第 8 回 LACE 研究会 2003.Sept.20-21

比喩の自動判定に向けて
- - - 既存の比喩文との意味ベクトル比較の場合

鳥取大学工学部
安井 敏　徳久 雅人　池原 悟　村上 仁一
{yasui,tokuhisa,ikehara,murakami}@ike.tottori-u.ac.jp

1　はじめに

機械翻訳において，比喩文の翻訳が問題の一つになっている．比喩とは，ある事柄を表現するために，その事柄となんらかの関係がある事柄で表現することである．比喩文を直接的に翻訳しても目的言語の表現から喩える関係が見い出せなければ，話し手の意図は，伝わらない．そこで翻訳の前段階として，比喩判定を行なう必要があり，比喩である場合は独自の比喩解釈により翻訳を行う必要がある．

比喩の下位分類は，言語学者により様々であるが，中村明は，語と語の関係に異常性がみられるものを結合比喩と分類している [1]．比喩の自動判定は，名詞と名詞の異常性から比喩を判定することは，桝井らや，内海により行われている [2][3]．しかし，名詞と動詞の異常性から比喩を判定をすることはなされていない．

語と語の関係が正常でない文は，非文といわれている．しかし，全ての非文が結合比喩の文であるとは限らない．つまり，比喩文とみなすには，ある特定の条件が存在すると考えられる．比喩となる条件は様々であり，それら全てを明示的に列挙することは困難である．

そこで，本稿では，既存の比喩文から，格要素と動詞の間の異常な意味的結合性を収集し，その異常性が類似している文を比喩文として自動的に判定する．

2　ベクトル空間を用いた比喩判定方法

2.1　類似文検索

ある 1 つの文が既存の比喩文と類似するかどうかを判定するために，類似文検索の方法を用いる．その検索方法の 1 つにベクトル空間法がある．ベクトル空間法では，文中に含まれるある要素を基底とする特性ベクトルで文を表し，ある 1 つの文とデータベースの文の類似度を，特性ベクトルの内積値で計算する．なお，内積値が高いほど類似度が高い．

2.2　特性ベクトル定義

名詞と動詞の異常性をとらえるため，特性ベクトルは，格要素と用言の共起関係により定義する．

この共起関係から，格助詞が付属している名詞の意味が用言を用いる際に適切であるかがわかる．たとえば「太郎がコップを壊す」という文には,「が格」と「を格」があり,「が格」には動作可能な名詞（主体）が,「を格」には具体物がそれぞれ対応している.「壊す」という用言において，これらの組合せは正常である．

さて，結合価パターンを参考にすると，格助詞は 13 個，名詞の意味属性[1]は 2,710 個である．特性ベクトルは，格助詞毎の意味属性の使用可否を 0 / 1 で表すと，$13 \times 2,710$ 次元になる．そこで，次の考えから次元を縮退させる．

まず，格助詞については,「が」,「は」は主格,「を」は対象,「から」,「より」は起点,「へ」,「まで」,「に」は終点,「も」,「の」,「と」,「で」,「その他」は『その他』として 5 分類にする．次に，意味属性は，木構造であり，上位ノードの意味属性を下位ノードが継承していることから，最下位ノードだけに注目する．したがって，名詞が中位ノードを指定している場合には，それを継承している最下位ノード全てがベクトル化に使われる．こうして，$5 \times 1,921$ 次元のベクトルを用いる．

以下に特性ベクトルの例を示す.

原文:苗木 (673(樹木),689(苗)) を植える

ベクトル (縮退前):

$V^{\text{苗木を}}_{\text{植える}} = [v_i]$ ただし $\begin{cases} v_{673\ \text{を}} = v_{689\ \text{を}} = 1 \\ \text{その他の } v_i = 0 \end{cases}$

ベクトル (縮退後):

[1]ここで意味属性は，日本語語彙大系の一般名詞意味属性を示す.

$V_{\text{植える}}^{\text{苗木を}} = [v_i]$ ただし $\begin{cases} v_{674\ \text{対}} = v_{675\ \text{対}} = v_{689\ \text{対}} = 1 \\ \text{その他の } v_i = 0 \end{cases}$

「$v_{673\ \text{を}} = 1$」は，意味属性が「673(樹木)」，かつ格が「を格」に対応するベクトルの成分が 1 である．本稿では以降，縮退後の特性ベクトルを単にベクトルと呼ぶ．

内積計算

本稿のベクトル同士の内積計算は，同じ用言を持つベクトル同士のみで行えるものとする．よって，用言は特性ベクトルの見出し語となり，内積計算は，実質的に格要素の特性ベクトルで計算する．

2 つの例文「苗木 (673(樹木),689(苗)) を植える」および「庭 (889(庭),423(庭園)) に椿 (樹木 (675(その他)) を植える」をベクトルで表すと，次のようになる．

