1S-08

テキストデータを対象とした文章における特有表現に着目した意見抽出方式

池上　藍羽†　　石井　雄大†　　北　椋太†　　中西　崇文†
武蔵野大学　データサイエンス学部　データサイエンス学科†

1. はじめに

近年，新聞の電子化や，SNS の発達などにより，Web 上には多くのテキストが数多く見られるようになり，テキストマイニングなどにより，これらの活用が進んでいる．

一般に，文章の構成は主に意見と事実の 2 つからなり，文章の形式により，意見と事実の割合や書かれている内容は異なり，意見が多い記事は発信者側の意向に沿った内容となる傾向にある．そのため，これらを活用するためには，様々なテキストデータを対象とし，テキストの種類ごとに意見を抽出することが重要である．

意見抽出に関する先行研究として，小林ら[1]の研究が挙げられる．小林ら[1]は，意見を「ある対象(商品，サービス，会社など)もしくは対象のある側面に対する，記述者の主観的な評価を表す記述」と定義し，「意見を表す典型的な文型の一つで，かつ個々の要素について定義が可能と思われる〈対象，属性，評価値〉の 3 つ組」で表すことができる意見を抽出した．この際，「属性とはある対象(商品)のある側面を表す表現を指し，評価値はその属性の値や記述者の好悪に関する心的な態度を表す表現を指す．」としている．

本稿では，テキストデータを対象とした文章における特有表現に着目した意見抽出方式について示す．本方式では，ニュースやレビューに関する記事を対象とし，意見が述べられている文における特有表現や心情表現に着目し，意見を抽出することを可能とする．本稿では，「思う」，「感じる」，「考える」など主観的な考えや感想を述べる際に用いられる表現と，それらの表現における類語を意見が述べられている文における特有表現と定義する．また，長岡技術科学大学の自然言語処理研究室が公開している日本語感情表現辞書[2]を心情表現として用いた．

本方式は，様々なテキストデータから意見を抽出し，テキストの形式ごとに意見の割合および，意見の主なトピックについての可視化をする．これにより，文章内の意見と事実を分別し，その文章内にどのような考え方が込められているのかを客観的に把握することが可能である．

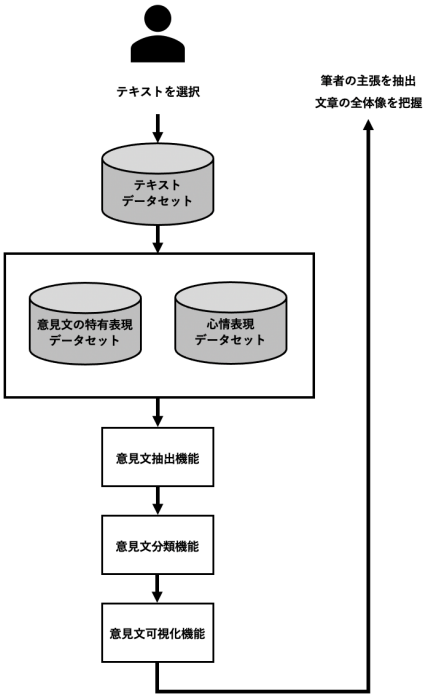

図 1：本方式の全体像

2. 文章における特有表現に着目した意見抽出方式

2.1 提案手法の全体像

本節では，本研究における提案手法の概要を述べる．提案システムの全体像を図 1 に示す．

本システムは，テキストデータセット，意見文の特有表現と心情表現データセット及び，意見文抽出機能，意見文分類機能，意見文可視化機能で構成される．

2.2 テキストデータセット

テキストデータセットは，新型コロナウイルスに関するニュース記事とレビューに関する記事を 100 件ずつ収集したものである．

2.3 意見文の特有表現と心情表現データセット

意見文の特有表現データセットは，「思う」，

Opinion extraction methods focusing on specific expressions for text data
Aiha Ikegami†, Yuta Ishii†, Ryota Kita†, Takafumi Nakanishi†
† Musashino University, Department of Data Science

「感じる」，「考える」など主観的な考えを述べる際に用いられる表現と，それらの類語を収集した．また，心情表現データセットは，日本語感情表現辞書[2]を用いた．

2.4 意見文抽出機能

意見文抽出機能は，テキストデータと意見文の特有表現と心情表現データセットをマッチングすることで，筆者の意見が述べられている文を抽出する機能である．

2.5 意見文分類機能

意見文分類機能は，抽出された意見を肯定的な意見と否定的な意見に分類する機能である．本手法では，意見文を分類する際に oseti[3]を用いた．oseti は，日本語評価極性辞書[1][4][5]を用いて，単語や文がポジティブな内容か，ネガティブな内容か判定するライブラリである．

本方式では，抽出された意見文を形態素解析し，形容動詞語幹を用いて肯定的な意見と否定的な意見を分類する．

2.6 意見文可視化機能

意見文可視化機能では，WordCloud を用いて，肯定的な意見と否定的な意見を可視化する．WordCloud で可視化することにより，文章内において，どのようなトピックが多く見られたか，視覚的に表現することを可能とする．この機能により，ポジティブな表現を含む文と，ネガティブな表現を含む文において，多く見られた意見が比較可能となる．

3. 実験結果

本実験では，ニュースとレビューに関する記事を対象とし，テキスト間における抽出される意見の割合の差異を明らかにする．また，意見文分類機能により，肯定的な文と，否定的な文に分類し，意見文可視化機能により，それぞれ可視化をし，抽出された意見を比較する．実験結果は表 1 と図 2 に示す．

抽出された意見の割合において，表 1 にあるように，意見のレビューに関する記事はニュース記事よりも多く意見が抽出された．ニュースは客観性が求められるのに対して，レビュー記事は個人的な文章であり，対象物に対しての評価や感想を述べられていること多いため，このような結果になったと考えられる．

意見文の可視化において，意見がより多く抽出されたレビュー記事を用いた．図 2 より，肯定的な意見では，「高級」や「安全」，否定的な意見では，「可能」や「必要」の出現頻度が高かった．否定的な意見に「可能」や「必要」が多く見られたことから，レビュー記事において，対象が不足している要素や，要望が多く述べられていることがわかる．

表 1：テキスト別抽出された意見の割合

文章内容	意見文の割合
ニュース	17.0262
レビュー	28.3083

図 2：レビュー記事におけるポジティブな意見(左)とネガティブな意見(右)

4. おわりに

本方式では，意見文における特有表現に着目した意見抽出方式を実現した．意見とは，筆者の主張であるため，意見を抽出することで文章の核となる要素を抽出可能とする．

また，本方式は，様々なテキストデータに対して有効であり，小説から人物の心情を抽出することも可能である．あらゆる記事における筆者の主張や，小説における心情表現を抽出することにより，その文章において重要な要素の抽出を可能する．

参考文献

[1]小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：意見抽出のための評価表現の収集，自然言語処理，Vol.12，No.3，pp.203-222(2005)．
[2]長岡技術科学大学 自然言語処理研究室：日本語感情辞書，〈http://www.jnlp.org/SNOW/D18〉(参照 2020-11-23)．
[3]oseti，〈https://pypi.org/project/oseti/〉(参照 2021-1-3)．
[4]東山昌彦，乾健太郎，松本裕治，述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，pp.584-587，2008．
[5] 日本語評価極性辞書，〈http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FJapanese%20Sentiment%20Polarity%20Dictionary〉(参照 2021-1-3)．