5S-07

トーナメントモデルを用いた属性と意見の対の抽出

小田 智己 †,a　　深川 大路 †,b
† 同志社大学文化情報学部

1　はじめに

近年，通販サイトやデジタル化されたコンテンツが普及したことで，取得できる商品・サービスの情報量は増大している。一方で，各商品を比較するには情報が膨大すぎるため，意思決定に必要とする情報が得づらくなってしまった。その結果，消費者は商品レビューなどの他人の評判を頼りにするようになっている。そこで，レビューのテキストから有用な情報を取得する研究として，機械学習を用いて，商品の側面や特徴を表す表現である「属性」，その属性に対しての評価値や主観評価を表す表現である「意見」の対を抽出する研究が行われている。

2　関連研究

中野ら [1] は SVM で作成した分類器を用いて，属性と意見を抽出し，その後，係り受け関係をもとに属性と意見の対を作成した。しかし，再現率に課題がみられた。

飯田ら [2] は属性と意見の対の抽出の手法に，1 つの意見と対となる属性を，複数の候補で勝ち抜き戦を行うことで，対として最尤の属性を同定するトーナメントモデルを用いた。彼らの手法では属性候補の辞書を分野ごとに作成したが，辞書の作成・再現は容易ではない。

そこで，本研究では属性候補の絞り込みに辞書を必要としないトーナメントモデルを提案し，中野ら（2015）と抽出性能を比較することでその有効性を検証する。

3　分析

3.1　研究方法

本研究では楽天市場の商品レビューデータ[1]を使用した。分析に使用したデータは 2019 年 1 月のレビューデータから掃除機に関するレビューを 500 件ランダムに取得したものである。レビュー文は GiNZA を用いて形態素解析で文節に区切り，係り受け解析を行った。

3.2　提案手法の概要

(1) レビュー文を区切った文節から，各文節の主辞の品詞をもとに属性の候補および意見の候補を抽出する。それらに対して，人手で属性もしくは意見であるかのラベルを与える。
(2) 中野ら [1] と同様に SVM を用いて，与えられた意見の候補が意見かどうかを推定する（3.3 節）。
(3) (2) で抽出した意見ごとに，その意見と対をなす属性を推定する。トーナメントモデルを用いる（3.4 節）。

再現率・適合率・F 値によって手法の評価を行う。

3.3　意見の抽出

区切られた文節のうち，主辞の品詞が形容詞・形容動詞・形状詞可能な名詞であるものを意見の候補とする。

各候補が意見かどうかを判別するために SVM を用いる。分類器の素性には (1) 候補が文末であるか，(2) 候補が評価値表現辞書 [3] に含まれるか，(3) 係り元文節末尾の単語の品詞，(4) 係り元文節の主辞の品詞，を用いる。

3.4　属性と意見の対の抽出

主辞が名詞もしくは未知語である文節を属性の候補として扱う。また，一つの意見に対する属性の候補は，同一文内から一文前までの文節を対象とする。

トーナメントモデル [2] では，1 つの意見に対して存在する複数の属性候補の中から 2 つを比べ，どちらが属性らしいかということを勝ち抜き戦の形式で行い，意見と対となる属性を推定する。トーナメントモデルで属性の判定に用いる分類器は SVM を用いる。

SVM の素性には (1) 候補の末尾の単語の品詞，(2) 候補がストップワードを含むか，(3) 係り先文節の主辞の品詞，(4) 直近係り元文節の主辞の品詞，(5) 意見と候補の文節間の距離，を用いる。

4　分析結果と考察

レビューデータ 500 件には属性の候補となる文節は 4966 個，意見の候補となる文節は 1398 個あった。属性とラベルを与えた文節は 314 個，意見とラベルを与えた文節は 303 個であった。

Extraction of Pairs of Attributes and Opinions using the Tournament Model
[a]Satoki Oda
[b]Daiji Fukagawa
[†]Faulty of Culture and Information Science, Doshisha University
[1]https://doi.org/10.32130/idr.2.0

