E_005

# WWWからの概念属性値獲得と属性値集合の洗練

## Concept property acquisition from World Wide Web and refinement of property set

森田 悠介† 桝井 文人† 河合 敦夫† 井須 尚紀†
Yusuke Morita Fumito Masui Atsuo Kawai Naoki Isu

## １．はじめに

本論文では比喩解釈を目的とした概念ベース構築について述べる．比喩は様々な分野で研究対象として重要視されており[1]～[3]，自然言語処理においても例外ではない．文章や会話において柔軟な言語処理を実現するためには，比喩表現の理解は不可欠である[4,5]．例えば，対話処理では，比喩理解によって発話数の削減や，より柔軟な対話制御が期待できる[4]．また，質問応答では，以下のような回答を提示することで，効果的に的確な回答をすることが可能となる[5]．

質問　：「中村俊介はどんな選手ですか？」
回答１：「バッジオのような選手です」
回答２：「中田英寿のような選手です」

上記のような比喩解釈を実現するためには，属性比喩の解釈過程が必要である．例えば，「リンゴ」という概念から「リンゴのような頬」という属性比喩を考えた時，その意味は「赤い頬」や「丸い頬」を指す．これらはリンゴの持つ「赤い」，「丸い」といった典型的な特徴（顕現属性値）から連想される．そこで，概念を顕現属性値の集合で表現した知識ベース（以下，概念ベース）を構築することができれば，計算機上での属性比喩解釈に寄与するものと考えられる．

従来，概念ベースの構築には，被験者を用いた心理学実験に基づいた手法を用いることが主流であった[6]．この手法は精度の高い知識が得られる一方，多数の人員の確保や膨大な時間がかかることから，コスト面において問題があった．この問題に対し，桝井らは新聞コーパスから連体修飾関係を利用して概念の特徴を抽出し[7]，World Wide Web（以下，WWW）を利用して知識を精緻化する手法を提案し，低コストでの概念ベースの構築に成功している[8]．しかしながら，彼らの手法では，基本的な知識獲得を新聞記事という知識源に依存しているために，知識の網羅性が新聞記事の内容以上に確保できないという問題があった．

そこで本論文では，WWW を知識源として概念ベースを構築する手法を提案する．WWW を知識源として利用する場合，WWW 中の文章は文体の自由度が高いため，高い精度で WWW から直接知識抽出を行うには，形態素解析や係り受け解析の解析誤りの影響を回避する手段が必要となる．しかし，WWW は非常に多量の情報を有しているためそれらすべてに対しこの様な処理を施していては膨大な計算量が必要となる [8]．提案手法では WWW から必要とする表現を含むテキスト情報を確保しておくことで，計算量を抑えつつ高い網羅性を持つ知識ベースを構築する手法を提案する．具体的には定型表現を含むテキスト情報を文字列照合により WWW から収集し，収集されたテキスト情報に対し形態素解析を施すことで連体修飾関係を抽出し，その連体修飾関係に尤度付けすることで知識ベースを自動構築する．更に，この知識ベースに対し WWW を用いた精緻化を行うことで概念ベースを構築する．

以下，２章で提案手法について述べ，３章で提案手法の妥当性を確認するための先行研究との比較実験と評価を行い，４章ではそれに対して考察する．

## ２．提案手法

本章では，提案手法について詳述する．本手法は，基本知識ベースの構築部と知識ベースの精緻化部の２つの処理部から構成される．以下，各処理部について説明する．

### 2.1 基本知識ベース構築部

本節では，基本知識ベースの構築部について詳述する．知識ベース構築部は(1)フレーズ生成・検索，(2)snippet 取得，(3)形態素解析，(4)概念－属性値情報取得，(5)頻度情報取得，(6)属性値ランキング，からなる（図１）．

以下，各処理過程について順を追って説明する．

**(1) フレーズ生成・検索**

形容詞・形容動詞と名詞の連体修飾関係では，多くの場合，「～な＋名詞」，「～い＋名詞」というパターンで出現する．そこで，入力クエリ（名詞）に対して「な＋クエリ」「い＋クエリ」という定型表現を生成し，これを WWW 上で検索する(注１)．例えば，クエリとして「リンゴ」が与えられた場合，「いリンゴ」「なリンゴ」という定型表現を生成し，WWW検索を行う．

