5ZB-4

Weblogから共起関係を利用して評判情報を把握する手法の提案

鈴木健之† 丸山広† 中村太一†
東京工科大学†

1 まえがき

製品やサービスを開発する企業にとって顧客の評判を把握するために, Web 上の情報を活用することが，重要となってきている [1].

Web 上の文章を収集し，調査したい製品やサービスの名称（以下，対象名と記す）とその対象の評価を表す語の辞書を用いて評判情報を抽出する研究がある[2]．この方法では，辞書を人手で構築するため膨大な時間を要する．

辞書構築のコストを削減するために，共起関係を利用して仕様や評価を表す語を半自動で抽出し，辞書を構築する研究がある[3]．しかし，この方法では，人手による辞書登録作業が必要である．

本研究は，この問題に対処するために，自動で辞書を構築する手法を提案する．だが，一般的に自動構築した辞書は，不要な語を多く含む．

そこで，“対象名と仕様を表す語”，“対象名と評価を表す語”，“仕様を表す語と評価を表す語”の共起関係を利用し，対象名の類義語と仕様を表す語を Weblog から交互に抽出し，対象辞書と仕様辞書に登録することで，辞書の網羅性を高め，構築した対象辞書と仕様辞書を用いて，評価を表す語を Weblog から抽出し，評価辞書に登録することで，不要な語を含まない高精度の辞書を自動で構築することを目指す．

2 辞書の自動構築

本研究は，対象名の付属品，機能，性能，及び価格に関する語を対象名の仕様として定義し，仕様を表す語の必要性，評価，要求，不満，及び評判を表す語を評価であると定義する．

定義した仕様を表す語を登録した仕様辞書と評価を表す語を登録した評価辞書を，“対象名と仕様を表す語”，“対象名と評価を表す語”，“仕様を表す語と評価を表す語”の共起関係を利用し，不要な語を含まない高精度の辞書を，以下の手順で Weblog から自動的に構築する．

i. Weblog の情報を形態素解析した結果から，”【対象名】+の+【仕様を表す語の候補】“の共起パターンを抽出し，更に【仕様を表す語の候補】語を抽出する．
ii. 抽出した候補語のうち “名詞- 一般”，“未知語”, 及び“複合名詞”を仕様辞書に登録する．
iii. 同じ形態素解析した結果から，“【対象名の類義語の候補】+の+【仕様辞書の語】”の共起パターンを抽出し，更に【対象名の類義語の候補】語を抽出する．
iv. 抽出した候補語のうち“複合名詞”, “名詞- 一般”，“名詞-形容動詞語幹”，“名詞-サ変接続”及び“未知語”を対象辞書に登録する．
v. 同じ形態素解析した結果から，“【対象辞書の語】+の+【仕様を表す語の候補】”の共起パターンを抽出し，更に【仕様を表す語の候補】語を抽出し，候補語のうち “ii” の手順と同じ品詞の語を，新たに仕様辞書に登録する．
vi. 同じ形態素解析した結果から，“【対象名の類義語の候補】+の+【仕様辞書の語】”の共起パターンを抽出し，更に【対象名の類義語の候補】語を抽出し，候補語のうち “iv” の手順と同じ品詞の語を新たに対象辞書に登録する．
vii. “v” から “vi” の手順を繰り返し，新たな共起パターンを抽出しなくなるまで，対象辞書と仕様辞書を交互に拡張する．
viii. 同じ形態素解析の結果から,“【対象辞書の語, 及び仕様辞書の語】+(が|は|も|に|を)+【評価を表す語の候補】”の共起パターン，及び“【評価を表す語の候補】+【仕様辞書の語】”の共起パターンを抽出し，更に【評価を表す語の候補】語を抽出する．
ix. 抽出した候補語のうち “形容詞-自立”,” 名詞-形容動詞語 “,” 接続品詞”,” 動詞-自立 “,” 名詞- 一般 “，及び” 複合名詞” を評価辞書に登録する．

以上の手順の中で，抽出した評価を表す語の前後に “接頭語”，“接尾語” がある場合は，評価を表す語に “接頭語”，“接尾語” を結合させる．

また，仕様辞書に登録する毎に，同じ形態素解析した結果の語尾に” 円” と” 機能” が付く語を仕様の候補語として抽出し，候補語のうち “ii” の手順と同じ品詞の語を仕様辞書に登録する．

同様に，評価辞書に登録する毎に，同じ形態素解析した結果の語尾に “感じ”，“的” が付く語を評価の候補語として抽出し，候補語のうち “形容詞-自立”，“名詞-形容動詞語幹”，“接続品詞”，

