コーパスに基づく直喩表現の理解
– 被喩詞の属性名・属性値の同定 –

三島　達夫　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{t-mishi, kshirai}@jaist.ac.jp

1 はじめに

本研究は「N_a のような N_b」といった直喩表現を理解することを目的とする．直喩表現の理解とは，喩詞 N_a によって示唆される被喩詞 N_b の属性名ならびに属性値を決定することと定義する．ここで属性値は形容詞，属性名は形容詞のカテゴリ名である．例えば「氷のような刃」という直喩表現が与えられたとき，刃は「温度」という属性名に対して「冷たい」という属性値を持つと判定する．この属性値は喩詞「氷」が典型的に持つ属性 (顕現属性) である．本研究では，直喩表現を喩詞 N_a の顕現属性が被喩詞 N_b に移される現象と捉え，移される属性名ならびに属性値を決定するモデルを提案する [6]．なお，いわゆる「父のような人物」といったリテラルな表現は本研究の対象外とする．

2 関連研究

比喩理解に関する先行研究のひとつに岩山らの研究がある [2]．彼らの手法では，ある概念の性質を属性名と属性値集合の対の集合で表現する．岩山らが示した概念の性質の例を図 1 に示す．「色」「外形」などが属性名，「赤」「球状」などが属性値に相当する．#の右の数値は，概念がその属性値を持つ確率である．そして，属性値集合の確率分布のエントロピーによって属性値の顕現性を計算し，顕現性の高い属性値，すなわちその概念が持つ典型的な性質が喩詞 N_a から喩詞 N_b に移されるとして比喩表現を理解する．

岩山らは上記のような比喩理解の手法を提案したが，図 1 における属性値の確率の具体的な求め方については

$$*(\text{リンゴ}) = \left\{ \begin{array}{ll} \text{色}: & \left\{ \begin{array}{l} \text{赤}\#0.8 \\ \text{緑}\#0.15 \\ \text{茶色}\#0.05 \end{array} \right\} \\ \text{外形}: & \left\{ \begin{array}{l} \text{球状}\#0.95 \\ \text{円柱状}\#0.05 \end{array} \right\} \\ \text{表面}: & \left\{ \begin{array}{l} \text{滑らか}\#0.9 \\ \text{ざらざら}\#0.1 \end{array} \right\} \end{array} \right\}$$

図 1: 岩山らのモデルによる概念の性質の表現

述べていない．心理実験によって属性値の確率，あるいは顕現属性を獲得する試みはいくつかあるが [1, 7]，その人的コストから大規模な解析システムを構築するのは困難である．そこで，本研究は属性値の確率をコーパスから得られる名詞と形容詞の共起データから推定し，比喩理解を行う手法を提案する．比喩理解や比喩性判定のために名詞と形容詞の共起データを利用する研究は過去にも行われている [4, 5]．しかし，名詞と形容詞が共起するとき，その形容詞が名詞の顕現属性値である場合とそうでない場合がある．本研究では，名詞の顕現属性値を抽出しやすいパタンを新たに発見し，それを用いて名詞と形容詞の共起データを獲得することで比喩理解の精度向上を目指す点が先行研究と異なる．

また，岩山らは喩詞 N_a から被喩詞 N_b に移される性質を決める際，喩詞 N_a の属性値の顕現性しか考慮していない．しかしながら，例えば「火のような風呂」では「(温度が) 熱い」という性質が，「火のような情熱」では「(感情が) 激しい」という性質が移されるように，喩詞 N_a が同じでも被喩詞 N_b によって移される属性名や属性値が異なることは十分考えられる．本研究では，特に属性名を決定する際に，被喩詞 N_b と属性名との関連性を考慮に入れる．

3 提案手法

3.1 属性名と属性値

本研究では，岩山らのモデルと同様に，名詞が持つ性質を属性名 a_i とその属性値集合 V_i の集合として表現する．また，V_i における個々の属性値を v_{ij} とする．

属性名ならびに属性値集合の具体的な設定は以下の通りである．属性名は分類語彙表 [3] の形容詞 (相の類) の意味クラスとした．例えば，意味クラス 3.5020「色」や 3.1912「広狭・大小」などが属性名に該当する．一方，属性名に対応する属性値集合はその意味クラスに属する形容詞とした．ただし，属性名ならびに属性値としてふさわしくないと思われる意味クラスや形容詞はあらかじめ人手で除いた．その結果，100 の属性名ならびに属性値集合を定義した．属性値集合に含まれる属性値の数は

