

色・形状情報を用いた比喩生成

小澤宏也, 岡本紘幸, 斎藤博昭
慶應義塾大学理工学部

1 はじめに

近年ではさまざまなワープロソフトが存在しているが、現時点ではまだ文章校正をする程度であり、文章の作成支援を行うにまでは至っていない。文章の作成支援の一つとして比喩の自動生成が可能となれば、ユーザが文章を作る際に表現を豊かにすることができる。

また現在比喩に関連した研究では、福本らにより「AのようなB」が比喩であるか例示であるかを統計的に判断する研究[3]が行われているが、比喩の生成に関する研究は行われていない。

比喩には直喩、隠喩、換喩、提喩などの種類があり[2]、そのすべてを1つの手法でカバーすることは不可能である。したがって本研究では、比喩の中でも物の持つ色や形状から連想できる直喩表現¹に範囲を限定し、比喩の生成を行う手法を提案した。

2 提案手法

「AのようなB」といった比喩を生成する際に、Aから、またBから連想されるイメージや一部の属性が近ければ、被修飾語Bに対する修飾語としてAが適しているのではないかと考えた。

そこで本研究では、まず被修飾語と色と形状が同じ名詞を修飾語の候補とし、さらに候補と被修飾語が持つイメージ同士を比較して、イメージが近いものほど比喩として適する、という手法をとった。

そのためにはイメージや色、形状の情報が記述された辞書が必要である。イメージ情報を取得するために石崎らによる連想概念辞書[1]を、また、連想概念辞書だけでは物の色と形状を取得するには足りないため、物の色と形状を記述した辞書を手作業により作成した。

これら連想概念辞書、色・形状辞書を用いて色や形状、イメージから妥当な比喩表現を生成することを目的としたシステムを作成した²。

3 辞書の概要

3.1 連想概念辞書

連想概念辞書とは、石崎らによって連想実験の結果作成されたものである。

ある語（刺激語）と七つの課題に対して連想する語（連想語）が記述されており、それぞれをノードとした意味ネットワーク構造を成している。またそれぞれの対に連想時間、連想順位、連想頻度、連想距離の4つの数値が付与されている。

連想概念辞書は次のように記述されている。

あご, 3, 上顎, 0.200, 4.000, 0.100, 9.180

あご, 4, 弱い, 0.475, 4.000, 0.200, 5.130

...

左から順に刺激語、課題、連想語、連想時間、連想順位、連想頻度、連想距離となっており、本研究では語から連想されるイメージの情報を利用するため、課題4番：属性概念を、またそのイメージの強度として連想距離を使用している。

3.2 色・形状辞書

刺激語に対して連想する色と形状を、459語について記述した。

くちばし 黄色 円錐

大根 白 円錐

かんづめ 銀 円柱

...

作成に当たり、色や形状が連想できないもの、また色や形状が複雑すぎて単純化できないものは不適当とし、リストから削除した。

4 システムの概要

システムの処理の流れを、図1に示す。

¹りんごのような頬、石のような拳といったもの。

²例：「白い温泉」から「牛乳のような温泉」を生成。

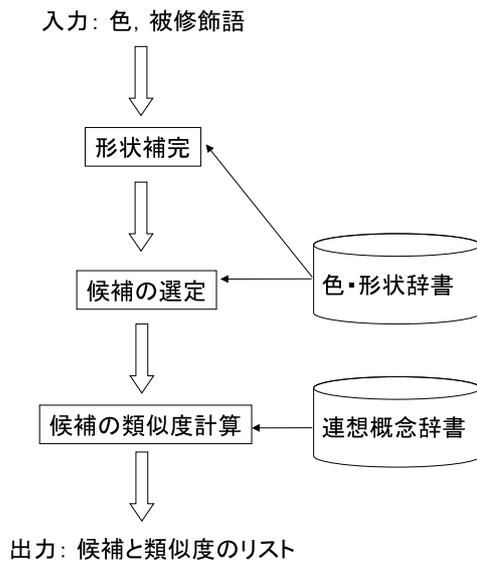


図 1: システムの全体図

4.1 入力

色, 被修飾語の順番に入力する. 入力として受け付けるのは色と被修飾語のみで, 形状情報は候補選定の際に内部情報として用いるのみである.

例として「白い泡」から「~のような泡」という比喻表現を求めるために, (白, 泡) と入力したとする.

4.2 形状補完

色・形状辞書を用い, 入力から受け取った色と被修飾語の対に, 被修飾語の持つ形状情報を付加する.

