修士論文

商品レビューの複数の観点からの有用性の評価

曽田 颯人

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
情報科学

令和３年３月

Abstract

Recently, customer reviews about products and services become popular due to the rapid spread of online shopping. User's decision on choice of a product to buy is greatly influenced by customer reviews posted by other users who have already used that product. However, there exists both useful and non-useful customer reviews. When a huge amount of reviews are posted in online shopping web sites, it is rather costly and time-consuming to find useful reviews among them. Therefore, it is necessary to develop a technique to automatically evaluate the usefulness of reviews and show the results to users to help their purchase.

In previous work on estimating the usefulness of customer reviews, major methods are training a classifier using a sentence length and part-of-speech as features to determine whether a given review is useful or not. Another studies aim at evaluating the usefulness of customer reviews from a specific point of view, such as extracting comparative opinion and identifying an entity mentioned by a reviewer. However, different users may think what useful reviews are differently. It is insufficient to simply classify whether a review is useful or to evaluate the usefulness of a review from a single viewpoint in order to satisfy various users who have their own demands. Therefore, it is necessary to evaluate the usefulness of reviews on not a single viewpoint but multiple viewpoints.

The goal of this thesis is to develop a system that can not only classify if customer reviews are useful but also clarify what is useful in them or how useful for users they are. More precisely, we propose a system that evaluates customer reviews from multiple viewpoints and displays the results of the evaluation to users. We propose seven viewpoints for evaluating the usefulness of reviews. Our proposed system analyzes reviews in diversified ways by evaluating them from each viewpoint separately. The final system will be designed to provide many useful functions to users, such as to display a list of reviews in order of a score of the viewpoint specified by a user, or to display only reviews that are highly rated from the viewpoint. Our system makes it easier for users to find helpful customer reviews that meet personal preference or requirements of individual users.

Based on the findings of the previous studies on evaluating the usefulness of reviews as well as our past experiences and insight, seven viewpoints for evaluating the usefulness of reviews are proposed: "A reviewer shows reason for his/her opinion," "A reviewer explains a product in detail," "A reviewer compares a product with others," "A reviewer may (or may not) actually use a product," "A reviewer shows reason for his/her rating," "A review is long" and "A review is easy to read." For a given set of reviews written about a certain product, the system evaluates the usefulness of them from each of the seven viewpoints separately. That is, the system consists of seven subsystems used for evaluation of each viewpoint. In this

thesis, we focus on three of seven viewpoints and propose methods to automatically evaluate the usefulness of reviews from them.

To evaluate reviews from Viewpoint1(A reviewer shows reason for his/her opinion), we aim to detect sentences that contain evaluation of a product and reason of it. A rule-based method is designed by considering opinion words (such as "便利です"(*benri-desu*;convenient), "実用的です"(*jitsuyô-teki-desu*;practical)) and keywords (such as "ので"(*node*), "ため"(*tame*)) that are conjunctions indicating reason of something. Precisely, after dependency analysis of a given sentence, the system judges that the sentence includes an opinion to a product and its reason by checking either of the following requirements: (1) a chunk including a word in the form of *renyôkei*(the conjugation indicating that its head is a predicate) modifies another chunk including an opinion word, (2) a chunk including the above keyword (e.g. *node*, *tame*) modifies another predicative chunk. In the evaluation experiment, the test data was constructed by retrieving reviews from the online shopping web site and annotating them with the label indicating whether they express the reviewer's opinion and reason for it. The performance of the proposed system was evaluated on this dataset. The recall, precision, and F-measure were around 0.8, 0.45, and 0.6, respectively. It was found that the recall was relatively high, while the precision was low since the first requirement was often fulfilled in sentences not including reason for a reviewer's opinion.

To evaluate reviews from Viewpoint2(A reviewer explains a product in detail), we define "degree of explanation", a score that represents how detailedly a reviewer explains a product, and propose a method to calculate it. First, for each category of products such as "PC" and "book", keywords relevant to the category are obtained. Nouns and compound nouns are retrieved as the keywords from descriptions of products in the dataset of the e-commerce site "Rakuten Ichiba". In addition, significance of the keyword for the category is measured by using TF-IDF for each keyword. The keywords and their significant scores are stored in the lexicon. Next, for a given review, keywords are extracted by looking up the lexicon, and their significance scores are summed up. Finally, the degree of explanation is calculated by the weighted sum of the total of the significance scores and the length of the review. In the evaluation experiment, for a given pair of reviews written for the same product, the proposed system judged which explains the product in more detail by comparing the degree of explanation of two reviews.

The accuracy of the proposed method was around 0.77, which was better than the baseline that simply selected a longer review.

To evaluate reviews from Viewpoint3(A reviewer compares a product with others), we propose a method to classify a review if it includes comparison among products. In this study, the method is designed to detect sentences that explic-

itly represent comparison. The following three types of rules are developed. The first one is a rule to check whether a sentence contains both a keyword indicating comparsion (such as “比べる”(*kuraberu*;compare) and “他のメーカー”(*hoka-no-mêkâ*;other maker) and an opinion word. The second one is a rule using keywords indicating that the reviewer bought a new product to replace old one, such as “買い替え”(*kaikae*;buy to replace). If such a keyword is found in the beginning of the review, it is regarded that the reviewer compare old and new products in the whole review. The third one is a rule to check whether an opinion word is the head of the conjunction “より”(*yori*;than) that often indicates comparison. Results of the experiment showed that the precision of the detection of reviews which contain comparison was sufficiently high for several rules. On the other hand, many rules could detect only a few reviews or no review including comparison, since the number of reviews with comparison was a quite few in the test data.

概要

近年，オンラインショッピングの利用の急速な拡大に伴って，商品やサービスに関するカスタマーレビューの投稿も盛んになってきている．先に商品を利用したユーザが投稿したカスタマーレビューは，ユーザの商品選択に大きな影響を及ぼすと考えられる．しかし, カスタマーレビューの中には役に立つレビューと役に立たないレビューが混在する．ひとつの商品に対するレビューの数が非常に多いとき，その中から有用なレビューを見つけ出すのは多大な労力がかかるという問題がある．このことから自動的に有用性を評価してユーザに掲示する技術が求められている．

カスタマーレビューの有用性を予測する既存の研究では，文長や品詞などを素性とした機械学習により，レビューが有効であるか否かを判定する分類器を学習する手法が主流である．また，比較意見文の抽出やレビューが言及している対象の分類など，特定の観点でレビューの有用性を評価する研究もある．しかし，どのようなレビューが有用であるかはユーザーによって異なると考えられ，単に有用か有用でないかを判定したり，ひとつの視点から有用性を判定したりするだけでは，ユーザの多様なニーズに対応しきれないと考えられる．そのため，レビューの有用性を 1 つの尺度だけでなく，複数の尺度で多角的に評価することが求められる．

本研究では，カスタマーレビューが単に有用か有用でないかではなく，どの点がどのように有用か有用でないかをユーザに示すことを目的とする．具体的にはカスタマーレビューを複数の観点から評価し，その評価結果をユーザに示すシステムを提案する．レビューの有用性を評価する観点を 7 つ提案し，それぞれの観点についてレビューを評価することで，レビューの有用性を多角的に評価する．また，最終的なシステムとして，ユーザが重視する観点を入力することで，その観点のスコアが高いレビューを優先して表示する，あるいはその観点について高く評価されたレビューのみを表示するフィルタリング機能をユーザに提供することを目指す．このシステムによって，ユーザは各々の嗜好にあった有用なカスタマーレビューを見つけ易くなる．

レビューの有用性を判定する先行研究の知見や著者らによる経験などを踏まえ，レビューの有用性を評価する観点として,「評価表現に対する根拠がある」,「商品に関係のある言及が多い」,「他の商品との比較をしている」,「実際に商品を使用した (あるいはしていない) と推測できる」,「評価 (レーティング) に対しての根拠がある」,「分量が多い」,「読みやすい文である」の 7 つを提案する．特定の商品について書かれたレビュー文の集合を入力として，個々のレビューの有用性を 7 つの観点でそれぞれ評価する．このとき有用性を評価するシステムは観点ごとに独立している．本研究では，7 つの観点のうち 3 つに焦点を当て，それぞれの観点からレビューの有用性を自動評価するサブシステムを構築する．

観点 1(評価表現に対する根拠がある) を評価するシステムを実現するために，商品に対する評価とその根拠が書かれている文を検出することを目指す．評価表現 (「便利です」「実用的です」など) と根拠を表すキーワード (「ので」,「ため」な

ど) を用いたルールベースの手法を提案する．具体的には，レビュー文を構文解析し，(1) 連用形を含む文節の係り先が評価表現を含む文節である，もしくは (2) 根拠を表すキーワードで終わる文節の係り先が用言を含む文節であるとき，その文を評価の根拠を含む文と判定する．評価実験では，提案手法によって評価用のレビュー集合の中から評価表現に対する根拠を含むレビューを検索した．実験の結果，検索の再現率，精度，F 値は 0.8，0.45，0.6 程度となった．再現率は比較的高いが，精度は低かった．精度が低い主な要因は，評価表現に対する根拠がない文もしばしば (1) の条件を満たしていたことであった．

観点 2(商品に関係のある言及が多い) を評価するシステムについて，レビューが評価対象の商品についてどの程度言及しているかの度合を「商品言及度」と定義し，これを定量化する手法を提案する．まず,「PC」「本」などの商品カテゴリ毎に，その商品カテゴリに関連するキーワードを取得する．EC サイトの「楽天市場」における商品説明文のデータセットから，名詞と複合名詞をキーワードとして取得する．さらに，商品カテゴリに対するキーワードの重要度を TF-IDF により算出する．次に，レビューに対し，それに出現するキーワードを検出し，その重要度の総和を求める．最後に，重要度の総和と文長の重み付き和により言及度を算出する．提案手法の評価のため，同じ商品に対する 2 つのレビューを比較し，どちらのレビューが商品に言及しているかを判定する実験を行った．提案手法による判定の正解率は 0.77 程度であり，単に文長の長いレビューを選ぶベースラインを上回ることを確認した．

観点 3(他の商品と比較している) を評価するシステムでは，比較が明示的に示されているレビューを検出することに焦点をあて，そのためのルールベースの手法を提案する．具体的には以下の 3 種類のルールを用いる．1 つ目は，比較を表すキーワード (「比べる」「他のメーカー」など) と評価表現の両方を含むとき，レビューは比較を含むと判定するルールである．2 つ目は，レビューの冒頭に他の商品からの買い替えであることを示唆するキーワード (「買い替え」など) が存在するとき，レビュー全体で他の商品と比較していると判定するルールである．3 つ目は，比較を示唆する接続詞「より」が評価表現に直接係るときに比較を含むと判定するルールである．評価実験の結果，いくつかのルールについて，比較を含むレビューの検出の精度が十分に高いことがわかった．一方，評価用データにおける比較を含むレビューの数が少なかったため，検知数が少数または 0 のルールも多かった．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

近年，オンラインショッピングの利用が急速に拡大しており [1]，それに伴って商品やサービスに関するカスタマーレビューの投稿も盛んになってきている．オンラインショッピングでは消費者が商品やサービスを購入前に体験することは困難であることから，先に商品やサービスを利用したユーザが投稿したカスタマーレビューは，ユーザの商品選択に大きな影響を及ぼすと考えられる．しかし, カスタマーレビューの中には役に立つレビューと役に立たないレビューが混在する．商品について投稿されたレビューが膨大な場合には，その中から有用なレビューを見つけ出すのは多大な労力がかかり，ユーザの商品選択の障害となり得る．

この問題に対する取り組みとして，e-コマースサイトの中には，図 1.1 の波線部に示すようにユーザによるカスタマーレビューに対しての有用性投票機能を実装しているものもある．有用性投票を多く受けているレビューから優先して表示することで，有用だと思われるレビューをユーザに閲覧されやすくしている．しかし，この手法には古いレビューが新しいレビューよりも投票を受けやすいという問題がある．そのため全てのレビューを公平に評価するために，自動的に有用性を評価してユーザに掲示するような技術が求められている．

レビューの有用性を予測するという問題に対する現在までの取り組みでは，レビュー文やレビュー対象の商品に関連した情報を素性として，機械学習によってレビューの有用性を判定する分類器を学習する手法が主流である．これに関する多くの先行研究では，単にレビューが有用か有用でないかを予測している．しかし，どのようなレビューが有用であるかはユーザーによって異なると考えられる．例えば商品を使用した体験が役に立つと考えるユーザもいれば，他の商品との比較を重視するユーザもいるだろう．このことから単に有用か有用でないかを予測するだけでは，ユーザの多様なニーズに対応しきれないと考えられる．そのためレビューの有用性を 1 つの尺度だけでなく複数の尺度で多角的に評価することが求められる．

図 1.1: Amazon のヘッドホンへのレビューに対する有用性投票の例

1.2　目的

本研究では，カスタマーレビューが単に有用か有用でないかではなく，どの点がどのように有用か有用でないかをユーザに示すことを目的とする．具体的にはカスタマーレビューを複数の観点から評価し，その評価結果をユーザに示すシステムを提案する．いくつかの先行研究の知見も踏まえ，レビューの有用性を評価する際の「評価表現に対する根拠がある」,「商品に関係のある言及が多い」などの観点を 7 つ定義し，それぞれの観点についてレビューを自動評価することで，レビューの有用性を多角的に分析する．