$V_{\text{植える}}^{\text{苗木を}} = [a_i]$ ただし $\begin{cases} a_{674\ \text{対}} = a_{675\ \text{対}} = a_{689\ \text{対}} = 1 \\ \text{その他の } a_i = 0 \end{cases}$

$V_{\text{植える}}^{\text{庭に椿を}} = [b_i]$ ただし $\begin{cases} b_{889\ \text{終}} = b_{423\ \text{終}} = b_{675\ \text{対}} = 1 \\ \text{その他の } b_i = 0 \end{cases}$

特性ベクトルの内積 prd を求める．

$$prd(V_{\text{植える}}^{\text{苗木を}}, V_{\text{植える}}^{\text{庭に椿を}}) = \sum_i a_i b_i$$

$$= 0 + \cdots + a_{674\ \text{対}} \cdot 0 + \cdots + a_{675\ \text{対}} \cdot b_{675\ \text{対}} + \cdots = 1$$

よって内積値は 1 である．

2.3 判定用 DB

比喩判定に用いるデータベース (DB) を「判定用 DB」と呼ぶ．本稿の実験では,「正常文 DB」と「比喩文 DB」の 2 種類を用いる．正常文 DB における特性ベクトルは，日本語語彙大系における結合価パターンから作成され，名詞と動詞の関係が正常な文の特性ベクトルといえる．ベクトル数は，11,481 である．比喩文 DB は，[1] で紹介されている 5,537 文の結合比喩文のうち，学習データとして，3,750 文を使用する．

類似度計算

ある 1 つの文（入力文）[2]と判定用 DB との類似度は，入力文の特性ベクトルを求め，そのベクトルと判定用 DB 内の各特性ベクトルとの内積の最大値とする．用言が異なる場合は，必ず内積値が 0 になるので，実際には，用言に着目して計算する．類似度 sim を求める式は次の通りである．

$$sim(V_{input}, DB) = \max_{S \in DB} prd(V_{input}, V_S)$$

[2]本稿では，単文のみを対象としている．

2.4 アルゴリズム

本稿では，比喩判定に 2 つのアルゴリズムを試みる．

2.4.1 絶対値による比喩判定

入力文と比喩文 DB の類似度を求め，それと閾値を比較する事により，入力文を正常文か比喩文かのどちらかに判定する．

入力文と判定用 DB の類似度は，用言が同じ場合，仮に 2 個の格要素の意味属性が適合すれば，類似度が 2 となる．このことから，正常文であっても比喩文 DB との類似度は必ずしも 0 にはならない．多くの正常文が比喩文と類似ありと判定されないためにも，正常文と比喩文の類似度を正常文の頻度と組にして求め，頻度が低くなる類似度を，類似ありと判定する閾値に定める．

2.4.2 相対値による比喩判定

入力文と正常文 DB，および，入力文と比喩文 DB の類似度をそれぞれ求める．比喩文の類似度より正常文の類似度の方が高いならば正常文，その逆なら比喩文と判定する．

3 実験

比喩判定のオープンテストを行う．

3.1 絶対値による比喩判定

3.1.1 閾値調査

正常文は [5] から 1,141 文用意し，比喩文は [1] から 965 文用意する．以下に入力文の例を示す．

- 正常例文
  - 彼女は親から財産を受けた。
- 比喩例文
  - スリルを味わう

図 1 に結果を示す．グラフの横軸は類似度，縦軸は頻度である．なお，図中の慣用句については 4.2 節で説明する．

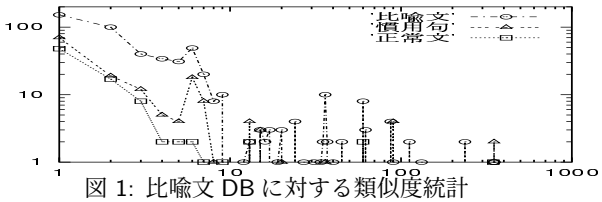

図 1: 比喩文 DB に対する類似度統計

類似度 3 を越えた地点において正常文があまりヒットしていない．したがって閾値を 3 に設定する．

3.1.2 比喩判定実験

(1) 入力

[5] より 100 文，[1] より 100 文を用いる．

(2) 判定結果

正常文は，全て正常文と判定された．比喩文は，82 文が正常文に，18 文が比喩文と判定された．

3.2 相対値による比喩判定

(1) 入力

3.1.2 節と全く同一の 200 文を用いる．

(2) 判定結果

正常文は，99 文が正常文に，1 文が比喩文と判定された．比喩文は，76 文が正常文に，24 文が比喩文と判定された．

3.3 総合結果

以上の結果を表 1 にまとめる．総合結果は,「上述の 2 種類の判定方法のうち，少なくともどちらかの方法で比喩文と判定されたならば比喩文に分類する」とした場合の結果である．

表 1: 判定結果

入力	閾値判定結果		相対的判定結果		総合結果	
	正常文	比喩文	正常文	比喩文	正常文	比喩文
正常文	100	0	99	1	99	1
比喩文	82	18	76	24	73	27

