二重伝播を用いた日本語感情極性辞書の自動作成

-製品レビューを事例として-

2018SE01 後藤　駿

指導教員：張　漢明

1　はじめに

インターネットの発達により，多くの消費者は評判を元に意思決定をする．それに伴い大量のテキストから製品やサービスの評判を分析する評判情報分析は重要な研究分野である．感情極性分析は意見がポジティブ，ネガティブのどちらを示すか分類する分析手法である．感情極性分析に利用される極性辞書は評価語と呼ばれる極性を持つ単語を収集したものである．収集するべき評価語は対象領域によって異なるため，特定の領域に特化した極性辞書が必要である．

本研究の目的は製品領域に特化した日本語極性辞書を自動的に作成することである．異なる製品領域に対する極性辞書を自動的に作成し評価を行う．

2　先行研究と既存技術

高村ら [1] は語彙ネットワークとスピンモデルを利用して英語と日本語の単語極性辞書を自動的に作成した．しかし，高村らの極性辞書は対象とする領域によって起こる極性の変化が考慮されていない．

金山ら [2] は統計的手法を利用して，特定の領域に特化した日本語の感情極性辞書の自動作成手法を提案した．金山らの手法では大規模なコーパスを必要とする．

Qiu ら [3] は大規模なコーパスを必要としない二重伝播による製品領域に特化した英語の単語極性辞書の自動作成手法を提案した．二重伝播では評価語に加え，製品を表すような属性表現を収集している．

3　極性辞書自動作成の方針

評価語と属性表現の二重伝播 [3] を用いて製品領域に対応した日本語極性辞書を自動的に作成する．二重伝播手法を採用した理由は，大規模なコーパスを必要とせず，新語に対しても有効であると考えたためである．

Qiu らのシステムは英語で構築されているため日本語に適応する必要がある．日本語へ適応する方針は大きく 2 点ある．

- 日本語の否定表現による極性の反転を行う．これは二重伝播を再現するために必要な処理である．
- 日本語のストップワードによる単語の除外を行う．頻出単語による誤った極性の伝播を防ぐことで，極性判定の精度が向上すると考えている．

4　システムの設計

4.1　システムの概要

極性を持つと予想される評価語と製品の特徴を表すと予想される属性表現を自動的に収集する．評価語は形容詞，属性表現は名詞に限定する．いくつかの評価語を初期極性辞書として用意する．ここで用意する単語は，領域による極性変化のない単語を選ぶ必要がある．例えば”良い”,”悪い”などである．属性表現の初期単語数は 0 である．

図 1 はシステム構成を表した図である．

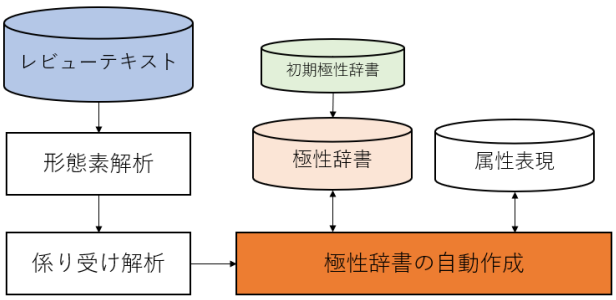

図 1　構成図

提案手法では二重伝播による極性付与と文脈による極性付与がある．

4.2　前処理

伝播に利用する係り受け関係の特定には，形態素解析と係り受け解析が必要となる．本研究では形態素解析器に MeCab，係り受け解析器に CaboCha を採用する．

4.3　二重伝播による極性付与

評価語と属性表現の 2 種類を伝播させることで極性の付与と新たな単語の獲得を同時に行う．伝播は次の 2 つの条件を満たす場合に行う．

1. 評価語または属性表現の 2 単語が文法的な係り受け関係にある
2. 片方が収集済み，もう一方が未収集の単語である

伝播が行われた未収集の単語は極性辞書または属性表現として新たに収集され，伝播元の極性を継承する．評価語は一貫した極性で使用されると考えられるが，人によって製品に対する意見は異なるはずである．そのため属性表現はレビューごとに極性のリセットを行う．

評価語の周囲に存在する否定表現の回数，極性の反転が行われる．

出現頻度が高く，誤った極性の伝播が行われると予想される単語はストップワードとして除外する．

4.4　文脈による極性の付与

極性のリセットが行われた属性表現から伝播を受けた評価語には極性が付与されない．このような単語は文内に存在する他の評価語がもつ極性値の合計から極性を決定する．

5　実験と結果

5.1　実験

提案手法により各製品領域に対する極性辞書を作成する．環境構築には Python を利用した．初期極性辞書は 4 単語,10 単語,20 単語の 3 種類で実験を行った．

