コーパスデータからの単語の感性イメージの獲得手法
Kansei image learning method for words based on text corpus

山下　正雄　　目良　和也　　黒澤　義明
Masao YAMASHITA　Kazuya MERA　　Yoshiaki KUROSAWA

広島市立大学　情報科学部
Faculty of Information Sciences, Hiroshima City University

A lot of words have their Kansei images and they help to arouse the emotions and the feelings from text data such as novels, blogs, and so on. In this paper, we propose a method to extract common images of words from corpus data and express the images by using multidimensional vector. The Kansei images of the words are extracted from text corpus considering the grammatical features and they are classified into various axes based on the meaning of the adjectives.

1 はじめに

個人サイトや Blog 普及により，現在非常に大量かつ多様なテキストが Web 上に存在している. 人間は, 個々の単語が持つ感性イメージを用いることで, このようなテキストからでも感性情報を読み取り処理することができる.

近年，こうした単語の感性イメージを Blog 等から収集することが試みられている[1]．しかし，そこで扱われている感性イメージは，あくまで対象語のイメージの一面に過ぎず，一般的なイメージを表現しているとは言いがたい．そこで名詞が持つ一般的なイメージを獲得するためには，大規模コーパスに基づいた多様な特徴記述を考慮する必要がある．これにより，比喩などの言語表現の検出[2]や画像・動画に対する感性検索[3]等に応用が可能となる．

そこで，本研究では感性イメージの多様性を表現するため, 単語が持つイメージの多次元化を試みる. 処理の手順としてはまず，コーパスから対象語と形容詞の対を得る．次に，得られた形容詞を感性語群対に振り分ける．そして，各感性語群対にポジティブ，ネガティブのラベル付けを行うことで，多次元化した対象語のイメージ情報を得る.

本稿は以下のような構成である．まず２節で関連研究について説明する．３節では対象語の多次元化手法について説明する．４節では，実験について述べる．最後に５節に結論を述べる.

2 関連研究

感性イメージを収集する研究として，矢野ら[1]は Web 文書を対象に, 評価語辞書や評価対象語辞書を用いずに評価文および評価情報の抽出を（対象，属性，評価，程度，信頼度）の 5 要素で行う手法を提案している．この手法では特定の対象物に関する属性情報が得られるが，それらの情報を統合した一般的な物概念の持つ属性情報については得ることが出来ない.

それに対し, 名詞が持つ一般性を利用する研究として，桝井ら[2]は，テキスト中に出現する比喩表現を認識するために，知識ベースを自動で構築する手法を提案している．知識ベースは，対象とするテキストコーパスを形態素解析し，得られた結果から，“修飾語－名詞”の共起関係とその共起頻度を抽出することで構築している．その結果, 単語ごとにその単語を表す様々なイメージ，つまり多次元で表現された，名詞が持つ一般性と，その生起確率を記述した知識ベースが構築されている.

また同様に，三浦・中川ら[3]は，ドラマ映像を用い，人物の動作記述とシーンの雰囲気という情報をシナリオ解析システムにより抽出し，シーン検索システムを試作している. この手法では，場面の雰囲気（ムード）を抽出するために，ドラマにおけるセリフの解析を行っている．セリフの解析では，30 対の感性語を作成し，セリフ一つ一つについて，複数人のユーザーにどの感性に当てはまるか選んでもらうことにより，セリフとムードの対応付けを行っている．その結果，セリフごとにそのセリフが表す様々なイメージ，つまり多次元で表現されたイメージが得られる.

3 対象語が持つイメージの多次元化

3.1 本研究の流れ

本研究では，対象語が持つ一般的なイメージを多次元表記するために，大規模なコーパスから，対象語と形容詞の対を得る．その際，得られた形容詞のガ格・ヨリ格に注目する．また，スパースネスを避けるため，類義する感性語を一つのグループとし，多次元を形成する軸として，対極的な概念を持つ対義語対を用いる．具体的には，作成

した各感性語は，その対義語対を軸とした多次元空間で表現される．さらに，軸となる各感性語群対にポジティブ，ネガティブのラベル付けを行うことにより，複数の対象語を比較できるようにする．

3.2 使用コーパス

本研究では使用するコーパスを Web 文書全般とする．2 節で紹介した桝井らの研究[2]では“修飾語－名詞”の共起関係から，多次元で表現された，名詞がもつ一般性を獲得している．そこで，本研究では使用コーパスからより大規模に対象語に対する一般的イメージを獲得するために，河原・黒橋ら[4]の格フレーム辞書を用いた．河原・黒橋らは，まず，コーパスを構文解析し，確信度の高い述語項構造のみを抽出・クラスタリングすることにより，1 次格フレーム辞書を得る．次に，2 次格フレーム辞書を用いてコーパスを格解析し，新たに分かる確実な情報を抽出することで、格フレーム辞書の自動構築を行っている．格フレーム辞書の例を表 1 に示す．

