# 単語の意味の類似性判別のための大規模概念ベース

グエン・ベト・ハー[†] 帆 苅 譲[††]
石 川 勉[†] 笠 原 要[†††]

概念を表す基本単位である単語間の意味的な類似性を判別するため，単語（以下，概念と呼ぶ）に関する知識ベース（概念ベースと呼ぶ）の構築を進めてきた．この概念ベースでは，各概念は，国語辞書の語義文から獲得される自立語を属性，その出現頻度を属性値として表現され，概念間の類似度は共通属性の比較により算出される．これまで，約 4 万の日常語からなる概念ベースを構築してきたが，具体的な応用においては概念数の不足や性能的な不十分さが問題となった．本論文では，語義文における属性の重要性およびシソーラス上の属性間の関係を考慮した属性値算出手法を提案し類似性判別能力の向上を図る．また，新語や造語を含めたあらゆる概念に対し類似性判別を可能にするため，一般的な単語に関する概念ベースのほか，同義語および漢字概念ベースからなる大規模概念ベースの構築を行った．また，これに対し，類似性判別能力等を評価した結果，従来の概念ベースに比べ大幅な性能向上が確認された．

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

Nguyen Viet Ha,[†] Yuzuru Hokari,[††] Tsutomu Ishikawa[†]
and Kaname Kasahara[†††]

A method for measuring the semantic similarity between words using a type of knowledge bases was proposed. Each word in this knowledge base is represented by a list of weighted keywords that automatically acquired from machine-readable dictionaries. A prototype knowledge base of about 40,000 Japanese words was constructed. However, experiments showed that the number of words and the ability of similarity judgment are not enough for practical applications. This paper describes a large-scale knowledge base, which contains a large number of common words, Chinese characters and synonyms in order to deal with every word including both new words and coined words. Moreover, new methods to acquire and modify the weights of keywords are proposed in order to improve the judgment-ability. In these methods, weights of keywords are calculated considering the relationship between keywords on thesaurus and the importance of keywords. Experimental results showed that the ability of similarity judgment of the proposed knowledge base is superior to the prototype.

## 1. はじめに

最近，コンピュータ上で柔らかな処理を実現しようとする研究が注目されている．実世界では，すべてのデータが整った問題ばかりでなく，不完全なデータを許容する技術の重要性が高まってきたためである．すなわち，従来，データや処理法がきちんと整理された問題を対象にしてきたのに対し，この枠を打ち破ろうとするものである．

このような柔らかい処理の実現を目指し，我々は，不完全な知識の下でも不完全さに応じた概略的な判断を行える推論法について研究を進めている[1),2)]．具体的には，知識が欠落して厳密な解を得られない場合でも，それと類似した知識を補完することにより推論を進め，概略的な解を導くものである．すなわち，この方式では，知識間の類似性判定が必要となる．知識は基本的には単語（概念を表す基本単位）を用いて表現されるので，知識間の類似性はそれを構成する単語間の類似性に基づいて決定されることになり，これを定量化する技術が必要となる．

単語に関する知識ベースとしては，日本では EDR 辞書[3)] や IPAL 辞書[4)] が，米国では CYC[5)] が代表的

† 拓殖大学工学部情報工学科
Department of Computer Science, Takushoku University
†† NTT アドバンステクノロジー株式会社
NTT Advance Technology
††† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

であるが，これらは用途を限定せず，汎用性を意識し，人手で作成されている．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，機械的に知識ベース（以下，概念ベースと呼ぶ）の構築を進めている．概念ベースでは，各単語（以下，概念と呼ぶ）は複数の属性と属性値のペアで表現される．属性はその概念に関連した概念であり，属性値はその関連の度合いである．この場合，概念間の類似性の度合い（以下，類似度と呼ぶ）は，属性の共通性の度合いを利用して定義すれば単純に計算することが可能となる．

これまで，日常用語約 4 万語からなる概念ベースを構築し，概念間の類似性判別について一定の性能を確認した[6)~8)]．しかし，新聞記事検索等への具体的な応用においては概念数が不足し，また，類似性判別能力も十分とはいえなかった[9)]．さらに，本来意味的に同一な同義語間の類似度についても，必ずしも人間の感覚に合った値とならない場合もみられた[10)]．

本論文では，これらに対処した大規模概念ベースの構築について述べる．本概念ベースは，一般語（約 26 万語），漢字（約 6,000 字），同義語（約 3,000 組）の各概念ベースにより構成され，前述した概念数不足に対しては，語数を拡大した一般語概念ベースと漢字概念ベースで対処する．ここで，漢字概念ベースは，概念が一般語概念ベースにない場合に，漢字単位の情報をもとにその概念を近似するために用いる．これにより，「激安」等といった漢字から構成される新語や造語に対しても，ある程度類似性判別が可能となる．また，類似性判別能力については，シソーラス上での属性間の関連性や各概念における語義文長を考慮した属性値付与の適正化を行うとともに，各概念ごとの属性数の均一化を行い，その向上を図る．さらに，同義語間の類似性については，専用的に設けた同義語概念ベースにより判定する．

また，パーソナルコンピュータ等での利用を考慮して，概念ベースから主要概念を選定し，それに基づいた概念ベースのコンパクト化を行う．さらに，構築した大規模概念ベースについては，その類似性判別能力と概念獲得能力を，従来の概念ベースおよび共起辞書（新聞記事から評価用に作成）と比較し，定量的に評価する．

