E-40

連想システムのための概念ベース構成法 ―概念属性の精錬と拡張

A Method of Concept-Base Construction for an Association-System - Concept Attribute Refinement and Expansion

小島　一秀†，渡部　広一†，河岡　司†
Kazuhide Kojima, Hirokazu Watabe and Tsukasa Kawaoka

1. はじめに

本研究の目的は人間らしい常識的な判断を実現する常識的判断システムの開発である．この常識的判断システムの中核には概念を定義する概念ベースが存在し，概念間の関連の強さを定量化する関連度[1]の計算などに用いられている．本稿では，この概念ベースの構築に用いられる，概念ベースの精錬と拡張について述べる．

2. 概念ベース

概念ベースとは，語Aの意味を語 a_i とその重み u_i $(0 < u_i \leq 1)$ の集合として定義する知識ベースである（表1）．語Aの意味の定義に使われる語 a_i を属性と呼ぶ．語 a_i は必ず概念ベースに属する語でなければならない．なお，属性はその重みが0になった時点で概念ベースから論理的に存在しなくなる．

表1　概念ベース

語	属性と重み
雪	(雪, 0.61), (白い, 0.30), (下る, 0.27), …
白い	(雪, 0.16), (白地, 0.14), (色, 0.14), …
下る	(低い, 0.23), (雪, 0.21), (雨, 0.20), …

基本概念ベース（基本 CB）は複数の国語辞書の見出し語を概念，説明文の自立語を属性として自動構築されている[2]．また，属性には辞書の説明文における語の出現頻度から決められた重みが付加されている．概念数は約3万，1概念当たりの平均属性数は約45となっている．

3. 概念ベースの精錬

3.1 概念ベース精錬の手順

提案する概念ベース精錬方式は，属性信頼度の計算，属性の分類，重みの決定からなる．基本 CB に精錬処理をした概念ベースが精錬概念ベース（精錬 CB）である．

属性信頼度の計算では，様々な手がかり毎に属性信頼度を取得し，複数ある属性信頼度を一つの属性信頼度に合成する．属性信頼度とは語と属性の間にある関係の確からしさを定量化した 0%から 100%の値である[3]．このとき用いる各手がかり，属性信頼度の合成については後の節で述べる．

属性の分類では，属性信頼度により表2のように属性を分類する．基本的に属性は属性信頼度によって分類されるが，さらに詳細に重み付けを行うために属性信頼度 100%の信頼度1クラスについてはより細かく分けている．なぜなら，信頼度1クラスに分類された属性は，それが定義する語と同義，類義，上位下位の3種類の論理的関係（表3）にある属性で構成しており，これらの重みは当然異なることが想定できるからである．このような理由から，信頼度1クラスの属性は，重み付けのクラス分けとして更に同義クラス，類義クラス，上下クラスに分類している．

表2　属性信頼度による分類

クラス	属性信頼度（%）	クラス	属性信頼度（%）
信頼度1	100	信頼度4	40以上60未満
信頼度2	80以上100未満	信頼度5	20以上40未満
信頼度3	60以上80未満	信頼度6	0以上20未満

表3　関係データ

語	語	関係
書籍	辞書	上位下位
字引	辞書	同義
雪	吹雪	類義

重みの決定では各クラスの重みを学習用データを使って実験的に決定する．これは，どのような関係がどのような重みになるかは人間にもわからないためである．今回は重みを決定するために，次のような試行実験を行った．信頼度2クラスを常に基準値1とし，信頼度3，4，5，6のクラスの重みには，1，0.5，0.25，0を試行した．ただし，試行数を抑えるため信頼度クラスの試行では属性信頼度が高いクラスの重みより低いクラスの重みが大きくならないような試行を行った．

3.2 属性信頼度を導く手がかり

属性信頼度は様々な手がかりを用いて求めるが，その手がかりから導かれる各属性の属性信頼度は，人間による属性の評価を用いて次のように決定する．まず，基本 CB から 100 語をサンプル語として選び出し，サンプル語の属性であるサンプル属性が適切かどうかを人間が判定を行い，各手がかりとサンプル属性の適切な率との関係を調べる．

例えば，手がかりの一つである関連度の場合，サンプル属性を関連度により何グループかに分ける．その後，各グループにおいて人間に適切と判断された属性の率を参考に，そのグループの関連度から導かれる属性の属性信頼度（関連度）を決める．

以下では，個々の手がかりとそこから導かれる属性信頼度について述べる．なお，ここでは語 a_i は語Aの属性としている．

(1) 語と属性の一致

語Aと属性 a_i が等しければ語Aと属性 a_i は関係があることは間違いないため，属性 a_i の属性信頼度は 100%である．同時に，語Aと属性 a_i が同義であることがわかる．

(2) 関係データ

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

語 A と属性 a_i の間に関係データにおいて論理的関係が定義されている場合，属性 a_i の属性信頼度は 100%になると同時に，語 A と属性 a_i の間にある論理的関係も明らかになる．関係データとは，電子化国語辞書の解析[4]により機械的に作成した語間の論理的関係のデータである．

