令和 4 年 5 月 24 日 (火) ゼミ資料　　M2 楠本 祐暉

進捗報告

1　データセット

1.1　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, 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データ群にした. 総データ数は 50211 である.

2　今週やったこと

東北大学の乾研究室が公開している日本語評価極性辞書[*1]に記載されている評価語を含むデータを用いて BERT によるポジティブまたはネガティブの 2 値分類をした. 評価語を含む文章全体を形態素解析する場合と, その文章から評価語のみを抽出した場合の 2 つを紹介する. 表 1 に実験に用いた各クラスのデータ数と, TF-IDF に用いた単語数を示す. また, 表 2 に各カテゴリのポジティブとネガティブのデータ数を示す.

表 1: カテゴリに属するデータ数とそれらのデータの単語数

	朝食	夕食	風呂	サービス	立地	施設	部屋
カテゴリに属するデータの数	11924	8814	7406	15159	5213	10695	8353
全文の総単語数	8213	7448	6615	10943	5915	9815	7109
評価語のみの総単語数	1214	1122	1050	1084	873	1461	721

表 2: カテゴリに属するデータに含まれるポジティブとネガティブの数

	朝食	夕食	風呂	サービス	立地	施設	部屋
データ数	11924	8814	7406	15159	5213	10695	8353
ポジティブ	10125	7497	5823	11277	4571	6898	6559
ネガティブ	1799	1317	1583	3882	642	3797	1794

表 1 の単語数についての 2 種類のデータを用いて BERT による各カテゴリのポジティブとネガティブの 2 値分類をした. 表 3 に実験で用いたパラメータを示す. また, 表 4 に SVM, Random Forest, BERT のそれぞれの分類結果を示す.

いずれのカテゴリにおいても, BERT の評価語データの分類精度が全文データを用いた場合よりも低下している. この原因として 2 つの要因を考える. 1 つ目は, データのポジティブとネガティブのラベルの割合が不均衡であるカテゴリが多いことである. 2 つ目は, 評価語のみのデータは全文データと比較して Attention を計算するためのトークンが不足していたことである. 表 5 に各カテゴリのトークン数の平均を示す. 実際に全文データと評価語データではトークン数

[*1] https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/

表 3: 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	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	15

表 4: 全文データと評価語データで 2 値分類をした正解率

	朝食	夕食	風呂	サービス	立地	施設	部屋
全文データの SVM	0.879	0.861	0.841	0.829	0.886	0.810	0.852
評価語データの SVM	0.885	0.878	0.868	0.876	0.873	0.816	0.878
全文データの Random Forest	0.878	0.877	0.869	0.844	0.895	0.825	0.853
評価語データの Random Forest	0.879	0.881	0.873	0.869	0.895	0.832	0.871
全文データの BERT	0.928	0.913	0.916	0.931	0.918	0.908	0.910
評価語データの BERT	0.895	0.870	0.878	0.857	0.879	0.831	0.869

に大きな差があることがわかる. 先ほど述べた 2 つの要因が重なることで, 全てのカテゴリにおいて全文データの正解率を下回ったと考えられる.

表 5: 全文データと評価語データでのトークン数の平均

	朝食	夕食	風呂	サービス	立地	施設	部屋
全文データ	23.4	23.9	24.2	26.8	23.1	26.1	24.5
評価語データ	5.1	5.0	4.9	5.1	4.6	5.1	5.1