6S-3

文書群からの単語のイメージを抽出する手法の提案

土橋　哲[†]　岸　義樹[‡]

[†]茨城大学大学院理工学研究科情報工学専攻　　[‡]茨城大学工学部情報工学科

1　はじめに

人間は言語処理において辞書のような明文化された情報以外に多数の知識を利用している．このため計算機による自然言語の意味理解にもこのような情報は不可欠である．

人間はある単語において社会的に共有されているイメージというものをもっていると考えられる．たとえば日本において「リンゴ」といえば「赤い」などが思い浮かぶ．このような情報のおかげで「リンゴのような顔」と聞いて「赤い顔」であることが推測できる．このイメージの利用も意味解析において不可欠であると考える．しかし，イメージは多くの場合論理的ではなく，また正しいとも限らない．そのため従来の知識では対応できない．

また，イメージというものは時代・地域など，集団によって様々であり静的なデータを元にすることは望ましくない．

本稿ではこのイメージを形容詞の形に限定しこれを抽出する手法を提案する．本手法は特定の静的な情報に依存しないよう，事前知識にはコーパスのような偏りのない多量の文章群のみを利用する．

2　本手法により出力される情報

本手法を用いたシステムの出力例を表 1 に示す．本手法ではある単語のイメージとして形容詞とそれを補助する名詞のペアが出力される．

本手法ではイメージを形容詞に限定している．よって本手法はすべてイメージと考えられる形容詞を出力する．たとえば．表 1 における男性の場合はイメージとして「荒々しい」という形容詞が出力されている．

また，同時にその形容詞の意味を補助するものとして名詞を 1 つ出力する．これは形容詞だけでは意味が通らない場合，たとえば日本のイメージとして「乏しい」が出力されても意味が分からないが,「資源が」という補助名詞があれば意味が通るからである．補助名詞は形容詞の前につくもの（表 1：女性=「姿は美しい」）と後につくもの（表 1：海=「青い空」）に分けられる．形容詞の前につく場合は助詞が，後につくものは終端記号（“+”）がそれぞれ名詞に付随する．

単語	形容詞（イメージ）	補助名詞
男性	荒々しい	なし
女性	美しい	姿は
海	青い	空+

表 1: 出力例

3　提案手法

本手法の流れは以下の通りである．4,5,6 について詳細を後述する．

1. 文書 DB を用意し，単語が含まれる文章群の収集
2. 係り受け解析器による各文章の解析
3. 単語と係り受け関係にあるすべての形容詞を取り出す.
4. 解析結果および文書 BD、検索エンジンを利用し、各種情報の抽出およびノイズの除去を行う
5. 各形容詞に対し、得られた情報を元にスコア付けを行う
6. 補助名詞の探査を行う
7. スコアの高い上位形容詞を出力する

3.1　処理 4．における処理内容

3.1.1　抽出する情報

抽出するデータについて説明する．

平均距離　形容詞と単語の距離を得る．距離とは係受関係より，形容詞と単語との階層差の絶対値を用いる．抽出された同じ形容詞の距離の平均が平均距離である．ただし，距離には一定の閾値をもうけこれを超えたものは加えない．

出現数　3. において取り出した形容詞において，同じ形容詞の合計を表す．ただし，上述の距離により弾かれたものはカウントしない．

A proposal of method for extracting a image of word from a great deal of sentence

Satoru Dobashi[†], Yoshiki Kishi[‡]
[†‡]Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

検索エンジンを利用したシンプソン係数　形容詞と単語をキーワードとし，検索エンジンの検索数を利用してシンプソン係数を求める．

文書群を用いたシンプソン係数　形容詞と単語をキーワードとし，利用している文書 DB の直接の検索数を利用してシンプソン係数を求める．

3.1.2　ノイズの除去

不必要な形容詞を除外する．

1 つは「出現割合の低い形容詞の除去」である．出現数/形容詞総数 に閾値をもうけ，これを満たない形容詞を除外する．

もう 1 つは「不要リストによる形容詞の除外」である．単体で意味をなさないもの（ex. ”無い ”，”っぽい ”），一般的すぎて利用できないものや多義的なもの（ex. ”多い ”，”よい ”，”高い ”）がある．このような形容詞は結果として適さないが本手法においてはスコアが高くなる傾向にある．そのため，これら形容詞については手作業でそのリストを作り除外する．

3.2　処理 5．における形容詞のスコアの算出

上述の情報を用いてスコアの算出を行う．スコアには検索エンジン・文書群を用いたシンプソン係数（以下，SSE，SDB）および出現数スコア AS，平均距離を用いたスコア DS の 4 つを用いる．

$$AS = \frac{\text{出現数}}{\text{総形容詞数}}, DS = 1 - \frac{\text{平均距離} - 1}{1 - \text{距離閾値}}$$

これらのスコアをすべて足し合わせたものが形容詞スコアとなる．ただし，各スコアがおおよそ均一になるようにそれぞれ重み付け（w, x, y, z）をする必要がある．

$$\text{形容詞スコア} = wSSE + xSDB + yAS + zDS$$

3.3　処理 6．における補助名詞の抽出

それぞれの形容詞の補助となる名詞の抽出を行う．抽出法は上述の形容詞の抽出と同様の処理を行う．結果もっともスコアの高い 1 つをその形容詞の補助名詞として出力する．候補として抽出する名詞は（1）助詞・助動詞を挟まず形容詞直下にある名詞.（2）「単語+名詞+形容詞」の順に並んでいる名詞.（3）「形容詞+名詞+単語」の順に並んでいる名詞．である．

4　手法の評価

4.1　内容

上記手法を実証するためのシステムを構築し，その評価のためのアンケートを実施した．

評価	完全	部分	中間
イメージ通り	117	61	0
どちらともいえない	3	18	19
イメージと異なる	2	10	7
意味不明	1	12	0

表 2: 実験結果

4.1.1　実装内容

文書群として国立国語研究所の「『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）」[2] の中から書籍・白書のデータを利用した．検索エンジンは YahooAPI を利用した．

4.1.2　評価手法

それぞれのイメージに対し「イメージ通り」「どちらともいえない」「イメージと異なる」「意味がおかしい」の 4 つを選んでもらう．これを 1 つに対して 3 人が人手で行う．

今回は単語 50 語に対してそれぞれ 5 個ずつイメージの抽出を行った．評価は 10 人で行い，各自が 75 個のイメージ（$50 \times 5 \times 3/10$）に対してこれを行った．

4.2　結果および考察

結果をまとめたものを表 2 に示す．3 人とも同じ評価のものを「完全」，2 人が同じ評価だったものを「部分」で表している．3 人全員が分かれた場合はその中間の人の評価をとり,「中間」と表した．

全員が「問題なし」としたイメージは 117 であり，全体の約 47%である．これに「部分」も加えると約 71%となる．この結果よりある程度はイメージにふさわしい形容詞がとれていると考えられる．ただし 3 人の評価の完全に一致しているのが約 49%であり，人間的な見方であっても画一的な判断ができない情報が多いことが伺える．

5　まとめ

本研究ではイメージとして形容詞を取り出す手法を提案し，アンケートを用いた評価実験を通してある程度のイメージが抽出できることを確認した．

参考文献

[1] 渡部広一，堀口敦史，河岡司：常識的感覚判断システムにおける名詞からの感覚想起手法，人工知能学会論文誌 Vol. 19 , No. 2 pp.73-82，2004

[2] コーパス整備計画 KOTONOHA
http://www.ninjal.ac.jp/kotonoha/