知能と複雑系　115 － 4
（1999. 1. 11）

言葉の意味に関する階層型大規模概念ベースの構築

帆苅 譲† 石川 勉† 笠原 要††

†拓殖大学工学部情報工学科 ††NTTコミュニケーション科学研究所

†〒193－8585 東京都八王子市館町 851－1
Tel: 0426－65－1441(内)5623
E-mail: yhokari@cs.takushoku-u.ac.jp

††〒619－0237 京都府相楽郡精華町光台 2-4
Tel: 0774－93－5353
E-mail: kaname@cslab.kecl.ntt.co.jp

あらまし

言葉の意味の類似性を判別するための知識ベース(概念ベース)の構築を進め、これまで、約4万語の単語からなるシステムを試作、評価してきた。この結果、新聞記事検索等への具体的な応用においては概念数の不足が問題となることが分かった。このため、今回、この概念ベースを数十万語規模へ拡張するとともに、漢字から構成されるあらゆる単語についても近似的な類似性判別を可能とした。さらに、同義語判定を行うための同義語辞書を構築した。本概念ベースは、従来の4万語の部分、広辞苑等から作成した26万語の部分、漢和辞典をもとに漢字一字ごとの意味をあつめた部分の3階層構造からなる

キーワード　概念ベース、類似性判別、階層構成、同義語

A Hierarchical Large-scale Knowledge Base for Measuring Semantic Similarity between Words

Yuzuru Hokari†, Tsutomu Ishikawa†, Kaname Kasahara††

†Department of Computer Science,
Faculty of Engineering,
Takushoku University

††NTT Communication
Science Laboratories

†851-1 Tatemachi Hachiouji Tokyo, 193-8585
Tel: 0426-65-1441
E-mail: yhokari@cs.takushoku-u.ac.jp

††2－4 Hikaridai SeikacyouSagaragun Kyouto
Tel: 0774－93－5353
E-mail: kaname@cslab.kecl.ntt.co.jp

Abstract

We have constructed a hierarchical large-scale knowledge base of words for measuring semantic similarity between words. This system consists of the basic knowledge base (GB1) of 40,000 Japanese daily-used words, the extended knowledge base (GB2) of 260,000 words based on Japanese-Dictionaries (*KOUJIEN*, *DAIJIRIN*) and the knowledge base (GB3) of Chinese characters based on a Chinese-Japanese dictionary, and the synonym dictionary (SD) of 7,000 words based on a Japanese dictionary. In the GB1, each word is represented by a series of weighted keywords. The keywords have some relationship with the word, and the weights of the keywords represent the degree of the strength of the relationship between the word and keywords. In the GB2, each word is represented by the word that is the most similar word to it in the GB1. The GB3 is used to form the series of weighted keywords of the word which is not contained in the GB1 and GB2. The SD consists of the number of groups used for judging whether or not two words are synonymous each other.

key words　knowledge base , similarity , hierarchical structure , synonym

1 はじめに

知識が不完全でも、その不完全さに応じた概略的な判断が行えるシステムの実現を目ざし、研究を進めている。ここでは、不足知識を常識知識で補足することを想定しているが、その一つとして単語の意味の類似性を判別するための知識ベース(以下、概念ベース[1])の構築を行ってきた。単語に関する知識ベースとしては国内ではEDR辞書[2]やIPAL辞書[3]が、米国ではCYC[4]が代表的であるが、これらは用途を限定せず汎用性を意識し人手により作成されている。これに対し、我々は用途を単語の意味の類似性判別に限定し、電子化された国語辞書等の文書から機械的に概念ベースを構築してきた。

現在、この概念ベースは約4万語の単語(以下、概念と呼ぶ)で構成されており、内蔵する単語間の類似性判別については一定の性能が確認されている[1]。しかし、新聞記事検索等への具体的な応用においては単語数の不足が問題となることが分かった。従って、これら4万語以外のあらゆる単語に対しても近似的な類似性判断が行えるように、大規模化を図ることとした。具体的には、新たに26万語規模の概念ベースと、新語や造語への対応を考慮した漢字概念ベースを作成した。さらに、類似性判別能力の向上を図るために、同義語辞書を作成した。これらを従来の4万語の概念ベースと統合し、階層構造化することで一つの大規模概念ベースを構築した。また、本概念ベースでは以上の大規模化に加え、幅広い応用、WSからPCへの移植等を考慮し、コンパクト化と高速化を図ることとした。