3.3 語に対する一般的イメージの獲得

対象語のイメージを多次元で表記するためには，対象語の持つ様々なイメージと各々のイメージがどれだけ強いかを示す特徴量が必要となる．そこで，本研究では，前述した格フレーム辞書を用いて得られた結果の，形容詞及び形容動詞のガ格及びヨリ格に注目することで，単語イメージを獲得することとした．その理由は，対象語のイメージを多く取り出せるであろうと推測されるからである．なお，イメージの表現に用いられる形容詞・形容動詞を本研究では感性語と呼ぶ．また，Web 文書中での感性語の出現回数も格フレーム辞書から得られるため，その出現回数を特徴量として扱うこととする．

表 1 ：格フレーム辞書を用いて，名詞「月」を検索した結果

格フレーム ID	格	頻度
見る:動 1	ヲ格	2107
見上げる:動 4	ヲ格	1055
なる:動 1	ニ格	988
なる:動 2	ト格	322
きれいだ:形 11	ガ格	318
みる:動 64	ヲ格	284
綺麗だ:形 8	ガ格	280
:	:	:

3.4 イメージの表現

対象語のイメージは，複数の属性概念を軸とする多次元ベクトルによって構成される．

抽出された感性語それぞれを一つの軸とみなした場合，対象語が持つイメージがスパースネスとなってしまうと考えられる．スパースネスを回避するため，本研究では類義する感性語を一つのグループとしてまとめて表現することとした．この時，まとめられたグループを感性語群とする．なお，感性語群は，類語例辞典[5]を用いて構築した．表 2 に感性語群「美麗」と「醜悪」についての例を示す．

表 2：感性語の分類例

感性語群「美麗」	感性語群「醜悪」
美麗，美しい，綺麗， 華やか，艶やか，etc	醜悪，醜い，汚い， 不細工，薄汚い，etc

また，三浦・中川ら[3]は，感性語対を作成し，単語イメージの評価を行っている．そこで本研究では，多次元を形成する軸として，対極的な概念を持つ対義語対を用いる．具体的には，各感性語は，その対義語対を軸とした多次元空間で表現される．ここで，対義語対となる感性語群を感性語群対と呼ぶ． 本研究においては８２対用意した．その例を表 3 に示す．“⇔”は両側にある感性語群が対義関係にあることを表す．

表 3：感性語群対（一部）

強⇔弱	洗練⇔荒削り
明⇔暗	動的⇔静的
上品⇔下品	和やか⇔とげとげしい
美麗⇔醜悪	ドライ⇔ウェット
かわいい⇔憎い	深い⇔浅い
鋭敏⇔鈍重	のどか⇔あわただしい
優しい⇔荒々しい	軽やか⇔重々しい
沈着⇔軽躁	日常的⇔非日常的
てっとりばやい⇔まわりくどい	親しみやすい⇔親しみにくい
能弁⇔訥弁	仲良し⇔対立

3.5 軸の極性

前述した感性語群対は，対となる感性語群のポジティブ，ネガティブのどちらに働くかは決められていない．軸には「善⇔悪」のように，極性が一定である軸もあれば，「多い⇔少ない」のように極性が一定でない軸もある．そこで，単語間の比較を行いやすくするため，軸にポジティブ，ネガティブを設定する．ここで，対象語により極性が一定でない軸を不定軸，どのような対象語がきても極性が一定である軸を固定軸と呼ぶ．本研究では，感性語群対の両方に出てきた形容詞，形容動詞がいずれも「程度」,「価格」,「数量」,「温度」,「速度」，五感を意味する「〇覚」の属性（IPAL 辞書[6]）を持つ場合に不定軸とし，それ以外を固定軸とする．不定軸を以下に示す．

強⇔弱、暖⇔冷、柔らかい⇔固い、深い⇔浅い、
熱い⇔寒い、細い⇔太い、長い⇔短い、広い⇔狭い、
厚い⇔薄い、大⇔小、多い⇔少ない、高い⇔低い、
軽い⇔重い、洗練⇔荒削り、緩い⇔厳しい、
速い⇔遅い、鋭敏⇔鈍重、赤い⇔青い、
賑やか⇔静か、白⇔黒