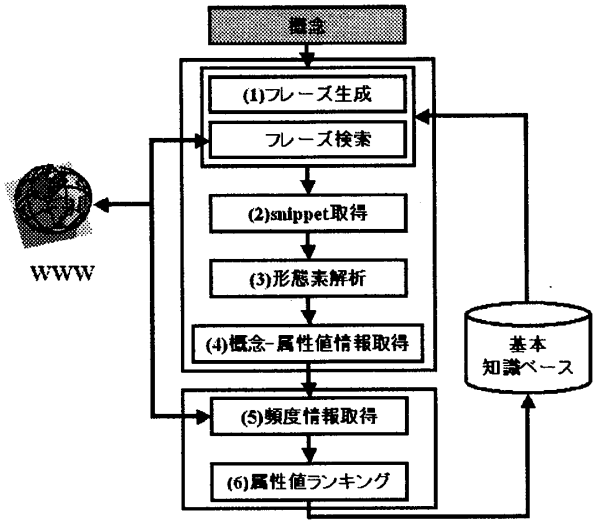


図１　基本知識ベースの構築手法概要

† 三重大学工学情報工学科人工知能講座　津市

(注１)：　検索には，検索エンジン，yahoo! Japan, http://www.yahoo.co.jp を用いた．

**(2) snippet取得**

検索した結果を走査し，検索したフレーズを含む snippet を収集する．

**(3)形態素解析**

収集した snippet を形態素解析し，形容詞・形容動詞（以下，連体修飾語句）と名詞の連接を連体修飾関係として抽出する．「リンゴ」の例では，「赤いリンゴ」や「大きなリンゴ」などの連接が取り出せる．

**(4)概念－属性値情報取得**

(3) で得られた連体修飾関係による語の共起を，概念－属性の関係として抽出する．神崎らは，連体という用法の観点から，連体修飾関係の解析・抽出を人手で行っている[9,10]．これは，連体修飾語句が必ずしも概念の特徴を表すものではなく，顕現属性か否かの判定を自動で行うことは難しいためである．

提案手法では，後述する適合性判定およびフィードバックによってこの問題を解決する．

**(5)頻度情報取得**

(4)で抽出した「属性値＋対象概念」を用いて，再び連体修飾関係「連体修飾語句＋名詞」を生成する．生成した連体修飾関係を再度検索エンジンで検索することにより得られた検索件数を頻度情報として，各属性値に付加する．その結果構築された属性値集合から，頻度情報が閾値以下となった属性値を削除することで，形態素解析誤りなどで抽出してきた連体修飾語句を削除することができる．

**(6)属性値ランキング**

最後に，その頻度情報を各属性値の尤度とし，尤度に基づき概念ごとに属性値集合のランキングを行う．その結果得られたものを基本知識ベースとする．

## 2.2 基本知識ベースの精緻化部

基本知識ベースの精緻化部は，桝井らが提案している適合性判定並びにフィードバックを用いた概念ベース構築手法[8]を拡張することで実現する．基本的には，2.1 で構築した基本知識ベースに対し，「リンゴのように赤い」「蜂のように忙しい」といった比喩や例示を意味する比較表現を生成し，これらの表現が一般的に利用されているか否かを，WWW を用いて調べることによって，連体修飾としての妥当性を自動的に検証・補正するという考え方に基づいている．

以下，基本知識ベース精緻化の流れ（図２）を説明する．

### 2.2.1 属性値の適合性判定

適合性判定の処理は，桝井らの手法と同様，2.1 で構築した知識ベースからの(1)要素の取得，(2)比較表現生成，(3)表現検索，(4)適合性判定，の４つの処理で構成される．

