情報工学実験Ⅱ レポート

B3　1201201120　堀本　隆誠

November 15, 2022

1　学習内容

日本語 BERT の使用方法を学んだ.

トークナイザーが存在するものについてもあらかじめ分かち書きをしておいたほうがよい. また, 京都大学の BERT を使っている場合は juman++ , 東北大学の BERT を使っている場合は mecab を用いている.

今回, 京都大学が公開する BERT 日本語モデルを用いて fine-turning なしの場合とありの場合において実行した. input_ids は, 入力した文に対して一対一に対応した単語の id の行列を表している.

2　苦労した点

BERT 日本語モデルを入手してそれを Google Colab で使用することに苦戦した. しかし, これにより, どのファイルを使用すればよいのか, それらのファイルの中身は何なのかを知る機会になった. config.json はパラメータのサイズなどを記したファイル, pytorch_model.bin は中間層やその他のパラメータをまとめたファイル, vocab.txt はさまざまな単語を行ごとに分けて書かれたファイルである.

3　理解が不十分な点

全体的に理解が不十分に感じるが今回特に不十分だと感じた点は, pytorch における tensor に関する理解である. これにより, 本来理解できるところも理解できていないように感じる. そのため, 今後はまず pytorch やその他の基礎を学び直す.