また，最終的なシステムとしてユーザーが重視する観点を入力することで，その観点の評価値を他の観点の評価値よりも大きな比重で評価し，ユーザーが重視する観点のスコアが高いレビューを優先して表示すること目指す．あるいは，ある観点について高く評価されたレビューのみを表示するフィルタリング機能をユーザに提供する．このシステムによって，ユーザーは各々の嗜好にあった有用なカスタマーレビューを見つけ易くなる．

上記のシステムの実現により，ユーザーが有用なレビューを見つける作業を支援し，ユーザが商品選択をより円滑に進められるようにすることを狙う．

1.3　本論文の構成

本論文の構成は以下の通りである．2 章では本研究の主要な関連研究について述べ，本研究の立場を明らかにする．3 章では有用性を評価する 7 つの観点の案について述べ，それに基づくカスタマーレビューの閲覧システムの構想ついて述べる．4 章では，評価だけでなくその根拠を示しているかという観点からレビューの有用性を評価する手法について述べる．5 章では，商品に関係のある言及が多いかという観点からレビューの有用性を評価する手法について述べる．6 章では，他の商品との比較があるレビューは有用であるという観点から，比較を含むレビューを検出する手法について述べる．4,5,6 章では，提案手法の評価実験についても報告する．7 章では上記 3 つ以外の観点について述べる．最後に 8 章では本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では本研究に関連した研究について述べる．2.1 節ではカスタマーレビューの有用性の判定に関連した研究を紹介する．続く 2 つの節ではレビューを特定の観点から評価している研究について紹介する．2.2 節ではレビューが言及している対象に関連する研究について，2.3 節では複数の評価対象を比較した文に関する研究について述べる．最後に，2.4 節では本研究と先行研究の違いについて論じる．

2.1　レビューの有用性の判定

山澤らは，Amazon[1]のレビュー文を対象に，書き手の性質や趣向が分からなくても，ユーザーが内容を信用して利用できる文 (有用文) を自動抽出する手法を提案した [19]．有用なレビューを「ユーザの購入の意思決定に役立つレビュー」と定義し，人手によって有用か否かを文に対してタグ付けしたデータセットを作成した．レビュー文の形態素情報を素性として Support Vector Machine(SVM) を学習し，約 2000 件のレビューのデータセットで実験した結果 82%の正解率を示した．また品詞の出現頻度による素性選択を行った．有用な文とそうでない文のそれぞれについて一方で出現頻度が上位でありかつもう一方では出現頻度が上位でない品詞を素性として選択することで，分類精度が向上することを示した．

佐々木と関は，機械学習を用いて有用なレビューを判別する手法を提案した [14]．有用なレビューを判別する 8 つの基準を定義し，それらの基準を参考にレビューが有用か有用でないかを人手によって分類した．この分類結果から有用なレビューを判別する際に有力な情報となり得る基準は「評価の根拠がある」,「レビュー投稿者が商品の使用者であると判断できる」であると推測した．これを踏まえ，有用なレビューは「評価の根拠や評価対象となる商品について詳しく述べられており，文章がしっかりしているレビューである」と定義した．以上の分析結果に基づき，レビューに出現する形態素情報を素性としてレビューが有用か有用でないかを分類する SVM 分類器を学習した．評価実験の結果，助詞のみを素性に用いた分類器の F 値が 0.86 であった．また形態素情報に加えて文字数や文の数などの構造的情報も素性に加え，これらの素性による分類結果への影響も調査した．調査の結果，レビューの文字数や，キーフレーズ数などを組み合わせることで判別の精度が向上することがわかった．

[1] https://www. amazon.co.jp/

木浪らは，レビューの主観的視点と有用性との間の関係性について分析した [7]．形態素レベルでの主観，客観表現の存在とレビュー文章の有用性との間の関係性を調査した結果，高有用性群より低有用性群のレビュー方が主観語の出現頻度が高いことが分かり,「主観語の出現が少ないレビューほど有用である」という仮説を立てた．また，商品レビューでは客観語はあまり用いられず，主観語を用いて記述される傾向が確認され，有用性が特に高いとされるレビューでは主観語，客観語ともに出現頻度が比較的低い傾向が見られた．

Fan らは，製品のメタデータ (タイトル，ブランド，カテゴリ，商品説明文など) とレビューテキストの両方を入力として，レビューが有用か有用でないかを分類するディープニューラルネットワークを提案した [4]．図 2.1 のアーキテクチャの概要図に示すように，レビュー文と商品のメタデータの分散表現を素性に，Bi-LSTM による RNN 層を含むニューラルネットワークを学習し，有用なレビューの識別と有用性投票における「有用である」の投票の比率を予測した．また有用性予測タスクに関する様々なアプローチの性能比較を公平に行うため，大規模なベンチマークデータセットを構築した．実験の結果，提案したモデルは全ての主流のアプローチの性能を上回った．

Rodak らは，レビューの長さ，全て大文字で表記されている単語の数などの構造的特徴，レビューの評価値などのメタデータ的特徴，ユニグラムや文章の可読性などの語彙的特徴を素性として，ナイーブベイスや SVM などでレビューが有用か有用でないかを判定するモデルを学習した [13]．評価実験の結果，Radial basis function と多項式カーネルを SVM に使用することで，製品レビューの有用性について約 70%の予測精度を達成した．

Yang らは，レビューの有用性はレビュー本文からのみ得られる特性であるという仮説を立て，レビューの意味的特徴から有用性を予測する手法を提案した [20]．言語学的･心理学的辞書を活用し，単語を意味的次元で表現する LIWC と INQUIRER という辞書から得られる特徴を素性として，SVM 分類器を学習した．評価実験の結果，既存の関連研究で使用された特徴を使用するモデルより高い性能を示した．また，LIWC と INQUIRER が持っている意味的特徴のカテゴリと有用性の相関関係を検証することで，有用なレビューは推論や経験についての記述を多く含み，感情的表現が少ないことを示した．

Hong らは，レビューの有用性に関する既存の 42 件の研究を分析し，これらの研究で指摘されているレビューの有用性の決定要因について，その有効性に関する研究毎の結論の違いを調査した [5]．例えば，多くの研究で一貫して有効であると結論づけられた要因は真にレビューの有用性判定に有効であるのに対し，研究によって有効であったりなかったりと結論づけられている要因はその有用性に疑問が残る．調査の結果から，各研究の間で有用性に与える影響に一貫性がないことが分かった決定要因についてメタアナリシスを実施し，混在する様々な研究の結論を統合した．メタアナリシスの結果，レビューの長さ，レビュー投稿日の古さ，レビュワーの情報開示，レビュワーの専門性が有用性にプラスの影響を与え，

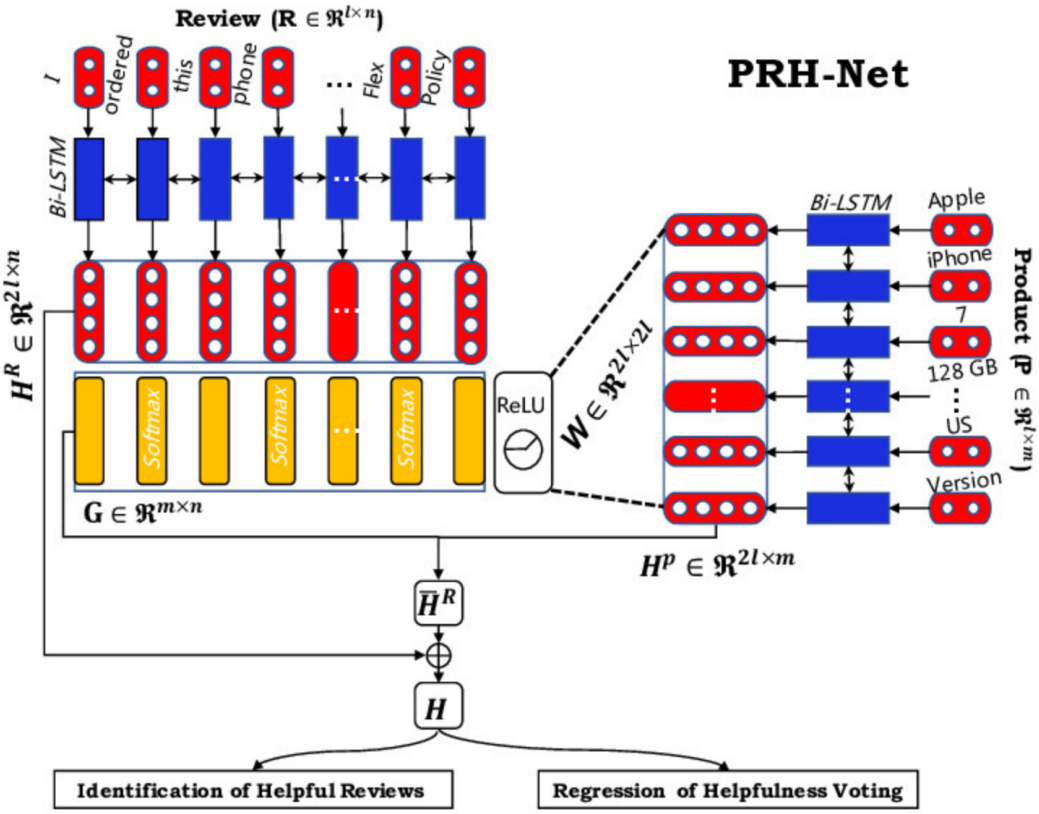


図 2.1: Fan らのモデルの概要図 [4]

レビューの可読性，レビューの評価値は有用性に大きな影響を与えていないと結論付けた．また有用性を計測する尺度の違い，レビュープラットフォーム運営者の立場の違い，検索商品と経験商品という製品のタイプの違いが，各研究の結論の一貫性のなさに影響を与えていると主張した．

2.2 レビューが言及している対象の分析

一般に，商品レビューでは，商品に対して意見を述べている文もあれば，ストアや配送業者など商品以外について言及している文や，その商品を買った動機を説明している文などが混在する．つまり，レビュー文が言及している対象は様々である．本節ではレビュー文が言及する対象に焦点を当てた研究について述べる．

山下らは，商品レビューにストアへの言及が含まれているか否かを判別するために，クラウドソーシングを利用した機械学習に基づく手法と，既存のストアレビューを用いた半教師あり機械学習に基づく手法を提案した [18]．前者の手法では，レビューの商品またはストアへの言及度合いを 5 段階で評価するタスクをクラウドソーシングで行い，計 15 万件回答を得て，その結果を用いて SVM 分類器を学習した．後者の手法では，商品レビューに加えてストアレビューも入力とした SVM 分類器を学習した．評価実験の結果，前者の手法では 0.9730 の判別精度を示し，後者の手法では 0.9386 の判別精度を示した

新井と佐藤は，評判情報が記述された個々の意見文を，言及している評価視点 (デザイン，携帯性など) に基づいて分類する意見文分類手法を提案した [2]．評価視点の語と共起する語を関連語とし，関連語が評価視点と共起する確率と，全体での出現確率を使って，関連語ごとに関連度を登録した辞書を作成した．その関連語辞書を用いて，文に現れる関連語の関連度の総和をその評価視点に対する言及度とし，文に対する言及度が一番高い評価視点がその文が持つ意見の評価視点と分類した．評価実験の結果，分類精度は約 75%であり，SVM を用いた手法と比較して 7%程度精度が向上できることが分かった．またこの手法によって，評価視点を表す語を明示的に含まない文も正しく分類されていることを確認した．

2.3 比較文の検出

複数の製品を比較したレビューは，ユーザにとって参考になる情報を含むという点で有用である．本節では比較文 (複数のものを参考にしている文) に関する研究を紹介する．

Jindal と Liu は，レビュー中から比較文を検出するタスクを提案した [6]．比較文を異なるタイプに分類し、テキスト文書から比較文を識別するために Class Sequential Rule(CSR) と教師あり学習を組み合わせたアプローチを示した．ニュース記事，消

費者レビュー，インターネットフォーラムの投稿の 3 種類の文書を用いた実験の結果，精度 79 %、再現率 81 %という結果が得られた．

Varathan らは，比較オピニオンマイニングに関する研究のサーベイを報告している [16]．比較オピニオンマイニングとは，複数の評価対象を比較している意見を集約し，その傾向を明らかにする技術である．機械学習やルールベースの比較意見の分類などの手法的側面と，比較意見に含まれる特徴などの要素的側面の 2 つの異なる角度から比較オピニオンマイニングに関する研究を分類し，個々の研究を紹介した．また論文調査の結果，これまで英語で発表された比較意見のマイニングの研究は，英語，中国語，韓国語についてのみであると報告した．

2.4 本研究の特徴

従来の研究ではレビューが単に有用か有用でないか，または特定の観点についてのみ有用性の判定もしくは評価をしていたのに対して，本研究では複数の観点について有用性を評価することで，多角的にレビューの有用性を評価した情報をユーザに提供する点に特徴がある．

また，従来の研究ではシステム自体がレビューの有用性の判定を行っているのに対して，本研究ではシステムはあくまで各観点による評価結果をユーザに提示するのみであり，有用性の判定自体はユーザ自身に委ねることを想定している．

第3章　提案手法

本章では，カスタマーレビューの有用性を複数の観点で評価しその評価結果をユーザに提示するシステムを提案する．3.1 節では，本研究で用いる有用性の観点を 7 つ提案し．それぞれについてその詳細な定義を述べる．3.2 節では，各観点の評価結果をユーザに提示するシステムの概要を示す.

