程度・信頼度表現を考慮したWeb文書からの評判情報抽出手法

矢野　宏実†　　目良　和也‡　　相沢　輝昭‡

†広島市立大学大学院情報科学研究科
‡広島市立大学情報科学部

E-mail: † yano@nlp.its.hiroshima-cu.ac.jp,　　‡ {mara,aizawa}@its.hiroshima-cu.ac.jp

1 はじめに

近年，個人サイトやblogの普及により，Web上に存在する情報は非常に大量かつ多様になっている．この中には対象の評判に関する情報も多数存在するが，これらを人手で一つ一つ評価することは困難である．そのため，評判の自動収集および自動評価に関する研究が行われている．

小林は意見を(対象，属性，評価)の3つ組で定義し，それらを自動収集する手法を提案している[1]．しかし，属性の程度や意見の信頼度は考慮されていない．そのため，例えば，2つの意見「Aは**少し**高い」，「Aは**非常に**高い」があった場合，どちらも3つ組で表現すると(対象:A，評価:高い)となり，2つの意見のニュアンスの違いを考慮できない．また，「Aは壊れやすい**かも知れない**」と「Aは**絶対**壊れやすい」のような意見自体の信頼度も考慮できない．

本研究では飲食物に関する評判を記述したWeb文書を対象として，小林らが定義した3つ組に程度，信頼度の2要素を加えた，(対象，属性，評価，程度，信頼度)の5要素からなる評価情報を，評価情報抽出ルールを適用することで抽出する手法を提案する．

本研究ではまず，程度表現と信頼度表現について定式化し，評価情報を定義する．そして，提案手法で用いる評価情報抽出ルールと抽出結果の絞込みルールを作成する．そして，抽出精度を向上させるために対象／属性／評価表現リストを作成し,評価節および評価情報の抽出に用いた．最後にこれらの妥当性を評価するための実験を行う．

2 評価情報抽出手法

2.1 程度表現と信頼度表現

評判文をもとに人間が対象を評価する際，特に飲食物など五感に訴える属性を評価する際には，属性の強度を表す程度表現および情報の確かさを表す信頼度表現といった,微妙なニュアンスをも加味していると考えられる．

程度表現とは，「とても」「少し」といった程度副詞を中心とする強調表現である[2]．信頼度表現とは「おそらく～だろう」「～とは限らない」といったモダリティ表現である．本研究では，程度表現として副詞を考え，信頼度表現として[3]で用いられているモダリティ表現12種類を用いる(表1)．

表1　モダリティ表現

肯定モダリティ表現	太郎は結婚するだろう
	太郎は結婚するかもしれない
	太郎は結婚すると信じています
	太郎は結婚すると思っています
	太郎が結婚すると考えています
	太郎が結婚するということを知っています
	太郎は結婚すると感じています
否定モダリティ表現	太郎は結婚しない
	太郎が結婚するとは限らない
	必ずしも太郎は結婚するということはない
	太郎が結婚するかどうかわかりません
	太郎は結婚するかどうか知りません

2.2 程度・信頼度情報を含む意見の抽出

入力文から最終的に抽出される評価情報は(対象，属性，評価，程度，信頼度)の5要素である．各要素の候補として，本研究ではそれぞれ以下の品詞を対象とする．

- 対象と属性･･･名詞一般，サ変名詞，未知語
- 評価･･･形容詞，形容動詞，動詞，名詞一般，サ変名詞
- 程度･･･副詞
- 信頼度･･･モダリティ表現

また、各要素の抽出について，以下の3つの制限を加えた．

- 代名詞と非自立名詞は評価対象・属性・評価候補としない
- “＜副詞＞(と)している”または“＜副詞＞(と)した”の形をとっている副詞を評価候補とみなす
- “＜副詞可能名詞＞の｜に”の形をとっている場合，これを副詞とみなす

このような評価情報を抽出するために，本研究では評判文の記述パターンに注目した．このパターンを形態素解析結果に基づいて記述し，評価情報抽出ルールとしてまとめた．さらに，抽出結果に含まれる過剰抽出例を解析して，抽出結果絞り込みのためのルールを作成した．

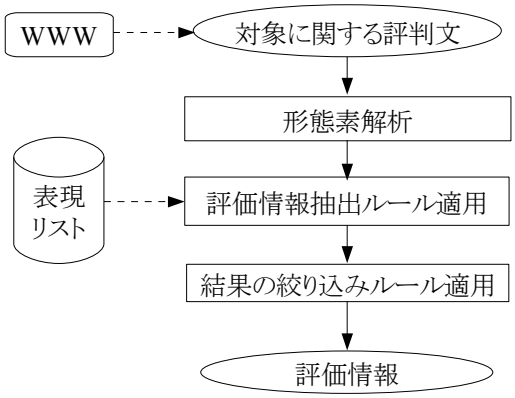

図 1　本手法の処理の流れ

図1に本手法の手順を示す. まず, 入力文に対して形態素解析を行い, その結果に対して評価情報抽出ルールを適用する. 次に, 過剰抽出を除外するための絞込みルールを適用する. また, 評価情報抽出ルール適用時に表現リスト(3.3節参照)を併用することにより, 抽出精度をより向上させることができる.

