進捗報告

1 今週やったこと

- Box Embedding の資料を読んだ.
- BERT のサンプルを実行した.
- BERT のモデルを東北大学のものに変更しサンプルを実行した.

2 内容

2.1 Box Embedding の資料について

Word2Vector と Word2Box の違い, Box Embedding の考え方や学習方法について概観を理解することができた. また, Gumbel Box や Soft Box などの Box Embedding を行う方法の利点・欠点を把握することができた.

2.2 BERT のサンプル実行

BERT のサンプルはローカル環境で実行を行ったが, 実行に必要な PyTorch などのライブラリがインストールされていなかったため事前準備が必要だった. また, サンプル内の BERT_DIR が機能していなかったのでファイルパスを直接指定する必要があった. サンプルを一通り実行した後にサンプルで利用していた京都大学のモデルから 東北大学のモデルへと切り替えて実行した. ともに日本語を対象としたモデルであるものの数値表現に違いが見られ, fine tuning の example では 2 つの文章を分類する二値問題で京都大学のモデルでは loss 合計が 0.00851310882717371 , 東北大学のモデルでは 0.06271297391504049 となる変化が存在した. マスク問題でも 2 つのモデルで結果に相違が発生することを確認した.

3 まとめ

自然言語処理で必要とされる BERT の PyTorch での基本的な使い方をサンプルを通じて学んだ. BERT に関する知識が不足しており関数の意味などを完全に理解することはできなかった. しかし, ファインチューニングなど既存のモデルを活用し組み込む方法を実際のコードから具体的にどのように行うかを把握することはできた. また, 学習モデルの違いによって引き起こされる学習結果や識別結果の変化をサンプルの実行を通して確認することができた. Box Embedding に関しては単語の表現をどのように表し, その表現をどのような形で効率よく学習するかの概念を知ることができた.