本報告は、この大規模概念ベースの構成と評価について述べたものである。以下、2章で概念ベースの基本的な構成と機能について、3章で大規模概念ベース構築の考え方およびそのシステム構成と動作について、4章で各部の具体的な構成と構築法について、5章で類似性判別能力についての評価と収容に必要な記憶容量および速度性能について述べる。

2 概念ベースの基本構成

概念の知識表現としては、その属性と属性値の集合として表現する方法が一般的である。たとえば、「りんご」という概念は、{(形:丸い),(色:赤い),(味:すっぱい),……}と表現される。このような属性と属性値の集合をあらゆる概念について獲得できれば、より強力な概念ベースとなる。しかし、実際にこのような概念の集合を機械的に獲得することは現在の技術では極めて困難である。そこで、我々は、電子化された辞書等からの機械的な獲得を前提に、できるだけ単純な形式で概念を表現することとした。

具体的には、国語辞書等の見出し語(概念とする)の語義文について形態素解析を行い、そこに含まれる自立語を抽出し、それを日英翻訳システムALT-J/E[5]上のシソーラスにおける2715のカテゴリに対応させ属性とする。属性値は、その自立語が語義文中で多く使われる程、対象となる概念との関連が強いと考え、基本的にはその出現頻度をもとに算出する。上記のように構築された各概念gは以下のように表現される。

$$g=\{(p_1,q_1),(p_2,q_2),\cdots,(p_i,q_i),\cdots(p_m,q_m)\} \quad (1)$$

$$\text{ただし} \quad 0\le q_i\le 1 \quad \sum_{i=1}^{m} q_i^{\,2}=1$$

ここで、p_iは属性、q_iはその属性値である。また、mは、属性数である。属性値は概念ごとにその二乗和が1になるように正規化している。たとえば、概念「りんご」は、以下のように表現されている。

りんご={(果実:0.328),(花:0.277),(高木:0.272),……}

以上の形式で表現しているため、各概念はそれと関連しない属性の属性値を0とみなせば、2715次元のベクトルとして扱うことができる(以後(1)式の形式の表現を概念ベクトルと呼ぶことにする)。そのため、二つの概念間の類似度はこれらベクトル間と何らかの演算をほどこすことにより容易に演算できることになる。また、類似性を考えるとき、何らかの視点を想定することがある。たとえば、「馬」に対して「豚」と「自動車」のどちらが類似しているかでは、視点が「動物」であれば「豚」が、それが「乗り物」の場合は「自動車」が似てくる。このような視点(以下観点と呼ぶ)を考慮した類似度についても、その二つの概念ベクトルを観点によって変形することにより容易に算出することが可能となる。

3 大規模概念ベースのシステム構成と動作概要

3.1 設計方針

大規模概念ベースは2章の考え方を基本に、さらに以下の設計方針のもとに構築する。

(1)大規模化

具体的な応用を想定して大規模化を図るが、単に単語数を増やすだけでなく新語や造語にも対応可能し、あらゆる単語に対して類似性判別を行えるようなシステムの

実現を目指す。

(2)類似性判別能力の向上

従来の4万語概念ベースの構成の見直しを行うとともに、意味的に同一の概念である同義語の判定を可能にする。

(3)記憶容量のコンパクト化

広範囲な分野で利用可能とするために、さらにはWSだけでなくPC上にも収容可能となるようにコンパクト化を図る。特に主記憶への常駐を可能とするように、総容量として数十MB程度に収めることを目指す。。

