人工知能学会研究会資料
SIG-KBS-B403-05

人間の連想傾向を基にした 属性の重み補正による概念ベースの精錬

Refining Concept-Base by weight correction of attributes based on human associative tendency

小泉 政弥[1*] 吉村 枝里子[2] 土屋 誠司[2] 渡部 広一[2]

Masaya Koizumi[1], Eriko Yoshimura[2], Seiji Tsuchiya[2], Hirokazu Watabe[2]

[1] 同志社大学大学院理工学研究科
[1] Graduate School of Science and Engineering, Doshisha University
[2] 同志社大学理工学部
[2] Faculty of Science and Engineering, Doshisha University

Abstract: In recent years, the realization of a robot that can conversation like humans using a natural language is required. It is essential that a robot can judge human common sense and associate the word from the other word. Concept-Base is a knowledge base that gathers such knowledge in specific forms. Concepts, the meanings of words, are defined by sets of attributes with weights as the significance of attributes. However, as the strong implications for the concept, larger weight is given for the attribute. Then increasing the weight of attributes which human is likely to associate, is considered to be able to approach the human associative. So this paper proposes a refining Concept-Base by weight correction of attributes based on human associative tendency.

1 はじめに

近年，情報処理システムの高性能化，高機能化に伴いその操作方法は複雑化を辿る一方であり，ユーザが特別な知識や技能を必要としないシステムの構築が求められる．そこで人間が日常会話で用いる自然言語を用いて，人間同士が会話を行うように情報処理システムを扱うことができれば，ユーザの負担が軽減されると考えられる．そのためには人間の常識を理解し，単語間の関連性を判断できるシステムが必要となり，概念ベース[1]や関連度計算方式[2]によりこれを実現する手法が提案されている．

概念ベースの属性は概念の意味を特徴付ける語であり，それらは人間が実際に連想する語とは必ずしも一致しない．概念ベースを用いた関連度計算の際，重みが大きい属性が優先して使用されるため，人間が連想しやすい語の重みを大きくすることで，人間らしい常識判断を行うシステムの開発に繋がると考えた．

そこで本稿では，実際の人間の連想傾向について調査し，NTT シソーラス[3]における傾向を探る．その傾向を基に，概念ベースの属性の属性追加手法及び重み補正手法を提案する．

* 連絡先：同志社大学大学院理工学研究科
〒610-0394 京都府京田辺市多々羅都谷 1-3
E-mail:dun0115@mail4.doshisha.ac.jp

2 関連技術

2.1 概念ベース

概念ベースとは複数の電子国語辞書や新聞記事等から機械的に構築された大規模な知識ベースである．1 つ 1 つの語が概念として定義されており，その意味特徴を表す語である属性と，属性の重要性を表す重みの対の集合によって構成されている．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表現される．

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

属性の数 m は各概念によって異なる．概念ベースの特徴として，属性は概念ベースの中で概念としても定義されている．このとき属性 a_i を概念 A の一次属性と呼ぶ．属性 a_i を概念とみなしさらに属性を導くことができ，概念 a_i の持つ属性を元の概念 A の二次属性と呼ぶ．このように概念ベースは，任意の次元までの属性連鎖集合により定義されている．概念

とその属性の具体的な例を表 1 に示す.

表 1: 概念ベースの具体例

概念	(属性，重み)
医者	(患者,0.63)，(医師,0.58)，(治療,0.35)，…
治療	(診療,1.07)，(手術,0.64)，(怪我,0.07)，…
治す	(癒やす,1.13)，(治療,0.41)，(薬,0.10)，…
…	…

概念ベースには,「林檎」と「リンゴ」のように表記が違うだけで同じ概念を指す言葉がそれぞれ定義されているが，これらを 1 つの概念として数えた場合，概念の総数は約 8 万 7 千概念となる

概念ベースの属性への重み付けには *tf・idf* [4]の考え方を用いる．概念ベースを 1 つの文書空間と見なし，*tf* と概念ベース *idf* による重み付け手法により重みを算出する．概念ベース *idf* 値は以下の式を用いて定義する．

$$CV_N(X) = \log_2 \frac{V_{all}}{df_N(X)} \tag{2}$$

$CV_N(X)$は N 次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る全概念表記数，$df_N(X)$は概念 X を N 次属性内に持つ概念数である．これらの値を利用し，ある概念 A の属性 a_i の重み $w(A, a_i)$を以下の式で付与する．

$$w(A, a_i) = tf_n(a_i) \times CV_N(a_i) \tag{3}$$

2.2 関連度計算方式

関連度計算方式とは，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，概念間の関連が強いほど数値が大きくなる．以下に，関連度計算方式を用いるために必要な一致度及び，重み比率付き関連度計算方式について述べる．