3.1　有用性の観点

レビューの有用性を判定する先行研究 ([19, 17, 7]) の知見や著者らによる経験などを踏まえ，以下の 7 つの観点からレビューの有用性を評価することを提案する．

1. 評価表現に対する根拠がある
2. 商品に関係のある言及が多い
3. 他の商品との比較をしている
4. 実際に商品を使用した (あるいはしていない) と推測できる
5. 評価 (レーティング) に対しての根拠がある
6. 分量が多い
7. 読みやすい文である

以下，各観点について，詳細な定義を述べる．

観点 1: 評価表現に対する根拠がある

レビューの中には商品に対して評価をしていても，なぜそのような評価になったかまで示していない文がある．このような文は評価の根拠まで示した文より有用性は低いと考えられる．そこで単に商品の良し悪しについて評価するだけでなく，評価の根拠となるような事実も合わせて書いているとき，そのレビューの有用性は高いと評価する．図 3.1 のレビュー例では，1 文目の「大きさもちょうど良いので便利です」という文において「便利です」という評価に対して「大きさもちょうど良いので」という評価の根拠が示されている．

観点 2: 商品に関係のある言及が多い

レビューの中には，商品について言及している文もあれば，ショップの対応や配送業者の対応など，商品以外のことについて言及している文もある．そこで，商品に言及している文が多ければ多いほどレビューの有用性は高いと評価し，一方で商品に関係のない文が多いほどレビューの有用性は低いと評価する．図 3.1 のレビュー例では，1 文目と 2 文目は商品について言及しているが，3 文目の「発送までが遅かったのが残念でした」はショップに対する言及である．

観点 3: 他の商品と比較している

レビューが評価対象の商品と他の商品を比較しているとき，ユーザにとって参考になる情報を提供している可能性が高いため，有用性が高いと評価する．図 3.1 の例では，2 文目の「商品 B に比べて軽そう」が比較表現である．

観点 4：実際に商品を使用した (あるいはしていない) と推測できる

商品を実際に使用していないユーザの評価は商品の品質や特徴を正確に把握した評価とは考えにくいため，レビューに書かれている事実の信頼性は低くなる．よって，商品を実際に使用していないと分かる人のレビューの有用性は低いと評価する．図 3.1 の例では,「まだ実際に使用していませんが」という句から，ユーザが実際に商品を使っていないことが推測でき，レビューの信頼性が低くなる．

観点 5：評価 (レーティング) に対する根拠がある

多くの EC サイトや口コミサイトでは，ユーザは商品に対して星の数などでレーティングをつけることができる．このレーティングが高いにも関わらずレビュー文ではネガティブな評価が多い場合や，反対にレーティングが低いにも関わらずレビュー文ではポジティブな評価が多い場合は，レビューとレーティングの一貫性がなく，ユーザの混乱を招く可能性がある．一方でレーティングに対する根拠が示されている場合，例えばレーティングが 5 段階中 4 で，商品を概ねポジティブな評価をしている一方，商品の一部の側面についてのみネガティブな評価をしているレビューは，レーティングとの一貫性が取れており，商品のどこが良くてどこが悪いのかを明確にユーザに伝えることができている可能性が高い．よって，レーティングの根拠となる文が示されているとき，有用性は高いと評価する．図 3.1 の例では，レーティングは 5 段階中 4 の評価だがレビュー中の「便利です」や「残念でした」といった評価表現が，レーティングは高いが最高点ではないことの根拠とみなせる．

観点 6：分量が多い

一言二言のごく簡単な感想を書いた短いレビューより，商品に対する意見や評価を詳細に書いた長いレビューの方が情報量が多い可能性が高いため，有用性が高いと評価する．

観点 7：読みやすい文章である

レビューの内容が詳細で多くの情報を含んでいたとしても，文章が難解であればユーザの理解を妨げる可能性があり，結果として有用なレビューではないと考えられる．よって読みやすい文章で書かれている文章ほど有用性が高いと評価する．

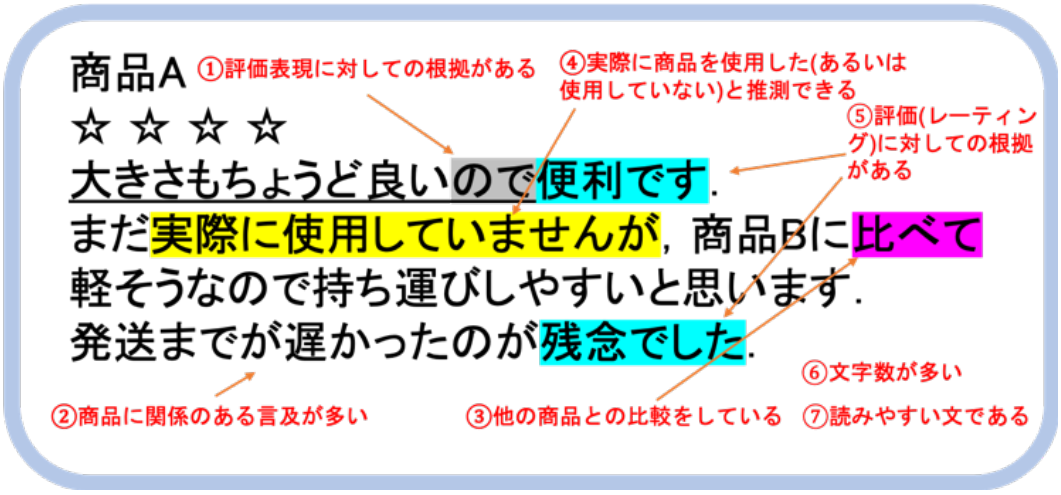


図 3.1: レビューの例

3.2 提案システムの概要

カスタマーレビューを複数の観点で評価して，ユーザに評価結果を提示するシステムの概要図 3.2 に示す．

特定の商品に付いたレビュー文の集合を入力として，個々のレビューの有用性を 7 つの観点でそれぞれ評価する．有用性を評価するシステムは観点毎に独立している，それぞれの観点に合わせた評価方法を考案し，システムを構築する．全てのレビュー文の評価結果を取得した後，その評価結果とユーザが重視する観点を入力としてレビューのフィルタリングを行い，ユーザの重視する観点に合わせてレビューの最終的な有用性の評価を行う．例えばユーザが商品の感想だけでなく根拠を示していることを重視する場合，観点 1 の評価結果が高いレビューを優先的に有用性が高いレビューと評価するようなフィルタリングを行う．最後にフィルタリングによって有用性が高いと評価したレビューをユーザに表示する．

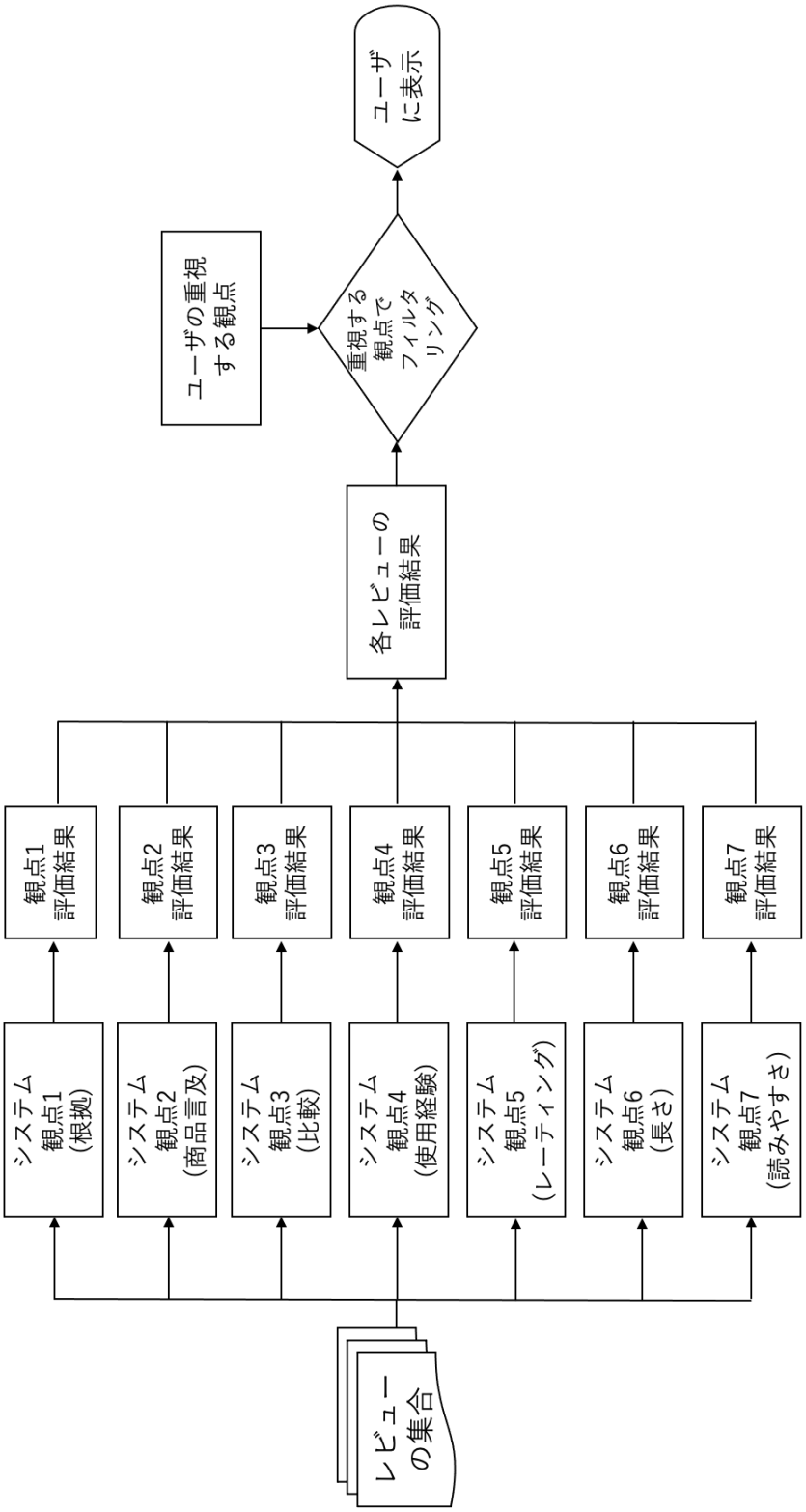


図 3.2: システムの概要図

第4章　評価の根拠を含む文の検出

本章では，3 章で提案した観点 1(評価表現に対する根拠がある) に基づいてレビューを評価する手法について述べる．4.1 節では，観点 1 からみて有用なレビューとは何か考察し，本研究で観点 1 からの有用性をどのように評価するかについて方針を述べる．4.2 節では，4.1 節で述べた方針に基づいて，観点 1 を評価する具体的な手法について述べる．最後に 4.3 章では，提案した手法の評価実験について述べる．

4.1　評価に対する根拠を含むレビューの考察

本節では，観点 1(評価表現に対する根拠がある) に基づいてレビューを評価するために，どのような文が評価に対する根拠を含むか，あるいは含まないのかを考察する．

実際に，評価に対する根拠があると考えられる例文を以下に示す．

例文 4.1: 明るいし温かい色なので満足です

例文 4.2: LED らしくなくて仲間から好評でした

例文 4.1 では「満足です」という評価表現に対して,「明るいし温かい色なので」という根拠が示されている．このレビューの商品に対する評価は「満足」であり，そのような評価になった理由は「明るいし温かい色」だからであるということが読み取れる．例文 4.2 でも「仲間から好評でした」という評価表現に対して,「LED らしくなくて」という根拠が示されている．このレビューの商品に対する評価は好意的であり，好意的評価をした理由は「LED らしくない」ところであるということが読み取れる．

以上の考察から，観点 1 から見て有用なレビューが満たすべき条件として以下の 2 つが考えられる．

(条件 4.1): 商品を評価する表現がある

(条件 4.2): (条件 4.1) の評価に至った理由が述べられている

これらの条件は論理積である．つまり (条件 4.1) と (条件 4.2) の両方を満たすとき，そのレビューは評価に対する根拠が示されていると言える．

次に，評価に対する根拠が示されていないと考えられる例文を以下に示す．

例文 4.3: 電気スタンドに使いましたがやっぱりこれにしてよかったです。

例文 4.4: 電動ドライバーでは外れないネジも簡単に取ることが出来ました

例文 4.3 では「よかったです」という評価表現が記述されているが，これに係る文は「電気スタンドに使いましたが」であり，レビュワーが商品をどう使ったかという事実を述べているが，商品が良かった理由を述べていない．例文 4.4 は「ネジも簡単に撮ることが出来ました」という事実が述べられているが，その事実に対するレビュワーの評価が述べられていない．