(4)処理速度の短縮

(3)で述べた主記憶への常駐等により、処理時間の短縮を図る。処理速度としては、類似度の計算をミリ秒のオーダで行えることを目指す。

3．2 システム構成

本システムは、従来の約4万語の概念ベースから再構成した基本概念ベース(GB1)、新たに構築した約26万語の拡張概念ベース(GB2)、単語の構成要素である漢字一字ごとの約6千字の漢字概念ベース(GB3)からなる3階層構造をとる。さらに、同義語判定を行えるように、約7千の概念(単語)から構成される同義語辞書(SD)を付加する。図1にシステム全体の構成を示す。これらの各概念ベース及び同義語辞書では、各概念は次のように表現する。

GB1：2章の(1)式の形式で表現。

GB2：$(g_i : g_i')$という組で表現。ここで、g_iはGB2中の概念で、g_i'はg_iに最も類似するGB1中の概念(代表概念と呼ぶ)である。

GB3：漢字一文字ごとに、GB1と同様(1)式の形式で表現。

SD　：同義語どうしを、一つのグループとして表現。

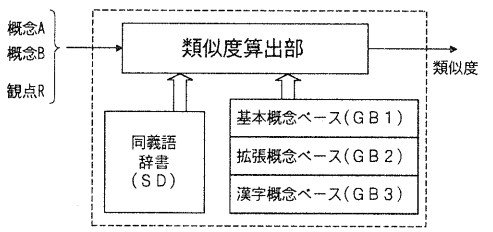


図1　階層型大規模概念ベースの構成

3．3 動作

本システムは、入力として概念A,Bおよび観点Rを与えて、Rに着目したときのAとBの類似度を出力する。Rを入力しないときは単純にAとBの類似度を出力する。出力形式は$(d_1 : d_2 : d_3$：類似度)とし、d_1, d_2, d_3はA,B,Rがどの概念ベースの概念ベクトルを用いたかを示す。具体的には、これらはそれぞれ、GB1のときは“1”、GB2のときは“2”、GB3のときは“3”とする。さらに、SDのときは“S”、見つからなかったときは“−”とする。類似度は0〜1の実数で表現し、数値がより1に近いほど類似していると判断する。

類似度Sは、概念A,Bが同義語辞書内の同一グループであるときは1とし、その他の場合は、概念ベクトルを利用して算出する。具体的には、以下に示すようにA,Bをn次元ベクトルとして表現し、その内積として算出する。

$$S = A \cdot B = \sum_{i=1}^{n} q_{Ai} * q_{Bi} \qquad (2)$$

ここで　$A = (q_{A1}, q_{A2}, \cdots, q_{Ai}, \cdots, q_{An})$

$B = (q_{B1}, q_{B2}, \cdots, q_{Bi}, \cdots, q_{Bn})$

また、観点を入力したときは、観点Rの属性中でその属性値が一定値以上の属性と一致する、概念A,Bの属性の属性値を強調(これを変調と呼ぶ)して算出する。具体的には、観点$R(r_1, r_2, \cdots\cdots, r_i, \cdots\cdots, r_n)$とベクトル表現したとき、Rで変調した概念AのベクトルA'は、以下のように求める。

$$A' = (q'_{A1}, q'_{A2}, \cdots, q'_{Ai}, \cdots, q'_{An}) \qquad (3)$$

ただし、

$$q'_{Ai} = \begin{cases} \omega * q_{Ai} & (r_i \geq Q(R) \quad \text{のとき}) \\ 1 * q_{Ai} & (r_i < Q(R) \quad \text{のとき}) \end{cases}$$

$$Q(R) = \frac{s}{\sqrt{atnum(R)}}$$

上式で、$atnum(R)$は、Rの属性値r_iの中で0でない属性の数であり、sとωは、実験的に決定される値であり、ここでは$s = 0.5$，$\omega = 5$とした(詳細は文献[6])。概念Bについても同様の方法で変調後のベクトルB^vを算出し、(2)式のようにA^vとの間で内積をとり類似度とする。