3 評価情報抽出ルール

3.1 評価情報抽出ルール

評価情報抽出ルールとは, 評価節(評価情報を含む節)を特定し, 評価節から評価情報を抽出するためのルールである. これらのルールは評価節の記述パターンを形態素の並びで表現している. これらは小林の研究で使用されたテンプレート[1]と矢野の研究で使用されたテンプレート[2]をもとに作成した.

評価情報抽出ルールを図 2 に示す. 〈評価節〉は, 対象表現や評価表現を記述した部分である〈評価節_0〉とその評価の確かさを表現する〈信頼度〉からなる. 〈評価節_0〉は〈対象〉と〈属性〉と〈評価句〉 の関係を表したものであり, 全部で 10 パターンある. 〈対象〉および〈属性〉の構成は, 「A や B」のように〈対象表現〉あるいは〈属性表現〉が並列に表現されている場合を考慮して, 〈並列表現〉を含んでいる. 〈評価句〉は, “〈評価表現〉は直前に〈程度表現〉を伴うことが多い”という特徴を考慮した構成となっている.

ルールを適用しての評価情報の抽出例を「ゆず胡椒の量が少し多いと思う。」という評判文を用いて示す.

まず、この入力文を形態素解析する. そして, 入力文の形態素の並びを図 1 に示したルールと比較して, 適用可能なものがあるか調べる. この例文は図 2 のルールに即して表現すると, 図 3 のような構成になっている. これにより, 評価情報(ゆず胡椒, 量, 多い, 少し, 思う)を得る.

```
＜評価節＞::=(＜信頼度＞)＜評価節 _0＞(＜信頼度＞)
＜評価節 _0＞::=＜対象＞の＜属性＞
                        が|は|も|に|を＜評価句＞
        | ＜対象＞の＜評価句＞な ＜属性＞
        | ＜対象＞の＜評価句＞
        | ＜評価句＞(の|な)＜対象＞|＜属性＞
        | ＜評価句＞のある＜対象＞
        | ＜属性＞が|は|も|に＜評価句＞
        | ＜評価句＞＜接尾辞＞の＜属性＞
        | ＜評価句＞＜対象＞
        | ＜評価句＞＜属性＞の＜対象＞
        | ＜評価句＞
＜対象＞::=＜対象表現＞(＜並列表現＞＜対象＞)
＜属性＞::=＜属性表現＞(＜並列表現＞＜属性＞)
＜評価句＞::=(＜程度表現＞)＜評価表現＞
＜接尾辞＞::=的|風|性|系|感
＜並列表現＞::=や|と|また|または
＜信頼度＞::=モダリティ表現
```

図 2　評価情報抽出ルール

```
＜評価節＞
    |－(＜信頼度＞)
    |－＜評価節 _0＞
    |        |－＜対象＞－ゆず胡椒
    |        |－の
    |        |－＜属性＞－量
    |        |－が
    |        |－＜評価句＞
    |                 |－＜程度表現＞－少し
    |                 |－＜評価表現＞－多い
    |－＜信頼度＞－と思う
```

図 3　評価情報ルール適用例

3.2 絞り込みルール

評価情報抽出ルールの抽出結果に含まれる過剰抽出例を解析して, 抽出結果絞り込みのための2つのルールを作成した.

a) 評価情報抽出において, 対象・属性・評価のうち, 抽出できたのが評価のみになった場合, 評価の品詞が名詞一般または動詞であれば, この評価節および評価情報を棄却する
b) 名詞一般が読点で並列に記述されている場合, 読点を並列表現とみなす

(a)については、「ちょっと立ち寄った」の(NIL, NIL, 立ち寄る, NIL, NIL)のように, 動詞単体しか抽出できない場合は, そのほとんどが記述者の行動を表していたことに起因する.

