進捗報告

1　今週取り組んだこと

アンサンブル学習を利用して作成したマルチラベル分類モデル (mpm+T) から得られた特徴量ベクトルを活用した BERT+MLP のモデルで, 特徴量ベクトルの次元数を複数のパターンに分けてマルチラベル分類をした.

2　データセット

データセットは楽天グループ株式会社が公開している「楽天トラベルレビュー : アスペクトセンチメントタグ付きコーパス」[*1]を使用した. 楽天トラベルの日本語レビュー文章とそれぞれの文章について, 立地, 部屋, 食事等の 7 項目のカテゴリに対するポジティブまたはネガティブのタグが付与されている.「朝食, 夕食, 風呂, サービス, 施設, 立地, 部屋」のポジティブ, ネガティブの 14 個のカテゴリに分類される. 今回は 14 のいずれのカテゴリにも属さないデータを除くことで, 少なくとも 1 つのラベルに属し, 語彙数が 10 以下と 100 以上のデータを取り除いた. 総データ数は 50211 である.

3　実験

図 1 に現在の実験で用いているモデルの概要図を示す. 圧縮する次元数は, 28, 56, 112, 224, 448 と変化させたが, いずれにおいても精度向上は見受けられなかった. 具体的には, 損失はわずかに低下するが, 精度 (micro-F1) がほぼ一定のままで, 予測値も固定化されてしまっていた. Transformer 層の学習時の特徴量ベクトルを圧縮する工夫することを検討している.

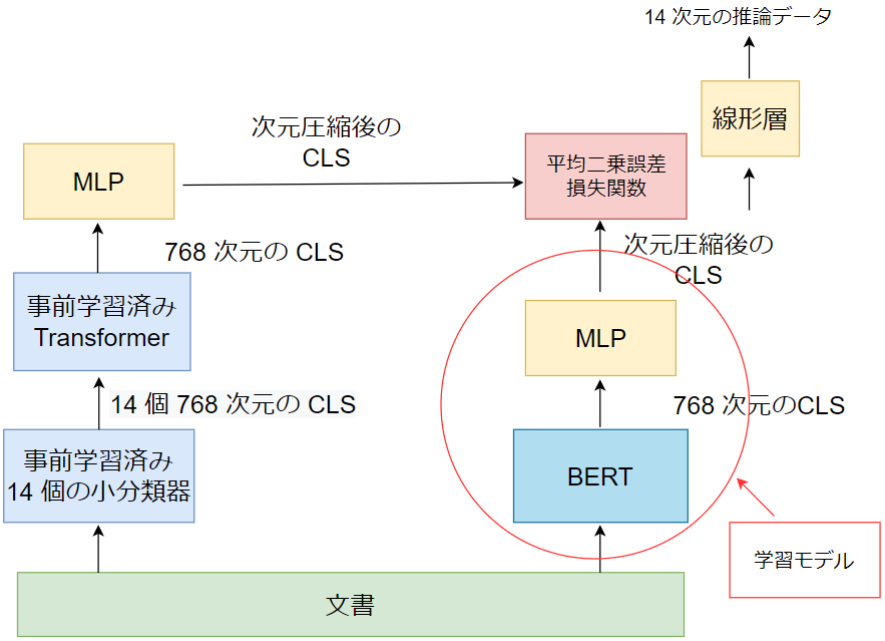

図 1: mpm+T を含めた実験モデルの概要図

[*1] https://www.nii.ac.jp/dsc/idr/rakuten/

4　次に取り組むこと

Optuna を用いてシンプルな BERT+MLP のマルチラベル分類モデルのハイパーパラメータチューニングをする.