3．4 概念ベクトルの取得法

類似度計算において概念xの概念ベクトルは、以下の

手順で取得する(観点に対しても同様)。なおx_sは、同義語辞書でxが属するグループに含まれる同義語とする。

[概念ベクトル獲得手順]

i) $x \in$GB1のときは、GB1中のxのベクトル。

ii) $x \notin$GB1 ∧ $x \in$SD ∧ $x_s \in$GB1のときは、GB1中のx_sのベクトル。

iii) $x \notin$GB1 ∧ $x \notin$SD ∧ $x \in$GB2のときは、代表概念x'のベクトル。

iv) $x \notin$GB1 ∧ $x \notin$SD ∧ $x \notin$GB2のときは、概念xを構成する各漢字のベクトル(GB3)より合成。

以上の手順で、概念ベクトルを得るので、観点を用いた類似度計算に必要な処理速度Tは、以下の式で表現される。

$$T = 2t_s + 3\{(1-\alpha_s)t_1 + (1-\alpha_s-\alpha_1)(t_s+t_1)m + (1-\alpha_s-\alpha_1-\alpha_1')t_2 + (1-\alpha_s-\alpha_1-\alpha_1'-\alpha_2)t_3\} \quad (4)$$

ここで、t_1, t_2, t_3は、それぞれGB1, GB2, GB3での概念ベクトル取得のための平均処理時間、t_sはSDでの概念検索のための平均処理時間、α_1, α_2は、それぞれGB1, GB2から概念ベクトルが取得される確率、α_sは概念がSDに存在する確率、α_1'はx_sがGB1から取得される確率、mはii)においてx_sを得るまでのSDおよびGB1に対する平均的な検索回数である。なお、(4)式において$2t_s$は概念A,BのSDでの検索時間であり、中括弧内での各項は、概念ベクトル取得手順の①～④にそれぞれ相当する時間である。

4 各部の構成

4．1 基本概念ベース(GB1)

従来の基本概念ベースでは、各概念の属性数(2章の(1)式のm)は1～533(平均69)とばらついている。ここでは、概念ベースのコンパクト化、処理時間の短縮のために属性数を均一化する。

このとき類似度計算において生じる誤差について以下に述べる。対象となる概念をg_1, g_2とする。まず一方の概念g_1のベクトル要素q_i (i =1～m)をソートした結果を、

$$q_1' \geq q_2' \geq \cdots\cdots q_i' \geq \cdots\cdots \geq q_m'$$

とする。このとき属性数をiで均一化したときの誤差δは、類似度をベクトルの内積で計算しているため下式で表せる。

$$\delta = \sum_{j=i+1}^{m} (q_j' * r_j) \quad (5)$$

ここで、r_jは概念g_2の概念ベクトルの要素で、q_j'の属性と同一の属性の属性値である。この誤差δは、g_2において$r_{i+1}=1$で、それ以外が0になるとき最大となる。すなわち、誤差の最大値(δ_{MAX})は$\delta_{MAX} = q_{i+1}$となる。

一方、概念ベースの応用について考えたとき、類似度の高い概念に対して何らかの処理を行うことがほとんどであると考えられる。実際、この概念ベースでは、互いに類似する概念間では、その類似度が0.5程度となる。従って、類似度に0.1程度の誤差が生じたとしても、問題は少ないと考えられる。

図2に均一化属性数と誤差の関係を示す。これは、属性数をiに均一化したときのGB1中の全概念(4万語)についてそれぞれδ_{MAX}を実測し、それらのδ_{MAX}の中で最大のものを示した結果である。すなわち、この値はGB1中の全ての概念に対して類似度を求めたときに、誤差はこれよりも大きくならないことを意味する。

この結果から、i = 50と設定して均一化した。このとき誤差の最大値は0.07である。

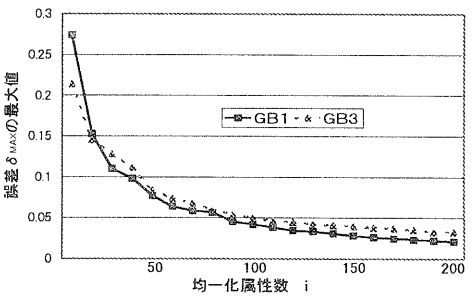


図2 属性数の均一化における類似度の誤差

4．2 拡張概念ベース(GB2)

