

楽天市場の商品レビューを使用した 評価軸と評価表現辞書の同時構築手法

金兵 裕太[†] 沼尾 雅之[‡]

[†] 電気通信大学情報・通信工学科沼尾研究室 〒182-0021 東京都調布市調布ヶ丘 1-5-1

[‡] 電気通信大学大学院情報工学専攻 〒182-0021 東京都調布市調布ヶ丘 1-5-1

E-mail: [†] yk1111046@gmail.com, [‡] numao@cs.uec.ac.jp

あらまし CGM(Consumer Generated Media)は、消費者の購買傾向や商品の統計的な評価を知る上で有用な情報であり、これを自動的に分析する研究が盛んになっている。CGM分析は、単語の肯定・否定の情報を登録した評価表現辞書の構築や、複数の単語を意味ごとに分類するカテゴリ分類など多岐にわたるが、これらを組み合わせて活用した研究は少ない。本稿では、ユーザの商品選択支援システムの作成を目的として、楽天市場の「みんなのレビュー・口コミ情報」をコーパスとした、評価軸とそれに対応する評価表現辞書の同時構築手法を提案する。

キーワード テキストマイニング, CGM分析, 評価表現辞書, クラスタリング

1. はじめに

SNSやブログ等の普及により、ライトユーザであっても簡単にネット上に情報を載せることが出来るようになった。これに伴い、ネット上に書いた消費者の声、CGM(Consumer Generated Media)に注目が集まっている。CGMを分析することで、消費者の購買傾向や商品に対する統計的な評価を調べることが出来るため、消費者と企業双方にとって有益である。そのため、CGMを自動的に分析する試みが盛んになっている。

CGM分析の例として、ユーザのレビューが肯定と否定どちらの評価であるかを判定する評判分析と、コーパスを生成する単語を意味的なカテゴリ集合に振り分けるカテゴリ分類がある。評判分析を行う際には、対象を評価する言葉が肯定/否定どちらであるかという情報を登録した評価表現辞書を用いるのが一般的であり、それを用いた文書の分析結果を特定の形式で可視化するというプロセスを踏む。評価表現辞書の構築手法については多くの研究がなされているが、評価表現辞書の活用に関するものは比較的少なく、他のCGM分析と組み合わせた研究もあまりされていない[1]。

本研究において目標とするのは、楽天市場におけるユーザの商品選択支援システムの作成である。大量のレビューがある場合、全てに目を通すことは困難である。また、レビューには「価格」や「機能」など複数の評価軸が内在するが、一般的なショッピングサイトにおいては予め決められた評価軸上での点数でしか現れない。大量のレビューを、「何について」の「どのような」評価であるかで分類することができれば、ユーザは自分の知りたい評価情報を少ない労力で知ることができる。そこで本稿では、楽天市場の商品レビューをコーパスとした、評価軸および評価表現辞書の同時構築手法を提案する。

2. 関連研究

2.1. 評価表現辞書の構築に関する研究

Kampsらが提案した手法では、類義語関係の形容詞同士がリンクで結ばれた語彙ネットワークを利用する。判定したい形容詞が、ネットワーク内で「good」と「bad」どちらに近いかを計算することで、肯定/否定の極性を付与するというものである。しかし、語彙ネットワーク内の言葉しか登録できないため、新語などの未知語には対応できないといった欠点がある[1]。

那須川らは周辺文脈の情報を利用した手法を提案している。文章中に評価表現が存在すると、同じ極性の文脈がその周辺に形成されることが多いという仮定をもとに、種表現からブートストラップ的に評価表現候補を収集する。その際に、種表現の極性と逆接の接続詞の有無に基づいて、評価極性の付与を行う[2]。

2.2. 特徴語の分類に関する研究

特徴語を指定されたカテゴリに分類するカテゴリ分類では、係り受け関係と相互情報に基づいて分類を行う研究が存在する[3]。しかし、カテゴリの種類や初期の集合を人手で設定する必要があるほか、構築したカテゴリ情報を用いた分析システムの作成や、評価表現辞書との紐付けについては言及されていない。

2.3. 複合名詞に関する研究

自然言語処理における形態素解析では、テキストを形態素という小さい粒度で分解するため、「転送速度」等の語句は「転送」「速度」のように分解されてしまう。しかし単に名詞を結合するだけでは誤った複合名詞を生成する可能性もあるため、複合名詞の妥当性を判定するために頻度情報を用いる場合がある[4][5]。

3. 提案手法

3.1. 用語の定義

本研究で扱う用語について整理しておく。

評価表現辞書とは、あるものを評価する語句(**評価表現**)を、その肯定/否定の情報(**評価極性**)とともに登録した語彙集である。また、学習開始時に初期条件として与える既知の評価表現を**種表現**と呼ぶ。評価表現の抽出対象は「形容詞」である。

