中国語感情極性辞書補完手法の検討

楊　鵬　　松本 忠博

岐阜大学　大学院　工学研究科

1. はじめに

インターネットの普及や IT 技術の進歩により，ユーザーが膨大な情報量を入手する事が可能になっている．我々もインターネット上にテキストを投稿する機会が増えてきている．しかし，そのようなテキスト内の感情を発見することや特定することに対しては，様々な場面の活用に重要な課題がある[4]．例えば，ある会社で発売の製品に関する評価を，インターネット上でのコメントを通じて行うことが可能になれば，その後の対策や需要に製品の開発に役立つと考えられる．

音声による言葉や表情に比べ，文書中では感情が直接表現されていることよりも，間接的に表現されている事が多い．そのため，感情を読み取る事が直接表現よりも困難で，複雑である[1]．感情極性値を特定することは，テキストに含まれる感情の判断に役に立つと考える．

現在公開されている中国語の感情極性辞書に NTUSD がある．収録単語がポジティブとネガティブに分類されているが，その程度については数値化されていない．例えば，优秀（優秀）と良好（良好）は同じポジティブの単語であるものの，表現する感情極性の程度には一定の差が存在する．このような差を表すため，感情極性値の導入が課題になる．感情極性値は、語彙ネットワークを利用するなどして自動的に計算されたものであり，もともと二値属性だが，−1から+1の実数値を割り当て，−1に近いほどネガティブ，+1に近いほどポジティブと考える[3]．

本研究では，強い感情極性を持つ単語を「標準単語」として設定し，その単語との類似度から目的の単語の極性値を算出し，中国語感情極性辞書を補完することで，中国語テキストの極性判定の精度向上を図る．単語間類似度の算出には Word2Vec を用いた．

2. 標準単語の設定

中国語によく使われる感情極性を極端に表現する単語を標準単語として設定する．

2.1 設定方法

標準単語設定方法は文献[6]を参考にし，中国語感情極性辞書 NTUSD を使用する．NTUSD は，台湾大学で編集された中国語単語の感情極性辞書である．辞書には 11,086 の単語が含まれ、その中には、ポジティブの 単語 2,810 個とネガティブの 単語 8,276 個が含まれる．データはテキスト形式であり，無料で利用可能な感情極性辞書である．辞書中の単語を対象に，30 件のアンケートを実施，その内から感情極性を極端に表現する単語を標準単語として設定した．

アンケートの実施方法は次のとおりである．NTUSDに含まる 11,086 の単語から極端な感情極性を持つ可能性のある単語 307 個を標準単語候補とする．アンケートの回答者は異なる年齢層，性別を含む中国語の母語話者である．回答者は標準単語候補に 1（ネガティブ）から 5（ポジティブ）の値を付け，結果をまとめ，点数 140 以上の候補単語 63 個をポジティブ標準単語，点数 40 以下の候補単語 56 個をネガティブ標準単語とした．

2.2 標準単語設定の例

標準単語の例を以下に示す．

ポジティブ標準単語の例

圣洁（神聖で清潔な）
上好的（上等な）
至高（最高）
壮丽雄伟（壮麗雄大）
完美（完璧な）
…

感情極性値：1

ネガティブ標準単語の例

污秽（不潔な）
变态（変態）
丑恶（醜悪な）
肮脏（汚れた）
荒谬（でたらめ）
…

感情極性値：-1

3. 極性曖昧な単語の感情極性重みの計算

極性曖昧な単語とは，ポジティブやネガティブな感情極性を表現できるが，表わされる感情極性の程度が標準単語より極端とは言えない単語である．これらはテキストの中で，感情極性の計算に一定の影響を与えるが(例:不错(悪くない))，もし，標準単語と同じ重みとして計算すると，判断結果の精度が下ることが想像できる．そのために，標準単語を設定した上で，極性曖昧な単語の重みを確定することが必要と考える．

3.1 計算方法

設定された標準単語を基に，極性曖昧な単語と標準単語との類似度を算出する．本研究では，文献[2]を参考に，word2vec を利用して類似度を求める．具体的には python gensim を用い，Wikipedia 中国語版を訓練コーパスとして word2vec のモデルを構築する．得られたモデルは標準単語と目標単語との類似度算出に使用し，目標単語の感情極性重みを定める．

