

ソーシャルメディアからの商品の利用目的の発見

村本 直樹[†] 湯本 高行^{††} 大島 裕明[†]

[†] 兵庫県立大学 大学院応用情報科学研究科 〒650-0047 兵庫県神戸市中央区港島南町 7-1-28

^{††} 兵庫県立大学 大学院工学研究科 〒671-2280 兵庫県姫路市書写 2167

E-mail: [†]{aa18c508,ohshima}@ai.u-hyogo.ac.jp, ^{††}yumoto@eng.u-hyogo.ac.jp

あらまし EC サイト上での商品の購入時、ユーザは様々な「利用目的」で商品の購入を検討する。ここで、ある商品に対して利用目的は複数存在しており、ある商品に対する利用目的を網羅することはユーザの商品購入支援に役立つと考えられる。本研究ではある商品に対する利用目的を、「～用」や「～のために」といった構文パターンを考慮したテキストマイニングにより、網羅的に発見する手法を提案する。コーパスとしては、商品に対する商品レビューとともに、「ランニング用のシューズについて質問です」といった利用目的に関する質問が存在する Q&A サイトを用い、それらの性質の違いについての分析を行う。

キーワード テキストマイニング, 知識発見, 利用目的

1. はじめに

インターネットの普及に伴い、消費者の商品購買行動は変化している。特に、EC サイトでの商品購入の機会は増加しているといえる。ここで、商品を購入する際に多くのユーザは、何らかの利用目的を持っていると考えられる。たとえば、シューズを買う場合、「マラソンで走りたい」、「週末に軽く走りたい」、「普段履くのに使いたい」など様々な利用目的が考えられる。ここで、シューズであれば「ランニングシューズ」や「ウォーキングシューズ」、「カジュアルシューズ」といった目的別のシューズが存在しており、ユーザが求めるものを見つけやすいと考えられる。しかし、単純に「ランニング」といった場合でも、週末に軽く走るだけなのか、マラソンで 3 時間切りを目指すのかによって適したシューズは異なる。それらをシューズに詳しくないユーザが、EC サイト上で購入する際に把握することは難しいと考えられる。また、EC サイト上ではそれぞれのシューズについて、サイズや機能が書かれているが、機能については、メーカー独自の名前が使用されていることが多く、その機能が自分の目的にあっているかを判断することも難しい。

カメラのような商品の場合においても、体育会で子供の様子を撮りたいのか、夜景を撮りたいのかなど、様々な目的が存在することが考えられる。カメラにはスペック表が存在しており、性能が数値化されて表されているが、自分の目的に適したカメラにはどの性能が大切なのかを理解することは容易ではない。

商品の利用者が、商品の製造者とはすこし異なる目的でその商品を利用していることもある。そのため商品の利用目的は、商品の利用者、商品の製造者の両者とも把握しているわけではない。たとえば、Q&A サイトを見ると、「バスケットボール」をするための「バスケットシューズ」が、靴のクッションが良い点や、足首のサポートが強いことより「ウェイトトレーニング」に おすすめであると書かれている。他にも、本来「登山用」のメーカーが出すウェアなどが、機能性とデザイン性の面から、若い世代を中心に普段着として着用されている。上記の例から

も、図 1 のように、1 つの商品に対して利用目的は複数あると考えられる。

このような中、既存の EC サイトの商品カテゴリ分けでは、利用目的別には分類されておらず、利用目的を第一に考えて商品を検索することはできない。そのため、その商品がどのような目的で使用されるのかわからないことが多い。

そこで、本研究では Q&A サイトや商品レビューといったテキストコーパスから、テキストマイニングにより、商品に対するユーザの利用目的を発見する手法を提案する。商品レビューだけでなく Q&A サイトも利用するのは、「ランニング用のシューズについて質問です」といった利用目的に関する質問や、質問に対する回答が存在しているからである。ここで提案する手法は、クエリに対する同位語や関連語を、両方向の構文パターンにより発見した大島ら [1] の手法を応用し、「ランニング用」の「用」や、「ランニングのため」の「のため」といった、構文パターンを用いて、利用目的を発見するものである。

実際に発見されたある商品に対する「利用目的」について、正しいものであるかの評価実験を行った。

2. 関連研究

2.1 本手法の基本となる研究

本研究で用いる手法は大島らの研究 [1] を応用したものである

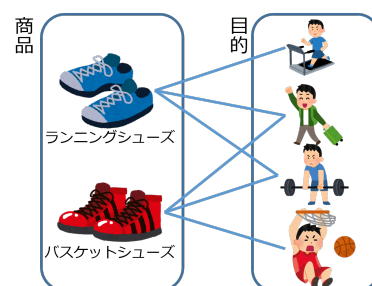


図 1 商品と目的の関係

る。大島らは Web 検索エンジンから同位語を始めとするクエリとして与えられた語に対して様々な関係性にある語を発見する手法を提案している。まずはじめに同位語を発見するために並列助詞「や」に注目した。クエリとクエリに対する同位語が「や」で接続されることから、「や」が前後に出現する 2 つの構文パターンを利用している。たとえばクエリ“ランニング”の同位語を発見する場合、「“ランニングや”」, 「“やランニング”」の 2 つの検索クエリで Web 検索を行い、検索結果から同位語を発見している。この手法を拡張し、関連語を発見する手法についても述べている。本手法ではこのような両方向の構文パターンを用い、「利用目的」の抽出にあった手法に応用する。