一方、**評価軸**とは、対象の評価指標を表したものである。ある評価軸に特有な語句を、その評価軸の**特徴語**と呼ぶ。本研究では、評価軸は特徴語を分類することによって実現される。特徴語の抽出対象は「名詞」と「未知語」とした。

なお、評価表現辞書と評価軸は商品ジャンル毎に個別に構築する。**商品ジャンル**とは、商品を種類別にまとめる「掃除機」や「プリンタ」といったグループであり、楽天市場で定義されているものを利用している。

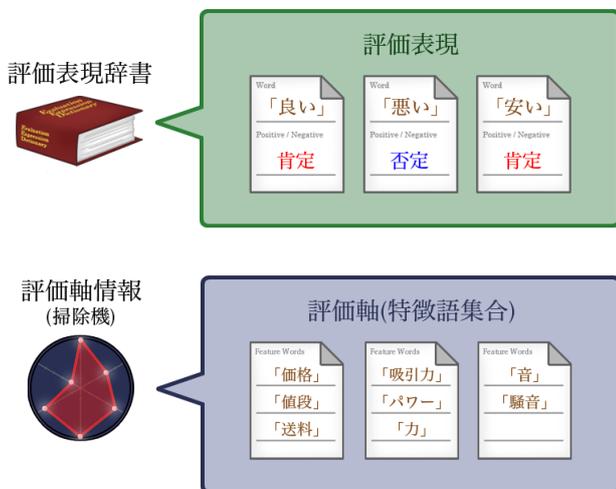


図1 評価表現辞書と評価軸

3.2. 商品選択支援システムの概要

本研究では、ユーザの商品選択支援システムの作成を目的とした、楽天市場での各商品ジャンルに対応した評価軸と評価表現辞書の同時構築を行う。

目標とする商品選択支援システムは、図2のような形式で情報を提供する。



図2 商品選択支援システムの使用例

ユーザが楽天市場のある商品を選択すると、予め学習した評価軸と評価表現辞書のデータベースを用いて、その商品のジャンル(図の例では「プリンター」)に応じた評価軸ごとの評価点を知ることができる。評価軸とは図2における「画質」や「価格」を指し、評価点とはレーダーチャートの値のことである。またデータベースには各評価軸に属する特徴語も記憶されているため、その商品のレビューの中から特定の評価軸に対する評価を探索し、具体的な評価内容を知ることができる。

3.3. 全体構想

システムの実現にあたって、コーパスから肯定/否定情報を持つ評価表現と、評価軸を構成する特徴語を抽出してデータベースに保存する必要がある。図3に、学習から商品選択支援システム利用までの、システム全体の概要を示す。