(3) 基本 CB における属性の重み

基本 CB 構築時の出現頻度に基づく重みが大きければ属性 a_i が語 A の属性として適切である可能性が高い．

(4) 関連度

関連度とは語と語の関連の深さを定量化した 0 から 1 の値で，関係が深いほど大きな値となる[1]．語 A と属性 a_i の関連度が高ければ，属性 a_i が語 A の属性として適切である可能性が高い．

(5) 相互属性

属性 a_i は語 A の属性であるが，さらに語 A が語 a_i の属性として使われている場合，属性 a_i を語 A の相互属性と呼ぶ．属性 a_i が相互属性であるとき，語 a_i の属性 A の重みが大きければ属性 a_i が適切である可能性が高い．

3.3 属性信頼度の合成

手がかりは複数あるため一つの属性に対して複数の属性信頼度が求められるが，属性信頼度は確率のような値であるため，計算により一つに合成できる．ある属性 a_i に対して統計的に独立した 2 つの手がかり 1，2 から属性信頼度 p_1，p_2 が得られたとき，このときの属性 a_i の属性信頼度 P は次のようになる[3]．

$$P = p_1 p_2 / \{p_1 p_2 + (1-p_1)(1-p_2)\} \quad (1)$$

4. 概念ベースの拡張

概念ベースに新しい属性を追加する処理が概念ベースの拡張である．基本 CB に拡張処理を行ったのが，拡張概念ベース（拡張 CB）である．

概念ベースの拡張は，(1)概念ベースからの雑音削減，(2)相互属性化による属性追加，(3)2 次属性の追加の 3 段階からなるが，具体的には次のようになっている．

雑音削減では不適切な属性である雑音を削除し，適切な属性を残す．具体的には，次の条件のいずれにも当てはまらない属性を基本 CB から削除する．ここでは，語 a_i は語 A の属性であるかまたは，属性として追加する候補とする．

- 属性 a_i の基本 CB における重みが 0.06 以上
- 属性 a_i と語 A との関連度が 0.2 以上
- 属性 a_i が相互属性である．
- 属性 a_i が語 A と同じ漢字を含んでいる．
- 属性 a_i と語 A がシソーラスにおいて上位下位関係か同じカテゴリに属している．
- 属性 a_i と語 A の関係が関係データで定義されている．

相互属性化による属性追加では，属性 a_i が語 A の属性であるとき，語 a_i の属性として語 A を追加する．2 次属性からの属性追加では属性を重みの上位 30 個までに制限して，語 A の 2 次属性の中から，上で述べた条件のうち漢字の条件，シソーラスの条件，関係データの条件のどれかを満たす属性を追加する．追加されたものを含めた属性の重みとしては関連度を用いている．

5. 精錬と拡張の評価

順序正解率はテストデータを用いて求める概念ベースの評価値である．テストデータは 4 語で 1 組をなすデータで，この語の組は，基準語 X，語 X と同義または類義の語 A，ある程度関係のある語 B，関係のない語 C からなっている（表 4）．テストデータは 590 組のデータから成り，人手によって作られている．テストデータの語間で関連度計算[1]を行い，求まった関連度の値を比較して求める．基準語 X と A，B，C の関連度をそれぞれ R_a，R_b，R_c とする．これらの値は $R_a > R_b > R_c$ という大小関係が期待される．テストデータの全ての語の組の中で，このような順序になり，かつ各関連度の差が全 R_c の平均より大きい率を順序正解率として概念ベースの評価に用いる．

表 4　テストデータ

X	A	B	C
樹木	木	木の葉	頭
天気	天候	雨	写真
時刻	時間	時計	消しごむ

基本 CB，精錬 CB，拡張 CB の順序正解率はそれぞれ，49%，64%，77%であった（図 1）．したがって，順序正解率は精錬により 15%，拡張により 28%改善され，概念ベースの精錬と拡張がともに有効であることがわかる．

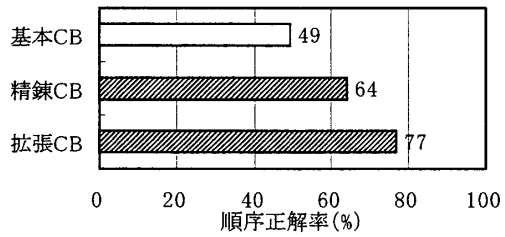

図 1　各概念ベースの順序正解率

6. おわりに

概念ベースの精錬と拡張の両方の有効性を，評価実験により示した．概念ベースの精錬では主に属性の選別と削除を行い，拡張では主に属性の追加を行った．したがって，両者を組み合わせることにより，属性の追加と削除の両方の側面に対応できる．今後は両者の整合性を取り，組み合わせた概念ベースの構成法を組み立てたい．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

[1] 渡部広一，河岡司：“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001

[2] 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

[3] 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

[4] 小島一秀，渡部広一，河岡　司：“常識判断のための概念ベース構成法：概念間論理関係を用いた概念属性の重み決定法”，信学技報，AI2000-80／KBSE88(2001-3)，pp.57-64，2001