通販サイトのレビューを用いた日本語の感情分析

Sentiment Analysis of Using Reviews on Online Shopping Sites

劉雨楠
指導教員　　亀田　弘之

東京工科大学大学院　バイオ・情報メディア研究科
コンピュータサイエンス専攻　亀田研究室

キーワード: 自然言語処理,感情表現,ベクトル化

研究背景と動機

NLP(自然言語処理Natural Language Processing)は人間が使っている自然言語をコンピュータに処理させる技術である。例えばGoogleの研究者が開発したWord2vecは数十万の単語を200次元くらいの空間内にベクトルとして表現することで、単語同士の類似度が計算できて、単語の意味を捉えられるようになった［1］。これらの技術を用いて、ニュース、報道のような客観的な自然言語文に対する分類と認識の精度が大幅に向上した。

しかし、主観的な文章には人の感情が包み込まれており、それをコンピュータが理解することはまだ困難である。人の気持ちと意見を分かるために、一般的には文章に現れた各単語のポジティブ・ネガティブの程度を判断してから、文章全体の感情の把握をする。さらに文章の態度を評価することによって、感情の所有者、対象者と感情のタイプを検出できる［2］。この中で、単語の極性辞書が利用される。

英語では、Linguistic Inquiry and Word Count、Senti Word Netなどの極性辞書が作成されている。日本語にもPNTable、日本語評価極性辞書などの辞書が存在している。そして人手で単語の極性が付与されることも多い。しかしこのような方法は効率が低く、辞書の単語数量が1万から2万ぐらいでは、十分とは言えない。一方、日本語の表現方式は多様で、ネットワーク上でも新しい表現がどんどん増えているので、単純に編集した極性辞書の単語を用いて感情分析を行っても、汎用性は悪い。

研究目的および研究意義

米国研究者がこういう実験を行った。Twitter から選挙とか消費者指数に関するユーザーのコメント内容を分析したところ、その結果が現実の表現とほぼ同じ傾向であった。本研究では言語の表現と感情を分析するにあたり、統計的な結果よりもっと高い精度と汎用性を求めたいので、通販サイトのレビューを対象として解析する。その理由は、SNS 上のコメントと比較して、通販サイトのレビューには単純に商品に対する気持ちが表現されており、政治とか経済とかの外部事件に影響を受けにくく、安定性高い感情表現と考えられるからである。

本研究により、日本語の主観的な感情表現を解析して、意見マイニングの効果が向上されることを目指す。

研究方法

1.ECサイトのユーザーレビュー取得

日本市場には、アマゾン、楽天、ヨドバシカメラなどの EC サイトが存在する。アマゾンは2017年の売上高が約1.2兆円であり、引き続きネット販売ランキングが第一位、アマゾンの商品ページのレビュー情報を取得し利用する。

2.自然言語の前処理

まずはノイズクリーニング。取得し

たレビュー情報は構造化されていない文字列であり、ウェブサイト特有のタグ、記号などを削除する。

次は形態素解析。本研究では MeCab と Neologdを利用する。MeCab とは日本語形態素解析ライブラリである。Neologdとはウェブサイトから得た新語に 対応して、毎週更新される MeCab 用のシステム辞書である、これを追加すれば、レビュー情報に現れる最新の単語にも対応した形態素解析ができる。

最後はストップワード除去。これらの単語は出現頻度が高い割に役に立たず、計算量や性能に悪影響を及ぼすため除去する。除去する方法は、ストップワード辞書を利用する。

3.ベクトル表現化

レビュー情報を直接Word2vecなどの手法で処理すると、悪い結果になってしまう可能性が高いと考えられる。例えば“値段高い”と“解析度高い”という表現を例にして、一般的に“解析度高い”はポジティブ“値段高い”はネガティブである。すなわち商品の属性により、言語“高い”の極性は逆になる。表現を正しく区別できるために、本研究ではFastTextを利用してデータを処理する。

FastTextはFacebookが開発した単語のベクトル化とテキスト分類をサポートした機械学習ライブラリであり、subword分割という仕組みを採用している、単語の中の一部が近いと近くなる特性をもつ［3］。これを利用すれば、Word2vecとは異なり商品の属性と商品の評価極性を合わせて評価できると予想される。

4.ベクトル表現データと感情を繋ぐ

単語の出現が評価ポジティブかネガティブかを直接に反映するとは必ずしも言えない。本研究ではこれを回帰問題とみなし、ユーザーの評価数値を参考にしてベクトル空間の各単語のベクトルデータを繰り返しトレニンーグして、もっとユーザーの感情を反映できるベクトル値を取得する。

一般的な回帰問題と違って、レビューに単語の出現順番も結果に影響を与える。例えば“xx は高いけど、品質は良い”のように、前者より後者のほうを表現したい場合があるので、CNNではない LSTMなどを利用する。

検証実験

レビュー内容を取得したら、一部分を上記の研究方法で分析する。分析した結果のベクトルを用いて、残っているデータで検証する。

検証データのレビューに対して、システムが予測した評価点数と実際にユーザーが評価した点数を比較する。最後に感情分析の効果が良いか悪いかを統計手法を用いて評価する。

さらにその結果の汎用性を把握するために、SNS とブログのコメントデータも取得して検証してみる。

参考文献

[1] Goldberg,Yoav;Levy, Omer. "word2vec Explained:Deriving Mikolov etal.'s Negative.Embedding,Method".arXiv:1402.3722.

[2] 鳥倉広大・小町守・松本裕治:“Twitte·を利用した評価極性辞書の自動拡張”,言語処 理学会年次大会発表論文集, 18th ROMBUNNO.A3-2 2012.

[3] Armand Joulin. Edouard Grave. Piotr Bojanowski Tomas Mikolov:“Bag of Tricks for Efficient Text Classification″, Facebook AI Research. c02017 Association for Computational Linguistics, April 3-7, 2017.