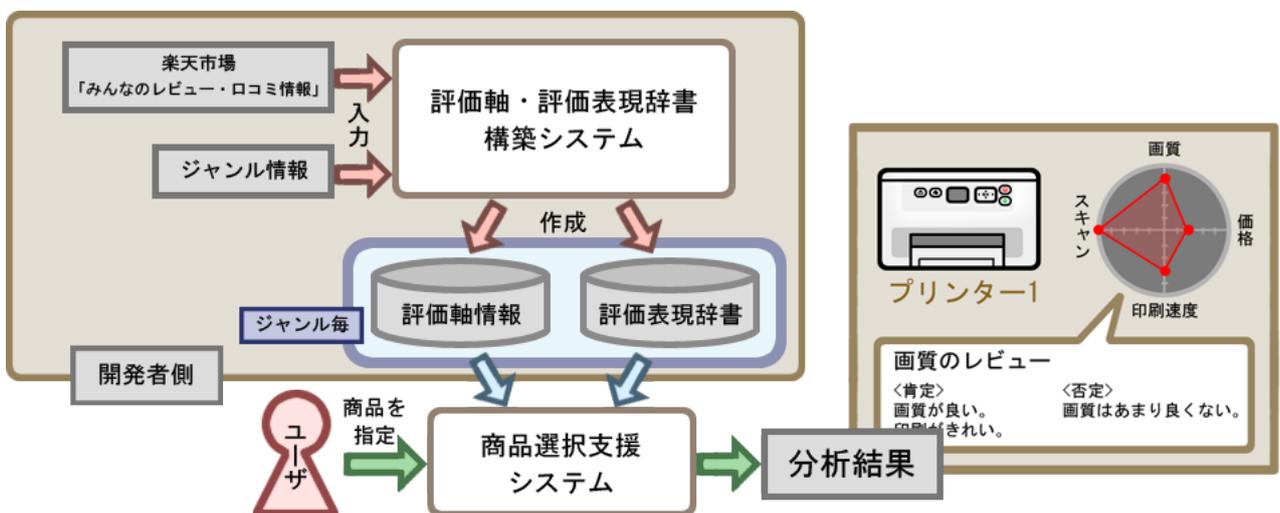


図3 システム全体の概要

本研究では、実際の商品レビュー集合をコーパスとして学習を行う。各商品ジャンルのレビューから、そのジャンルに応じた評価軸と評価表現辞書をそれぞれ構築していく。一度の学習ステップにおいては、当該ジャンルのレビュー集合全体を対象に、既に登録されている評価表現の情報を利用して、新たな特徴語と評価表現の獲得を行う。学習を繰り返し、新たな特徴語と評価表現が得られなくなった時点で学習が終了したとみなす。なお、初期条件として与えておく必要があるのは、少数の種表現だけである。

次に、一度の学習ステップにおける処理について具体的に述べる。始めに、新たな評価表現と特徴語を獲得するための下準備として、全商品レビューに対して以下の処理を行う。

- ・形態素解析
- ・複合名詞の作成
- ・既知の評価表現と特徴語の同定

これらの処理の後、新たな評価表現と特徴語の候補をそれぞれ特定の手法で収集する。手法の詳細な説明は「3.4.1 評価表現候補の抽出」と「3.5.1 特徴語候補の抽出」の節で行う。

全商品レビューから評価表現候補と特徴語候補を抽出した後、頻度情報等を用いたフィルタリングを行い、条件を満たした語句を新たな評価表現と特徴語としてデータベースに保存する。また、この際に特徴語の分類を行い、評価軸の情報も構築する。評価表現と特徴語の登録条件については、それぞれ「3.4.2 評価表現の判定」と「3.5.2 特徴語の判定」で述べる。特徴語の分類に関しては、「3.5.3 特徴語の分類」で説明する。

以上が一度の学習ステップにおける全工程である。図4にそのフローチャートを示す。

3.4. 評価表現辞書の構築

この節では、図4における「評価表現候補の抽出」と「評価表現の判定」について説明する。

3.4.1. 評価表現候補の抽出

評価表現辞書に新たに加える評価表現の候補を、その評価極性ととも抽出する。抽出の手法としては、那須川らの提案した「周辺文脈の情報を利用した手法」[2]をもとに行う。「文書中に評価表現が存在すると、その周囲に評価表現の連続する文脈が形成されることが多く、その中では、明示されない限り、好不評の極性が一致する傾向がある」という仮定に基づき、既に評価表現辞書に登録されている評価表現を中心に、新たな評価表現候補を抽出する。

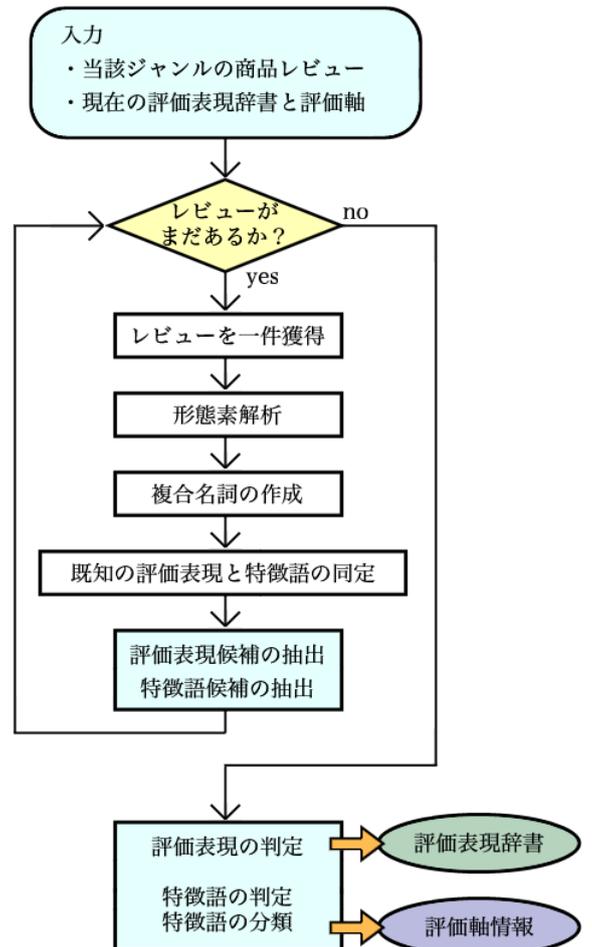


図4 一度の学習におけるフローチャート

具体的には、文章中に評価表現が現れると、接続詞や接続助詞を伴って評価表現と隣接する場合に限り、隣接した表現である以下の3つを、評価表現候補として抽出する。

- ・一つ前の文章の主節
- ・同一文中の、評価表現と並列な用言句
- ・一つ後の文章の主節

また、評価表現候補を抽出する際には、その評価極性と共にカウントする必要がある。評価極性は以下の条件を考慮して決定する。

1. 既知の評価表現の極性が肯定か否定か
2. 既知の評価表現の極性が文中で反転しているか
3. 接続詞(接続詞)が逆接の意味を持つか
4. 評価表現候補の極性が文中で反転しているか