例文 4.3，例文 4.4 における考察から，観点 1 から見て有用でないレビューが満たす条件として以下の 2 つが考えられる．

(条件 4.3): 商品の評価だけ述べられていて，その理由がない

(条件 4.4): 商品に関する事実のみ述べられていて，評価がない

これらの条件は論理和である．つまり (条件 4.3) か (条件 4.4) のどちらか一方が満たされれば，そのレビューは根拠を伴う評価を含まないと言える．

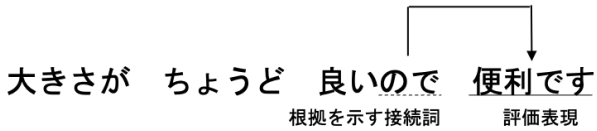
上記の考察を踏まえ，本研究における観点 1(評価表現に対する根拠がある) からのレビューの有用性の評価は,「入力のレビュー文が単に商品に関する事実，または評価のみを述べるだけでなく，商品を評価する表現があり，かつその評価に至った理由が述べられているレビューかどうか判定すること」とする．これを実現するために，評価表現とその根拠を示す表現の両方を含む文をレビューの中から検出し，検出に成功すれば，そのレビューは観点 1 から見て有用であるとみなす．

最終的なシステムでは，個々のレビューに対し，それが根拠を伴う評価表現を含むか否かの情報をユーザに提供する．商品に対する他者の評価の詳細を知りたいユーザは，観点 1 から見て有用であると判定されたレビューを優先的に閲覧することにより，有用なレビューを見つけやすくなる．

4.2 根拠文の検出手法

この節では，4.1 節で述べた方針に基づいて，観点 1 を評価する手法について述べる．具体的には，レビュー文が商品に対する評価とそれに対する根拠の両方を含むか判定する．以下，簡単のため，(例文 4.1) や (例文 4.2) のように評価表現とそれに対する根拠が示されている文を単に「根拠文」と呼ぶ．

基本的な考えとしては，評価表現とその根拠をつなげる役割を持つ可能性が高いと考えられる接続詞のキーワード (「ので」,「ため」など) で終わる文節が評価表現 (「便利です」など) を含む文節に係る時，そのレビュー文は根拠を含むと判定する．以下に例を挙げる．



この文では,「便利」が評価表現であり,「ので」が根拠を示す接続詞である,「ので」を含む文節『良いので』が評価表現を含む文節『便利です』に係るため,「ので」より以前の節が評価表現の根拠を示しているとみなせる.

根拠を示す接続詞のキーワードのリストはあらかじめ人手で作成した．作成したキーワードを図 4.1 に示す.

のが，ので，のは，為，ため，点が，くて，のも，ところが，ところも

図 4.1: 根拠を示す接続詞

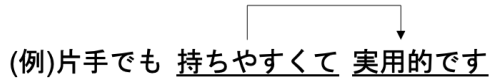
また評価表現の検出には評価表現辞書や感情分析 API を用いる．詳細は後述する.

上記の条件を満たす文を検出するシステムを簡易的に実装し，実際に根拠文を検出する予備実験を行った．その結果，根拠を示すキーワードのリストや評価表現のリストが不十分であり，実際に評価表現に対する根拠を含む文を十分に検出できないことがわかった．そこで，上記の条件を緩和し，以下のいづれかの条件を満たす文を根拠文として検出する.

条件 1：連用形 → 評価表現

用言の連用形 (連用接続) を含む文節が評価表現を含む文節に係る.

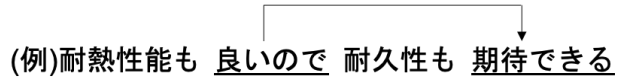
以下はこの条件を満たす文の例である.『持ちやすくて』は用言「やすい」の連用形を含む文節であり，それが評価表現「実用的」を含む文節『実用的です』に係っている.



条件 2：キーワード → 用言

図 4.1 の根拠を示す接続詞で終わる文節が用言を含む文節に係る.

以下はこの条件を満たす文の例である．根拠を示す接続詞「ので」で終わる文節『良いので』が用言「できる」を含む文節『期待できる』に係っている.



上記を簡単に説明すると，根拠を示す文節と評価表現を含む文節の両方を含むという条件ではなく，どちらか一方を含めば根拠文として検出するというように条件を緩和している.

評価表現の検出は以下の 2 通りの方法で行う.

1. 日本語評価極性辞書 (用言編)[8] に含まれる評価表現との一致

   日本語評価極性辞書 (用言編) は用言を中心に収集した評価表現約 5 千件からなる辞書である．この辞書では評価表現はフレーズと呼ばれる．表 4.1 に示すようにそれぞれのフレーズに「経験」,「評価」と「ポジティブ」,「ネガティブ」のラベルが付いている．ただし，根拠文の検出にはこれらのラベルは用いずに，文節中に辞書内のフレーズと一致する文字列があるかどうかで評価表現を検出する．

表 4.1: 日本語評価極性辞書 (用言編) の抜粋

ラベル	フレーズ
ネガ（経験）	くやむ
ネガ（評価）	分かりづらい
ポジ（経験）	助かる
ポジ（評価）	ちょうどいい

2. COTOHA API[3] による評価表現の検出

   COTOHA API は構文解析，固有表現抽出，類似度算出などの様々な自然言語処理機能を提供している API である．今回は感情分析 API を使用する．文を入力すると，図 4.2 に示すように書き手の感情 (Positive,Negative,Neutral), 0 から 1 までのセンチメントスコア (1 に近づくほど判定結果が確からしいことを示す), 感情語のリストを返す．根拠文の検出には，書き手の感情，センチメントスコアは用いず，感情語のリスト (図 4.2 における emotional_phrase) から感情語を抽出し，これを評価表現とする．

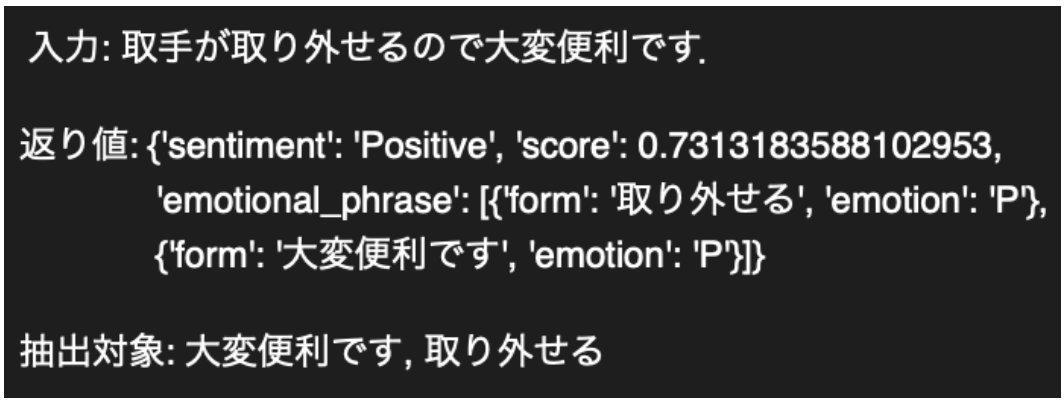


図 4.2: COTOHA 感情分析 API の実行例

上記の説明のまとめとして，入力のレビュー文に評価の根拠が含まれているかを判定する手順を示す．

1. レビュー文を文単位で分割する
2. CaboCha[10] を用いて文に対して文節の係り受け解析を行う．CaboCha は形態素解析も同時に行うため，文中の単語の品詞や活用形も解析される．
3. 評価表現もしくは用言 (動詞，形容詞，形容動詞) を含む文節 E を検出する．
4. 3 で検出した文節 E を直接の係り先とする文節 R を抽出する．
5. 4 で検出した文節 R の中に図 4.1 の根拠を示すキーワードが含まれるかどうかを判定する．
6. 文節 R と E が上記の**条件 1**，**条件 2** のいずれかを満たす時，評価に対する根拠を含む文と判定する．

4.3 根拠文検出の評価

4.3.1 実験の手順

Amazon にて LED 電球，ロボット掃除機, インパクトドライバー，洗顔料の製品について投稿されたレビューをランダムで 50 件取得した．データの内訳を表 4.2 に示す．

これらのレビューに対して，評価の根拠を含む文が現れるかを作業者 2 名が独立に判定した．このとき，レビュー中にひとつでも評価の根拠を含む文があればそのレビューは評価の根拠ありとして評価した．また，提案手法は評価とその根拠を含む文 (根拠文) を抽出する手法であるので，商品に対する評価とその根拠が複数の文に書かれている場合は評価の根拠ありと判定せず，一つの文に書かれている場合のみ評価の根拠ありと判定した．評価者 2 人の判定の対応関係を表 4.3 に示す．2 者の判定の一致率は 0.76，κ 係数は 0.56 であった．

表 4.2: 根拠文検出のテストデータの内訳

製品	レビュー数
LED 電球	22
ロボット掃除機	8
電動ドライバー	10
洗顔料	10
合計	50

この評価データに対し，4.2 節で述べた手法で評価の根拠を含む文を検出した．本実験は評価の根拠を含むレビューを検出するタスクであるので，評価基準は精度，再現率，F 値とする．さらに，この実験はレビューを「根拠あり」と「根拠

表 4.3: 評価者 2 人の評価の分割表

		評価者 B	
		根拠あり	根拠なし
評価者 A	根拠あり	17	8
	根拠なし	3	22

なし」に分類する 2 値分類タスクと見なすこともできるため，2 値分類の正解率も評価基準とした．実験結果は表 4.4 の混同行列のように表すことができる．ここで Positive は「根拠あり」の判定, Negative は「根拠なし」の判定を表す．この混同行列から，精度，再現率，F 値，正解率はそれぞれ以下の式 (4.1),(4.2),(4.3),(4.4) で表される．

表 4.4: 根拠文検出の実験における混同行列

		予測された判定	
		Positive	Negative
評価者による判定	Positive	真陽性 (True Positive)	偽陰性 (False Negative)
	Negative	偽陽性 (False Positive)	真陰性 (True Negative)

$$\text{精度} = \frac{\text{真陽性}}{\text{真陽性} + \text{偽陽性}} \tag{4.1}$$

$$\text{再現率} = \frac{\text{真陽性}}{\text{真陽性} + \text{偽陰性}} \tag{4.2}$$

$$F \text{ 値} = 2 \times \frac{\text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}} \tag{4.3}$$

$$\text{正解率} = \frac{\text{真陽性} + \text{真陰性}}{\text{真陽性} + \text{真陰性} + \text{偽陰性} + \text{偽陽性}} \tag{4.4}$$

4.3.2 結果と考察

システムの予測と，評価者 A,B それぞれの判定との混同行列を以下の表 4.5,4.6 に示す．評価者 A,B のそれぞれについて，それを正解とした時の精度，再現率，F 値，正解率を表 4.7 に示す．再現率は比較的高いが，精度は低い傾向が見られる．特に評価者 B の精度が 0.45 であり，F 値も 0.6 程度に留まっている．

次に，根拠文の有無の判定に正解した例，しなかった例を分析する．表 4.8 に正解または不正解となったレビューの例を示す．

正解例の 1 つ目は条件 2(キーワード→用言) に当てはまる．1 番目の文において，根拠を示す接続詞のキーワード「ので」を含む文節が文節『暑くない』に係ってお

表 4.5: システムの予測と評価者 A の判定の混同行列

		予測された判定	
		Positive	Negative
評価者 A による判定	Positive	20	5
	Negative	13	12

表 4.6: システムの予測と評価者 B の判定の混同行列

		予測された判定	
		Positive	Negative
評価者 B による判定	Positive	15	5
	Negative	18	12

り，この文節は『暑く (形容詞) ない (助動詞)』の様に形態素解析され，用言を含むので，根拠を含む文と判定できている．また，2 番目の文もキーワード「ので」を含む文節が文節『ない』に係っており,「ない」は形容詞 (用言) なので，根拠を含む文と判定できる．正解例の 2 つ目は条件 1(連用形→評価表現) に当てはまる．文節『明るくて』における「明るく」は連用テ接続であり，この文節の係り先の文節『良いです』の「良い」が評価表現として検出されたため，根拠を含む文と判定できている．

一方，不正解例の 1 つめは評価者による判定は根拠ありだったが，システムによる予測では根拠なしと判定した例 (偽陰性の例) である．評価者による判定では,『いい感じです』が評価表現を『レトロで』がその根拠を表すと解釈し，根拠ありと評価していたが，構文解析の結果によると,『レトロで』の「レトロ」は名詞と判定され，用言ではない．また，根拠を表す接続詞のキーワードのリストに「で」はないため,『いい感じ』という評価表現があっても『レトロで』を根拠だと検知することができなかった．

