F-007

Web および二次属性を用いた属性追加手法の提案

The Method of Attribute Addition by Web and Secondary Attribute

芋野 美紗子† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Misako Imono Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

１．はじめに

人間は自然言語によるコミュニケーションにおいて，柔軟な理解を行うことができる．自然言語は様々な表現，言い回しを持つが，人間は自分自身が持つ言葉の知識を用いることでそれらの曖昧性を適切に処理することができる．このような人間の機構を機械に持たせるためには，人間が持つ言葉の知識のモデル化や，それを用いた連想メカニズムの構築が必要となる．

人間が持つ言葉の知識には，ある１つの言葉に対して様々な分野の知識が付随している．例えば「煙草」という言葉に対して，「吸う」という動作や，分類を表す「嗜好品」，副産物である「ニコチン」など，人間はそれぞれの言葉に対して様々な知識を持つことで，言葉から自然な連想や言葉の関連性の理解を行うことができる．

このような，人間が言葉に対して持つ知識を機械上にモデル化したものが概念ベース[1]である．一つの言葉を概念とし，概念の特徴を表す語である属性，属性の重要度を表す重みの集合によって定義されている．概念の意味定義は属性の集合によってなされており，人間が言葉に対して持つ知識を属性として多く付与することで概念に常識的な意味をもたせることが可能となる．この概念ベースは国語辞書などから語を抽出することで自動的に構築されており，基本は国語辞書の見出し語を概念，その説明文から属性を抽出している．その他にも百科事典や新聞などの媒体から概念ベースを作る技術[2]も研究されている．この概念ベースを用いて言葉の関係性を定量化する手法が関連度計算方式[3]である．概念同士の持つ属性を比較し，その類似度合いから“関連度”と呼ばれる値を算出する．

概念ベースでは属性の集合体によって概念の意味定義を行う．つまり，属性へ様々な知識が付与されることで，概念は広い意味を持つことができる．概念の持つ意味が多岐に渡ることで，人間のような自然な連想や，曖昧性を解消することのできる概念の意味定義が可能となる．また，関連度計算方式においては比較する属性の選択肢が増し，柔軟な属性比較が可能となる．

そこで本稿では概念ベースで既に定義されている概念に対して，属性を追加付与する手法について述べる．属性として適する語を自動的に獲得し，選別・重み付けをした上で概念に追加付与する．属性の追加により概念の持つ意味を自動的に増やすことで，概念の意味定義を自動的に充実させることが可能となる．本稿では国語辞書とシソーラスを用いて作成された概念ベース[4] に対して属性追加を行い，属性追加手法の有効性を検証した．

†同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

2．概念ベース

概念ベースは単語一つ一つを概念として定義し，人間が持つ概念への常識的な知識をモデル化したものである．概念の意味定義は，属性という単語群と属性それぞれの重要さを表す重みによってなされている．ある概念 A は n 個の属性 a_i と重み w_i の対によって次のように表現される．

$$A = \{(a_1, w_1), \cdots, (a_i, w_i), \cdots, (a_n, w_n)\} \quad (1)$$

ここで属性 a_i を概念 A の一次属性と呼ぶ．概念ベースの具体例を表 1 に，概念ベースの構造を図 1 に示す．

表 1　概念ベースの具体例

概念	属性
医者	(医師,0.34)(患者,0.11)(病院,0.08) …
病院	(医院,0.25)(手術,0.18)(施設,0.04) …
治す	(治療,0.43)(医療,0.21)(病気,0.13) …
…	…

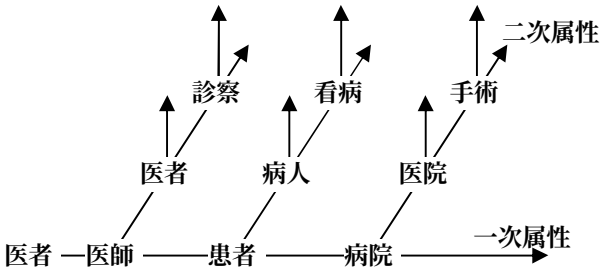


図 1　概念ベースの構造

属性は概念ベース中で概念定義された語のみで構成される．つまり属性を概念と見なし，更に属性を導くことができる．例えば概念「医者」の一次属性「病院」から「医院」や「手術」といった語群導くことができる．これを元の概念「医者」の二次属性と呼ぶ．同様に属性は任意の次元まで導くことが可能であり，概念ベースはこのような属性の連鎖集合によって定義されている．

