A Japanese Dataset and Efficient Multilingual LLM-Based Methods for Lexical Simplification and Lexical Complexity Prediction

Adam Nohejl[1] Akio Hayakawa[2] Yusuke Ide[1] Taro Watanabe[1]
[1]Nara Institute of Science and Technology [2]Universitat Pompeu Fabra
{nohejl.adam.mt3, ide.yusuke.ja6, taro}@is.naist.jp akio.hayakawa@upf.edu

Publication Information

Volume 32, issue 4, pp. 1129–1188.
DOI: https://doi.org/10.5715/jnlp.32.1129

Abstract

Lexical simplification (LS) is the task of making text easier to understand by replacing complex words with simpler equivalents. LS involves the subtask of lexical complexity prediction (LCP). We present MultiLS-Japanese, the first unified LS and LCP dataset targeting non-native Japanese speakers, and one of the ten language-specific MultiLS datasets. We propose methods for LS and LCP based on large language models (LLMs) that outperform existing LLM-based methods on 7 and 8 of the 10 MultiLS languages, respectively, while using only a fraction of their computational cost. Our methods rely on a single prompt across languages and introduce a novel calibrated token-probability scoring technique, G-Scale, for LCP. Our ablations confirmed the benefits of G-Scale and of concrete wording in the LLM prompt. We made the MultiLS-Japanese dataset available online under a CC-BY-SA license, including detailed metadata.[1)]

1) The complete dataset is made available at https://huggingface.co/datasets/naist-nlp/multils-japanese

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).