E-019

概念の多義性を考慮した属性構造化による概念ベースの構築
Developing the Concept Base with the Attribute Structuration Considering Concept Polysemy

小川 真路† 芋野 美紗子† 土屋 誠司‡ 渡部 広一‡
Shinji Ogawa Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

人は，ある語から関連性のある語を連想する能力があり，この連想機能を日常の会話で役立てている．コンピュータ上に連想能力を実現することができれば，言葉を理解し，人のように返答できる会話システムの実現に近づくと考えられる．そのためには，コンピュータが語と語の関連性に関する知識を大量に保持しておく必要があり，それらを一定形式で集約し保持した知識ベースとして概念ベース[1]がある．

概念ベースは複数の電子国語辞書や新聞記事から機械的に構築された知識ベースである．概念ベースには様々な語（概念）が，それを特徴付ける語（属性）とその重要度を表す数値（重み）の対の集合によって定義されている．

既存の概念ベースは，属性の品詞情報，概念と属性における同義や類義などの語関係に関する情報は保持していない．概念ベースに語関係を定義することで，概念と各属性との関係性を詳しく把握することが可能となる．また，多義である概念の属性には多義のいずれかの意味で関連のある語が混在する形で登録されている．多義概念を意味合いに沿って区別して登録することで，文脈に応じて多義概念を一つの意味に特定することが可能となる．

そこで本稿では，概念と属性間に語関係を定義し，さらに品詞情報を持たせ，多義性を解消した概念ベースの構築手法を提案する．

2. 概念ベース

ある概念Aはm個の属性a_iと重みw_i（>0）の対によって次のように定義される．

$$概念A=\{(a_1, w_1),\ (a_2, w_2),\ \cdots,\ (a_m, w_m)\} \quad (1)$$

例を挙げると，概念「夏」は式(2)のように定義される．

概念「夏」＝{(夏場,0.34),(夏休み,0.11),･･･,(海,0.08)} (2)

既存の概念ベースの概念数は約 9 万語，平均属性数は約 37 個である．

3. 関連度計算方式

関連度計算方式[2]とは，ある 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の実数値で算出され，関連が強いほど高い数値となる．

4. 概念ベースの構築

構築する概念ベースは，属性を名詞，形容詞および動詞ごとに品詞別でグループ化し，各グループの中で概念と属性の語関係（同義や類義など）を属性に持たせた構造とする．構築する概念ベースの概念の定義を図 1 に示す．名詞，形容詞，動詞は単一で意味を持つことができると考えたため，それら 3 つの品詞の語を概念とする．なお，重みについては 4.3 節で述べる．

概念 属性

N	$\{ n_{sa}, n_{si}, n_t, n_i, n_r, a_{sa}, a_{si}, a_o, a_i, a_r, v_{sa}, v_{si}, v_o, v_i, v_r \}$

- N ：名詞概念
- n ：名詞属性
- a ：形容詞属性
- v ：動詞属性
- sa：同義語
- si ：類義語
- t ：上位語
- i ：反意語
- r ：共起語
- o ：名詞化

図 1 概念の定義

概念Nの属性n_{sa}には，概念Nの同義かつ名詞である属性を追加する．図 1 の概念の定義に従って属性を追加することで，語関係や品詞情報を属性に付与することができる．なお，図 1 では名詞概念の定義を示したが，形容詞概念および動詞概念も同様の定義で属性を追加する．

4.1 概念の登録および共起属性の取得

概念の登録および共起属性の追加の際における情報源として，「基本語データベース−語義別単語親密度−[3]」を用いる．情報源の例を表 1 に示す．

表 1 基本語データベースの例

読み	見出し語	語義文（意味文）	品詞
ウスイ	薄い	厚さがわずかである。	形----
ウスイ	薄い	物の濃度や密度が少ない。	形----
⋮	⋮	⋮	⋮

情報源の特徴として，多義である見出し語は意味により区別して登録されている．従って，見出し語を概念とすることで，多義概念は意味により異なる概念として登録することができる．また，共起属性は，形態素解析器，茶筅[4]を用いて見出し語の語義文を単語に分割し，それを属性として追加する．

4.2 関係語辞書による属性の取得

図 1 に示す概念の定義に沿って同義，類義，上位，反意，名詞化の属性を追加する．属性取得では，辞書から各関係語を抽出して構築した関係語辞書を用いて，概念に追加する各々の関係に従った属性を取得する．この過程で概念が多義の場合，ある概念から取得した関係語が，他の意味の概念に対する関係語である可能性がある．そのため，文と文の関連の深さを定量的に表現する文間関連度計算方式[5]を用いて，概念と属性候補の語義文から文間関連度を算出し，値が閾値以上であれば，概念の意味に従った属性とみなして概念に追加する．

4.3 重み付け

属性の重み付けにはtf･idf [6]の考え方を概念ベースに応用した概念ベースidf [7]を用いる．概念と属性の関連度と概念ベースidf の積を属性の重みとする．

† 同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡ 同志社大学理工学部
Faclty of Science and Engineering, Doshisha University

4.4 構築結果

本稿では，2 つの概念ベースを構築した．概念ベース 1（*CB1*）は，前節までで示した方法で構築したもので，概念ベース 2（*CB2*）は既存の概念ベース[1]（*CB*①）から属性を取得し，*CB1* に対して属性追加を行って構築したものである．構築した 2 つ概念ベースの概念数は約 4 万語で，*CB1* および*CB2* の平均属性数はそれぞれ 6 個，15 個となった．構築した概念ベースの概念の例を図 2 に示す．