3．関連度

関連度とは，概念ベースに定義されている概念間の関連性を定量的に表現した値である．関連度は 0.0 から 1.0 に間で値が変動し，概念間の関連が強いほど大きな値を示す．表 2 に関連度の具体例を示す．

表 2　関連度の具体例

概念 A	概念 B	関連度
自動車	車	0.912
	飛行機	0.130
	学校	0.012

関連度は概念がもつ属性同士の対応によって算出される．過去研究[5]より概念が持つ属性のうち 30 個を使用した場合の関連度が最も良い精度とされている．

以下に関連度の算出に用いる一致度および，関連度計算方式について述べる．

3.1 一致度

ある概念 A，B について，その一次属性を a_i，b_j，重みを w_i，v_j とする．それぞれが持つ属性数が M 個と N 個とすると，概念 A，B はそれぞれ

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_M,w_M)\} \quad (2)$$

$$B=\{(b_1,v_1),(b_2,v_2),\cdots,(b_N,v_N)\} \quad (3)$$

と表現される．このとき概念 A，B の一致度 $DoM(A,B)$ は以下のように定義される．

$$DoM(A,B)=\sum_{a_i=b_j}\min(w_i,v_j) \quad (4)$$

$a_i=b_j$ は属性同士が表記的に一致した場合を示している．つまり一致度とは概念 A と概念 B 双方が共通して持つ属性のうち，小さいほうの重みを足し合わせたものとなる．

3.2 関連度計算方式

関連度を算出する概念同士の一次属性全ての組合せに対して一致度の計算を行い，一致度の高い属性の組み合わせから順に対応を決定する．(2)式の概念 A の属性順を固定した上で，一致度が最大となる組み合わせに概念 B の属性を並べ替えたものを以下のように定義する．

$$B=\{(b_{x1},v_{x1}),(b_{x2},v_{x2}),\cdots(b_{xN},v_{xN})\} \quad (5)$$

これら概念 A,B についての関連度を以下のように定義する．

$$DoA(A,B)=\sum_{i=1}^{L}DoM(a_i,b_{xi})\times\frac{(w_i+v_{xi})}{2}\times\frac{\min(w_i,v_{xi})}{\max(w_i,v_{xi})} \quad (6)$$

$DoA(A,B)$ は対応の決定した属性の一致度に，属性それぞれが持つ重みの平均と重みの比率を掛け合わせることで算出される．

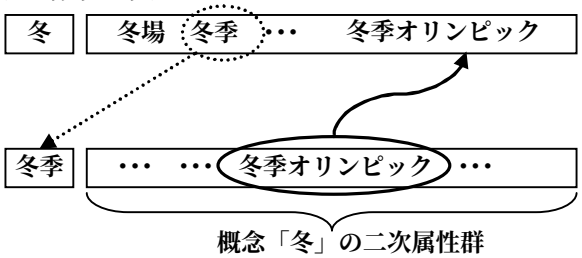
4．属性候補の取得

概念へ追加する属性の候補となる語を 2 つの手法を用いて取得する．以下に具体的な属性候補の取得手法について述べる．

4.1 二次属性からの属性候補の取得

概念ベースは 2 章に示した通り，属性の連鎖的な構造によって定義されている．ある概念 X の意味は X が持つ一次属性によって定義されている．同様に一次属性それぞれの意味も自身が持つ属性，つまり概念 X の二次属性によって定義されていることになる．このことより，ある概念 X から導出される二次属性の中には，概念 X の意味定義に直接関連する語が含まれているのではないかと考えられる．よってこの二次属性を概念への属性候補として取得する．

図 2 に概念「冬」に対して二次属性から新たな属性を取得する様子を示す．



概念「冬」の二次属性群

図 2　概念「冬」への属性追加例

概念「冬」の一次属性である「冬季」からさらに属性を導く．これが概念「冬」の二次属性となる．他の一次属性からも同様に属性を導き，得られた二次属性群から概念「冬」と関連が強い語，例えば「冬季オリンピック」を概念「冬」の新たな属性とする．

4.2 Web からの属性候補の取得

Web からの属性候補の取得はオートフィードバック（以下 AF）[6]を用いて行う．AF とは，ある任意の語について Web 上から属性を付与し，半自動的に概念化を行う技術である．概念化したい語で Web 検索を行い，その検索結果のページ群から自立語を抜き出して属性とする．この AF を既に概念ベースに定義済みに概念について行い，Web 上から属性候補の取得を行う．図 3 に AF からの属性候補取得の具体例を示す．

