An analysis of

Lower Perplexity is Not Always Human-Like

**Tatsuki Kuribayashi[1,2], Yohei Oseki[3,4], Takumi Ito[1,2], Ryo Yoshida[3],
Masayuki Asahara[5], Kentaro Inui[1,4]**

[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN [5]NINJAL

Acer Jagrup

CS4395.001 – Human Language Technologies

Dr. Karen Mazidi

University of Texas at Dallas

Problem

This paper addresses the issue of how language models are evaluated for human likeness. Specifically, the paper focuses on the concept of lower perplexity implying that a model is more humanlike to determine if that conclusion, which was made based on research on English, is generalizable to other languages. Human-like computational models, in this context, are ones which “better predict human reading behavior” by performing better at next-word prediction. The authors conclude that less perplexity is not universally associated with higher human-likeness, and so universal human-like models will require analysis through the lens of multiple languages for adequate evaluation.

Prior Work

The study builds on and reexamines prior work related to human reading behavior. Most notably, the paper revisits a report on the trend of lower perplexity correlated with human-likeness and makes many references to studies on the widely-adopted surprisal theory, which “suggests that the processing difficulty of a segment is determined by how predictable the segment is in its preceding context”. The study also analyzes the results of studies done using eye-tracking data corpora, namely the BCCWJ-EyeTrack corpus and the Dundee Corpus for Japanese and English data, respectively. Additional supplemental references are made throughout the paper which are listed in the references section, including prior work done by the authors themselves.

Unique Contributions

The report considers trends outlined by previous reports and analyzes them against Japanese eye movement data to demonstrate “a surprising lack of universality”. For that reason, this report serves to prompt more discussion on cross-lingual evaluation of language models and more similar reports with other languages which haven’t had the same focus as English due to a lack of readability data (such as eye movement recordings), and so contributes a unique perspective to contrast with the English-speaking perspective when evaluating language models.

Self-Evaluation

The authors acknowledge that the differences between English and Japanese on a linguistic level could contribute some to the disparity of results that were achieved and call for more tests of other languages to help assess how much of an impact those differences caused. They also recognize the smaller size of the Japanese eye tracking dataset and call for more studies on more diverse texts and the “creation of a large-scale corpus of human reading behaviors in diverse languages”.

Citations and Importance

The authors of this report have received numerous citations as follows (according to Google Scholar) [1]:

Tatsuki Kuribayashi	217
Yohei Oseki	231
Takumi Ito	211
Ryo Yoshida	13,480
Masayuki Asahara	2,888
Kentaro Inui	8,046

Ryo Yoshida received the most citations by far, but each author has received hundreds to thousands of citations, though not necessarily in an NLP-related field—Takumi Ito has received many citations for driving assistance devices, for example.

I think this report was important because it serves as a stepping stone for further examination into the universality of the trends we gather when looking at data in just one language, like English. It is important to recognize that correctly-identified trends in one language may not translate well if at all to other languages, and our analysis of natural language should extend to all human languages, within reason. I appreciate that this report encourages more study into both Japanese texts and texts in other languages, including forms of reading tracking like eye trackers.

Bibliography

[1] Google, "Google Scholar," Google, 8 November 2022. [Online]. Available: https://scholar.google.com/schhp?hl=en. [Accessed 8 November 2022].

[2] T. Kuribayashi, Y. Oseki, T. Ito, R. Yoshida, M. Asahara and K. Inui, "Lower Perplexity is Not Always Human-Like," in *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.