Consumer Generated Mediaからの購入商品推定

平野 徹　　牧野 俊朗　　松尾 義博
日本電信電話株式会社 NTT サイバースペース研究所
{hirano.tohru, makino.toshiro, matsuo.yoshihiro}@lab.ntt.co.jp

1 はじめに

マーケティング分析のひとつに，購買データに基づく購入者像の把握や購買トレンドなどの分析がある．この購買データを取得する方法としては，消費に直接関与しているスーパーやコンビニなどから POS(販売時点情報管理) データを購入する方法と，バーコードスキャナーをモニターに配布し，モニターは購入した商品に付属するバーコードをスキャンしてデータを入力するという方法がある．前者の方法は，実際の販売データに基づいた傾向把握などが行えるが，全ての販売店からデータの購入ができないため，購入した販売店で取り扱っている商品に限られたデータとなる．一方，後者の方法は，どの販売店で商品を購入してもデータが得られるが，モニターの負担が大きいため，多くても 2，3 万人の消費者の購買データを収集しているのが現状である．

そこで我々は，数千万人が利用している CGM (Consumer Generated Media) から，購買データを収集することを目指している．CGM から購買データが収集できれば，どの販売店で商品を購入してもデータが得られ，かつ，数千万人以上という規模の購買データとなる．加えて，CGM には消費者の趣味や商品に対する意見などが存在するため，これらの情報と商品購入情報を組み合わせたサービスも実現できる．例えば，意見抽出技術 [2] によって抽出された意見と組み合わせることで，購入者と未購入者の意見を比較して分析ができるようになる．

CGM から購買データを収集するために，本研究では，消費者自身が商品に対して投稿したテキストから，その商品を購入済か否か判定する研究に取り組む．具体的には，所与の商品名を含むテキストを入力として，投稿者がその商品を「購入済」か「未購入」,「不明」のいずれかを出力する問題とする．例えば，以下のテキストでは商品「健康飲料 H」(本稿では，実在の商品名の代わりに商品カテゴリとアルファベット 1 字で商品名を示す) に対して,「購入」や「飲みます」と書かれており，投稿者が「健康飲料 H」を購入したことがわかるため「購入済」となる．

> 会社で健康飲料激安で販売してる!
> わたしは**健康飲料 H** を割引期間中
> 摂取しようと今日も購入．明日も飲みます．

比較的簡単な実現手法としては，学習データを作成し単語表記を素性とした分類器を構築することである．しかし，自動車なら「運転」，飲料なら「飲む」のように対象となる商品によってその使われ方が異なるため，複数の商品を混ぜた学習データから構築した分類器では精度が低くなってしまう．また，商品毎に学習データを作成し分類器を構築することは，作成コストおよび運用コストを考えると現実的でない．

そこで，本稿では，各商品の使われ方を示す単語を大規模コーパスから自動獲得することで，複数の商品を混ぜた学習データから構築した分類器でもあらゆる商品に対して精度良く分類できる手法を提案し，その有効性を議論する．また，そもそも人は所与の商品名を含むテキストを読んで，投稿者がその商品を「購入済」か「未購入」かを判断できるのか調査しタスクの妥当性についても議論する．

以下，2 節でタスクの妥当性について，3 節で提案手法について述べる．4 節で評価実験の結果を報告し，5 節で関連研究について述べる．最後に 6 節でまとめる．

2 タスク検証

そもそも人は所与の商品名を含むテキストを読んで，投稿者がその商品を「購入済」か「未購入」かを判断できるのかを調査するために，以下の 18 種類の商品名を含むブログ 854 記事を収集した．

- デジタルカメラ L，デジタルカメラ P，スマートフォン I，スマートフォン X，自動車 F，自動車 P，テレビ R，テレビ V，電子レンジ H，ホームベーカリ G，焼酎 M，シャンパン V，健康飲料 H，健康飲料 K，洗濯用洗剤 S，洗濯用洗剤 A，スキンケア用品 N，ヘアケア用品 A

作業者には，対象商品名とその商品名を含むテキストを提示して，下記の作業を指示した．

表 1: 人手での判断結果

購入済	未購入	不明	合計
401	124	329	854

- 記事を読み，その記事を書いた人が対象商品を購入したとわかれば「購入済」を，購入していないとわかれば「未購入」を，判断できない場合は「不明」を選択する

結果を表 1 に示す.「購入済」が 401 記事,「未購入」が 124 記事,「不明」が 329 記事であった.「不明」と判断された記事には，ネット販売をしている業者のブログが 269 記事含まれおり，消費者の投稿は 60 記事であった. 販売業者のブログ記事には，商品の売り文句や入荷情報などが記載されており，この内容からこの記事を書いた人がその商品を購入したかはわからないため不明と判断された．一方で，消費者が投稿した 60 記事には，自動車のリコール問題などのニュースの引用と感想が記載されており，不明と判断された．このことから，消費者が投稿したテキストに限定すると，83.6%ものテキストに対して，人は投稿者がその商品を「購入済」か「未購入」かを判断できることがわかった.