不定軸にポジティブ，ネガティブを設定する手法としては，まず感性語群対の各固定軸においてポジティブ，またはネガティブのどちらの極性の特徴量がより多いかを調べる．そしてポジティブ寄りの固定軸，ネガティブ寄りの固定軸の個数を比較することにより，対象語イメージのポジティブ，ネガティブ，0 の判定を行う．次に，各不定軸の両極性の特徴量を比較し，特徴量のより多い側の極性を対象物のポジティブ／ネガティブイメージと一致させる．固定軸の差が 0 の場合は，表 3 の左にある感性語群をポジティブ，右にある感性語群をネガティブとする．

4.実験

4.1 方法

あらかじめ「ポジティブ」，「ネガティブ」のイメージが人手で判断されている 80 個（ポジティブ：40，ネガティブ：40）の名詞を実験データとして評価実験を行う．本研究では，目良らが質問紙調査によって収集したデータを用いる [7]．目良の研究では「ネガティブ」の名詞が 71 個あったが，「ポジティブ」のデータ数と等しくするため，71 個の名詞の中から，ランダムに 40 個選んだ．

正解不正解の判断は，まず感性語群対の全ての軸においてポジティブ，またはネガティブのどちらの極性の特徴量がより多いかを調べる．そしてポジティブ寄りの軸，ネガティブ寄りの軸の個数を比較することにより，対象語イメージのポジティブ，ネガティブ，0 の判定を行った結果と実験データで割り振られたイメージとが一致するかにより行った．

図 1 に名詞「桜」の多次元化したイメージを示す．この図の感性語群対の左に“◎”が付いている感性語群対が不定軸である．名詞「桜」のイメージとしては，ポジティブ方向に伸びた“美麗⇔醜悪”軸の感性語群“美麗”が特に強い．感性語群“美麗”には“美麗”，“美しい”，“綺麗”などの感性語が含まれている．また，不定軸に関しては，名詞「桜」がポジティブと判断されているため，“多い”，“鋭敏”などの感性語群がポジティブに設定されている．

図 2 に名詞「注射」の多次元化したイメージを示す．この図を見ると，名詞「注射」のイメージとしては，ネガティブ方向に伸びた“好⇔嫌”軸の感性語群“嫌”が特に強い．感性語群“嫌”には“嫌”，“嫌い”，“忌まわしい”などの感性語が含まれている．また，不定軸に関しては，名詞「注射」がネガティブと判断されているため，“多い”，“固い”の感性語群がネガティブに設定されている．

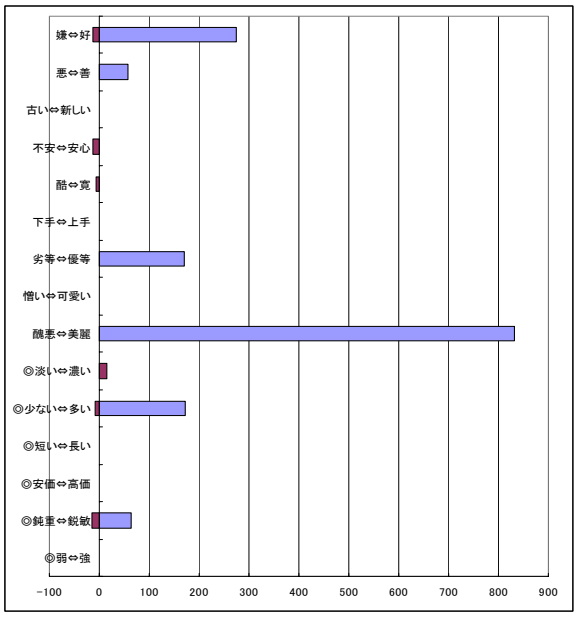

図 1：名詞「桜」の多次元化例

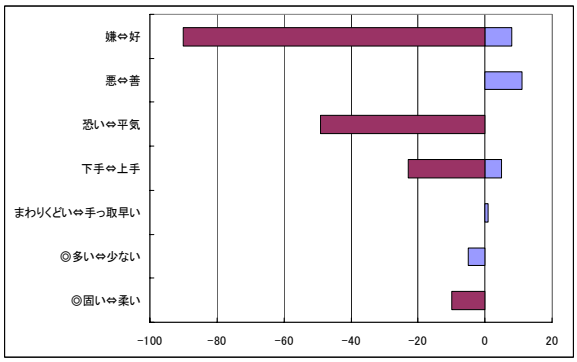

図 2：名詞「注射」の多次元化例

4.2 結果

表 4 に実験結果を示す．表 4 は，縦軸に実験データ，横軸に本手法の出力を示している．この表をみると，ネガティブの実験データに対して，ポジティブと誤って出力された失敗が多かった．これは，不定軸の「ポジティブ」，「ネガティブ」を設定する際に，固定軸で対象語のイメージが「0」と判断されてしまう場合，また，固定軸で対象語が「ポジティブ」と判断されてしまう場合に影響が出ると考えられる．