概念：メニエール病
属性：**めまい**，症状，耳鳴り，**難聴**，内耳，病気，**治療**，**耳**，**原因**，**吐き気**，**ストレス**，**純**，**炎**，**降板**，発作，**診断**，**回転**，**加護**，**薬**，**水腫**，**嘔吐**，**舞台**，**病名**，**悪化**，**リンパ液**，**医学**，**リンパ**，**検査**，**病院**，**突発**

図 3　AF からの属性候補取得の具体例

「メニエール病」は現在の概念ベースに定義されている既存概念であり，太字で示した属性が現在の概念に属性として登録されていない語となっている．既存概念に AF を行うことで，Web 上の自立語から概念に関連のある新たな語を取得することができる．

5．属性候補の選別

二次属性および AF から得られる属性候補は自動的に語を取得しているため，概念にとって重要な語とそうでない語がともに存在する．また，二次属性からの属性候補取得では平均属性数で算出しても 37 の二乗で 1369 個もの膨大な属性候補が得られることなる．

そこで二次属性および AF から得られる属性候補の内，概念にとって重要と考えられる語を多く追加するために属性候補の選別を行った．選別には，属性候補の概念ベース idf，概念と属性候補の関連度，属性候補の重みの三つを閾値として用い，この閾値が一定以上の属性候補が概念にとって重要ではないかと考えた．そこで各選別を様々な閾値で行い，閾値以上の属性候補を新たな属性として概念ベースへ追加して複数の概念ベースを作成し，最も精度の良い選別方法および閾値の選択を行った．それぞれの選別方法について，以下に詳しく述べる．

5.1 概念ベース idf による選別

概念ベース idf とは，文書処理でよく用いられる tf・idf[7]の考え方を概念ベースに適用したもので，概念ベース内での各概念の価値基準の一つである．概念ベース全体を一つの文書空間として捉えることで算出し，値が大きいほど概念の意味定義に重要な語となる．

概念ベースは N 次の属性連鎖集合によって構成されている．この N 次まで属性を展開した空間内で，ある概念 X を属性として持つ概念数から概念ベース idf を算出する．例えば概念「色」を N 次属性空間内で属性として持つ概念数と，概念「紫色」を属性として持つ概念数を比べた場合，後者のほうが少ない．つまりこの N 次属性空間に

おいて，概念「紫色」の方が概念「色」よりも概念を強く特徴付けていると言える．

概念ベース *idf* の算出式は以下のように定義される．

$$CV_N(X) = \log_2 \frac{V_{all}}{df_N(X)} \quad (7)$$

$CV_N(X)$は N 次属性空間内における概念 X の概念ベース *idf* である．V_{all} は概念ベースに定義されている全概念数，$df_N(X)$は N 次属性集合内において概念 X を属性として持つ概念の数である．

概念ベース *idf* はその値が大きいほど概念ベース内での出現頻度が少ないということになる．そのため属性として出現した際には，その概念にとって重要な意味を持つと考えられる．よって得られた属性候補の概念ベース *idf* を算出し，一定の閾値以上ならば属性として重要であると判断して追加を行う．なお本稿では過去研究[8]より精度が最も良いとされている三次概念ベース *idf* を用いた．

5.2 関連度による選別

属性とは概念の意味定義を行う語であるため，概念との関連性も高くなると考えられる．そこで概念と属性候補の関連度を計算し，その値が一定以上ならば属性として追加を行う．

5.3 重みによる選別

概念ベースの属性には，その重要性を意味する重みが付与されている．そこで属性候補への重み付与を行った後，その重みに閾値を定めて一定以上ならば属性として追加する．

ある概念 A に対する属性 B の重み *w(A,B)* について，

$$w_1(A,B) = DoA(A,B) \times CV_3(B) \quad (8)$$

$$w_2(A,B) = DoA(A,B) \times CV_3(B) \times w\_ori(B) \quad (9)$$

$$w_3(A,B) = DoA(A,B) \times \sqrt{CV_3(B)} \quad (10)$$

$$w_4(A,B) = DoA(A,B) \times \sqrt{CV_3(B) \times w\_ori(B)} \quad (11)$$

という 4 種類の重み付与手法について評価を行った．$DoA(A,B)$ は概念 A，B の関連度，$CV_3(B)$ は属性 B の三次概念ベース *idf* である．ここで w_2，w_3，w_4 による選別は二次属性からの属性追加手法においてのみ使用した．$w\_ori(B)$ は二次属性 B の展開元の属性の重みを指しており，w_2，w_3，w_4 は，追加する属性を展開した一次属性の重みを掛け合わせたものとなっている．二次属性は，展開元となる一次属性が存在してこその属性であるため，元の一次属性の重みは重要ではないかと考えたためにこの重み付けについても選別手法として使用した．