不正解例の 2 つめは評価者による判定は根拠なしだったが，システムによる予測では根拠ありと判定した例 (偽陽性の例) である．システムが「楽しみ」を評価表現として検知し,『まだ実作業をしていませんが』を根拠だと誤判定した.「楽しみ」を評価表現と検知したことは問題ないが，評価表現を含む文節を係り先に持つ文節『していませんが』の「い」が「いる」の連用形であることから条件 1(連用形→評価表現) に当てはまり，根拠ありと判定してしたことが問題である．

以上に示した不正解のレビューの例から誤りの要因を分析すると，偽陰性の対策として，評価の根拠が用言ではなく名詞で表されている場合に対応する必要があると考えられる．また，表 4.8 に示したレビューの例の他に，商品を評価していると思われる表現が評価表現として検出されないことが誤りの要因となった例もいくつかあり，評価表現の検出方法についてもさらに検討する必要があると考えられる．

表 4.7: 評価の根拠を含む文の検出の評価結果

	精度	再現率	F 値	正解率
評価者 A	0.61	0.80	0.68	0.64
評価者 B	0.45	0.75	0.57	0.54

表 4.8: 正解, 不正解のレビュー例

正解例	1	発熱しないので夏は暑くない。LED が見えないので節電になっているか実感がない。
	2	明るくてとても良いです
不正解例	1	直視すると目がチカチカしますが、遠目に見るととってもレトロでいい感じです。
	2	ドリルドライバーがとても使いやすかったので、インパクトも購入。 まだ、実作業をしていませんが、使うのが楽しみです。

偽陽性の対策として，条件 1(連用形→評価表現) における根拠部の検出が連用形の用言があるかどうかのみでは不十分な例がいくつかあったため，この条件を再考する必要がある．また，条件 2(キーワード→用言) について根拠を表すキーワードのリストについて再検討する必要がある．キーワードの見直しや，単にキーワードの出現だけでなくその文脈も考慮して接続詞が根拠を表す場合を厳密にチェックする必要がある．

第5章　商品への言及度の算出

本章では，3 章で提案した観点 2(商品に関係のある言及が多い) に基づいてレビューを評価する手法について述べる．5.1 節では，商品への言及が多いレビューの特徴を分析し，これを踏まえて観点 2 からレビューの有用性を評価する方針を示す．5.2 節では 5.1 節で述べた方針に基づいて，観点 2 からレビューの有用性を評価する具体的な手法について述べる．最後に 5.3 節では，提案した手法の評価実験について述べる．

5.1　レビューにおける商品への言及に関する考察

本節では，観点 2(商品に関係のある言及が多い) に基づいてレビューを評価するために，どのようなレビューが商品に関係があるか，反対にどのようなレビューが商品に関係がないのかを考察する．

商品に関係がある言及を含む可能性が高いと考えられるレビューの特徴として以下の 2 つがあげられる．

(特徴 5.1): 商品自体の性能や性質について言及している

(特徴 5.2): 商品によって何らかの影響を与える，または与えられる人やもの，行動，状態などについて言及している

特徴 5.1 を満たすものの例として，パソコンに対する「CPU」,「メモリ」,「ハードディスク容量」「重量」などについて説明しているレビューが挙げられる．商品自体がどのようなものかを説明している文章は，商品に関係があるレビューである可能性が高い．また，特徴 5.2 を満たすものの例としてドッグフードに対する「犬」,「食いつき」,「健康状態」などについて書かれているレビューが挙げられる．商品自体の説明ではないが，商品によって何かしらの影響が与られるものについて書かれている文章は商品に関係があるレビューである可能性が高いと考えられる．

一方で，商品に関係がない言及を含む可能性が高いと考えられるレビューの特徴として以下の 2 つが考えられる．

(特徴 5.3): 評価の対象が商品ではない

(特徴 5.4): 言及している性能や性質と商品の間の関連性が低い

特徴 5.3 に当てはまるレビューの例として配送に関する言及やショップに関する言及などがある.「商品に関係のある言及が多い」という観点から見ると，商品と直接関係のない事柄への評価は有用な評価であるとは考えにくい．特徴 5.4 について，例えばパソコンに対するレビューの中で「味」や「栄養価」について言及している場合，その言及は商品に関係のない言及である可能性が高いと考えられる．

以上の考察を踏まえ，観点 2(商品に関係のある言及が多い) からのレビューの有用性の評価は,「レビューがどれだけ商品自体の性能や性質，または商品によって影響を与える，または与えられる事について言及しているか，加えて言及している事がどれだけ商品との関係が深いかを測定すること」であるとする．これを実現するために，レビューにおいて，商品に関する言及がどれだけ多いかを定量化し，スコアとして表現する．以下，これを「商品言及度」と呼ぶ．商品言及度及びその算出方法の詳細は次節で述べる．

最終的なシステムでは，個々のレビューに対して商品言及度を算出し，レビューを商品言及度の順にソートしたり，ある閾値以上の商品言及度を持つレビューのみ表示するフィルタリング機能をユーザに提供する．商品に関する説明や意見を重点的に知りたいユーザは，商品言及度が高いレビューを優先的に閲覧することで，有用なレビューを見つけやすくなる．また，商品以外のこと，例えば配送業者や EC サイト自体の評判も知りたいユーザは，商品言及度を参考にしないこともできる．

5.2 言及度の算出

この節では，5.1 節で述べた方針に基づいて，観点 2 を評価する具体的な手法について述べる．既に述べたように，本研究では商品レビューが評価対象の商品について言及している度合いを「商品言及度」(以下，単に言及度と記す) と定義し，これを推定する．

言及度を算出するために，レビュー中にレビューが属する商品カテゴリと関係が深い単語がどれほど出現するかを計算する．レビューが属する商品カテゴリと関係が深い単語とは，例えばペットカテゴリであれば,「大型犬」,「フード」,「食いつき」などが該当する．これらの単語はペットに関連する商品に言及するときによく使われると考えられる単語である．このような単語が多く出現するレビューほど商品に対する言及度が高いと言える．

レビューの言及度は，あらかじめ商品カテゴリ毎にキーワードの重要度を計算した辞書を用いて算出する．キーワードの重要度は TF-IDF に基づいて計算する，以下の式 (5.1) に製品カテゴリ c におけるキーワード k の重要度 ($sig(k,c)$ と記す) を定義する．

$$sig(k,c) = tf_{kc} \cdot \log \frac{N_c + 1}{cf_k} \tag{5.1}$$

ここで，tf_{kc} は商品カテゴリ c 内の文書に出現する全てのキーワードの出現頻度に対する商品カテゴリ c 内の文書に出現するキーワード k の出現頻度の比 (相対出現頻度) である．N_c は商品カテゴリの総数である．cf_k はキーワード k が出現する商品カテゴリの数であり，商品カテゴリ内の文書の中に一度でもキーワードが出現するカテゴリの数をカウントする．

重要度の算出には楽天データ [11] における楽天市場データセットの商品説明文の文書の集合を用いる．商品カテゴリとして，Amazon における商品カテゴリを基に,「本・コミック・雑誌」「DVD・ミュージック・ゲーム」など，18 のカテゴリセットを定義した．また，Amazon における商品カテゴリと楽天市場における商品カテゴリは異なるため，両者のカテゴリの対応表を作成した．表 5.1 に，本手法における 18 のカテゴリのリストと，それに対応する楽天データセットのカテゴリを示す．楽天データセットにおける商品カテゴリは階層的になっており，最上位カテゴリとその子カテゴリについて対応表を作成した．

楽天データセットにおける商品説明文を形態素解析し，名詞および複合名詞 (連続する 2 つ以上の名詞を連結したもの) を抽出する．これらが重要度の辞書の登録単語 (キーワード) となる．個々のキーワードに対し，18 のカテゴリの説明文の集合における出現頻度や，そのキーワードが 1 回以上出現するカテゴリの数を求める．最後に，それぞれのカテゴリ毎に，キーワードの重要度を式 (5.1) にしたがって計算する．以上の方法で商品カテゴリ毎の単語の重要度の辞書を作成した．

次に，言及度の計算方法を説明する．商品カテゴリ c に属するレビュー r の言及度を以下の式 (5.2) のように定義する．

$$Ex(r,c) = w \sum_{k \in K_r} sig(k,c) + (1-w) \log len(r) \tag{5.2}$$

ここで，K_r は入力のレビュー中に出現するキーワードの集合であり $len(r)$ はレビューの長さ (文字数)，w は重みである．すなわち，レビュー内に出現する全てのキーワードの重要度の総和とレビューの長さの重み付き和を言及度とする．また，重み w は実験的に決定する．

まとめとして，入力のレビュー文の言及度を算出する手順 (フローチャート) を図 5.1 に示す．

5.3 評価

5.3.1 実験の手順

Amazon に実際に投稿されたレビューを用いて評価データセットを作成した．言及度算出に用いた商品カテゴリーと同じ 18 のカテゴリからそれぞれ 20 件ずつ計 360 件のレビューを取得した．次に，同じ商品についてのレビュー同士でペアを作り，計 180 組のレビューの組を作成した．このデータセットに対して作業者 2 名が

表 5.1: 本手法のカテゴリと楽天データセットのカテゴリの対応表

本手法のカテゴリ	楽天 最上位カテゴリ	楽天 子カテゴリ
本・コミック・雑誌	本・雑誌・コミック	
DVD・ミュージック・ゲーム	CD・DVD・楽器	
	おもちゃ・ホビー・ゲーム	テレビゲーム
家電・カメラ・AV 機器	家電	
	TV・オーディオ・カメラ	
	美容・コスメ・香水	美容機器・脱毛
	インテリア・寝具・収納	ライト・証明
パソコン・周辺機器	パソコン・周辺機器	
PC ソフト	パソコン・周辺機器	PC ソフト
文房具・オフィス用品	日用雑貨・文房具・手芸	
	家電	電子辞書・FAX・電話
	インテリア・寝具・収納	デスク
		イス・チェア
		オフィス家具
		インテリア・寝具・収納
	本・雑誌・コミック	カレンダー・ポスター・パンフレット
ホーム&キッチン	インテリア・寝具・収納	
	キッチン用品・食器・調理器具	
	日用品雑貨・文房具・手芸	タオル・バス用品
		日用品・生活雑貨
		洗剤・柔軟剤
		防災関連グッズ
		手芸・クラフト・生地
	おもちゃ・ホビー・ゲーム	アート・美術品・骨董品・民芸品
	家電	住宅設備家電
DIY・工具・ガーデン	花・ガーデン・DIY	
ペット	ペット・ペットグッズ	
食品&飲料	スイーツ・お菓子	
	水・ソフトドリンク	
	ダイエット・健康	健康食品
	食品	
お酒	日本酒・焼酎	
	ビール・洋酒	
ドラッグストア	ダイエット・健康	
	医薬品・コンタクト・介護	
ビューティストア	美容・コスメ・香水	
	ダイエット・健康	デンタルケア
	日用品雑貨・文房具・手芸	アロマ・癒しグッズ
ベビー・おもちゃ・ホビー	おもちゃ・ホビー・ゲーム	
	キッズ・ベビー・マタニティ	
	本・雑誌・コミック	絵本・児童書・図鑑
	CD・DVD・楽器	楽器
服・シューズ・バッグ・腕時計	靴	
	メンズファッション	
	ジュエリー・アクセサリー	
	バッグ・小物・ブランド雑貨	
	レディースファッション	
	インナー・下着・ナイトウエア	
	腕時計	
	キッズ・ベビー・マタニティ	
スポーツ&アウトドア	スポーツ・アウトドア	
	ダイエット・健康	リラックス・マッサージ用品
		矯正グッズ
	靴	レディース靴
		メンズ靴
車&バイク・産業・研究開発	車用品・バイク用品	
	車・バイク	
	花・ガーデン・DIY	
クレジットカード	学び・サービス・保険	