**(1)要素の取得**

基本知識ベースから，比較表現生成の対象となる概念（名詞）と，その概念の属性値（連体修飾語句）を取り出す．

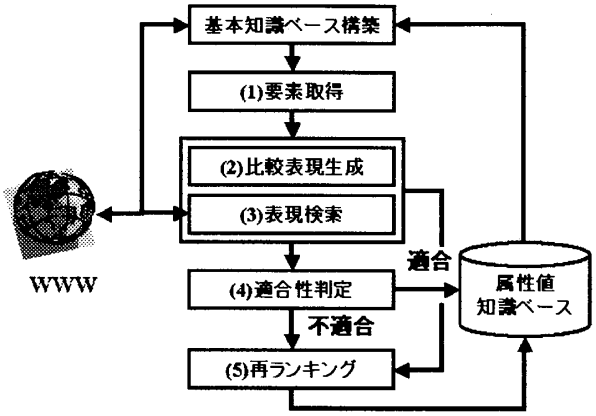


図２　基本知識ベースの精緻化手法概要

**(2)比較表現生成**

まず「X_ground_γ」という定型パターンを用意しておく．このときの ground は指標表現であり，指標表現とは，比較表現を構成する「～のように」といった定型表現を指す．また，X には概念（名詞）が，γには属性値（連体修飾語句）をそれぞれ当てはめて「リンゴのように赤い」のような比較表現を生成する．

使用した指標表現(ground)は，中村の調査による「比喩指標実現形」35 種[11]のうち，比較表現を生成できるもの９種類(注2)の中から，桝井らが使用した「のように」，「のような」に加え，「みたいな」「みたいに」の計４種類を使用した．

**(3)表現検索**

(2)で生成した比較表現をクエリとして WWW 検索を行う．検索の結果得られた件数を，(2)で生成された比較表現の頻度として取得する．

**(4)適合性判定**

(3)の結果，得られた頻度が閾値α以上であれば，属性値は適合と判定される．頻度がα未満であれば，属性値は不適合と判定される．このとき，閾値αは，対象属性が顕現性の高い属性値として適合か否かを判定する基準である．本研究では閾値α＝15 に設定している．閾値をこれ以下にすると概念ベースの精度が失われ，これ以上にすると必要以上に属性値が削除され，網羅性が損なわれる可能性がある．

### 2.2.2 判定結果のフィードバック

フィードバックは，(2)比較表現生成，(3)表現検索，(5)再ランキング，から成る．(2)，(3)は 2.2.1 で示した処理であり，ここで使用される頻度情報は 2.2.1 で取得したものを使用する．

**(5)再ランキング**

属性値集合の再ランキングを行う．WWW から取得した比較表現の頻度情報を用いて，属性値の尤度の再計算を行

(注2)：上位から順に「のような」，「のように」，「みたいな」，「みたいに」，「に似て」，「に似た」，「のごとく」，「のごとき」，「じみた」[8]

う．属性値の尤度計算は，適合性判定の時点で閾値α以上となったものとそうでないものにそれぞれ異なった計算式を用いることで，基本知識ベース構築の際に行った属性値ランキングでうまれたランキングの歪みが補正できる．計算方法の詳細については桝井らの論文[8]を参照されたい．

## 3．実験と評価

提案手法の有効性を検証するために，実験評価を行った．以下，実験と評価について詳述する．

### 3.1. 実験環境

実験では，提案手法および，ベースラインとして，桝井らの手法を用いた．

まず，提案手法およびベースラインを用いて概念ベースを構築する．構築した概念ベースは，提案手法によるものをCBw，ベースラインによるものをCBnとする．次に，10組の概念（表1）[12]について，それぞれの概念ベースを参照し，参照された属性値集合を人手で評価し，その結果を比較検討した．

表1　実験に用いた概念

| 概念 |
|---|
| 部屋，番犬，風船，チーター，湖<br>鏡，冷蔵庫，鬼，風，流れ星 |

両知識ベースから参照した属性値集合は，以下のように評価した．5名の被験者（成人男性）に対して10組の属性値集合を提示し，3段階の評価ポイントを付与してもらった．

- 属性値が概念の顕現属性値として妥当である場合：1pt
- 属性値が概念の属性値としては妥当であるが，顕現属性値としては妥当でない場合：0 pt
- 属性値が概念の属性値として妥当でない場合：-1pt

ポイント付与終了後，各属性値の合計ポイントを求め，合計ポイントが閾値β以上である場合，その属性値が顕現属性値であると見なし，属性値集合における顕現属性値の割合を適合率とした．

ところで，提案手法は，新聞記事と比較してはるかに大規模なWWWを知識源とすることで，獲得可能な顕現属性値の網羅性を高めようとするものである．したがって，本手法の優位性を検証するためには，従来手法に対する網羅性を評価すればよい．

一般に，網羅性を評価するためには，正解データと処理結果を比較する必要がある．ところが，本論文で議論するような属性値集合の正解データを理想的な状態であらかじめ作成することは難しい．そこで，適合率に基づいて網羅性を考慮できる相対再現率（Relative Recall）という尺度を用いた．

