消費者の語彙と販売者の語彙の類似性を考慮した商品検索

村本　直樹† 　橋口　友哉† 　藤田　澄男†† 　申　吉浩††† 　山本　岳洋††††
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
†† ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3 東京ガーデンテラス紀尾井町 紀尾井タワー
††† 学習院大学　〒 171–8588 東京都豊島区目白 1–5–1
†††† 兵庫県立大学 社会情報科学部　〒 651–2197 神戸市西区学園西町 8 丁目 2–1
††††† 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,aa19j508,ohshima}@ai.u-hyogo.ac.jp, ††sufujita@yahoo-corp.jp,
†††yoshihiro.shin@gakushuin.ac.jp, ††††t.yamamoto@sis.u-hyogo.ac.jp, †††††yumoto@eng.u-hyogo.ac.jp

あらまし　本研究では，EC サイトにおける商品検索において消費者の語彙と販売者の語彙のギャップを考慮した手法を提案する．EC サイトなどでは販売者によって，商品情報がテキスト情報として提供されている．そこに書かれている語彙をクエリとして用いれば検索は成功する．しかし，消費者は，販売者が用いる語彙とは別の語彙を用いて検索を行うことがしばしばある．たとえば，ランニングシューズを買いたいときに「初めてのマラソン用」といったように要求を表現する場合があると考えられる．しかし,「初めてのマラソン用」という表現は販売者が提供する商品情報には現れず，代わりに,「エントリーモデル」といった表現が用いられるということが考えられる．本研究では，そのような消費者の語彙と販売者の語彙のマッチングを行う手法を開発し，消費者が知る語彙を用いた商品検索の実現を目指す．

キーワード　商品検索，特徴抽出，機械学習

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おり，消費者は商品を購入する際に，EC サイト上で提供される商品の情報を参考にすると考えられる．例えば，カメラであれば機能の有無や，センサーサイズなどのスペック表が与えられている．また，ランニングシューズなどでは図 1 のように，素材やそのシューズの特徴的な性能の情報が提供されている．ランニングシューズに詳しい消費者であればこれらの商品情報を基に，自分の欲求に適したシューズを購入することが可能であると考えられるが，詳しくない消費者にとっては簡単ではない．EC サイト上で適した商品を検索するためには，商品名をあらかじめ調べるか，多くの商品ページ上の商品情報を見る必要がある．そのため，あまり知識のない消費者にとっては非常に手間がかかる．また，商品情報に含まれる販売者側の語彙をクエリとして用いれば検索は成功するが，多くの消費者は販売者の語彙とは別の語彙を用いた検索欲求を持つことが考えられる．例えば，あるランニングシューズを購入するときに「初マラソン用」のシューズがほしいといったように要求を表現することがある．仮にこの「初マラソン用」という語彙が商品名や商品説明に含まれていれば，検索できる可能性があるが，多くの場合でこのような具体的な表現が使われることはない．商品説明は多様なの消費者に対して情報を提示するためである．

図 1　ランニングシューズの商品ページ上の商品説明の例

例として,「ゲルカヤノ」というランニングシューズの実際の商品説明の一部を以下に示す．

> GEL-KAYANO シリーズの 26 代目モデルです．反発性に優れた FLYTEFOAM[TM] PROPEL を採用．**マラソンエントリー**から使いやすく，レースまで活用できる汎用性が特長です．

「ゲルカヤノ」は「初マラソン用」に適したランニングシューズであるが,「初マラソン用」という語彙は出現しない．しかし，ほぼ同じ意味の「マラソンエントリー」と言う語彙が存在する．また，メーカー独自の商品属性として「FLYTEFOAM」という表現や,「反発性に優れた」という，商品独自の表現も存在しており，これらの特徴的な表現は「初マラソン用」に適したシューズであることを，暗に示しているという事が考えられる．

このように消費者が用いる語彙と販売者が提供する語彙には，同じ商品のことを指していてもギャップが存在している．本研究ではこのギャップに注目し，それらの語彙のマッチングを行い，消費者が知る語彙を用いた商品検索の実現を目指す．語彙のマッチングに関して，上記の例であれば，消費者が表現する「初マラソン」といった語彙と販売者が用いる「マラソンエントリー」といった語彙のマッチングを行う．これらの語彙のマッチングを行うことで，消費者がより簡単に商品検索を行うことができると考えられる．

本研究では，販売者の語彙が含まれるデータとして EC サイト上の商品説明を利用する．また，消費者の語彙が含まれるデータとして，Q&A サイト上に投稿された以下のような商品に関する質問と，その質問に対する回答のテキストデータを利用する．

> 最近は月 200 キロ前後で 3ヶ月後に**初マラソン**の予定です．ハーフは 1 時間 40 分程度で，目標はサブフォーの 35 歳です．アシックスの**ゲルカヤノ**を履いて練習してますが，本番用には向かないのでしょうか？