ラーメン, スープ, 麺, チャーシュー, 具, 中華そば, 海鮮, トンコツ, 味噌, 塩, 醤油, サービス, (待ち)時間, (お)店, 料理, (サイド)メニュー

図4　対象表現リスト(一部)

味わい, ダシ, 量, コク, コシ, 隠し味, 好感, パンチ, 印象, 味, 香り, 感じ, 風味, 雰囲気, におい, しつこさ, 辛さ, うまみ, しゃきしゃき感, お得感

図5　属性表現リスト(一部)

あっさり(系), つるつる(系), ある, 利く, すごい, 限定, 普通, きれい, 洗練された, 損なう, 絶妙, 薄味, 珍しい, 重視, 大量, たくさん, 柔らか(い)

図6　評価表現リスト(一部)

また, (b)については,「チャーシュー, ネギ, モヤシがのっている」のような記述がある場合でも,「チャーシューとネギとモヤシがのっている」と同様に処理をするためのものである.

3.2 表現リスト

評価情報抽出ルールと絞込みルールを適用した場合の評価情報抽出結果は, 部分的にでも抽出できたものを含めた場合,再現率 74.3%,精度 51.1%であった. 再現率が高く,精度が低いということは,所定のものをよく抽出しているが,不要なものもまた抽出していることを示している. そこで、さらに精度を向上させるために、ドメインをラーメンに限定して対象・属性・評価表現リストを作成し, 評価情報抽出ルールと併用することにした. これらの表現リストは矢野の研究で用いられた評価対象語辞書と評価語辞書[2]をもとに作成した. 対象表現リストに登録された語は 30 語, 属性表現リストに登録された語は 64 語, 評価表現リストに登録された語は 110 語である. 各リストに登録されている語の一部を図 4, 図 5, 図 6 に示す.

リストを使用しての評価節および評価情報の抽出手順は以下のようになる. まず, 形態素と評価表現リストから評価要素のある場所を特定する. これにより評価節候補を獲得する. 次に, 評価節候補に図 2 のルールを適用して他の要素を抽出する. その抽出結果が評価のみとなり, 品詞が動詞であった場合, リストに登録されている語であったとしても破棄する. 他の要素も抽出できた場合, 抽出結果とリストを照らし合わせることで,不適切な評価節および評価情報を破棄したり,抽出された要素を対象として抽出すべきか属性として抽出すべきかを判断したりする.

4 実験と考察

評判に関する記述を含む Web 文書 289 文(評価節 261 箇所)に対して, 3.3 の表現リストを使用しなかった場合と使用した場合での評価実験を行った. 結果を表 2 に示す.

表2 適用結果

リスト		入力文書 → 評価節	評価節 → 評価情報	入力文書 → 評価情報
不使用	再現率	74.3%	74.3%	74.3%
	精度	50.9%	51.1%	49.2%
使用	再現率	91.6%	75.1%	75.1%
	精度	89.9%	73.7%	71.3%

(野菜ベースのラーメンは味の印象が弱くなりがちで, 他店の野菜ベースのラーメンではそれを補う工夫がなされている.)
しかし, この店ではそれもない.
→ 評価情報 (店, NIL, ない,NIL,NIL)

図7　抽出結果に問題がある例

リストを使用しなかった場合の評価節および評価情報の抽出結果は,人手で与えた正解との部分一致で評価すると, 表 2 のようになった. しかし, 表現リストを適用することで, 抽出できなかったものが抽出できるようになり, 過剰抽出の大部分を除去することもできた. そのため, 正しくルールを適用し, 評価情報を適切に抽出できるものが増え, 再現率・精度ともに向上した.

しかし, もともと適用可能なルールが存在しない場合や, 抽出すべき表現がリストにない場合は抽出できないという問題がある. また, 図 7 のような, 現在は誤りとしていないが抽出結果に問題があると思われる事例もある. そのため, 先行研究を参考に, 効率的なルール獲得やリスト作成方法を考えたり, 照応解析手法などの導入を検討することが今後の課題となる.

参考文献

[1] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一:テキストマイニングによる評価表現の収集, 情報処理学会研究報告NL-154-12, pp.77-84(2003)
[2]吉江誠:真偽疑問文に対する返答発話の肯定／否定意図解析手法の改良, 平成14年度広島市立大学大学院情報科学研究科修士論文(2003)
[3]青山広:真偽判定と確信度, 計量国語学, 第21巻, 第1号, pp1-10(1997)
[4]矢野宏実, 目良和也, 相沢輝昭:嗜好を考慮した評判情報検索手法, 情報処理学会研究報告NL164-28, pp.165-170(2004)
[5]矢野宏実:程度・信頼度表現を考慮したWeb文書からの評判情報抽出手法, 平成17年度広島市立大学大学院情報科学研究科修士論文(2006)