2.2.1 一致度

ある概念 A，B において，その属性を a_i，b_j，対応する重みを u_i，v_j とし，それぞれ属性が L 個，M 個 $(L \leqq M)$とすると，概念 A，B はそれぞれ

$$A = \{(a_1, u_1), (a_2, u_2), \cdots, (a_L, u_L)\} \tag{4}$$

$$B = \{(b_1, v_1), (b_2, v_2), \cdots, (b_M, v_M)\} \tag{5}$$

となる．このとき，概念 A と概念 B の属性一致度 $DoM(A,B)$を以下のように定義する．

$$DoM(A, B) = \sum_{a_i = b_j} min(u_i, v_j) \tag{6}$$

$$min(u_i, v_j) = \begin{cases} u_i (u_i \leq v_j) \\ v_j (u_i > v_j) \end{cases} \tag{7}$$

ここで, a_i=b_jは属性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 に共通する属性の重みの内，小さい方の重みを総和した値となる．これは，小さい方の重みは互いの属性の重みの共通部分となっていることから概念 A と概念 B に共通して有効な重みだと考えられるためである．このとき各概念の重みの総和は 1 になるように正規化する．よって，一致度は 0.0 から 1.0 の値をとる．

2.2.2 重み比率付き関連度計算方式

2.2.1 項で述べた概念 A，B において，属性数の少ない方の概念 A を基準とし，その属性の並びを固定する．その上で概念 B の属性を概念 A の各属性との一致度の和が最大になるように並び替える．概念 B の属性と重みを(b_{Xi}, v_{Xi})として次のように定義する．

$$B = \{(b_{X1}, v_{X1}), (b_{X2}, v_{X2}), \cdots, (b_{XM}, v_{XM})\} \tag{8}$$

概念 A と概念 B についての重み比率付き関連度 $DoA(A,B)$を次の式で定義する．

$$DoA(A, B) = \sum_i DoM(a_i, b_{Xi}) \times \frac{(u_i + v_{Xi})}{2} \times \frac{\min(u_i, v_{Xi})}{\max(u_i, v_{Xi})} \tag{9}$$

関連度 $DoA(A,B)$は，属性の一致度に，属性間の重みの比率と平均値を乗じた値となる．

2.3 NTT シソーラス

NTT シソーラス(以下，シソーラス)は単語の上位-下位関係，全体-部分関係を木構造で体系付けた辞書である．日本語語彙体系[3]から作成されている．2710 個の意味属性(ノード)の関係が木構造で表され，ノードに属する名詞として約 13 万語のリーフが登録されている．シソーラスの一部を図 1 に示す．下線の語がノード，括弧内の語がリーフである．

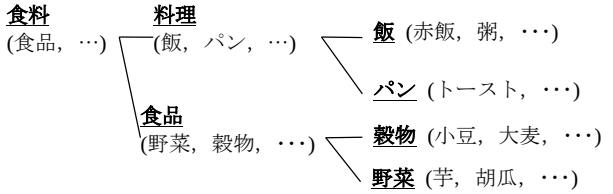

図 1: NTT シソーラスの一部

リーフ「トースト」に対し，ノード「パン」は所属ノード，ノード「料理」やノード「食料」は上位ノードと呼ぶ．また，リーフ「小豆」とリーフ「大麦」は仲間関係，ノード「穀物」とノード「野菜」や，リーフ「小豆」とリーフ「胡瓜」はそれぞれ兄弟関係にあると呼ぶ.