まず、大辞林 [7]と広辞苑 [8]を利用し2章で示した方法で、GB2のもととなる26万語の概念ベースを構築した。次に、この概念ベース中の各概念(26万語)について、GB1の約4万の概念との間の類似度を求め、その中で最も類似度の高い概念を代表概念として抽出し、3.2節で示したような表現の代表概念を用いた概念ベースとした。GB2で、このような代表化の手法を用いたのは、3.1節で述べたコンパクト化のためである。

4．3 漢字概念ベース(GB3)

漢和辞典 [9]をもとに GB1 とほぼ同様の方法で構築した。ただし、従来は語義文のみを用いたのに対し、GB3 では約 6 千字について、語義文だけでなく説明文や例文などすべてを用いた。このように作成した概念ベースでは、属性数は平均 493(4～1260)となった。従って、4.1 節の GB1 と同様に属性数の均一化を行った。ただし、概念ベクトルの合成に伴う誤差の増大を考慮して、均一化する属性数は、GB1 の二倍の 100 とした。図 2 の破線に、4.1 節と同様に実測した漢字一文字の類似度の誤差と均一化属性数の関係を示す。

GB3 では概念ベクトルは以下のように合成する。まず、概念(単語)を漢字に分解し、各漢字毎の概念ベクトルを GB3 より抽出する。次に、抽出した各漢字の概念ベクトルの中で属性が同一のものどうしの属性値を加算することで、概念ベクトルの合成を行う。最後に、この合成ベクトルに対して、属性数を 50 に削減して正規化し、単語の概念ベクトルとする(図 3)。

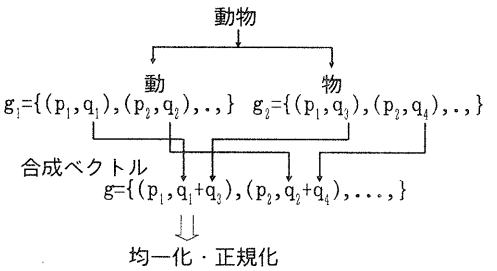


図 3 漢字の概念ベクトル合成法

4．4 同義語辞書(SD)

国語辞典 [10]を利用し、見出し語と語義文の関係に着目して構築した。具体的には、辞書内の語義文には、単語一語からなる文が存在している。すなわち、各見出し語に対する語義文中での各文を形態素解析したときに、それ以上、品詞分解できない文がある。例えば、図 4 に示すように「いぎ【意義】」という見出し語に対して「わけ」「意味」、また「いみ【意味】」という見出し語に対しては、「意義」である。

このような単語は、一語でその見出し語を説明しているはずなので、同義語である可能性は高いと考えられる。そこで、このような組を抜き出し、同義語の候補群(図 4)とした。

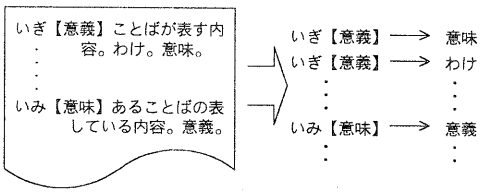
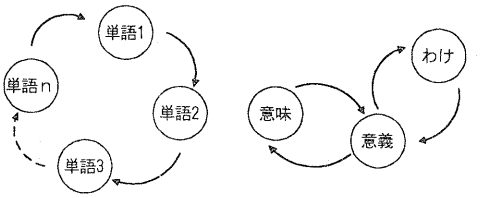


図 4 辞書からの抽出

次に、こうして得た同義語の候補群に対して、図 5(a)に示すようなループとなる組み合わせをすべて抜き出す。図 4 の例では、「意義」と「意味」の関係がそれにあたる。このようにループとなる組み合わせの中で、共通の単語を含むループ(図 5b)どうしをまとめることで、一つの同義語のグループとする。このグループ化をすべてのループに対して行った結果、2986 のグループが得られた。



(a)多重ループ (b)グルーピング

図 5 同義語の取得

次に、人手により、これらの中から実際には同義語関係にないグループを削除し、最終的に 2657 のグループ(平均概念数 2.4)の同義語辞書 SD を作成した。なお、この同義語辞書は、同義の判別だけでなく、3.4 節の概念ベクトル取得の ii) に示したように利用され、実効的に GB1 の規模を拡大させるよう機能する。

