進捗報告

1　今週やったこと

BERT 層と Transformer Encorder 層を組み合わせたモデルの BERT 層において 5 分割交差検証をして 7 個の各カテゴリ毎に 2 値分類器を作成した. 目的は, 現在用いているデータセットである, 楽天トラベルレビューのアスペクトセンチメントタグ付きコーパスによるマルチラベル分類精度向上のためである. 各カテゴリ毎に今回作成した 5 個のモデルを用いて最もシンプルな Max Voting をして 2 値分類をする. そしてそれらの 7 個の分類結果を binary relevance により和集合として予測する. 各ラベルでの 2 値分類では以下のパラメータを用いた.

表 1: 各カテゴリの 2 値分類のパラメータ

パラメータ	値
BERT 層の入力次元数	768
BERT 層の出力次元数	2
バッチサイズ	40
最適化関数	Adam
学習率	0.0001
損失関数	CrossEntropyLoss
エポック数	10

2　データセット

2.1　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

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とは, 楽天グループ株式会社が提供しているデータセットである. 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 項目のアスペクト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. 総データ数は 76624 で, 朝食, 夕食, 風呂, サービス, 施設, 立地, 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データ群にした. 総データ数は 50211 である.

2.2　実験

各カテゴリ毎に訓練データ数が 24000 であり, 検証データ数が 6000 となるような 5 分割交差をした. 作成した 5 つのモデルはデータ数が 6000 のテストデータを用いて評価した. いずれのカテゴリにおいても高い精度で分類ができている事に加えて, 精度のばらつきも非常に少ないことがわかる.

2.3　次にすること

これらのモデルを各カテゴリで統合して, ポジティブとネガティブという文章の極性分類を除いた 7 個のカテゴリの分類をする. 具体的には各カテゴリ毎に今回作成した 5 個のモデルを用いて最もシンプルな Max Voting をして 2 値分類をする. そしてそれらの 7 個の分類結果を binary relevance により和集合として予測する. その予測精度の比較対象として, ランダムフォレストによる分類や, BERT の出力次元数を 7 にして単一のモデルを作成して分類する手法を検討している.

表 2: BERT-Transformer モデルとベースラインの正解率の平均と標準偏差

朝食	夕食	風呂	サービス	立地	設備	部屋
0.916 (0.00146)	0.923 (0.00618)	0.943 (0.00598)	0.888 (0.00538)	0.951 (0.000677)	0.868 (0.00960)	0.919 (0.000844)