これは初マラソンに向けて自分の持っているシューズが適しているかどうかを聞いたものである．先程の商品説明の例で上げた「ゲルカヤノ」という商品名を含んでおり，このような商品名を含んだ質問や回答では，その商品に関する利用目的や興味のある商品の観点について，消費者の知る語彙により表現されている．これらのある商品に対する Q&A 上のテキストを消費者テキスト，EC サイト上の商品情報のテキストを販売者テキストとして，それらのペアデータを用いることで語彙のマッチングを行い，それらを利用して検索モデルを作成する．

2 関 連 研 究

本研究では，ある商品に対して消費者が用いる語彙に注目して，販売者が提供する商品に関する語彙との類似性に基づく商品検索を目的としている．

一般的に，EC サイトでは協調型の商品推薦システムを利用している事が多い [1] [2]．近年では深層学習を利用した商品検索の研究も行われている．深層学習を用いた情報検索に関する研究として，Huang らの研究 [3] が挙げられる．Huang らは Web 検索クエリが与えられたときに，クリックされたドキュメントに関して，クリックスルーデータを用いて，条件付き尤度を最大化するような深い構造化セマンティックモデルを構築している．また，Kalloori ら [4] は，ユーザのアイテムに対する評価などの絶対的な評価と，アイテムのクリック情報などの暗示的な評価から，一対比較法を利用したランキング手法を提案している．これらの研究では，テキストデータ以外のデータをモデルに組み込むことで，ユーザに適した商品検索を行っている．このように様々な情報を用いて商品検索や推薦を行う研究として，Van Gysel [5] らの研究がある．Van Gysel らはテキストベースの検索を行う先行手法に，検索ログなどを用いて，ユーザの好みを考慮した情報を組み込むことで，検索精度を向上させた．McAuley ら [6] 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．

また，消費者の語彙として利用するテキストには，消費者がその商品をどのように利用したかという情報が多く含まれるが，テキストデータから人の行動を抽出する研究も行われている [7]．たとえば，馬縁ら [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取得している．Kozareva ら [9] は，「医師の義務は何ですか？」という質問に答えるような動詞関係を学習している．

本研究では Q&A サイトのテキストデータを利用しているが，Q&A サイトの分析を行う研究も行われている．相川ら [10] は，Q&A サイト上の質問を，質問者が主観的回答か客観的回答のどちらを期待しているかの 2 種類に分類している．石川ら [11] は，Q&A サイトにおけるベストアンサーを推定できるかについて検証している．ここでは，人手による評価と機械学習による推定を「恋愛相談」「パソコン」「一般教養」「政治」の 4 つのカテゴリで行っており，「恋愛相談」以外のカテゴリでは人手と同等以上の推定精度を機械学習により実現している．

3 語彙のマッチングと商品検索

3.1 消費者の語彙と販売者の語彙

本研究では 2 つの課題に取り組む．第一に，ある商品に対する消費者の用いる語彙と販売者の用いる語彙のマッチングを行う．例えば，消費者は「初マラソン用」のシューズがほしいと検索欲求を表現するが，販売者が提供する情報には「初マラソン用」という表現ではなく，ほぼ同じ意味の「マラソンエントリー」や「エントリーモデル」といった表現が出現している．このような，ある商品ジャンルにおける，同じ商品かどうかを特徴づけるような，消費者側と販売者側の語彙のマッチングを行う．なお，本研究では消費者の語彙を含む文書を消費者テキスト，販売者の語彙を含む文書を販売者テキストとし，それぞれ Q&A サイトと EC サイト上のテキスト文書を利用する．

第二に，語彙のマッチングの結果を利用して，商品について書かれた消費者テキストをクエリとした，対象商品上における商品検索を行う．

ここで，本研究で対象とする語彙について述べる．例えば，以下のようなある商品に対する消費者テキストが存在した場合で説明する．

> 初心者にもおすすめです。
> ジョグやマラソンでも利用できると思います。
> 底が厚くて、クッションがあります。

この時，商品を特徴づけるような消費者の語彙として，まず「初心者」,「ジョグ」,「マラソン」,「底」,「クッション」が取得できる．このような単語レベルでの語彙と,「底が厚い」と

「クッションがある」のような，修飾非修飾関係にあるような表現も，本研究では語彙として扱う．これは，「底が厚い」と「底が薄い」のような「底」だけでなく，「底」が「厚い」のか「薄い」のかによって，商品としての特性が変わってしまう重要な表現も，重要な語彙であるといえるからである．

これらの語彙を消費者テキストと販売者テキストの各集合に対して取得する．取得には係り受け解析を用いる．詳細については 4.2 節で述べる．

3.2 提案アプローチ

本節では本研究で提案するアプローチについて述べる．本研究では，前節で述べたように以下の 2 つの問題に取り組む．

- 消費者の語彙と販売者の語彙のマッチング
- 消費者テキストをクエリとしたときの商品のランキング

語彙のマッチングでは，消費者テキストと販売者テキストのペアデータを作成し，そのペアデータの 2 値分類問題を解く過程で，マッチングが成立すると考えられる語彙のペアを発見する．詳細については次章で述べ，本節では簡単な例を用いて，どのように 2 値分類問題化を行い，語彙マッチングを行うのかについて述べる．