2.2 商品推薦や目的に関する研究

本研究では、ソーシャルメディア上のテキストデータから、ある商品に対するユーザの利用目的を収集することを目的としている。ユーザの目的という観点について研究されているものに、高見らの研究 [2] が挙げられる。高見らは Web 検索結果において、検索クエリからユーザの検索目的にあったスニペットについて分類をすることに取り組んだ。また、徳田ら [3] は、観光地検索において、明確な目的がないユーザに対し、発見思考型のナビゲーションシステムを提案している。

利用目的に注目し、商品の推薦システムを構築する研究も行われている。本田ら [4] は、EC サイトにおける商品カテゴリに注目して、カテゴリ特徴ベクトル、商品特徴ベクトル、レビュー特徴ベクトルを作成し、類似度を算出することで、ユーザの使用目的にあった代替商品の推薦システムを提案した。McAuley ら [5] は、ある商品を購入する際に推薦すべき商品について、その商品について代替利用できる代替品と、その商品と同時に買うことが推薦される補完品の 2 種類に注目し、代替品について推薦システムを提案している。

利用目的とはその商品を使った人の行動と言えるが、テキストコーパスから人々の行動を抽出する研究も行われている [6] [7] [8]。たとえば、馬縹ら [8] は、「薬剤師が薬を調合する」のような、ある職業における行動を、対象の職業が主語になっている主語ベースの文と、対象の職業に従事するユーザによって書かれた著者ベースの文から抽出している。

本研究では Q&A サイトのテキストコーパスを利用しているが、Q&A サイトの分析を行う研究も行われている [9]。

また、ユーザの意思決定に注目し、商品の推薦システムを構築した研究も存在する。小池ら [10] は、実際に買い物をする際の女性と男性の差に注目し、キーワードを可視化した対話型推薦アルゴリズムを提案した。益田ら [11] は、実際の家電量販店の販売員の商品販売過程を、「販売員思考」、「対象ユーザ」、「対象商品」、「システムの特徴とコンセプト」の 4 段階で構成しモデル化することで商品推薦システムを考案している。

3. 構文パターンを用いた目的抽出

3.1 取得する利用目的とコーパスについて

本研究では、ソーシャルメディア上のテキストデータから、ある商品に対するユーザの「利用目的」を発見する。たとえば、「サッカー用のシューズを買いたいのですが」のような文章の

場合、「シューズ」に対して利用目的は「サッカー」をすることになる。また、「サッカーを見に行くときのシューズは何がよいですか」のような場合は、「サッカーを見に行く」ことが目的となる。

こういった商品に関するテキストマイニング手法においては、多くの場合で商品レビューが利用されるが、本研究では商品レビューだけでなく、Q&A サイトのテキストコーパスも用いて利用目的の発見を行う。商品レビューでは「納得の行く性能でした」のような、対象となる商品が存在し、その商品に対する購入者の感想や意見が述べられる。一方、Q&A サイトでは何らかの利用目的が先に存在し、その利用目的にあった商品や、商品の特性について質問することがある。たとえば今回利用する Yahoo!知恵袋では、

「一眼レフカメラで体育館でのミニバスの撮影と運動会や子供のサッカーの撮影を考えています。そこで、このくらいはそろえたほうがよいという組み合わせを教えてください。できましたら、最低限はこのくらい。とお金に余裕があればこのくらいまでという感じでアドバイスを頂ければありがたいです。」

のような利用目的にあった商品を知るための質問や、質問に対する回答が存在している。よって、商品レビューだけでなく Q&A サイトのテキストコーパスからも利用目的の発見を行う。

3.2 目的表現発見のための文書の前処理

本手法では取得する目的表現の候補となる形態素列は一文内に存在する形態素の列からなるものとする。そのためにまず与えられた文書を以下の手順で前処理を行う。

- (1) 「.」, 「!」, 「?」以外の記号の削除
- (2) 英数文字の半角化
- (3) 「.」, 「!」, 「?」で文単位に分解する。
- (4) 形態素単位に分かち書き

このような処理を行い、半角スペースで形態素単位に分かち書きされた文に対し、正規表現で表される構文パターンについて、それぞれ一致するかの判断を行う。なお、形態素解析器には MeCab^(注1) を用いており、MeCab の辞書については新語に対応している mecab-ipadic-NEologd^(注2) (2019 年 1 月 31 日更新分) を追加している。

3.3 両方向の構文パターンについて

本節では 2.1 節で述べた両方向の構文パターンについて、本手法ではどのようなパターンを用い、どのように形態素列を取得するかについて述べる。ここで、本研究ではある商品に対する利用目的について書かれている形態素列を「目的表現」とする。Q&A サイトや商品レビュー上のテキストにおいて、商品の「利用目的」についての記述に注目し、目的表現の前後に出現する構文パターンを見つける。取得したい目的表現の前に出現する形態素列からなる構文パターンを「前方パターン」、同様に、目的表現の後に出現する形態素列からなる構文パターン

