Examining readability of text generated through GPT-3

Aron Winkler
University of Tübungen
winkler.aron5@gmail.com

Abstract

This paper examines readability-related metrics of AI-generated text, specifically texts produced by OpenAI's GPT-3. Generation was performed for the English language across 10 topics and the 6 proficiency levels defined by the CEFR framework, and outputs were compared with Simple Wikipedia and (standard) Wikipedia articles on the same topics. 7 commonly used complexity metrics have been employed, targeting chiefly lexis and syntax. In addition, a classifier is also trained on Wikipedia data and used to evaluate the generated texts. In the limited scope of this experiment, GPT-3 is observed to be at least partially aware of what constitutes reading difficulty.

1 Introduction

AI text generation has experiences a massive rise in popularity with the media reach achieved by the release of OpenAI's GPT-3 (Brown et al., 2020) and its chat GUI interface ChatGPT[1]. While early public opinion focused on how - in the school setting - it would be students who would benefit from the new system by having their schoolwork done for them, little was said about how AI text generation could become an asset for didactitians in generating teaching material.

AI text generation might become an important tool for teachers, allowing them to generate learning material about topics and events that engage their students that might not be available on the market. An interesting setting for such a phenomenon is language learning, where it is necessary to produce learning material for students at different proficiency levels: whether to create ex-novo or simplify existing text material, AI text-generation tools can easily be imagined as having an enabling effect.

It is however not obvious that AI generation tools reach the desired quality or adherence to the didactic goals that teachers set out. Owing to the increased popularity of these tools, more research is desirable on their applicability in a number of different settings. For example, GPT-3 has been observed to "make up" information about some real-world events or phenomena, a fact that makes it hard to rely on for teaching. In the specific context of this experiment, employing GPT-3 for language-learning material is only valuable insofar as the generated content is adequate in difficulty to the audience.

This work aims to make a contribution to these questions by verifying the readability level of text content generated by GPT-3 with regard to 10 topics and the 6 proficiency levels defined by the CEFR framework. A number of complexity metrics are computed on the generated texts and compared with Simple Wikipedia[2] as well as Wikipedia[3] articles on the same topics in order to gauge the true readability levels of the model outputs.

2 Dataset and methods

2.1 Dataset

OpenAI's REST api was employed to produce texts with the "text-davinci-003" model (a GPT-3 variant) across 10 topics[4] and 6 CEFR proficiency levels[5]. Topics were selected without any particular guiding principle other than a general goal of variety between events, entities, and phenomena. The prompts fed to the model followed the template:

> *Write a text for learners of English at the [level] level about [topic].*

In addition to the 60 texts obtained in this way, the dataset was enriched by Simple Wikipedia and

[1] https://openai.com/blog/chatgpt

[2] https://simple.wikipedia.org/wiki/Main_Page

[3] https://www.wikipedia.org/

[4] 'Color Blindness, The Great Depression, Butterflies, Dogs, Semantics, The Internet, The Moon, Dinosaurs, Economics, Quantum Mechanics'

[5] A1, A2, B1, B2, C1, C2

Wikipedia articles corresponding to each of the 10 topics, resulting in an additional 20 dataset entries.

2.2 Preprocessing

Preprocessing of the dataset comprised enrichment with POS and dependency annotation. Part-of-speech tagging was achieved with nltk (Loper and Bird, 2002) and dependency relations were added through nltk in conjunction with CoreNLP models (Manning et al., 2014).

2.3 Metrics

Metric	Abbreviation
Sentence length	ASL
Noun to verb ratio	NVR
Type Token Ratio (first 100 tokens)	TTR
Clauses Per Sentence	CPS
Length of Longest Dependency Link	LLDL
Parse Tree Depth	PTD
Ratio of words in top 3000 English words	WIT

Table 1: Readability metrics

Table 1 lists the readability metrics used in this experiment. This specific selection was inspired by (Venturi et al., 2015), but many more examples of their usage can be found.

This feature set almost exclusively targets lexis and syntax, with little regard for morphology outside of the noun-to-verb ratio feature. While this is a limitation of this set, the language of this experiment is solely English, a language not necessarily known for a complex morphological system.

