D-38 シード表現を用いたマイクロブログに存在する観光情報の極性判定

前田 裕† 井上 俊右‡ 嶋田 和孝† 遠藤 勉†
†九州工業大学大学院情報工学府 ‡九州工業大学情報工学部

1 はじめに

近年，観光は様々な地域において基幹産業の 1 つになっている．観光の活性化は，地域の活性化に繋がると考えられる．そのような環境下で，Web に存在する多くの観光情報の分析は重要な役割を持ってきている．

本論文では，Twitter と呼ばれるマイクロブログ上に存在する観光情報を対象にして評判分析を行う．Twitter は，他のメディアと比べ，現地性やリアルタイム性，強い主観情報といった特徴を持っている．こうした特徴は，観光的なイベントや事象の分析に対して有用であると考えられる．

評判分析の方法としては，極性判定を行う．極性判定とは，その文が肯定的な意見 (P) なのか，もしくは否定的な意見 (N) なのかを分類することとする．自動的に獲得された訓練データに基づく機械学習を適用し，その有効性を検証する．

2 提案手法

テキスト極性判定については，様々なアプローチがある．その一つは機械学習を用いた手法である．Pang ら [1] は映画のレビュー記事を対象とした極性判定において，いくつかの機械学習に基づく手法の有効性および比較を行っている．機械学習を用いれば，高い精度が見込めるが，一方で高い精度を得るためには十分な訓練データが必要となる．しかしながら一般に，訓練データとなるタグ付きコーパス作成のための作業は高コストである．

本論文では，人手によるタグ付きコーパスなしで，高精度な極性判定を行う枠組みを導入する．Turney[2] は，語やフレーズの極性値を自動的に算出するために，“ excellent ”や“ poor ”といった肯定および否定の極性をよく表す語との共起度を利用する手法を提案した．本手法もこれに倣い，シード語 (種語) から自動的に訓練データを作成する．

まず，大量の tweet(Twitter 上での投稿文) からシードを含む tweet を抽出する．シードと共起した単語群を肯定の表現もしくは否定の表現として仮定し，機械学習に適用する．なお，本論文では機械学習器として SVM を用いる．

3 実験・考察

本節では提案手法の有効性を検証する．評価データとして，人手で 116 tweets をラベル付けした．そのうち，P の数は 64 であり，N の数は 52 であった．

本手法はシードを用いて訓練データを獲得する必要がある．そのためのタグなしデータとして，100 万 tweets を利用した．この 100 万 tweets に対して，P または N のシードがマッチしたものがそれぞれ P または N の訓練事例である．

シードは P や N を象徴する単語や顔文字，記号を主観的に選択した．その内いくつかのシードパターンにおける極性判定結果を表 1 に示す．なお，複数のシードの場合は，それらのうち，いずれかが含まれる tweet を訓練データとして用いるものとする．

表 1: シード毎の極性判定の精度

P のシード	N のシード	精度
♪	ToT	0.61
♪	残念	0.68
♪	orz	0.73
♪	orz，残念	0.72
♪	orz，＃），ない	0.75
良	orz，＃），ない	0.64
♪，＾▽＾)	orz，＃），ない	**0.78**
♪，＾▽＾)，≧▽≦)	orz，＃），ない	0.77

様々なシードパターンを試したところ，P のシードに「♪，＾▽＾)」，N のシードに「orz，＃），ない」を用いた場合に最も高い判定精度 0.78 を得た．精度を変化させるいくつかの原因を以下のように考察する．

「ToT」をシードに用いると極端に精度が低くなった．これは，「ToT」という顔文字が嬉し泣きや感涙といった Positive 表現としても用いられることが比較的多いためであると考えられる．

「orz」は人間が跪く様を表したアスキーアートであり，落胆，失意，挫折の心理状態を表現する際に用いられる．「残念」も似たような意味を持つと考えられるが，「残念」と「orz」の精度に着目すると「orz」の方が有効に働いた．これは「残念」に比べ「orz」の方がより多様な Negative な発言を網羅できるためだと考えられる．例えば「orz」を含む tweet では，「食べ過ぎた orz」「起きれるかな…orz」のように，主観的疲労感や微弱な否定的感情を表現する際にも用いられる．それに比べると，「残念」を含む tweet が表現し得る範囲は限られている．

また，Twitter の文書には様々な文体が存在し，文体のドメイン毎に表現が異なる．たとえば，硬派な文体を使うユーザの間では顔文字はあまり用いられない．そのため，シードとして用いる語と，極性判定の対象となる文書の文体が一致していなければ期待する精度が得られない．

4 おわりに

本実験では Twitter 上の観光性のある文書に対して極性判定を行った．今後は感情や文体のドメイン毎に分割統治的に極性判定を行うことで精度の向上を目指す．

参考文献

[1] B. Pang, L. Lee, and S. Vaithyanatan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of EMNLP, pp. 79-86, 2002.

[2] P. D. Turny. Thumb up? or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL, pp. 417-424, 2002.