### 3.2. 相対適合率

相対再現率（Relative Recall）は，二つの概念ベースA，B間の適合率を利用して求められ[13]，網羅性を間接的に評価することができる．したがって，相対再現率は，正確な再現率を把握できない状況では有効である．相対再現率Ra,bは以下の式で求めることができる．

$$Ra,b = \frac{Ra}{Rb} = \frac{Ca}{Cb} \qquad (1)$$

Ra,b　：概念ベースBから得られる概念ベースAの相対再現率
Ra　：Aの再現率　　Rb　：Bの再現率
Ca　：Aに含まれる顕現属性値数
Cb　：Bに含まれる顕現属性値数

ここでの，Ca，Cbは概念ベースCBw，CBnに含まれる顕現属性値の総数を指す．ここで，それぞれの属性値の適合率Pを利用することで，$C \approx P \times |A|$　と示すことが可能となる．したがって，上記の式(1)を(2)に書き換えることができる．

$$Ra,b = \frac{Ca}{Cb} = \frac{Pa \times |A|}{Pb \times |B|} \qquad (2)$$

Pa　：概念ベースAの適合率
Pb　：概念ベースBの適合率
|A|　：Aの属性値総数　　|B|　：Bの属性値総数

相対再現率Ra,bを求めることで，比較対象の網羅性の性能比較が可能である．

### 3.3. 評価結果

提案手法およびベースラインについて，閾値βを0～5まで変動させた場合の相対再現率の変化を図3に示す．(◆)は提案手法，(▲)はベースラインを示す．提案手法の相対再現率は全ての閾値（β）において1を上回り，ベースラインよりも網羅性が高いことを示した．提案手法の相対再現率は，β=2の時に最大で1.7143であり，ベースラインの相対再現率はβ=4の時に最大0.9167であった．

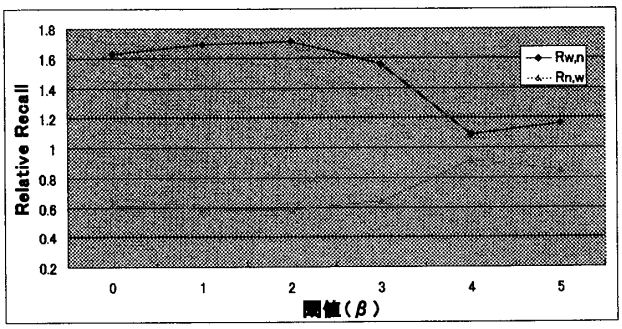


図3　閾値の変化による相対再現率の推移

## 4．考察

評価結果について考察する．

$\beta$が 0〜3 では，相対再現率の差が大きい．$\beta$の値が小さいということは，一般性が大きくない顕現属性値（最も強いとは言えない典型的特長）を考慮することを意味する．

したがって，この結果は，提案手法が，ベースラインと比較して，属性値獲得性能においてより高い網羅性を確保していることを示す．

$\beta$が 4〜5 では相対再現率の差は小さくなっている．$\beta$の値が大きいということは，より一般性の高い顕現属性値（最も強い典型的特徴）であることを意味する．したがって，この結果は，一般性の高い顕現属性値の獲得性能については，提案手法とベースラインの間に大きな差がないことを示している．このことは，一般性の高い顕現属性値については，新聞記事を知識源とした場合でも十分獲得可能であることも意味する．

また，適合率についてみると，提案手法（Pw）では 0.52，ベースライン（Pn）では 0.42 であり，意外にも適合率においても若干の優位性が見られた．提案手法が名詞－形容詞・形容動詞の連体修飾関係に基づいて知識を抽出する点はベースラインと同じであるから，基本的な適合率に差は生じないはずである．にもかかわらず適合率に差が生じた理由として，ベースラインと比べて基本知識ベースの網羅性が高いため，ベースラインでは得られなかった一般性の低い顕現属性値が獲得できるようになったことが適合率にも反映されたと考えられる．

