進捗報告

1　今週取り組んだこと

ポジティブのみ, またはネガティブのみのラベルが付与されている, 文章全体のポジネガのラベルを指定したデータセットと, ポジティブとネガティブの両方のラベルが付与されているデータセットを作成した.

2　元のデータセット

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」[?] 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　作成したデータセット

- 「データセット 1」ポジティブのみ, またはネガティブのみのラベルが付与されている, 文章全体のポジネガのラベルを指定したデータセットを作成した. データ数は 47308 である.
- 「データセット 2」ポジティブとネガティブの両方のラベルが付与されているデータセットを作成した. データ数は 2903 である.

4　実験

東北大学の乾研究室が後悔している BERT モデルに新たに作成したデータセット 1 の 30000 のデータを用いて 5 分割交差検証をして 2 値分類モデルを作成した. 正解率は 0.926 ± 0.0341 であった. このモデルに新たに作成したデータセット 2 を入力として, ポジティブかネガティブどちらに分類される可能性があるかを確認した.

5　次に取り組むこと

多クラス分類タスクを解くためのライトなモデルの分類精度を向上するための手法の検討.