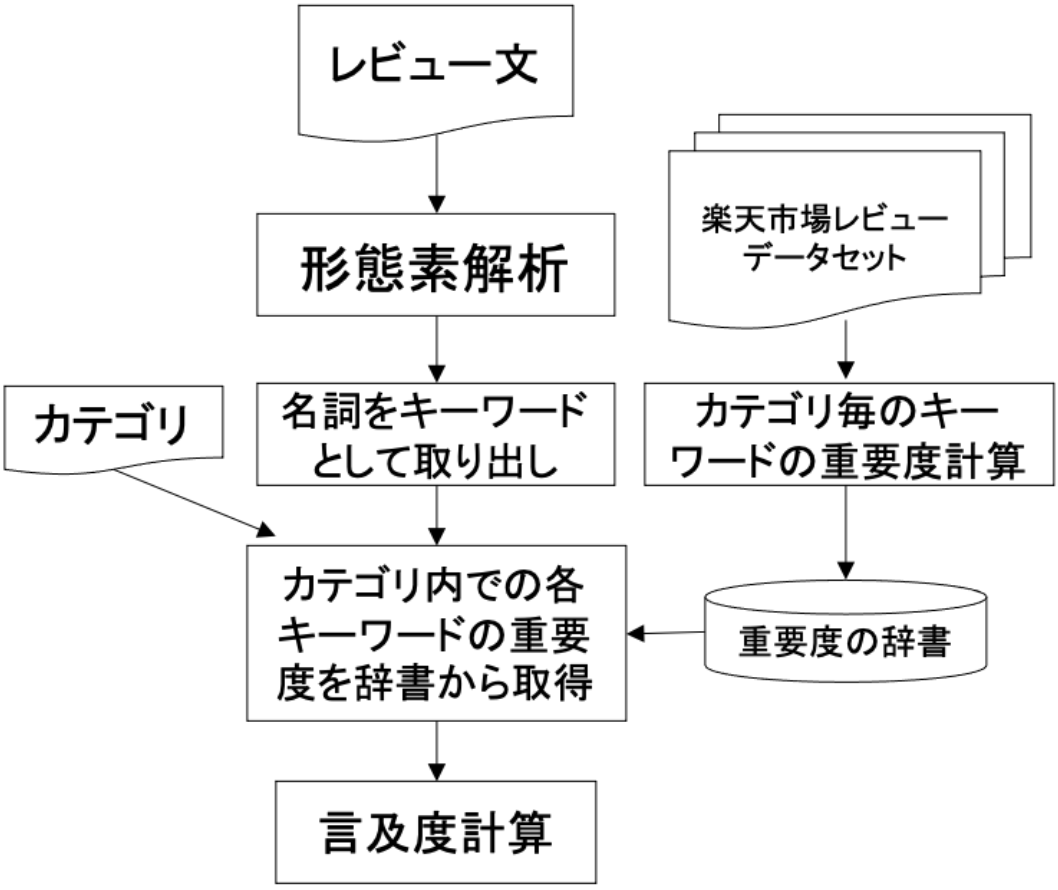


図 5.1: 商品への言及度の算出手順

独立に，レビューの組のどちらがより多く商品に言及しているかを判定した．ただし，商品への言及が多いレビューがどちらかの判定が難しい場合は「不明」とした．

表 5.2 にレビューの組とその判定の例を示す．同じ行にあるレビュー A 列とレビュー B 列のレビュー同士が比較の対象となる．最初のレビューの組では，評価者 1 はレビュー A の方が言及度が高いと判定しているが，評価者 2 はレビュー B の方が高いと判定しており，判定が分かれている．2 番目のレビューの組では，評価者 2 はレビュー B を選んでいるが，評価者 1 は 2 つのレビューの言及度は同程度と判断している．

2 者の判定の結果を表 5.3 に示す．表 5.3 のデータの内，評価者のいづれか 1 名が「不明」とした 24 組のデータを除き，残りの 156 組のレビューの組を評価データセットとした．評価データセットにおける 2 者の判定の一致率は 0.904，κ 係数は 0.807 であった．表 5.2 では 2 者の判定が一致していない例を紹介したが，実際には一致率は高く，判定が分かれることは少ない．

表 5.2: レビューの組と判定の例

レビュー A	レビュー B	評価者 1	評価者 2
塗りやすくて消臭力があってとても使えます。制汗性については微妙。	手放せません。消臭効果抜群。一日中匂いません。ずっと使い続けます	レビュー A	レビュー B
甘いだけの梅酒ではなく穂のかな木の香り。とても美味しく　頂けました。	少し甘めで芳醇な香り、そこらへんの梅酒とは一線を画す美味しさです	不明	レビュー B

表 5.3: 評価者による言及度が大きいレビューの判定結果

		評価者 2 の判定		
		レビュー A	レビュー B	不明
評価者 1 の判定	レビュー A	76	7	2
	レビュー B	8	65	3
	不明	8	5	6

この評価データセットに対し，5.2 節で提案した手法で言及度を算出し，言及度の大きい方のレビューを商品への言及が多いレビューと予測した．この予測タスクは A か B のどちらかを選ぶ 2 値分類問題であることから，評価指標として正解率を用いた．このタスクの実験結果は表 5.4 の混同行列のように表すことができる．ここで，正解数はレビュー A, B のいづれかで予測と正解ラベルが一致した数である．また，不正解数 BA または AB は予測がレビュー A で正解ラベルがレビュー B または予測がレビュー B で正解ラベルがレビュー A であった数を表す．この混同行列から，正解率は式 (5.3) で算出される．

表 5.4: 言及度の大きいレビューの判定の混同行列

		予測された判定	
		レビュー A	レビュー B
評価者による判定	レビュー A	正解数 (A)	不正解数 (AB)
	レビュー B	不正解数 (BA)	正解数 (B)

$$\text{正解率} = \frac{\text{正解数}\,(A) + \text{正解数}\,(B)}{\text{正解数}\,(A) + \text{正解数}\,(B) + \text{不正解数}\,(AB) + \text{不正解数}\,(BA)} \tag{5.3}$$

提案手法をベースラインと比較した．ベースラインは，常に文字数が多いレビューを選ぶ手法とした．これは式 (5.2) の w を 0 と設定した提案手法に相当する．

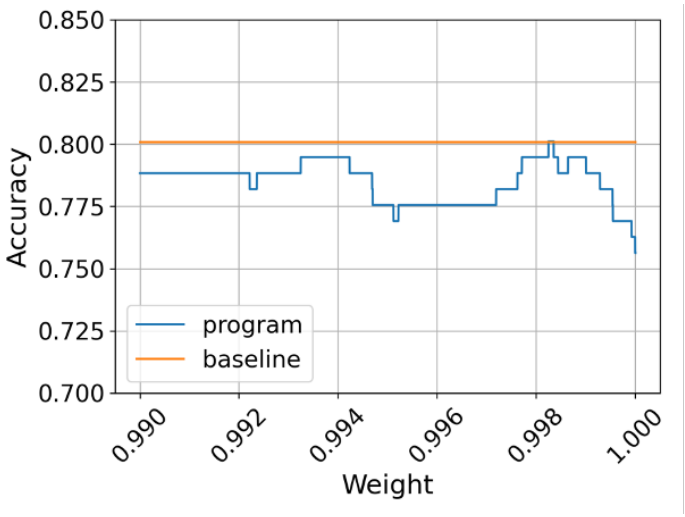
5.3.2 結果と考察

重み w を 0.990 から 1.000 まで 10^{-6} 刻みで変動させた時の正解率の変化を図 5.2 に示す．図 5.2 のオレンジ色の直線がベースラインの正解率で，青い線が提案手法の正解率である．また，ベースラインの正解率，提案手法の正解率の最大値，正解率が最大のときの重み w の値を表 5.5 に示す．提案手法の正解率は，評価者 1，評価者 2 のいずれについても，$w = 0.9983$ のときに最大となった．$w = 0.9983$ の時のシステムの予測と評価者 1,2 による判定の混同行列を表 5.6 と表 5.7 に示す．提案手法の正解率は，0.801(評価者 1) もしくは 0.769(評価者 2) となった．評価者 1 についてはベースラインの正解率と同じだが，評価者 2 についてはベースラインを 0.029 ポイント上回った．提案手法による観点 2(商品に関係のある言及が多い) からのレビューの評価の正確性がベースラインと同等もしくはそれ以上であることを確認した．ただし，パラメタ w の最適化は本来はテストデータとは別の開発データで行うべきである．開発データを用いた w の最適化は今後の課題である．

表 5.5: 商品への言及度の評価結果

	ベースライン	提案手法	正解率が最大時の重み w
評価者 1	0.801	0.801	0.9983
評価者 2	0.740	0.769	0.9983

(a) 評価者 1



(b) 評価者 2

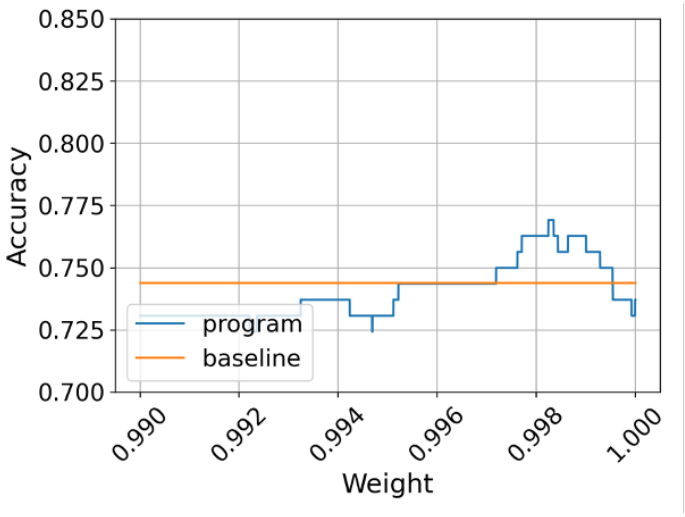


図 5.2: 重み w に対する正解率の変化

表 5.6: システムの予測と評価者 1 の判定の混同行列

		予測された判定	
		レビュー A	レビュー B
評価者 1 による判定	レビュー A	64	20
	レビュー B	11	61

表 5.7: システムの予測と評価者 2 の判定の混同行列

		予測された判定	
		レビュー A	レビュー B
評価者 2 による判定	レビュー A	61	22
	レビュー B	14	59

次に，商品への言及が多いレビューの予測が正解した事例，不正解だった事例を分析する． 表 5.8 に正解または不正解のレビューの組の例を示す．これらの例は，正解例，不正解例いづれもペットカテゴリーに属するレビューの組である．

正解例のレビューについて，2 つのレビューに出現したキーワードとその重要度，レビュー毎の重要度の合計を表 5.9 に示す．評価者による判定ではレビュー A,B どちらとも商品に関係のない言及がなく，全ての文が商品に言及しているとみなせたが，レビュー A の方が文数が多く言及している事柄が多いため，レビュー A の方が商品への言及が多いという判定であった．システムの予測でも，キーワード数がレビュー A の方が多いこともあり，重要度の合計はレビュー A の方が高くなっている．また文字数もレビュー A の方が多い．したがって，言及度もレビュー A の方が大きく，正しい判定ができている．しかし，抽出したキーワードの中には，「さ」,「よう」などキーワードとしてはふさわしくないと思われるものも含まれており，またそれらのキーワードの重要度が比較的高く算出されているという問題が見つかった．

次に不正解例のレビューについて，2 つのレビューに出現したキーワードとその重要度，レビュー毎の重要度の合計を表 5.10 に示す．評価者による判定ではどちらのレビューも全ての文が商品に関係のある言及であったが，レビュー A の方がフケや油への効果など様々なことに言及している印象を受けたため，このレビューがより商品に言及しているという判定であった．システムの予測では，レビュー A について,「フケ」や「油ギッシュ」などレビュー A で言及されている単語がキーワードとして検出されない，これらは重要度の辞書に登録されていないとことがわかった．例えば,「フケ」は形態素解析によって動詞と解析されたためキーワード (本研究では名詞に限定している) として検出されなかった．一方レビュー B では「犬」という単語の重要度が比較的大きく，このキーワード一つでレビュー A の重要度の合計を上回っている．また，キーワード数についてもレビュー B の方がレビュー A よりも多い．その結果，最終的な判定がレビュー B となり，誤った

判定をしている．

以上の事例の分析から，現時点での提案手法の問題点として，キーワードの検出方法が不十分である点と，重要度の計算においてカテゴリの中で重要な単語の重要度を高く，重要でない単語の重要度を低く計算することができていないことが多いという点が挙げられる．キーワードの検出方法については，名詞をキーワードとするだけでなく，形容詞などの他の品詞もキーワードとして採用し，言及度の計算に用いることが考えられる．重要度の計算については，商品の説明文からキーワードを抽出して重要度を計算するだけではなく，データセット中の商品レビューでよく使われる単語を選択し，これを重要度の辞書に加えるなど，重要度辞書を拡充することを検討する必要がある．また，キーワード間での言及度の差が大きく，突出して重要度の大きいキーワードに重要度の計算が大きく影響されるという問題については，何らかの方法で重要度を正規化するなどの対策も必要である．

表 5.8: 提案システムによる言及度が大きいレビューの判定の例

	レビュー A	レビュー B	正解	予測
正解例	今までホームセンターでペットシーツを購入していましたが枚数が多いものがなく、今回この商品を購入してみました。厚さは非常に薄いと思います。吸水性は用を足してすぐには吸い込まないように感じます。この枚数でこのお値段なので良しとしましょう。	質より量という人向けのペットシーツです。値段と量で満足しました。	A	A
不正解例	脂漏症+フケが多く困っていましたがこのシャンプーで洗い、まずフケが一回目で全くでなくなった。一週間は油ギッシュにならなくなった。普通のシャンプーは 3 日でギッシュ、フケは当日から。。。それを考えるとこのシャンプー良いと思う	皮膚が弱く湿疹ができやすい老犬のシャンプー用に買いました。何回か使っているうちに痒がる仕草が減ってきたように思われます。効果が徐々に表れてきているようです。もう少し使ってみます。	A	B

