Classification of Human and Machine Translation by BERT

Yifei Zhou

May 14, 2021

1 Abstract

This paper handles a classification task in the context of machine learning. We are given a reference sentence in Chinese, a gold translation by human to predict whether another translation is produced by human or by machine. Our classification model is an end-to-end model without the need of feature engineering. It makes use of the pretraining model bert-base-uncased and bert-base-chinese to successfully attain a final macro F-1 result of 0.878. Note that this model is a fairly quick and straightforward one, future directions and possible improvements are also articulated.

2 Data preprocess and feature extraction

Our training data contains 584 pairs of sentences. By preliminary inspection of the training data, we discovered that the machine translation and human translation do not differ much in their BLEU score. Although some of the machine translations are somewhat incoherent for human beings, there are no outstanding features to distinguish them from human translations. We inspected the confidence value given by a part-of-speech tagger provided by nltk, but the results are not very informative either. Additionally, the data size is actually pretty small to carry out semantic analysis on its own. Therefore, we decide to include the semantic features from BERT for the classification task.

We include in our feature vector the bert embeddings of the Chinese referrence (although it actually does not contribute a lot in this context), the bert embeddings of the gold translation and the bert embeddings of the translation to be tested. We made use of a PCA of .93 for fast training and decreasing overfit.

3 Experiment results

We feed the PCA transformed features into a dense neural network for the classification. Our neural network architecture is 400(input layer)-32-8-2. We split the training data into training set and validation set by a factor 8:2. After that, we calculate the macro F-1 score on the testing set, which is 0.878 (the macro F1 on the training set is 0.982).

4 Conclusions

This is a basic implementation to classify human and machine translations from a semantic perspective. Many improvements are still to be made. For example, trying POS features, implementing other classification models, and increasing the training size. But it can already be concluded that semantic classification for human and machine translation is pretty promising.

5 Github repository

`https://github.coecis.cornell.edu/yz639/translation_classifier`