Project Proposal: Smart Docx Translator

Jakob Bergman | jabergma@kth.se
Xiaoying Sun | xsu@kth.se
2021.11.16

Background

The subject we are working with is language translation. We want to create a machine that can translate Chinese to English. It is based on PaddleHub and we will be able to do free and local text translation.

Problems

1. Which programming language is going to be used? Python
2. Which dataset is going to be used for training? wmt19_translate/zh-en
3. Which library is going to be used?
    - python-docx
    - PaddleHub
4. Which model is going to be used?
   Trasnformer[1] is the mainstream model in the field of machine translation, including model training, prediction, and the use of custom data. Users can build their own translation models based on the published content.

Goal

This is designed as a two-person project. At present, we use the online translation capabilities of major manufacturers to translate some paragraphs (at least the author is like this). However, based on PaddleHub, we can achieve free and local text translation capabilities. At the same time, this project also provides the possibility of forming a translation capability that can more accurately match your own data sets.

Contributions

Our work will help bring language barriers down. We expect to be able to translate Chinese to English accurately and freely.

Reference

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).