自然言語処理 141－20
（2001．1．26）

関連度評価のためのルールによる概念ベースの自動精錬

浦 政博　小島 一秀　渡部 広一　河岡 司

同志社大学大学院工学研究科
〒610—0394 京都府京田辺市多々良郡都谷 1—3

本稿では“柔軟な判断や推測”に基づく知的判断をコンピュータにさせる上で中核となる汎用的な知識資源「概念ベース」の自動精錬方法について提案している．概念ベースとは概念をその他の概念（属性と呼ぶ）集合で表した知識ベースであり，電子化辞書等から機械的に作成される．このため不適切な属性も多く含まれる．提案方式は概念ベース内における概念の適切な関係をルール化して，自動的に概念ベース内から不適切な属性を除去，適切な属性の追加を行う．提案手法を用いて自動精錬を行った概念ベースはサンプルによる目視判断においても，また関連の近さを示す評価用テストデータを用いた実験においても精錬前の概念ベースと比較して品質が向上していることを示した．

Automatic Refinement of Concept Base with Rules for Evaluation of The Degree of Association between Concepts

Masahiro Ura,　Kazuhide Kojima,　Hirokazu Watabe,　Tsukasa Kawaoka

Graduate School of Engineering, Doshisha University
Kyotanabe, Kyoto 610—0394

Human can judge intelligently based on imperfect information than computer. Concept Base is one of the main elements to realize the intelligent judgment by computer. This paper shows that Concept Base becomes closer to human judgment by automatic refinement. Concept Base is a knowledge base in which each concept consists of a set of other concepts(attributes). Concretely, Concept Base is refined automatically with removing improper attributes and adding proper attributes by using the rules which are derived from the relationship between concepts in Concept Base. The improvement of Concept Base is shown by the experimental result using the degree of association.

1. はじめに

従来の情報処理システムは全てのデータや処理方法がきちんと整理された問題だけを対象としてきた．しかし，現実の世界では常に全てが整った問題ばかりを対象とするわけではなく人間が得意とする“柔軟な判断や推測”を行えるような高度な知的判断を伴う情報処理システムが必要とされる．

人間の“知的”や“知能”については古くから哲学，生物学，認知科学，理学，工学など幅広い領域で各種の研究が行われてきた．工学的な観点での“知的”に限っても，人工知能の実現可能性をめぐり多くの議論，研究がなされてきたが，未だに“知的”に関する明快な定義が与えられているとはいい難い．

そこで我々は，複雑で厄介な人間の“知的”

の本質の解明というよりは，情報処理の高度化につながるような現実的なメカニズムの創出を目的とした概念ベースを基盤とする“知的判断メカニズム”の実現に取り組んでいる．

知的判断メカニズムを実現するにはその中核機構として，広範囲な様々な判断に対応するための汎用的な知識資源が必要となってくる．本研究ではこの汎用的な知識資源として，実世界における様々な事象の概念を格納した概念ベースを想定している．

このような「概念ベース」を構築する際に問題になるのは格納するべき概念の数の膨大さである．従って全ての概念とその属性を手作業で入力するのは現実的ではない．そこで概念ベースを機械的に構築可能かどうかが重要となるが，現時点では機械的に構築された概念ベースは手作業で構築された概念ベースに比べ概念の質の面では大幅に劣っている．

そこで，概念ベースの質の向上を目的とした概念ベースの品質向上，特に機械的な品質向上の有無が重要になってくる．このような概念ベースの質の向上を目的とした操作を概念ベースの精錬と呼んでいる．精錬の一つの方法は，概念そのものと属性の関係の強さを定量的に評価した値『関連度』を用いて行われるのだが，精錬前に各属性の適切な関連度を期待することは困難である．このため本稿では関連度を用いずに概念ベース内における概念間の直接的な表記一致関係，間接的な相互関係（ルール）を用いて概念ベース内の知識を精密化する機械的な精錬手法の提案を行う．そして，精錬後に関連度を定義しこれを用いた評価実験で提案する精錬手法が有効であることを示す．

2. 概念ベース

2.1 機械構築された概念ベース

概念ベースにおける概念とは，属性語 a_i と重み $w_i(>0)$の組で定義された属性の集合のことを指す．概念ベースはこのような概念を多数格納している．

本稿で対象とする概念ベースは，複数の国語辞書等から機械的に構築されている[1]．この概念ベースの各属性と重みは，複数の国語辞書等の語義文から自立語の出現頻度に基づいて獲得している．更に，自己参照による新たな属性の追加，及び不要な属性の統計的な除去という操作も行っている．なお，この概念ベースの概念数は約 5 万語，属性数は約 150 万語となっている．