のべ 3,006，異なりで 2,313 となった．

3.2 共起属性抽出パタン

ここでは，コーパスから得られる名詞と形容詞の共起データから，ある名詞が属性名 a_i について属性値 v_{ij} を持つ確率を推定する．

まず，図 2 に示す抽出パタンを用いて名詞と共起する形容詞 (属性値) を抽出する．図 2 において，A は形容詞を，N は名詞を，→は文節の係り受け関係を表わす．以下，図 2 のパタンを共起属性抽出パタンと呼ぶ．

A → N	(ex. 赤いリンゴ)
N + が → A	(ex. リンゴが赤い)
N + は → A	(ex. リンゴは赤い)

図 2: 共起属性抽出パタン

名詞と共起する形容詞を抽出するコーパスとして，毎日新聞の 13 年分の新聞記事と青空文庫[1]からダウンロードした小説 3,466 編を用いた．このコーパスに図 2 の抽出パタンを適用することで以下のデータを得る．

- $O_a^{\mathrm{co}}(v_{ij})$
  喩詞 N_a と共起する属性値を図 2 のパタンで取り出したときの属性値 v_{ij} の頻度．
- p_{ij}^{co}
  $O_a^{\mathrm{co}}(v_{ij})$ から推定される属性値 v_{ij} の確率．属性値集合 V_i 毎に $\sum_{v_{ij}\in V_i} p_{ij}^{\mathrm{co}} = 1$ という制約を満たすように $O_a^{\mathrm{co}}(v_{ij})$ を正規化して推定する．
- $O_b^{\mathrm{co}}(v_{ij})$
  被喩詞 N_b と共起する属性値を図 2 のパタンで取り出したときの属性値 v_{ij} の頻度．

共起属性抽出パタンのマッチングには文節の係り受け解析を行う必要がある．本研究では CaboCha[2] を用いた．

3.3 顕現属性抽出パタン

図 2 の共起属性抽出パタンによって得られた形容詞は，名詞と共起はしているが，必ずしもその名詞の顕現属性値であるわけではない．例えば「青い机」という句があるとき,「青い」は「机」と共起しているが,「机」の典型的な色が青であるというわけではない．したがって，共起属性抽出パタンによって抽出されたデータから属性値の確率 p_{ij}^{co} を推測しても，名詞の顕現属性値に対して高い確率が得られるとは限らない．本研究では，顕現属性値となる形容詞を高い信頼度で抽出するパタン (以下，これを顕現属性抽出パタンと呼ぶ) を新たに発見することでこの問題を解決することを試みた．

顕現属性抽出パタンは以下の手続きで求めた．まず，名詞 N とその顕現属性値 sv の組を 40 個，人手で作成した．その例を以下に示す．

$$(N, sv) = (\text{火, 熱い}), (\text{血, 赤い}), (\text{糸, 細い}),...$$

次に，N と sv がともに現われる例文を青空文庫のコーパスから抽出し，N を中心とする KWIC で表示した．抽出した例文の数は 4,144 である．その例を図 3 に示す．

私の背中と胸へ、何か　火　のように (熱い) ものが触
額に手を当ててみたら　火　のように (熱い) というの
していた－突然、空が　血　のように (赤く) なった－
窓から見れば－ (赤い)　血　のような無数の星の流れ、

図 3: 名詞とその顕現属性値を含む例文

これらの例文から名詞の顕現属性値がよく現われると思われる言語表現を人手で抜き出し，顕現属性抽出パタンとした．例えば，図 3 中には「N のように sv」という表現が頻出するため，これを顕現属性抽出パタンとした．最終的に得られた顕現属性抽出パタンを図 4 にまとめる．

(1)	N + のように → A	(ex. 氷のように冷たい)
(2)	(N + のような)(A) → N’	(ex. 氷のような冷たい感情)
(3)	A → N のような	(ex. 冷たい氷のような)
(4)	N + の → A[さ]	(ex. 氷のような冷たさ)

