進捗報告

1　今週取り組んでいること

図 1 に提案モデルの概要図を示す. 図 1 のモデルにおいて, CLS トークンの次元圧縮をして精度向上を試みている. BERT+MLP モデルと mpm+T モデルのパラメータを再度チューニングしている. 目的は, BERT+MLP モデルのパラメータチューニングと, チューニング後の BERT+MLP モデルと mpm+T モデルに同じテストデータを入力して分類精度を厳密に比較することである. BERT+MLP モデルのパラメータのチューニングは完了した.
mpm+T モデルのパラメータはチューニング中で, そちらのパラメータや分類精度についても再度アップロードする.

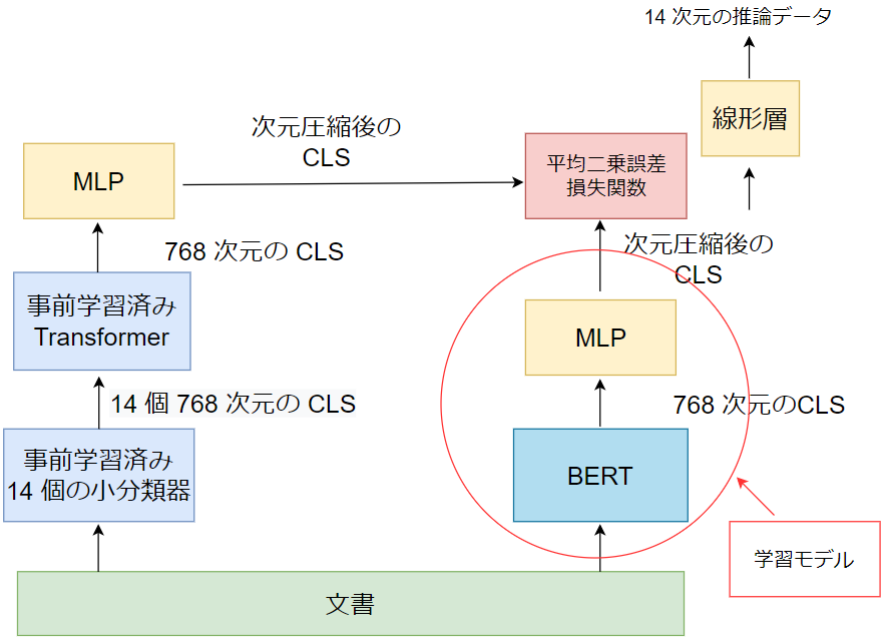

図 1: mpm+T を含めた実験モデルの概要図

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.

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

3 実験

BERT+MLP モデルの学習率をチューニングした. 表 1 に BERT+MLP モデルでのマルチラベル分類パラメータを示す. また, 表 2 にチューニングして得た学習率で訓練データ数が 24000 であり, 検証データ数が 6000 となるような 5 分割交差検証をした場合の分類精度を示す.

表 1: BERT+MLP でのマルチラベル分類パラメータ

パラメータ	値
BERT 層の入力次元数	768
BERT 層の出力次元数	768
MLP 層の層数	3
MLP 層の入力次元数	768
MLP 層の出力次元数	14
バッチサイズ	40
最適化関数	Adam
学習率	0.0000002
損失関数	BinaryCrossEntropyLoss
エポック数	20

表 2: 実験結果の評価指標の平均と標準偏差

評価指標	Precision	Recall	micro-F1
BERT+MLP	0.683 $\pm$ 0.00468	0.772 $\pm$ 0.00871	0.724 $\pm$ 0.00462