Human vs Machine Translation (GitHub)

Kabir Walia (ksw63)

The following methods were based on experiments performed by **Bhardwaj et al** [COLING 2020]. Before diving into the models, here are the final results:

Model	Stacked bi-LSTM	**BERT (monolingual)**	mBERT (bilingual)
F1	0.712	**0.906**	0.882

Method 1: Monolingual model: stacked bi-LSTM (layers = 2): I developed a stacked (2 layer) bi-directional LSTM which was trained solely on the *candidate* translations. The translations were vectorized using **GloVe embeddings (300d)** and fed into the model using a custom PyTorch Dataset and Dataloaders. The intuition behind the bidirectionality was that structural cohesion of the machine translated sentences might be worse. I added a **Dropout (p=0.3)** considering the small size of the dataset. This reduced overfitting and improved the validation performance. I trained for 10 epochs with a learning rate of 0.0001, batch size of 32, and optimizer AdamW after varying these parameters to find a local optimum on eval.

The next two transformer-based methods were developed using the **simpletransformers** library – which adds an abstraction on top of **Huggingface.** I finetuned the pretrained models (which had sequence classification heads or linear layers on top – see BertForSequenceClassification) using the **ClassificationModel** class – which comes with train and eval methods.

Method 2: Monolingual Model – BERT: I finetuned the **bert-base-cased** pretrained BERT model – a 12-layer transformer model trained on purely English text. I finetuned the model on the candidate translations for **5 epochs**, with a **max_seq_length** of **256** and **AdamW**. BERT employs Multi-head attention (12 heads) which would produce far richer representations of the translation than the bi-LSTM. Furthermore, BERT's bidirectionality and masked language modeling aides its ability to learn contextual patterns. Although Bhardwaj et al attempt to use camemBERT on the bilingual formulation as well, a majority of the time multilingual BERT outperformed it, so I decided to try this model on only the candidates.

Method 3: Bilingual Model – mBERT (multilingual BERT): I finetuned the mBERT model on **both** the source Chinese as well as the candidate English sentences. Although similar in configuration the bert-base-cased, mBERT is trained on **104 languages**. This would enable it to learn a representation for the Chinese sentence as well. Apart from added information, feeding Chinese sentences with the translations also allow mBERT's attention mechanisms to better capture **word-level correspondence**. Further, due to its pretraining on Chinese and English, its theoretically expected to better handle **spurious words** as well as words with different **fertility** (language to language). I train for 7 epochs with similar hyperparameter values at method 2.

Results: Surprisingly, the best performance comes from the monolingual BERT model – with a (macro) f1 score of **0.906**. I had hypnotized that with the added information, the bilingual model would perform the best. Both, however, significantly outperform the bi-LSTM model. The monolingual BERT model produces an **equal number of FP/FN (5)** whereas the bilingual BERT gets 9 FN and 3 FP (FN = machine translated classified as human). This shows that the slight imbalance could be impacting the model performance. However, when solely focusing on the English translations, that effect is reduced. Fair to say that (these) human and machine translations can be considered fairly distinguishable based on these results.

Future Work/Alternative Methods: 1) ensemble of these, **2)** feature-based (different individual and cumulative BLEU scores or even NIST values, **3)** customized architectures with attention mechanisms to incorporate reference translations.