まず，図 2 のように，語彙「レース」を含む商品 A に対する消費者テキスト X と，語彙「エリートランナー」，「モデル」を含む商品 A に関する販売者テキスト Y，語彙「モデル」を含むが「エリートランナー」を含まない商品 B の販売者テキスト Z が存在したとする．この時，ペア (X, Y) とペア (X, Z) の 2 つのペアデータを作成する．ペア (X, Y) は両方とも商品 A に関するテキストペアである．このようなペアには「同一」のラベルを与える．一方ペア (X, Z) は違う商品に関するテキストペアである．このようなペアには「相違」のラベルを与える．また，消費者テキストに「レース」，販売者テキストに「エリートランナー」を含む語彙ペア 1 と，消費者テキストに「レース」，販売者テキストに「モデル」を含む語彙ペア 2 が存在する．ここで，ペア (X, Y) では語彙ペア 1 と語彙ペア 2 の両方が出現するが，ペア (X, Z) では語彙ペア 2 のみ出現する．これらの語彙ペアが出現するかどうかを基に，それぞれのペアデータを特徴ベクトル化する．具体的には表 1 のように，語彙ペアが特徴次元となり，出現するかどうかの 0/1 が与えられ，各ペアデータに対してラベルが与えられる．ここで，語彙ペア 2 に関してはラベルを特徴づけておらず，語彙ペア 1 がラベルを特徴づけることがわかる．この消費者の語彙「レース」，販売者の語彙「エリートランナー」からなる語彙ペア 1 が同じ商品に関するペアかどうかを表すラベルを特徴づける語彙のペアと言える．本研究ではこのような語彙のペアを発見する語彙マッチングを行う．しかし，このようなラベルを特徴づける特徴次元である語彙ペアを発見するには，不要な特徴次元を削除していく必要がある．そこで本研究では特徴選択と呼ばれる手法を用いて，不要な特徴次元を削除する．

検索への応用については，語彙マッチングを用いて得られた語彙ペアを用いて，ペアデータを特徴ベクトル化し．Support Vector Machine（SVM）を用いて，そのペアデータが同一商品であるかどうかの確信度を出力する．このモデルより，ある商品に対する消費者テキストを与えた時に，対象商品集合の各販売者テキストとのペアデータを作成することで，各確信度を出力し，商品のランキングを出力する．

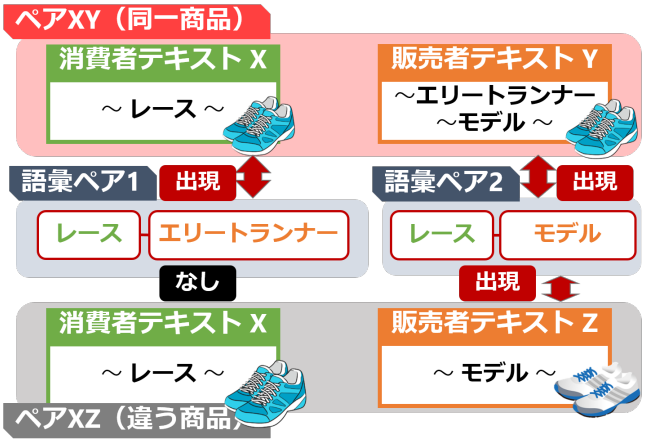

図 2　ペアデータの 2 値分類問題化

表 1　特徴ベクトルの例

	語彙ペア 1	語彙ペア 2	ラベル
ペア (X, Y)	1	1	同一
ペア (X, Z)	0	1	相違

4　係り受け解析による語彙取得と語彙マッチング

4.1 販売者テキストと消費者テキストの前処理

本節では，本研究で対象としているテキストデータに対する前処理について述べる．

本研究では，次節で述べる係り受け解析結果を用いることで，消費者テキストと販売者テキストから語彙を取得する．これらのテキストデータには多くのノイズが出現する．例えば，販売者テキストは EC サイト上の商品説明であるため，以下のような文書が存在する．

> 【最新モデル】
> 素晴らしい履き心地で、
> すべてのランナーにおすすめ！
> 軽量、加速、素足感覚を追及した勝負靴。
> ■重量：約 200g ■目標タイム：4 時間

このような文書に対して，そのまま係り受け解析を行うと，文の区切りがなかったり，記号が間に入ることから係り受け解析が正常に機能しない可能性がある．そのため以下の二点が前処理として必要となる．

- ノイズの除去
- 文書を文単位に分割

ノイズの除去において重要な処理は 4 点存在する．まず，販売者テキストには，隅付き括弧（【】）を用いたタグのような表現が頻出する．これらの隅付き括弧内の表現を 1 文とする．上記の例であれば「最新モデル」が 1 文となる．続いて，改行が文

