レビュー文書からの省略された属性の推定を含めた意見情報抽出

柏木 潔　　小町 守　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{kiyoshi-k, komachi, matsu}@.is.naist.jp

1　はじめに

ブログやレビューサイト，Web掲示板などには電化製品や衣服などの商品に関する個人の意見や評価が数多く存在する．このような意見や評価は，顧客が商品の購入を検討する際であったり，企業など商品開発を行う側が商品の改良や新商品の開発を行う際の重要な情報源となる．そのため，評判情報抽出や意見情報抽出に関する研究が数多く行われてきた．

たとえば，[1] では日本語ブログのデータを対象にして，＜評価の書き手，評価の対象，属性値，評価値＞という四つ組を商品に対する評価データとして抽出した．ここで**属性値**とは評価の対象において何らかの性質を示す側面のことを言い，**評価値**とは対象または対象の属性に対する書き手の意見のことを言う．例を挙げると，「動作が遅い」という評価がされていた場合，「動作」が属性値で「遅い」が評価値となる．[1]

しかし，これらの研究ではレビュー文書中に属性値が明記されている場合のみを抽出の対象としている．日本語では述語に対する項が省略されることが多く，レビュー文書においても，評価値に対する属性値が省略されているものが見受けられる．例を挙げると，下記のレビュー文において「抜群」に対しては「デザイン」，「使いやすさ」という属性値が明記されているが「シンプル」，「わかりやすい」に対しては属性値が省略されている．レビュー文書500件に対して調査を行ってみたところ，1,458件の評価値に対して388件（26.6%）の属性値が省略されていた．

> デザイン/キーボードの使いやすさが抜群です。仕事ではWINをつかっていますが、いらないソフトも入ってないし、シンプルだし、わかりやすいです。

そこで，本研究ではレビュー文書中に明記されている属性値だけでなく，省略された属性値も抽出対象とすることで，より網羅的な意見情報抽出を行うことを目的とする．省略された属性値の推定を，評価値抽出と属性値抽出の2つのフェーズに分け機械学習の手法を用いることで意見情報抽出を行った．本研究での貢献は，属性値が省略された場合を考慮した際に，どの程度の精度で意見情報抽出ができるのか調査したことである．

[1] 波線は属性値，下線は評価値を表している．

2　関連研究

Kobayashiら [1] は日本語のブログ記事を対象に意見情報抽出を行った．彼女らの手法では，評価値と属性値の候補となる表現を登録した評価値辞書と属性値辞書をあらかじめ用意しておき，それらの辞書を基に対象となる文書から評価値候補と属性値候補を網羅的に抽出する．そして，抽出した評価値候補と対になる属性値の同定をトーナメントモデル [3] を用いて行う．彼女らは，抽出した評価値候補と属性値の対に対して意見性判定を行い，最終的な抽出結果とした．

本研究ではレビュー文書中の属性値の抽出に関してKobayashiらの手法をベースにしているが，文書外の属性の推定も行う点が異なっている．

Yuら [2] はWebショッピングサイトのレビュー記事データを対象にし，属性値のランキング付けをすることで重要な属性値を特定し，レビュー文書の意見極性分類に応用した．

Yuらの研究と本研究とでは，属性値抽出を主体にしている点で類似しているが，Yuらの研究がレビュー文書の意見極性分類という応用のために極性分類と関連の強い属性値の抽出を目的としているのに対して，本研究では最終的に属性値によるレビュー文書の検索やカテゴリ分類の応用を考えているので，網羅的に属性値を抽出することを目的としている点で異なる．また，Yuらがレビュー文書中の属性値抽出を対象としているのに対し，本研究ではレビュー文書外の属性値の推定も対象としていることも相違点である．

表 1: アノテーションデータの規模

文書数		500
文数		1,420
評価値-属性値の対	文書内	1,070
	文書外	388

3 レビュー文書への意見情報のアノテーション

本研究では，対象とするレビュー文書データとして，楽天のカスタマーレビューデータ[2]を用いる．以後このデータを楽天データと呼ぶ．

本研究では楽天データの中で，ノート PC のドメインのデータを用いる．ノート PC のドメインは約 4,600 件のレビューで構成されている．その中からランダムに選んだ 500 件のデータに対して以下のように，評価値にあたる単語とその評価値に対する属性値にあたる単語をアノテーションした．

