国語辞書を利用した日常語の類似性判別

笠原 要[†] 松澤和光[†] 石川 勉[††]

人間は，単語の意味を表す「概念」間の類似性を，その単語が扱われる文脈や状況の変化に応じて柔軟に判別する．本論文では，属性とその重みより構成した概念の知識ベース（「概念ベース」）を用い，文脈や状況等を表現する単語（「観点」）を指定したときに，観点に応じて概念間の類似性を判別する方式を提案する．この方式は，概念どうしの類似性判別を行う前に，概念中の属性の重みを観点に基づいて強調することを特徴とする．概念ベースは，まず国語辞書等の語義文から自立語の出現頻度に基づいて属性とその重みを獲得し，次いで得られた概念ベースの自己参照による新たな属性の追加，および不要な属性の統計的な除去からなる精錬を行うことによって，完全に機械的に構築した．実際に作成した約4万の日常語に関する概念ベースで方式評価を行った．この結果，提案の類似性判別方式がシソーラスを用いる従来の方式に比べて有効であり，また，この判別において観点が効果的であることを明らかにした．

A Method for Judgment of Semantic Similarity between Daily-used Words by Using Machine Readable Dictionaries

KANAME KASAHARA,[†] KAZUMITSU MATSUZAWA[†] and TSUTOMU ISHIKAWA[††]

We propose a method for measuring the semantic similarity between words using a large-scale knowledge base that is automatically constructed from machine-readable dictionaries and is self-refined. This method of measuring similarity takes into consideration the fact that similarity changes depend on situation or context, this is what we call a 'viewpoint.' A feature of this method is that certain parts of the overall concept of measured words, compared with each other, are emphasized by using the viewpoint when calculating the degree of similarity. An experimental knowledge base, which contains knowledge of 40,000 Japanese daily-used words, was employed in order to evaluate the proposed method of measurement. The similarity measurements with the proposed method were closer to those decided by human judges than were similarity measurements made using the conventional way of using a thesaurus. Moreover, it was found that consideration of the viewpoint was effective when measuring the semantic similarity.

1. はじめに

従来の情報処理技術は「コンピュータ世界」の問題，つまりすべ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「現実世界」ではつねにすべてが整った問題ばかりを対象とするわけではない．したがって，不完全な知識の下でも問題解決を可能とする技術が，今後の情報処理分野の大きな課題の1つとなっている[1]．

そこで我々は，人間が不完全な知識の下で行う概括的な判断（俗にいう「アバウトな判断」）に着目し，これをコンピュータで実現する処理方式を「アバウト推論[2]」と名付けて研究を進めてきた．この方式は，知識が欠落していても類似した常識の補完によって推論を進めることを特徴とする．このため，常識を構成する種々の基本的な概念について，それらの間の類似性を人間と同じように判別する技術が必要となる．

我々はこれら基本的な概念を「日常用いる単語（日常語）が表す意味」ととらえ，数万語規模の日常語に対し実際に類似性判別が行える技術の確立を目標とした．また，人間は状況や文脈等（以下，「観点」と呼ぶ）に応じて類似性を柔軟に判別しており，このような観点に応じた判別が行えることも同時に目標とした．これは，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判別する際，「動物」が話題であれば「豚」が，「乗り物」の話題では「自動車」

† NTTコミュニケーション科学研究所
NTT Communication Science Laboratories
†† 拓殖大学工学部
Faculty of Engineering, University of Takushoku

が，各々「馬」により似ていると判別することに相当する．

このような観点に応じた類似性判別の研究としては，認知心理学[3),4)]や比喩・連想に関するもの[5),6)]等がある．しかしこれらでは，具体的な概念知識の獲得法について考察されておらず，多数の単語の類似性判別を実際に実現することはできない．

一方，自然言語処理分野では，単語を意味に基づいて分類したシソーラス（たとえば文献7)等）を用いたり，コーパスでの構文的な共起関係から類似性判別を行う方法[8)~10)]等が提案され利用されてきた．これらは確かに多数の単語を対象としているが，単語間の関係は固定的であり，観点に応じた判別は考慮されていない．

これに対し最近の研究[11),12)]では，機械可読辞書から獲得した大規模な語彙知識を用い，観点を考慮した類似性判別方式も提案されている．ただし，観点をどのように類似性判別に反映すべきか，十分な研究は進んでいない．また，機械可読辞書といっても，あくまで人間が読むためのものであり，計算機で処理することを考えて作成されたものではない．したがって，辞書から十分な品質の知識を獲得することは難しい．これは特に日本語の辞書を対象とした場合，形態素解析の困難さ等により顕著に現れる問題である．このため従来の同種の研究では主として英語の辞書が用いられてきた．

本論文では，日常用いられる数多くの単語に対し，観点に応じた柔軟な類似性判別を行える新しい方式を提案する．この方式の特徴は，まず国語辞書の語義文から容易に獲得可能な知識だけを用いて概念知識ベース（概念ベース）を構築し，次いで「精錬」と名付けた自己参照的な手法により，概念ベースの知識を精密化することにある．この2段階の処理により，日本語の辞書からでも高品質の概念知識を獲得することができる．

また，概念ベースに含まれる概念の中から観点に相当する概念を指定することにより，「変調」と名付けた処理によって，概念ベース中の任意の2概念間の類似性を観点に応じて判別することを可能としている．この方式は，既存の観点を考慮した方式に比べ，観点をより一般的にとらえる手法である．

