進捗報告

1　今週取り組んだこと

提案モデルで得た特徴量を正解データとして, BERT モデルと楽天トラベルレビューのデータセットを用いたマルチラベル分類をした. 以前ではデータ数の不足が精度が向上しないと考えたので, 訓練に使用するデータ数を 10000 から 30000 に増加することで問題の解消を試みた.

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. また, いずれのカテゴリにも属さないデータと, 1 つのカテゴリにポジティブとネガティブの両方が付与されたデータを取り除いた. 総データ数は 48354 である.

3　実験

1. BERT 1: 正解データをデータセットから得られる 0,1 の 14 クラスのラベルとする場合.
2. BERT 2: 正解データを, 事前学習済みのモデルにデータセットの文章を入力して得られる特徴量とする場合.　特徴量である正解データの次元数は 768 次元であるが, 今回は 14 次元に圧縮して用いた.

1. と 2. の場合の分類精度の比較と, Attention を一部可視化して差異を確認した. ただし, 交差検証はしていない. いずれの場合も訓練データ数 30000, テストデータ数 6000 で実験をした.

表 1 に実験時のパラメータを示す. 表 2 に実験結果を示す.

表 1: 2 値分類の実験時のパラメータ

パラメータ	値
BERT 層の入力次元数	100
BERT 層の出力次元数	14
バッチサイズ	4
最適化関数	Adam
学習率	0.00001
BERT 1 の損失関数	BCELoss
BERT 2 の損失関数	MSELoss
エポック数	15

データ数を大幅に増加することで精度の向上は確認できた. 14 クラスの 0, 1 で表現される正解データを用いる場合よりも特徴量を正解データとする場合の方が精度が高く出ているため, 本報告では記載できていないそれぞれのモデルの Attention の可視化をしてその結果を再度報告する. ポスター発表では, 提案モデルから得られた特徴量を用いることで, よりアスペクト情報に基づいた分類がなされていると言えることができれば良いと考えている.

表 2: 実験結果の評価指標

評価指標	Precision	Recall	micro-F1
BERT 1	0.633	0.794	0.709
データ数 10000 の場合	0.588	0.785	0.6723
BERT 2	0.703	0.810	0.753

参考文献

[1] 楽天グループ株式会社. 楽天データセット（コレクション）, aug 2010.