の途中で入る場合において，読点後に改行されているものを，改行後の文と連結する．上記の例であれば「素晴らしい履き心地で、すべてのランナーにおすすめ！」が 1 文となる．次に，記号を用いて箇条書きのように表現されることも多く存在する．そこで，本研究では「○●□■△▲▽▼」で文分割を行う．上記の例であれば,「■重量：約 200g ■目標タイム：4 時間」が「重量：約 200g」と「目標タイム：4 時間」に分割される．最後に，販売者テキストにおいては「重量：約 200g」のように，商品の属性とその値や評価をコロン（：）でつなぐ表現が頻出する．そこで，このコロンを「は」に変更する．上記の例であれば,「重量は約 200g」,「目標タイムは 4 時間」のようになる．

これらの処理とその他のいくつかの前処理を以下にまとめる．

（1） 隅付き括弧の中身を一文とする．
（2） 読点後に改行されている文を読点前の文と連結．
（3） 記号を用いた箇条書き表現をそれぞれ一文とする．
（4） コロンを「は」に変更．
（5） URL を削除し，連続する記号を 1 つに統一．
（6） カタカナを全角に，英語や数字を半角小文字に修正．
（7） 「。」「!」「?」で文分割．

上記の例において前処理を行うと,「最新モデル」,「素晴らしい履き心地で、すべてのランナーにおすすめ」,「軽量、加速、素足感覚を追及した勝負靴」,「重量は約 200g」,「目標タイムは 4 時間」の 5 文が得られる．

4.2 係り受け解析結果を利用した語彙の取得

本節では，3.1 節で述べた語彙について，係り受け解析を用いてどのように取得するか述べる．

例として,「薄底シューズから厚底シューズのものに変えて速く楽にジョギングできた」という文に対して係り受け解析を行い，どのように語彙を取得するかを述べる．まず，係り受け解析から得られる情報を表 2 に示す．なお，本研究で用いる係り受け解析には CaboCha [12] [13] を利用しており，辞書については mecab-ipadic-NEologd [1] （2019 年 1 月 31 日更新分) を追加している．

係り受け解析では，対象文が形態素列に分解され，それらの形態素列が文節毎に分解されて，分節間の係り受け関係も得ることができる．これらの解析結果から語彙を取得するが，文節単位で語彙を取得するものを文節 Uni-gram，係り受け関係より語彙を取得するものを係り受け Bi-gram とする．

文節 Uni-gram

まず「文節 Uni-gram」について述べる．この文節 Uni-gram では，文節毎に重要となりえる表現を語彙とする．基本的に語彙とするのは「名詞」,「動詞」,「形容詞」,「形容動詞」,「連体詞」,「副詞」からなる形態素列を，基本形に修正したものである．この語彙を取得するための処理は大きく分けて以下の 2 点である．

（1） 文節内の形態素列の連結
（2） 不要な形態素の削除

1 : https://github.com/neologd/mecab-ipadic-neologd

表 2 係り受け解析結果の例

文節	係り先文節	表層形	品詞	詳細品詞	基本系
0	3	薄	接頭詞	名詞接続	薄
		底	名詞	一般	底
		シューズ	名詞	一般	シューズ
		から	助詞	格助詞	から
1	2	厚底シューズ	名詞	固有名詞	厚底シューズ
		の	助詞	連体化	の
2	3	もの	名詞	非自立	もの
		に	助詞	格助詞	に
3	6	変え	名詞	自立	変える
		て	助詞	接続助詞	て
4	6	速く	形容詞	自立	速い
5	6	楽	名詞	形容動詞語幹	楽
		に	助詞	副詞化	に
6	-	ジョギング	名詞	一般	ジョギング
		でき	動詞	自立	できる
		た	助動詞	-	た

最初に，形態素列の連結について述べる．例えば「薄底シューズから」という文節においては，助詞の「から」を除いて「薄底シューズ」を語彙とする．このとき，文節内の連続する形態素列が，名詞や名詞接続する接頭詞からなっていたときにそれらを連結する．連結する条件としては，品詞が「名詞」で詳細品詞が「一般」,「固有名詞」,「自立」,「副詞可能」，または品詞が「接頭詞」のものである.「軽運動」なども形態素レベルで見ると「軽（接頭詞)」と「運動（名詞)」になるが，この処理により,「軽運動」として特徴化することができる．しかし,「お休み」や「お弁当」における「お」などは，重要な表現でないと考えられるため接頭詞であるが除く．除く条件はひらがな 1 文字からなる接頭詞とする．このように連続する形態素を連結させることで，辞書に含まれない様々な表現にも対応する．上記の例においても,「厚底シューズ」は辞書に登録されており，一つの固有名詞となっている一方で,「薄底シューズ」は登録されていないため，上記の処理で一つの語彙として取得可能となる．