2.2 基本概念ベース

本稿では概念ベースの属性の重み情報を取り除きこれを基本概念ベースと呼ぶ．属性の重み情報を捨てた理由は，機械的に抽出した重み情報の信頼性が乏しいこと，さらに自動学習（概念の機械的な追加と変更）の容易性の面から出来るだけ単純な構造にすることが望まれるためである．もし重み情報を削除しなければ，ある概念に新たな属性を追加する場合，その属性の重みの決定や既に付与されている属性の重みの再調整を行わなければならない．しかし，例えば情報源 A から統計的に得られた属性の重みと情報源 B から別の処理で得られた属性の重みを適切に扱うのは非常に困難な問題となる．そこで，概念ベース内の統計的に求められた重みは保持しない方が妥当と判断した．

3. 基本概念ベースの精錬

機械構築された基本概念ベースには多くの雑音が含まれており人間の感覚とは適合しない部分が多い．そこで概念ベース内から不適切な属性の抑制，適切な属性の追加を機械的に行う方法が不可欠となる．

3.1 精錬操作の必要性

概念ベースの質を評価する方法として，概念の各属性を人間が見て判断する方法が挙げられる．そこで，サンプルとして 50 の概念を抽出し，それぞれの概念の各属性について以下のような評価を行った．

表 1. 属性の評価基準

評価	評価基準	
◎	関連あり	概念の属性として望ましい属性
○		十分関連はあるが絶対に必要というほどではない属性
△	関連？	全く関連なしとはいえないがどちらかというと不適切な属性
×	関連なし	概念と全く関連のない属性

この表 1 の評価基準を用いて，基本概念ベースの 50 概念の属性を評価した結果が図 1 である．この図から分かるとおり，基本概念ベースには多くの不適切な属性が含まれているといえる．以下に精錬操作の詳細について述べる．

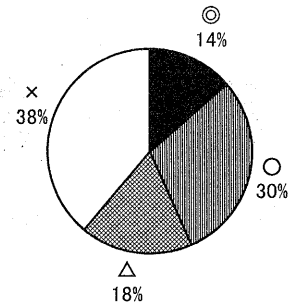

図 1. 基本概念ベースの属性評価

3.2 不適切な属性の削除（雑音除去）

概念ベース内の概念—属性関係には一方の概念が他方を属性として持ち，その逆で他方を概念として見たときに一方を属性として持っているような特殊な関係が存在する．この関係を両概念間における相互リンクと呼ぶ．相互リンクの関係を持つ両概念は基本概念ベースを構築する上で属性としてお互いを獲得しているので，両概念の関係は非常に深いと考えられる．よって，相互リンク関係を持つ概念—属性関係を残存させる（図 2）．

この相互リンク関係は両概念間の一つの特徴を示している．雑音除去ではこれに加えて概念—属性間の表記特徴も利用する．概念—属性の関係となっている両概念において一方の一部の表記漢字が他方の表記の一部に含まれている場合その概念—属性関係は適切な属性と考えられるのでこの概念—属性関係についても残存させる（図 3）．

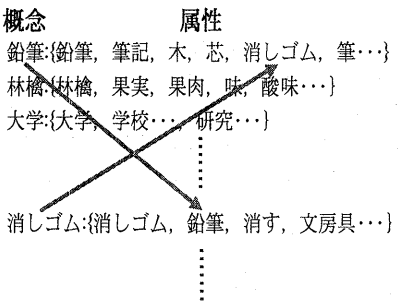

図 2. 相互リンクの例

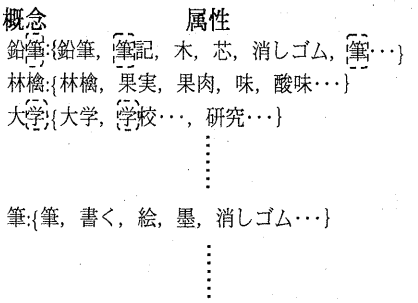

図 3. 表記漢字を用いた雑音除去

すなわち不適切な属性の削除はこの 2 つのルールによって行う．この 2 つのルールに該当しない属性は全て雑音とみなし削除する．この雑音除去の効果については 4 章で述べる．