以下，国語辞書からの概念知識の獲得および精錬を2章で，得られた概念ベースを用い観点に応じた類似性判別を行う変調方式を3章で，これら概念ベースと類似性判別方式の評価を4章で，他研究との比較を5章で述べる．

2. 概念ベースの構築手法

人間の行うような概念の類似性判別を実現するため，以下の方針に従って概念ベースを構築する．

- 方針1：**辞書からの概念知識の自動獲得**
  多数の概念について，人が直接知識を記述して概念ベースを構築することは困難である．そこで，国語辞書や百科辞典等を概念の知識源と見なして自動的な獲得を行う．
- 方針2：**概念の簡略化**
  現状の自然言語処理の技術レベルでは，辞書の語義文の意味を正確に解釈することは困難である．そこで概念ベース構築の第1段階として，大規模な知識を容易に獲得し得る単純な概念表現を採用する．
- 方針3：**概念ベースの精錬**
  人間は，辞書の語義文から見出し部分の単語の概念を理解できない場合，辞書の関連する部分の情報を総合して理解することができる．そこで，まず個々の語義文のみから対応する見出し部分の単語の概念を獲得し，すべての概念に関する粗い情報を持った概念ベースを最初に構築する．次に，この構築された概念ベースに対し，概念間の関係等に基づいて個々の概念を精密化する「精錬」の操作を行い，概念ベースの質を向上させる．

2.1 辞書よりの概念知識の獲得[13),14)]

概念の知識表現は様々提案されているが，属性と属性の値の集合として表現する方法が一般的である[15)]．「りんご」の概念表現を例としてあげる．

‘りんご’：{(‘形’, ‘丸い’), (‘色’, ‘赤い’),
(‘味’, ‘すっぱい’), ···}.

しかし，このような構造の知識を語義文を処理して自動獲得するのは，現状の技術レベルでは困難である．そこで，我々は，辞書から獲得し得る単純な形式の概念の定義を用いた．

概念ベース K における1つの概念を，意味特徴を表す属性 p_{ij} $(j = 1, 2, \cdots)$ と，概念と属性 p_{ij} の関連の深さを表す重み（$q_{ij} \geq 0$）の対集合で表現する．

‘概念’：$\{(p_{i1}, q_{i1}), \ldots, (p_{ij}, q_{ij}), \ldots\}$. (1)

国語辞書等の語義文を用いてこのような概念知識を獲得する．辞書の見出し部分にある単語を概念とし，その語義文中の自立語を形態素解析で抽出してその概念の属性とする．また，見出しごとの語義文中での属性の出現頻度をその属性の重みとし，式(1)の概念の知識が獲得される．

たとえば，「馬」に関する辞書の語義文が，

『馬（うま）』家畜の一．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ような「馬」の概念が獲得される．

‘馬’ : {(家畜, 1), (一, 1), (たてがみ, 1),
(長い, 1), …, (動物, 2), …}

ここで，概念ベースを構成する概念および属性を，ともに同じ n 種類の単語の中から選ぶものとする．この場合，概念ベース中の j 番目の概念が「馬」であるとき，その j 番目の属性は同じ「馬」となる．このような概念ベースの構造において，式 (1) の概念を表す $Word_i$ は，以下のようなベクトルとして表現される．

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

この表現を用い，概念ベース K を n 種類の概念の行ベクトルからなる，n 行 n 列の属性の重みの行列として表現する．

$$K \stackrel{\text{def}}{=} \begin{pmatrix} Word_1 \\ \vdots \\ Word_i \\ \vdots \\ Word_n \end{pmatrix} = \begin{pmatrix} q_{11} & \cdots & q_{1j} & \cdots & q_{1n} \\ \vdots & & & & \vdots \\ q_{i1} & \cdots & q_{ij} & \cdots & q_{in} \\ \vdots & & & & \vdots \\ q_{n1} & \cdots & q_{nj} & \cdots & q_{nn} \end{pmatrix}. \quad (3)$$

K の (i, j) 要素 q_{ij} は，概念 $Word_i$ の j 番目の属性の重みを表す．また，式 (2) は，K を用いて以下のように表される．

$$Word_i = e_i K. \quad (4)$$

ここで e_i は，i 成分のみ 1 で，他の成分が 0 の行ベクトルを意味する．

2.2 概念ベースの精錬[16),17)]

概念ベースの精錬は，**図 1** で示すように，2 種類の概念ベースの参照法，参照した概念の属性と元の概念の属性の線形結合，そして，属性の重みの調整よりなる．

(1) 概念ベースの再帰的参照

人間が単語の意味を辞書で調べるとき，語義文中に不明な単語があっても，孫引き，つまり不明な単語の意味をもう一度辞書を引いて調べることにより，元の単語の意味を理解することができる．たとえば，単語「馬」の意味を辞書で調べる際に，辞書の語義文にある単語「たてがみ」の意味が理解できなければ，人間は，「たてがみ」の意味を辞書で再帰的に調べる．そして，「たてがみ」の意味を理解したうえで，「馬」の意味を理解する．この方法を概念ベース中の概念の新たな属性を獲得する方法として用いる．

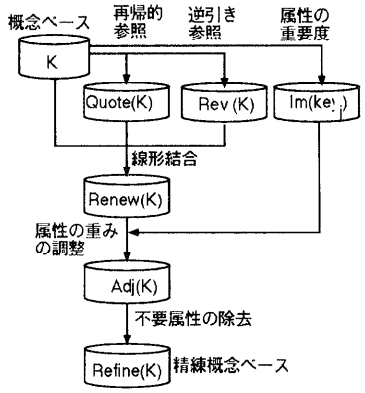


図 1 概念ベース精錬方式
Fig. 1 An overview of refining the knowledge base of word concepts.