以下，2 章で類似性判別能力の向上という観点からみた概念ベース構築手法，3 章で大規模化の観点からみた概念ベースの構成，4 章で主要概念の選定と概念ベースのコンパクト化，5 章で概念ベースの性能評価について述べる．

## 2. 概念ベースの構築手法

本章では，従来の概念ベース構築法[6),7)]について述べた後，その構築法における類似性判別能力の問題点と改良手法について述べる．

### 2.1 従来の構築法

概念の知識表現法としては，その属性と属性値の集合を基本とする方法が典型的である[11)]．たとえば，「林檎」という概念は，{(形:丸い),(色:赤い), ···} のように表される．このような知識を獲得すれば，汎用的かつ強力な概念ベースが実現されるが，現在の技術ではこれを機械的に行うことは不可能といえる．

したがって，我々は機械的な構築を第一義として，‘tf-idf’[12)] の考え方に基づいて国語辞書から概念知識を獲得することとした．‘tf-idf’ は，情報検索の分野でよく用いられる文書中のキーワードに重み付けをする方法で，その文書中に多く出現するキーワードほど（tf），また他の文書に出現する確率が少ないほど（idf），その文書の識別に重要であるとの考え方を基本としている．ここでは，この考え方に基づき，基本的には単語（概念に相当）の語義文に含まれている自立語を抽出してそれを属性とし，その出現頻度を属性値とする．さらにこの属性値は，その自立語が他の語義文に含まれる確率を考慮して再設定される．すなわち，この手法では概念は検索キーワードに相当し，その語義文は新聞記事等の検索対象に相当する．概念ベースの構築手順を以下に示す．

1）辞書からの属性と属性値の獲得

単語の語義文に含まれる自立語をその単語と何らかの関係があると見なして，属性とする．また，その自立語の出現頻度が高いほど関連性が高いと考え，‘tf’ の考えに基づいて出現頻度を属性値とする．ここで，語義文における自立語の重みはすべて 1 とする．たとえば，「馬」に関する辞書の語義文が，

『馬（うま）』家畜の一つ．たてがみが長い
草食の動物で ··· 動物 ···

と書かれているとき，この語義文を元にして以下の属性と属性値の集合が獲得される．

「馬」= {(家畜:1),(一つ:1), ···(動物:2),··· }.

2）孫引き・逆引き参照

語義文からは十分な属性が獲得できない場合がある．これに対して，概念の孫引きおよび逆引き参照で属性数を増やす．孫引きでは，属性自体も概念であることに着目し，それの属性も元の概念の属性とする．また，逆引きでは，概念 $g$ に属性 $g'$ があるとき，$g$ と $g'$ の間には関連性が存在するので，概念 $g'$ にも属性 $g$ を

付加する．

3）情報量による属性値の修正

概念ベースでは，すべての属性の重要度は一様ではない．たとえば，辞書特有の言い回しに由来する属性「もの」や「こと」は多くの概念に含まれているため，概念に対する意味的な重要性は低いと考えられる．したがって，‘idf’ の考えに基づいて出現頻度から属性の情報量を計算し，これを用いて属性値を修正する．

4）カテゴリへの属性変換

この段階で概念ベースは，すべての概念が属性になりうるので，総概念数を $N$ とすれば，$N \times N$ の行列で表現されることになる．この場合，$N$ は辞書の総語彙数であり非常に大きいため，同義あるいは類似等互いに関連する属性が多く含まれることになる．すなわち，属性は互いに独立であるとはいえない．このため，ここでは意味の近い属性をシソーラス上の同一のカテゴリ（比較的意味的に独立といえる）にまとめている．なお，これには 2,715 種のカテゴリが 12 段の木構造に分類されている，NTT の日英翻訳システムのシソーラス[13] を用いている．

以上のように構築された概念ベースでは，概念 $g_i$ は以下のように表現される．

$$g_i = \{(p_{i1}, q_{i1}), (p_{i2}, q_{i2}), \cdots\}. \quad (1)$$

ここで，$p_{ij}$ は属性，$q_{ij}$ はその属性値である．シソーラスのカテゴリ（属性）の総数を $n$ $(= 2715)$ とすると，$g_i$ は以下のように表現できる．

$$g_i = \{q_{i1}, q_{i2}, \cdots, q_{in}\}. \quad (2)$$

ここで，2）までで獲得されなかった属性 $p_j$ に対しては，属性値 $q_{ij}$ は 0 とする．したがって，概念は $n$ 次元の意味空間で表現され，その空間でのベクトルであると見なすことができる．

このようにベクトル表現される概念間の類似度は，各属性が独立（直交）と仮定すれば，単純にそのベクトルがなす角度 $\theta$ の余弦（正規化すると内積）として，以下のように定義することができる．

$$R(g_i, g_j) = cos(\theta) = \sum_{k=1}^{n} q_{ik} q_{jk}. \quad (3)$$

したがって，類似度は 0〜1 の間の値をとることになる．なお，同式において属性値 $q_{ij}$ はその二乗和が 1 になるように（$\sum_{j=1}^{n} q_{ij}^2 = 1$）正規化した値である．

また，類似度は文脈や見方（以下，観点と呼ぶ）によって変化する．たとえば，観点「家畜」における「豚」と「馬」の類似度は，観点「乗る」におけるそれらの類似度より高いといえる．このような観点を考慮した類似度は，観点の主な属性に対応する概念の属性の属性値を増幅した後，式 (3) を適用して算出する．具体的には，観点 $g_v$ で増幅された概念 $g_i$ の属性値 $q_{ik}^v$ は以下のように求める．

$$q_{ik}^v = \begin{cases} mq_{ik} & : q_{vk} > \lambda \\ q_{ik} & : q_{vk} \leq \lambda \end{cases}. \quad (4)$$

