共起情報と品詞情報を利用した高再現率コーパスの構築手法の提案

鈴木健之† 丸山広† 中村太一†
東京工科大学 バイオ・情報メディア研究科†

1. まえがき

企業にとり，自社あるいは他社の商品の評判を把握することは重要である[1].

評判は，対象表現・仕様表現・評価表現で構成される(以下，3つを評判構成要素と呼ぶ)[2].

評判構成要素毎の辞書の精度(適合率) を重視し，構文パタンを用いて半自動で辞書を構築する手法がある[3]. しかし，評判獲得のための分析手法は複数あり，全ての分析手法に適用可能な取りこぼしのない評判構成要素毎の辞書を用意することが重要となる.

Web の文書から取りこぼしなく専門用語を抽出する研究がある[4]. しかし，この手法は特定の文書と文書集合全体への偏差を考慮して専門用語を抽出するため，出現頻度の少ない評判構成要素を取りこぼしてしまう. また，抽出対象の文書から評判構成要素を取りこぼしなく抽出すると評判構成要素以外の語(以下，不要語)が多く抽出される問題がある.

本研究は，評判構成要素の候補語を抽出対象の文書から抽出し，取りこぼしのない評判構成要素毎の辞書を構築する. 次に辞書毎の登録語から品詞情報を用いて明らかな不要語を除外し，残った辞書毎の登録語から助詞と共起しない不要語を除外することで，取りこぼしがなく不要語の少ない評判構成要素毎の辞書を構築する手法を提案する.

2. 再現率を重視した辞書構築手法

2. 1 段階的に不要語を除外する手法

品詞情報を利用して評判要素毎の辞書登録語から明らかな不要語を除外し，除外結果から更に助詞と共起しない辞書毎の登録語を除外する.

このように段階的に不要語を除外していくことが，取りこぼしのない評判構成要素毎の辞書の構築に有効であると考える.

Methodology for building high recall ratio corpus using co-occurrence relation and part of speech.
†Kenshi SUZUKI †Hiroshi MARUYAMA
†Taichi NAKAMURA; Tokyo University of Technology Graduate School of Bionics, Computer and Media Sciences

この段階的な除外処理の流れを図1に示す.

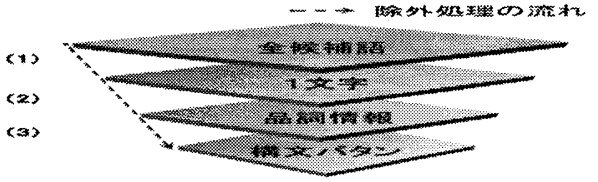

図1 段階的な不要語の除外処理の流れ

2. 2 今回実装した手法

本提案手法では，品詞情報を獲得するために，日本語形態素解析器 MeCab を用いた. 形態素解析時に，名詞が連続，同じ品詞が連続，動詞の直後に助動詞が存在，語の前後に接頭語・接尾語が存在した場合，その文字列を結合し，語解析結果とする.

(1)語解析結果を評判構成要素毎の辞書に登録し，全辞書の登録語から漢字以外の1文字の登録語を除外する.

(2)漢字以外の1文字を除外した結果から，表1の評判構成要素毎の辞書に対応する品詞情報以外の登録語を除外する.

表1 辞書と対応する品詞情報

辞書	対応する品詞情報
対象表現	名詞-, 接頭詞-名詞接続
仕様表現	名詞-, 接頭詞-名詞接続, 動詞-, 形容詞-
評価表現	名詞-, 接頭詞-, 動詞-, 形容詞-, 副詞-, 連体詞

(3)品詞情報で除外した結果から表 2 のいずれかの構文パタンにも該当しない辞書毎の登録語を除外する.

表2 構文パタン

No	構文パタン
1	仕様表現 + 語[0,1] + 助詞 + 語[0,1] + 評価表現
2	対象表現 + 語[0,1] + 助詞 + 語[0,1] + 仕様表現
3	対象表現 + 語[0,1] + 助詞 + 語[0,1] + 評価表現
4	対象表現 + (、|, |・) + 対象表現

3. 評価実験

3. 1 実験方法

携帯電話に関する Weblog 記事 500 件に含まれる 4,357 文から携帯電話の評判構成要素の候補語を抽出した. 抽出した候補語は，異なり語数

14,626 であり，この候補語から人手で対象表現 80 語，仕様表現 356 語，評価表現 433 語を正解として抽出した．

3．2　実験結果

全ての辞書の登録語から漢字以外の 1 文字を除外した結果，残った登録語は，全て 14,502 であった．この結果から更に品詞情報により除外し残った辞書の登録語は，対象表現辞書 10,414，仕様表現辞書 13,668，評価表現辞書 14,155 であった．品詞情報による除外時点での再現率は全て 100%である．