概念ベース K（式 (3)）中では，j 番目の概念と j 番目の属性は同じ単語を指す．したがって，概念 $Word_i$ における j 番目の属性 p_{ij} の重み q_{ij} が 0 でないとき，$Word_i$ は概念 $Word_j$ の各属性と関連性があり，これを $Word_i$ の新たな属性と見なすことができる．たとえば，「馬」の概念 $Word_{馬}$ において属性「たてがみ」の重みが 0 でないとき，$Word_{たてがみ}$ の属性を $Word_{馬}$ の属性と見なす．そこで，$Word_i$ 中の各属性を再帰的に参照して新たな属性とその重みを獲得する再帰的参照 $quote(Word_i, K)$ を以下のように定義する．

$$\begin{aligned} & quote(Word_i, K) \\ & \stackrel{\text{def}}{=} \sum_{j=1}^{n} q_{ij} Word_j \\ & = \sum_{j=1}^{n} q_{ij} e_j K \\ & = e_i Quote(K) \quad (Quote(K) = K^2). \end{aligned} \quad (5)$$

ここでは，概念 $Word_i$ の属性 p_{ij} と概念 $Word_j$ は同じであり，概念 $Word_i$ の属性 p_{ij} の再帰的参照は，$Word_j$ を属性 p_{ij} の重み q_{ij} の強さで参照することを意味する．これにより，$Word_i$ 中で，属性の重みに従って属性に対応する概念が再帰的に参照される．また，再帰的参照は，複数回行うことが可能である．$Word_i$ について，s 回概念ベース K を参照すると，その結果 $quote^s$ は，

$$quote^s(Word_i, K) = e_i K^{s+1} = e_i Quote^s(K) \quad (Quotes(K) = K^{s+1}).$$

と表される．

(2) **概念ベースの逆引き参照**

単語の意味を辞書で理解する方法としては，辞書の語義文を読んで見出し部分の単語の意味を理解する以外に，語義文中の単語から見出し部分の単語の意味を推定する方法が考えられる．たとえば「寝る」という単語を辞書で引くと，主として寝る動作に関する記述が語義文に書かれている．一方，辞書で語義文中に「寝る」が頻出する「枕」や「寝巻」などの単語は，「寝る」と関連性が深く，「寝る」の意味の理解に役立つ．このような辞書の逆引きのアナロジーとして，概念ベースを逆引きして新たな属性を獲得する方法を提案する．

概念ベース K 中の属性 p_{ij} の重み q_{ij} が 0 以外の値をとるときには，$Word_i$ と j 番目の属性が関連性があることを示している．これは逆に，概念 $Word_j$ と i 番目の属性が関連性があると見なすことができる．そこで，概念ベース K における，概念 $Word_i$ の逆引き参照による属性獲得 $rev(Word_i, K)$ を以下のように定義する．

$$\begin{aligned} rev(Word_i, K) &\stackrel{\text{def}}{=} (q_{1i}, q_{2i}, \cdots, q_{ni}) \\ &= e_i K^T \\ &= e_i Rev(K) \\ &\quad (Rev(K) = K^T). \end{aligned} \tag{6}$$

ここで，K^T は概念ベース行列 K の転置行列を表す．この逆引き参照によって，たとえば，概念 $Word_{枕}$ の属性「寝る」の重みが 0 でないとき，その重みが概念 $Word_{寝る}$ の属性「枕」の重みとなる．

(3) **獲得属性の線形結合**

概念 $Word_i$ について，辞書から機械的に構築される概念ベース K を用いて，再帰的参照による属性獲得 $Quote(K)$（式 (5)）と，逆引き参照による属性獲得 $Rev(K)$（式 (6)）によって，新たな属性を獲得した．これら新たな属性は，元の概念ベースを補足する情報と考えることができる．そこで，これを元の概念ベース K にある重みで加えて自己参照概念ベース $Renew(K)$ とすることによって，概念ベースの質の向上が期待できる．加える方法としてはいろいろ考えられるが，ここでは，最も単純な線形結合を用いる．ただし，$Quote(K)$ では，属性の重みの値が K の要素の 2 乗のオーダになっているので，各重みの平方根をとった後に結合する．

$$\begin{aligned} Renew(K) &= \alpha K + \beta Route(Quote(K)) + \gamma Rev(K) \\ &= \alpha K + \beta Route(K^2) + \gamma K^T. \end{aligned} \tag{7}$$

$Route$ は行列の各要素ごとに平方根をとる演算子であり，α，β，γ は実験的に決定する正数である．

(4) **属性の重みの調整**

概念 $Word_i$ に対し，属性 p_j がどれほど意味を持つかは，その重み q_{ij} として表されている．しかし，概念ベース全体で見た場合，これら属性 p_j の重要性は一様ではない．たとえば，「～すること」「～するもの」等，辞書特有の言い回しに由来する属性「こと」や「もの」の類は，数多くの概念に含まれているがために，概念に対する意味的な重要性は低いと考えられる．反対に，ごく少数の概念だけに含まれる属性は，他の概念と比較して特に重視すべき属性であろう．以上の考えに基づき，概念ベース全体から見た属性 p_j の重要性に従って，重み q_{ij} の値を調整する．