3.3 適切な属性の追加

3.2 の 2 つのルールを用いて基本概念ベース内の雑音除去を行った．この雑音除去によって得られた概念ベースを以後中間概念ベースと呼ぶ．基本概念ベースにおいて雑音除去を行った後の中間概念ベースの属性を人手で検証してみると雑音の低減が見られ属性数の減少が顕著となる．

中間概念ベース内においてより適切な属性を追加するため以下のような関係を満たす 2 つの概念について新たに概念—属性関係を適正とする．

漢字表記の特徴ルールを満たす概念—属性関係において相互リンクの関係になっていなければ相互リンクを実現する．この操作を相互リンクの拡充と呼ぶ．漢字表記の特徴ルールにおいて残存した概念—属性関係は相互リンクでない場合が多い．表記特徴を満たすのであれば関係が深いと考え相互リンクの拡充を行う．属性追加はこのルールによって行う（図 4）．

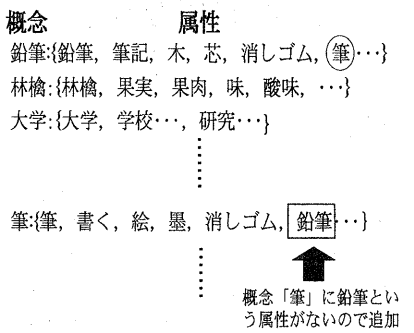

図 4. 相互リンクの拡充

4. 精錬手法の評価

4.1 属性分布から見る精錬手法の検証

図6に基本概念ベース，基本概念ベースに相互リンクルールを適用した概念ベース，表記特徴ルールを適用した概念ベース，中間概念ベース，中間概念ベースで相互リンクを拡充した精錬後概念ベースについて平均属性数を示した．この図からわかるように基本概念ベースと比較して中間概念ベース，精錬後概念ベースは著しく属性数が減っている．

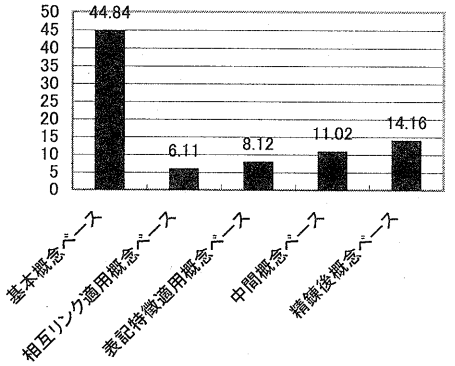

図 5. 平均属性数

また，図7には精錬後概念ベースの属性分布を示すが精錬後概念ベースにおいてはほとんどの概念が属性数10個以内とわかる．

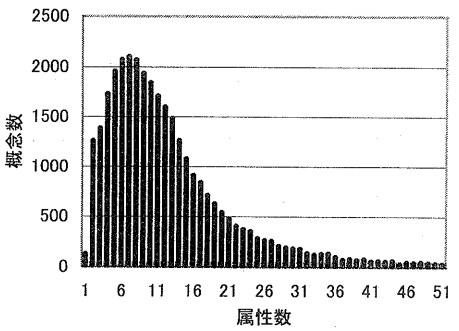

図 6. 精錬後概念ベースの属性分布

4.2 人間の判断による精錬手法の検証と評価

精錬操作によって，概念が持つ属性は図6を見ても分かるとおり大きく変化した．雑音除去，属性追加の影響を見るため表1を利用して精錬後概念ベースについての属性評価を行った．

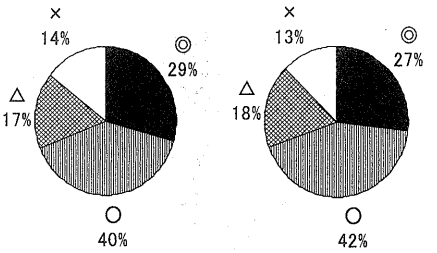

図 7. 中間概念ベース（左）と精錬後概念ベース（右）の属性評価

図 8 より精錬後概念ベースは精錬前の基本概念ベースと比較して，不適切な属性が約半分に適切な属性が約7割を占めている．

4.3 関連度

概念間の関連性を定量的に評価する方法として関連度を定義する．例えば「米」と「飯」の関連性と「米」と「機械」の関連性では人間ならば前者の方を関連性が強いと判断する．この判断をコンピュータ上で実現するために，関連性を関連度と呼ぶ値で定義している（図5）．

