Vol.2013-IFAT-109 No.1
Vol.2013-NL-210 No.1
2013/1/11

原因表現抽出のマーケティング支援への応用

定政邦彦[†1]　細見格[†1]　赤峯享[†1]
中澤聡[†1]　石澤善雄[†1]

製品のマーケット分析には、ユーザが製品を選択する理由の把握が重要である。原因表現抽出技術を利用して製品の購入理由を Web コーパスから自動抽出することで、従来の評判分析技術では収集できない理由が収集可能となる。車の購入理由を対象として提案手法を試行し、評価表現以外の購入の理由やより関心度の高いユーザの意見が抽出できるようになったことを示す。

Applying Causal Phrase Extraction to Product Market Analysis

KUNIHIKO SADAMASA[†1]　ITARU HOSOMI[†1]　RYO AKAMINE[†1]
SATOSHI NAKAZAWA[†1]　YOSHIO ISHIZAWA[†1]

We applied causal-phrase extracting method to product marketing analysis. The method can extract a wide variety of reasons for purchasing product compared with sentiment analysis. We applied the method to extract reasons for purchasing cars, indicating that the method extracts reasons which are not sentiment, and reasons from users with higher interest.

1. はじめに

近年、Twitter や Facebook に代表される SNS の流行の影響もあり、以前にも増して個人が Web 上で意見を発信するようになった。特に、企業が販売する製品やサービスについて言及した意見は、販売元企業が製品企画や販売計画などのマーケティングを行う材料として有用なため、従来から個人の意見を自動的に収集し、マーケティング支援に用いる取り組みが行われている。

製品に関する意見収集方法として広く用いられているのが、評判分析技術である。評判分析技術とは、対象物について Web 上で個人が述べた肯定的な意見と否定的な意見を収集し分析するための技術である。評判分析技術を用いることにより、製品がどの観点で好評・不評であるかが概観できるようになる。一方で、製品マーケティングの目標は「自社製品を買って貰うこと」であるため、最も重要なのは、「何が決め手となって製品を買って貰えた/貰えなかったか」の把握となる。その観点に立つと評判分析技術を用いた意見収集には、以下に挙げる課題が存在する。

(1) 購入者の境遇など、製品の評価以外の外部要因が製品の購入理由となることも多い。また、一般に好評・不評を表す訳ではない観点が購入者の好みの問題で購入理由となることもある。つまり、対象物の評価表現を抽出する評判分析技術では、分析の母集団としての購入理由に抽出漏れが生じる。
(2) 評価表現の有無のみに注目する評判分析技術では、例えば製品の販促を目的とした製品紹介ページの記述と製品購入者の記述を区別できない。つまり、購入を行った人から、購入には興味がない人まで様々な関心レベルの意見が混在して抽出されるため、得られた購入理由がどの程度購入の決め手となりうるのかの判断ができない。

そこで本稿では、マーケティング支援を想定した意見抽出方法として、分析対象製品やそのライバル製品を購入したと明言した記述のみを処理対象とし、その購入の理由を原因表現抽出の枠組みで自動抽出するアプローチを提案する。原因表現抽出では、出来事や行為の原因となる表現をテキストから自動抽出する。本手法では、購入を明記した記述のみを対象とすることで、２つめの課題に対応、原因表現抽出の枠組みで意見を収集し、製品自体の評価に依らない購入理由も抽出することで１つめの課題に対応する。本稿では「車の購入理由」を対象として提案手法の試行を行い、得られた結果についての考察を報告する。

以下、２章では関連研究について述べる。続いて３章では提案手法について述べ、４章でその評価・考察を行う。最後に、５章でまとめと今後の課題について述べる。

2. 関連技術

評判表現を収集する既存研究としては、まず、立石ら[12]の研究が挙げられる。立石らは Web コーパスを対象とし、肯定的な意見を表す表現と否定的な意見を表す表現を予め人手で評価表現辞書として用意し、評価表現と対象物の文内共起を用いて対象物に対する意見を収集している。しかし、事前に評価表現辞書を構築するのは容易ではないため、辞書を用いず、意見を表す文と事実を表す文の分離や肯定