> ＜デザイン/asp/01>/キーボードの＜使いやすさ/asp/02>が＜抜群/eval/(01) デザイン,(02) 使いやすさ/01,02> です。仕事では WIN をつかっていますが、いらないソフトも入ってないし、＜シンプル/eval/*機能/03> だし、＜わかりやすい/eval*操作/04> です。

上記の例では「抜群」と「シンプル」,「わかりやすい」が評価値であり,「抜群」に対する属性値は「デザイン」,「使いやすさ」である.「シンプル」,「わかりやすい」に対する属性値は文書中になく，3.1 節で述べる属性値候補集合から「機能」,「操作」がそれぞれ選ばれる．アノテーションしたデータの規模と評価値，属性値の出現傾向を表 1 に示す．

3.1 文書外の属性値候補集合の作成

文書外の属性値を推定する際に，あらかじめ推定する属性値の候補のリストを作成する．今後この推定候補のリストを**属性値候補集合**と呼ぶ．

属性値はドメインごとに大きく異なってくるので，ドメインに依存しない属性値候補集合を作成することは困難である．ゆえに，本研究ではドメインに特化した属性値候補集合を作成した．

属性値候補集合の作成において，以下の 2 つの方法で取得した属性値を基に 107 個の属性値を選出した．

[2]http://rit.rakuten.co.jp/rdr/index.html

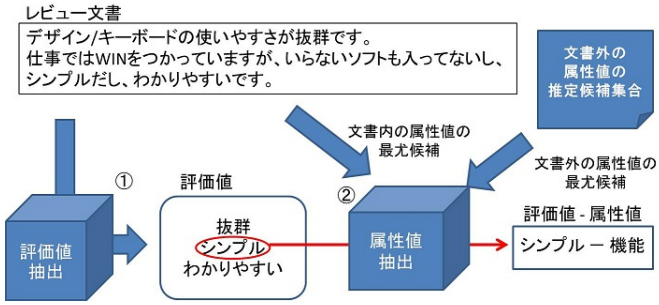

図 1: 提案手法の全体の流れ

文書外の属性値を推定する際は，この属性値候補集合の中から適切な属性値を選択する．

1. ノート PC，アクセサリなどの製品 HP の仕様表から取得した項目
2. 属性値候補集合なしの状態でレビュー文書のアノテーションを行い，その際に取得した属性値

3.2 評価値辞書と属性値辞書の作成

レビュー文書内から評価値候補，属性値候補の検出を行う際に用いる**評価値辞書**と**属性値辞書**を作成する．それぞれの辞書はアノテーションした 500 件の楽天データに出現した全ての評価値，属性値を辞書の項目として登録し作成した．

4 省略された属性値の推定も含めた意見情報抽出手法

本研究の提案手法では，図 1 で示す以下の 2 つのフェーズで意見情報の抽出を行う．

1. **評価値抽出**: レビュー文書から，評価値を抽出する
2. **属性値抽出**: 評価値と対となる属性値を抽出する

図 1 は提案手法の一連の流れを示している．
以降でそれぞれのフェーズについて説明する．

4.1 評価値抽出

評価値辞書を用いて，レビュー文書内から辞書の項目に該当する評価値候補を検出する．次に検出した評価値候補に対して実際に評価値であるかどうかの判定を行う．

評価値であるかどうかの判定には，検出した評価値候補が評価値であるかどうかの 2 値分類を行う機械学習を用いる．機械学習に用いる訓練事例は，検出した

表 2: 評価値抽出と属性値抽出に用いる素性

素性	説明
表層文字列	単語の表層
単語 3-gram	前後の 1 単語を含む 3 単語の列
品詞 3-gram	前後の 1 単語の品詞を含む品詞の列
文書中の位置	単語が文書中の何文目にあるのか
係り受け関係	評価値と属性値の間に係り受け関係があるかどうか

