研究指導 中澤 真 教授

質問投稿サイトにおける比較表現抽出手法

―対話型メディアのデータ形式に着目して―

佐藤 柚月

1. はじめに

近年，企業がマーケティングや商品開発のためにSNSなどの情報を活用することが増えている．総務省が行った2020年の調査[1]によると，マーケティングにSNSやブログのデータを活用していると答えた大企業の割合は約47％と約半数を占めている．クチコミ情報には消費者の生の声として，商品に対する魅力や不満などが書かれているため，ユーザのニーズや自社製品への改善点を把握するのに有用だからである[2]．さらに，競合製品との比較意見には，自社製品との優劣に関する比較表現も含まれているため，業界での立ち位置を読み取ることもできる．このようなクチコミの有用性が注目され，膨大なクチコミデータの中から効率的にマーケティングに利用できる情報を抽出するための研究が行われている[3][4].

しかし，これらの研究はすべてのソーシャルメディアに対応できているわけではなく，質問投稿サイトのような対話型データ[2]の情報から比較表現を十分な精度で抽出することはまだできていない．質問投稿サイトでは特定の商品に限らず情報共有が行われるため，競合製品との比較が行われやすい．そのため質問投稿サイト内のクチコミを分析することで，より効率的に有用な情報を収集することができると考えられる．

ゆえに本研究では対話型のデータに含まれる比較表現抽出に有効な特徴量の考察，および精度向上を目指す．

2. 情報共有サイトにおける比較表現抽出の課題

2.1 商品改善と比較表現

近年，インターネット上の情報共有サイトでは，商品に対するクチコミの投稿が盛んである[5]．それらの中には商品の評価や不満などが記述されているため，企業側も自社製品の改善やマーケティングに活用している．クチコミの情報には様々なタイプのものが存在するが，その中でも他社製品と自社製品の比較をしているクチコミは，企業にとって有用性が高い．例えば「商品Aは保湿力が高いけど，商品Bは低い」というような他社製品と比較しているようなクチコミから，商品Bの商品開発者は他社製品と比較した具体的な改善ポイントを読み取ることができる．

このような2つの商品を比較して，差異や優劣を記述している比較表現はマーケティング的に有用であるが，膨大なクチコミの中から比較表現だけを手作業で抽出するのは現実的ではない．このため，比較表現をシステムによって自動抽出する研究が取り組まれている[3][4].

2.2 比較表現抽出の手法

比較表現の抽出手法についてはいくつかの先行研究が存在する．石澤ら[3]は競合する商品に対し優劣表現をもつ比較表現を対象とした抽出手法を考案した．この手法ではあらかじめ肯定表現と否定表現を対応付けた辞書を作成し，それらに該当する表現を持つ文を比較表現として抽出している．山崎ら[4]は，テキスト中の比較対象がどのような関係にあるかを判定する手法を考案した．この手法では，比較表現によく見られる「より」や「ほうが」などの判別の手がかりとなる語(以下手がかり語とする)で，なおかつ判別精度の高い語のみを特徴量として用いた．その特徴量を用いて比較表現候補の抽出を行ったのち，並列表現を持つ文は比較表現ではないと仮定し除外することで，判別精度向上を図っている．

しかし，石澤らの研究では辞書に存在する語でしか評価表現を抽出できないため，検索漏れの可能性が高くなるという課題を残した．また，山崎らの手法も一般的なレビューサイトに対しては有効であるが，質問投稿サイトのような対話型データに対しては十分な精度を得られない可能性がある．

2.3 質問投稿サイトの概要と特徴

比較表現の抽出対象となる情報共有サイトには，質問投稿サイトのように対話を前提とした対話型データのものと，一般的なレビューサイトのように独話型が基本となるものがある[2]．代表的な質問投稿サイトとしては，「Yahoo!知恵袋[1]」や「教えて!goo[2]」のようなサービスがあり，自分の求める特定の情報を気軽に尋ねることが魅力となり，幅広いユーザに利用されている．これらの質問には商品の特徴や優劣を尋ねるものも多くあるため，マーケティングに有用な情報源となりうる．

しかし，対話型データの場合は質問と回答に文章が分割されているため，文体も独話型の場合と異なることが多い．このため，山崎らが用いていた手がかり語が含まれないケースも増えると考えられ，レ

