Lexical Simplification: From a Japanese Dataset to Multilingual Methods Leveraging Large Language Models and A YouTube Subtitle Corpus

Name: Adam Nohejl

Laboratory's name: Natural Language Processing Laboratory

Supervisor's name: Taro Watanabe

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 the first unified LS and LCP dataset targeting non-native Japanese speakers, MultiLS-Japanese, one of the 10 language-specific MultiLS datasets. We propose methods for LS and LCP based on large language models (LLMs) that outperform existing LLM-based methods on 9 of the 10 MultiLS languages, while using only a fraction of their computational cost. Our methods use the same prompt across all languages, and for LCP we employ a novel calibrated token-probability-based scoring technique, G-SCALE. Our ablations confirm the benefits of G-SCALE and of concrete wording in the LLM prompt.

We use YouTube subtitles to build TUBELEX, a corpus approximating spoken vocabulary for five diverse languages, Chinese, English, Indonesian, Japanese, and Spanish. We evaluate the correlation of frequency norms based on the corpus with lexical decision time, word familiarity, and lexical complexity, showing that they provide an approximation comparable to, and often better than, the best available resources. We use lexical frequency and dispersion from TUBELEX to further improve our LLM-based LCP method.

We provide further insights into how non-native speakers perceive Japanese lexical complexity. By analyzing a partial reannotation of MultiLS-Japanese by native Chinese speakers, we demonstrate that their perception of lexical complexity in Japanese differs from the perception of speakers of other languages due to Sino-Japanese vocabulary. We also show that there are considerable differences in lexical complexity perception between individual non-Chinese speakers, making personalization desirable but difficult using the currently available data.