6T-06

感情辞書を用いたWebテキストの多言語感情抽出

藤平　啓汰†

崇城大学　情報学部　情報学科†

1. はじめに

SNS に代表される Web サービスの普及に伴い，トレンド予測や評判分析が国や地域を問わず行われるようになった．Web ブログやツイートなどの Web テキストに含まれているユーザの感情を抽出し利用することで，より詳細な分析と予測を行うことが可能となる．しかし，テキストの構造は言語によって異なるため，単一の言語のみを想定した感情抽出手法を他の言語へ用いても，正確な感情抽出を行うことはできない．そのため，多言語に対応する感情抽出手法の研究が進められており，機械翻訳と機械学習を組み合わせた手法が主に研究されている[1,2]．これらの研究では，学習に使用する各言語のデータセット不足とそれに伴う分類精度の低下が課題となっている．

本研究では，目的言語の感情辞書とルールにもとづく多言語対応の感情抽出手法を提案する．既存手法では，各言語のデータセット，特にラベル付きデータが必要となる．一方，ラベル付きデータとして，目的言語のデータのみを必要とする点が本研究の特徴であり，データセットが少ない言語に対しても精度を落とすことなく，感情分類を行うことが可能となる．

2. 多言語対応の感情抽出手法

本研究で扱う感情の値は，ポジティブ，ネガティブ，ニュートラルの3種類とし，Webテキストを対象に感情抽出を行う．以下では，提案する多言語対応の感情抽出手法について解説する．

2.1 提案手法の概要

提案手法は，形態素解析，テキスト中の各単語の感情値算出，各単語の感情値をもとにしたテキストの感情値算出の3段階で構成される．

図1は，単語の感情値を算出するアルゴリズムである．また，図1中の表現対候補の獲得方法は，那須川らが考案した手法を参考にしている[3]．まず，原言語テキストを形態素解析後，単語Aの分散表現と原言語コーパスを用いて，単語Aと類似度の高い原言語の単語 10 個のリスト（原単語リスト）を作成する．次に，汎用辞書を用いて原単語リストを目的単語リストへ変換する．この目的単語リストの各単語から分散表現の平均値を求め，分散表現の平均値に対して類似度の高い目的言語の単語 10 個のリストを作成する．このリストがもとの単語Aの表現対候補リストとなる．続けて，感情辞書より各表現対候補の感情値を参照する．その結果得られるポジティブの個数，ネガティブの個数，ニュートラルの個数を要素とする3次元の感情値ベクトルを用いて単語Aの感情値を算出する．例えば，感情値ベクトルが（8,0,2）である場合，ポジティブを表す単語の個数が最大であるため，単語Aの感情値はポジティブとなる．このアルゴリズムを形態素解析後の各単語に適用し，テキスト中の単語がどのような感情を表現しているか判断する．

次に，テキスト中の各単語が持つ感情値から，単語の感情値算出と同様の方法で感情値ベクトルを作成し，テキストの感情値を算出する．

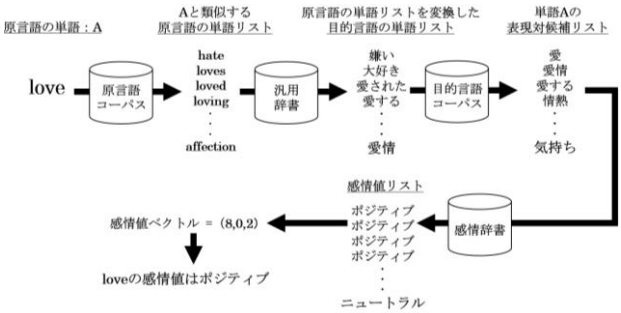

図1　単語の感情値算出アルゴリズム

2.2 実装

感情抽出の第一段階である形態素解析は，Stuttgart 大学の Helmut Schmid 博士によって開発された形態素解析ツール「TreeTaggr」を用いて行う．この際，普通名詞，形容詞，副詞，一般動詞以外の品詞は形態素解析後に除去する．次に，第二段階である単語の感情値算出における単語間の類似度計算は，分散表現・テキスト分類学習ライブラリの「FastText」を用いて行う．分散表現学習モデルは，FastText の Web サイトで公開されているものを使用する．また，本研究では，感情辞書として高村らの単語感情極性対

Multilingual Sentiment Analysis for Web Text using Sentiment dictionary

†Keita Fujihira
Department of Computer and Information Science, Sojo University, Japan.