ここで，$m$ $(> 1)$ は増幅度，$\lambda$ はしきい値である．

なお，以上の概念ベースの構築法および類似度計算法の詳細については文献 7) を参照されたい．

### 2.2 改良手法

前節に述べた構築手法では，

（1） 属性が意味的に独立でない，

（2） 概念あたりの属性数のばらつきが大きい，

（3） 語義文中の重要な属性が強調されていない，

等の問題があり，適切な類似度が算出されない場合がある．これらに対し，それぞれ，シソーラスの情報を用いた上位属性の獲得，不要属性を削除した属性数の均一化，語義文の長さに応じた属性値の付与で対処する．以下，これらの手法について述べる．

#### 2.2.1 上位カテゴリ付与

前述したように，概念間の類似度はベクトルの内積で求めている．しかし，このベクトルの基底である属性が完全に意味的に独立（直交）していないため，適切な類似度が算出されない場合がある．

たとえば，以下のように概念「ダイニングテーブル」（$g_a$ とする）が属性 “机” と “材木”，概念「踏み台」（$g_b$ とする）が属性 “台” と “板” を持つと仮定しよう．

$g_a$ = {(机:0.7),(材木:0.7)}.

$g_b$ = {(台:0.7), (板:0.7)}.

この場合，類似度 $R(g_a, g_b)$ を求めると “0” となる．これらの属性は互いに意味的に関連するにもかかわらず，独立（直行）すると扱っているからである．しかし，実際 “机” と “台”，“材木” と “板” は，シソーラス的に表現すると，たとえば**図 1** のように関連する．したがって，本来 $g_a$ と $g_b$ は意味的にある程度類似し，$R(g_a, g_b) > 0$ となるべきである．このため，こ

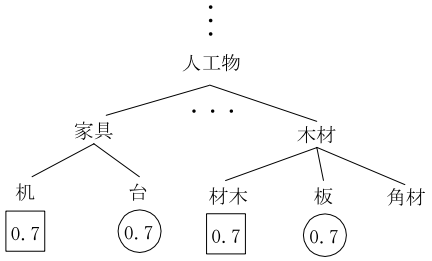



**図 1** 属性間の関係
Fig. 1 Relationship between categories.

のような直交でない意味空間を考慮した類似度算出が必要となる．

直交する意味空間を構築する方法としては，行列の直交変換や主成分分析が利用されている[14),15)]．たとえば，宮原らは辞書から獲得したデータに対して，固有値分解を行い，直交する意味空間を構築している[14)]．しかし，これらの方法では構築された空間の基底はデータに依存して決定されるため，データが不完全な場合には得られた基底は真に直行しているとはいえない．また，たとえ完全なデータを獲得できたとしても，類似度を前述したベクトルの内積で求める場合，直交変換された後の空間での類似度は，変換前の空間での類似度と同じ値となる．

ここでは，シソーラス上の関係に着目して，属性の非直行性を補完する手法，具体的には直行しない属性を直行すると見なすことにより生じる問題に対処する手法を提案する．概念はある属性を持ったとき，シソーラス上では下位属性と上位属性との間には is-a 関係等が存在するため，その属性の上位属性に対しても関連性を持つ[16)]．この関係を利用し，もとの属性からその上位の属性を獲得し，概念ベクトルを修正する．たとえば，先の例では上位属性「家具」や「木材」を以下のように概念 $g_a$，$g_b$ の属性とする．

$g_a$ = {(机:0.7),(家具:$q_1$),(材木:0.7),(木材:$q_2$)}.

$g_b$ = {(台:0.7),(家具:$q_3$),(板:0.7),(木材:$q_4$)}.

これから分かるように，$g_a$，$g_b$ は共通な属性を持つことになり，類似度 $R(g_a, g_b) > 0$ となる．すなわち，意味的に独立でない属性間からその共通属性を抽出し付加したことになり，実効的に属性の非直行性が補完されたと見なすことができる．この場合，上位属性に対し，どの程度の値 $q_1$〜$q_4$（付加値と呼ぶ）を付加するか問題となるが，付加値は下位属性の属性値とシソーラスの構造に依存すると考えられる．したがって，ここでは，シソーラスの構造をもとに算出された属性の情報量を利用して付加値を決定する．すなわち，属性が上位の位置にあるほど下位属性の数が多くなり情報量が小さくなるので，付加値を小さくする．具体的には，ある属性 $j$ に属性値が存在した場合の，その上位属性 $k$ に対する付加値 $q_{ik}^{j}$ を以下のように求める．

$$q_{ik}^{j} = q_{ij}(\frac{I_k}{I_j}). \tag{5}$$

ここで，$I_j$，$I_k$ は，それぞれ属性 $j$，$k$ の情報量であり，$q_{ij}$ は属性 $j$ の属性値である．また，その上位属性 $k$ には複数の下位属性が存在したり，それ自体に初めから属性値が存在したりする場合もある．これらを考慮し，ここでは，最終的な属性 $k$ の属性値 $\hat{q}_{ik}$ を以下のように求めることとする．

$$\hat{q}_{ik} = max\ \{q_{ik}, q_{ik}^{j1}, q_{ik}^{j2}, \dots\}. \tag{6}$$

ここで，$q_{ik}$ は初めからの属性 $k$ の属性値であり，$q_{ik}^{j1}$，$q_{ik}^{j2}$，... は $k$ の下位属性から算出された付加値である．たとえば，前述の例で概念「踏み台」が属性「角材」と「木材」も持つとする（図 1）．この場合，属性「木材」の付加後の属性値は，「木材」の属性値と「板」，「角材」の属性値から算出された付加値の最大値となる．