概念「バス」＝{浴室(名詞 同義),風呂(名詞 類義),…,西洋風,風呂,浴室(名詞 共起),…}

図 2　概念「バス」の例

5. 評価方法

5.1 目視評価

100 個の概念を無作為に選び，概念に追加された属性が妥当か否かを 3 名の目視評価により判断した．3 名中 2 名以上が正しいと判断した属性の割合を精度とする．

5.2 *X-ABC* 評価

X-ABC 評価では，任意の基準概念を *X* と置き，*X* と高関連な概念 *A*，*X* と中関連な概念 *B*，全く関連のない概念 *C* で構成された 4 つの概念の組を 300 組用意する．例えば，概念 *X* が「飲食店」であれば，*A* に「食堂」，*B* に「客」，*C* に「得意」として概念が与えられている．この評価用データを用いて，次の条件式を満たすものを正解とし，正解となった組の比率を概念ベースの精度として評価を行う．

ここで，概念 *X* と概念 *A* との関連度を *DoA*(*X*,*A*)とし，評価セット全体での *DoA*(*X*,*C*)の平均を *AveDoA*(*X*,*C*)とする．

$$DoA(X,A) - DoA(X,B) > AveDoA(X,C) \quad (3)$$

$$DoA(X,B) - DoA(X,C) > AveDoA(X,C) \quad (4)$$

5.3 多義用 *X-ABC* 評価

多義の評価では，基準概念 *X* に対してある意味で高関連な概念 *A*，*A* と同じ意味で *X* と中関連な概念 *B*，*A* とは異なる意味で *X* と高関連な概念 *C* の組を 214 組用意する．例えば，*X* が「バス」であれば，*A* に「風呂」，*B* に「湯」，*C* に「自動車」が与えられる．風呂の意味の概念「バス」に自動車の意味の属性が含まれず，多義が区別されていれば，*DoA*(*X*,*A*)は高い数値となり *DoA*(*X*,*C*)は低い数値となる．この評価セットにおいて次の条件を満たすものを正解とし，その比率を概念ベースの精度とする．なお，*MedDoA*(*X*,*C*)は評価セット全体における *DoA*(*X*,*C*)の中央値である．

$$DoA(X,A) - DoA(X,B) > MedDoA(X,C) \quad (5)$$

$$DoA(X,B) - DoA(X,C) > MedDoA(X,C) \quad (6)$$

6. 評価結果および考察

6.1 目視評価の結果と考察

目視評価の結果を表 2 に示す．

表 2　目視評価による精度（%）

*CB*①	*CB1*	*CB2*
49.3	69.9	61.9

*CB*①より *CB1*，*2* の方が精度が高いことから，雑音となる属性が少ないことがわかる．よって，提案手法により，雑音を抑制することができると考えられる．

6.2 *X-ABC* 評価の結果と考察

X-ABC 評価の結果を表 3 に示す．

表 3　*X-ABC* 評価による精度（%）

*CB*①	*CB1*	*CB2*
82.0	57.3	76.0

CB1，*2* は *CB*①と比べて精度が低い．これは関連度計算の際，属性が少ないために概念の特徴を上手く表現できないことが原因と考えられる．しかし，精錬を繰り返し行った *CB*①と *CB2* の精度は 6%の差であることから，概念に対して相応しい属性の追加により精度の向上が期待できる．

また，属性への語関係の有効性を見るため，精度が高い *CB2* に対して，同義，類義・名詞化，上位，反意，共起属性の順に重みが小さくなるように倍率を設定し，それぞれの属性の重みに積算し補正を行った．補正後の *CB2* を評価した結果，77.3%の精度となり，語関係を用いることで 1.3%の精度向上が見られた．このことから，属性への関係性の付与は有益であると考えられる．

6.3 多義用 *X-ABC* 評価の結果と考察

多義用 *X-ABC* 評価の結果を表 4 に示す．

表 4　多義の評価による精度（%）

CB1	*CB2*
54.2	44.4

多義の評価と *X-ABC* 評価の結果から，*CB1* は同程度の精度であったが，*CB2* では 31.6%精度が低下した．これは，*CB*①からの属性追加において，多義概念の属性に他の意味の概念に対する属性が追加されたことが原因である．今後，多義を考慮した属性追加手法の提案が必要である．

7. まとめ

本稿では属性の品詞や多義，同義や類義の関係をあらかじめ定義した概念ベースを構築した．これにより，属性の雑音を抑制でき，さらに多義を考慮した高品質な概念ベースを構築できた．

謝辞

本研究の一部は，科学研究費補助金（若手研究(B) 24700215）の助成による．

参考文献

[1] 北川晋也, 渡部広一, 河岡司, “連想のための大規模概念ベース構築における重み付け手法の一般化“, 信学技報，AI2007-52, pp.49-54 (2008).
[2] 井筒大志, 渡部広一, 河岡司, “概念ベースを用いた連想機能実現のための関連度計算方式”, 情報科学技術フォーラム FIT2002, pp159－160 (2002).
[3] NTT コミュニケーション科学基礎研究所, “基本語データベース−語義別単語親密度−”, 株式会社学習研究社 (2008).
[4] 形態素解析器, http://chasen-legacy.sourceforge.jp/, 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室) (2013/1/10).
[5] 藤江悠五，渡部広一，河岡司，“概念ベースと Earth Mover's Distance を用いた文書検索”，自然言語処理，Vol.16, No.3, pp.25-49 (2009).
[6] 徳永健伸, “情報検索と言語処理”, 東京大学出版会 (1999).
[7] 芋野美紗子, 吉村恵理子, 土屋誠司, 渡部広一, “概念ベース精錬のための属性追加手法の提案”, 信学技報，AI2010-58, pp.1-6 (2011).