既知の評価表現と評価表現候補が順接関係であれば同じ極性を、逆接関係であれば逆の極性を付与するということである。2と4の「極性が文中で反転している」とは、極性反転子「ない」等が評価表現に続くことで、

「面白くない」のように本来とは逆の極性を示す場合である。極性反転や逆接の接続詞は予め人手で設定しておく。3の接続詞とは、抽出するのが並列用言句である場合は評価表現と候補の間の接続詞であり、一つ前の文章の場合は元の文の先頭の接続詞、一つ後の文章の場合はその文章の先頭の接続詞である。図5に肯定の種表現として「満足」を設定した場合の抽出例を示す。

一つ前の文章：軽くて使いやすいです。

種表現の登場した文章：それに音もうるさくないので、**満足**です。

一つ後の文章：でも値段は少し高かったです。

<p style="text-align: center;"><< 抽出される評価表現候補 >></p> <p style="text-align: center;">肯定表現：使いやすい</p> <p style="text-align: center;">否定表現：うるさい, 高い</p>

図5 評価表現候補の抽出例

3.4.2. 評価表現の判定

抽出された評価表現候補の中から、以下の条件を満たしたものを、新たな評価表現として評価極性付きで辞書に登録する。

1. 出現頻度が一定(評価実験では10)以上
2. 候補として抽出した極性の割合が、肯定/否定のいずれかで一定(評価実験では0.9)以上

3.5. 評価軸の構築

評価軸の構築は、獲得した特徴語を分類することによって実現される。そのため本節では、図4における「特徴語候補の抽出」と「特徴語の判定」、「特徴語の分類」について説明する。

3.5.1. 特徴語候補の抽出

本研究における特徴語とは、特定の評価指標を代表した言葉である。そのため、「この価格は嬉しいです」や「吸引力は良いけど、少し音がうるさいです」等の表現のように、文章中で評価表現と共に出現する頻度が高いと推測できる。そこで、特徴語候補の抽出も既知の評価表現を中心に行うのが妥当である。

具体的には、レビュー中に既知の評価表現が現れると、その評価表現の周囲に共起する特徴語候補があるかどうかを探索する。共起情報の取得には、ある単語とN語以内の距離にある単語が共起関係にあるとする「ウィンドウサイズ」を用いた手法を用いる。評価実験におけるウィンドウサイズは3とした。特徴語候補の抽出対象は名詞と未知語であるが、実際に抽出するのはその内「代名詞」「非自立」「接尾」「特殊」「副詞可能」を除いたものとする。

3.5.2. 特徴語の判定

抽出された特徴語候補のうち、出現頻度が一定(評価実験では100)以上のものを対象とする。また特徴語は主語や目的語として現れる可能性が高く、候補として抽出された特徴語の前後の品詞を考慮するのが有益だという報告がある[4]。そこで、特徴語候補の前後いずれかに助詞が出現する頻度を計算し、その割合が閾値(評価実験では0.5)に満たないものは、特徴語候補から除外する。

最後に、特定の助詞としか共起しない特徴語候補のフィルタリングを行う。形態素解析を行うと、「それなり」や「割」といった語句も名詞となるが、これらは勿論特徴語とするべきではない。特徴語は主語や目的語どちらにもなり得るため、格助詞や係助詞が同程度共起する傾向にあるが、これらの語句は「それなりに」や「割と」といった形で特定の助詞としか共起しないため、その出現頻度が偏る傾向にある。それを踏まえて、共起した助詞の回数を比較して、格助詞と係助詞のうち一方が8割以上を占める場合は特徴語候補から除外し、残った候補を特徴語として確定する。

3.5.3. 特徴語の分類

既知の特徴語を、既知の評価表現を用いて分類する。本研究では特徴語の分類手法として教師なし学習であるK-means法を採用し、要素間の距離を測る尺度としてTF-IDFを利用する。