3.2 目標単語の感情極性値の設定

テキストを感情極性分析する時，極性値未知の単語が現れた場合は，命令 model.similarity により，設定された標準単語と極性値未知の目標単語の類似度を求め，標準単語の感情極性値:1×類似度値＝目標単語感情極性値とする．なお，標準単語と目標単語は種類が同じである必要がある．

標準単語:完美（完璧な） 目標単語:完好（完全な）

入力： model.similarity(u"完美",u"完好")
出力：0.20424200324437293

標準単語と目標単語の類似度により，標準単語の感情極性重みを 1 とすると，目標単語の感情極性重みは 0.204 となる．

3.3 標準単語から類似度近い単語の感情極性値の特定

3.2 の方法で極性値未知の単語の極性値を取得することができるが，辞書を補足するために，効率的な極性値算出法も必要である．命令 model.most_similar により得られた標準単語の周りの類似度近い単語も一度に取得する．

標準単語：丑恶（醜悪な）

入力：model.most_similar(u"丑恶")
出力：[(u'怪异', 0.5321423411369324), (u'蛮横无理', 0.5272585153579712), (u'表露出', 0.5247347950935364), (u'放不下心', 0.5239706039428711), (u'馊主意', 0.5227030515670776), (u'乖僻', 0.5165132284164429), (u'害臊', 0.5157655477523804), (u'大刺刺', 0.5118492245674133), (u'尖酸', 0.5116422772407532), (u'捉摸不定', 0.5108243227005005)]

出力の結果から感情極性表現でない単語の削除が必要である．

今回は分析実験の実行のため，2,316 単語の感情極性重みを辞書に登録した．

4. テキスト感情極性分析の実験

本研究で作成した感情極性辞書が中国語テキスト感情極性分析にどの程度貢献するかを検証するために分析実験を行った．

感情極性分析実験の流れは次のとおりである．中国語テキストは Web から収集し，形態素解析システムにより副詞や形容詞などの品詞を抽出する．その後，作成した中国語感情極性辞書により極性を調べ，結果を判断する．

4.1 形態素解析

本研究では，中国語形態素解析システム ICTCLA を利用した．ICTCLA2011 は，Java・C・C#・Python などのプログラミング言語から利用でき

る．テキストを入力し解析させると、単語（形態素）ごとに句切り，POS タグをつけた結果が返される．品詞分類は，中国科学院計算技术研究所の《计算所汉语词性标记集》に従っており（ICTCLA2011 にも「ICTPOS 汉语词性标记集.doc」のファイル名で収録），合計 99 項目が 22 の大分類に収められている．