3 連想傾向のデータ取得

文献による調査として，連想語頻度表[5]を用いて連想傾向を調査した．連想語頻度表とは延べ 934 名に対し，ある語（刺激語）から連想された語をその出現頻度とともにまとめたものである．刺激語と連

想語の組，計 15048 組のデータを得た．

4　連想傾向の分析

4.1　シソーラス上での位置関係の定義

シソーラスにおいて，語と語の間に定義されたシソーラス距離と類似度，及び位置関係名の 3 点に着目する．シソーラス距離とは，2 語の位置の近さを示した値であり，類似度は意味の近さを示した値である．位置関係名とは，2.3 節で一部述べたようなノードやリーフの関係の呼称である．シソーラス上の深層ごとにノードに付与されている段数を用い，以下の式で距離と類似度を求める．なお本稿では，リーフの段数は所属ノードの段数に 1 を加えた値とする．

$$R_D(x,y) = (d_x - d_z) + (d_y - d_z) \quad (10)$$

$$R_L(x,y) = \frac{d_x \times 2}{d_x + d_z} \quad (11)$$

刺激語 x と連想語 y に対し，$R_D(x,y)$は距離，$R_L(x,y)$は類似度である．d_x，d_yはそれぞれ x，y の段数であり，d_zは x と y の共通上位ノードの段数である．また，特定の位置関係名においてその距離は一定となる．位置関係名と距離の関係を表 2 に示す．

表 2: 位置関係名における距離

位置関係名	距離
直接の親子	1
間接の親子	$\mid d_x - d_y \mid$
兄弟ノード	2
兄弟リーフ	4
上位ノード	$\mid d_x - d_y \mid$
下位リーフ	$\mid d_x - d_y \mid$
所属ノード	1
直リーフ	1
仲間リーフ	0
関係不明	式(10)

シソーラス上では，ある 2 語間には必ず距離と類似度が算出できる．そこで各距離や類似度，及び位置関係名に対しての連想されやすさを求め，重み補正手法に適用する．本稿ではその連想されやすさを信頼度と定義する．以下 4.2 節，4.3 節，4.4 節に，距離別信頼度の算出手順を述べる．

4.2　平均語数の算出

連想語頻度表の刺激語 284 語に対し，ある刺激語 x_iから距離 D に出現する語数$N_D(x_i)$の平均値$AveT_D$を以下の式で算出する．

$$AveT_D = \frac{\sum_{i=1}^{k} N_D(x_i)}{k} \quad (12)$$

なお k は刺激語数 284 である．同様に連想語頻度表での語の出現頻度$AveR_D$も式(12)により算出する．

4.3　連想語正解率の算出

各距離において，シソーラス上の平均語数における連想語頻度表での平均語数の割合を連想語正解率と定義する．この値が大きいほど，概念から任意の距離に存在する語の内，概念から連想できる語である可能性が高い．距離 D における連想語正解率P_Dを以下の式で算出する．

$$P_D = \frac{AveR_D}{AveT_D} \quad (13)$$

4.4　連想率の算出

連想語頻度表において，刺激語 284 語に対する平均連想語は 52.9 語であった．この 52.9 語に対する，連想語頻度表での平均語数のもつ割合をその距離ごとの距離別連想率と定義する．

4.5　信頼度の算出

以下の式により，距離別の連想語正解率と距離別連想率を掛け合わせた値を距離別信頼度と定義する．

$$RE_D = P_D \times \frac{AveR_D}{52.9} \quad (14)$$

距離別信頼度の算出に用いた値のまとめを表 3 に示す．なお距離 21 以降は信頼度の値が 0 となったため省略する．

表 3: 距離別信頼度の算出

距離	平均語数(シソーラス)	平均語数(頻度表)	連想語正解率(%)	連想率(%)	信頼度
0	128.4	3.19	2.56	6.03	0.2307
1	13.04	1.46	11.18	2.75	0.4602
2	7.59	0.44	5.80	0.83	0..0721
3	21.52	0.53	2.47	1.00	0.0371
4	307.29	2.08	0.68	3.92	0.0397
5	431.54	2.07	0.48	3.90	0.0279
6	877.74	2.70	0.31	5.10	0.0235
7	1581.40	2.68	0.17	5.06	0.0128
8	2687.44	3.01	0.11	5.67	0.0095
9	2845.60	3.07	0.12	5.79	0.0093
10	2965.18	3.41	0.06	6.45	0.0111
11	4102.58	4.29	0.08	8.10	0.0127
12	4991.94	4.54	0.09	8.57	0.0117
13	5675.54	4.98	0.09	9.40	0.0124
14	5905.95	4.79	0.08	9.05	0.0110
15	5611.31	4.19	0.07	7.91	0.0088
16	4509.23	2.77	0.06	5.24	0.0048
17	2934.72	1.85	0.06	3.48	0.0032
18	1228.56	0.69	0.06	1.30	0.0011
19	382.98	0.22	0.06	0.41	0.0004
20	127.38	0.24	0.02	0.05	0.00001