また作業者間の一致率を調べるために，別の作業者にも同内容の指示を与えた．作業者間の判断が一致したのは，854 記事中 753 記事であり，その　値は 0.806 となり，非常に相関があることがわかった.

上記の結果より，所与の商品名を含むテキストを読んで，投稿者がその商品を「購入済」か「未購入」かを判断する問題は，人が共通の指針を持って判断できることを確認でき，妥当なタスク設定であると考えられる.

3 提案手法

所与の商品名を含むテキストを入力として，投稿者がその商品を「購入済」か「未購入」,「不明」のいずれかを出力する問題を実現する簡単な手法としては，学習データを作成し単語表記を素性とした分類器を構築することである．しかし，分類の重要な手がかりとなる商品が使われたことを示す単語は，自動車なら「運転」，飲料なら「飲む」のように対象となる商品によって異なるため，複数の商品を混ぜた学習データから構築した分類器では精度が低くなってしまう.

精度の低下を防ぐには，商品毎に，購入したことや使用したことを示す単語のリストが必要だと考えられる. このリストがあれば，当該商品のリストと一致する入力テキスト中の単語を，購入を示す単語のクラス “⟨BUY⟩” や使用を示す単語のクラス “⟨USE⟩” に抽象化することで，複数の商品を混ぜた学習データから構築した分類器でもあらゆる商品に対して精度良く分類できると期待される.

また，対象商品に対して購入や使用を示す単語が記載されなくても，人は投稿者がその商品を「購入済」か否かを判断できるテキストがある．例えば，以下のテキストでは，商品「焼酎 M」を購入したことや使用したことを示す単語は明記されていないが，一緒にチョコレートを食べたり，もうすぐ寝ると述べていることから,「焼酎 M」を飲んでいるとわかり「購入済」と判断された.

> やっとのことで帰宅．いつものように
> **焼酎 M** とチョコレートを一緒に食べてます.
> この組み合わせが最高です.
> いい気分のまま寝ます．おやすみ.

このようなテキストに対して正しく推定するためには，各商品の購入や使用を示す単語リストと関連のある単語のリストが必要だと考えられる．上記の例では,「飲む」という「焼酎 M」の使用を示す単語に関連する単語として「食べる」や「寝る」がわかれば，これらと一致する入力テキスト中の単語を，関連する単語のクラス “⟨REL⟩” に抽象化することで正しく推定可能な分類器を構築できると期待される.

そこで，商品毎に購入や使用を示す単語リスト，および，それらと関連する単語リストを大規模コーパスから事前に自動獲得し，テキストの単語表記を素性とする際に，自動獲得された単語リストと一致した単語を “⟨BUY⟩” や “⟨USE⟩”，“⟨REL⟩” のクラスに抽象化して分類器を構築する方法を提案する．提案手法は，大きく 3 つのステップに分かれる.

STEP1:購入や使用を示す単語の自動獲得 大規模コーパスから対象となる商品名を含むテキストを抽出し，各テキストにおいて商品名が道具格 (で)，もしくは対象格 (を，に) をともなって係っている先の動詞を獲得する．ここで道具格や対象格に限定するのは，消費者が商品に対して行う行動を獲得したいからである．このようにして得られた頻度上位 10 個の動詞を単語リストとする．ここで，商品を購入したことを示す単語は，商品によって異ならないので，事前に「買った」「購入」「ゲット」などの単語は人手で準備しておく．そして得られた 10 個の単語のうち，購入を示す単語と一致しない単語は使用を示す単語とする．例えば，約 7 億ページからなるブログコーパスから,「焼酎 M」を入力として単語リストを獲得すると「飲む」「呑む」「はまる」などが,「自動車 P」を入力として単語リストを獲得すると「乗る」「借りる」「運転」などが使用を示す単語として獲得された.