次に，不要な形態素の削除について述べる．表 2 における「ものに」という文節においては，品詞が「名詞」であるが詳細品詞が「非自立」のである「もの」と，助詞の「に」しか含まれていない．本手法では，この非自立名詞や非自立動詞は語彙として利用しない．このため，この文節からは語彙を取得しない．

最後に,「ジョギングできた」という文節を見ると,「ジョギング（名詞 - 一般)」と「できる（動詞)」が文節内に混在している．他にも「運動する」のような,「運動（名詞 - サ変接続)」と「する（動詞)」が混在する例が挙げられる．この時，本手法では名詞のみを語彙とする．つまり,「ジョギング」や「運動」をその文節の語彙とする．これは「できる」や「する」といった語は重要ではなく，目的語が重要であるからである．また,「あげられる」のように「あげる（動詞 - 自立)」と「られる（動詞 - 接尾)」のように，複数の動詞が存在する場合は，文節内で一番初めに出現する自立動詞のみを語彙とする．

以上の各文節に対する処理を以下に示し，実際に生成される語彙を表 3 に示す．

表 3　取得される語彙の例

特徴の種類	語彙	
文節 Uni-gram	薄底シューズ 変える 楽	厚底シューズ 速い ジョギング
係り受け Bi-gram	薄底シューズ → 変える 速い → ジョギング	変える → ジョギング 楽 → ジョギング

（1） 名詞，動詞，形容詞，形容動詞，連体詞，副詞を対象とし基本形を利用．

（2） 名詞（自立，一般，固有名詞，副詞可能）や接頭詞を連結．

（3） 名詞（一般，サ変接続）と動詞が混在する場合は名詞のみを扱う．

（4） 動詞が連続する場合は 1 つ目の自立動詞のみを扱う．

係り受け Bi-gram

係り受け Bi-gram では，1 つの文節とその係り先の文節の組み合わせから語彙を取得する．表 2 の例では，以下の 6 つの文節の組み合わせが作成される．

- 薄底シューズから → 変えて
- 厚底シューズの → ものに
- ものに → 変えて
- 変えて → ジョギングできた
- 速く → ジョギングできた
- 楽に → ジョギングできた

このとき，各文節に対して，文節 Uni-gram の時と同様の処理を行う．この処理により，文節「ものに」が削除されるため，「ものに」を含む文節の組み合わせは削除する．結果として得られる語彙を表 3 に示す．表を見ると様々な語彙が抽出されることがわかるが，これらの語彙には，商品を特徴づけるのに貢献する語彙と，貢献しない多くの語彙が含まれる．

4.3　語彙ペアの作成と特徴選択による語彙マッチング

本節では，前節で述べた語彙の取得により，消費者テキストと販売者テキストから得られた語彙を用いて，語彙のマッチングを行う手法について詳細を述べる．

はじめに，全体的な流れを以下にまとめる．

（1） 消費者テキストと販売者テキストからの語彙の取得．

（2） 全語彙ペアの作成．

（3） ラベルを付与したペアデータの作成と特徴ベクトル化．

（4） 特徴選択による語彙マッチング．

（5） （3）（4）の繰り返し．

まず，語彙のマッチングを行うために，前節で述べた語彙の取得を消費者テキストと販売者テキストの両方で行う．続いて，得られた語彙を用いて消費者と販売者の語彙の全語彙ペアを作成する．そして，3.2 節で述べたように，消費者テキストと販売者テキストのペアデータの 2 値分類問題を解く過程で，語彙マッチングを行う．なお，ペアデータを特徴ベクトル化する際に，語彙ペアが特徴次元となり，その語彙ペアが出現するかどうかの 0/1 を値として持つ．また，同一商品についてのペアデータか違う商品についてのペアデータかの 2 種類のラベルをもつ．ここで，本研究で得られる語彙ペアは約 1 億程度存在しており，有用な語彙ペアのみを残す必要がある．本研究では特徴選択と呼ばれる次元削減手法を用いて，不要な語彙ペアを削除する．

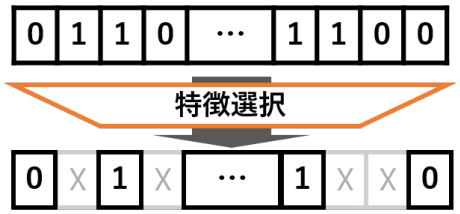

図 3　特徴選択の例

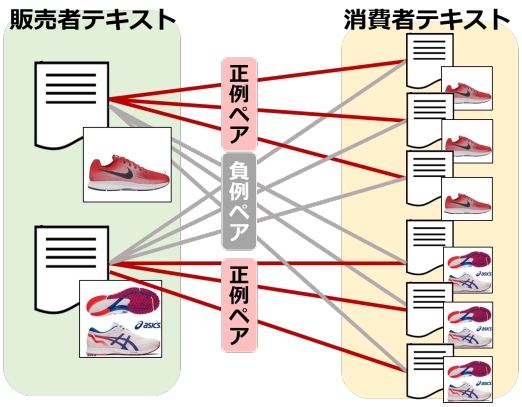

図 4　特徴選択を行う 2 商品によるバッチの例

本研究で用いる特徴選択は，図 3 のように，分類に効果的な特徴次元のみを値を変更せずに残し，ラベルを特徴づけない特徴次元や，冗長な特徴次元を削除する手法である．本研究では特徴選択の手法として，最新の手法の一つである申らの Super-CWC [14] を用いる．