図 4: 顕現属性抽出パタン

パタン (4) の A[さ] は形容詞の派生名詞 (形容詞語幹+語尾「さ」) を表わす．また，パタン (2) は「N のような」という文節と「A」という文節が同じ名詞 (N’) に係ることを表わす．図 4 の顕現属性抽出パタンのうち，(1) は直接的な比喩表現であるため，顕現属性値が頻出する言語表現のパタンが新たに得られたとは言い難い．これに対し，(2),(3),(4) は今回の調査で初めて明らかになった顕現属性値の頻出出現パタンであると考えている．ただし，我々はもっと多くのパタンが獲得できることを期待したが，結果的には 4 つのパタンしか得ることができなかった．今後，より多くの名詞とその顕現属性値を含む例文を調査し，顕現属性抽出パタンを更に獲得したい．また，テキストマイニングの手法などを用いて顕現属性抽出パタンを自動的に獲得することも今後検討するべき課題のひとつと考えている．

3.2 項で用いたコーパスに図 4 の抽出パタンを適用することで以下のデータを得る．

[1] http://www.aozora.gr.jp/
[2] http://chasen.org/%7Etaku/software/cabocha/

- $O_a^{\mathrm{sa}}(v_{ij})$
  喩詞 N_a と共起する属性値を図 4 のパタンで取り出したときの属性値 v_{ij} の頻度.
- p_{ij}^{sa}
  $O_a^{\mathrm{sa}}(v_{ij})$ から推測される属性値 v_{ij} の確率.
- $O_b^{\mathrm{sa}}(v_{ij})$
  被喩詞 N_b と共起する属性値を図 4 のパタンで取り出したときの属性値 v_{ij} の頻度.

3.4 比喩理解モデル

本項では,「N_a ような N_b」という直喩表現が与えられたとき，喩詞 N_a から被喩詞 N_b に移される属性名と属性値の組 (a_i, v_{ij}) を決定する手法について述べる.

まず，あらかじめ定義した 100 個のそれぞれの属性名 a_i に対し，以下のように属性値 v_{ij} を選び，N_a から N_b に移される属性名と属性値の組 (a_i, v_{ij}) の候補とする.

- 式 (1) の条件を満たすなら，V_i の中から最大の p_{ij}^{co} を持つ v_{ij} を選ぶ.

$$\sum_{v_{ij}\in V_i} O_a^{\mathrm{co}}(v_{ij}) \geq 15 \tag{1}$$

- 式 (2) の条件を満たすなら，V_i の中から最大の p_{ij}^{sa} を持つ v_{ij} を選ぶ.

$$\sum_{v_{ij}\in V_i} O_a^{\mathrm{sa}}(v_{ij}) \geq 5 \tag{2}$$

確率が最大の属性値はその属性名の顕現属性値であると考えられるため，N_a から N_b に移される属性値の候補とする．また，式 (1),(2) の左辺は，それぞれ確率 p_{ij}^{co} ならびに p_{ij}^{sa} を最尤推定したときの分母に相当する．すなわち，これらの値が十分大きくなければ，コーパスから推測した確率の信頼度が高くないとみなし，(a_i, v_{ij}) を N_a から N_b に移される属性の候補には加えない.

次に，このようにして得られた全ての候補 (a_i, v_{ij}) に対し，式 (3) のスコア $S(a_i, v_{ij})$ を計算し，それが最大となる組を 1 つ選択する.

$$S(a_i, v_{ij}) = \alpha \cdot SA_a(a_i, v_{ij}) + (1-\alpha)\cdot REL_b(a_i) \tag{3}$$

式 (3) の $SA_a(a_i, v_{ij})$ は，属性値 v_{ij} の喩詞 N_a に対する顕現度を測る指標である．一方，$REL_b(a_i)$ は，被喩詞 N_b と属性名 a_i の関連度を測る指標である．また，α は両者に対する重みである.