†1 日本電気株式会社
NEC Corporation
E-mail: {k-sadamasa@az, i-hosomi@ay, s-akamine@ak, s-nakazawa@da, y-ishizawa@bq}.jp.nec.com

Vol.2013-IFAT-109 No.1
Vol.2013-NL-210 No.1
2013/1/11

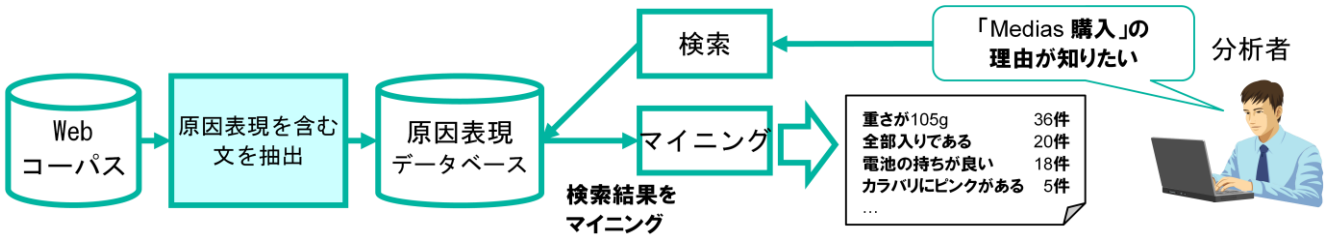

図 1　マーケティング支援システムの全体像

意見と否定意見の分離を機械学習に基づいて行うアプローチが現在主流である[13][15][15]。ただし、これらの方法では、製品の評価表現以外の意見の抽出は困難であった。

原因表現の抽出技術は、抽出対象となる原因表現の単位によって、単語レベル、文レベルの大きく２つに分類される。単語レベルを対象とする原因表現抽出では、ある出来事や行為(=結果)と、その原因の両方を単語(=名詞句)の単位で扱う[1][2][3][4]。例えば、「逮捕」という名詞句の結果表現に対応して、「犯罪」という名詞句の原因表現を抽出する。但し、単語レベルの情報から元々の原因の全貌を把握することは一般に困難であるため、本研究では利用していない。

文レベルの原因表現抽出では、原因と結果のいずれかまたは両方を述語の単位で扱う。

乾ら[5]は新聞記事を対象とし、因果関係を表しやすい手掛かりとなる表現の「ため」に着目、「ため」を含む複文の抽出後、更に述語の意志性の有無を機械学習によって判定することで細分化し、原因表現を抽出している。坂地ら[7]は、同じく新聞記事に対し「ため」以外にも多様な助詞を手掛かり表現として利用、何らかの手掛かり表現を含む文を対象として、構文木の構造を素性に用いた機械学習によって、原因表現を抽出している。

一方で乾ら・坂地らの方法では、予め定められた手掛かり表現や構文構造の文のみが抽出対象となり、それ以外の原因表現を抽出することはできない。東中ら[6]は、手掛かり表現を事前に列挙するのではなく、各文が原因を表すか否かを人手で付与した学習データから統計的に抽出している。具体的には、bact[10]という学習器を用い、ある文が原因を表すか否かを判別する際に、文の係り受け木の任意の部分木構造の各々がデータの判別に有用か否かを計算することで有用な手掛かり表現を自動抽出している。

3. 提案手法

3.1 システム全体像

想定するマーケティング支援システムの全体像を図 1 に示す。本手法では、分析対象とする製品やそのライバル製品に対しユーザが行った行為と、その行為を行った理由とを Web コーパスから自動抽出し、原因表現データベースとして蓄えておく。１つの原因表現の内部では、製品名と製品に対する行為、行為の理由を紐づけて記録しておく。製品に対する行為とは、ユーザが製品を入手したり手放したりする行為のことで、購入する、購入しない、解約する、(購入を)迷う、などが典型例となる。製品マーケティングの担当者は、蓄積された原因表現データベースに対し、分析したい対象の製品名と行為によって絞り込みを行い、絞り込み結果を直接閲覧ないしテキストマイニングすることより、製品購買行動の分析を行うという想定である。