表 5.9: 正解例のレビューにおけるキーワードと重要度

レビュー A		レビュー B	
キーワード	重要度	キーワード	重要度
商品	0.00149	ペットシーツ	0.00124
ペットシーツ	0.00124	量	0.000196
もの	0.000580	満足	0.000132
さ	0.000437	値段	2.50×10^{-5}
よう	0.000435	質	1.22×10^{-5}
吸水性	0.000200		
ホームセンター	0.000179		
購入	0.000168		
今	9.81×10^{-5}		
非常	7.41×10^{-5}		
用	5.61×10^{-5}		
枚数	5.13×10^{-5}		
値段	2.50×10^{-5}		
今回	4.82×10^{-6}		
重要度合計	0.00503298	重要度合計	0.00160276

表 5.10: 不正解例のレビューにおけるキーワードと重要度

レビュー A		レビュー B	
キーワード	重要度	キーワード	重要度
シャンプー	0.000201	犬	0.00236
脂	7.75×10^{-5}	皮膚	0.000704
それ	4.91×10^{-5}	効果	0.000540
当日	3.66×10^{-5}	よう	0.000435
普通	3.01×10^{-5}	湿疹	5.51×10^{-5}
一週間	8.26×10^{-7}	うち	4.76×10^{-5}
		仕草	9.54×10^{-6}
		何回	2.48×10^{-6}
重要度合計	0.000395374	重要度合計	0.00415326

第6章　比較文の検出

本章では，3 章で提案した観点 3 に基づいてレビューを評価する手法について述べる．観点 3 とは，他の商品との比較を含んでいるレビューは有用であるという観点である．6.1 節では観点 3 からレビューの有用性を評価する方針を述べる．6.2 節では 6.1 節で策定した方針に基づいて，観点 3 を評価する具体的な手法について述べる．最後に 6.3 章では，提案した手法の評価実験について述べる．

6.1　レビューにおける比較の考察

本節では，観点 3(他の商品と比較している) に基づいてレビューを評価するために，どのようなレビューが比較を含むかを分析する．

予備調査として，Amazon の家電とホームカテゴリの商品レビューを 1,201 件取得し，複数の商品を比較しているレビューを人手で探した．170 件のレビューに複数の商品の比較に関する記述があった．次に，これらのレビューから比較を表すと考えられるキーワードを調査し，抽出した．抽出したキーワードの中で出現頻度が多いものを表 6.1 に示す．次に，該当したレビューにおける比較を含む文の特徴を調査した．その結果，比較を含む文やレビューには以下のような特徴があることがわかった．

表 6.1: 比較を表すと考えられるキーワード

キーワード	出現頻度
より	46
比べ	35
方が	16
比較	16
買い替え	14
違い	13

(特徴 6.1): 比較を表す特有のキーワードが現れている．

「比べると」「比較」など，他の商品との比較を示唆するキーワードがある．例文として「A に比べて B の方が良い」や「A より B の方が大きい」などが

挙げられる．これらの例文における「比べて」「方が」といったキーワードは，複数の商品を比較していることを示唆すると考えられる．この 2 つ以外にもこのような比較を表すキーワードが数十個程度見つかった．

ただし，比較を表す文は必ずしも比較を示唆するキーワードを含むわけではない．(特徴 6.1) に当てはまらない例文として「A 社の製品は匂いが強いが，B 社の製品はほぼ匂いがありません」が挙げられる．この文は前述のような比較を示唆するキーワードは出現しないが，明らかに A 社の製品と B 社の製品を比較している．ただし，予備調査の結果では，このような文は少なく，概ね比較文はキーワードを含んでいることがわかった．

(特徴 6.2): 比較を示唆するキーワードの後に評価表現が出現する．

複数の商品を比較する文には，(特徴 6.1) で述べたキーワードの後に評価表現が出現することが多かった．先程の例文「A より B の方が大きい」を見ると,「方が」というキーワードの後に「大きい」という評価表現が出現している．もちろん，評価表現を含まない比較文も存在する．例えば,「A より B だと思う」は，比較を示唆する「より」というキーワードの後に評価表現は出現しない．この文の場合，A より B が何であるのか，どちらを高く評価しているかが不明である．このように比較のキーワードを含んでいるが評価表現がない場合，どちらの評価が高くてどちらの評価が低いのかが分からない，あるいはわかりにくく，有用性を評価するという観点に置いて，比較をしていても有用な文であるとは考えにくい．

(特徴 6.3): レビューの最初の文に「買い替え」などの単語が出現する．

レビューの最初の文に，レビュワーが他の商品からの買い替えで商品を買ったことを明示している場合があり，そのレビュー全体で買い替え前の商品と後の商品を比較していることが多かった．例として,「旧式の B からの買い替えです」という文から始まるレビューがあった．このレビューは，全体として，最初の文に現れる製品 B と現在レビューしている製品を比較していた．

(特徴 6.4): 名詞 + 「より」を含む文節が評価表現に係る．

「より」は比較を表す特有のキーワードの 1 つであるが，商品を比較していない文にも出現する．例えば「何より」「というより」「以前より (from の意味)」といった表現は比較を示唆しない.「より」が商品を比較する文に出現するときは,「A より B が大きい」や「A は B より大きい」のように，評価表現 (この場合は「大きい」) と直接の係り受け関係があることが多かった．

以上の考察を踏まえ，観点 3(他の商品と比較している) からの有用性の評価は，レビューが比較を含むか否かを判定することにより行う．上記で考察した特徴を踏まえ，レビューの中に比較文がある，またはレビュー全体で他の商品との比較をしているとき，レビューが比較を含むと判定する．

今回の調査では，主に比較が明示的に示されている文に対してその特徴を分析した．しかし，他の商品との比較を行う際には，比較を示唆するキーワードを伴わずに暗黙的に比較する場合や，複数の文によって他の商品との違いを説明する場合も見られた．このような比較は検出が難しいと考えられる．本論文では，まず比較が明示的に示されているレビューを検出することに焦点を当てる．暗黙的な比較を含む網羅的な比較の検出は今後の課題とする．

最終的なシステムでは，個々のレビューに対し，それが比較を含んでいるかどうかの情報をユーザに提供する．他の商品との比較を重視するユーザは，比較ありと判定されたレビューを閲覧することにより，有用なレビューを見つけやすくなる．

6.2 比較を含むレビューの検出

本研究では，キーワードマッチングによる簡単なルールベースの手法でレビューが比較を含むかを判定する．策定したルールは以下の 3 つである．

(ルール 1): 比較を表すキーワード + 評価表現

文中に比較を表す特定のキーワードを含み，キーワードの後に評価表現が出現している場合，その文を比較文と判定する．比較を表すキーワードとして図 6.1 に示す 34 個のキーワードを使用した．

文がキーワードを含むかは文字列マッチングにより行う．図 6.1 に示すキーワードの中には「方/が」「が/違う」(/は単語区切りを表す) など，複数の単語で構成されているものもある．単語単位のマッチングでは，このような複数の単語から構成されるキーワードを検出することは難しい．一方，単純な文字列のマッチングでは，動詞などの活用語のマッチングに失敗する可能性がある.「比べる」「違う」などは，文中で「比べて」「違って」のように活用している場合でもマッチさせたい．そこで，キーワードとのマッチングを行うときは，文を形態素解析し，各単語を原型に変換したものを繋げた文字列に対してキーワードの検出を行う．同様に，キーワードも単語毎に原型に変換しそれを繋げて一つのキーワードとする処理を行う．例えば，図 6.1 の「変わるぬ」は，元のキーワードは「変わらぬ」だが，原形「変わる」と「ぬ」を連結した文字列をキーワードとして登録している．

一方，評価表現の検出には日本語評価極性辞書 (用言編)[8] を用いる．文字列のマッチングにより，辞書中の評価表現と一致するものがあるかを検出する．

(ルール 2): レビューの最初の文に特定のキーワードが出現

くらべる，較べる，比べる，比較，違う，方が，ほうが，反面，買い替え，違い，型違い，型落ち，同じ，代替え，変わるぬ，変わるない，と違う，以前より，優れるて，匹敵する，が違う，同程度，良いほう，劣ます，遜色ない，匹敵するた，他社さんは，他の店，他のメーカー，格が違う，今までで一番，他の製品より，では苦手だた，比較になるないほど

図 6.1: 比較を表すキーワード

「買い替え」や「今まで」などのキーワードがレビューの最初の文に出現している場合，レビュー全体で他の商品と比較していると判定する．ルール 2 で使用するキーワードの一覧を図 6.2 に示す．

買い替え, 買替え, 買い換え, 買換え, 以前, 今まで

図 6.2: レビューの 1 文目に出現する比較を表すキーワード

(ルール 3): 「より」 + 評価表現

「名詞 + より」を含む文節の直接の係り先が評価表現を含む文節の場合，その文を評価文と判定する．この時，評価表現の有無の判定にはルール 1 と同じく日本語評価極性辞書 (用言編)[8] を用いる．評価表現の検出は (単語列ではなく) 文字列マッチングにより行う．ルール 1 も評価表現が存在することをチェックしているが，ルール 3 では単に評価表現が存在するだけでなく,「より」を含む文節と直接の係り受け関係があることをチェックする．6.1 節の (特徴 6.4) で述べたように,「ので」は比較を示唆するときもあればそうでないときもあるので，評価表現との係り受け関係を確認することで,「より」が比較を表していることを厳密にチェックする．

判定対象のレビューに対し，上記のルールを用いて比較の有無を判定する．3 つのルールのうちいづれか 1 つでも比較があると判定されれば，そのレビューは比較を含むと判定し，それ以外は比較を含まないと判定する．

6.3 比較検出の評価

6.2 節で述べた比較を含むレビューを検出する手法を評価する．

まず，評価用データを作成する．レビューに対し，それが比較を含むかどうかを人手で判定し，正解ラベルを付与する．本実験では 2 つの評価データを用意した．一つは，比較を表す文の特徴の調査や比較を検出するルールの設計に使用したレビューを用いた評価データである．Amazon の家電カテゴリとホームカテゴリの商品に投稿されたレビューを用いている．この評価データはルールの作成に用いたレビューと同じデータを使うため，クローズドの評価となる．以降，このデータ

を「評価データ C」とよぶ．もう一つは Amazon のベビーカテゴリの商品に投稿されたレビューからなる評価データである．このデータはルールの作成に用いたレビューとは異なるため，これを用いた評価はオープンテストとなる．以降，この評価データを「評価データ O」と呼ぶ．どちらのデータセットもレビューが他の商品と比較をしているかどうかを作業者 1 名が判定した．各データセットのカテゴリ，レビューの件数，比較ありと判定されたレビューの件数を表 6.2 に示す．

表 6.2: 比較の有無の判定手法の評価データ

データセット	カテゴリ	レビュー数	「比較あり」の数
評価データ C	家電，ホーム	1201	176
評価データ O	ベビー	557	37

作成した 2 つの評価データに対して，6.2 節で提案した手法でレビューの中に他の商品との比較があるかを判定し，その結果を評価した．評価基準は精度とした．精度は，提案手法によって比較ありと判定されたレビューにおける正解の割合であり，式 (6.1) のように定義される．既に述べたように，本研究では，まず比較が明示的に示されているレビューを検出することに焦点を当てている．そのため，精度を評価基準とし，再現率は評価基準としない．

$$\text{精度} = \frac{\text{比較ありと正しく判定したレビューの数}}{\text{比較ありと判定したレビューの数}} \tag{6.1}$$

比較を含むレビューの検出精度を表 6.3 示す．検出精度はルール毎に図っている．この表におけるルールは 3 つのグループに別れている．最初のグループはルール 1(キーワードによる検出)，2 番目のグループはルール 2(レビューの最初の文に出現するキーワードによる検出)，3 番目のグループ (最後の行) はルール 3(「より」による検出) の評価結果である．クローズドテストでは，大部分のルールで検知数が 1 以上であった．また，いくつかのルールについては精度が十分に高いことがわかった．これに対し，オープンテストでは検知数が少数，または 0 のルールが多かった．これはベビーカテゴリのデータセットでは正例の数が 37 件しかないことによると考えられる．オープンテストの中で比較的検知数が多いルールについて,「比べる」は高い精度で判定ができているが,「同じ」「より」は 0.4 から 0.5 と精度はあまり高くなく,「違う」については検知数 11 件ある中で 1 件も人手の判定と一致しなかった．とはいえ,「比べる」「比較」など，精度が高いルールもいくつかあった．

以上の結果から，今後の課題として，今回よりも大きなデータセットを作成し，各ルールの精度をさらに正確に評価することが挙げられる．また精度の低い，または検知数が少ないルールを削除することも必要である．加えて表 6.1 で示したような出現頻度が高キーワード中で検知数が低いキーワード，例えば「方が」などについては,「より」のように個別に検知するルールを策定すること必要である．オー

プンテストでの精度が低いのは，比較を示唆するキーワードがドメインによって異なることも原因のひとつであると考えられるので，ドメインに依存しないルールやドメインごとに固有にルールを設計する必要もある．