The feature "Words in Top 3000 English words" uses unigram frequencies from the Google Web Trillion Word corpus[6] to isolate the 3000 most frequent words of the English language. Many research approaches to readability employ similar ideas to gain insight into the lexical complexity of texts, such as the BIV (basic Italian vocabulary) employed by (Dell'Orletta et al., 2011).

All feature values outside of type-token-ratio (which ignores sentence boundaries) are computed at the sentence level. When a dataset entry comprises multiple sentences, the output for that entry is the average of the values for the sentences that compose it.

[6] https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

2.4 Text classifier

A classifier was trained on Wikipedia and Simple Wikipedia data, specifically on a train dataset of approximately 240k paragraphs, evenly distributed across sourced from Wikipedia and Simple Wikipedia articles. A test set of 40k entries, sourced and distributed in the same manner as the train set, was also obtained to evaluate the data. The task of the classifier was to determine whether a given paragraph had originally been a part of Simple Wikipedia or (standard) Wikipedia - $P(standard)$ is the derived readability metric reported in the results.

The model takes BERT sentence embeddings as input, which are fed through a single-layer LSTM with hidden size 126. LSTM outputs are then used to compute a distribution over the two labels in the dataset, namely "simple" (paragraph obtained from Simple Wikipedia) and "standard" (paragraph obtained from standard Wikipedia). Training was performed with the Adam optimizer and a batch size of 50 for 5 epochs. After evaluating the model at each epoch, it was observed that after the third epoch there are signs of overfitting, therefore the final parameters were those obtained after 3 epochs of training. Hardware-wise, the free tier of Google Colab was used - do to time limits set on the platform, several days were necessary for the full training process. After the third epoch, the model achieves 87.8% macro-averaged F_1-score on the test set. While this is not state-of-the-art performance, it is likely good enough for the scope of this project.

3 Results

3.1 Metrics

Table 2 details the average output for each difficulty level and metric. For each metric, the 3 values most indicative of low readability are highlighted.

With regards to the selected metrics, GPT-3 does a good job generating texts for the target readability level in the sense that prompts requesting C1 and C2 level material on average yield the least readable texts, whereas lower proficiency levels tend to have scores indicating higher readability.

While no rigorous checks are carried out in this paper in reference to the correlation between these metrics and article readability levels, it's still possible to examine some of the results to gain some

Level	ASL	NVR	TTR	CPS	LLDL	PTD	WIT
S Wiki	18.145	**1.918**	**0.661**	1.629	7.868	4.629	**0.749**
A1	18.539	1.681	0.657	1.932	8.0	5.113	0.83
A2	20.717	1.477	0.648	**2.104**	8.843	5.415	0.828
B1	19.124	1.607	0.65	**2.114**	8.416	5.383	0.826
B2	20.766	1.694	0.632	**2.064**	**9.148**	5.64	0.821
C1	**21.24**	**1.848**	0.646	1.967	9.147	**5.549**	0.804
C2	**21.94**	1.836	**0.658**	1.94	**10.189**	**5.427**	**0.791**
Wiki	**26.131**	**2.342**	**0.667**	1.876	**11.323**	**5.681**	**0.678**

Table 2: Average metric results for each difficulty level and metric. The 3 highest values are highlighted for each metric. For WIT, the lowest 3 are highlighted instead.

insight into the black box that is GPT-3.

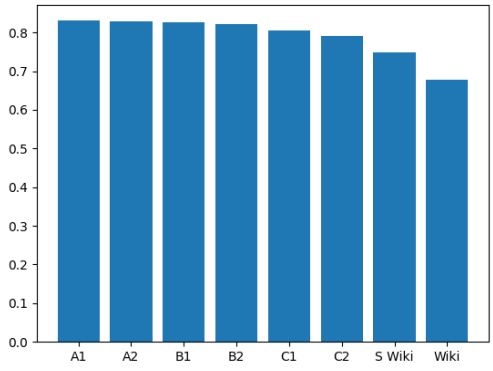

Figure 1: Mean metric values for WIT.