(注1) : <http://taku910.github.io/mecab/>

(注2) : <https://github.com/neologd/mecab-ipadic-neologd>

表 1 主な後方パターン

構文パターン	例
用	ランニング用に
のため	ランニングのため
時に	ランニング時に
[でに] 使 [あ-ん用] ※正規表現	ランニングに使う

を「後方パターン」とする。

- ・ <前方パターン> 目的表現
- ・ 目的表現<後方パターン>

後方パターンについて注目したのが、「ランニング用」や「トレーニング用」といった、目的を表す表現に「用」を接続した形である。同様に前方パターンにも注目して前方パターンのリストを作成し、目的表現を抽出する。

目的表現を抽出するには、両方向の構文パターンよりそれぞれのパターンについて、目的表現の候補となる形態素列を取得する必要がある。詳細な手法については次節で述べ、本節では、基本的な構文パターンによる取得方法を述べる。

まず、後方パターンを「用」としたときの例を示す。たとえば、「ランニングマシンも使っているのでランニング用です」という文の場合、分かち書きが行われているので「ランニング_マシン_も_使_っ_て_い_る_の_で_ランニング_用_です」となる。ここで、後方パターン「用」が出現するため、形態素列「用」の前に出現する形態素を設定した個数分取得する。ここで、取得する形態素の数について、パラメータ n として設定する。今回仮に、 $n = 5$ とすると、「使_っ_て_い_る_の_で_ランニング」が取得対象となる。

次に「ランニングマシンも使っているのでランニング専用です」という文の場合について述べる。まず、分かち書きを行うと「ランニング_マシン_も_使_っ_て_い_る_の_で_ランニング_専用_です」とようになる。ここでは、形態素列として「用」は出現するが、「専用」の「用」に一致しており、形態素列「用」が存在しているわけではない。この場合は一致していないと判断し形態素列は取得しない。

前方パターンについても同様な処理を、前方パターンの後ろ側に出現する形態素列に対して行う。

次に具体的な構文パターンについて述べる。まず、後方パターンについて最も基本となるパターンとして、上述の通り「ランニング用」のようなパターンが最もわかりやすい例である。他にも「ランニングのため」や「ランニング時に」、「ランニングにします」のようなパターンが出現しやすいと考えられる。このような代表的な後方パターンについて、表 1 にまとめる。なお構文パターンは正規表現で表現されている。

次に、前方パターンについて見てみると、「～はランニング」や、「～の体育会」のような助詞が多く見られた。代表的な前方パターンについて表 2 にまとめる。これらの両方向の構文パターンを利用して、目的表現の候補となる形態素列を取得する。

3.4 各条件を用いた目的表現候補の抽出方法

本手法では、以下の 3 種類の方法で目的表現を抽出する。

- ・ 正規表現のみを用いた単純な抽出手法

表 2 主な前方パターン

構文パターン	例
は	目的はランニング
の	来年からの通学
に	本格的にランニング

- ・ 両方向構文パターンによる挟み込み抽出
- ・ 独立した両方向の構文パターンによる抽出

それぞれの詳細について次項以降で述べる。

3.4.1 正規表現のみを用いた単純な抽出手法

ベースとなる手法として、正規表現のみを用いた単純な抽出手法について述べる。商品に対する利用目的を表すもっとも単純な表現として、「ランニング用」、「旅行用」といった「○○用」がある。この形態素「用」に注目し、漢字かカタカナの連続から「用」につながる形態素列を目的候補表現として取得する。分かち書きされたテキストに対し、正規表現を用いて検索を行う。

3.4.2 両方向構文パターンによる挟み込み抽出

本項では両方向の構文パターンの挟み込みにより、目的表現を抽出する手法について述べる。まず、前方パターンにより目的表現の候補となる形態素列を取得する。そして取得された形態素列に対し、後方パターンを用いて目的表現の抽出を行う。ここで、前方パターンとして「は」、「の」、後方パターンとして「用」を用い、取得する形態素数を $n = 5$ とした場合に、2 つの文書に対して実際にどのような流れで目的表現を取得するか示す。まず、「ランニングマシンに使うシューズはランニング用のものを準備したほうがよいです」という文書に対し実際に行う手順を以下に示す。

- (1) 前方パターン「は」に一致するため、「ランニング用のものを」を取得
- (2) 前方パターン「の」に一致するため、「ものを準備した」、「を準備したほう」を取得
- (3) 取得した 3 つの形態素列に対し、後方パターン「用」が一致するため、「ランニング」を目的表現として抽出

次に、「娘の体育会用に購入しました」という文書に対し、実際に行われる手順を示す。

- (1) 前側の構文パターンの「の」に一致するので「体育会用に購入し」を取得
- (2) 取得した形態素列に対し、後方パターン「用」が一致するため、「体育会」を目的表現として抽出

以上の例について、最初の前方向パターンによる形態素列の取得を図 2 に示す。次に前方パターンにより取得された形態素列に対し、後方パターンを用いて目的表現の取得を行う様子を 図 3 に示す。これらの手法で抽出された目的表現の出現回数をスコアとし、上位 r 件の目的表現をその商品に対する利用目的とする。