各評価値候補に対して，評価値か否かの 2 値のクラスと評価値に関する素性で作成する．評価値であるかどうかの判断には，文書内に属性値らしい表現が出現しているかどうかが手掛かりになると考えられるため，属性値辞書を用いて検出した属性値に関する素性も用いる．素性に関して表 2 にまとめた．

これにより評価値候補がどのような文脈で出現したのかということ，また係り受け関係にある属性値候補が存在する場合には評価値候補と属性値候補との位置関係を考慮することができる．

4.2 属性値抽出

4.1 節で抽出した評価値に対して，文書内の属性値の最尤候補と文書外の属性値の最尤候補を選出し，どちらか適切な最尤候補と評価値との対を意見情報として抽出する．

4.2.1 文書内の属性値の最尤候補の選出

文書内の属性値候補からの最尤候補の選出には，Kobayashi ら [1] と同様にトーナメントモデル [3] を用いる．評価値を言語表現，文書内の属性値候補を先行詞とみなすことでトーナメントモデルを適用し，属性値候補間での比較を行い，勝ち抜き方式で評価値に対する文書内の属性値の最尤候補を決定する．

評価値，属性値候補それぞれに対し，表 2 の素性を用いる．トーナメントモデルの学習に用いる素性は評価値に関する素性と属性値に関する素性を合わせたものになる．

4.2.2 文書外の属性値の最尤候補の選出

文書外の属性値候補からの最尤候補の選出は，3.1 節で作成した文書外の属性値候補集合から選出する．選出には多値分類の機械学習の結果を用いる．訓練事例には属性値が省略されている場合のもののみを用いているので，訓練事例に対象とする評価値が出現しない場合が考えられる．その場合は分類器は結果の出力ができないので，訓練事例の中で最も多く出現した文書外の属性値を最尤候補として選出する．

機械学習に用いる訓練事例は，クラスに評価値に対して省略された属性値を割り当て，それに評価値に関する素性を加えて作成する．また省略された属性値候補には係り受け関係がないので，素性には表 2 の素性から係り受け関係を除外したものを用いる．属性値を推定するべきクラスとするため，属性値に関する素性は用いることができない．

4.2.3 意見情報の抽出

4.1 節で抽出した評価値に対して，4.2.1 節で選出した文書内の属性値の最尤候補と 4.2.2 節で選出した文書外の属性値の最尤候補のどちらが適切なのか判別し，評価値と属性値の対を意見情報として抽出する．

判別には 2 値分類の機械学習を用い，訓練事例は，文書内か文書外かの 2 値のクラスと評価値・文書内の属性値候補の素性で作成する．素性には，評価値と属性値それぞれに対し表 2 の素性を用いる．

5 省略された属性値の推定実験

この節では，提案手法に関して行った性能評価実験について述べる．この実験では提案手法でどの程度文書内の属性の抽出，文書外の属性の推定が正しく行えるのかを確認することを主な目的としている．

5.1 実験データ

この実験のデータには 3 節で述べた楽天データを用いる．アノテーションしたデータの中には評価値が文書中に存在しないデータがあるので，それらを除いてランダムに選び出した 400 件のアノテーション済み楽天データを使用し，5 分割交差検定を行った．

5.2 評価尺度

評価値抽出と属性値抽出のそれぞれを適合率，再現率，F 値（適合率と再現率の調和平均）によって評価する．4 節で述べたように，提案手法は 2 つのフェーズで構成されているので，この実験ではそれぞれのフェーズに関する精度を確認した．各フェーズでの入力データは前のフェーズでの推定結果を用いている．