また，属性の情報量はその属性に属する下位属性の数をもとに算出することとし，以下のように設定する．

$$I_k = -log_2(\frac{n_k + 1}{n}). \tag{7}$$

ここで，$n_k$ はシソーラスにおける属性 $k$ の下位属性の数であり，$n$ は総属性数である．

#### 2.2.2 属性数均一化

概念ベースの応用の 1 つである概略推論法[2)] では，2 つの概念が類似するか否かの判断を行う．すなわち，この判断のための固定したしきい値が必要となる．また，$g_a$，$g_b$ が類似し，かつ $g_c$，$g_d$ が類似しないとすれば，$R(g_a, g_b) > R(g_c, g_d)$ が成り立つ必要がある．

一方，国語辞書では語義文の記述量が各概念で同じではなく，その量が多い概念は獲得される属性の数が多くなる．概念間の類似度は平均的には，共通属性の比較に基づいているので，属性数の多い概念のペアではその値が高くなり，属性数の少ない概念のペアではそれが小さくなる傾向がある．属性が平均的に分布する概念ベースモデル[8)]（総属性数 $M$，平均属性数 $r$）を用いると，すべての概念間の平均的な類似度 $\bar{R}$ は以下で与えられる（属性値は均一で $1/\sqrt{r}$ と仮定）．

$$\bar{R} = \sum_{k=0}^{r} \frac{{}_M C_k \times {}_{M-k} C_{r-k} \times {}_{M-r} C_{r-k}}{({}_M C_r)^2} \times \frac{k}{r}. \tag{8}$$

したがって，$r$ が大きければ大きいほど $\bar{R}$ が大きくなる．たとえば，上式では $M = 12$，$r = 3, 4, 5$ の場合，$\bar{R}$ はそれぞれ，0.25，0.33，0.42 となる．このため，属性数が多ければ実際には類似しない概念間の類似度が，類似する概念間の類似度より大きくなる場合が生じうる．

一方，前述した孫引きや逆引き参照で属性を増やすため，概念に関連しないノイズとなる属性が付与されることがある．属性数が多ければ多いほどノイズも多くなると考えられる．しかし，ノイズとなる属性は，本来その概念と関連しないので，語義文中の出現頻度が低く，属性値が小さいと考えられる．

以上を考慮して，ここでは属性値の小さい属性をノイズと見なし，これらを削除して属性数の均一化を行う．なお，このときの属性数は実験的に決定する．

#### 2.2.3 語義文解析手法

従来の手法では，属性値を単純に語義文における属性（自立語）の出現頻度で決定している．すなわち，すべての自立語の重みは同じとしている．しかし，語義文中の個々の文においては，その文が短いほど自立語の持つ意味が大きいと考えられる．見出し語（概念）を少ない語数で説明しているため，自立語と見出し語の関連性が高いといえるからである．したがって，ここでは文の長さ（自立語の数）の逆数を重みとし，属性値を出現頻度とその重みで算出する．具体的には，1 文の中の自立語の数を $m$ とすると，自立語の重要度 $d$ を以下のように求める．

$$d = 1/m. \tag{9}$$

属性 $p$ の属性値 $q$ は，それが語義文に $n$ 回出現し，重要度がそれぞれ，$d_1, \ldots, d_n$ のとき，下式で求める．

$$q = \sum_{i=1}^{n} d_i. \tag{10}$$

たとえば，「空（そら）」の語義文が「地上に高くそびえる空間．天.」であったとする．この場合，「空」の各属性値は以下のように算出される．

「空」= {(地上:0.25),(高く:0.25),(そびえる:0.25),
(空間:0.25),(天:1)}.

なお，従来の手法を用いると，この例では属性値はすべて 1 となる．

## 3. 大規模概念ベースの構成

2 章に述べた手法を用いて概念ベースの構築を行う．しかし，あらゆる概念に対して，この手法で概念ベクトルを生成することは，新語や造語の存在を考えると不可能である．したがって，ここでは，大規模な一般語の概念ベースとともに漢字概念ベースを構築し，一般語概念ベースに存在しない概念に対し，漢字概念で近似することとする．また，同義語概念ベースを構築して同義語判定を可能にする[17]．本概念ベースの構成と類似性判別の動作を**図 2** に示す．

本概念ベースでは，観点 $g_v$ における 2 つの概念 $g_i$，$g_j$ 間の類似度計算は以下のように行う．

1) $g_i$，$g_j$ が同義語であれば，すべての観点において類似度を 1 とする．
2) $g_i$，$g_j$ が同義語でなければ，概念ベクトルを取得し，2.1 節の手法で類似度を算出する．