特徴選択を行うことで，与えられたデータセットの特徴ベクトルとラベルを基に，そのデータセットにおけるラベルを特徴づける特徴次元（語彙ペア）を得ることができる．この特徴選択を本研究に応用するためには，先述の通り，特徴ベクトル化されラベルが与えられたデータが必要となる．ここで，本研究で対象となる商品集合から 2 商品を選択して，2 つの商品に関する販売者テキストと消費者テキストの全ペアデータを作成する．このとき同一商品に関するペアデータを正例，そうでないものを負例とする．例えば，2 商品について販売者テキストが 1 件ずつ，消費者テキストが 3 件ずつ存在した場合は，図 4 のように 12 通りのペアデータが作成される．このペアデータ集合に対して語彙ペアを基に，表 1 のような語彙ペアを特徴次元に持つ特徴ベクトル与える．そして，このペアデータ集合に対して特徴選択を行うことで，そのペアデータ集合における特徴的な語彙ペアが得られる．このように，商品を選択してペアデータ集合を作成してから，特徴選択を行うまでの流れを 1 バッチとして，対象商品数 n に対して ${}_nC_2$ 通りのバッチを作成し，それぞれのバッチにおいて特徴選択を行う．このように，繰り返し特徴選択を行い，一度でも選択された語彙ペアを語彙マッチ

ングの結果とする．

5 消費者と販売者の語彙を用いた検索モデルの作成

本章では語彙マッチングの結果を基に商品検索モデルを作成する手法について述べる．

はじめに，4 章で述べた語彙マッチングにより得られた語彙ペアを特徴次元とし，全消費者テキストと全販売者テキストの特徴ベクトル化されたペアデータを作成する．そして，それらのペアデータに対し，同一商品かそうでないかのラベルを与える．このラベルと特徴ベクトルを基に SVM を用いてモデルの学習を行い，新たな消費者テキストと検索対象となる各商品の販売者テキストとのペアデータが，同一商品であるかどうかの確信度を出力する．この確信度を基に商品のランキングを行う．

なお本研究で用いた SVM のパラメータは，scikit-learn[2] でのデフォルト設定を用いた以下の通りである．

- $kernel = \mathrm{RBF}$
- $C = 1.0$
- $gamma = 1/12127$

6 評 価 実 験

6.1 対象データと実験について

本研究では消費者の語彙を含んだテキストデータとして Q&A サイトの投稿を，販売者の語彙を含んだテキストデータとして EC サイト上の商品説明を利用する．評価実験で利用する Q&A サイトは Yahoo!知恵袋 [3]，EC サイトは Yahoo!ショッピング [4] である．Yahoo!知恵袋には，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．

本研究では，ある商品に対する Q&A サイト上の投稿と EC サイト上の商品説明を利用するため，実験に利用する商品を選択する必要がある．選択する商品は Yahoo!ショッピング上で販売されている商品とする．また，ここで，Q&A サイト上の商品に対する投稿を収集するときに，商品名が質問か回答どちらかに含まれているかを確認し，含まれている場合にそれらを消費者テキストとして収集するため，商品名により商品を断定できるものに限定する．例えば以下のような，商品名だけでは商品を特定できないものは選択しない．

ランニング 通学 シューズ

また，一部商品に関してはシリーズ化しており，例えば,「ターサージール」というランニングシューズが存在しているが,「ターサージール 5」と「ターサージール 6」のようにモデルが違うシューズがある．Q&A サイト上ではこれらを区別することが確実には行えないため，本研究ではこのような同一シリーズの旧モデル，新モデルの関係にある商品は同様のものとして扱い，そのシリーズの代表商品として最新の商品の商品情報を利用する．また，Yahoo!ショッピング上の商品カテゴリを利用し，以下のカテゴリに属する商品を対象とする．

- シューズ : スポーツ–マラソン，ランニング–シューズ–メンズ

対象商品はこれらの条件を満たす 56 商品を選択し，それぞれの商品に対する商品情報を 1 件ずつ収集している．消費者テキストに関しては Yahoo!知恵袋より，商品毎にベストアンサーに商品名を含むか，質問に商品名が含まれるもののうち，文書量が多い 11 件を取得し，その 11 件のうち，最も文書量が少ないものをテストデータ，それ以外を訓練データとする．また 11 件存在しなかったものに関しては存在する文書のみを利用する．知恵袋に存在する全データのうち，文書量が多いものを選択した理由は，その商品に関する情報が多いと考えるからである [15]．また，収集後に，各テキスト中の収集に用いた商品名を「シューズ」に変換する処理を行う．利用するデータの量について表 4 に示す．