表2　獲得された顕現属性値の比較

| 概念 | CBw に含まれる顕現属性値 | CBn に含まれる顕現属性値 |
|---|---|---|
| 部屋 | プライベートな | |
| 番犬 | | |
| 風船 | 軽い | 軽い |
| チーター | 速い | 優美な |
| 湖 | 静かな，穏やかな(おだやかな)，深い，青い，美しい，綺麗な(きれいな)，巨大な，波静かな | 大きい，静かな，深い，青い，広い |
| 鏡 | 美しい，綺麗な(きれいな，キレイな)，不思議な，透明な | 美しい，きれいな |
| 冷蔵庫 | | 重い，大きい |
| 鬼 | 強い，恐い，悪い，強力な，大きい，冷酷な，醜い，巨大な，真っ赤な，凶悪な | 怖い，恐ろしい，悪い，赤い |
| 風 | 優しい(やさしい)，爽やかな(さわやかな)，自由な，軽やかな，冷たい，気まぐれな，清々しい，自然な，爽快な，涼しい，涼やかな，気持ちいい，心地よい | 優しい(やさしい)，爽やかな(さわやかな)，自由な，心地良い，爽快な，清々しい（すがすがしい），気まぐれな，涼しい |
| 流れ星 | | |

次に，提案手法の処理誤りについて述べる．

基本知識ベースとしては抽出できたものの，適合性判定よって顕現属性値として獲得できなかった例が存在した．表 2 では，「番犬」や「流れ星」の顕現属性値は全く獲得できなかったことがわかる．しかし，これらの概念に対応する基本知識ベースを調べてみると，「流れ星」に対して「ロマンチックな」，「綺麗な」，「番犬」に対して「賢い」，「忠実な」など，顕現属性値として妥当と思われる属性値が含まれていた．これらの属性値を獲得するためには，適合性判定における判定基準の設定において，WWW 以外にも判定用知識源を用いる方法や，適合性判定の閾値の動的決定，比較表現の記述頻度以外にも根拠を求めるなどの工夫などが必要である．

## 5．おわりに

本論文では，WWW を知識源として概念ベースを構築する手法を提案した．WWW から必要とする表現を含むテキスト情報を確保しておくことで，計算量を抑えつつ高い網羅性を持つ知識ベースを構築することに成功した．実験による検証の結果，獲得できる顕現属性値については，従来手法に対して高い網羅性を確保できることが確認でき，さらに，適合率についても優位性が認められた．

今後は，複数の知識源の利用や適合性判定の判定基準における高度化による提案手法のさらなる精緻化の追求，比喩解釈に対する実質的な効果の検証などを進める予定である．

### 参考文献


[1] G. Lakoff and M. Johnson: ``Metaphors We Live by", The University of Chicago Press, Chicago, IL, 1988.

[2] 山梨正明:``比喩と理解", 東京大学出版会, 1998.

[3] 芳賀純, 安増生:``メタファーの心理学", 誠信書房, 1990.

[4] 池ヶ谷有希, 野口靖浩, 鈴木夕紀子, 伊藤敏彦, 小西達裕, 近藤真, 高木朗, 中島秀之, 伊藤幸宏:``対話文脈を利用した構文・意味解析手法の検討", 人工知能学会第 18 回全国大会講演論文集, pp.3E2-10, 2004.

[5] 桝井文人, 森田あすか, 福本淳一:``比喩指標を利用した曖昧な質問への応答", 信学技法 NLC2002-39, vol.102, No.414, pp.7-12, 2002.10.

[6] 岡本潤, 石崎俊:``概念辞書の構築と概念空間の定量化---連想実験による概念空間の抽出", 情処研報, NL130-11, pp81-88, 1999.

[7] 桝井文人, 福本淳一, 椎野努, 河合敦夫:``確率的尺度を用いた比喩性検出手法", 自然言語処理, Vol9, No5, pp.71-92, 2002.

[8] 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手法とそのフィードバック", 電子情報通信学会論文誌, Vol.J89-D, No.4, pp.860-870, 2006.

[9] 神崎享子:連帯修飾関係を結ぶ形容詞類と名詞, 軽量国語学, Vol.21, No.2, pp.53-68, 1997.

[10] 神崎享子, 井佐原均:形容詞類の連体用法にみられる連用的な意味, 軽量国語学, Vol.22, No.2, pp.51-65, 1999.

[11] 中村明:``比喩表現の理解と分類''，共立出版, 1997.

[12] 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.

[13] Patrick Pantel, Deepak Ravichandran, and Eduard Hovy : ``Towards terascale knowledge acquisition", In Proceedings of 20th International Conference on Computational Linguistics: COLING-2004, pp. 771–777, 2004.