6．概念ベースの精度評価

属性追加を行ったことによる概念ベース全体の精度評価を行った．以下に評価に用いた *X-ABC* 評価および精度評価結果を示す．

6.1 X-ABC 評価

X-ABC 評価は関連度の値を比較することで概念ベースを評価する方法である．この評価は表 3 に示す様なテストセットを用いて行う．

表 3　*X-ABC* 評価用データ例

X	*A*	*B*	*C*
飲食店	食堂	客	得意
飲み物	飲料	液体	選択
病人	患者	治療	磁石

ある基準概念 X と，この概念 X と関連が非常に強い概念 A，概念 A ほどではないが関連があると思われる概念 B，まったく関連のない概念 C によって構成している．この 4 つの概念を一組の(*X-A,B,C*)として，人手により人間の常識に沿っていると判断した 500 組を用いて評価を行った．

概念 X と概念 A との関連度を *DoA(X,A)*，概念 X と概念 B との関連度を *DoA(X,B)*，概念 X と概念 C との関連度を *DoA(X,C)*とする．そして表で示したテストセット 500 組での *DoA (X,C)*の平均を *AveDoA (X,C)*として，次の条件によって評価を行う．ここで(14)式の m はテストセットの総数であり，本稿では m=500 となる．

$$DoA(X,A) - DoA(X,B) > AveDoA(X,C) \quad (12)$$

$$DoA(X,B) - DoA(X,C) > AveDoA(X,C) \quad (13)$$

$$AveDoA(X,C) = \frac{\sum_{i=1}^{m} DoA(X_i, C_i)}{m} \quad (14)$$

概念 X と関連がない概念 C との関連度 *DoA(X,C)*は，本来 0.0 となるのが理想である．しかし関連度計算方式の特性上，概念 X と概念 C に一つでも共通した属性が存在すれば微小な値が算出されてしまう．そこで概念 C との関連度を誤差とみなし，その平均 *AveDoA (X,C)*をテストデータ全体での平均誤差とする．そして *DoA(X,A)*，*DoA(X,B)*，*DoA(X,C)*それぞれの関連度の間に平均誤差以上の差が存在していれば，正当な数値ができているとして正解と見なす．この評価を全ての組に対して行った上で，正解となったテストデータの組の比率を概念ベースの精度とした．

6.2 精度評価

二次属性および Web から得た属性候補に対して 5 章で述べた各選別を行った．表 4 に，それぞれの選別において得られた最も高い精度を示す．なお，統合は二次属性からの属性追加と Web からの属性追加の両方の手法を用いて属性の追加を行った場合の結果である．

表 4　*X-ABC* 評価結果

	概念ベース *idf*	関連度	重み (w_1)
追加前	83.6%		
二次属性	84.4%	83.3%	82.0%
Web	84.2%	84.0%	84.6%
統合	85.6%		

二次属性からの属性追加においては，5.3 節で述べたように 4 種類の重み付与について評価を行った結果，重み w_1 の場合が最も高い精度となった．

最終的な結果として，二次属性からの属性追加と Web からの属性追加の両手法を統合して属性を得た場合に，属性追加前と比べて最大 2.0%の精度向上を得た．

7．考察

7.1 属性数の変化

概念が持つ属性数の変化について調査を行った，表 5 にひとつの概念が持つ属性数の平均と，属性数 30 個以下 の概念の割合を示す．なおこの値は表 4 で示した，個々の手法における最大精度を得た場合の結果である．

各属性追加手法とも，属性追加前と比べて概念が持つ属性の数が増加している．また，属性数が 30 個以下の概

念の割合は，両手法の統合時に 19.9%となり，属性追加前と比べて大幅に削減されている．これにより，過去研究[5]において関連度が最も良い精度となる属性数 30 個に満たない概念が減り，関連度計算方式の精度向上を得られたと考えられる．

表 5　属性追加による属性数変化

	平均属性数	属性数 30 個以下の概念の割合
追加前	37.6	51.0%
二次属性	41.0	42.6%
Web	43.5	38.9%
統合	57.6	19.9%

7.2　追加属性例

処理を行うことで実際に追加される属性の調査と，それぞれの手法によって得られる属性の特徴について検証と考察を行った．両手法で新たに追加される属性の成功例を表 6 に示す．

表 6　属性の追加例

概念	二次属性	Web
	追加属性	追加属性
雨	天気雨，暴雨	傘，予報，観測
紙	ケント紙	リサイクル，パルプ
嬉しい	嬉し泣き	応援，土産，サービス

二次属性からの属性追加手法では概念の具体的な語が多く取得される傾向があった．例えば概念「雨」には，具体的な雨の種類が属性として追加されている．二次属性とは，概念がもつ一次属性それぞれの意味定義をしている語群である．*N* 次の属性を辿っていくことは，語の意味をより詳しく展開していくことになる．