4.1 意見の抽出結果と考察

意見の候補を学習データ 1129 件（意見ラベル 253 件）とテストデータ 269 件（意見ラベル 50 件）に分割して実験を行った。実験の結果，再現率：0.591，適合率：0.568，F 値：0.579，となった。

実験結果より考察を行う。偽陰性は 20 件あり，その多くが係り受け関係が要因となっていた。意見の素性には 4 項目中 2 項目で係り元に関する素性を用いており，そのことがこの誤りに影響を及ぼしていると考えられる。

偽陽性は 21 件あり，21 件中 9 件は他商品に対しての意見を述べたものであった。そのうち半数以上が他商品との比較について言及したものであったため，「より」などの比較の際に用いられる助詞や「前の」「他の」といった出現がした際に検出しないという制約をつけることなどを検討する必要がある。

4.2 属性と意見の対の抽出結果と考察

学習は 253 件の意見に対し，951 件の属性候補を用いて行った。解析の際は，意見の抽出において意見であると判断された 51 件それぞれに対して，属性候補を収集し解析を行った。実験の結果，再現率：0.588，適合率：0.577，F 値 0.582 となった。

再現率が 0.588 にとどまったことは，意見の抽出段階で正解ラベルを付与された 51 件中 20 件を抽出できなかったことに起因している。

適合率は 0.577 であったが，抽出精度に影響を与えた要因のほとんどは，トーナメントモデルが必ず 1 つの最尤な属性を対とする性質である。すなわち，意見の抽出段階で誤りであったものに対しても対を作成してしまう。誤りが 23 件あるうちの 21 件がこの誤りであった。

これらの結果から，属性と意見の対の作成の精度を向上させるためには，意見の抽出の精度を向上させる工夫を行う必要であることが確認された。

一方で，意見の抽出において，正確に抽出できた意見に対して着目すると，対となる属性を 31 件中 30 件抽出できており，高精度であったことに加え，掃除機などの分野において，特有の表現を抽出できた。これは，属性候補を抽出する際に，その分野について事前に学習した辞書を使用しない場合でもトーナメントモデルの手法が有効であることを示す結果となったと言える。

4.3 関連研究との比較

提案手法と従来手法の結果を表 1 に示す。提案手法は従来手法と比べ，再現率は向上したが，適合率は低下する結果となった。中野らら [1] の手法では 22 件の対が抽出され，提案手法では 30 件の対が抽出された。

中野ら [1] の手法で抽出できたが，提案手法で抽出できなかった対は 1 件あった。これは，提案手法で非名詞を名詞と判定したために起こった誤りであった。

逆に，提案手法で抽出できたが，中野ら [1] の手法で抽出できなかった対は 9 件あった。中野ら [1] の手法では，属性も SVM で抽出しており，これら 9 件は，属性の抽出を行った際に，誤って属性ではないと判断されたため，抽出できなかったものであった。

表 1　属性と意見の抽出結果

	中野ら	提案手法
再現率	0.431	0.558
適合率	0.786	0.577
F 値	0.557	0.582

5 おわりに

本研究では，商品レビューから商品の側面や特徴を表す「属性」と，属性に対しての評価値や主観評価を表す「意見」の対を機械学習により抽出した。その際に，属性が事前に学習した辞書によって抽出されない場合でもトーナメントモデルの手法が有効であるかを明らかにすることを目的に研究を行った。提案手法では，SVM によって抽出した意見に対し，トーナメントモデルを用いて対となる属性の特定を行った。

今後の課題として，意見の判別に用いる素性の追加や，対の同定の際の制約の追加の検討が望まれる。

謝辞

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」を利用した。

参考文献

[1] 中野, 湯本, 新居, 上浦：機械学習による商品レビューの属性-意見ペアの抽出, 情処研報, Vol. **2015-DBS-162**, No. 14, pp. 1–8 (2015).

[2] 飯田, 小林, 乾, 松本, 立石, 福島： 意見抽出を目的とした機械学習による属性-評価値対同定. 情処研報, Vol. **2005**, No. 1(2004-NL-165), pp. 21–28 (2005).

[3] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一： 意見抽出のための評価表現の収集. 自然言語処理, Vol. **12**, No. 3, pp. 203–222, (2005).