Methodology for Extracting opinion information using co-occurrence pattern from the Weblog.
†Kenshi SUZUKI †Hiroshi MARUYAMA
†Taichi NAKAMURA; Tokyo University of Technology

及び“動詞-自立”を評価辞書に登録する．

仕様辞書に登録する候補語に，対象名と同じ語がある場合は，この語を削除する．

3 評価実験

3．1実験方法

本研究は，生成した辞書に登録されている語のうち，被験者が１人でも確からしいと判定した語を辞書の精度として定義する．生成した辞書の精度の式を以下に示す．

$$\text{辞書の精度} = \frac{\text{被験者が，1人でも確からしいと判 断した語数}}{\text{辞書に登録されている 語の総数}}$$

提案する手法を基に，対象名を“PS3”に限定し，“PS3”に関する“はてなダイアリー”からシステムが生成した仕様辞書と評価辞書に含まれる語を，被験者が判定することにより本手法の有効性を検証する．

また，被験者毎に１人だけが確からしいと判定した語数の標準偏差を算出し，実験結果が正確であるかを検証する．

被験者は，本学の学生 19 名，大学院生１名，教員 1 名である．評価実験の事前準備として，“PS3”を被験者が，知らないことがないように，“Wikipedia”の“PS3”の項目を被験者に提示した．

被験者は，“対象名の類義語”，“対象名の仕様を表す語”，“仕様の評価を表す語”の説明を受けてから，仕様辞書，評価辞書に順で辞書に含まれる語を判定した．仕様辞書から判定することにより，対象名の仕様を表す表現を理解して，評価を表す表現を判定することができる．

3．2 実験結果

提案手法を基に生成した仕様辞書には 256 語，評価辞書には 768 語が登録された．被験者が１人でも確からしいと判定した語は，仕様辞書では，234 語，評価辞書では 594 語である．

そのため，仕様辞書の精度は 91.4%であり，評価辞書の精度は，77.3%である．

被験者毎に，その被験者だけが確からしいと判定した仕様を表す語と評価を表す語の語数を図１に示す．

図１から，14 人目の被験者は，他の被験者よりも評価を表す語を多く選択している．また，仕様，及び評価辞書に登録されている語から，被験者のうち１人だけが確からしいと判定した語の標準偏差を算出したところ，仕様辞書の標準偏差は 1.4 であり，評価辞書の標準偏差は 11.8 である．

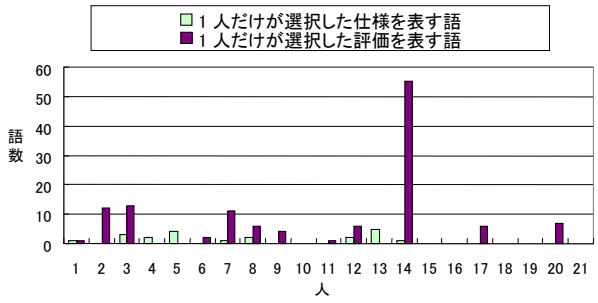

図１　被験者毎の結果

14 人目の被験者だけが選択した語は 55 語あり，この語数は他の被験者のみが選択した語数に比べ極めて多く，特異であるといえる．そこで，この被験者の評価結果を除外し，算出した評価辞書の精度は 70.2%である．

評価辞書に登録されている語から，被験者のうち１人だけが確からしいと判定した語の標準偏差は 4.3 である．

4． 考察

提案した辞書構築方法により Weblog から自動的に高精度の仕様辞書を生成することが可能となった．

他方，評価辞書の精度は，仕様辞書に比べ約 20%低い．これは，被験者が，仕様を表す語に比べ評価を表す語を，幅広く捉えたため，発生したと考えられる．

5 まとめ

本研究では，共起関係を利用し，Weblog から高精度の仕様辞書と評価辞書を自動で構築する手法を提案し，生成された辞書に含まれる語を，人により判定し，有効性を示した．

今後は，誰も確からしいと判定しなかった語を削除することにより，構築する辞書の精度をより向上させ， 構築した辞書を基に対象名の評判情報を分析することを目指す．

参考文献

[1]産業構造審議会 新成長政策部会 中間報告:創造的産業組織の構築

[2]立石健二,石黒義英,福島俊一 ： インターネットからの評判情報検索,情報処理学会研究報告,NL-144-11, pp.75-82 ,2001

[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