品詞情報による除外結果から構文パタンを利用して除外した結果を表 3 に示す．

表 3 構文パタンによる除外結果

評判構成要素	残った辞書登録語	再現率
対象表現	9,111	91.3%
仕様表現	12,538	97.5%
評価表現	10,622	94.7%

この結果,対象表現から 37.7%，仕様表現から 14.3%，評価表現から 27.4%の不要語を除外した．

4．　考察

表 3 から仕様表現辞書に残った登録語は，他の 2 つの辞書に比べ多い．これは，商品名の省略により表 2 の構文パタン No.2 で不要語が多く共起したことが原因であった．

そこで，表 2 の構文パタン No.2 が該当した場合，仕様表現を除外するようにした．その結果，仕様表現辞書には，残った辞書登録語が 10,777 であり，再現率は，変更前に比べ 2%低下したが，不要語を 12.0%多く除外できた．この結果から，評判構成要素を抽出対象とする文書に合わせて構文パタンを厳選する必要があると言える．

また，構文パタンの制約の程度を変化させた結果を調査した．構文パタンの制約を厳しくするために，表 2 の構文パタン No.1 と No.3 の助詞を(は|が|も|に|を)に，構文パタン No.2 の助詞を“の”に変更した．他方，制約を緩くするために表 2 の構文パタンを表 4 のように変更した．

構文パタンの制約を変更，および変更しない場合の除外語数を図 2 に，再現率を図 3 に示す．

表 4 制約が緩い構文パタン

No	構文パタン
1	仕様表現 + 語[1,2] + 評価表現
2	対象表現 + 語[1,2] + 仕様表現
3	対象表現 + 語[1,2] + 評価表現
4	対象表現 + 語 + 対象表現

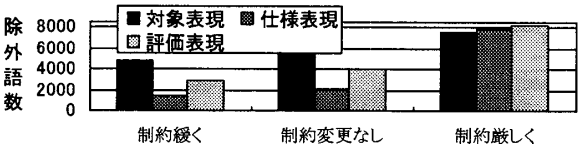

図 2 構文パタンの制約の違いによる除外語数

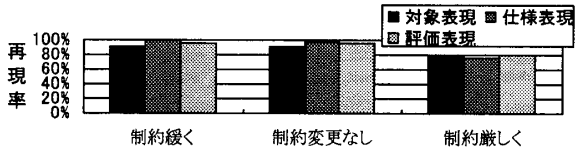

図 3 構文パタンの制約の違いによる再現率

図 2 と図 3 から制約が厳しくなるほど，除外語数は増加するが，再現率は低下する．制約の厳しい構文パタンを適用した場合の取りこぼした語を拾い上げるために，抽出精度を重視した手法が利用できると考えている．

再現率重視の提案手法の検証のために除外手法毎の再現率と適合率の加重平均を以下に示す式で求めた．

加重平均　＝　(再現率×α) + (適合率×β)

再現率重視ということで，α を 0.9 に，β を 0.1 に設定した評価表現での結果を図 4 に示す．

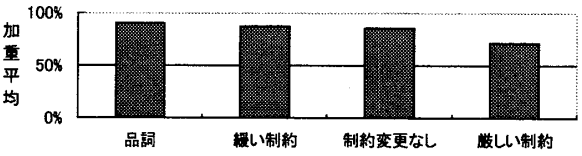

図 4 評価表現の加重平均

図 4 から再現率重視の除外手法では，品詞情報の利用が最も効果があった．このため，品詞情報は，除外手法の基礎になると考えられる．

5．　まとめ

品詞情報および助詞を利用した構文パタンで不要語を除外することで，高い再現率を得ることができた．今後は，抽出精度を重視した手法と本提案手法を組み合わせて高い再現率を維持しつつ不要語を除外する手法を検討する．

参考文献

[1]石井哲 :テキストマイニング活用法　顧客志向経営を実現する ,リックテレコム,263pp.,2002
[2]乾孝司 ,奥村学 :テキストを対象とした評価情報の分析に関する研究動向,自然言語処理, Vol.13,No3,pp.201-241,2006
[3]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
[4]池野篤司,濱口佳孝,山本英子,井佐原均 :Web 情報抽出のための専門用語獲得,言語処理学会 第 12 回年次大会,2006