意見情報の推定における特徴量選択についての考察

情報科学科メディア情報コース　藤巻直也　　指導教員：山村 毅

1　はじめに

近年では、インターネットの普及により、Twitter などの SNS や Amazon などの通販サイトのレビューなどで様々な人々が自分の意見を投稿する機会が非常に増えている。こういった文章から、意見情報を取り出すことが出来れば、ものごとに対する人々の考えや、製品や、作品についての生の評価を得ることができ、極めて有用であると考えられる。ただ。これらの文章は、通常莫大であるため、機械で取り扱うことが前提となる。

本研究では、意見情報推定において、どのように特徴量を選択すればよいかを考察する。実験の対象とするデータには新聞社説を用いる。新聞社説とは、各社が最新のニュースや時事問題に解説や主張を加えた記事のことである。新聞社説から意見情報を得ることは日々の政治問題や国際問題、経済問題等の世の中の変化を追っていく上で非常に有効であると考えられる。

2　意見情報の推定方法

本研究で用いる意見情報推定は、三浦ら[1]が用いていたものと同様のものを用いる(表 1)。

表 1:分類カテゴリ

肯定	支持や同意を示しているもの
期待	良い結果や状態の予期を示しているもの
疑問	疑念を示しているもの
助言	忠告・アドバイスしているもの
否定	反対を示しているもの
非メッセージ	上記5カテゴリに属さないもの

意見情報の推定は、これらの情報を人手でラベル付けした文章データを用いて、文(の特徴表現)から意見情報への写像(分類器)を求めることによって行うことが出来る。この際、どんな特徴表現を用いるのか、どんな分類器を用いるのかが、推定性能を左右する鍵となる。

本研究では、分類器にナイーブベイズ分類器を、特徴表現にBag-of-Words表現を用いるが、特徴の選択の仕方が、推定性能にどう影響するかについて実験・考察を行う。

3　特徴量の選択手法

Bag-of-Words表現は、文を単語の集合で特徴表現する方法である。語順情報が失われるものの、簡便であるため、自然言語処理の分野では広く使われている。自然言語では語彙が莫大であるため、又、単語の中には、意見情報を推定するのに有効でないものも存在するため、Bag-of-Words表現に用いる単語を取捨選択する必要がある。

本研究では、文全体でなく、1)係り受け解析を用いて、修飾句を削った(核となる)文に変換したもの(短縮文)を用いる方法(短縮処理)、2)文中の括弧表現を除去したもの(括弧削除文)を用いる方法(括弧削除処理)で語彙を減らすことで、Bag-of-Words表現に用いる特徴(単語)の削除を試みる。又、3)情報利得値を用いて全体としてではなく各意見情報推定に有用なもの(カテゴリ特徴)を特徴(単語)として選択する。

4　実験方法

毎日新聞2006年度前半の社説344記事9492文に対して人手で意見情報を付与したもの[1]をデータとして用いて、3. で述べた短縮処理、括弧削除処理、カテゴリ特徴の効果を調べた。具体的には、これらの処理の有無によって、意見情報推定を行った場合の正解率、F値マクロ平均を比較した。これらの値の評価には、10分割交差検定を用いる。

5　実験結果とまとめ

各手法の最大の正解率及び F 値マクロ平均(Fmacro)についての結果を表 2 に示す。表中の short は短縮処理の有無、kakko は括弧削除処理の有無、mul はカテゴリ特徴の有無を示しており、org はカテゴリ特徴でなく、全カテゴリに有用な特徴(単語)を用いた方法を指す。

表 2:各手法の最大の正解率と F 値マクロ平均

番号	記号	正解率	Fmacro
1	org	75.9	58.2
2	org-short	78.4	61.9
3	org-kakko	76	58.2
4	org-kakko-short	78.4	62.2
5	mul	77.4	61.4
6	mul-short	79	63.7
7	mul-kakko	77.5	61.5
8	mul-kakko-short	75.7	57

この表から、短縮処理及びカテゴリ特徴については、正解率、F 値マクロ平均が、それらを用いていない場合に比べて向上しているので(例えば、表中 1 と 2、5 と 6、1 と 5)有効であるが、括弧削除処理については、有効ではない(例えば、表中 1 と 3、5 と 7)と言うことが出来る。全体としては、短縮処理とカテゴリ特徴を行った時が、一番性能が良かった。

今後の課題として、特に判定精度の悪かった「肯定」の判定精度向上が挙げられる。これにはカテゴリ毎のデータ数をある程度統一することや、特徴として Bag-of-Words ではない単語間の依存関係を考慮したものを用いることなどが考えられる。

参考文献

[1] 三浦弦太, 石井絵理香, 山村毅: ″ナイーブベイズ分類を用いた新聞社説の感性的情報の分類″, 電気関係学会東海支部連合大会講演論文集, L1-5, 2014