対象語のイメージが「0」と判断される場合，不定軸の「ポジティブ」，「ネガティブ」の設定を，表 3 の左にある感性語群を「ポジティブ」，右にある感性語群を「ネガティブ」としたことにより，正しい設定が行われていないと思われる．

一方「ポジティブ」と判断される場合，コーパス

から得られる感性語が，人間が思っているイメージと異なることが考えられる．例えば，名詞「不幸」に対して，“お前の不幸がすきなの”や“不幸っぷりがいい”などの原文から“好き”，“善”などポジティブな感性語が多くみられた．また，目良らは，対象語の「ポジティブ」，「ネガティブ」の判断を学生に対してのアンケートから導き出している．そのため，対象語が正しく一般性を表しているとはいえない．

また，本実験は各軸の特徴量を全く考慮していない．そのため，正しく対象語のイメージを捉えていないことが考えられる．

表 4：実験結果

	出力：ポジティブ	出力：0	出力：ネガティブ
データ：ポジティブ	37	0	3
データ：ネガティブ	13	0	27

その他の失敗として，感性語を感性語群に割り振る際にどの感性語群にも当てはまらず，対象語のイメージに反映されない感性語が存在した.図 2 で示した名詞「注射」の場合，感性語「痛い」がどの感性語群にも当てはまらなかった．改善案として，実験の結果どの感性語群にも当てはまらなかった感性語を再分類することで，感性語群を新たに追加することが挙げられる．

5 おわりに

本研究では，大規模コーパスから対象語と形容詞の対を獲得し，得られた形容詞のガ格・ヨリ格に注目することによりイメージの多次元化を行う手法を提案した．また，感性語群対を用いることや，ポジティブ･ネガティブのラベル付けを行うことにより，複数の対象語の比較を行いやすくした．

実験の結果, ポジティブの実験データに対しては，精度が 92.5%と信頼のおける結果となっている．しかし，ネガティブの実験データに対しては，精度が 67.5%と，まだまだ考慮することが必要である．原因としては，不定軸の「ポジティブ」，「ネガティブ」を設定する際に，固定軸で対象語のイメージが「0」と判断されてしまう場合，また，固定軸で対象語が「ポジティブ」と判断されてしまう場合に影響が出ると考えられる．

対象語のイメージが「0」と判断される場合については，不定軸の「ポジティブ」，「ネガティブ」の設定を，表 3 の左にある感性語群を「ポジティブ」，右にある感性語群を「ネガティブ」としたことにより，正しい設定が行われていないと思われる．この点では，対象語のイメージが「0」の場合に，ポジティブ／ネガティブ設定の考慮をする必要がある．一方「ポジティブ」と判断される場合，コーパスから得られる感性語が，人間が思っているイメージと異なることが考えられる．

また，ポジティブ／ネガティブの設定に関しては，現段階では軸の特徴量を考慮しておらず，考慮が必要であること．そして，対象語のイメージを表している感性語が，どの感性語群にも当てはまらない感性語をなくすことが必要である．

今後の課題としては，軸のポジティブ／ネガティブの設定の際に特徴量を考慮すること，対象語のイメージが「0」の場合の不定軸の設定，感性語群の改良が挙げられる．

参考文献

[1] 目良和也，矢野宏実，市村匠：“信頼度表現を考慮した評価文からの属性情報の抽出”，第 21 回ファジィシステムシンポジウム講演論文集，p.3，2002

[2] 桝井文人，福本涼一，椎野努，河合敦夫：“確率的判定尺度を用いた比喩性検出手法”，自然言語処理学会，vol9，No5，pp71-91，2002

[3] 三浦健二，中川祐志：“シナリオを用いたドラマのシーン検索システム”，情報処理学会論文誌, Vol40, No. SIG3 (TOD 1), pp144- 151,1999

[4] 河原大輔，黒橋禎夫：“格フレーム辞書の漸次的自動構築”，自然言語処理，Vol2，No2，pp109-131, 2005

[5] 遠藤織枝, 中川祐史, 三井照子, 村木新次郎, 吉沢靖：“使い方の分かる類語例辞典”，小学館，2001

[6] “計算機用日本語基本形容詞辞書 IPAL”, “計算機用日本語基本形容動詞辞書 IPAL”，1989

[7] 目良和也，市村匠，山下利之，三木睦明：“言語真理値を用いた情緒生起手法の連体修飾への応用”，日本知能情報ファジィ学会誌，Vol.15，No.4，pp.465-473，2003

問い合わせ先

〒731-3194
広島市安佐南区大塚東 3-4-1
広島市立大学情報科学部
黒澤 義明
TEL&FAX： 082-830-1581
Email: kurosawa@its.hiroshima-cu.ac.jp