表 4 データセット

商品数	販売者テキスト	消費者テキスト
56	56	594

これらのデータセットに対して，4.2 節で述べた語彙の取得を行い，訓練データのみを用いて，4.3 節で述べた語彙のマッチングを行う．この語彙マッチングの結果より，5 章で述べた検索モデルを作成する．そして，各商品に対して存在するテストデータの消費者テキストと，各商品に対するペアデータを作成し，訓練データで得られた語彙マッチングの結果より特徴ベクトル化を行う．このテストペアデータを用いて検索モデルの評価を行う．

以下，6.2 節では，語彙マッチングによりどのような語彙のペアが得られたかについて述べる．続いて，6.3 節では，商品検索モデルのランキング結果について述べる．

6.2 語彙マッチングの結果と考察

本節では，販売者の語彙と消費者の語彙のマッチングの結果について述べる．

6.1 節で述べた，販売者テキストと消費者テキストの両方で，4.2 節で述べた語彙の取得を行い，得られた語彙数を表 5 に示す．語彙を取得する際に，非自立名詞や助詞などを除いたのにも関わらず，少ないテキストから様々な語彙が取得されていることがわかる．

表 5 取得された語彙数

	語彙数
販売者テキスト	2,754
消費者テキスト	31,442

続いて，語彙マッチングの結果について述べる．本研究では，4.3 節で述べたように，消費者テキストと販売者テキストのペアデータに対して，データの選択の仕方を変えながら複数のバッチを作成し，そのバッチ毎に特徴選択を利用することで語彙マッチングを行う．今回 56 商品を対象としているので，56 商品の中から 2 商品ずつを選択する ${}_{56}C_2 = 1540$ 通りのバッチ

2 : https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

3 : https://chiebukuro.yahoo.co.jp/

4 : https://shopping.yahoo.co.jp/

表 6　語彙のマッチングができている語彙ペア

	販売者の語彙	消費者の語彙
語彙ペア 1	語彙ペアエネルギーロス	ジョギング
語彙ペア 2	アッパー → 合成繊維メッシュ	マラソン
語彙ペア 3	軽量 → 加速	駅伝
語彙ペア 4	上級マラソンシューズ	スピード
語彙ペア 5	仕事 → 運動	ダイエット

で特徴選択を行う．そこで，一度でも出現した語彙ペアを本手法における語彙マッチングの結果とする．表 5 より全語彙ペアは 8,691,268 次元となるが，この手法を利用することで 12,127 次元まで削減した．実際に選択された語彙ペアのうち，語彙のマッチングができていると考えられる例を表 6 に示す．

表を見ると消費者が「ジョギング」という語彙を用いている商品に対して，販売者テキスト上の商品情報での「エネルギーロス」が，語彙ペアとして選択されている．商品情報では「エネルギーロスを軽減」とあるが，利用目的である「ジョギング」に対して,「エネルギーロス」という商品の特徴が対応していると考えられる．また,「マラソン」という消費者の語彙に対して,「アッパー → 合成繊維メッシュ」という商品の属性がマッチングしている．他にも消費者の「ダイエット」に対して，販売者テキスト上の「仕事 → 運動」がマッチングしている．この語彙を含む販売者テキストの一部を見てみると，以下のような内容が書かれており，ダイエットに適した販売者テキストであると言える．

> 通学、仕事用、軽い運動。クッション性と通気性が魅力。ジョギングや軽い運動にオススメの使い勝手のよい幅広ランニングシューズです。

このように選択された語彙ペア見るだけで，消費者の語彙と，販売者の語彙のマッチングを行えているものが存在する一方，販売者の語彙「軽量化」と消費者の語彙「色」のペアや販売者の語彙「履ける → スニーカー」と消費者の語彙「ナイキ」のペアといった，あまり関係のない語彙ペアも選択されている．他にも，販売者の語彙「ズーム」と消費者の語彙「使う」のペアや販売者の語彙「やや → 硬い」と消費者の語彙「考える」のペアといった，消費者の語彙としてあまり意味のないものが選択されている．これらは，特徴選択の際の各バッチ内において，これらの語彙ペアが出現するペアデータが同一商品であったことが原因である．つまり，そのバッチにおいてのみ，この語彙の組み合わせで商品を特徴づけることができるということである．

これらの問題を解決するために，語彙取得時に消費者テキストにおいて，ストップワードを用意することを検討する．また，特徴選択時のバッチの作り方について，今回は 2 つの商品に関する全ペアデータを用いたが，その他のバッチの作成手法を検討する．例えば，選択する商品の数を増やしたり，消費者テキストにおいて，ある一定の語彙を含む消費者テキストと，全販売者テキストの組み合わせなどを検討している．