表 3 における距離別信頼度の値は，距離 0 から 24 それぞれに対する信頼度の合計が 1 となるように正規化した値を示す．信頼度は，各距離に存在する語が，ある語から連想される語である可能性，及びどの距離の語がより連想されやすいのかを表現している．この信頼度を重み補正手法に用いる．

同様に類似度に関して，類似度別信頼度を算出する．類似度別信頼度の算出に用いた値のまとめを表 4 に示す．

表 4: 類似度別信頼度の算出

類似度	平均語数 (シソーラス)	平均語数 (頻度表)	連想語正解率 (%)	連想率 (%)	信頼度
0.9~	41.27	2.19	5.31	4.13	0.4914
0.8~	359.82	4.71	1.31	8.89	0.2607
0.7~	361.83	2.07	0.57	3.91	0.0501
0.6~	1266.32	3.52	0.28	6.64	0.0414
0.5~	5283.16	5.90	0.11	11.13	0.0279
0.4~	3395.84	3.81	0.11	7.19	0.0181
0.3~	4172.02	3.38	0.08	6.38	0.0116
0.2~	12354.19	11.60	0.09	21.89	0.0461
0.1~	19576.98	15.58	0.08	29.40	0.0524
0.0~	558.83	0.24	0.04	0.45	0.0004

同様に位置関係名に関して，位置関係名別信頼度を算出する．位置関係名別信頼度の算出に用いた値のまとめを表 5 に示す．なお本稿では，概念との関係が不明な属性に対して補正は行わない．

表 5: 位置関係名別信頼度の算出

位置関係名	平均語数 (シソーラス)	平均語数 (頻度表)	連想語正解率 (%)	連想率 (%)	信頼度
直親	0.10	0.03	29.63	0.05	0.0168
直子	0.15	0.05	34.09	0.10	0.0363
間親	0.91	0.10	10.47	0.18	0.0201
間子	0.94	0.09	9.77	0.17	0.0180
兄弟 N	0.57	0.12	20.86	0.23	0.0504
仲間 L	124.10	3.19	2.57	6.03	0.1657
兄弟 L	236.46	1.35	0.57	2.55	0.0156
上位 N	4.81	0.41	8.57	0.78	0.0712
所属 N	0.98	0.42	42.81	0.79	0.3617
下位 L	43.74	0.46	1.05	0.86	0.0097
直 L	11.81	0.92	7.78	1.73	0.1442
関係不明	46945.70	45.85	0.10	86.52	0.0903

5　属性追加及び重み補正

本稿で重み補正対象となる語は，概念と属性ともにシソーラスで定義されていることが条件である．そのため補正が行われない属性が多く存在し，補正が概念ベースの精度に与える影響が少ない可能性がある．そこで概念の二次属性を参照し，シソーラス距離や類似度を用いて選別し，一次属性に追加を行う．2.1 節で述べた *tf・idf* により重み付けを行った概念ベースを初期 CB とする．初期 CB に対して属性追加及び重み補正を行う．

5.1 距離及び類似度を用いた属性追加手法

シソーラス距離と類似度を用いて，以下の閾値により属性追加を行う．4.5 節で算出した信頼度が比較的大きい値を示したものを採用する．

① シソーラス距離が 0
② シソーラス距離が 1
③ シソーラス距離が 1 以下
④ シソーラス距離が 2 以下
⑤ シソーラス距離が 3 以下
⑥ シソーラス距離が 4 以下
⑦ 類似度が 0.9 以上
⑧ 類似度が 0.8 以上

以上の 8 種類の概念ベースに対し，重み補正手法を適用する．

5.2 信頼度を用いた属性の重み補正手法

属性を追加後，信頼度を用いて属性の重み補正を行う．属性に対する重みは以下の式で付与する．

$$W_N(A, a_i) = w(A, a_i) \times (1 + S) \quad (15)$$

ここで$w_N(A, a_i)$は概念 A に対する属性a_iの補正後の重みである．$w(A, a_i)$は補正前に属性a_iに付与されていた重み，S は信頼度を表す．本稿では，シソーラスに定義されている属性に対してのみ補正を行う．5.1 節で述べた 8 種類の概念ベースに対し，距離別，類似度別，位置関係名別それぞれの信頼度を用いて重み補正を行い，計 24 種類の概念ベースを作成する．