総合結果より，比喩文と判定する上での適合率は 96%，比喩文と判定する上での再現率は 27%である．

4 考察

4.1 結合価パターンのベクトル化における問題

予備実験として，結合価パターンと比喩文のマッチングを本特性ベクトルを用いて試みた．その結果，約 6 割がマッチしていた．その原因の内訳として，中村の結合比喩と語彙体系の収録用言の重複，判定方法の問題が考えられる．基本的には，結合価パターンでは，使用する格要素全ての条件が満たされなければならないが，ベクトルを用いた場合は，少なくとも一つの要素で条件が満たされればマッチしたことになる．

今後の課題として，結合価パターンと比喩文の厳密なマッチングアルゴリズムの追加があげられる．

4.2 慣用句を判定用 DB に用いる効果

慣用句の多くは，単語単独で眺めた際と異なる意味で単語が使われている．これは結合比喩に近い．そこで，慣用句を判定用 DB に用いた際，比喩判定においてどれだけの判定率向上の効果が得られるかを調査する．

[6] で紹介されている 1,550 の文のうち 725 文の慣用句から慣用句 DB を作成する．

4.2.1 閾値による比喩判定

3.1.1 節で用いた正常文および比喩文，それから，慣用句 DB の作成に用いなかった 725 文を慣用句 DB に対して，3.1.1 節の方法で，閾値を求める．図 2 に結果を示す．

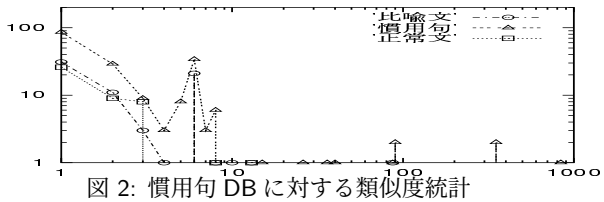

図 2: 慣用句 DB に対する類似度統計

3.1.1 節の閾値調査の時と同様に，閾値 3 を越えた地点において正常文がヒットしなかった．閾値による判定は，3.1.2 節で用いた 200 文と慣用句 DB の類似度を求め，3.1.2 節の方法で判定を行った．結果は，3.1.2 節と同様であった．

4.2.2 相対値による判定

3.1.2 節の入力文と 3 つの DB(正常文 DB，比喩文 DB，慣用句 DB) の類似度を求め，入力文と正常文 DB の類似度が一番大きかったら正常文，逆なら比喩文とした．結果は，3.2 節の場合より 3 文多く比喩文が正常文と多く判定され，3.2 節の場合より 1 文多く比喩文が比喩文と判定された．

4.2.3 総合結果

以上の結果を表 2 にまとめる．総合結果は，3.3 節の方法で求める．

表 2: 慣用句を用いた際の判定結果

入力	閾値判定結果		相対的判定結果		総合結果	
	正常文	比喩文	正常文	比喩文	正常文	比喩文
正常文	100	0	96	4	96	4
比喩文	82	18	75	25	73	27

総合的結果の再現率は，3.3 節の 27%で同様であったが，適合率は，3.3 節の 96%から 87%に減少している．

判定用 DB に慣用句を用いる効果は，3 章の総合結果と比較すると適合率減少などから薄い．原因は，慣用句の位置付けが正常文と比喩文の中間的なものであるためだと考えられる．

5　おわりに

本研究では，名詞と動詞の関係における比喩の自動判定という課題に取り組んだ．その方法として，既存の比喩文との類似性に着目し，その判定方法として，類似文検索の一つである，ベクトル空間方法を用いた．そしてその特性ベクトルに用言と格要素をとり入れ，結合比喩における語と語の異常性をとらえた．

比喩の判定実験では，2 つの方法を行った．1 つは，類似性判定の閾値を定めた絶対的な方法，もう 1 つは，入力文が正常文のベクトル空間に近いか比喩文のベクトル空間に近いかによる相対的な方法である．2 つの方法で総合的に判定すると，適合率は 96%，再現率は 27%となった．

本稿での判定方法は，判定結果で適合率が高いことから比喩文抽出の 1 つのフィルタとして用いることができる．今後の課題は，特性ベクトルの作り方の改良，比喩文のさらなる追加，そして，正常文と判定された比喩文の抽出，などである．

参考文献

[1] 中村: 比喩表現の理論と分類, 共立出版, 1977.

[2] 桝井, 福本, 椎野, 河合: 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.72-92, 2002.

[3] 内海: 比喩理解モデルの諸相, 思考と言語研究会 79-TL-97-5, 信学技報, 1997.

[4] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林: 日本語語彙大系, 岩波書店, 1997.

[5] 計算機用日本語名詞辞書 IPAL 解説編, 情報処理振興事業協会技術センター, 1996.

[6] 宮地: 慣用句の意味と用法, 明治書院, 1982.

[7] 木本: 単語意味属性を用いたベクトル空間法, 鳥取大学修士論文, 2000.