3.2 「車の購入理由」の抽出

上述のアプローチで有用な購入理由が抽出可能かを検証するため、実際に「車の購入」行為について購入理由の抽出を行った。以下ではその手順を述べる。

まず、抽出元となる Web コーパスは、上述の評価を簡単に行うため、今回は Web 検索エンジンを用いて収集した。具体的には Google 検索を用いて、具体的な車名と、車の購入理由が多く含まれ得る絞り込みキーワードとの AND 検索を行った。用いた車名は、プリウス、ムーヴなど 18 車名、絞り込みキーワードは、「ので購入しました」「ので購入を決め」「ので買いました」「ので決めました」「ので迷っています」「決め手は」の 6 種類である。結果、9000 ページを収集した。

次に「車の購入」表現の抽出エンジンと、「車の購入理由」表現の抽出エンジン用の正解データを人手で作成した。前出の 9000 ページのうち 3393 ページに含まれる各文に対し、「車の購入」を表す箇所と「車の購入理由」を表す箇所に人手でタグを付与した。収集したコーパス中には車のパーツの購入など、車以外の購入に関する記述も出現したが、それらには購入行為、購入理由のどちらのタグも付与しなかった。結果、「車の購入」タグを含む文が 1827 文、含まない文が 212076 文、「車の購入理由」タグを含む文が 1026 文、含まない文が 212877 文、得られた。

続いて、作成した正解データを元に抽出エンジンを学習した。本稿では、抽出元となる文書が Web コーパスであり、新聞よりも原因記述の方法が幅広く、予め定められた手掛かり表現や構文構造のみを対象とした抽出では抽出漏れが多数生じてしまう。そこで「車の購入理由」の学習には、東中ら同様、手掛かりとして有用な部分木構造を自動抽出可能な bact[10]を用いて学習した。言語解析には NEC で開発している JAna[11]を用い、判定の単位は文単位とした。「車の購入」表現の出現の有無も同様に bact を用いて文単位で学習した。bact を用いた理由は、実データを調べた結果、必ずしも車名と購入を表す用言が近傍に記述されるわけではなく、柔軟な素性選択が必要であったためである。

最後に、学習に用いなかった 5607 ページ中の各文に対し、上述の抽出エンジンを適用し、購入行為と購入理由の記述箇所を特定した。うち、近接して出現する購入行為と購入理由を対応すると見なし、購入表現の最も近くに出現する車名を対象車名と見なした。

以上によって、車名と、購入行為と、購入理由の３つ組を自動収集した。

4. 評価・考察

まず、抽出された購入理由数は 1602 文となった。実験的に用意した 9000 ページという比較的小規模なコーパスからの抽出結果であるので、抽出元の文書を増加させることで更に大量の購入理由が自動抽出できるようになると考えられる。また、抽出精度を確認すべく、抽出結果からランダムに選択した 150 の購入理由を人手で評価したところ、73%の文について正しく抽出できていた。更に抽出スコア上位の 100 件にしぼると精度は 78%となる。原因表現抽出は、原因表現の記述に用いられる内容語をも考慮しないと真に原因を表すか判断できないことが多いため、ドメインを限らない場合、一般的には精度が出にくい部類の抽出タスクである。しかし今回のようにドメインを絞ると、原因表現に用いられる内容語のバリエーションが、少量の学習コーパスでも機械学習可能な範囲に近づくため、実用上有用な精度が出せることが分った。

次に、先に述べた評判分析技術での２つの課題に対処可能な抽出結果となっているかを確認した。まず１つめの「対象物の評価表現以外が抽出できない」点については、期待通り多様な表現が抽出されていた。例えば、「一度も３列シートは使いませんでした」という表現は単独では事実文でしかないが、３列シート付きの車を買わなかった理由として抽出されていた。一方で「カッコ良かったから」などの評価表現的な購入理由も抽出できている。

２つめの「抽出された意見の重要性が分りにくい」点については、必ず購入行為との共起を見るようにしているため、購入したことまで偽装した書き込みでなければ、比較的購入に直結した理由が抽出できていると考えられる。

最後に、製品アンケートによる意見収集との差についても述べる。一般にアンケートは自社製品の購入時または解約時にユーザに記述して頂くため、他社製品の購入理由や、そもそも自社製品が選ばれなかった場合にその理由を知るのは困難だった。提案手法では、他社製品の購入理由を自動抽出できる上、購入理由には製品を選ばない理由についての記述が含まれるため、上述の意見も抽出可能となる。