5.3 ベースラインの設定

本研究の提案手法の精度を測る比較対象として，以下のような 2 つのベースラインを設定する．

表 3: ベースラインと提案手法の比較実験結果

手法，素性	評価尺度	評価値抽出	属性値抽出	
			文書内	文書外
ベースライン 1 (最頻出)	適合率	0.59	0.15	0.36
	再現率	0.83	0.21	0.52
	F 値	0.69	0.18	0.42
ベースライン 2 (表層文字列のみ)	適合率	0.83	**0.24**	0.50
	再現率	0.88	**0.24**	0.51
	F 値	0.85	**0.24**	0.47
提案手法	適合率	**0.87**	**0.24**	**0.55**
	再現率	**0.91**	**0.24**	**0.55**
	F 値	**0.89**	**0.24**	**0.55**

評価値に対する最頻出の属性値の選出　訓練データに出現する各評価値に対して，属性値との共起の回数を計測しておく．テストデータの評価値に対して，訓練データに出現した評価値の中に一致する評価値がある場合，その評価値と最も多く共起した属性値を出力とする．もし訓練データに出現しなかった評価値であれば，評価値でないと判断する．

提案手法と同様の手法で用いる素性が最も単純なもの　各フェーズの機械学習において表層文字列の素性のみを使用する．

5.4　実験結果

表 3 にベースラインと提案手法との比較結果を示す．比較実験の結果から，評価値抽出，属性値抽出の両方において最頻出の属性値を出力するベースライン 1 を上回り，評価値抽出，文書外の属性値抽出において表層文字列の素性のみを用いたベースライン 2 も上回る性能を示した．この結果から，評価値抽出，属性値抽出に関して頻度を用いる手法よりも機械学習を用いる手法が有効であり，特に文書外の属性値の抽出に関して有効であることが確認できた．

6　考察

今回，機械学習の分類結果を用いる提案手法と頻度を用いるベースラインとの比較を行ったが，それぞれの属性値の出力を確認してみると，212 個の出力の中で 61 個と 3 割程度の出力が一致していた．また，それぞれの正解数を確認してみると，提案手法とベースラインがともに正解していた数が 47 個，ベースラインのみ正解していた数が 11 個，提案手法のみが正解していた数が 25 個であった．このことから，機械学習ベースの手法に属性値に関する事前知識を追加することが今後の課題となる．

また，文書外の属性値に関して正しく推定できた事例，誤った推定をした事例を調査してみると，いくつか特徴がみられた．誤った推定をしてしまう要因として 2 つほど特徴があり，それは曖昧性とデータ数の問題である．曖昧性がある，つまり色々な属性値と共起する評価値ではそれだけ学習事例も分散してしまい機械学習による出力が定まらないので誤った推定をしてしまいやすい．こういった問題に対応するためには，その評価値がどういう話題で出てきているのかなど他の特徴語の情報が必要になってくると考えられる．誤った推定のもう一つの問題であるデータ数に関しては，特に訓練事例に評価値と正解の属性値が出現しない場合問題になってくる．機械学習ではそもそも学習ができなければ正しい推定を行うことができないので，訓練事例を増やし，より正解の属性値が出現しやすくすることで今回の実験で推定できなかったものも推定できるようになると考えられる．

7　おわりに

本論文では，意見情報抽出において評価値に対する属性値が省略される場合を考慮し，省略された属性値の推定を含めた意見情報抽出の手法を提案した．省略された属性値の推定実験の結果から，評価値抽出と属性値抽出，特に文書外の属性値抽出に関して 2 つのベースラインを上回る精度を示すことが確認できた．

今回はドメインをノート PC に限定し，ドメインに特化したアノテーションを行い属性値の推定候補集合を作成することで約 6 割の精度で文書外の属性値を推定することが可能になった．しかし，様々なドメインで今回のようなアノテーションや推定候補集合を作成することは高コストであるので，今後の課題として今回行ったアノテーションを他のドメインに適応していき，低コストでのアノテーションや推定候補集合の生成を目指していくことが考えられる．

参考文献

[1] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proc. of EMNLP*, pp. 1065–1074, 2007.

[2] Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. Aspect ranking: Identifying important product aspects from online consumer reviews. In *Proc. of ACL*, pp. 1496–1505, 2011.

[3] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌, Vol. 45, No. 3, pp. 906–918, 2004.