3.4.3 独立した両方向の構文パターンによる抽出

本項では、独立した両方向の構文パターンにより目的表現を抽出する手法について述べる。両方向の構文パターンそれぞれで目的表現を抽出し、その出現回数の掛け算によるスコアを与

え、上位 r 件の目的表現を、その商品に対する利用目的とする。

ここで、前方パターン「は」、後方パターン「用」を用いて、取得する形態素数を $n = 5$ とした場合に「ランニングマシンに使うシューズはランニング用のものを準備したほうがよいです」に対して、具体的に行われる手順を示す。

まず、前方パターンにより目的表現の候補を取得する。前方パターン「は」が存在しており、このとき、3.4.2 項で述べた、両方向構文パターンによる挟み込み抽出手法では、5 つの形態素からなる形態素列「ランニング用のものを」を取得していたが、今回は、取得する形態素数を 1 つから 5 つまで変化させながら目的表現として抽出する。つまり今回の場合であれば、下記の 5 通りの形態素列を目的表現として抽出する。

- (1) ランニング
- (2) ランニング用
- (3) ランニング用の
- (4) ランニング用のもの
- (5) ランニング用のものの

対象データすべてに対して、前方パターンのみを用いて上記のように目的表現を取得し、同時に前方パターンにおける出現回数を記録する。

続いて、後方パターンにより目的表現を取得する。後方パターン「用」が出現しているので、同様に後方パターンにより目的表現の候補を取得する。このとき、取得される目的表現の候補は下記の 5 通りになる。

- (1) ランニング
- (2) はランニング
- (3) シューズはランニング
- (4) 使うシューズはランニング
- (5) に使うシューズはランニング

このように、対象データすべてに対して後方パターン用いて目的表現を取得し、前方パターンと同様に後方パターンにおける出現回数として記録する。前方パターンと後方パターンそれぞれでこれらの処理を行った後に、出現した目的表現について、

前パターン	取得する文字列
は	ランニング用～
の	ものを準備した
の	を準備したほう
の	体育会用に～

図 2 前方パターンによる形態素列の取得

目的表現	後パターン
ランニング	用
ものを準備した	
を準備したほう	
体育会	用

図 3 後方パターンによる目的表現の取得

表 3 掛け算によるスコア付け例

目的表現	前方パターン	後方パターン	スコア
ランニング	1	1	1
ランニング用	1	0	0
…	1	0	0
はランニング	0	1	0
…	0	1	0

前方パターンにおける出現回数と後方パターンにおける出現回数をかけたものを、その目的表現のスコアとする。上記の例であれば表 3 のようになる。表のようにどちらか一方にしか出現しない場合はスコアが 0 になるため、除去することができる。得られたスコアよりランキングを作成し、上位 r 件の目的表現をその商品に対する利用目的とする。

また、これらの手法では「前方パターンまたは後方パターン」、「取得する形態素数： n 」、「スコアランキングから取得する件数： r 」の 3 つのパラメータが存在しており、これらを調整することである商品に対する利用目的を発見するための最適化を行う。

3.5 抽出された目的表現の後処理

3.4 節で述べた 3 つの手法により、目的表現の候補となる形態素列には、「、ランニング」のような、明らかに「、」が不要な場合も含んでいる。このようなノイズを減らすために以下の手順で後処理を行う。

- (1) 先頭以外に読点を含む場合は削除し先頭に含む場合は読点を除く
- (2) 抽出された利用目的を形態素解析
- (3) ジャンル依存のストップワードを形態素単位で除く
- (4) 非自立と形容動詞語幹を除く名詞、または動詞を含んでいる場合のみ処理を続ける
- (5) 先頭の形態素が「記号」、「助動詞」、「助詞」、「接続詞」、「非自立名詞」でなくなるまで除く
- (6) 末尾の形態素についても 4 と同様し、末尾の形態素が動詞の場合は原形に戻す
- (7) 1 つの形態素の場合に形容詞であれば削除する
- (8) 1 文字のカタカナや英字の場合は削除する
- (9) ひらがなの連続か数字の連続から成り立つ場合は削除する

(10) 共通ストップワードに含まれている場合は削除する
ここで「ジャンル依存のストップワード」とは、ジャンルごとに設定したストップワードである。たとえば、商品「シューズ」の場合、目的表現として「ランニングシューズ」が抽出される。このとき、商品そのものを表す「シューズ」をジャンル依存のストップワードとし、形態素単位で除き、「ランニング」を目的表現とする。続いて「共通ストップワード」は、一人称や、国名、「好き」といった利用目的になりえない動詞などを、すべてのジャンルに共通するストップワードとしている。共通ストップワードに関しては、上記の手順 (9) までにより処理された目的表現に対して一致判定を行う。つまり、目的表現として「ドイツ」が抽出された場合は、国名となるためストップワードに

より削除されるが、「ドイツに旅行に行く」は削除対象とならない。

この後処理を、3.4 節で述べた 3 つの手法において、文単位で目的表現の抽出を行う際に行う。「ランニングに使います」と言う文から、「ランニング」、「ランニングに」が抽出された場合、後処理を行うと、「ランニングに」が「ランニング」となる。このとき、「ランニング」が重複するため、「ランニング」の出現回数は 1 回となる。