表 1 に購入理由の抽出例を示す。

表 1　購入理由の抽出例

車名	購入理由
スイフト	オデッセイは７人乗りでしたが、買ってから１度も３列シートは使いませんでしたのでコンパクトカーを候補に検討した結果このスイフトを選びました。
ヴィッツ	なんせミラは古いの軽だったので通勤だけで１日１１０キロ走る私としては疲れて疲れてしょうがなかったのでヴィッツにしました。
レクサス	この車の存在感と、安心して乗ることができる安全装備の数々が気に入ったので購入しました。
ステップワゴン	去年ステップワゴンスパーダを新車購入しました☆決め手はまだ若者なのでたまにしか使わない三列目を簡単に、完全にしまえたことと、Mサイズミニバンの中でピカイチでカッコ良かったからです。
アイシス	アイシスは見た目的にはミニバンの中ではまずまず、そして３列目が一番使えそうだったので決めました
セレナ	我が家はヴォクシーとセレナで迷いましたが、決め手はセレナの多様なシートアレンジでした。
ランエボ	インプレッサと迷ったんですけど、インプは乗ってる人が多いのとランエボはインタークーラーが前置きだから、という理由からランエボを選びましたw

5. まとめ

本稿では、製品のマーケティング支援を目的とした個人の意見の収集方法として、製品の購入を表す表現とその理由表現の組を抽出するアプローチを提案した。車の購入に関して上述のアプローチを試行した結果、従来の評判抽出技術では収集の難しかった、製品の評判表現以外の購入理由を抽出できることや、購入した上での意見に絞って収集できることを確認した。

一方で現状の方法では、抽出対象の製品の種類が変化するごとに正解コーパスを作成する必要があり、高コストである。今後は、複数の抽出対象の学習コーパスを有効活用するなどして、必要となる正解コーパスの削減に取り組む。

参考文献

[1] Girju: Automatic detection of causal relations for question answering, ACL Workshop on Multilingual Summarization and Question Answering, pp.76-83 (2003).
[2] Marcu: An Unsupervised Approach to Recognizing Discourse Relations, Association for Computatoinal Linguistics (2002).
[3] Saeger: Large Scale Relation Acquisition Using Class Dependent Patterns, Ninth IEEE International Conference on Data Mining (2009).
[4] Tsuchida: Toward Finding Semantic Relations not Written in a Single Sentence: An Inference Method using Auto-Discovered Rules, IJCNLP (2011).
[5] 乾: 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得, 情報処理学会論文誌, vol.45, No.3 pp.919-933 (2004).
[6] Higashinaka: Corpus-based question answering for why-questions, IJCNLP, pp.219-425 (2008).
[7] 坂地:新聞記事からの因果関係を含む文の抽出手法, 電子情報通信学会論文誌 Vol. J94-D No.8 (2011).
[8] 大友:述語項構造の共起情報と節間関係の分布を用いた事態間関係知識の獲得, 言語処理学会第 17 回年次大会(2011).
[9] Surdeanu: Learning to Rank Answers to Non-Factoid Questions from Web Collections, Association for Computational Linguistics(2011)
[10] 工藤:部分木を素性とする Decision Stumps と Boosting Algorithm の適用, 自然言語処理研究会 SIGNL-158 (2003).
[11] 佐藤: CRM 分野へ向けた日本語処理機能のミドルウェア化, 言語処理学会第９回年次大会 (2003).
[12] 立石: インターネットからの評判情報検索, 人工知能学会誌,

Vol.2013-IFAT-109 No.1
Vol.2013-NL-210 No.1
2013/1/11

pp.317-323, (2004).
[13] Yu: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Proceedings of the 2003 conference on Empirical methods in natural language processing, pp129-136 (2003).
[14] Hu: Mining and Summarizing Customer Reviews, Proceedings of the tenth ACM SIGKDD international conference, pp168-177 (2004).
[15] 藤村: 文の構造を考慮した評判抽出手法, 電子情報通信学会第 16 回データ工学ワークショップ (2005)