ここでは，属性 p_j の重要性尺度 I_j は，$i = 1 \sim n$ において $q_{ij} \neq 0$ である概念の数を N_j とすると，確率 N_j/n から計算される情報量に比例すると考え，以下のように定義する．

$$I_j = \begin{cases} -\log \dfrac{N_j}{n} & (\text{if } N_j \neq 0) \\ 0 & (\text{if } N_j = 0) \end{cases} \tag{8}$$

すべての概念に含まれるような属性は，実は概念どうしを比較する際には意味がなく，その重要度は 0 となる．また，概念ベースに含まれない属性の重要性尺度も同じく 0 とした．

さて，概念ベース K に対する属性 p_j の重要性尺度 I_j を用いて，自己参照概念ベース $Renew(K)$ のすべての重み q_{ij} に対し重要性尺度 I_j の値をそのまま乗じることによって値を調整し，概念ベース $Adj(K)$ を得る．

$$Adj(K) = Renew(K)\ IM(K)$$

$$IM(K) = \begin{pmatrix} I_1 & & 0 \\ & \ddots & \\ 0 & & I_n \end{pmatrix}. \tag{9}$$

(5) **不要な属性の除去**

前項までの各種の精錬操作によって，各概念中の 0 でない重みの数が増加する．このとき，重みの相対的に小さな属性は，その概念にとってあまり意味のない属性，つまり不要な属性と考えられる．そこで，実際にデータを調べて，不要と思われる属性の重みの最大値の基準を設定し，この値以下の重みはすべて 0 にす

る（具体的な基準値は，4.1節(3)で詳述する）．

3. 観点に応じた類似性判別方式

前章で述べたように作成された概念ベースを用い，観点に応じた概念の類似性判別を行う．2つの概念の意味の近さを表す類似度 S は，概念ベース中の2つの単語の概念 $Word_1$，$Word_2$ と，判別の観点となる概念（「観点」と呼ぶ）$View$ より計算される（**図2**）．S は次の4つの手順に従って決定される．

- **STEP 1：属性のシソーラス圧縮**
  類似度の計算は，同じ属性どうしの重みの比較によって行う．このため，表記が異なるが意味の等しい属性は等しく扱う必要がある．そこで，シソーラスを利用し，属性をシソーラスのカテゴリーに変換する．
- **STEP 2：重みの正規化**
  個々の概念について，重みの0でない属性の数や個々の重みの合計値は語義文の分量によって異なり，一律には比較できない．そこで，概念ごとに重みの正規化を行う．
- **STEP 3：観点に応じた重みの変調**
  観点に応じた類似度の計算とは，観点中の属性を重視して行うものと仮定する．そこで，比較する概念中に観点と共通する属性が含まれるとき，その属性の重みを大きくして概念を変調し，状況に応じた概念を生成する．この概念を変調概念と呼ぶ．
- **STEP 4：類似度の計算（S）**
  類似度 S を，属性空間上の2つの変調概念がなすベクトルの角度から計算する．

以下，類似性判別方式の詳しい内容について説明する．

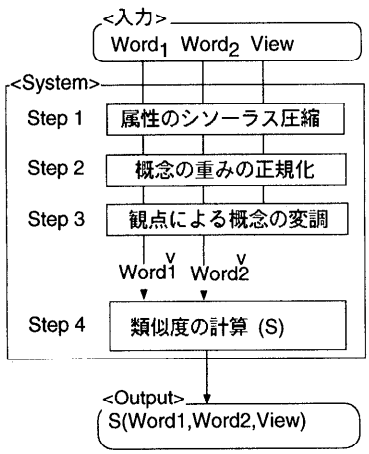


図2 類似性判別方式の全体図
Fig. 2 Similarity measurement scheme.

(1) 属性のシソーラス圧縮

類似度の決定は，同じ属性どうしの重みの比較を基本とする．したがって，'太陽'，'お日様' のような意味の等しい属性は，同じ属性と見なす必要がある．そのために，種々の単語を意味に基づいて k 種類のカテゴリーに分類したシソーラス T を用いる．

$$T = \{c_1, c_2, \cdots, c_m, \cdots, c_k\}$$
$$c_m = \{p_{m1}, p_{m2}, \cdots p_{ml}, \cdots\}. \quad (10)$$

c_m は，カテゴリーを表し，そこに含まれる単語 p_{ml}（単語数はカテゴリーによって異なる）は，すべて意味的に等しいものとする．概念の属性をカテゴリーに変換することによって，属性数を圧縮する．

$$\mathbf{Word_i} = (Q_{i1}, Q_{i2}, \cdots, Q_{ij}, \cdots, Q_{ik})$$
$$Q_{ij} = \sum_{l=1 \ p_l \in c_j}^{n} q_{il}. \quad (11)$$

この圧縮は，属性を単語からカテゴリーに一般化することを意味する．属性どうしを独立であると見なすと，概念 $\mathbf{Word_i}$ は，k 次元の空間（意味空間）上のベクトルとして表現される．

