Abstract

Even with the recent advancements in machine translation quality, one of the remaining challenges of machine translation is the presence of neologisms in short texts. Neologisms are newly coined words or expressions that are not yet accepted by the mainstream language but are gaining wider usage. Having these words in a sentence lowers the quality of translation since systems often treat these words as proper nouns, leading to a fragmented output. In this project, we develop a pipeline that can automatically detect potential neologisms and find possible replacements for the unknown word. The replacements are found by passing through a fill mask layer of a widely used large language model, BERT, which incorporates surrounding context to estimate the word that will fit into the sentence.

The detection pipeline can separate relatively new words from all unique words in the dataset, but the performance still suffers from being unable to distinguish between neologisms and named entities. By replacing these detected neologisms using the fill mask task of BERT, the system obtained a result of 0.249 BLEU score compared to 0.202 BLEU score from the raw translation. Translation through replacement shows to perform better qualitatively; while raw-translated sentences are highly fragmented, sentences translated after the replacement are more structured. However, even with using a large language model, the system is not capable of detecting the exact meaning of the neologism since the training data of the language model are based on formal text like newspaper articles. This thesis emphasizes the need to conduct further research regarding the identification and processing of neologisms in text and the necessity to develop large language models that are trained on short informal texts which are becoming more dominant in modern days.