学習によって得られた特徴語は、似たような意味を持つ特徴語同士が集合することで評価軸となる。正しく評価軸を構築するためには分類を正しく行う必要がある。類似した意味の特徴語同士が近くなるように距離を定義することが重要である。本研究では意味的な類似度を表現するためにTF-IDFを流用する。TF-IDFとは本来「文書集合において、ある単語がそれぞれの文書内でどれだけ重要であるか」を表す尺度であり、しばしば文書の分類などにも利用されるものである。TF-IDFは以下の式で表される。

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{d: d \ni t_i\}|}$$

$n_{i,j}$ は単語 t_i の文書 d_j における出現回数で、 $|D|$ は文書の総数、その分母は単語 t_i が登場する文書数を表している。tfはTerm Frequencyであり、単語の出現頻度を表す。idfはInverse Document Frequencyであり、逆文書頻度と呼ばれる。「今日」等のどんな文書にでも頻繁に登場するような一般語は、tfの値は大きくなるが、idf

の値は小さくなるため、idf は一種のフィルターとして機能する。

本研究ではこの TF-IDF を「特徴語と評価表現の関連度」を表す尺度として用いる。これは、「同じ評価軸に属する特徴語は類似した意味を持っており、それらは類似した評価表現と共起する可能性が高い」という仮定に基づいたアルゴリズムである。TF-IDF は次式で求められる。

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{F}{F_i}$$

分類対象は特徴語であるため、特徴語を上記の文書、評価表現を上記の単語に対応させる。したがって、 $n_{i,j}$ は評価表現 e_i と特徴語 f_j のコーパス中での共起回数、 F は特徴語の総数、 F_i は評価表現 e_i と共起する特徴語の数となる。よって TF-IDF は、「ある特徴語とある評価表現がどれだけ強い相関関係にあるか」の指標となる。これによって各特徴語と各評価表現の距離を定義することができるため、特徴語の分類を行うことができる。なお、実際に分類に用いる値は、TF-IDF を特徴語ごとに正規化したものとする。

特徴語のクラスタリングには K-means 法を用いる。K-means 法は非階層的クラスタリングと呼ばれる分割手法の一つである。予め学習データを与える必要のない教師なし学習であり、評価関数を用いて対象を K 個のクラスタに分割する。K-means 法では、クラスタの重心であるセントロイドをクラスタの代表点として扱う。対象のクラスタへの割り当ては、対象とセントロイドの距離が最も小さくなるようなクラスタを選択することによって行われる。対象のクラスタへの割り当てと、セントロイドの再計算を繰り返し、クラスタに変化が起らなくなった時点で収束したとみなし、終了する。

本研究でのクラスタリングも同様にして行い、特徴語を分類して構成された集合が評価軸となる。分類に用いる特徴ベクトルは前節の TF-IDF であり、次元数は既知の評価表現の数に等しい。初期クラスタの割り当ては一般的な K-means 法と同様、基本的にランダムに行うが、前回の学習ステップで構築された評価軸は繰り返し初期クラスタとして用いる。これには、前回正しく分類された特徴語がランダムな割り当てにより誤った分類をされる可能性を低くし、局所最適解に陥ることを防ぐ目的がある。

しかし、これには「前ステップで構築された評価軸が必ずしも正しいとは限らない」という問題点がある。

K-means 法に用いる距離関数は評価表現の TF-IDF を利用しているため、その次元数は学習が進むにつれて増加する。また、学習初期では次元数が少なく、十分な分類精度を出すことは困難であると考えられる。そこで、今回の学習ステップで信頼できる評価軸情報だけを用いるため、クラスタリング後にフィルタリングの処理を行うこととする。K-means 法は有限個のクラスタのうち最適なものを適宜選択していく手法であるため、分類の途中で局所最適解に陥ってしまうことが多々ある。しかし、そうして出来たクラスタが完全に誤ったものであるかという点必ずしもそうではなく、ほぼ完成しているクラスタに少数のノイズが加わっている場合や、複数個のクラスタが結合した状態になってしまっている場合などが考えられる。そのようなクラスタを適切な状態にするために、「セントロイドから離れた特徴語の除去」を行う。セントロイドから除外するための閾値を設けて、以下の手順で処理を行う。

1. セントロイドを計算する
2. 各特徴語からセントロイドの距離を計算する
3. 最長距離が閾値以上ならその特徴語を除外する
4. 特徴語を除外されなくなるまで 1~3 を繰り返す

以上の処理を全てのクラスタで行い、最終的にクラスタの要素が 2 つ以上のものを評価軸として保持し、次の学習ステップで初期値として利用する。

4. 評価実験

4.1. 楽天データ

学習用のコーパスには、楽天の公開データセットである「みんなのレビュー・口コミ情報」を使用する。これには楽天市場の商品レビューの情報が収められており、2010 年から 2012 年までの 3 年間分のレビューが公開されている。本研究ではこのデータセットから、「レビュータイトル」と「レビュー内容」のカラムを合わせて、一つのレビューとして扱う。