なお、同義語として判定できなかったグループは、多義語を介してグループ化されたものが多かった。一例として、「音盤」「レコード」「記録」というグループがあげられる。「音盤」と「レコード」は同義であり、「レコード」と「記録」もまた同義である。しかし、「レコード」が多義語であるために、これらは同義語のグループとはならない。

5 評価

5．1 評価法

ここでは、文献[11]で提案した類語辞典[12]を利用する方法により評価した。この評価法は、概念ベースの特性として以下に述べることが重要になることを考慮した方法である(図 6)。

i) 類似する概念との間の類似度と全く類似しない概念との間の類似度の差が大きい。

ii) 二つの候補概念があったとき、どちらが対象概念に似ているかを識別可能。

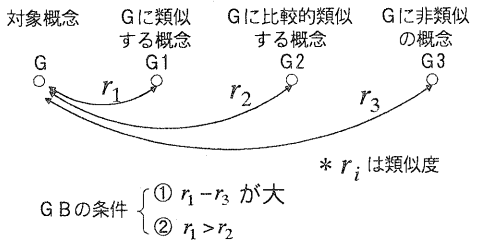


図6 概念ベースに必要な特性

具体的には、対象概念 G(サンプル概念)に対し、これと類似する概念G1、比較的類似する概念G2、非類似概念G3の組を用い、以下の評価指数F_dで評価する。

$$F_d = F_1 * F_2 \qquad (6)$$

ここで $F_1 = (R_1 - R_3)/(\sigma_1 + \sigma_3)$

$F_2 = 1/(1 + wg)$

ここで、R_1, R_3は、図6におけるr_1, r_3の全サンプル概念に対する平均値であり、σ_1, σ_3は、同じくr_1, r_3の標準偏差である。また、wgは、全サンプル概念でr_1とr_2の類似度の大小関係が反転した数である。上式のF_1, F_2は、それぞれi), ii)の特性に対応するので、F_dの値が高いほど、類似性判別能力が高いことになる。

5．2 評価結果

サンプル概念として、以下の表1に示すような100組のデータを用い評価した。この場合、概念ベースに類似性判別能力が全くないとすると、$R_1 = R_3, wg = 50$であるから、F_1, F_2はそれぞれ0，0.02となる。

表1 サンプル概念

G	G1	G2	G3
全力	総力	人力	鉱物
視野	視界	視線	戦死
歓声	歓呼	叫ぶ	貯蓄
近道	早道	通行	学友
⋮	⋮	⋮	⋮

(1) 基本概念ベース (GB1)

GB1の属性数を50に均一化したことの妥当性について実験的に検討した。具体的には、まず、従来の属性数が不均一な4万語の概念ベースから、それぞれ属性数10,20,……,190,200で均一化した概念ベースを作り、5.1節のF_dで評価した。結果を図8に示す。属性数を50で均一化したときF_dは最大となり、4.1節で設定した値が適切であることがわかった。

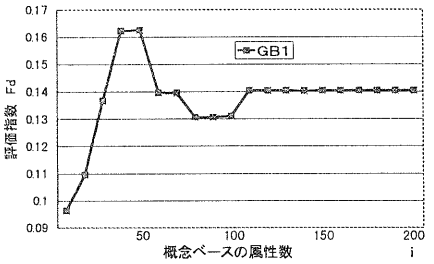


図7 属性数における評価指数の推移

また、従来の属性数が不均一な概念ベース(GB1のもととなった概念ベース)と属性数50で均一化を行ったGB1の性能は、それぞれF_dの値で0.14，0.16であった。後者の性能が若干向上しているが、原因は、均一化により、属性値が小さい属性(ノイズともなりうる)を削除できるためであると考えられる。

(2) 拡張概念ベース (GB2)

GB2のもととなった概念ベースと代表概念を用いるGB2について性能評価を行った。

具体的には、まず、各サンプル概念をGB2中から検索し、その概念の代表概念(GB1中の概念)を得る。次に、その代表概念をGB1中から検索し、概念ベクトルを取得する。類似度は、このように取得した概念ベクトルを使うことで計算した。その結果、F_dの値は、もとの概念ベースで0.08、GB2で0.04となった。F_dはかなり低下しているが、F_2についてみると0.03(wg=29)であった。また、r_1とr_3の大小が反転した数、r_2とr_3の大小が反転した数は、それぞれ、100組中10，11組であった。従って、ある程度の類似度判別能力は得られていると言える。