喩詞 N_a に対する属性値の顕現度 $SA_a(a_i, v_{ij})$

顕現度 $SA_a(a_i, v_{ij})$ は式 (4) と定義した.

$$SA_a(a_i, v_{ij}) = \beta\cdot SA_a^{\mathrm{co}}(a_i, v_{ij}) + (1-\beta)\cdot SA_a^{\mathrm{sa}}(a_i, v_{ij}) \tag{4}$$

SA_a^{co} ならびに SA_a^{sa} は，$p_{ij}^{\mathrm{co}}, p_{ij}^{\mathrm{sa}}$ を基に計算される v_{ij} の顕現度である．また，β は両者の重みである．SA_a^{co}, SA_a^{sa} の定義を式 (5) に示す.

$$SA_a^X(a_i, v_{ij}) = \begin{cases} (p_{ir_1}^X - p_{ir_2}^X)\cdot R_a^X(a_i) & (j = r_1) \\ 0 & (j \neq r_1) \end{cases} \tag{5}$$

但し, $R_a^X(a_i) = \dfrac{\sum_{v_{ij}\in V_i} O_a^X(v_{ij})}{\sum_{\forall v_{mn}} O_a^X(v_{mn})}$

X は co または sa のいずれかを表わす．r_1 と r_2 は，確率 p_{ij}^X の大きい上位 1 位と 2 位の属性値の番号を表わす．すなわち，1 位の属性値の確率と 2 位の属性値の確率の差が大きければ大きいほど，v_{ij} は喩詞 N_a の典型的な性質を表わすと考え，高いスコアを与える．一方，$R_a^X(a_i)$ は，N_a と共起する全ての属性値 v_{mn} の頻度の総和に対する属性値集合 V_i に属する属性値の頻度の和の割合である．すなわち，N_a と V_i 中の属性値がよく共起すれば，属性名 a_i に対する候補 (a_i, v_{ij}) に高いスコアを与える.

ここでは，共起属性抽出パタン (図 2) を用いたときと顕現属性抽出パタン (図 4) を用いたときとで属性値の確率を独立に推定し，顕現度 SA_a^X も独立に計算する．顕現属性抽出パタンから得られる属性値は，喩詞 N_a の顕現属性値である可能性が高い反面，パタンマッチに成功する回数が少ないため，データの過疎性の問題が生じやすい．一方，共起属性抽出パタンから得られる属性値は必ずしも N_a の顕現属性値ではないが，パタンとしてはより一般的であるために獲得される属性値の頻度が高く，データの過疎性の問題は生じにくい．両者を独立に計算し，重み付き和によって最終的に $SA_a(a_i, v_{ij})$ を求めているのは，上記のような両者の特徴を考慮したためである.

被喩詞 N_b と属性名との関連度 $REL_b(a_i)$

1 節で述べたように，N_a から N_b に移される属性値を決定するプロセスには被喩詞 N_b も深く関係する．本研究では被喩詞 N_b と属性名 a_i の関連度 $REL_b(a_i)$ を式 (6) のように定義し，全体のスコアに反映させる.

$$S_b(a_i) = \frac{\sum_{v_{ik}\in V_i} O_b^{\mathrm{co}}(v_{ik}) + O_b^{\mathrm{sa}}(v_{ik})}{\sum_{\forall v_{mn}} O_b^{\mathrm{co}}(v_{mn}) + O_b^{\mathrm{sa}}(v_{mn})} \tag{6}$$

すなわち，$REL_b(a_i)$ は被喩詞 N_b と共起する属性値の頻度の総和に対する属性値集合 V_i に属する属性値の頻度の和の割合である．割合の計算には共起属性抽出パタンと顕現属性抽出パタンの両方から得られた頻度を用いる．これは，被喩詞 N_b がある属性名 a_i の属性値とよく共起していれば，その属性名に対する顕現属性値が喩詞 N_a から移されやすいという考えに基づく.

4 評価実験

本節では提案手法の評価実験について述べる．まず，青空文庫の小説から「A のような B」という 47 個の直喩表現を選別し，テストデータとした．その一部を図 5 に示す．

火のような情熱, 氷のような声, 花のような姫君,
絹糸のような髪の毛, 焔のような愛, 粉のような雪

図 5: テストデータ (抜粋)

これらの直喩表現に対して，提案手法によって喩詞 N_a から被喩詞 N_b に移される属性名ならびに属性値を決定した．ただし，式 (4) における重み α は 0.8，式 (6) における重み β は 0.4 とした．これらの値はテストデータの正解率が最大となるように決めた[3]．結果を表 1 に示す．

表 1: 実験結果

	BL	M1 ($\alpha = 1$)	M2 ($\beta = 1$)	M3 (提案手法)
a_i,v_{ij} が正解	–	21	12	25
v_{ij} のみ正解	15	28	14	30
不正解	32	14	18	12
出力なし	–	5	5	5