6　概念ベースの精度評価

属性追加及び重み補正を行った概念ベースに対し，*X-ABC* 評価により精度を算出する．基準概念を X と置き，この概念 X と関連が非常に強い概念 A，ある程度関連がある概念 B，まったく関連のない概念 C によって構成された 4 つの概念の組を 689 セット用意する．本稿では属性追加が行われた名詞概念を概念 X とし，評価セットを作成した．ここで $DoA(X,A)$，$DoA(X,B)$，$DoA(X,C)$をそれぞれ X と A，X と B，X と C の関連度とする．そして $AveDoA(X,C)$を *X-ABC* 評価セット全体における $DoA(X,C)$の平均とする．このとき式(16)及び(17)を満たすものを正解とする．

$$DoA(X, A) - DoA(X, B) > AveDoA(X, C) \quad (16)$$

$$DoA(X,B) - DoA(X,C) > AveDoA(X,C) \quad (17)$$

$$AveDoA(X,C) = \sum_{i=1}^{n} (X_i,\ C_i) / n \quad (18)$$

689 セットの内，正解となったセットの組の比率を概念ベースの精度とする．5.1 節で作成した 8 種類の概念ベースに対し，最も精度の高くなった重み補正手法とその評価結果を表 6 に示す．なお属性追加及び重み補正なしの概念ベースの精度は 56.7%であった．

表 6: 概念ベース評価結果

CB	補正手法	精度
①	位置関係名	59.4%
②	位置関係名	74.6%
③	位置関係名	57.5%
④	位置関係名	56.2%
⑤	位置関係名	56.7%
⑥	位置関係名	49.6%
⑦	類似度	72.4%
⑧	類似度	48.4%

表 6 より，シソーラス距離が 1 の二次属性を追加し，位置関係名別信頼度を用いて重み補正を行った概念ベースが最も高い精度を示した．また，類似度 0.9 以上の二次属性を追加し，類似度別信頼度を用いて重み補正を行った概念ベースも同様に高い精度を示した．

7 考察

最も精度が高くなった概念ベースは，シソーラス距離を用いて属性追加を行ったのち，位置関係名を用いて重み補正を行った概念ベースであった．属性追加手法と重み補正手法が異なる理由として，補正対象の属性が異なることが挙げられる．シソーラス距離による重み補正の場合，距離が遠くても一定の補正を掛けている．つまり，本来補正すべきでない属性にも補正が行われている．それに対し位置関係名別信頼度による補正では，関係不明の場合は補正を行っていない．加えて表 2 に示すように，位置関係名が定義されている語は距離も近いことから，追加対象となった属性に対して補正が行われ，精度向上に繋がったと考える．

類似度を用いた属性追加手法では，類似度による補正が最も精度が向上した．これにより，連想されやすい語に対し，補正が行われたといえる．

8 むすび

本稿では人間の連想傾向を，シソーラスにおける位置関係名，シソーラス距離，及び類似度を基に信頼度という値で定義した．その信頼度を用いて二次属性からの属性追加手法，及び属性の重み補正手法を提案した．その結果，概念とのシソーラス距離が 1 の二次属性を追加し，位置関係による重み補正を行った手法により，最大 17.9%の精度向上を得られた．

これにより，人間の連想傾向を反映した概念ベースの構築を実現した．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

[1] 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).

[2] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74(2006).

[3] NTT コミュニケーション科学研究所監修，“日本語語彙体系”，岩波書店，東京，(1997).

[4] 徳永健伸，辻井潤一 (編)，“情報検索と言語処理”，東京大学出版会，東京(1999).

[5] 水野りか (編)，“連想語頻度表－3 モーラの漢字・ひらがな・カタカナ表記語－”，ナカニシヤ出版，2011.

[6] 長尾眞，佐藤理史，黒橋禎夫，角田達彦，“自然言語処理(岩波講座 ソフトウェア科学 15)”，岩波書店(1996).