次に、代表概念を一つでなく、複数個(n個)用いる場合について評価した。この場合には、対象概念に類似する概念を類似度の高いほう方からn個選び代表概念とする。nが二以上のときは、n個の代表概念について、4.3節と

同様の方法で合成して概念ベクトルを得ている。n を 1～10 と変化させたときのF_d の値を図 8 に示す。

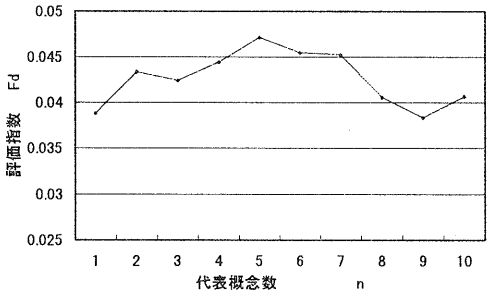


図 8　代表概念数による評価指数の推移

同図から、代表概念を複数化した方が若干性能が向上することがわかる。しかし、設計方針の一つである概念ベースのコンパクト化、高速化を考慮し、最終的には代表概念数は 1 のままとしている。

（3）漢字概念ベース（GB 3）

GB3 の属性数を 100 に均一化したことの妥当性について検討する。それぞれのサンプル概念に対して GB3 を用い、4.3 節で示したように概念ベクトルの合成を行い評価した。図 9 に、均一化する属性数を変化させたときの評価指数F_d を示す。属性数を 100 で均一化したときのF_d は 0.024 となり、それ以降、属性数を増やしてもF_d の値は変化しないことから、属性数 100 は妥当であると判断できる。

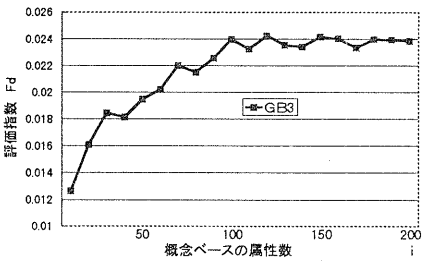


図 9　属性数における評価指数の推移

また、GB3 のもととなった属性数が不均一な漢字概念ベースと均一化した GB3 の性能は、それぞれ、F_d の値で 0.020，0.024 であった。後者のF_d が高い原因としては、GB1 と同様に、不要属性が削除されたためであると考えられる。

GB3 のF_d の値は GB1 と比較するとかなり数値的には低くなっているが、F_2 についてみると 0.029(wg =33)であり、さらに、均一化した属性数が 100 のときのr_1 とr_3 の大小が反転した数、r_2 とr_3 の大小が反転した数は、100 組中それぞれ、13，27 であった。従って、ある程度の類似性判別能力は得られていると言える。

以上のF_d による評価では感覚的な性能を把握しづらいため、一例として、概念“夢想”に対して、各概念ベースを用いて最も類似する概念の上位 20 個を抽出した。具体的には、最初に GB1,GB2,GB3 の各概念ベースから、概念“夢想”の概念ベクトルを 3.4 節に示したようにそれぞれ取得する。次に、GB1 の約 4 万の概念との間で類似度を求め、GB1 から類似度の上位 20 概念をそれぞれ抽出した。

この結果を表 2 に示す。GB1,GB2,GB3 の順で非類似の概念が多くなっていることが分かる。

表 2　各ＧＢにおける上位 20 概念

GB1	GB2	GB3
夢想	夢	想う
夢見る	襲われる	不了見
白昼夢	夢見	念
そら夢	正夢	所懐
夢	そら夢	心事
空想	夢幻	所在
夢見	夢路	憧れる
襲われる	白昼夢	考え
夢幻	逆夢	夢見る
正夢	悪夢	思う様
夢路	迷夢	思慮
逆夢	夢見る	抱懐
夢語り	夢現つ	会心
想像	夢想	唯心
夢現つ	夢語り	感懐
悪夢	夢心地	疑う
迷夢	夢見心地	迷夢
幻想	初夢	心から
理想的	夢中	低意
理想主義	空想	縦横無尽