評価実験で扱うジャンルは以下の 2 ジャンルとする。これらのジャンルは楽天市場で定義されているものを利用しており、これをもとに商品进行分类してそれぞれのコーパスを生成している。なお大量の商品レビューの中には同一ユーザによる同一内容のものもしばしば見られるため、同一内容のレビューは 1 つしか登録しないようにした。

このようにして、それぞれのジャンルで 10 万件を超える商品レビューを収集した。

表 1 ジャンルごとのコーパス情報

番号	ジャンル名	レビュー数
1	外付けドライブ・ストレージ	104596
2	掃除機	123511

4.2. 評価軸と評価表現辞書の同時構築実験

前述の2つのジャンルにおいて、評価軸と評価表現辞書の構築実験を行った。各ジャンルで抽出された評価表現と、各ジャンルで構築された評価軸の情報を以下に示していく。

なお、初期条件として与える種表現は、肯定表現は「良い」「満足」、否定表現が「悪い」「不満」とした。

4.2.1. 評価表現辞書の構築結果と考察

各ジャンルにおける評価表現辞書の構築結果を以下に示す。

表2 「外付けドライブ・ストレージ」ジャンルの評価表現抽出結果

種表現	
肯定表現	良い, 満足
否定表現	悪い, 不満
1回目の学習結果	
肯定表現	速い, 小さい, 多い, 安い, 欲しい, 早い, 可愛い, かわいい, よい
否定表現	安っぽい
最終的な学習結果(4回目)	
肯定表現	速い, 小さい, 多い, 安い, 欲しい, 早い, 可愛い, かわいい, よい, ちいさい, かつこよい, ちっちゃい, 薄い, 軽い, ほしい, うれしい, やすい, はやい
否定表現	安っぽい

表3 「掃除機」ジャンルの評価表現抽出結果

種表現	
肯定表現	良い, 満足
否定表現	悪い, 不満
1回目の学習結果	
肯定表現	長い, 安い, 欲しい, 早い, 強い, 可愛い, かわいい, 軽い
否定表現	うるさい, 重い
最終的な学習結果(4回目)	
肯定表現	長い, 安い, 欲しい, 早い, 強い, 可愛い, かわいい, 軽い, 白い, 賢い, 丸い, 易い, 小さい, 薄い, 素晴らしい
否定表現	うるさい, 重い, 古い, でかい, 大きい, ごつい

抽出された語句の評価極性を、人手で判断した極性と比較した際の適合率は以下ようになった。

表4 評価表現辞書の適合率

ジャンル	適合率
外付けドライブ・ストレージ	94.7%
掃除機	85.7%

いずれのジャンルも適合率が85%を超えたほか、明らかに逆の極性で登録された語句もないため、評価表現の抽出が高い精度で行えていることが確認できた。

4.2.2. 評価軸の構築手法の検証

次に、評価軸の構築結果を示す前に、評価軸の構築において「前回の学習ステップで構築した評価軸情報を繰り返し初期値として用いる」ことに優位性があるかを検証する。

評価軸情報を初期値として利用する場合と利用しない場合とでそれぞれ評価軸の構築を行い、完成したクラスタを分析する。K-means クラスタリングに用いるKの値は8とした。クラスタの分析には参考文献[6]による2通りの指標を用いる。

1つ目はクラスタ内距離二乗和と呼ばれる、クラスタ内の凝集性を測る尺度であり、次の式で求められる。

$$P_k = \sum_{i=1}^k \sum_{x \in C_i} (d(x, c_i))^2$$

ここで、 k はクラスタ数、 C_i は*i*番目のクラスタ、 $d(X, Y)$ は*X*と*Y*の距離、 x はクラスタの要素、 c_i は*i*番目のクラスタのセントロイドである。

2つ目の指標はPseudo Fという尺度であり、次式で求められる。

$$PseudoF = \frac{(T - P_k)/(k - 1)}{P_k/(n - k)}$$

T は全データの距離二乗和(全データの平均と各データの距離の平方和)である。前述の指標がクラスタ内の凝集性のみを見ているのに対し、Pseudo Fでは複数のクラスタ間の分散性も考慮している。クラスタ同士は疎、クラスタ内は密であることが望ましいとし、値が大きいほどクラスタリングとして良い結果だと言える。

それぞれのジャンルにおいて、前回の学習ステップで構築された評価軸情報を繰り返し初期値として「利用する」及び「利用しない」場合の各指標の値は以下ようになった。

表5 前学習ステップの評価軸情報の利用の優位性

ジャンル番号	利用	P_k	$\frac{P_k}{[\text{特徴語数}]}$	PseudoF
1	する	3.526	0.1306	29.67
	しない	3.261	0.1254	27.70
2	する	3.680	0.1187	45.14
	しない	2.960	0.1287	22.37

