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, 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 SVM と Random Forest 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. なお, 文章のベクトル化には TF-IDF 手法を用いた. 2 つの場合に分けた目的は, 学習データを評価語のみに絞ったとしても, 絞らない場合と比較して同等以上に分類できるのかということを確認するためである. 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.

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
評価語以外も含むデータの単語数	8213	7448	6615	10943	5915	9815	7109
評価語のみのデータの単語数	1214	1122	1050	1084	873	1461	721

表 1 の単語数についての 2 種類のデータを用いて SVM と Random Forest 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. なお, SVM と Random Forest のハイパーパラメータは Grid Search によって最適化した. 表 2 に評価語以外も含むデータと評価語のみのデータで 2 値分類をした正解率を示す.

表 2: 評価語以外も含むデータと評価語のみのデータで 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
評価語以外も含む SVM	0.879	0.861	0.841	0.829	0.886	0.810	0.852
評価語以外も含む Random Forest	0.878	0.877	0.869	0.844	0.895	0.825	0.853
評価語のみ SVM	0.885	0.878	0.868	0.876	0.873	0.816	0.878
評価語のみ Random Forest	0.879	0.881	0.873	0.869	0.895	0.832	0.871

表 2 からは, 評価語のみのデータを用いた 2 値分類のほうが, より高い正解率で分類できていることがわかる. 評価語以外の単語が分類には不要であるとは言えないが, 評価語のみであっても同等以上の精度で分類することが確認できた.

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