(白, 泡) の例では, 泡の形状が「球」であることを取得し, (白, 球, 泡) となる.

4.3 候補の選定

色・形状辞書を参照し, 色と形状が一致した項目を修飾語としてリストアップする.

(白, 球, 泡) から出力された候補の例を次に示す.

ご飯粒, 牡蠣, 玉ねぎ, 卵, 饅頭

4.4 候補の類似度計算

先ほどの候補それぞれと被修飾語の類似度を計算する. ここでは卵と泡の類似度計算の例を示す.

まず卵と泡に, それぞれが持つ連想語の和集合を次元とするベクトルを設定する. 値としては連想概念辞書の連想距離を用いるが, 辞書中に出てくる最大距離は 15.12 であるため, 15.12 からそれぞれの距離を引いたものがそれぞれの要素の値となる. 次元が和集合となっているため, その語が持っていないイメージに該当する数値がなく, その場合は 0 とする. 例として, 卵及び泡が持つイメージを表すベクトルを表 1 に示す.

表 1: 卵と泡のベクトル

要素名	卵	泡
丸い	13.577	12.060
白い	12.690	10.665
美味しい	11.700	0.000
壊れやすい	6.750	6.750
脆い	6.750	0.000
...
軽い	0.000	12.060
フワフワ	0.000	11.790
テカテカ	0.000	6.750
儂い	0.000	6.750
球状	0.000	6.480

二つのベクトルの類似度の計算には, コサイン類似度を用いる. 卵のベクトルを \vec{a} , 泡のベクトルを \vec{b} とすると,

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} \quad (1)$$

と表せる.

この式を用いて計算した結果, 卵と泡の類似度は 0.504 という結果を得た. この類似度が高いほど候補と被修飾語のイメージが近い, すなわち修飾語として適している可能性が高い, ということになる.

4.5 出力

計算の結果, 類似度が 0 となったものは不適であるとして結果から削除する. 最後に類似度順にソートして出力する.

卵のような泡	0.504
饅頭のような泡	0.394
ご飯粒のような泡	0.281

玉ねぎのような泡

0.275

5 評価実験

5.1 問題作成

入力データとなる色と名詞の対を、辞書にある語の組み合わせで問題作成者に自由に選んでもらい、30の対を用意した。一つの対に対し、システムが出力した修飾語の候補一覧とダミーデータを合わせ、合計数が10になるように候補を提示する。

このダミーデータは色・形状辞書から、色は一致するが形状は一致しない名詞³をランダムに取得してきたものである。そしてそれぞれの候補について、その候補が比喩の修飾語として適切かどうか可、不可をつけてもらう。

■白色の温泉

- (×) おにぎり
- (×) しおり
- (×) ハツカネズミ
- (○) ミルク
- (×) ヨーグルト
- (×) 救急車
- (○) 牛乳
- (○) 乳
- (○) 乳飲料
- (×) 冷蔵庫

このうち、ミルク、ヨーグルト、牛乳、乳、乳飲料がシステムが提示したものであり、それ以外はダミーである。○×はその名詞が比喩の修飾語として適切かどうかという被験者による回答であり、この場合は「ミルクのような温泉」、「牛乳のような温泉」は良いが、「おにぎりのような温泉」等は適切ではない、ということになる。

このような問題を30対、19人の被験者に回答してもらった。

5.2 評価方法

5.2.1 1候補あたりの○の割合

色と形状を用いた候補選定が、色だけを基にしたダミーデータに対して優れているかどうか、また、ある対について出力候補がどれだけ比喩に適切であったかを評価するためのものでもある。

まず一つの対において、10候補をシステムの出力とダミーデータとに分け、それぞれ○の合計数

³候補数が合計10に満たない場合は、同じ色ではなく近い色の名詞を使用した。

を計算する。次にそれぞれを候補数で割り、1候補あたり平均で何人が○としたのかを計算する。最後に被験者数で割り割合とする。これを各対について計算する。

5.2.2 相関係数

連想概念辞書を用いて計算した類似度が、適した数値であるかどうかをこれにより評価する。

各対について、システム出力の候補において、○の割合と類似度にどれだけの相関があるかを、相関係数を用いて計算する。この値が1に近いほど、両者には相関があり、類似度計算が適切に行われたと言える。