(2) 属性の重みの正規化

概念中の重みが正の属性の数や重みの合計は，語義文の分量によって変動し，類似計算に影響を与える．たとえば，数語の語義文からなる概念について，同じような概念と数ページの語義文からなる概念と比べると，意味空間上の原点付近に存在する数語の語義文からなる概念どうしが，つねに類似してしまう．そこで，意味空間上で概念 $\mathbf{Word_i}$ のベクトルの長さが一定（$=1$）になるように，属性の重みを正規化する．

$$\mathbf{Word_i} = (\widehat{Q}_{i1}, \widehat{Q}_{i2}, \cdots, \widehat{Q}_{ij}, \cdots, \widehat{Q}_{ik})$$
$$\widehat{Q}_{ij} = \frac{Q_{ij}}{|\mathbf{Word_i}|} = \frac{Q_{ij}}{\sqrt{\sum_{m=1}^{k} Q_{im}^2}} \quad (12)$$

(3) 観点に応じた属性の重みの変調

類似性を判別する際の観点としてはいろいろなものが考えられるが，ここでは，概念ベースに含まれるすべての概念が観点 $View$ となりうると考える．たとえば，ユーザから観点として「動物」が与えられたときには，概念 $Word_{動物}$ が観点 $View$ となる．

$$\mathbf{View} = (\widehat{Q}_{v1}, \widehat{Q}_{v2}, \cdots, \widehat{Q}_{vj}, \cdots, \widehat{Q}_{vk}). \quad (13)$$
$$(|\mathbf{View}| = 1)$$

観点中の属性を類似性判別において重視すべきものと仮定する．具体的には，概念 $\mathbf{Word_i}$ 中に $\mathbf{View}$ と同じ属性が含まれている場合，その属性の重みを強調

し，変調概念 $\mathbf{Word_i^v}$ を生成する．

$$\mathbf{Word_i^v} = (Q_{i1}^v, Q_{i2}^v, \cdots, Q_{ij}^v, \cdots, Q_{ik}^v)$$
$$Q_{ij}^v = \widehat{Q}_{ij} \cdot M(\widehat{Q}_{vj}). \quad (14)$$

M は観点中のどの属性が判別で重要かを決定する変調関数であり，様々な関数が考えられる．ここでは，観点の重みがあるしきい値を超えたとき，概念の属性の重みを定数倍する矩形関数を用いる．強調のしきい値としては，重みが0でない属性の数 $atnum(\mathbf{View})$ の関数を用いる．$atnum(\mathbf{View})$ が0でないとき，仮に $View$ の属性の重みがすべて等しいとする．属性の重みは，正規化条件より，$atnum(\mathbf{View})$ の平方根の逆数となる．この値を変調のしきい値の基準として用いる．

$$M(\widehat{Q}_{vj}) = \begin{cases} r & \text{for } \widehat{Q}_{vj} \geq Q_{avg}(\mathbf{View}) \\ 1 & \text{for } \widehat{Q}_{vj} < Q_{avg}(\mathbf{View}) \end{cases}$$
$$Q_{avg}(\mathbf{View}) = \frac{s}{\sqrt{atnum(\mathbf{View})}}. \quad (15)$$

$r\ (>1)$ は，変調の倍率を表す変数であり，実験的に決定される．また，s は実験的に決定される定数である．このようにして，観点 $View$ から見た概念 $\mathbf{Word_i^v}$ が表現される．変調された概念は，重みの2乗和が1ではないので，再び重みを正規化する．

(4) 類似度の計算（S）

観点 $View$ における，概念 $Word_1$ と $Word_2$ の類似度 $S(Word_1, Word_2, View)$ は，$View$ によって変調された概念間の類似度として定義する．

$$S(Word_1, Word_2, View) = R(\mathbf{Word_1^v}, \mathbf{Word_2^v}). \quad (16)$$

R は，意味空間上での2つの概念の近さの度合を示す関数であり，以下のような条件を満たすものである．

条件1

$0 \leq R(\mathbf{Word_a}, \mathbf{Word_b}) \leq 1.$

条件2

$R(\mathbf{Word_a}, \mathbf{Word_b}) \equiv R(\mathbf{Word_b}, \mathbf{Word_a}).$

条件3

$R(\mathbf{Word_a}, \mathbf{Word_a}) \equiv 1.$

条件4

$R(\mathbf{Word_a}, \mathbf{Word_b}) < R(\mathbf{Word_c}, \mathbf{Word_b})$ のとき，$\mathbf{Word_b}$ は $\mathbf{Word_a}$ よりも $\mathbf{Word_c}$ に類似している．

ここでは，類似度自体を相対尺度と位置付ける．何故ならば，(似ている/似ていない) の絶対的な尺度を2つの概念間の類似度で与えることが目的ではなく，同じ概念に対する他の2つの概念のどちらかが類似しているかをもって類似性判別とするためである．

上記条件を満たす類似度関数 R としていろいろ考えられるが，ここでは，情報検索で代表的な SMART[18] 等で採用されているのと同様に，2つの変調概念のベクトルのなす角の余弦で計算する．

$$S = cos\theta = \mathbf{Word_1^v} \cdot \mathbf{Word_2^v} = \sum_{j=1}^{k} Q_{1j}^v \, Q_{2j}^v. \quad (17)$$

4. 評価結果

概念ベースの構築，および類似性判別の評価基準として，人間の行う類似性判別結果との“近さ”を採用する．過去の評価実験[13]では，概念の類似検索を行い，文献検索における再現率—適合率[19]の考えに基づいて評価を行った．しかし，人間の判定する類似概念以外に，概念ベース中には適切な類似概念が多くあり，評価値が人間の感覚と一致しない問題が見られた．そこで，概念ベースの評価としては，検索された上位の類似な概念について，人間による判別に基づく評価尺度を用いる．また，類似性判別における観点の効果の評価としては，シソーラスより作成できる多義語の類似性判別の精度を用いて行った．