また，概念 $g$ の属性ベクトルは以下のように取得する．

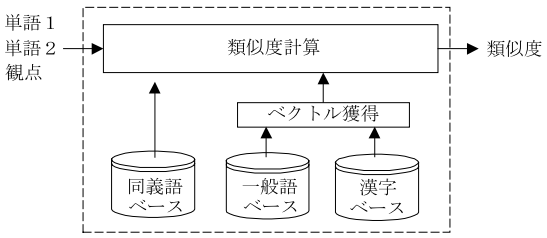



**図 2** 大規模概念ベースの構成と動作
Fig. 2 Configuration of large-scale knowledge base of words.

i) $g$ が一般語概念ベースに存在すれば，その概念ベースからベクトルを取得する．
ii) 一般概念ベースに存在しなければ，漢字概念ベースを用いてベクトルを合成する．

### 3.1 一般語概念ベース

3 つの国語辞典[18)~20)] を用い，2 章で示した方法により総数 257,905 の一般語概念ベース（一般語 GB）を構築した．また，属性数を 100 で均一化した．5 章に述べる評価手法で評価した結果，均一化する前と均一化した後の評価値は，それぞれ 0.609 と 0.611 となり，均一化することにより性能が若干向上した．評価値の差は誤差範囲内であるが，ノイズが削除され性能が向上したとも考えられる．なお，均一化前は属性数が 10 から 1,619 までばらつき，その平均は 229 であった．また，均一化後は，属性数 100 未満の概念数は 47,125 であり，これらの概念の平均属性数は 70 であった．

### 3.2 漢字概念ベース

漢和辞典[21)] と国語辞典を用いて総字数 5,651 の漢字概念ベース（漢字 GB）を構築した．国語辞典の場合，漢字 1 字の見出し語が存在するため，その語義文も利用することとした．この概念ベースでは，語義文中の自立語（属性）は，基本的には漢字 1 字ではないため漢字概念とはならない．したがって，属性も概念であるという特徴を利用した「孫引き・逆引き」参照は行わない．また，前節と同様に属性数を 100 で均一化した．均一化前は，属性数は 4〜1,499 までばらつき，その平均は 467 であった．また，均一化後は，属性数 100 未満の概念数は 443 であり，これらの概念の平均属性数は 59 であった．

漢字 GB を利用した概念ベクトルの合成は以下のように行う．

1) 概念（単語）を漢字に分解し，漢字ごとの属性ベクトルを抽出する．
2) 漢字のベクトルを線形結合し，属性値の二乗和

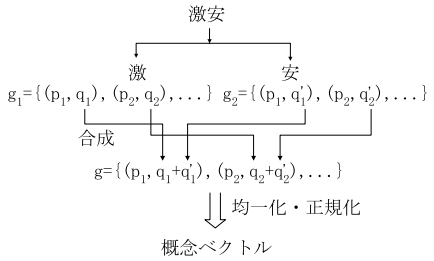



**図 3** 漢字 GB による概念ベクトル合成例
Fig. 3 Example of word-vector composition.

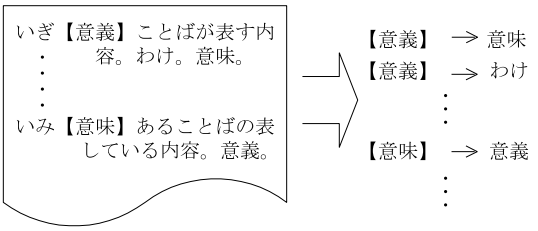



**図 4** 辞書からの抽出
Fig. 4 Synonym acquisition.

が 1 になるように正規化を行い，概念ベクトルとする．

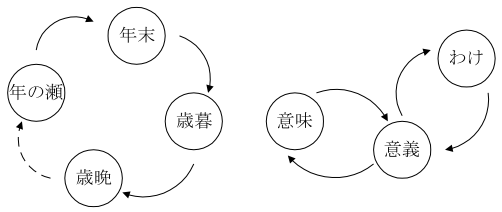
概念ベクトルの合成例を**図 3** に示す．このように，「激安」等の新語や造語に対して概念ベクトルを合成することができる．

### 3.3 同義語概念ベース

学研国語辞書[20] を利用し，見出し語と語義文の関係に着目して同義語概念ベースを構築した．辞書内の語義文には，単語 1 語からなる文が存在している．たとえば，**図 4** に示すように「いぎ『意義』」という見出し語に対しては，「意味」がそれにあたる．このような単語は，1 語でその見出し語を説明しているので，同義語である可能性は高いと考えられる．したがって，このような組を抜き出して同義語の候補群とした．

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合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合せの中で，共通の単語を含むループどうしをまとめることで，1 つの同義語グループとする（図 5 b）．このグループ化をすべてのループに対して行った結果，2,986 のグループが得られた．

さらに，人手により，実際には同義語関係にないグループを削除し，最終的に 2,657 のグループ（平均概念数 2.4）の同義語辞書を作成した．同義語として不適切なグループは，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」もまた同義といえる．しかし，「レコード」は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．




a) 同義ループ　　b) ループ結合

**図 5** 同義語グループ
Fig. 5 Synonym groups.

## 4. 主要語選定と概念ベースのコンパクト化

前章で述べた概念ベースの容量はテキスト形式で 500 MB 以上になり，パーソナルコンピュータ等の主記憶上での利用は困難である．したがって，このような利用を想定した，コンパクト版概念ベースの構築を行った．コンパクト化は，主要な概念を選定し，主要でない概念を主要概念で近似することにより実現した．以下，これらについて述べる．