3.6 構文パターン候補の取得

本節では 3.3 節、3.4 節で述べた、「前方パターン」、「後方パターン」を半自動的に見つける手順について説明する。

まず初めに、各商品ジャンルにおいてシードとなる目的表現を用意する。この目的表現の前後に出現する形態素列を取得し、包含案系にあるパターンを除いた出現回数を計算することでスコアを与える。得られたスコアから人手で追加するパターンを選択する。このとき、取得する形態素列は、目的表現を取得するときと同様に形態素の列からなるものとし、今回は 5 つの形態素を取得する。このときの形態素列の取得を行う。具体的に「ランニング用のものを準備したほうがよいです」という文に対して、シードとなる目的表現を「ランニング」としたときの後方パターンの候補は以下の 5 通りになる。

- (1) 用
- (2) 用の
- (3) 用のもの
- (4) 用のものを
- (5) 用のものを準備

ここで、仮に後方パターンにおいて「の」というパターンが 100 回、「のため」というパターンが 90 回出現したとする。この場合それぞれのスコアを与える際に、「の」自身の出現回数は 100 回であるが、スコアを 100 とするのではなく、「の」自身を含む形態素列の中でより長い形態素列のうち、最多の出現回数である「のため」の出現回数 90 を引いた 10 が「の」のスコアとなる。「のため」のスコアはこれ以上長い形態素列が出現しないため 90 となる。こうして得られたスコアランキングより、人手でパターンを選択する。手順を以下にまとめる。

- (1) 各商品ジャンルごとにシードとなる目的表現を決める
- (2) 目的表現の前後に出現する形態素列を形態素単位で取得
- (3) スコアを計算する
- (4) 得られたランキングから人手でパターンの追加

4. 評価実験

4.1 データセット

本研究ではソーシャルメディアから商品の利用目的を網羅的に発見する。ソーシャルメディア上のテキストデータをコーパスとして、テキストマイニングを行う。3.1 節で述べたように 2 種類のコーパスを用いる。Q&A サイトとして、「Yahoo!知恵袋」[12] と、商品レビューデータとして、「楽天市場」[13] の 2 つのデータを利用する。

表 4 Yahoo!知恵袋の各データ量

商品	質問データ数	回答データ数	合計
カメラ	41,826	118,740	160,566
シャツ	25,707	80,136	105,843
シューズ	3,951	10,275	14,226
ペン	30,309	86,036	116,345
リュック	2,256	7,779	10,035

表 5 楽天市場 商品レビューの各データ量

商品	レビュー数
カメラ	16,286
シャツ	280,230
シューズ	144,762
ペン	13,547
リュック	37,264

4.1.1 Yahoo!知恵袋

Yahoo!知恵袋データについては、研究用途に研究者に提供されている「Yahoo! 知恵袋データ（第 2 版）」を利用する。Yahoo!知恵袋は、ユーザが自由に質問を投稿し他のユーザが回答を行う Web サービスである。ここで 2004 年 4 月 1 日から 2009 年 4 月 7 日に投稿された質問のデータと、それに対する回答のデータを使用する。カメラ、シャツ、シューズ、ペン、リュックの 5 つの商品について、関連がある可能性があるデータを以下の条件で選択し本研究で利用する。

(1) 質問データから質問文の中にカメラ、シャツ、シューズ、ペン、リュックが、存在する場合に抽出。

(2) 抽出された質問データに対する回答を、回答データから抽出

上記の条件により抽出された各商品ジャンルにおける質問データ、回答データについて表 4 に示す。

4.1.2 楽天市場商品レビュー

「楽天データ公開」^(注3)にて公開されている、EC サイト「楽天市場」に投稿された商品レビューのデータを使用する。本稿で使用するデータは 2010 年の 1 年間に投稿されたレビューデータである。このレビューデータから、Yahoo!知恵袋と同様に、カメラ、シャツ、シューズ、ペン、リュックの 5 つの商品ジャンルについてのレビューを抽出し使用する。各レビューに与えられる商品ジャンル ID より、商品レビューの収集を行う。実際に使用するレビュー数について表 5 に示す。

4.2 構文パターン発見の実験

4.2.1 構文パターン発見の実験方法

まず、3.6 節で述べたように構文のパターンを収集し評価を行う。使用するデータは 4.1.1 項で述べた、Yahoo!知恵袋データを用いる。シードとなる目的表現から前後の形態素列を取得し、人手で選択したパターンを加える。シードとなる目的表現を商品ジャンルごとに表 6 に示す。

4.2.2 構文パターン発見の結果

両方向の構文パターンを追加するために前項で述べたシードとなる目的表現を用いて構文パターンを発見する。得られた結

(注3) : https://rit.rakuten.co.jp/data_release_ja/

表 6 構文パターン拡張のための各商品ジャンルのシードなる目的表現

商品ジャンル	目的表現				
カメラ	撮る	夜景	旅行	体育会	入門
シャツ	着る	部屋	ランニング	夏	男性
シューズ	履く	室内	ランニング	お出かけ	入門
ペン	書く	学校	デザイン	イラスト	子供
リュック	持つ	学校	登山	通学	子供