概念ベースが対象とする日常語は，4種類の辞書[7),20)~22)]の見出しを参考とし，表記揺れを考慮したうえで日常よく使われる単語34,000語を選定した．概念ベース中の概念と属性は，この日常語で構成される．語義文の分量の多い4種類の辞書[7),22)~24)]を用い，概念ベースの構築を行った．また，類似性判別における属性のシソラース圧縮（式(11)）のために用いるシソーラス T として，37万語を約3,000のカテゴリーに分類した当研究所作成の ALT-J/E シソーラス[25]を用いた．

4.1 概念ベース構築の評価

構築の各段階における概念ベースの評価は，3章で説明した類似性判別法を用いて行った．

- Step 1：34,000語の基本単語からランダムに50概念 $Word_i\ (i = 1, \cdots, 50)$ を評価対象として選択する．
- Step 2：評価を行う概念ベース G を用いて，$Word_i$ と，概念ベース中の全概念34,000との類似度を計算する．この場合，観点は考慮しない．
- Step 3：類似度の高い順に20概念 $Word_{ij}\ (j = 1, \cdots, 20)$ を選択し，概念ベースに基づいて得られた類似概念とする．
- Step 4：被験者が $Word_i$ と $Word_{ij}$ を比較し，**表1** にあげる判定値 $judge(Word_i, Word_{ij})$ を

表1 類似性判別の判定値
Table 1 Similarity judgment scores.

判定値 $judge$	$Word_i$ と $Word_{ij}$ の関係
1	類似する
0	類似していないが関連がある
−1	関連がない

表2 '台風' の類似検索結果の判定例
Table 2 A list of scores for words which are similar to $Word_i$ = 'taifū' (typhoon).

Rank(j)	$Word_{ij}$	$judge$(台風, $Word_{ij}$)
1	野分き	+1
2	具風	0
3	竜巻	+1
4	来襲	0
5	爽涼	−1
6	秋涼	+1
7	不揃い	−1
8	秋日和	0
9	春一番	+1
10	風速	0
11	涼風	0
12	嵐	+1
13	旋風	+1
14	秒速	0
15	競漕	−1
16	暴風圏	0
17	炎天	−1
18	襲来	0
19	真夏	0
20	全速力	−1

決定する．**表2** は，被験者による類似性判別結果の判定例である．

- Step 5：表2 のような，20 概念についての類似性判別の判定値 $judge(Word_i, Word_{ij})$ を用いて，類似度を順位の逆数で重みづけして平均した評価値 $eval(Word_i, G)$ を求める．

$$eval(Word_i, G) = \sum_{j=1}^{20} \frac{judge(Word_i, Word_{ij})}{j}$$
$$(-3.6 \leq Eval(Word_i, M) \leq 3.6). \quad (18)$$

- Step 6：サンプル 50 概念 $Word_i$ のついての評価値 $eval(Word_i, G)$ の平均 $Eval(G)$ を概念ベース G の評価値とする．

(1) **概念ベース作成における辞書の規模の効果**

概念ベースの知識源は辞書の語義文であるため，辞書の語義文量が多いほど，概念ベースの知識量は高まると考えられる．4 種類の辞書ごとの語義文それぞれと，4 つの辞書の語義文を合わせた語義文から個々に概念ベースを作成した．

表3 辞書のサイズと概念の属性数の関係
Table 3 Results.

辞書	平均属性数	辞書サイズ (MB)	比率
Dict 1	3.6	2.5	1.4
Dict 2	4.1	2.5	1.6
Dict 3	8.1	6.0	1.3
Dict 4	9.5	6.9	1.4
4 Dicts	16.2	20.8	1.3

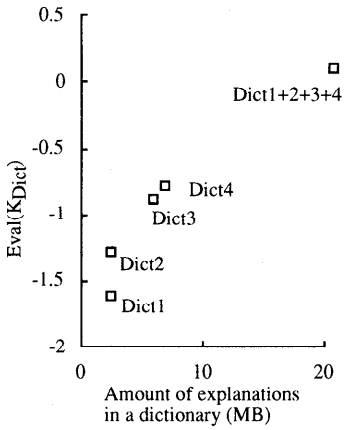


図3 辞書の記述量と獲得された概念ベースの関係
Fig. 3 Results.

概念ベース中の概念の平均属性数と語義文のサイズの関係は，**表3** のとおりである．記述量の多い辞書を用いたほど，概念ベースに獲得される属性数が増加している．

次に，それぞれの概念ベースについて評価を行った結果を**図3** に示す．辞書によって語義文の質は異なるが，語義文の量が多いほど概念ベースの概念ごとの属性が増し，品質を高めていると考えられる．また，4 種類の辞書の語義文をまとめて作成した概念ベースが最も良い評価を与えた．概念の属性を辞書から獲得するときには，できるだけ多量の語義文を用いることにより，属性数の多く質の良い概念ベースが得られることを示している．以降の構築および精錬の評価は，4 種類の辞書に基づく概念ベース（K）を用いて行った．

(2) **概念ベースの自己参照の効果**

概念ベースの自己参照による知識獲得法である再帰的参照（$Quote(K)$）と，逆引き参照（$Rev(K)$）を行い，式 (7) について，予備的な実験で決定した結合定数を用い，以下の式に従った線形結合を行って自己参照概念ベースを獲得した．

表4 概念ベースよりの知識獲得結果
Table 4 Results.