結果より、クラスタ内距離二乗和 P_k は、評価軸内の特徴語数で平均すると、「利用する」場合と「利用しない」場合との誤差はそれぞれおよそ5%と8%にとどまった。一方で、Pseudo Fはいずれも「利用する」場合に高くなり、特に掃除機においては2倍以上の値を出している。評価軸を構築する上で他のクラスタとの差別化が重要であることを踏まえて、以降の実験では「前回の学習ステップで構築された評価軸情報を繰り返し初期値として用いる」手法を採用する。

4.2.3. 評価軸の構築結果と構築

次に、「外付けドライブ・ストレージ」と「掃除機」のジャンルにおける評価軸の構築結果を以下に示す。なお、各評価軸における特徴語の順番はセントロイドとの距離の昇順であり、すなわち評価軸を代表する順番となっている。

表 6 「外付けドライブ・ストレージ」ジャンルの評価軸構築結果

評価軸 1	価格, 値段, お値段, 送料, 金額
評価軸 2	発送, 到着, 対応
評価軸 3	色, 形, デザイン
評価軸 4	持ち運び, 場所
評価軸 5	スピード, 速度, 書き込み速度, 読み込み速度, 読み書き
評価軸 6	使い勝手, 評価
評価軸 7	USB メモリ, メモリ
評価軸 8	本体, キャップ

表 7 「掃除機」ジャンルの評価軸構築結果

評価軸 1	値段, 価格, お値段, 送料
評価軸 2	配送, 発送, 到着, 対応
評価軸 3	見た目, デザイン, 動き, 色, 姿
評価軸 4	小回り, ヘッド, 階段, 重量, 持ち運び
評価軸 5	機能, 性能
評価軸 6	母, 自分, 妻, クリーナー
評価軸 7	ごみ, ゴミ, 収納, 印象
評価軸 8	毛, 髪の毛, コード

いずれの構築結果にも、「価格」「値段」等の「金額」に関する評価軸、「デザイン」「見た目」「形」といった「外見」に関する評価軸、「配送」「発送」「対応」等の「サービス」に関する評価軸が共通して現れた。これらの評価軸は、おおよそ全ての商品に対して存在するものであり、異なるジャンルのコーパスからこのような評価軸を構築できていることから、この手法の汎用性は高いと考えられる。

一方で、それぞれのジャンルに特有な評価軸も構築することができた。外付けドライブ・ストレージでは「書き込み・読み込み速度」に関する評価軸等が現れ、掃除機のジャンルでは「機能・性能」の評価軸のほか、「母」「自分」「妻」といった人のクラスタが生成された。これは「使用者」に関する評価軸だと考えられる。

以上の結果から、評価軸の構築が正しく行えているであろうことが直観的には言える。次節においては、構築した評価軸情報を定量的に評価するために新たなデータセットを用意し、その分析結果について述べる。

4.2.4. 評価軸の定量的な評価

楽天市場の商品レビューをコーパスとした評価軸情報は、正答データと言えるものがないため、定量的な評価基準を設けることは困難である。

定量的な評価を行うために、楽天が公開している「楽天トラベル：施設、お客様の声情報」をコーパスとした実験を行った。このデータセットにはホテルに対するユーザのレビュー情報が収められている。また、楽天トラベルでは「部屋」「食事」「風呂」「サービス」「設備・アメニティ」「立地」「料金」という7つのカテゴリが評価項目としてサイトに設定されているため、構築した評価軸情報と設定されているカテゴリ情報を比較することで、定量的な評価をすることができる。

正答データは参考文献[3]を元に作成する。この論文では同データセットを対象に、単語のカテゴリへの割り当てを行っており、その際にカテゴリ毎の種語集合と正当データを人手で作成している。

以下に構築された評価軸情報と、評価軸情報と正答データとの適合率を示す。前節と条件を揃えるため、コーパスのレビュー数は100000件とした。

表 8 「楽天トラベル」での評価軸構築結果

評価軸 1	部屋, お部屋, ロビー, 浴室, ユニットバス, トイレ, バスルーム, 室内, 客室
評価軸 2	朝食, 食事, 料理, バイキング, お料理, お店, パン, 夕食, ご飯, ボリューム, 店, レストラン
評価軸 3	ベッド, 浴槽, バスタブ, バス, 風呂, お部屋, 窓
評価軸 4	温泉, 対応, お湯, 露天風呂, 大浴場, 対応, 接客
評価軸 5	内装, 施設, 設備, 建物, 掃除
評価軸 6	立地, 立地条件, ロケーション, 場所, アクセス
評価軸 7	値段, 料金, お値段, 価格, 宿泊料金, 金額
評価軸 8	便, 愛想, 天気, 使い勝手