[1] https://chiebuKuro.yahoo.co.jp/

[2] https://oshiete.goo.ne.jp/

ビューサイトの場合に有用であった評価表現抽出方法では十分な精度が得られない可能性がある．逆に，質問と回答に分割されているという性質を活用することができれば，精度を高められる可能性もある．

そこで，本研究では質問投稿サイトを対象とした比較表現抽出について取り組み，比較表現判別のために山崎らの特徴量に新たな特徴量を追加することで精度の向上を図る．

3. 比較表現判別に用いる特徴量

本節では対話型データの傾向を明らかにし，比較表現の抽出手法について考察する．

3.1 比較表現の収集

まず，分析に使用する比較表現の候補を収集する．本研究では「Yahoo!知恵袋」の化粧品カテゴリの投稿質問文と，それに対するベストアンサー[6]を回答文とし，このペアをwebスクレイピングツールであるOctoparse[3]を用いて投稿単位で収集した．なお，検索に使用したキーワードは「ちふれ セザンヌ」，「ランコム ディオール」のように，2つのブランド名を含んだものを対象とした．商品名ではなく，ブランド名とした理由は，「セザンヌのアイシャドウ」というように，商品名よりもブランド名による表現が多く見られたからである．

次に収集したデータに対し，人手により比較表現か否かのラベル付けを行う．一般的に比較表現は，「対象＋属性＋評価」の形で構成されることが多い．ここで「対象」は商品やブランド名，「属性」は比較対象商品を評価する際の観点，「評価」は属性の量や質に関する表現を意味する[7]．例えば，「商品Aの方が保湿力が高い」という文の場合，「商品A」が対象，「保湿力」が属性，「高い」が評価となる．また，「商品Aの方が潤う」という文のように，「潤う」1語で属性と評価を表現しているケースもある．これを考慮して，以下の条件をすべて満たすものを本研究では比較表現と定義する．

① 1つの発言中に2つの比較されうる対象となる商品名，またはブランド名が含まれている．
② 少なくとも1つの比較対象商品に対して，属性と評価が記述されている．
③ 比較対象商品名のいずれかに対する優劣が明示的に読み取れる．

3.2 比較表現の分析と特徴量の決定

質問投稿サイトは対話型データであるため，その特性を考慮して特徴量を決定する必要がある．本研究では，先行研究の特徴量に加え，比較対象商品名の冒頭出現の有無，対象商品名の周辺語の品詞の出現傾向の2つを新たな特徴量候補とする．

1つ目の比較対象商品名の冒頭出現の有無は，回答文の1文目に比較対象商品名が含まれているか否かを情報とする特徴量である．これを特徴量の候補とした理由は，質問文で比較対象商品同士の優劣を尋ね，回答として最初に商品名を挙げた後に，詳細な理由を説明するパターンが多いからである．例としては以下のようなものである．

Q. 商品Aと商品Bではどちらが良いですか？
A. 断然商品A. 保湿力が高いうえに値段もお手頃です．

対象商品名の周辺語の品詞の出現傾向を特徴量候補としたのは，比較表現には定形パターンが多いため，対象商品名の周辺語にもパターンに伴う品詞の偏りがあると考えたからである．そこで，周辺語の品詞ごとの出現頻度を分析し，比較表現と非比較表現で傾向に差異が大きいものを明らかにし，その品詞の出現の有無を特徴量とすることを考えた．

4. 分析と結果

4.1 分析手順

今回の分析で使用するデータ数は，比較表現を含む投稿120件と，比較表現を含まない投稿471件の計591件である．それぞれの投稿データの集合を，比較表現投稿群，非比較表現投稿群と呼ぶ．ただし，両者のサンプル数の差異が大きく異なるため，判別精度を適切に評価するためには同程度のサンプル数で分析する必要がある．

そこで今回は，データ数が少ない比較表現投稿群に非比較表現投稿群の件数をそろえ，それぞれ120件で分析する．なお，信頼性の高い結果となるよう，比較表現投稿群を3つに分割してK=3としたK分割交差検証法[8]を変則的に用いて，3通りの分割群の組み合わせで数量化Ⅱ類[9]により分析した．なお，目的変数には比較表現か否かのラベル，説明変数には先行研究と本研究それぞれの特徴量を使用し，最終的な精度評価を分割交差検証法の平均値で行う．