Rel(米,飯)=0.51

Rel(米,水)=0.10

Rel(米,機械)=0.01

Rel(A,B)：概念Aと概念Bの関連度

（ただし，$0 \leqq Rel(A,B) \leqq 1$）

図 8. 関連度の例

本稿では，概念間の関連度を2つの概念における属性集合の一致度合で定義しており，計算方法として概念連鎖計算方式を用いている[2]．具体的には，概念A，Bの関連度として2次の関連度 $Rel_2(A, B)$ を次式で求めている．

$$A=\{a_1,a_2,\cdots,a_i,\cdots,a_m\}$$

$$B=\{b_1,b_2,\cdots,b_j,\cdots,b_n\}$$

$$Rel_d(A,B) \overset{def}{=} \begin{cases} \dfrac{m+n}{2mn}\sum_{k=1}^{n} Rel_{d-1}(a_k,b_k) & (if\ d \geq 2) \\ \dfrac{m+n}{2mn}eq & (if\ d=1) \end{cases}$$

（ただし，概念A，Bの属性は $Rel_d(A,B)$ が最

大になるように並び替えており，eqは概念A，Bの属性の一致数である）

4.4 関連度による精錬手法の検証と評価

本節では関連度を用いて概念ベースを評価する．評価方法は表2のように，4つの単語の組からなる評価尺度を用意し，その1つを基準概念M_xとし，残り3つの単語については基準概念と関係が深い概念M_a，関係のある概念M_b，関係のない概念M_cを用意する．基準概念とこれら3つの概念との関連度R_a，R_b，R_cの大小を比較し$R_a>R_b>R_c$となったとき，概念ベースはこの組を正しく解釈したと考える．

表2．評価尺度の例

Mx	Ma	Mb	Mc
ご飯	飯	米	青空
安易	簡易	気持ち	経済
意図	志向	内心	帰宅
飲料	飲み物	喉	反省
羽	翼	鳥	返還
⋮	⋮	⋮	⋮

M_a:概念M_xの同義または類義の概念

M_b:概念M_xと関係のある概念

M_c:概念M_xと全く関係のない概念

図9は，評価尺度559組に対する基本概念ベース，基本概念ベースに相互リンクルール，表記特徴ルールを適用して雑音除去した基本概念ベース，中間概念ベース，精錬後概念ベースの正解率である．図からわかるとおり，相互リンクルール，表記特徴ルールそれ単独では概念ベースの質が大幅に落ちている．だが両者の組み合わせによる中間概念ベースは基本概念ベースとほぼ正解率が同じである．これよりこの2つのルールは基本概念ベース内の雑音除去に大きく寄与していると言える．また，精錬後概念ベースは基本概念ベースと比較して平均属性数については70%以上減少しているのに約2%の品質向上が見られる．これは相互リンクの拡充により概念により適切な属性が追加されたためである．なお，基本概念ベースについては属性数を最大30個として関連度計算を行っている．これは概念ベースにおいて一概念の属性を重み順で最大30個採用して関連度計算を行ったときに正解率が一番高かったためである[3]．

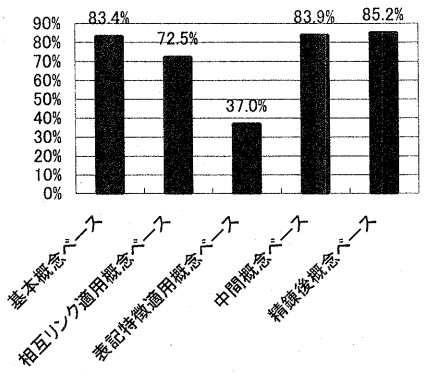

図 9．精錬手法に対する解釈成功率の推移

図10は評価尺度において，$|R_a-R_b| \leqq 0.06$となる，すなわち関連の度合の差が小さいと考えられる94組についての解釈成功率である．このような基本概念ベースにおいて関連度が接近している評価尺度の組を難領域と呼ぶ．精錬後概念ベースはこのような基本概念ベースでは正解率の上がらない組に対して高い解釈成功率を上げ，精錬の効果が大きく表れている．また．図11ではこの94組に対するR_a，R_b，R_cの平均を示すがR_a，R_bの差が精錬によって大きくなりより人間の判断に近づいている．