### 4.1 主要概念の選定

主要概念は国語辞書[20] の 10 万語の中から手作業で選定した．選定の基準としては，日常的な基本単語を取り出すことを目指して以下を設定した．

i) 複合語（たとえば「立体図形」）を除く．ただし，きわめてよく使うもの（たとえば，「株式会社」）は選択する．
ii) 慣用的な用語（たとえば，「舌を巻く」）を除く．
iii) 接頭辞，接尾辞のついた語（たとえば，「不確実」や「歴史的」）を除く．ただし，きわめてよく使うもの（たとえば，「不真面目」や「合理的」）は選択する．
iv) 複数の動詞が結合した単語（たとえば，「叩き出す」）を除く．ただし，よく使うもの（たとえば，「組み合わせる」）は選定する．

これらの基準はきわめて曖昧であるともいえるが，厳密な基準は，自然言語を対象としているためその設定自体が困難である．また，その設定ができたとしても，それを用いて人間が判定を行うため，適切にその基準に合致するとは考えにくい．さらに，判定に要する時間も膨大となる．したがって，ここではこのような基準を用いることとした．選定にあたっては，この曖昧性を補完し主要概念を取りこぼさないために，4

人の実験者で行い，それぞれ選定された概念の和集合を主要概念集合とした．その結果，47,441 の概念が選定された．なお，各実験者が選定した主要概念の数はそれぞれ，31,731，26,055，18,962，22,266 であった．また，4 人，3 人以上，2 人以上，1 人以上に選ばれた概念の数はそれぞれ，7,155，16,183，28,235，47,441 であった．

なお，主要概念集合として，従来 GB の概念集合をそのまま利用することも可能であったが，機械的に構築しているため，一部重要な概念が欠落していた．したがって，前述したように新たに主要概念の選定を行った．

### 4.2 概念の代表化

概念ベースを圧縮する 1 つの手法として，主要でない概念をそれと類似した主要概念で代替（以下，代表化と呼ぶ）することが考えられる．ここでは主要概念でない概念 $g$ に対し，主要概念の中からそれと一番類似する概念 $g'$（式 (3) で求めた類似度が一番高いもの）を検索し，これを $g$ の代表概念とした．

前節の主要概念を用いて代表化を行った結果，本 GB の容量は 100 MB となり，約 1/5 に圧縮された．なお，もとの概念（$g$）と代表概念（$g'$）との類似度の平均と分散の実測値は，それぞれ 0.85 と 0.13 であった．この結果から，このような単純な代表化でも，もとの概念ベースをある程度近似できているといえる．

## 5. 評　　価

構築した概念ベースに対し，概念獲得能力と類似性判別能力について評価した．また，相対的に性能を比較するために，新聞記事より共起概念ベースを構築し，同様の手法で評価を行った．以下，評価手法とその結果について示す．

### 5.1 概念ベクトル獲得能力

概念ベースを情報検索に適用する場合，検索のキーワードと対象文書の中の単語に対して類似度を算出し，それに基づいて文書を出力する等が考えられる[9]．したがって，このような応用では，概念ベースから単語のベクトルが獲得できる確率（すなわち，その単語が概念ベースに存在する確率）が高いほど望ましいといえる．

ここでは，規模という観点から単純に新聞に出現する単語のベクトル獲得率で評価することとする．評価データとしては，毎日新聞の記事 1 年分[22] を用いた．この記事には，延べ語数 10,238,204 語，総語種数 106,590 が含まれていた．これに対するベクトル獲得率を**表 1** に示す．同表から分かるように，延べ語数に対するベクトル獲得率，語種に対するベクトル獲得率ともに，大幅に向上している．なお，一般語 GB と漢字 GB を併用しても，獲得率は 100%にはなっていない．これは，新聞記事ではカタカナの固有名詞や，本来，漢字の語がひらがな表記されたものが多用されていたためである．

**表 1** ベクトル獲得率（%）
Table 1 Rates of word-vector acquiring.

| | 従来 GB | 本 GB（一般語） | 本 GB（漢字併用） |
|---|---|---|---|
| 獲得率（語数） | 62.5 | 84.0 | 91.3 |
| 獲得率（語種） | 22.3 | 43.4 | 63.1 |

### 5.2 類似性判別能力

#### 5.2.1 評 価 手 法

概念ベースの特性としては，

（1）類似する概念との間の類似度とまったく類似しない概念との間の類似度の差が大きい，

（2）類似する概念との間の類似度は関連する概念（類似しないが何らかの関連性を持つ）との間の類似度より大きい，

が必要といえる．これを考慮した評価法には類語辞典を用いた評価手法[8] があるが，ここではこれを評価値が最大 1 になるように改良する．具体的には，まず類語辞典の中からランダムに $n$ 個の単語を抽出し，サンプル概念（$g_a$）とする．各サンプル概念 $g_a$ の類語としてあげられている単語の 1 つを類似する概念 $g_b$，類語より 1 つ上のレベルの分類に含まれる単語の 1 つを関連する概念 $g_c$ とする．また，最大分類の異なった分類の中からランダムに 1 つの単語を選択し無関係の概念 $g_d$ とする．これらの $n$ 組の $g_a$〜$g_d$ を用いて，上記の（1）の評価指数として以下を設定する．

$$f_1 = (r_1 - r_3)/(1 + \sigma_1 + \sigma_3). \tag{11}$$

ここで，$r_1$，$r_3$ はすべての組の $R(g_a, g_b)$，$R(g_a, g_d)$ の平均値であり，$\sigma_1$，$\sigma_3$ はそれらの標準偏差である．同式の分母は，2 つの類似度分布の重なり具合を表しており，これが小さいほど判別能力が高いと考える．

