7W-07

Twitterにおける意見文の抽出と要約

曹 金煜† 杉本 徹‡

芝浦工業大学大学院 理工学研究科† 芝浦工業大学 工学部‡

1. 研究背景と目的

近年，インターネットの発展によりソーシャルメディアで人々の投稿が膨大な数となっている．Web 上で様々なトピックスに対する意見が述べた投稿が多数ある．例えば，Twitter 上でホットイベントに対する感想，食べログ上でグルメのレビュー，アマゾン上で商品の評価など色々な情報がある．これより，ある話題に対して皆がどのような感想や意見を持っているのかを知りたい場合がある．

先行研究として，池上ら[1]と小林ら[2]の研究を挙げる．池上らは意見文の特有表現と心情表現を用いて文章内の意見と事実を分別し単語ごとの可視化機能を作って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把握した。小林らは商品の意見を＜対象，属性，評価＞の 3 つ組の形で抽出した．この際，属性はある商品のある側面を表す表現として，評価は投稿者の好悪に関する心の態度を表す表現とした。

そこで，本研究では，Twitter に投稿されたツイート文から意見文を抽出し，要約する手法を提案する．本手法では，機械学習を用いて意見が含まれる文をまず抽出し，係り受け関係の分析により長い文を短縮することを目的とする．

2. 研究内容

2．1 提案手法

図 1 に研究の全般の流れを表す．

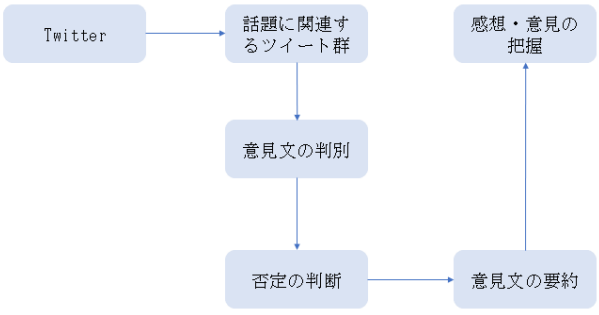

図 1 ：本手法の全体像

本研究では Twitter API を利用してツイート文を収集し学習コーパスを構築する．SVM を用いて機械学習を行って，ある話題に関連するツイート文の中から意見文を判別し抽出する．抽出した意見文を要約して，ある話題に対する皆の感想や意見を把握することができるようにする．

2．2 意見文の判別

本稿では投稿者自身の考えた主張，事件に対する感想や要望など，主観的に書かれた文を意見文と定義する．対象のツイート文から事実を述べる文と意見を述べる文の 2 種類に分類する．

2．3 否定の判断

意見文を短縮する前に，その文が否定の意味を持っているかを判別する必要がある．本研究では形容詞や助動詞の「ない」，接頭辞の「不・未・無」を否定のパターンとし，「仕方がない・…しなければならない」など特定の短句を肯定のパターンとして，ある意見文が否定の意味を持っているかを判断する．

2．4 意見文の要約

多数の意見文に含まれる感想や意見を簡単に把握できるようにするために意見文を要約する．要約の方法として池上ら[1]のように出現頻度が高い単語を求めて WordCloud で可視化する方法もあるが，抽出した意見の把握は見た目で主観的に偏る可能性がある．そのため，本稿では長いテキストを「何がどうした」のような短文に要約して，意見文の意思をもっと客観的に把握することを目指す．

3. 実験と結果

3．1 学習コーパスの作成

Twitter API を利用して，毎日のトレンドから話題語をランダムに選んでツイート文を収集した．一つのツイート文に対して，意見が含まれる場合にラベル 1，意見が含まれない場合にラベル 0 を人手で付ける．

ラベルを付けるときに意見文と判別する基準として，「〇〇速報」「…が現状」など事実だけが含まれる文にラベル 0 を付けて，「…と思う」「…ください」「…がおすすめ」，「…の方がいい」「…すべき」「…に賛成・反対」など感想や指示，禁止，推薦が含まれる文に意見文としてラ

Extraction and Summarization of Opinion Sentences on Twitter

† JINYU CAO ‡Toru Sugimoto

†Graduate School of Engineering and Science, Shibaura Institute of Technology

‡Faculty of Engineering, Shibaura Institute of Technology

ベル 1 を付けた.