果の各上位 5 件を表 7 と表 8 に示す。

表 7 より、「ランニングに使います」のように、文頭から利用目的が書かれるパターンが多く存在することがわかる。また、上位 100 件程度の結果を見ると、前方パターンには特徴的なパターンが見られず、「の」、「は」など一般的な助詞などが多く存在した。後方パターンにおいては前方パターンに比べると、特徴的なパターンを見つけることができた。ここで、追加する前方パターンを「、」、「(文頭)」, 後方パターンを「では」、「の場合」、「向け」とする。

4.3 定量的評価のための実験計画

得られた目的表現の評価方法について述べる。ここで、抽出された目的表現は商品に対する利用目的について表現された形態素列である。この目的表現における評価の軸は 2 つあると考えられる。

まず 1 つは、「利用目的として妥当かどうか」という点である。たとえば、「シューズ」を「コレクション」のために利用する人はいても、「絵を書く」ために利用する人はいないと言える。このように、発見された目的が妥当であるかの「妥当性」の評価を行う必要がある。

もう 1 つの軸は、「その利用目的がどれだけ認知されているか」という点である。たとえば、商品が「シューズ」の場合「MTB(マウンテンバイク)のためのシューズ」が存在している。普段から MTB に乗る人にとっては知られていることであると考えられるが、乗っていない人にとっては認知されにくいと考えられる。このように、商品に対しての様々な利用目的が、どの程度認知されているかの「認知度」を評価することで、「目的として正しいかどうか」、「意外な利用目的かどうか」といった評価を行うことも可能になる。

評価者には大きく分けて 2 つの質問に答えてもらう。まず、商品ジャンルと評価対象となった目的表現を表示し、以下の項目について 1 問ずつ、当てはまるか当てはまらないか、と「よくわからない」の 2 択で回答してもらう。

(1) 表示されている利用目的は商品に対する利用目的として正しいと思いますか？

(2) 表示された利用目的で商品を利用、選ぶことがあることを知っていましたか？

1 個目の項目において「妥当性」の評価を行い、2 個目の項目において「認知度」の評価を行う。妥当性が高ければ、得られた目的表現が利用目的として正しいことがわかり、認知度が高ければ、よく知られた利用目的であることがわかる。ここで、妥当性が高い場合において認知度が低いとき、「意外な利用目的」とであると言える。各商品ジャンルごとに 3 種類の手法、2 種類

のコーパスが存在するため、1 商品ジャンルにつき 6 つの結果が存在する。これらの上位 50 件の結果を抽出し、集合をとり発見された目的表現を評価してもらい、正解データセットとする。

4.4 目的表現の発見実験

4.4.1 実験方法と発見された目的表現

目的表現を、3.4 節と 3.5 節で述べたように、以下の 3 つの手法で目的表現を抽出する。

(1) 手法 1: 正規表現のみを用いた単純な抽出手法

(2) 手法 2: 両方向構文パターンによる挟み込み抽出

(3) 手法 3: 独立した両方向の構文パターンによる抽出

このとき、取得する形態素数を $n = 8$ とし、使用する後方パターンは表 1, 前方パターンは表 2 に示したものと、4.2 節で述べた構文パターンを追加したものとする。

ここで、各ジャンルに関して 3 つの手法を Yahoo!知恵袋と楽天市場の商品レビューの 2 つのコーパスに適用したときの上位 5 件の結果を表 9, 表 10, 表 11, 表 12, 表 13, 表 14 に示す。

4.4.2 評価結果

4.3 節で述べた、評価方法を用いて著者のうちの一名が評価を行った。商品に対する利用目的として妥当であると判断された数を表 15, 表 16 に示す。

ここで、妥当と判断されなかった利用目的について注目すると、以下の 3 点が挙げられる。

- (1) その商品に対しての利用目的としておかしい
- (2) そもそもその目的表現が利用目的になっていない
- (3) 利用目的として意味がわからない。

また、利用目的として正しいと判断され、知らなかったと判断された利用目的を以下にまとめる。

- ・ カメラ: ダイビング
- ・ シャツ: マタニティ, 授乳, アイロンプリント
- ・ シューズ: MTB, ファースト, エレクトーン
- ・ ペン: 左利き
- ・ リュック: タンデム

これらについて次項で詳細を述べる。

4.4.3 発見された利用目的についての分析と考察

抽出結果より、ある商品に対する利用目的には大きく分けて「シチュエーション」と「利用ユーザ情報」の 2 種類が存在することが考えられる。

「シチュエーション」型の利用目的には、その商品を何に使用するかという基本的なものと、より詳しく書かれたその商品をどの場所や場面で利用するかといったものが含まれる。まず、基本的な利用目的には「撮る」、「履く」、「走る」、といった動詞的な目的表現と、「ランニング」や「登山」、「ダイエット」といったサ変名詞の目的表現が存在している。また、両方向の構文パターンを用いた手法 2, 手法 3 については、「娘の通学」や「父の日のプレゼント」など複数文節からなる詳細な利用目的が発見できた。単純な正規表現を用いた手法 1 では、正答率は高かったものの、複数文節からなる目的表現はとれていなかった。次に、その商品をどの場所や場面で利用するかというのは「夜景を撮る」や「体育館で履く」、「河川敷を走る」という目的