概念ベース Kb	平均属性数	評価値 $Eval(Kb)$
K	16	0.10
$Quote(K)$	549	0.11
$Rev(K)$	22	0.28
$Renew(K)$	562	1.02

$$Renew(K) = K + 0.2Route(Quote(K)) + 0.2Rev(K) = K + 0.2Route(K^2) + 0.2K^T. \quad (19)$$

参照によって得られる属性の関連性は，参照される属性の関連性よりも下がると考えられるので，結合定数を小さくして属性の重みを全体的に下げた．**表4** は，自己参照のそれぞれの過程で生じた概念ベースについての評価結果である．再帰的参照（$Quote(K)$）によって，概念ベース中の単語が保有する属性数の平均が30倍に増加するが，参照による評価は元の概念ベース K とほとんど変化がない．これは，新しい属性の獲得と同時に，不適切な属性も多数獲得されたためと考えられる．一方，逆引き参照でも，評価値は元の概念ベースとあまり変化がない．それに対して，これらの参照結果を元の概念ベースと線形結合した自己参照 $Renew(K)$ では，評価値のかなりの向上が見られる．

(3) **不要属性の除去の効果**

概念ベース $Adj(K)$ 中の任意に抽出した30の概念について，人手で不要と思われる属性の重みの最大値を調べた．その結果，概念の属性数の平方根と不要属性の重みの最大値に，正の相関関係（相関係数0.78）があることが判明した（**図4**）．

この結果に基づき，概念 $Word_i$ の重みが0ではない属性の数を $Num(Word_i)$ としたとき，$Word_i$ 中の不要な属性の重みの最大値 $Maxkey(Word_i)$ を以下のように定義する．

$$Maxkey(Word_i) = 0.26\sqrt{Num(Word_i)} + 1.5 + m\sigma.$$

($Num(Word_i)$ は $Word_i$ の属性数)

σ は最適化の誤差の標準偏差であり，m は結合定数（除去パラメータ）である（図4）．m を変化させて不要属性の最大重みを決定した後，不要属性を $Adj(K)$ より除去した概念ベース $Refine(K,m)$ を作成して評価を行った（**表5**）．

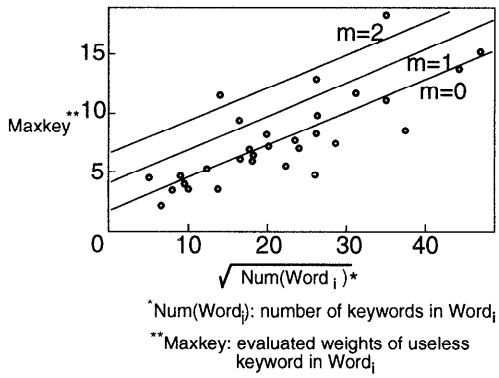


図4 概念の属性数と不要属性の重みの関係
Fig. 4 A correlation between the maximum weight of useless keywords and the number of keywords.

表5 不要属性除去の結果
Table 5 Results.

概念ベース Kb	平均属性数	$Eval(Kb)$
$Renew(K)$	562	1.02
$Refine(K,0)$	21	1.23
$Refine(K,1)$	44	1.26
$Refine(K,2)$	138	0.89

この場合，不要属性の最大重みを適切に見積もることにより（この場合は $m=1$），不要属性を除去することができる．

辞書より獲得された概念の具体的な例として，概念ベース K と精錬した概念ベース $Refine(K,1)$ とに含まれる単語「台風」の属性の一部を**表6** に示す．精錬化により，「多い」，「呼ぶ」，「最も」等の属性として不要なものが属性の上位から除去され，「風」，「低気圧」等の「台風」に必要な属性を含むことができた．

4.2 類似性判別方式の評価

過去に，1049語の概念からなる概念ベースを構築し，シソーラスを用いた類似性判別の既存技術に対して，観点に基づく類似性判別方式との比較評価を行った．評価は，類似語検索の再現率と適合率から行い，適合率の平均値が既存技術に比べて約2倍優れることを報告した[13]．ここでは，我々の提案した概念ベースによる類似性判別の比較を，観点を指定せずに行う．比較のための既存の類似性判別方式として，木構造シソーラス[25]上での概念間の距離より類似度を算出する方式を用いた．2つの概念が分類されるシソーラスのカテゴリーを隔てるカテゴリーの数を距離 l としたときに，以下の類似関数を既存手法とした．

表6 「台風」の属性（一部）
Table 6 An Example of $Word_i$ in the refined knowledge base.

順位	概念ベース K		精錬済み概念ベース $Refine(K,1)$	
	属性	重み	属性	重み
1	発生	6.0	発生	57.4
2	風速	3.0	風速	41.6
3	中心	3.0	最大	33.2
4	最大	3.0	起原	31.9
5	起原	3.0	嵐	31.6
6	多い	3.0	襲来	28.6
7	秋	3.0	風水害	28.4
8	呼ぶ	2.0	秋	27.6
9	風水害	2.0	襲う	24.4
10	夏	2.0	中心	23.3
11	襲来	2.0	九月	19.1
12	襲う	2.0	夏	18.9
13	起こる	2.0	来襲	18.8
14	起す	2.0	起こる	17.9
15	嵐	2.0	高潮	17.8
16	湾	1.0	低気圧	17.6
17	列島	1.0	台風	17.6
18	来襲	1.0	暴風	17.5
19	弱い	1.0	七月	16.6
20	最も	1.0	風	16.5

表7 概念ベース構築の評価結果
Table 7 Results.