最終的に 3003 個のツイート文を含むコーパスが作成できた.

３．２　SVM を用いた意見文判別の実装

３．２．１　特徴抽出

Sklearn ライブラリの CountVectorizer と TfidfTransformer 関数を利用して特徴要素を抽出して特徴単語リストを作成する. ここに TF-IDF で単語の重みを計算して特徴単語リストに対する特徴ベクトルリストを作る.

３．２．２　SVM の学習

３．１節の学習コーパスを用いて SVM の学習をすると正解率が 0.76, 適合率が 0.78, 再現率が 0.56, 特異率が 0.89, F 値が 0.65 になった.

正しく判別できた意見文と意見文でない文の例を表 1 に示す.

表 1：意見文の判別の例

ツイート文	判別結果
オミクロン株対応ワクチンを接種しましょう	1
そして今年やっと細胞性免疫をオミクロンは避けている論文が出ました	0

３．３　意見文の要約

長い意見文を短縮するために, 意見文の主語と対象格と述語を抽出して要約する. 日本語係り受け解析器 CaboCha を利用して,テキストを解析し, 文節ごとに分割することができる.

意見文に含まれる文節の中で「は・が・て・も」と繋がる文節を「主語」とし, 「を・が」と繋がる文節を「対象格」とし, 文の最後の文節を「述語」としてペアの形で抽出する. 抽出したペアを集計し, 出現頻度が高いペアを代表的な意見とする.

４．　評価実験

４．１　実験方法

話題語を 5 つ選んで関連するツイート文から提案手法を用いて意見文を抽出して,「主語＋述語」と「対象格＋述語」を抽出して評価を行った. 評価はアンケート形式で, 抽出したペアごとに以下の 2 つの観点で行った.

- 自然度（抽出した短句が自然かどうか）を 0～4 ポイントで評価
- 理解度（ペアを見て皆が持っている感想や意見を理解できるか）を 0～2 ポイントで評価

４．２　実験の結果

「主語＋述語」と「対象格＋述語」のそれぞれ 25 ペアの自然度と理解度を被験者 7 人が評価した結果を表 2 と表 3 に示す.

表 2：「主語＋述語」に関する被験者の平均得点

	最大値	最小値	平均値
自然度	3.14	0	1.26
理解度	0.57	0	0.17

表 3：「対象格＋述語」に関する被験者の平均得点

	最大値	最小値	平均値
自然度	3.57	0	2.45
理解度	1.57	0.85	1.20

評価点が高かったペアの例を表 4 に示す.

表 4：評価点が高かったペアの例

話題：「ワクチン」	（ワクチン，打つ）
	（ワクチン，接種する）
話題：「ヴィヴィアン」	（時代，終わる）
	（冥福，お祈りする）

５．　考察

抽出された「主語＋述語」のペアは「対象格＋述語」のペアより得点の平均値が低い. しかし「何がどうした」のように意見を要約するために,「対象格＋述語」のペアだけで抽出するのは良いとは言えない. 抽出されたペアにこの話題の重要な情報を追加して, もっと自然な短文に短縮することが必要である.

６．　まとめ

本研究では Twitter における意見文の抽出と要約の手法を提案して実験を行った. 今後の課題として, 意見文をより自然な文に短縮する手法の考案と学習コーパスや否定の判断の改良が挙げられる.

参考文献

[1] 池上藍羽, 石井雄太, 北椋太, 中西崇文, “テキストデータを対象とした意見抽出方式”, 情報処理学会第 83 回全国大会講演論文集, 2021(1), 541-542 (2021)

[2] 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)