4.2 先行研究の特徴量の精度検証

まず，山崎らの特徴量を用いて比較表現を抽出した場合の精度を検証する．抽出は手がかり語を含み，かつ並列表現を含まない投稿を比較表現と判別することによって行う．使用する手がかり語は先行研究と同様に「おすすめ」「より」「ほうが」の3語を，並列表現は「とか」「やら」「や」「か」「と」「も」「Aでも～Bでも」「Aでもなく～Bでもない」「Aだけでなく～Bも・・・」「Aであれ～Bであれ」「Aにしても～Bにしても」「Aにも～Bにも」の12語を用いる．

表 1は先行研究の特徴量を用いて3通りの数量化Ⅱ類を用いた判別結果である．

[3] https://www.octoparse.jp/

表 1 先行研究の手法による判別的中率

組み合わせ	判別的中率
分割群1	52.50%
分割群2	52.92%
分割群3	53.75%
平均値	**53.06%**

平均判別的中率が約50%との結果から，この方法では十分な精度が得られないことがわかる．原因としては質問と回答という対話型データの場合，手がかり語が含まれない場合があるため，この手法のみでは抽出漏れが生じてしまっていると考察できる．

4.3 本研究の特徴量の精度検証

次に，本研究で新たに追加する特徴量である比較対象商品名の冒頭出現の有無について，先と同じく分割交差検証法を用いた数量化Ⅱ類により検証する．ここでの冒頭出現の有無とは，回答文の1文目に比較対象商品名が含まれているか否かを意味する．なお，1文の区切りとしては，句点だけでなく顔文字，「!」，改行も区切り記号として処理した．表 2は対象商品名の有無を説明変数，比較表現か否かのラベルを目的変数として，数量化Ⅱ類にかけた判別結果である．

表 2 比較対象商品名の冒頭出現による判別的中率

組み合わせ	判別的中率
分割群1	67.08%
分割群2	70.00%
分割群3	75.00%
平均値	**70.69%**

平均判別的中率が70%を超えることから，比較対象商品名の冒頭出現の有無は特徴量として適切であると考えられる．

最後に対象商品名の周辺語の品詞の出現傾向を特徴量とすることについて検証する．まず，特徴量とするべき品詞の種類を決めるため，比較表現と非比較表現に現れる比較対象商品名の前後3単語以内を対象に，形容詞，動詞，助詞，助動詞の出現頻度を分析した．あらかじめ名詞を除外した理由は，「口紅」や「ファンデーション」などの商品種別には名詞が多く，これらの語は比較投稿群と非比較投稿群問わず出現し，判別には有用でないためである．なお品詞の解析には日本語形態素解析ステムであるMeCab[4]を用いた[10]．表 3，表 4は回答文と質問文それぞれの品詞の出現頻度である．

表 3 質問文中の品詞出現頻度

	形容詞	動詞	助詞	助動詞
比較表現投稿群	6	21	148	42
非比較表現投稿群	4	16	156	37

表 4 回答文中の品詞ごとの出現頻度

	形容詞	動詞	助詞	助動詞
比較表現投稿群	27	69	270	138
非比較表現投稿群	6	27	121	53

助詞の出現頻度に着目すると，質問文と回答文のどちらにも100件を超えて頻出しているため，判別の特徴量には適さないと判断できる．また，質問文の場合は比較表現投稿群と非比較表現投稿群での出現頻度について，いずれの品詞でも差異がないことから回答文の対象品詞の出現の有無のみを特徴量とする．

次に，上述で決定した品詞を用いて先と同じく分割交差検証法を用いた数量化Ⅱ類により検証する．また，対象語からの適切な分析範囲を知るために，前後3，4，5単語の範囲を対象とした場合の検証をそれぞれ行った．表 5は決定した品詞3つと対象語からの単語距離を，それぞれ独立した形で説明変数として数量化Ⅱ類にかけた判別結果である．

表 5 比較対象商品名からの距離別的中率

比較対象商品名からの距離	前後３単語	前後４単語	前後５単語
判別的中率	67.50%	72.92%	70.83%

前後4単語の範囲での判別的中率が最も高いことから，この距離の場合を特徴として用いることにする．そこで，対象商品名の前後4単語以内に形容詞，助詞，助動詞が含まれているか否かを特徴量とした場合の最終的な判別的中率について，分割交差検証法により確認する．