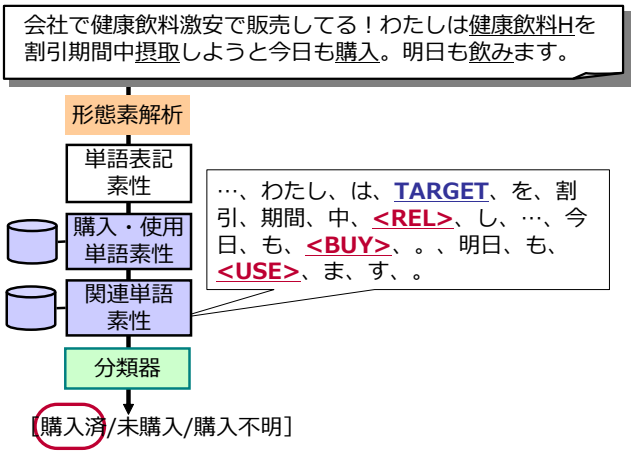

図 1: 処理構成

STEP2:購入・使用単語の関連単語の自動獲得 大規模コーパスから対象となる商品の購入や使用を示す単語を含むテキストを抽出し，各テキストにおいて該当単語が連用形で係っている先の動詞を獲得する．ここで連用形に限定するのは，購入や使用を示す行動よりも時間的に後に行われる行動を獲得したいからである．このようにして得られた頻度上位 10 個の動詞を関連単語リストとする．例えば，約 7 億ページからなるブログコーパスから,「焼酎 M」の購入・使用単語リスト「飲む」「呑む」「はまる」などを入力として，関連単語リストを獲得すると「食べる」「呑まれる」「寝る」などが,「自動車 P」の購入・使用単語リスト「乗る」「借りる」「運転」などを入力として関連単語リストを獲得すると「出かける」「帰る」「行く」などが獲得された．

STEP3:分類器の構築 ステップ 1，2 で自動獲得した，対象商品の購入・使用単語リストと関連単語リストを用いて，テキストの単語表記を素性とする際に，購入単語リストと一致した単語を “⟨BUY⟩”，使用単語リストと一致した単語を “⟨USE⟩”，関連単語リストと一致した単語を “⟨REL⟩” のクラスに抽象化して，複数の商品を混ぜた学習データから分類器を構築する．なお本稿では，ブースティングに基づく学習器 [1] を用いて分類器を構築した．また分類時には，one vs rest 法を用いたため,「購入済か否か」「未購入か否か」「不明か否か」の 3 つの分類器を構築した．

このように構築された分類器を用いて，所与の商品名を含むテキストを入力として，投稿者がその商品を「購入済」か「未購入」,「不明」のいずれかを出力する処理構成を図 1 に示す．分類器の構築時と同様に，まず入力テキストの単語表記を素性とし，次に購入単語リストと一致した単語をクラス “⟨BUY⟩”，使用単語リストと一致した単語をクラス “⟨USE⟩”，関連単語リストと一致した単語をクラス “⟨REL⟩” に抽象化する．図 1 では，購入がクラス “⟨BUY⟩”，飲むがクラス “⟨USE⟩”，摂取がクラス “⟨REL⟩” に抽象化された様子を示している．そして,「購入済か否か」「未購入か否か」「不明か否か」の 3 つの分類器から one vs rest 法で「購入済」か「未購入」,「不明」のいずれかに決定する．

表 2: 実験結果

	正解率 [%]
提案手法	78.0 (78.2)
ベースライン	68.3 (72.7)

4 評価実験

所与の商品名を含むテキストを入力として，投稿者がその商品を「購入済」か「未購入」,「不明」のいずれかを出力する問題において，提案手法の有効性を調査するために，ベースラインとしてテキスト中の単語表記をそのまま素性として用いる手法と比較した．本評価実験では，提案手法，ベースラインともに入力テキスト全てではなく，対象商品名の出現する文，および前後 1 文の単語表記を素性として利用した．提案手法では，この単語表記が自動獲得した単語リストと一致するか判定し，一致すればクラス “⟨BUY⟩” などに抽象化した．

4.1 評価データ

評価データには，2 節で述べた 18 種類の商品に対する 854 記事を用いて，10 分割交差検定を行った．なお交差検定では，商品単位でデータを分割し，評価対象となる商品は，学習データに一切含まないようにした．また比較のため，学習データに必ず評価対象となる商品のデータが含まれる交差検定も実施した．

4.2 実験結果

実験結果を表 2 に示す．なお正解率は次式の通りである．

$$\text{正解率} = \frac{\text{正しく分類できた記事数}}{\text{854 記事}}$$

実験結果から，提案手法は 78.0%とベースラインの 68.3%に比べて，9.7 ポイント向上したことがわかり提案手法の有効性が確認できた．また丸括弧の数値は，交差検定において，学習データに必ず評価対象となる商品のデータを含む場合の正解率である．これらの結果から，ベースラインでは，対象となる商品が学習データに含まれていれば 72.7%であるが，学習データに含まれなければ 68.3%まで下がってしまう．一方，提案手法は，78.2%が 78.0%とほとんど低下しないことがわかる．このことから，提案手法は，学習データにない商品に対しても頑健に推定できることが確認できた．つまり提案手

