

Title	構文構造を用いた感情極性分類の精度向上
Author(s)	中山, 貴樹
Citation	
Issue Date	2012-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/10421
Rights	
Description	Supervisor: 東条敏, 情報科学研究科, 修士

構文構造を用いた感情極性分類の精度向上

中山 貴樹 (1010045)

北陸先端科学技術大学院大学 情報科学研究科

2012年2月6日

キーワード: 機械学習, クラスタリング, 感情極性分類, テキスト分類.

近年、近年、インターネット上の大量のテキストが、様々な物事に関する情報を得るための重要な情報源となっている。テキストから著者の感情に関する情報を得るための技術として、テキスト感情極性分類と呼ばれる技術があり、企業による新製品の評判分析などに用いられている。感情情報処理技術とは、テキストを解析し包括的な調査をすることにより人の感性を客観的に評価することを目的とした技術である。その技術の1つとして感情極性分類があり、これを用いて新製品のサーベイなど様々な場で用いられている。感情極性分類は、人の手によって書かれたある対象について述べられている文章などを対象として、その文章の中に存在する言語表現が好意的、非好意的な極性であるかを自動的に判定する技術である。好意的または非好意的とは人が書いた文章を自動的に好意的または非好意的かに分類する為には、テキスト分類での Bag-of-Words モデルを使うのが一般的である。Bag-of-Words モデルとは、各単語が素性空間における一つの次元であると仮定する文書表現である。Bag-of-Words モデルを用いた感情極性分類の既存の手法は、文書中に出現する単語を素性とした機械学習によるアプローチの1つである。しかし、Bag-of-Words モデルでは、単語の表層的な感情極性を独立にとらえることしかできないため、高い精度を達成することは難しい。文書の中には、ほとんどの場合、好意的または非好意的な単語が複数含まれて書かれている。そのため、文書全体としての感情極性を正確に測ることが困難になる。さらに、逆説や否定などにより極性が反転することもある。

複数の単語列による表現の感情極性は、例えば“ 美しくない ”という複数語では “美しい” だけでは好意的な極性だが、“美しくない” という全体を考慮した場合、非好意的な極性として捉える必要がある。また、文書全体を捉えるために、重要な要素になるものの1つとして接続詞がある。接続詞は、前後の文脈の関係を表す語で、品詞の一つである。接続詞によって、文書は前後の文で感情極性が変化を捉えることが、評価文分類で重要であるといえる。そこで本研究では、構文構造での係り受けの關係に着目する。感情極性単語が係り受け元となり、係り受け先に存在する単語によって、感情極性単語の意味と単語の極性が反対になるという仮定のもと、その組み合わせを機械学習に用いる。この組み

合わせを学習することにより，感情極性単語の意味と極性が反転したかの有無を知ることが可能になる．さらに，この結果を用いて，Bag-of-Word モデルでの抽出結果との差を利用することで，各文書の本来の感情極性単語の総数を計ることが可能である．ユーザーによって書かれた文書集合に対し，Cabocha を用いることで，その文書の構文構造を解析する．その解析結果を用いて，感情極性単語と係り受け先の単語のセットを抽出する．それを学習データとし，サポートベクターマシンを用いて学習する．そして，感情極性分類に使用する文書に適応することで，各文書の感情極性単語の意味と極性の反対になる回数を抽出することが出来る．この結果を，Bag-of-word モデルを用いて抽出した感情極性単語の総数に影響を与えることで，本来文書が持つ感情極性単語の総数と仮定できる．そして，我々の実験は次のような手法を用いて実験を行なった．N-gram モデルを用いて，ユニグラム，バイグラム，トライグラムの3種類を使用し素性ベクトルを抽出した学習データ．次に，感情極性単語の好意的単語と非好意的単語の2種類の総数を数え上げる．この総数を，N-gram モデルと組み合わせた学習データとした．最後に，係り受け情報によって得られた感情極性単語の総数と N-gram モデルを組み合わせたものを学習データとした．これらの学習データをサポートベクターマシンの用いて，感情極性文書分類を行う．評価実験は，2010年に公開した楽天トラベルの内，ユーザ評価及びユーザレビューを使用した．先の手法を用いて，対象を文書の感情極性分類を行ない，各実験結果の比較を行う．実験の結果，手法の中では，Trigram と係り受け情報を組み合わせた学習データを用いたものが高い分類精度を示した．これらの結果から，係り受け情報を考慮し感情極性単語の意味と極性が反対になる点に着目することが，感情極性分類の精度向上に有用であることがわかった．