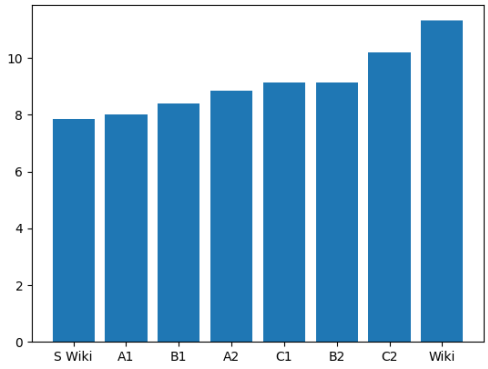

Figure 2: Mean metric values for LLDL.

Among the selected measures, WIT (ratio of words in the 3000 most frequent English words) appears to be the most well-behaved, insofar as the prompted CEFR levels mostly maintain their natural progression. Figure 1 shows the isolated scores for the WIT metric, which show a correlation with difficulty. Wikipedia and Simple Wikipedia articles have lower values, however this is possibly connected to those texts being several times longer.

LLDL (length of longest dependency link) is another metric that performs well in terms of level differentiation. It also reflects the common belief that the biggest difficulty spike in language learning - as shown by Figure 2. As with most metrics, however, the natural progression of the CEFR levels is lost, particularly in the middle portions.

Interesting discussions can be had around topics as well, since, depending on GPT-3's ability to generalize across domains, certain areas might constitute problem for it. For example, "semantics" and "quantum mechanics" would be hard to write about for A1 level students as they are likely to be children who have not interacted with those fields yet.

Generally, a difference in values is visible across the board, more so at the lower difficuly levels. At CEFR levels B2 and above, many of the metrics appear similar regardless of topic. Among the selected measures, NVR (noun to verb ratio) shows the highest cross-topic variation, maintaining strong differentitation even at the C2-level texts. Figure 3 shows metric values at the A1 and C2 levels, which indicate that topics commonly regarded as intrinsically have higher information density in GPT-3 texts. This is however a weak case, as few metrics differ so sharply as NVR, and bars in these figures correspond to a single text entry.

3.2 Classifier scores

Table 3 shows classifier outputs, with highlighted values for each topic and difficulty level associated to lowest readability. The Wikipedia entries that were reported on in previous sections are not included here, as they may have been part of the train set of the classifier.

	A1	A2	B1	B2	C1	C2	*Avg*
Color Blindness	0.020	0.022	0.056	**0.095**	0.015	0.030	0.04
The Great Depression	0.014	0.128	0.027	**0.890**	0.342	0.461	0.31
Butterflies	0.016	0.027	0.014	0.041	0.021	**0.281**	0.067
Dogs	0.028	0.069	0.013	**0.208**	0.088	0.041	0.075
Semantics	0.014	0.144	0.013	**0.541**	0.475	0.262	0.242
The Internet	0.012	0.027	**0.291**	0.043	0.031	0.041	0.074
The Moon	0.022	0.073	0.026	0.029	0.053	**0.812**	0.169
Dinosaurs	0.019	0.015	0.016	0.272	0.052	**0.384**	0.126
Economics	0.031	0.012	0.018	0.344	0.222	**0.697**	0.221
Quantum Mechanics	0.011	0.024	0.121	0.465	0.313	**0.628**	0.26
Avg	0.019	0.054	0.060	0.293	0.161	0.364	

Table 3: Classifier scores corresponding to $P(standard)$. In each row the highest value is bolded, and in each column the highest value is underlined.

These results show again that GPT-3 has learned what constitues readability, at least to some extent. For 4 of the topics presented in this study, the C2 prompt achieves the highest complexity (or lowest readability) level, with 3 more having B2 as the top entry.