表 6.3: 比較を含むレビューの検出結果

ルール	評価データ C(クローズド)			評価データ O(オープン)		
	検知数	正解数	精度	検知数	正解数	精度
くらべる	0	–	–	0	–	–
較べる	2	2	1.0	0		
比べる	14	12	0.857	9	7	0.778
比較	13	4	0.308	3	3	1.0
違う	19	8	0.421	9	0	0.0
方が	0	–	–	0	–	–
ほうが	0	–	–	0	–	–
反面	1	1	1.0	0	–	–
買い替え	2	1	0.5	0	–	–
違い	2	1	0.5	0	–	–
型違い	0	–	–	0	–	–
型落ち	0	–	–	0	–	–
同じ	20	12	0.6	8	3	0.375
代替え	0	–	–	0	–	–
変わるぬ	2	2	1.0	2	0	0
変わるない	4	3	0.75	0	–	–
と違う	6	3	0.571	3	0	0.0
以前より	0	–	–	2	1	0.5
優れるて	7	5	0.714	2	2	1.0
匹敵する	0	–	–	0	–	–
が違う	7	2	0.286	1	0	0
同程度	1	0	0.0	0	–	–
良いほう	3	1	0.333	0	–	–
劣るます	3	3	1.0	1	1	1.0
遜色ない	1	1	1.0	0	–	–
匹敵するた	0	–	–	0	–	–
他社さんは	1	1	1.0	0	–	–
他の店	0	–	–	0	–	–
他のメーカー	6	4	0.667	3	2	0.667
格が違う	0	–	–	0	–	–
今までで一番	0	–	–	0	–	–
他の製品より	1	1	1.0	0	–	–
では苦手だた	0	–	–	0	–	–
比較になるないほど	0	–	–	0	–	–
(先頭) 買い替え	24	14	0.583	1	1	1.0
(先頭) 買替え	0	–	–	0	–	–
(先頭) 買い換え	8	4	0.5	1	0	0.0
(先頭) 以前	16	7	0.438	7	1	0.143
(先頭) 今まで	15	5	0.333	0	–	–
より	21	16	0.696	8	4	0.5

第7章　その他の観点からの有用性の判定

4，5，6 章で述べてきた観点 1,2,3 以外の観点については，現時点では評価方法を検討している段階である．

観点 4

観点 4(実際に商品を使用したと推測できる) については，与えられたレビューに対し，それを書いた人が実際に商品を使用したか否かを自動的に推測することで，レビューの有用性を評価する予定であった．その手法を検討するための予備調査を行った．Amazon に投稿された本，電子機器，ホーム・キッチンカテゴリのレビューを 1,385 件取得し，商品を使用していないと思われるレビューを人手で判定し抽出した．その結果，該当したレビューは 11 件のみであった．その全てのレビューを表 7.1 に示す．これは全体の 1%に満たない割合である．商品を使用していないと推測できるレビューの数が予想以上に少なかったことから，この観点について自動評価する手法の検討は行わなかった．しかし，別の商品カテゴリーや Amazon 以外のサイトに投稿されたレビューについては，商品を使用していないと思われるレビューが多く存在している可能性もある．今後も調査を続けたいと考えている．

観点 5

観点 5(評価に対する根拠がある) については，ユーザのレーティングとレビューの内容が一致していないかどうかを評価する．例えば，レーティングが 5 点満点中 5 点なのに，レビューでは商品のことを酷評しているとき，レビューの信頼性が低く，有用でないとする．このために文書の極性判定の手法を利用する．極性判定とは，レビューで表明されている意見が肯定的か否定的か，またその度合 (すなわちレーティング) を予測する手法であり，近年盛んに研究が行われている．極性判定の結果得られたレビューの極性がユーザが与えたレーティングに近いかどうか，また極性判定信頼性が十分に高いかどうかなどの手法によって，この観点からの有用性の評価を行う予定である．

表 7.1: 商品を使用していないと思われるレビュー

日本語訳が下手で、非常に読みにくい。2〜3 ページ読んで、それ以上読む気がしなくなった。
流行り物、意識高い系本という事で洒落で買いましたが、漫画脳男にはつまらなかった。正確には読み飛ばしてます。バカは買ってはいけない本です。
何回も何回も同じ事を繰り返す内容で、読み終えるまでに大変でした。半分読み終えたところで、最後まで読み切る意欲がなくなった。今回は失敗でした。言いたい内容なまぁ納得できる内容です。
なんだか退屈で三分の一も読めなかった
本棚に並べておくだけで、部屋に来た友人から「インテリなんだね」と言われます。良本です。
以前より気になっていた本なので時間をかけて読みたいと思う。
期限通りに届けていただき満足です。まだ全て読んでいませんので星 5 つはつけづらいですが、おもしろいと思います。
ファミリーにプレゼントで読んでない。
文の表現も難しくなく、スラスラ読めます。まだはじめの方しか読んでいませんが、ゆっくりと読み進めたいと思います。
まだ全部は読んでいないけど、なかなか日本だけで子育てをしていて経験できない内容かなと思い、そういう世界もあるのだとは知って、子供にも教えてあげられたらいいなと思い、読み進めてます。
早々に届けて頂きありがとうございました。中身は確認しました、間違いないと思います。忙しくて、未だセットできていません、今度の休みでセットします。

観点 5 からの有用性をユーザに提示する際には，システムで予測した極性値と実際のレビューのレーティングをともに正規化し，極性値をレーティングで割ったスコアを提示することを考えている．このスコアが 1 に近いほどレビューが観点 5 を満たしていることを示すことができる．

観点 6

観点 6(分量が多い) については，文字数や単語数などで比較的容易に定量化できると考えられる．最終的なシステムでは，単純に商品についたレビューの中で評価対象のレビューの文字数が何番目に多いかを提示する方法などが考えられる．また，この観点については他の観点と組み合わせることでレビューの有用性をさ

らに多角的に評価できると考えられる．例えば，観点 2(商品に関係のある言及が多い) と組み合わせれば，1 文あたりの言及度を情報の密度と考えることができる．このように観点を組み合わせてレビューの有用性を評価し，ユーザに提供することも検討したい．

観点 7

観点 7(読みやすい文章である) について，文章の可読性，難易度を測定するいくつかの先行研究があり [9, 15, 12]，これらの技術を適用することで定量化できると考えられる．この観点についても他の観点と組み合わせることで，より有用なレビューを見つけることができると考えられる．例えば，観点 3(他の商品と比較している) からの有用性の自動評価では，専門的なことについて難解な言い回しで比較が行われているレビューを高く評価する可能性があるが，観点 7 による評価結果と組み合わせることで，より理解しやすい文章で他の商品との比較を行っているレビューを見つけ易くすることができると考えられる．観点 6 と同じように，観点 7 についても独立した評価だけでなく，他の観点と組み合わせた評価の有用性についても検討したい．

第8章　終わりに

8.1　まとめ

本研究では，商品レビューの有用性を複数の観点から評価し，その評価結果をユーザに示すシステムの構想を示した．有用性を評価するための 7 つの観点を定義し，その観点で入力のレビューをそれぞれ評価し，その評価結果を掲示することでユーザが自身の嗜好に合わせて有用な商品レビューを見つける作業をサポートすることを狙った．

7 つの観点のうち観点 1(評価表現に対する根拠がある)，観点 2(商品に関係のある言及が多い)，観点 3(他の商品と比較している) の 3 つの観点について，有用性を評価する手法を実現し，その有効性を実験により示した．

観点 1(評価表現に対する根拠がある) からのレビューの有用性を評価する手法として，商品に対する評価とそれに対する根拠の両方が書いてある文を検出する手法を提案した．「ので」など根拠を表す接続詞を含む文節が用言を含む文節に係る，用言の連用形を含む文節が評価表現を含む文節に係る，という 2 つの条件を策定し，レビュー文がこれらの条件を満たすか否かを判定するシステムを実装した．実験の結果，再現率は比較的高いが精度, F 値が低く，評価表現の検出方法や用言による評価部の検出について改善が必要であることがわかった．

観点 2(商品に関係のある言及が多い) からレビューの有用性を評価する手法として，レビューが商品に言及している度合の強さを表す「商品言及度」という新しい概念を提案し，これを算出する手法を提案した．あらかじめ，商品カテゴリ毎に，それに関連の深いキーワードとその重要度からなる辞書を構築した．この重要度辞書を参照し，レビューに含まれる全てのキーワードに対する重要度の和を求めた．最終的に，重要度の和とレビューの長さの重み付き和により言及度を算出した．評価実験では，同じ商品に対する 2 つのレビューの組を作り，どちらが商品により多く言及しているかを判定するタスクを設定し，提案手法で算出した商品言及度の比較により言及の多いレビューを判定した．2 名の評価者が独立に言及の多いレビューを選択し，これを正解として判定の正解率を測った．その結果，一方の評価者を正解としたときの提案手法の正解率はベースラインと同じであったが，もう一方の正解者を正解としたときはベースラインを上回った．

観点 3(他の商品と比較している) からレビューの有用性を評価する手法として，レビューから複数の商品を比較している文を検出する手法を提案した．比較を表すキーワードを計 40 個用意し，それらのキーワードの存在をパターンマッチによっ

てチェックするルールによって比較文を検出した．また，キーワード「より」については，比較文を正確に検出するために，詳細な条件をチェックするルールを設計した．評価実験の結果，クローズドテストではいくつかのルールについては精度が十分に高いことがわかった．これに対し，オープンテストでは検知数が少数，または 0 のルールが多かったが,「比べる」,「比較」などのルールについては精度が高いことがわかった．

8.2 今後の課題

本論文では，提案した有用性の 7 つの観点のうち 3 つについての有用性評価は実装したが，残りの 4 つの観点は現在未着手の状態である．したがって，これら 4 つの観点，すなわち観点 4(実際に商品を使用したと推測できる)，観点 5(評価に対する根拠がある)，観点 6(分量が多い)，観点 7(読みやすい文章である) からの有用性評価を実現することが今後の課題としてあげられる．

また，本研究で有用性評価を実装した観点について，観点 1 の評価では F 値が 0.6 前後，観点 2 の評価では正解率がベースラインをわずかに上回るか同じ程度であり，観点 3 の評価ではいくつか精度が高いルールがあるが全体としては精度が低く，またテストデータの正例が不足していることから，比較を含むレビューの検出の精度を正確に測ることができなかったルールも散見された．したがって，本論文の提案手法をさらに洗練させることも必要である．

以上に加えて，最終的に 7 つの観点の評価結果をユーザに提示して，ユーザのレビュー検索を補助するシステムをどのように実装するかも今後の重要な課題である．

関連図書

[1] 統計局ホームページ/統計 today　no.141 急拡大するネットショッピングと電子マネーの利用 家計消費状況調査 2018 年の結果から. https://www.stat.go.jp/info/today/141.html. (Accessed on 01/14/2021).

[2] 新井智也, 佐藤哲司. 評価視点別の言及度を用いた意見文の分類手法の提案. In *DEIM Forum, A2*, 第 2 巻, 2010.

[3] NTT コミュニケーションズ. COTOHA API. https://api.ce-cotoha.com/contents/index.html. (Accessed on 01/18/2021).

[4] Miao Fan, Chao Feng, Lin Guo, Mingming Sun, and Ping Li. Product-aware helpfulness prediction of online reviews. In *The World Wide Web Conference*, pp. 2715–2721, 2019.

[5] Hong Hong, Di Xu, G Alan Wang, and Weiguo Fan. Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems*, Vol. 102, pp. 1–11, 2017.

[6] Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 244–251, 2006.

[7] 木浪貴博, 小林亜樹. 商品レビュー文における主観的表現と有用性に関する検討. 第 77 回全国大会講演論文集, Vol. 2015, No. 1, pp. 159–160, 2015.

[8] 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.

[9] 近藤陽介, 松吉俊, 佐藤理史. 教科書コーパスを用いた日本語テキストの難易度推定. 言語処理学会第 14 回年次大会発表論文集, pp. 1113–1116, 2008.

[10] 工藤拓, 松本裕治, 立石健二, 福島俊一. チャンキングの段階適用による日本語係り受け解析. 自然言語処理学会論誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[11] Rakuten Institute of Technology. 楽天データ公開. https://rit.rakuten.co.jp/data_release_ja/. (Accessed on 01/19/2021).

[12] 李在鎬. 日本語教育のための文章難易度に関する研究. 早稲田日本語教育学, Vol. 21, pp. 1–16, 2016.

[13] J Rodak, M Xiao, and L Longoria. Predicting helpfulness ratings of amazon product reviews. technical project report, 2012.

[14] 佐々木優衣, 関洋平. 商品レビューを対象とした有用性の定義と判別. 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), 2014 年 3 月.

[15] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

[16] Kasturi Dewi Varathan, Anastasia Giachanou, and Fabio Crestani. Comparative opinion mining: a review. *Journal of the Association for Information Science and Technology*, Vol. 68, No. 4, pp. 811–829, 2017.

[17] Chau Vo, Dung Duong, Duy Nguyen, and Tru Cao. From helpfulness prediction to helpful review retrieval for online product reviews. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, pp. 38–45, 2018.

[18] 山下達雄, 東野進一. 商品レビューに含まれるストア言及の抽出. 第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 7–8, 2016.

[19] 山澤美由起, 吉村宏樹, 増市博. Amazon レビュー文の有用性判別実験. 情報処理学会研究報告自然言語処理 (NL), Vol. 2006, No. 53 (2006-NL-173), pp. 15–20, 2006.

[20] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 38–44, 2015.