なお、ALT - J/E[5]のシソーラスを用い類似度を求めたときの評価指数F_d を求めると 0.36 となった。類似度は 1/(1+distance)により求め、distance はシソーラス上での概念の距離である。シソーラスを用いたほうが、F_d の値は高くなっているが、シソーラスは人手で作成したものであり、一面では当然の結果といえる。しかし、単純な 2 概念間の類似性判別でなく、観点を用いた場合に対する評価では、従来の 4 万語の概念ベースでも約 2 倍優れていることが報告されている[6]。

5．3 容量と速度

表3に各概念ベースにおける容量と速度を示す。ここで、GB1,GB3 は属性数を均一化したときの値であり、GB2 は概念の代表化を行ったときの値である。なお、括弧内はこれらを行う前の容量と速度である。

全体の概念ベースの総容量はテキスト形式で約46MB(実際の主記憶上では約 20MB)であり、PC の主記憶上に常駐させることが可能となった。また、処理速度は、各概念ベースが一回の類似度計算に要する時間であり、PC9821-V200(pentium 200)上で実行した結果である。

表3　容量と速度

	容量(Mﾊﾞｲﾄ)	速度(秒)
GB1	27(57)	0.03(0.69)
GB2	5(62)	0.02(25.2)
GB3	14(66)	0.04(0.31)
SD	0.11(-)	0.01(----)

6 むすび

4万語の基本概念ベース、26 万語の拡張概念ベース、6 千字の漢字概念ベース、7 千語の同義語辞書 SD から構成される、階層型大規模概念ベースについて述べた。本システムは、概念の追加で大規模化を図るとともに、新語、造語に対しては、漢字概念ベースを用いることで、あらゆる単語に対する類似性判別を可能としている。

また、本システムの類似性判別能力について評価した結果、拡張概念ベースと漢字概念ベースは、従来の4万語の基本概念ベースに対して性能は若干劣るが、類似、非類似の大まかな判断については充分な能力があることが分かった。

本システムでは、属性数の均一化や概念の代表化により、収容のための記憶容量の削減を図り、PC の主記憶上に常駐させることを可能としている。

今後は、拡張概念ベースと漢字概念ベースの類似性判別能力を改善し、システム全体の性能向上を図っていく予定である。

[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, Vol38, No7, pp1272－1283(1997)

[2]横井 俊夫, 仲尾山雄, 荻野 孝野, 田中 裕一：概念レベルにおける電子化辞書の情報構造, 情報処理学会論文誌, Vol.38, No1, pp32－43(1997)

[3]青山 文啓, 橋本 三奈子：名詞の辞書記述, 情報処理学会自然言語処理研資, Vol.94－104, pp9－16(1991)

[4]Guha, R.V. and Lenat, D.B.：Cyc：A midterm report, AI Magazine, Vol11, No3, pp32－59(1990)

[5]池原, 宮崎, 横尾：日英機械翻訳のための意味解析辞書, 情報処理学会自然言語処理研資, Vol84－13, pp95－102(1991)

[6]笠原 要, 松澤 和光, 石川 勉, 河岡 司：観点に基づく概念間の類似性判別, 情報処理学会論文誌、Vol.35, No3, pp.505－509(1994)

[7]松村 明(編), 三省堂編修所：大辞林第二版, 三省堂 (1995)

[8]新村 出(編)：広辞苑第四版, 岩波書店 (1991)

[9]藤堂 明保, 松本 昭, 竹田 晃 (編)：新版漢字源, 学習研究社(1994)

[10]金田一春彦, 池田弥三郎(編)：学研国語大辞典第二版, 学習研究社(1988)

[11]石川 勉, 井澤 潤次朗, Nguyen Viet Ha, 笠原 要：単語の意味に関する概念ベースの類似性判別能力からの最適構成, 人工知能学会誌, Vol13, No3, pp470－479(1998)

[12]大野晋, 浜西正人：類語国語辞典, 角川書店(1990)