4.2 感情極性の計算

テキストの極性計算は形態素解析の結果を使用し，感情極性に影響を与える副詞や形容詞などの品詞の単語をパラメーターとして計算する．処理の流れを図 1 に示す．

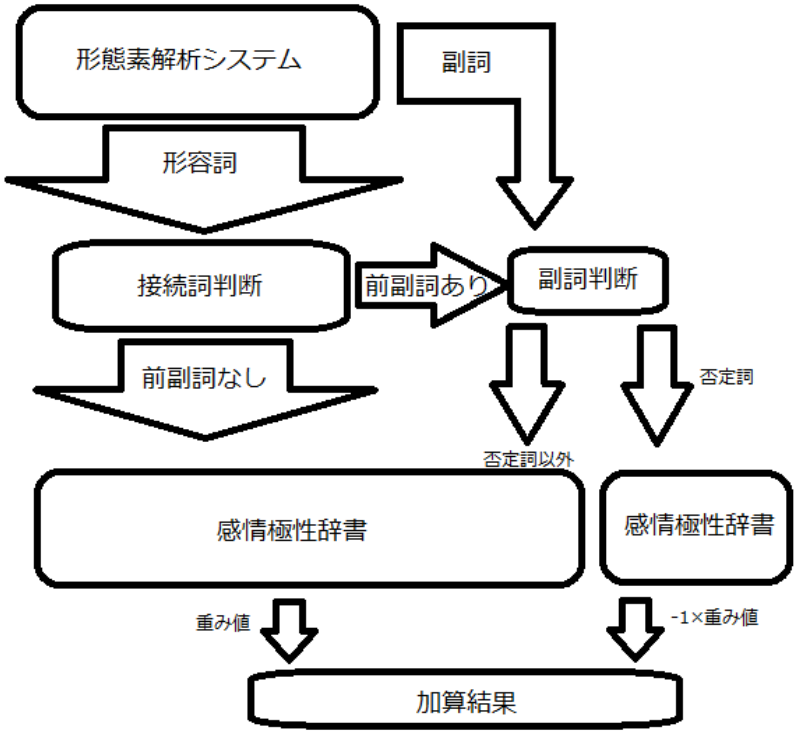

図 1．感情極性計算の流れ

中国語には，形容詞の前に副詞あることが多いため，副詞が否定詞の場合は，形容詞を表現する感情極性が逆転する．

4.3 実験と評価

Web 上にある中国語ニュースや商品レビュー等から抽出した感情を含む 100 例文を対象に実験を行い，その結果を評価する．表 1 は実験の結果である．具体例を表 2 に示す．

表 1 実験の結果

テキスト別	抽出数	正解数
ニュース・記事	50	39（78.0%）
レビュー	50	45（90.0%）

表 2 実験の例

番号	例文	極性値	評価
1	15 级风能吹成这么严重样子，想像一下，要是是 18 至 20 级不知道什么样子了。好可怕的风哦。	-0.751 negative	正解
2	许多关卡中，都出现了城墙、峭壁和高大的建筑物，但在	-1.676	正解

	玩家靠近时，这些 3D 物体却不能变为半透明的显示模式，在这种情况下，你只能按住左 Alt 键，同时晃动鼠标，以便切换到更容易观察的视角—这种做法实际相当蹩脚。更糟糕的是，视角问题可能导致角色做出错误的行动，甚至被警卫发现，这将直接给游戏体验带来负面影响	negative	
3	电池重量达到 9kg，拎起来上楼对于妹子来说还是比较重的。对于电车大家都比较关心续航里程，iTank 的续航能力还算不错，在 20km/h 的速度下，它的续航里程可以达到 100km，但是大家都知道这是理想状态。	0.492 positive	不 正 解 (negative)
4	美丽的月亮给大地撒下一片银辉，温柔的月光如同水一般平静，散落在人们的脸上。美丽的天空好似一张蓝色的地毯上，镶嵌着无数亮晶晶的“小钻石”美丽无比。	2.835 positive	正解
5	质地呈乳白色乳液状，流动性弱。仔细看可发现乳液外圈包着一层精油，待推开之后，质地又幻化成啫喱状，轻轻按压至完全吸收，感觉肌肤非常细滑柔软。说到这个质地，还有有一个非常厉害的技术在里面，将精华与精油按 6:4 的比例，糅合啫喱的清透质地，一触即融，极易吸收。	3.189 positive	正解

4.4 問題点の分析

例文 3 については，単語が主な感情極性を表すわけではなく，客観的なデータより常識の判断から感情極性を表現するテキストである．

Web 上にある中国語の単語は膨大な量が存在するため，コーパスとする Wikipedia 中国語版の内容では不十分と考えられる．この原因で，計算された一部の極性値と常識上表す感情極性に大きな差があることが明らかになった．標準単語の設定においてアンケートの収集を拡大すると，設定の精度が上昇すると考えている．

5. 今後の課題

本研究では，提案手法により中国語感情極性辞書の形容詞極性値を登録し，評価実験を行った．今後，感情極性の計算や辞書の登録を続行する．

テキスト感情極性分析手法については，感情極性単語の出現率の計算も導入し，分析精度の向上を目指したいと考えている．

動詞や名詞などの品詞もテキストの感情極性に影響を与えると考えられるので，動詞や名詞などの品詞の感情極性値特定方法を検討することも今後の課題と考えている．

参考文献

[1] Kim SM, Hovy E., “Automatic detection of opinion bearing words and sentences,” In Carbonell JG, Siekmann J, eds. Proc. of the IJCNLP 2005. Morristown: ACL, 2005.

[2] Radim Řehůřek, “Making sense of word2vec,” https://rare-technologies.com/making-sense-of-word2vec/, 2014

[3] 高村 大也「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.

[4] 赵 妍妍「文本情感分析」，哈尔滨工业大学，2010

[5] 志村 千尋「Web ニュースデータを用いた感情極性値推定」 法政大学，2016

[6] 陈 晓东「基于情感词典的中文微博情感 倾向分析研究」华中科技大学，2012