表 1 において,「a_i,v_{ij} が正解」は属性名と属性値の両方が正解のとき,「v_{ij} のみ正解」は属性値は正解したが属性名が誤りであったとき,「不正解」は属性値が不正解のときを表わす．また,「出力なし」はモデルが属性名・属性値を決めることができなかった場合を表わす．これは，式 (1),(2) を満たす組 (a_i, v_{ij}) がひとつも見つからなかった場合であり，名詞と形容詞の共起データがコーパスから十分に得られなかった場合に該当する．一方，BL, M1, M2, M3 は実験で比較したモデルを表わす．BL はベースラインモデルで，単純に頻度 $O_a^{\mathrm{co}}(v_{ij}) + O_a^{\mathrm{sa}}(v_{ij})$ の大きい属性値を選ぶモデルである．属性名の選択は行わない．M1 は重み α を 1 としたとき，すなわち $REL_b(a_i)$ を無視したモデルである．M2 は重み β を 1 としたとき，すなわち顕現属性抽出パタンから得られる共起データを用いないモデルである．M3 は本研究の提案手法である．

提案手法 M3 の精度は，属性名と属性値の両方が正解した場合で 60%，属性値だけが正解した場合で 71%である．後者の値はベースラインモデルを大きく上回り，提案手法がある程度有効であることを示している．

M1 と M3 との比較から，本研究で提案する被喩詞 N_b と属性名 a_i の関連度 $REL_b(a_i)$ によって精度が若干向上していることがわかる．また，M3 が属性値のみ正解した事例数は，M1 と比べて 2 つ増えたのに対し，属性名・属性値ともに正解した事例数は 4 つ増えた．属性名と属性値のうち，$REL_b(a_i)$ は属性名の選択に有効に働くことを期待して導入したが，この目的がある程度達成されていることがわかる．とはいえ，M1 と M3 の差はあまり大きいとは言えない．

M2 と M3 を比較すると，M3 の正解数は M2 を大きく上回る．本研究で発見した顕現属性抽出パタンを用いることで名詞の顕現属性値を正確に判定することが可能になり，そのことが直喩理解の精度向上にも大きく貢献したと考えられる．また，この実験結果は，単に名詞と形容詞の共起情報をコーパスから獲得するだけでは名詞の性質をうまく捉えることができないことも示唆する．

[3] 重みの調整をテストデータ上で行っているため，今回の実験は厳密な意味でのオープンテストではない．

5 おわりに

本研究では「N_a のような N_b」という直喩表現が与えられたとき，コーパスから得られる名詞と形容詞の共起データを基に N_a から N_b に移される属性名と属性値を決定する手法を提案した．現在のモデルで解析に失敗している主な原因は，属性値 v_{ij} の確率を正しく推定できていないことにある．特に，顕現属性抽出パタンによって抽出された属性値の数が少ないときには解析に失敗することが多かった．今後はより多くの顕現属性抽出パタンを発見するなどして，属性値の確率を正確に推定する手法を探究したい．

参考文献

[1] 今井豊, 石崎俊. 比喩理解における顕著な属性の発見手法. 自然言語処理, Vol. 6, No. 5, pp. 27–42, 1999.

[2] 岩山真, 徳永健伸, 田中穂積. 比喩を含む言語理解における顕現性の役割. 人工知能学会誌, Vol. 6, No. 5, pp. 674–680, 1991.

[3] 国立国語研究所（編）. 分類語彙表. 大日本図書, 2004.

[4] 桝井文人, 福本淳一, 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, Vol. 9, No. 5, pp. 71–92, 2002.

[5] 桝井文人, 福本淳一, 荒木健治. 比喩解釈を目的とする World Wide Web を利用した属性値の適合性判定. 言語処理学会第 11 回年次大会発表論文集, C2-2, pp. 420–423, 2005.

[6] 三島達夫. 名詞の属性に着目した比喩理解に関する研究. Master's thesis, 北陸先端科学技術大学院大学, 3 2007.

[7] Edward E. Smith, Daneil N. Osherson, Lance J. Rips, and Margaret Keane. Combining prototypes: A selective modification model. *Cognitive Science*, Vol. 12, No. 4, pp. 485–527, 1988.