表 7 発見された前方パターン上位 5 件

順位	カメラ	シャツ	シューズ	ペン	リュック
1 位	写真:1,391	(文頭):1,636	(文頭):167	(文頭)1,498	(文頭):349
2 位	(文頭):716	T シャツ:245	靴:65	字:276	て:37
3 位	の写真:213	で:218	シューズ:53	で:179	し:35
4 位	(文頭) 写真:166	シャツ:183	用:25	は:167	で:32
5 位	か:161	スーツ:178	の:22	し:161	自分:32

表 8 発見された後方パターン上位 5 件

順位	カメラ	シャツ	シューズ	ペン	リュック
1 位	の:1,268	の:2,267	シューズ:376	の:2,310	の:361
2 位	と:657	は:1347	の:113	が:1,415	が:261
3 位	に:506	に:979	シューズを:92	に:915	に:248
4 位	こと:416	と:824	と:83	を:815	は:146
5 位	機:305	が:559	用:81	と:600	を:117

語を含んだ目的表現で表されるものである、

次に「利用ユーザ情報」型の利用目的は、その商品を利用するユーザについて、「男性」や「女性」、「子供」といった誰が使用するかというものと、「プロ用」や「入門用」といったどのレベルのユーザが使用するかの 2 種類が見られた。

まとめたものを以下に示す。

- ・ シチュエーション
 - － バスケットボール、ランニング、登山、サッカー観戦
 - － 娘の通学、息子の遠足、父の日のプレゼント
- ・ ユーザについて
 - － 入門、初心者、子供

利用目的として正しくないと判断された目的表現の多くは、人によって解釈が分かれる物が多く見られた。例を挙げると、商品「シューズ」に対して「お店」など、お店で履くためのものなのか、お店で売るためのものなのか不明であった。次に多いのは商品に関する属性が抽出されていたことである。商品「カメラ」に対して「価格」や「機能」、「キャノン」といったものが例に挙げられる。これらはジャンル依存のストップワードや、共通ストップワードに加えて省いていく処理を検討している。本研究では Q&A サイトと商品レビューのテキストコーパスを利用したが、以下のような特性が見られた。

- ・ Q&A サイト
 - － メリット：様々な利用目的を発見できる
 - － デメリット：投稿が商品についてのべているのか判断できない
- ・ 商品レビュー
 - － メリット：商品に対するテキストであるため正しい利用目的が発見しやすい
 - － デメリット：多様な目的表現を発見しづらい

上記のように、Q&A サイトでは利用目的の網羅性が高く、正答率が低くなる傾向にある。商品レビューでは正答率が高くなりノイズが入らず、「娘の通学」などの詳細な利用目的が取得できる一方、Q&A サイトに比べるとシューズに対する「MTB」のような、一般的でない利用目的のスコアは低くなっていた。

また、評価者がその商品や利用目的について知識をどの程度

所持しているかにより、評価にばらつきがでる。「マタニティ期の方が着るシャツ」などは、女性や医療関係の知識がある人であればよく知られているが、男性では認知度が低下すると考えられる。このような問題を解決するために、評価者を増やす必要がある。

他の問題点としては、多くの人が知っているわけではない「ある商品に対する意外な利用目的」は、テキストコーパス内に多く出現しない。今回提案した手法では基本的に目的表現の出現回数により、スコアが与えられているので、意外な利用目的を発見しづらいという問題点があると考えられる。実際、商品「リュック」に対して、「ディズニーに行く時専用」や、「ブラジルに短期留学する息子」などを抽出できていたが、出現回数が低いためスコアが小さくランキング下位になっていた。利用目的を考慮した商品推薦アプリケーションなどを作成する場合は、このような下位になってしまう利用目的も考慮できるように改善する必要がある。

5. まとめと今後の課題

本研究では、両方向の構文パターンに基づくテキストマイニングにより Yahoo!知恵袋と楽天市場の商品レビューより商品の利用目的の発見を行った。提案手法により、ある商品に対して存在する様々な利用目的を自動的に発見することが可能である。また、目的表現を発見する手法を 3 種類提案し、同時に両方向の構文パターンを見つける手法についても提案を行った。今後は、構文パターンのさらなる最適化と、抽出できた目的表現を絞り込む際のスコア付け方法について改善する。さらに今回は評価者が 1 人であったため、目的表現の評価をクラウドソーシングにより行う予定である。また、商品に対する利用目的の類似性を計算可能な状態にすることで、商品の代替推薦などに活かしていく予定である。

謝 辞

本研究の一部は JSPS 科学研究費助成事業 JP16H02906, JP18H03494, JP16H01756, JP18H03243 による助成を受けたものです。また、本研究では国立情報学研究所の IDR データセット提供サービスによりヤフー株式会社から提供を受けた「Yahoo!知恵袋データ（第 2 版）」、ならびに、楽天株式会社から提供を受けた「楽天データセット」を利用しました。ここに記して謝意を表します。

文 献

- [1] 大島裕明, 田中克己, “両方向構文パターンを用いた web 検索エンジンからの高速関連語発見手法,” 日本データベース学会論文誌, vol.7, no.3, pp.1–6, 2008.