表 9 評価軸と評価項目の適合率(%)

	部屋	食事	風呂	サービス	設備	立地	料金	その他
評価軸 1	55.6	0	33.3	0	11.1	0	0	0
評価軸 2	0	84.6	0	0	0	0	0	15.4
評価軸 3	28.6	0	71.4	0	0	0	0	0
評価軸 4	0	0	57.1	42.9	0	0	0	0
評価軸 5	0	0	0	0	80.0	0	0	20.0
評価軸 6	0	0	0	0	0	100.0	0	0
評価軸 7	0	0	0	0	0	0	100.0	0
評価軸 8	0	0	0	25.0	0	0	0	75.0

表 9 より、半数以上の評価項目で適合率が80%を超えた。この結果より、特徴語の分類に評価表現との共起情報を用いることが有用であると言える。

5. おわりに

5.1. まとめ

本稿では、評価軸と評価表現辞書の同時構築手法を提案し、評価実験においてはその構築結果を検討し、提案手法の有用性を示した。

関連研究では、教師データや種語集合など、特徴語を抽出するための初期条件を人出で用意しなければいけない場合が多くあった[3][4][5]。しかしそれを一々行うのは手間であり、その初期条件によって得られる結果が異なれば、使い手にとっては不便であると考えられる。一方で、本研究で初期条件として与える必要があるのは少数の種表現だけであり、またその種表現は「良い」や「悪い」といった、肯定と否定を代表するような言葉であるため、ジャンルに対応させて新たに設定する必要がない。そのためショッピングサイトの商品レビューというコーパスの枠にとらわれずに、様々な形式のコーパスに対して汎用的に利用が可能であると考えられる。

5.2. 今後の課題

評価表現辞書の構築において、評価表現を誤って分類することはなかったが、肯定表現に比べて否定表現はあまり得られなかった。コーパスにおける否定表現が肯定表現と比較して非常に少ないことが主な原因だと考えられるが、既存手法における網羅性の低さも大きく影響していると考えられる[1]。実際にはコーパスでは「遅い」や「弱い」といった語句も現れているが、否定表現として抽出される割合が十分でなく、登録するまでには至っていない。これは逆接と単純接続の両方に使われる語が存在し、極性の判定が必ずしも正確ではないためである。本来そのようなケースはノイズとして無視される[2]が、コーパス自体への出現頻度が少ないと上手くいかないことがある。よって今後は、既存手法の精度を落とさずに網羅性を上げる方法について検討する必要がある。

評価軸の構築では前述の通り、評価表現との共起情報を用いて特徴語をクラスタリングすることができた。しかし、構築した評価軸が必ずしも商品の評価を左右するとは限らない。現状では、掃除機のジャンルで出現回数は多いが単語数が少ない「音」に関する評価軸がノイズとして扱われてしまっているという問題もある。どのような語句が商品レビューにおいて重要な意味を持つかに着目することで、より有益な評価軸が構築できる可能性があると考えられる。

今後は、上記の課題をもとに評価軸と評価表現辞書の構築精度を向上すると共に、商品選択支援システムで必要となる、レビューの分類や評価軸ごとの評判分析等の機能の実装を行っていく。

参 考 文 献

- [1] 乾孝司, 奥村学, “テキストを対象とした評価情報の分析に関する研究動向”, 自然言語処理 Vol. 13, Num. 3, pp. 201-241, 2006.
- [2] 那須川哲哉, 金山博, “文脈一貫性を利用した極性付評価表現の語彙獲得”, 情報処理学会自然言語処理研究会(NL-162-16), pp. 109-116, 2004.
- [3] グェン ファム タン タオ, 岡部誠, 尾内理紀夫, 林貴宏, 西岡悠平, 竹中孝真, 森正弥, “新たな弱教師付き型分類手法 Bautext”, 情報処理学会論文誌 Vol. 52 No. 1, pp. 269-283, 2011.
- [4] 杉浦広和, “議事録集合からの特徴語抽出とその応用に関する研究”, <http://www.nagao.nuie.nagoya-u.ac.jp/paper/11253.html>, 2009.
- [5] 峠泰成, 山本和英, “意見情報獲得のためのクエリー関連のドメイン特徴語抽出”, 言語処理学会第12回年次大会, pp. 85-88, 2006.
- [6] , クラスタリング結果の評価の尺度基準, <http://soonraah.hatenablog.com/entry/2014/05/06/192258>, 2014