5.2　結果

作成した極性辞書に含まれる評価語が，対応する製品領域において正しい極性を持つかどうか主観的な評価を行った．図 2 は異なる初期単語数における各製品領域に対する極性辞書の精度を比較したものである．図 3 は極性の反転とストップワードの使用による精度の変化を表したグラフである．

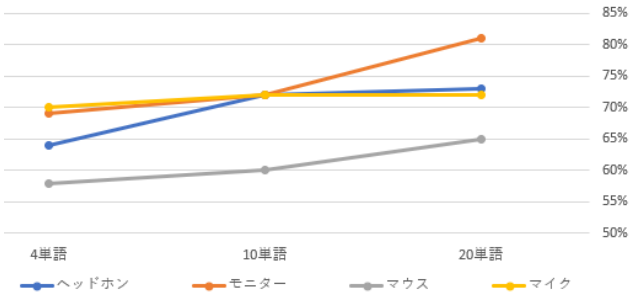

図 2　異なる初期単語数での精度比較グラフ

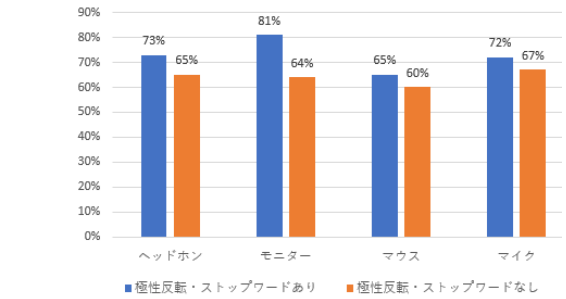

図 3　極性反転・ストップワードの有無を比較したグラフ

5.3　評価

対象とした製品領域によってばらつきはあるが初期単語数が増えると精度も高くなり，初期単語数を 20 とした場合に最も高い精度を得た．極性反転とストップワードはどの製品領域に対しても精度の向上が確認できた．

6　考察

二重伝播により製品領域に対応した評価語を収集することができた．例えばモニター領域では”明るい (+)”,”眩しい (-)”や，ヘッドホン領域では”うるさい (-)”などの単語を自動的に収集することができた．また”ダサい”,”しょぼい”といった口語的な表現も収集することができた．また，極性反転とストップワードはどの製品領域でも精度向上に寄与した．これらの結果は提案手法が製品領域に対する極性辞書の自動作成問題に有効であったといえる．課題点を以下に示す．

- 伝播による極性付与は機能していた一方，文脈による極性の推定が上手く機能しなかった (図 4)．想定よりも文脈の極性と登場する評価語の極性は一致しなかったことが大きな原因である．例えば”痒い所に手が届く”はポジティブな文脈で使われるため,”痒い”はポジティブな単語として収集された．このように句になることで極性の変化する単語に対して正しい極性が判定できなかった．また，一文内を対象として文脈による極性判定を行ったが，判定される単語数は少なくなってしまった．

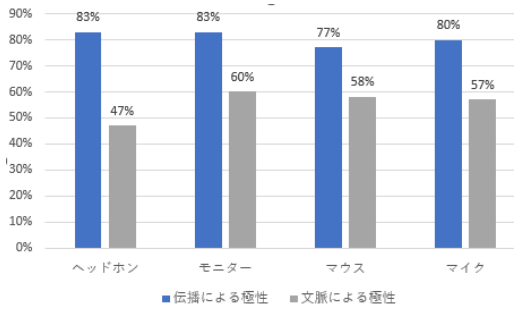

図 4　極性付与手法の精度比較グラフ

- 利用したレビューデータのノイズによって極性が正しく判定できないケースがあった．これは読点のない箇条書きや文末が記号で構成されるような一文の区切りが曖昧なレビューに対して，係り受け解析が正確に行われなかったことが原因だと考えられる．
- 形態素解析器に登録されていない単語に弱く，新語が思ったほど収集されなかった．例えば,”エモい”,”テクい”のような表現は形態素解析器に認識されず辞書に追加されることはなかった．

7　おわりに

本研究では，二重伝播を用いた日本語単語極性辞書の自動拡張手法を検討した．提案システムでは二重伝播における，日本語の否定表現による極性反転，伝播の妨げになるストップワードの除外を行い極性判定に対しての有効性を示した．一方，文脈による極性判定では多くの課題が残った．句になることで極性の変化する単語に対しても正しい極性を判定できるような手法が必要である．

参考文献

[1] 高村大也, 乾孝司, 奥村学”スピンモデルによる単語の感情極性抽出”, 情報処理学会論文誌ジャーナル, Vol.47 No.02 pp. 627–637, 2006.

[2] Hiroshi Kanayama and Tetsuya Nasukawa. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In Proc. ofEMNLP’ 06, pp. 355-363, 2006.

[3] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding Domain Sentiment Lexicon through Double Propagation. In Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), pp. 1199–1204, 2009.