed enhanced FKGL metric, we compared its performance against the traditional FKGL using the OneStopEnglish dataset. Specifically, the Spearman rank correlation coefficient was calculated against the dataset's manual three-point scale readability annotation labels: beginner, intermediate, advanced.

The traditional FKGL yielded a correlation coefficient of r=0.5776. In contrast, the enhanced FKGL, wherein the average word count component was supplemented with the BERT-derived sentence-boundary perplexity (scaled by a constant for value range alignment), demonstrated an improved correlation of r=0.5825. The scaling constant was computed as the reciprocal of the average perplexity across all texts.

A Wilcoxon signed-rank test was employed to test the statistical significance of the difference between the two metrics. The difference was found to be statistically significant, with a p-value less than 0.01.

4. Discussion

In this study, we demonstrate that the "Average Word Count" component of the Flesch-Kincaid Grade Level (FKGL) can be interpreted as the perplexity derived from a unigram language model. Similarly, the "Average Syllables per Word" can be understood in terms of the perplexity of word boundaries in a unigram language model designed for sequences of syllables. Consequently, the FKGL can be conceptualized as a linear combination of two distinct unigram language models: one considering words as units and the other considering syllables as units. Our findings can be immediately applied to other readability formulae that

utilize average word or syllable counts. Notably, these include formulae such as the Flesch Reading Ease (https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests) and the Automated Readability Index (https://en.wikipedia.org/wiki/Automated_readability_index).

5. Conclusions

Our experimental results suggest a noteworthy observation: while the conventional unigram model fails to capture the intricate complexities of sentence boundary determinations, integrating BERT's prowess in detecting sentence boundaries significantly enhances the performance of the FKGL metric. This improvement was determined to be statistically significant.

While this study primarily focused on the average word count in a text, there's potential to further investigate the average syllable count from a similar perspective. Specifically, the intricacies of determining boundaries within sequences of syllables can be viewed analogously to perplexities in the context of a unigram model. Thus, the FKGL can be characterized as a linear combination of the perplexities derived from unigram models.

The implications of this research are profound, particularly in lending theoretical validity to studies that employ FKGL for Japanese texts [5]. We anticipate that our findings will pave the way for more extensive research on the multi-lingual adaptation of FKGL and its tailored applications for English learners.

Acknowledgements

This study was supported by This work was supported by JST ACT-X, Grant Number JPMJAX2006 and JSPS KAKENHI 22K12287. We would like to thank all the people who reviewed this paper.

References

[1] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical report, Naval Technical Training Command Millington TNResearch Branch (1975).

[2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proc. of NAACL, pp. 4171–4186 (2019).

[3] Martinc, M., Pollak, S. and Robnik-Sikonja, M.: Supervised and Unsupervised Neural Approaches to Text Readability, Computational Linguistics, Vol. 47, No. 1, pp. 141–179 (online), (2021).

[4] Manning, Christopher, and Hinrich Schutze. Foundations of statistical natural language processing. MIT press, 1999.

[5] Shinya Akagi, Kazuhiro Notomi. Development of Text Evaluation System based on Readability Index for Japanese: A Comparison of Indices and Students' Evaluation. N-007. FIT2016 (in Japanese)