表 6 対象商品名の前後 4 単語以内の品詞出現傾向による判別的中率

組み合わせ	判別的中率
分割群1	72.92%
分割群2	67.50%
分割群3	72.92%
平均値	**71.11%**

表 6に示した最終的な平均判別的中率でも70%を超えていることから，この特徴量が適切であることが確かめられた．

[4] http://taKu910.github.io/mecab/

5. 特徴量の統合による精度検証と考察

前節では先行研究の仮説と，本研究の仮説の検証をそれぞれ独立した形で精度を検証した．本節ではこれら2つの仮説を統合した場合の精度について検証する．手順としては，先行研究の特徴量1つと，本研究の2つの特徴量を説明変数として，数量化Ⅱ類で分析をした．その判別結果を表 7に示す．

表 7 先行研究と本研究の特徴量を統合した場合の判別的中率

組み合わせ	判別的中率
分割群1	74.17%
分割群2	70.42%
分割群3	77.92%
平均値	**74.17%**

結果として，先行研究の特徴量のみを用いた手法の的中率よりも，これに提案した新しい特徴量を追加した場合のほうが20％以上向上しており，提案の有効性を確認することができた．要因としては，先行研究の手法では手がかり語を用いて抽出できていなかった検索漏れの部分を，本研究の手法を用いることで拾い上げることができているのではないかと推測される．

最後に，この特徴量を用いた場合の再現率と適合率[11]について確認した．表 8に示したように，再現率と適合率のどちらも70％を超えたことから，正しい比較文の抽出漏れも，比較文でないものを誤って拾い上げてしまう誤りも，バランスよく改善できていることを示すことができた．このことから，先行研究の特徴量と本研究の新たな特徴量を統合して用いる手法は，十分有効であるといえる．

表 8 先行研究と本研究の特徴量を統合した場合の再現率と適合率

組み合わせ	再現率	適合率
分割群1	71.67%	75.44%
分割群2	75.83%	75.83%
分割群3	76.67%	76.67%
平均値	**74.72%**	**75.98%**

6. むすび

本研究では，質問投稿サイト「Yahoo!知恵袋」を題材に，対話型データに含まれる比較表現を抽出する手法を提案した．新たな特徴量を対象商品名の冒頭出現の有無と，対象商品名の周辺語の特定品詞の出現の有無で構成した結果，対話型データの抽出の面では先行研究の手法よりも高い精度で比較文を抽出することができた．また，これらの特徴量と先行研究の手法を統合して検証することで，より高い精度で対話型データから比較文を抽出できることも示した．

一方で，本研究では化粧品分野のデータのみを用いて検証したため，本研究で用いた特徴量が，スポーツ用品などの化粧品以外の商品分野でも有効性を発揮するのかを今後検証する必要がある．

参考文献

[1] 総務省，デジタルデータの経済的価値の計測と活用の現状に関する調査研究，https://www.soumu.go.jp/johotsusintokei/linkdata/r02_05_houkoku.pdf(参照:2022-02-08)

[2] 松尾義博，富田準二，クチコミ分析システムの作り方，近代科学社，2019.

[3] 石澤ら，“商品改善に繋がる比較意見の抽出”，情報科学技術フォーラム講演論文集,No.11,pp.193-194,2012.

[4] 山崎ら，“競合事物間における比較関係認識”，情報処理学会研究報告.自然言語処理研究会報告,Vol.2011,No.5,pp.1-7,2011.

[5] 株式会社グルーブワークス，日本人の平均口コミ人数～ライフスタイル調査結果報告，https://grooveworks.co.jp/?p=6195(参照:2022-02-08)

[6] ヤフー株式会社，Yahoo!知恵袋ヘルプ，https://support.yahoo-net.jp/s/(参照:2022-02-07)

[7] 小林ら，“意見抽出のための評価表現の収集”，奈良先端科学技術大学院大学情報科学研究科修士論文，2003.

[8] 高木章光，鈴木栄太，最新データサイエンスがよ～くわかる本，秀和システム，2019.

[9] 石井俊全，意味がわかる多変量解析，ベレ出版，2014.

[10] 末吉美喜，テキストマイニング入門:Excel と KH Coder で分かるデータ分析，オーム社，2019.

[11] 金明哲，テキストアナリティクスの基礎と実践，岩波書店，2021.