一方，（2）の評価指数としては，単純に正しい判断（$R(g_a, g_b) > R(g_a, g_c)$）の割合が大きいほど良いとし，以下を設定する．

$$f_2 = m/n. \tag{12}$$

ここで，$m$ は正しい判断の数を表す．

これらを用い，概念ベースの類似性判別能力の評価指数を以下のように設定する．

$$F = f_1 \times f_2. \tag{13}$$

この総合評価指数 $F$ は，類似概念間の類似度が 1，無関係概念間の類似度が 0，また類似概念と関連概念と

**表 2** 評価結果
Table 2 Evaluation results.

| | 従来 GB | 本 GB（一般語） | 本 GB（漢字） |
|---|---|---|---|
| $f_1$ | 0.456 | 0.650 | 0.345 |
| $f_2$ | 0.880 | 0.940 | 0.770 |
| $F$ | 0.401 | 0.611 | 0.266 |

**表 3** 評価データの例
Table 3 Examples of evaluation samples.

| サンプル（$g_a$） | 類似（$g_b$） | 関連（$g_c$） | 無関係（$g_d$） |
|---|---|---|---|
| 全力 | 総力 | 人力 | 鉱物 |
| 視野 | 視界 | 視線 | 戦死 |
| 悪臭 | 異臭 | 臭い | 船 |
| 絶食 | 断食 | 試食 | 本 |
| 歓声 | 歓呼 | 叫ぶ | 貯蓄 |
| 跳躍 | 飛躍 | 快速 | 唾 |
| 熟睡 | 熟眠 | 仮眠 | 番号 |
| 近道 | 早道 | 通行 | 学友 |
| 行為 | 行動 | 言動 | 橋 |

の類似度の逆転がない，理想的な概念ベースの場合で 1 となる．

#### 5.2.2 評 価 結 果

1) 本 GB と従来 GB との比較評価

小学館の類語例解辞典[23] を用い，200 組のサンプル概念を選択し，各概念ベースについて評価を行った．結果を**表 2** に示す．同表から分かるように，本 GB の評価値は従来 GB の評価値を大きく上回った．すなわち，2.2 節に示した改良手法は有効であったといえる．また，評価値 $F = 0.611$ は，前述 $F = 1$ の理想的な場合に対し，かなり良い値といえる．また，漢字 GB のみを用いた場合でも，類似性判別がある程度可能であることが明らかになった．なお，サンプル概念の一部を**表 3** に示す．

2) 共起 GB との比較評価

また，このようなベクトル表現の概念ベースの 1 つとして，新聞記事データから共起概念ベースを作成した．ここでは，文書の中の共起情報に基づいて属性と属性値を決定する．たとえば，「予防接種の普及で発生数は減っている」という文があったとき，自立語である「予防」，「接種」，「普及」，「発生」，「数」，「減る」を抽出し，これらは互いに関連すると考える．具体的には，たとえば，「予防」の概念に対し，他の概念をこの概念の属性とする．また，属性値は，文の中の距離が短いほど関連性が高いとし，距離の逆数とする．たとえば，「発生」は「予防」から数え 3 番目の単語であるため，距離を 3 とし，「発生」の属性値を 1/3 とする．すなわち，「予防」の属性と属性値は以下のように算出される．

「予防」= {(接種:1.0), (普及:0.5), (発生:0.33), (数:0.25), (減る:0.2) }.

このように属性と属性値を新聞記事のすべての文で獲得した後，2 章で述べた他の手順を用いて共起 GB を構築した．なお，出現頻度の低い（3 以下）単語（概念）は，十分な属性数を確保することができないと考え，共起 GB から外すこととした．

**表 4**「台風」の上位類似概念
Table 4 List of words most similar to word ‘taifuu’.

| 本 GB（一般語） | 本 GB（主要概念） | 従来 GB | 共起 GB |
|---|---|---|---|
| 颱風 | 嵐 | 野分き | 低気圧 |
| 室戸台風 | 夕凪 | 来襲 | 曇り |
| 枕崎台風 | 野分 | 具風 | 南々西 |
| 野分き | 朝凪 | 竜巻 | 前線 |
| 嵐 | 暴風 | 爽涼 | 雨 |
| 夕凪 | 大嵐 | 不揃い | 降り |
| 野分 | 一荒れ | 秋涼 | 東南東 |
| 朝凪 | 凪 | 春一番 | 高波 |
| 夕凪 | 暴風雨 | 秋日和 | 見舞う |
| 暴風 | 強風 | 秒速 | 洪水 |

こうして構築した共起 GB に対し，同様に評価を行った結果，評価値 $F = 0.071$ が得られた．なお，共起 GB は 1 年分の記事データしか用いていないため，前述の 200 組のサンプルすべてが含まれてはいなかった．したがって，評価はこれに含まれていた 157 組について行った．また，この 157 組を用いて，本 GB（一般語）を再評価した結果，$F = 0.611$ が得られた．これらの結果から，類似性判別等の用途においては国語辞書を知識源とした概念ベースは，新聞記事等を知識源とした共起ベースより圧倒的に有効であるといえる．

3) 連想検索による評価

前述した機械的評価では感覚的な性能を把握しにくいため，一例として，概念「台風」に対して，最も類似する概念の上位 10 個を各概念ベースで抽出した．結果を**表 4** に示す．同表から本 GB では，従来 GB や共起 GB に比べ，適切な類似概念が多く検出されていることが分かる．また，本 GB の主要概念を用いた場合，日常的に多用される概念が多く検出されることが分かる．すなわち，類語検索等にはこれを利用した方がよいといえる．