応表[4]を使用し，辞書中の−1 から+1 の感情数値をもとに感情値を決める．ポジティブ，ネガティブ，ニュートラルを分類する感情数値 x のしきい値は以下のように設定する．

$$\begin{aligned} Negative &: -1.0 \leq x \leq -0.40 \\ Neutral &: -0.40 < x < -0.37 \\ Positive &: -0.37 \leq x \leq 1.0 \end{aligned} \quad (1)$$

3. 評価実験

提案手法を用いて，英語とドイツ語のツイートデータを対象とする感情分類実験を行う．テストデータには，英語ツイート 1000 件，ドイツ語ツイート 350 件を使用する．また，提案手法と比較するため，Natural Language Toolkit（NLTK）の感情分析パッケージと Google Cloud Platform（GCP）の感情分析ツールでも同様の実験を行う．

3.1 実験結果

図 2 は英語ツイート，図 3 はドイツ語ツイートの感情分類における正解率，各感情値の適合率の平均，再現率の平均，F 値の平均を比較したものである．英語ツイートの感情分類結果では，提案手法の正解率が GCP を上回っている．しかし，それ以外の評価指標はいずれも比較手法を下回っている．特に，提案手法の平均再現率は，比較手法に対して 10%以上下回っている．一方，ドイツ語ツイートの感情分類結果では，全ての評価指標で提案手法が NLTK を上回り，正解率と F 値については GCP も上回っている．

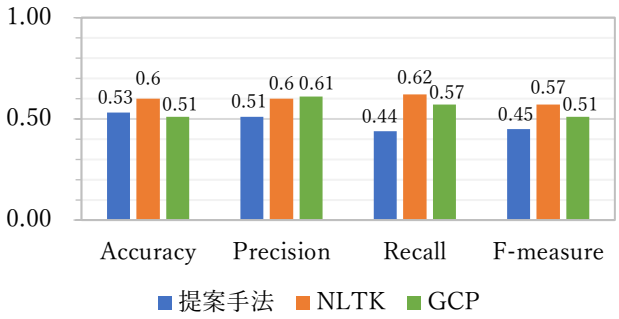

図 2　英語ツイートの感情分類結果

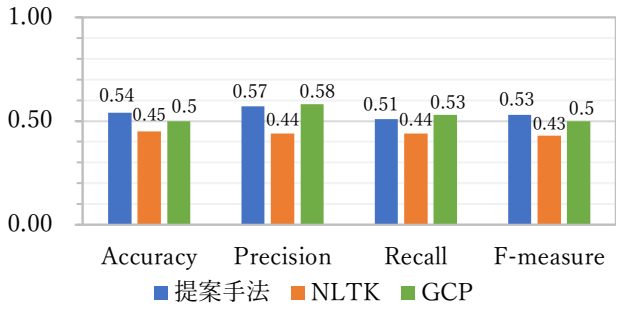

図 3　ドイツ語ツイートの感情分類結果

3.2 考察

どの評価指標の値も 0.5 付近となっており，既存手法と同等の精度であった．今回の実験では，感情辞書に収録されていない単語をニュートラルな単語と解釈していた．また，表現対候補は機械学習を用いてコーパスから抽出しているが，実際に抽出した単語には，“素晴らしかっ”や“大スキ”のように脱字や表記ゆれがあった．感情辞書には“大好き”又は“だいすき”といった表記しかないため，これらはポジティブな単語であるにもかかわらず，ニュートラルな単語と判定されていた．このような感情値の設定方法が，分類精度の低下につながったと考えられる．一方，言語の違いによる分類精度の大幅な増減はなく，多言語対応の感情抽出手法という点では，応用可能な手法であることを示すことができた．

4. おわりに

本研究では，言語の壁を越えた評判分析やトレンド予測を行うため，感情辞書とルールにもとづいた多言語対応の感情抽出手法を提案した．英語とドイツ語のテキストを対象とした感情分類実験では，実用化には不十分な精度ではあるものの，多言語のテキストからの感情抽出へ応用可能であると検証することができた．

今後は，分析対象となるテキストの種類を増やし，提案手法がどのようなテキストの感情抽出を得意または不得意としているのか検証していく必要がある．さらに，中国語やアラビア語など，テキスト構造が英語とは異なる言語に対して提案手法が有効であるか検証していきたい．

参考文献

[1] Alexandra Balahur, Marco Turchi, Comparative experiments using supervised learning and translation for multilingual sentiment analysis, Computer Speech and Language, Vol.28, Issue 1, January 2014.

[2] Ethem F. Can, Aysu Ezen-Can, Fazli Can, Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data, arXiv preprint arXiv:1806.04511, 2018.

[3] 那須川哲哉，Daniel Andrade，海野裕也，村松祐希，山本和英，言語横断テキストマイニングのための翻訳対抽出，言語処理学会第 15 回年次大会発表論文集，pp.108-111, 2009.

[4] Hiroya Takamura, Takashi Inui, Manabu Okumura, Extracting Semantic Orientations of Words using Spin Model, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005), pp.133-140, 2005.