As was observed with the metrics, there tends to be good differentiation between the extremes (i.e. between A1 and C2 texts), but this is less true in the middle portions. Interestingly, prompts for C1 level texts often do not yield particularly difficult texts according to the classifier. Such a claim is confirmed by looking at the average values for each level, where C1 does score higher than A1 and A2, but lower than B2. Whether this is caused by this classifier's failure to capture some aspect of text readability, ot by GPT-3's lessened ability to generate material for this target level, remains to be confirmed by more robust experiments.

On the matter of inherent topic complexities, there is unfortunately not much to talk about - no topic or set of topics emerge as consistently more difficult than others. However, *Economics* and *Semantics* do get the highest average score, indicating they are the least readable, while (more surprisingly) *Color Blindness* is the most readable on average.

4 Discussion and future research

This work attempted to gauge true readability levels of GPT-3 output texts with prompts across 10 topics and CEFR proficiency levels, optimising for the second language learning application environment. Outputs computed from the metric set suggest that GPT-3 is at least somewhat aware of syntax and lexis when performin generation for a target readability level.

Future research into this topic could explore a wider variety of textual metrics, as this work only selected a handful that lent themselves particularly well to interpretation and did not require steep hardware requirements to obtain.

Since recent research (e.g. Deutsch et al., 2020) has also focused on readability assessment through neural approaches without such metrics, this paper also attempted to adopt this strategy by training a classifier on Wikipedia and Simple Wikipedia data. Outputs derived from this model also support the theory of GPT-3 being at least somewhat aware of readability-related aspects of language. With regards to this section, it is nonetheless paramount to utilise a larger, SOTA readability model for exploring complexity - as the one employed here is only suitable for prototyping - as well as to perform the experiment on a richer repository of GPT-3 data.

In summary, future extensions of this experiment or related experiments would most benefit from higher volume. Simply put, more data is necessary across multiple domains to make educated conclusions about the quality of GPT-3 outputs in a teaching setting. While data is easy to generate, quality annotations on it are yet unavailable. Indeed, it is a good question whether making any is a fruitful enterprise, as the actual models behind GPT-3 and ChatGPT are constantly updated - thus making it hard to pinpoint what exactly is being studied.

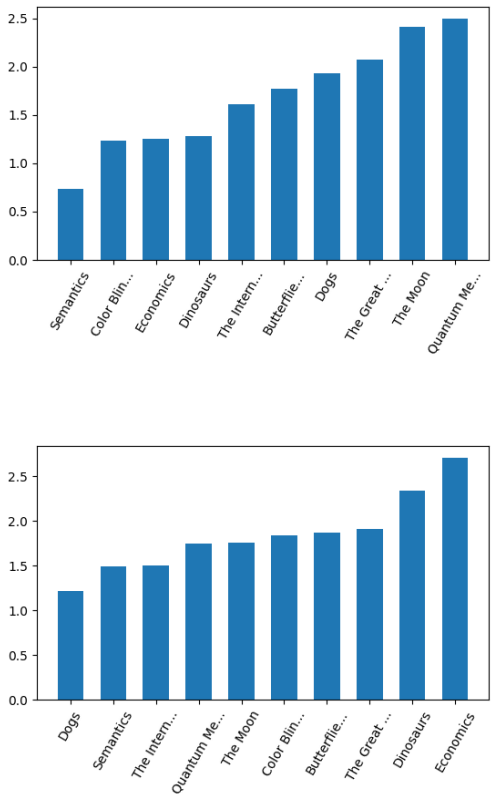

Figure 3: Text values for NVR. Top figure shows A1-level, bottom figure shows C2-level.

Acknowledgements

I thank Prof. Dr. Walt Detmar Meurers for exposing me to the topics of readability and automatic readability assessment, and for the excellent course he taught on the matter.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 62–69. Somerset, NJ: Association for Computational Linguistics. http://arXiv.org/abs/cs/0205028.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Giulia Venturi, Tommaso Bellandi, Felice Dell'Orletta, and Simonetta Montemagni. 2015. NLP–based readability assessment of health–related texts: a case study on Italian informed consent forms. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 131–141, Lisbon, Portugal. Association for Computational Linguistics.

Resources

All code for this paper is available at https://github.com/waron97/nlp_for_readability_term_paper.