## 6. む す び

本論文では，単語の意味の類似性判別のための大規模概念ベースの構築手法について提案した．本手法では，シソーラス上での属性間の関連性や語義文長を考慮した属性値付与，さらに属性数の均一化等を行い，

より適切な概念ベクトルの生成を実現している．構築した概念ベースは，総数約 26 万語の一般語概念ベース，約 6,000 字の漢字概念ベースと約 2,700 グループの同義語概念ベースからなる．漢字概念ベースは，新語や造語等，一般語概念ベースにない概念についての概念ベクトル生成に用いる．また，同義語概念ベースは同義語判定に用いる．

構築した概念ベースについて，概念数および類似性判別能力の面から評価した結果，従来の概念ベースに対し，大幅な性能向上が実現された．また，主要概念を選定して概念ベースのコンパクト化を行い，小規模のシステムでも容易に利用可能とした．

**謝辞** 本研究の一部は文部省科学研究費補助金（課題番号 11650412）および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」の研究助成によって実施された．

## 参 考 文 献

1) 松澤和光, 石川 勉, 河岡 司：アバウト推論とその類似性判別機構，AI 学会研究会資料，SIG–J–9401, pp.103–110 (1994).
2) Nguyen, V.H., 石川 勉, 阿部明典：知識の類似性を利用した概略推論法，電子情報通信学会論文誌，Vol.J84-D-I, No.4, pp.389–400 (2001).
3) 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, No.1, pp.32–43 (1997).
4) 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究会資料，Vol.94–104, pp.9–16 (1991).
5) Guha, R.V. and Lenat, D.B.: Cyc: A midterm report, *AI Magazine*, Vol.11, No.3, pp.32–59 (1990).
6) 笠原 要，藤本和則，松澤和光，石川 勉：精錬化に基づく概念ベース構成法，信学技報，DE95-7, pp.49–56 (1995).
7) 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).
8) 石川 勉，井澤潤一郎，Nguyen, V.H.，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，Vol.13, No.3, pp.470–479 (1998).
9) 熊本 睦，島田茂夫，加藤恒昭：概念ベースの情報検索への適用，情報処理学会研究報告，Vol.99, No.1, pp.9–16 (1999).
10) 入江 毅，渡部広一，河岡 司，松澤和光：知的メカニズムのための概念間の類似度定量化方式，情報処理学会全国大会，pp.201–202 (1999).
11) 安西祐一郎：認識と学習，岩波書店 (1989).
12) Salton, G. and Allen, J.: Text Retrieval Using the Vector Processing Model, *3rd Annual Symposium on Document Analysis and Information Retrieval* (1994).
13) 池原 悟，宮崎正弘，横尾昭男：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究会資料，Vol.84–13, pp.95–102 (1991).
14) 宮原降行，清木 康，北川高嗣：意味の数学モデルによる意味的連想検索の高速化アルゴリズムとその実現方式，情報処理学会論文誌，Vol.38, No.7, pp.1399–1411 (1997).
15) 小島秀樹，伊藤 昭：文脈依存的に単語間の意味距離を計算する一手法，情報処理学会論文誌，Vol.38, No.3, pp.482–489 (1997).
16) Nguyen, V.H., 石川 勉, 笠原 要：ベクトル表現された概念に対する類似度計算法，AI 学会ことば工学研究会資科，SIG–LSE–A001, pp.49–55 (2000).
17) 帆苅 譲，石川 勉，笠原 要：言葉の意味に関する階層型大規模概念ベースの構築，情報処理学会研究報告，Vol.99, No.1, pp.25–32 (1999).
18) 松村 明（編）：大辞林第二版，三省堂 (1995).
19) 新村 出（編）：広辞苑第四番，岩波書店 (1991).
20) 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).
21) 藤堂明保，松本 昭，竹田 晃（編）：新版漢字源，学習研究社 (1994).
22) 毎日新聞社：CD 毎日新聞 94 データ集 (1994).
23) 遠藤織枝ほか（編）：類語例解辞典，小学館 (1994).

(平成 13 年 2 月 23 日受付)
(平成 14 年 9 月 5 日採録)

**Nguyen Viet Ha**（学生会員）
1997 年拓殖大学工学部情報工学科卒業．1999 年同大学院博士前期課程修了．現在，同博士後期課程に在学中．現在の研究テーマは大規模知識ベースおよびそれを用いた推論法．

**帆苅 譲**
1997 年拓殖大学工学部情報工学科卒業．1999 年同大学院博士前期課程修了．同年 NTT–AT（株）に入社．現在，コンピュータネットワーク関連技術の開発に従事．人工知能学会会員．

**石川　勉**（正会員）

1970 年電気通信大学電気通信学部応用電子工学科卒業．同年電電公社入社．1995 年より拓殖大学工学部情報工学科教授．工学博士．柔らかな情報処理のための人工知能，並列処理の知識処理への応用等の研究に従事．人工知能学会，電子情報通信学会，IEEE 各会員．

**笠原　要**（正会員）

1991 東京工業大学大学院総合理工学研究科電子化学専攻修士課程修了．同年日本電信電話（株）入社．現在 NTT コミュニケーション科学基礎研究所研究主任．知識処理技術，特に大規模知識ベースの研究に従事．人工知能学会会員．