表 9 手法 1: Yahoo!知恵袋

カメラ	シャツ	シューズ	ペン	リュック
家庭	男性	練習	製図	登山
プロ	夏	室内	布	子供
撮影	子供	ランニング	子供	通学
写真	選択	ロード	家庭	旅行
オークション	冬	子供	車	遠足

表 11 手法 2: Yahoo!知恵袋

カメラ	シャツ	シューズ	ペン	リュック
撮影	男性	練習	学校	子供
家庭	夏	室内	子供	登山
初心者	洗濯	ランニング	製図	通学
プロ	子供	子供	普通	学校
写真	スーツ	初心者	初心者	旅行

表 13 手法 3: Yahoo!知恵袋

カメラ	シャツ	シューズ	ペン	リュック
撮影	着る	履く	書く	子供
撮る	行く	買う	使う	行く
写真	洗濯	練習	必要	使う
使う	スーツ	足	出来る	登山
必要	買う	ランニング	買う	買う

表 10 手法 1: 楽天市場

カメラ	シャツ	シューズ	ペン	リュック
プレゼント	仕事	仕事	プレゼント	通学
仕事	プレゼント	通勤	手帳	遠足
旅行	夏	子供	仕事	通勤
子供	主人	プレゼント	子供	子供
ブログ	息子	練習	勉強	息子

表 12 手法 2: 楽天市場

カメラ	シャツ	シューズ	ペン	リュック
プレゼント	仕事	仕事	プレゼント	通学
仕事	夏	通勤	手帳	遠足
旅行	プレゼント	子供	仕事	通勤
撮影	インナー	プレゼント	お店	子供
運動会	息子	練習	大事	旅行

表 14 手法 3: 楽天市場

カメラ	シャツ	シューズ	ペン	リュック
撮影	プレゼント	履く	プレゼント	通学
プレゼント	インナー	子供	書く	遠足
撮る	仕事	購入	手帳	娘
購入	値段	プレゼント	仕事	子供
写真	夏	通勤	購入	息子

表 15 目的表現上位 50 件における正答数: Yahoo!知恵袋

商品ジャンル	手法 1	手法 2	手法 3
カメラ	28	21	11
シャツ	43	33	16
シューズ	49	44	20
ペン	31	24	10
リュック	36	33	18

表 16 目的表現上位 50 件における正答数: 楽天市場

商品ジャンル	手法 1	手法 2	手法 3
カメラ	42	35	19
シャツ	47	35	21
シューズ	46	44	25
ペン	31	25	16
リュック	49	44	26

- [2] 高見真也, 田中克己, “検索目的に基づくスニペットの動的再生成によるウェブ検索結果の個人適応化,” 日本データベース学会論文誌 DBSJ Letters, vol.6, no.2, pp.33–36, 2007.
- [3] 徳田英準, 伊藤昌毅, 高汐一紀, 徳田英幸, “ぶらりナビ: 潜在的欲求を引き出す発見志向型ナビゲーションシステムの構築,” マルチメディア, 分散, 協調とモバイル (DICOMO2006) シンポジウム, pp.485–488, 2006.
- [4] 本田達也, 北山大輔, 角谷和俊, “オンラインショッピングサイトにおけるレビューを用いた商品対の目的判定による商品推薦,” 平成 24 年度 情報処理学会関西支部支部大会, pp.G–28, 2012.
- [5] J. McAuley, R. Pandey, and J. Leskovec, “Inferring networks of substitutable and complementary products,” Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794, 2015.
- [6] E. Filatova and J. Prager, “Tell Me What You Do and I’ll Tell You What You Are: Learning Occupation-related Activities for Biographies,” Proceedings of the Conference on Human Language Technology and Empirical Methods in

Natural Language Processing, pp.113–120, 2005.

- [7] Z. Kozareva, “Learning verbs on the fly,” Proceedings of COLING 2012: Posters, pp.599–610, 2012.
- [8] 馬縹美穂, 笹野遼平, 高村大也, 奥村学, “職業ごとの行動に関する知識の収集,” 情報処理学会論文誌データベース (TOD), vol.11, no.3, pp.12–22, 2018.
- [9] N. Aikawa, T. Sakai, and H. Yamana, “Community qa question classification: Is the asker looking for subjective answers or not?,” 情報処理学会論文誌データベース (TOD), vol.4, no.2, pp.1–9, 2011.
- [10] 小池恵里子, 伊藤貴之, 渡辺知恵美, “提示量を適正化した女性向け商品検索支援システム,” 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2013), pp.D2–1, 2013.
- [11] 益田怜央, 増田英孝, 山田剛一, 福原知宏, “家電製品の利用シーン, 要求機能, 製品体系の知識に基づく商品推薦システム: 実装と評価,” 情報処理学会 情報基礎とアクセス技術 (IFAT) 研究会, vol.2014-DD-93, no.6, pp.1–8, 2014.
- [12] ヤフー株式会社, “Yahoo! 知恵袋データ (第 2 版). 国立情報学研究所情報学研究データリポジトリ. (データセット),” 2011. <https://doi.org/10.32130/idr.1.2>.
- [13] 楽天株式会社, “楽天データセット. 国立情報学研究所情報学研究データリポジトリ. (データセット),” 2014. <https://doi.org/10.32130/idr.2.0>.