表 7　検索課題の MRR による評価結果

手法	MRR
ベースライン	0.146
提案手法	0.295

6.3　商品検索の結果と考察

本節では，6.2 節で得られた 12,127 次元の語彙ペアを用いて，消費者テキストを与えた際に，対象商品のランキングを行う検索モデルの評価結果について述べる．

本実験ではテストデータとして，消費者テキストを 56 商品に対して 1 つずつ用意し，その消費者テキストを入力した時に，56 商品のランキングを得る．このため評価指標として，ある商品 A に対する消費者テキストを入力した時に，その商品 A が何位に出現するかに注目した，Mean Reciprocal Rank（MRR）を評価指標とする．なお，ベースライン手法は，テストデータの消費者テキストと対象商品の販売者テキストを，形態素単位の Bag-of-Words（値を 2 値化）で特徴ベクトル化し，コサイン類似度を算出することでランキングしている．

56 商品それぞれのテストデータ（消費者テキスト）をクエリとした時の，MRR の結果を表 7 に示す．表より，提案手法がベースライン手法に比べて良い結果となっていることがわかる．ベースライン手法では，文書間のコサイン類似度をもとに検索を行っているため，含まれる語彙にギャップが存在した場合に類似度が低下する．しかし，本手法では語彙マッチングを行うことで，語彙のギャップを考慮することができるため，検索精度を向上させることができたと考えられる．

7　まとめと今後の予定

本研究では消費者の語彙と販売者の語彙の類似性に注目し，これらのマッチングを行い，消費者の語彙による商品検索を行った．消費者の語彙を含むものとして Yahoo!知恵袋，販売者の語彙を含むものとして Yahoo!ショッピング上のテキストデータを利用した．そして，これらのペアデータの 2 値分類問題を解く過程で，特徴選択を行うことで語彙のマッチングを行う手法を提案した．また，これらの語彙マッチングを基に，消費者の語彙をによる商品検索モデルについても提案し，ベースライン手法よりも良い結果を得ることができた．今後の予定として，語彙マッチングについて，語彙取得時の条件や特徴選択時のデータ選択のバッチの作り方などを再考することが挙げられる．また，対象とする商品ジャンルを増やすことと，語彙マッチングと検索モデルの両方について，十分な定量評価を行えていないため，再度実験を行う予定である．

謝　　辞

本研究の一部は JSPS 科学研究費助成事業 JP16H02906，JP18H03494，JP17H00762，JP18H03244，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）」を利用しま

した．ここに記して謝意を表します．

文　　献

[1] G. Linden, B. Smith, and J. York, “Amazon. com recommendations: Item-to-item collaborative filtering,” IEEE Internet computing, pp.76–80, 2003.

[2] 麻生英樹，小野智弘，本村陽一，黒川茂莉，櫻井彰人，“協調フィルタリングと属性ベースフィルタリングの統合について，” 電子情報通信学会技術研究報告.NC，pp.55–59，2006．

[3] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, “Learning deep structured semantic models for web search using clickthrough data,” Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2013), pp.2333–2338, 2013.

[4] S. Kalloori, T. Li, and F. Ricci, “Item recommendation by combining relative and absolute feedback data,” Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), pp.933–936, 2019.

[5] C. Van Gysel, M. deRijke, and E. Kanoulas, “Mix’n match: Integrating text matching and product substitutability within product search,” Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), pp.1373–1382, 2018.

[6] 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 (SIGKDD 2015), pp.785–794, 2015.

[7] 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 Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp.113–120, 2005.

[8] 馬緤美穂，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] Z. Kozareva, “Learning verbs on the fly,” Proceedings of the the 24th International Conference on Computational Linguistics (COLING 2012), pp.599–610, 2012.

[10] N. Aikawa, T. Sakai, and H. Yamana, “Community QA question classification: Is the asker looking for subjective answers or not?,” IPSJ Online Transactions, pp.160–168, 2011.

[11] 石川大介，栗山和子，酒井哲也，関洋平，神門典子，“Q&A サイトにおけるベストアンサー推定の分析とその機械学習への応用，” 情報知識学会誌，vol.20，no.2，pp.73–85，2010．

[12] T. Kudo and Y. Matsumoto, “Japanese dependency analysis using cascaded chunking,” Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002), pp.63–69, 2002.

[13] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp.230–237, 2004.

[14] K. Shin, T. Kuboyama, T. Hashimoto, and D. Shepard, “Super-CWC and super-LCC: Super fast feature selection algorithms,” Proceedings of the 2015 IEEE International Conference on Big Data (BigData 2015), pp.1–7, 2015.

[15] L.A. Adamic, J. Zhang, E. Bakshy, and M.S. Ackerman, “Knowledge sharing and yahoo answers: everyone knows something,” Proceedings of the 17th international conference on World Wide Web(WWW 2008), pp.665–674, 2008.