そのため概念をより具体的に表した語が属性として取得されやすくなっている．

一方 Web からの属性追加では現在の概念が持つ属性と繋がりのない語を獲得できる可能性が高くなっている．同じく概念「雨」から得られる属性を見ると「雨」という語から連想されるような語が得られているのが分かる．

7.3　他の概念ベースへの適用

1 章で述べたとおり，概念ベースは辞書や百科事典，新聞など，様々な媒体を元に作られており，使用する媒体によって各概念ベースに定義される概念や属性も違う．

そこで，本稿で述べた属性追加手法が他の媒体を元に作られた概念ベースに対しても有効であることを示すため，概念および属性が違う概念ベースに対して本稿で述べた属性追加を行い，追加前との精度を比較した．属性追加には 6.2 節の結果より最も精度の良い，二次属性からの属性追加と Web からの属性追加の両手法を統合した手法を用いた．比較に使用した概念ベースは「自動精練概念ベース」[9]と「新聞概念ベース」[10]である．自動精練概念ベースは国語辞書を元に作成した概念ベースに対して，複数のルールを用いて自動的に属性を精練した概念ベースである．新聞概念ベースは新聞記事から新たな概念および属性を追加した概念ベースである．詳細については参考文献[9] [10]を参照されたい．表 7 に結果を示す．全ての概念ベースにおいて比べて精度が向上した．これにより，本稿の属性追加手法は概念ベース作成の媒体にかかわらず，属性の追加が行えることを示した．

表 7　各概念ベースでの属性追加結果

	自動精練概念ベース	新聞概念ベース
追加前	81.0%	81.6%
追加後	81.4%	82.4%

8．おわりに

本稿では概念ベースに定義されている概念に対して新たな属性を追加する手法として，二次属性からの属性追加と *AF* からの属性追加の二つを提案した．提案手法の結果として，概念がもつ属性数の平均が 37.6 個から 57.6 個に増加し，概念ベース全体での属性数を増やすことに成功した．属性数が 30 個以下の概念割合も，追加前の 51.0%から 19.9%に削減され，これにより関連度計算方式の精度向上を得ることが出来た．また，概念ベースの精度は 83.6%から 85.6%となり，属性の追加前と比べて 2.0%の精度向上を得られた．このことより，属性を追加することで概念が持つ意味を広げ，概念ベースの精練が行われたことを示した．他の媒体を元にした概念ベースに対しても精度向上を得られており，本稿の属性追加手法は他の媒体を元に作られた概念ベースに対しても有効であることを示した．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）21700241）の補助を受けて行った．

文　献

[1] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41–64，2007.

[2] 橋本隆志，渡部広一，河岡司，“新聞記事等の文書を用いた概念自動学習による概念ベース構築方式”，情報処理学会自然言語処理研究会資料，2000-NL-148-13，pp.89-96，2002.

[3] 笠原要，松澤和光，石川勉，“国語辞書を利用した日常語の類似性判別”，情報処理学会論文誌，Vol.38，No.7，pp.1272-1283，1997.

[4] 北川晋也， 奥村紀之，渡部広一，河岡司，“シソーラスの分類情報を利用した概念ベースの属性追加手法”， 情報処理学会第 68 回全国大会講演論文集，4N-5，2006.

[5] 荒木孝允，奥村紀之，渡部広一，河岡司，“比較対象概念の共通属性を重視する動的関連度計算方式”，同志社大学理工学研究報告，Vol.48，No.3，pp.14-24，2007.

[6] 辻泰希，渡部広一，河岡司，“www を用いた概念ベースにない新概念およびその属性獲得手法”，第 18 回人工知能学会全国大会論文集，2D1-01，2004.

[7] 徳永健伸（編），“情報検索と言語処理”，東京大学出版会，1999.

[8] 小島一秀，渡部広一，河岡司，“概念ベースにおける概念属性の確からしさによる概念属性の重み決定法”，信学技報，AI2001-39，pp.39-46，2001.

[9] 広瀬幹規， 渡部広一，河岡司，“概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法”， 信学技報，NLC2001-93，pp.109-116，2002.

[10] 奥村紀之， 渡部広一，河岡司，“電子化新聞を用いた概念ベースの拡張と属性重み付与方式”， 情報処理学会研究報告，2005-NL-166-(8)，pp.55-62，2005.