6 結果と考察

6.1 1候補あたりの○の割合

各対についての、1候補あたりの○の割合を表2に示す。

表 2: 1候補あたりの○の割合

対	出力	ダミー
赤 マグマ	0.895	0.158
白 温泉	0.789	0.243
黒 コーラ	0.724	0.018
黒 排水	0.711	0.018
黄色 カボチャ	0.684	0.230
...
桃 桜	0.184	0.211
黒 ピアノ	0.158	0.053
黒 カラス	0.158	0.120
赤 リンゴ	0.135	0.193
茶色 チョコ	0.053	0.216
平均	0.413	0.164

まず平均を見た場合、ダミーデータの値0.164に対し、システム出力の値は0.413となっている。この結果から、色・形状辞書を用いた候補選定は有効であると言える。

結果の上位を見ると、マグマ、温泉、コーラ、排水、なみだの形状である「液状」や、かぼちゃ、たまねぎの形状である「球」が目立つ。これら上位にあるものは、共通して形状が単純であり、「球」、「液状」、「気体状」等の一般的な形状表現で表すことが可能である。

それに対し、下位にあるカラス、ピアノ、桜、森等は単純な形状で表すことは不可能である。これらのことから、形状が単純である場合に、色・形状辞書を用いた候補選定が特に有効であると言える。

割合が低い場合の問題点は、形状が合致しても連想するその他の属性が違うということが挙げられる。例えば(赤, リンゴ)の候補として「小豆」といったようなもので、この場合は色と形状は同じだが「大きさ」という属性が違うということになる。このような例では色・形状辞書だけでは解決不能であり、さらに属性を追加した辞書が必要となる。

また、(肌色, うで)のような場合には、手足、人体、指等の同じ上位概念を持ったものが多くリストアップされてくる。これらが「手足のような腕」のように修飾語となると、比喩としての意外性が無く、また比喩としても適切ではないため、シソーラスを用いて省くといった改善も必要である。

複雑な形状をむりやり定義したところにも問題がある。例えば「カラス」は「鳥型」に属するが、他に鳥型に属する候補として上がってきたのはつばめ、とんび、キツツキ等であり、すべて同じ概念の語である。これらは先ほどの例と同様比喩の修飾語として適さない。このことから、形状が複雑な場合は辞書に別の属性が必要になってくる。

6.2 相関係数

結果を表3に示す。

表 3: 相関係数

対	相関係数
白 温泉	1.000
白 バラ	1.000
赤 日の丸	1.000
茶 髪の毛	0.992
黒 排水	0.991
...	...
黒 ピアノ	-0.748
桃 桜	-0.883
緑 森	-1.000
黄色 タンポポ	N/A
透明 なみだ	N/A

1 候補あたりの○の割合と同じような傾向が見られ、上位にあるものは単純な形状のものが多い。

従って、先ほどと同様提案手法が、形状が単純な場合に有効である事が言える。

相関係数が低い場合は、連想概念辞書に記述された連想されるイメージが複雑すぎる事が考えられる。例えば「トマト」から連想されるイメージには「マヨネーズをかける」、「豊作」等が含まれており、このようなものを排除しシンプルにすることで、より正確に類似度が出せるのではないかと考えられる。

また相関係数を計算する際に、データ数が1つ2つと少ない場合があり、この場合に正しく評価できない可能性がある。1つの場合は相関係数自体を出すことができず、二つの場合少し数値が変わっただけで1から-1になってしまうこともありえるからである。より良い評価手法を考える必要がある。

6.3 まとめ

色・形状情報及び連想概念辞書を用いて比喩生成を行い評価実験を行った結果、色のみを用いて選定したダミーデータよりも、色と形状を利用したほうが良い候補を選定することが出来た。また、色・形状辞書を用いた候補選定、連想概念辞書を用いた類似度計算、いずれも形状が単純である場合に有効な手法であることがわかった。

7 今後の展望

今回は手作業で辞書を生成したが、それでは規模を大きくすることが出来ないため、今後はこれを自動生成させる必要がある。

類似度計算に関しては、連想されるイメージをよりシンプルにし、最終的には一つの辞書に合わせたい。

評価手法においては、類似度の評価をする際の相関係数では欠点も見られたため、より良い評価手法を考えていきたい。

参考文献

- [1] 岡本潤, 石崎俊 (2001). "概念間距離の定式化と既存電子化辞書との比較" 自然言語処理, V.8 巻, No.4 号, pp.37-54.
- [2] 山梨正明, 岩田純一 (1988). "比喩と理解" 東京大学出版会
- [3] 福本淳一, 榊井文人 (2002). "統計的手法に基づく形容詞分類を用いた比喩認識" 信学技報 NLC2002-42, Vol.102, No.414, pp.25-30.