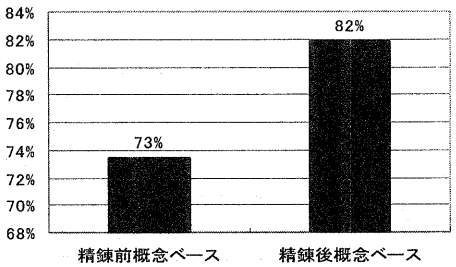

図 10．難領域での解釈成功率

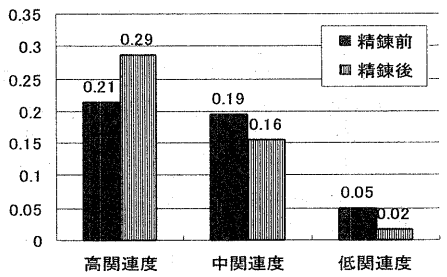

図 11．難領域での評価尺度の関連度

以上よりルールを用いた概念ベースの精錬はよりコンパクトな記憶容量とより正しい解釈がなされているという点でうまくいっているといってよい．しかし，今回の精錬では精錬前も判断できず精錬後にも改善されなかった失敗例，精錬前では判断が出来ているが精錬後になると判断できなくなる改悪例が表れた．以下にその例を示す．なお，評価尺度の組において上に示しているのが精錬前，下に示したのが精錬後の関連度である．失敗例については概念ベースの構築方法に問題があると考えられる．今回の精錬方法では精錬前に失敗して精錬後に成功した評価尺度の組もあるが，精錬だけではこれらの失敗を全て成功に変えることは出来なかった．表4ではいずれも精錬後の関連度 R_a，R_b が逆転してしまっている．これは M_x，M_a の共通属性が失われたと考えられる．これらについては以後の研究課題とする予定である．

表3．精錬失敗例

	M_x	M_a	R_a	M_b	R_b	M_c	R_c
精錬前	魚	鰯	0.229	水	0.070	蜜柑	0.151
	労働	勤労	0.306	徹夜	0.043	本	0.046
	道	道路	0.295	建設	0.027	海	0.051
精錬後	魚	鰯	0.145	水	0.049	蜜柑	0.127
	労働	勤労	0.402	徹夜	0.012	本	0.0198
	道	道路	0.527	建設	0.026	海	0.049

表4．精錬改悪例

	M_x	M_a	R_a	M_b	R_b	M_c	R_c
精錬前	疾患	疾病	0.334	慢性	0.267	体罰	0.047
	判別	判断	0.156	区分	0.120	信仰	0.054
	優秀	優等	0.182	受賞	0.082	設置	0.063
精錬後	疾患	疾病	0.186	慢性	0.318	体罰	0.028
	判別	判断	0.124	区分	0.147	信仰	0.012
	優秀	優等	0.061	受賞	0.078	設置	0.026

5．おわりに

本稿では，コンピュータにより人間に近い知的判断を行わせるという目的において，その中核をなす汎用的な知識資源「概念ベース」をいかに自動構築，精錬するかについて述べた．概念ベースは膨大な概念に対応しなければならないので，手作業で作成するのは現実的ではない．そのために機械的な構築が必要となってくるが，機械構築では人手に比べて質が劣るのは否めない．機械構築された質の悪い概念ベースを概念が相互に持つ本質的な特徴をルール化することで概念ベースの質向上が図れることを示した．して不適切な属性を削除，適切な属性を追加することで，より品質の高まった概念ベースが構築できることがわかった．概念間の関連性を数値で表す関連度を用いた評価実験でも品質の向上が確認できた．

今後は，概念ベースの知識をより精密化する方法，外部の知識資源を用いた概念ベースの拡充，様々な知識資源への対応の方法などを検討する予定である．

謝辞

本研究は文部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「知能情報科学とその応用」における研究の一環として行った．

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] 入江　毅，渡部　広一，河岡　司，松澤　和光：知的判断メカニズムのための概念間の類似度評価モデル，信学技報，Vol.98，No.499，AI98-75，pp.47-54(1999)

[3] 入江　毅，渡部　広一，河岡　司：概念ベースにおける属性数の検討と概念間の関連度計算方式，信学技報，Vol.99，No.534，AI99-82，pp.37-44(2000)