Unsupervised English-Korean Translation

CSCI 1460 Project Proposal

Jeanne Bang

March 22, 2021

Machine Translation has been showed impressive performance since machine learning techniques were applied. However, to train the machine translation models, a large and well-paired language data set is essential. This type of data is hard to build, and the amount of parallel multilingual corpora is absolutely smaller than monolingual corpus. G. Lample suggested a possibility of "unsupervised machine translation using monolingual corpora only" [1]. The main idea of the paper is building a shared latent space between source and target language, and generating a sentence from the latent space. The paper examined its method with English-French and English-German data. For this project, I will re-implement the paper, with English-Korean data [2]. This parallel corpora contains 94,123 sentences for training and 2,000 sentences for test. The performance of the model will be measured by perplexity and accuracy, and compared with the performance of standard supervised machine translation. If time allows, this project can be extended to the unsupervised model with BPE (byte pair encoding) or semi-supervised model.

References

[1] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

[2] Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. Korean language resources for everyone. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 49–58, Seoul, South Korea, October 2016.