法はあらゆる商品に対して精度良く分類できる手法であることがわかる．

4.3 誤り分析

提案手法が誤って分類した 188 記事を分析してみると，2 つのケースに大別できる．

1. 複数の商品名が記述されるケース
   入力テキストに複数商品名が記載されており，それらのうち 1 つを購入したという内容の場合，購入されなかった商品も購入済と判断してしまう誤り．下記の例では，投稿者が購入したのはスマートフォン X であるが，誤ってスマートフォン I も購入したと判断した．

   > スマートフォン X，**スマートフォン I** と悩みました．でも買っちゃったよ．

2. 著者以外の行動が記述されるケース
   入力テキストに投稿者以外の人物が記載されており，その人物が対象商品を購入した内容の場合，投稿者が対象商品を購入したと判断してしまう誤り．下記の例では，投稿者ではなく先輩がスマートフォン X を購入したという内容にも関わらず，誤って投稿者がスマートフォン X を購入したと判断した．

   > 先輩が**スマートフォン X** を買いました．

ケース 1，2 は買うという行動がどちらの商品のことを指しているのか (ケース 1)，投稿者もしくは他の人物の行動か (ケース 2) が判断できればよく，述語項構造解析の知見を利用することで，更なる性能改善が期待できる．

5 関連研究

所与の商品名を含むテキストを入力として，投稿者がその商品を「購入済」か「未購入」,「不明」のいずれかを出力する研究そのものはこれまで行われていない．関連研究としては，高野らのイベント名を含むテキストを入力として，投稿者がそのイベントに行ったか否かを分類する研究 [3] がある．この研究では，我々の研究と同様に動詞に着目して，イベントに行ったとわかる動詞を人手でリストアップし，入力テキスト中にその動詞があれば投稿者はそのイベントに行ったと判断する手法である．本研究との違いは，高野らはイベントに行ったとわかる動詞を人手でリストアップしているのに対して，本研究では大規模コーパスから自動獲得する点である．

CGM の投稿者の情報を推定する関連研究として，安田らの居住地を推定する研究 [4] がある．この研究では，居住地として 47 都道府県に対応付けられた学習データを準備し，テキスト中の単語表記を素性として分類器を構築している．4 節で述べた評価実験において，比較手法としてあげたベースラインはこの手法を我々のタスクに適用したものと位置づけられる．本研究との違いは，テキスト中の単語表記を素性とする際に，我々は自動獲得した単語リストに一致した単語を “⟨BUY⟩” などのクラスに抽象化した上で分類器を構築している点である．

6 おわりに

本研究ではマーケティング分析で利用されている購買データを CGM から収集することを目指し，所与の商品名を含むテキストを入力として，投稿者がその商品を「購入済」か「未購入」,「不明」のいずれかを出力するタスクを設定した．そして人手による判定作業を実施し約 83%ものテキストに対して,「購入済」もしくは「未購入」と判断できること，2 者間の判断一致率が κ 値で 0.806 と非常に相関が高いことから，タスクとして妥当であることを示した．

また，商品の購入や使用を示す単語とその関連単語を大規模コーパスから自動獲得することで，複数の商品を混ぜた学習データから構築した分類器でもあらゆる商品に対して精度良く分類できる手法を提案し，評価実験を通してその有効性を確認した．

今後の課題としては，誤り分析で言及した 2 つのケースの問題を改善して更なる精度向上を目指す．また，本稿では所与の商品名を含むテキストのみを対象としたが，より精度の高い購買データの収集には，商品名の同一性を判定する技術にも取り組む予定である．

参考文献

[1] Kudo, T. and Matsumoto, Y.: A Boosting Algorithm for Classi cation of Semi-Structured Text, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 301–308 (2004).

[2] Nakagawa, T., Inui, K. and Kurohashi, S.: Dependency tree-based sentiment classi cation using CRFs with hidden variables, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 786–794 (2010).

[3] 高野太希, 井上潮: 文章構造に基づいた Blog からの体験情報抽出方法の提案, 第 9 回日本データベース学会年次大会 (2011).

[4] 安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹: ブログ作者の居住域の推定, 言語処理学会第 12 回年次大会論文集, pp. 512–515 (2006).