判別方式	概念ベース Kb	評価値 $Eval(Kb)$
S(提案)	K	0.10
S(提案)	$Refine(K,1)$	1.02
S'(既存)	-	−0.13

$$S'(Word_1, Word_2) = \frac{l - L}{L} \tag{20}$$

ここで，L は，シソーラス間の距離の最大値を表し，この類似度は，先にあげた類似度関数の条件 1–4 を満たしている．

表7 は，この類似性判別判別方式に対する，概念ベースを用いた類似性判別の精度の比較結果である．評価方式は，概念ベース構築の評価方式を用いている．精錬後の概念ベース $Refine(K,1)$ を用いた提案手法ではもちろん，精錬前の4種類の辞書の語義文から構築した概念ベース K を用いた提案手法においても，既存手法より評価値が高くなっている．

4.3 類似性判別における観点の効果

変調の倍率 r（式(15)）に対する類似性判別の精度を評価した．評価法としては，シソーラス[7]を用い，類似性判別において概念の多義性の識別を必要とする方式を用いた．概念 $Word$ がシソーラスにおいて，2つのカテゴリー c_1，c_2 に属するとき，それぞれのカテゴリーの名前を観点 K_1，K_2，各カテゴリーに属する概念のうちからランダムに選んだ概念をそれぞれ $Word_1$，$Word_2$ とする（例 $(c_1, c_2, Word, Word_1, Word_2)$ = (動物, 乗り物, 馬, 豚, 自動車)）．その場合，以下のような判別が期待される．

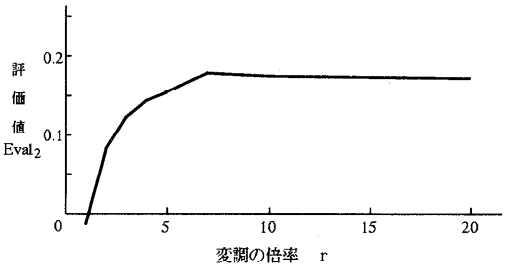


図5 観点の効果
Fig. 5 Result.

- 観点 K_1 においては，$Word$ に対して $Word_1$ が類似する．
  $S(Word, Word_1, K_1) > S(Word, Word_2, K_1)$.
  例）S(馬, 豚, 動物) > S(馬, 自動車, 動物)
- 観点 K_2 においては，$Word$ に対して $Word_2$ が類似する．
  $S(Word, Word_1, K_2) < S(Word, Word_2, K_2)$.
  例）S(馬, 豚, 乗り物) < S(馬, 自動車, 乗り物)

このような，2つの観点と3つの概念からなる評価データをシソーラスを用いて2927組作成した．このデータについて，類似度計算に基づく類似性判別を行い，判別結果が期待される結果と一致するかに基づいて評価値 $Eval_2$ を定めた．

$$Eval_2 = \frac{(\text{正しい判別結果})}{(\text{サンプル数})} = \frac{GG - BB}{2927} \tag{21}$$

ここで，GG とは，2つの観点とも期待された判別がなされた評価データ数を示し，BB は，2つの観点とも期待された判別がなされなかった評価データ数を表す．$Eval_2$ は，−1 から 1 の値をとり，値が大きなほど観点に応じた類似性判別を行っていることを示す．また，値が0のときには，ランダムな判別と同じ評価を意味する．

評価結果を**図5**に示す．評価値 $Eval_2$ は，変調の倍率が5倍付近において最大値をとり，観点が多義語の類似性判別に有効であることを示している．また，変調の倍率が1では観点がない状態であり，評価値は0付近を示し，観点に応じた類似性判別ができていないことを示している．

5. 類似研究との比較

ここでは，これまでに説明した概念ベースに基づく類似性判別方式について，関連する研究との比較を行う．

辞書から獲得した語彙知識を用いて単語をベクトル空間中で表現し，この単語間の類似性を判別する研究は種々行われてきたが，「文脈や状況」に対応した判別を行おうという研究は少ない．ここでは文献11)と12)の研究を取り上げ，違いを論じる．

まず文献12)の「動的シソーラス」では，「文脈や状況」を単語集合Cで表し，これらC中の単語がベクトル空間中でクラスタを成すように，つまり互いに類似するように，ベクトル軸のスケールを変換している．しかし，文脈を表す単語は必ずしも「その文脈において類似した単語」とは限らない．たとえば，その文脈の様々な側面を表す単語群である場合等がある．これに対して本論文では，文脈を「観点」と呼ぶ1つの単語で表し，この観点のベクトル中の成分の大きな軸について，類似性を比較するベクトルの成分を強調している．ある軸の成分を強調することは，その軸のスケールを変換することと等価であるが，上述のようにその軸の選び方がまったく異なる．また，ここでは観点は1つの単語だが，2つ以上の単語を観点とする場合にも容易に拡張が可能である．逆に文献12)の手法では，集合Cが1つの単語では定義できない．つまり，本手法の方が「文脈や状況」をより一般的に取り扱えると考えられる．

次に文献11)の「意味の数学モデル」であるが，文脈を同様に複数の単語で表し，これら単語に関連する軸だけからなる部分空間中で類似性を判別している．この手法は，いわば本論文の手法で「観点変調」を極端に大きくした場合に相当する．しかし，各々の単語で観点に関係しない部分をまったく考慮しないのは問題であり，事実，我々の予備検討[26]では，こうした方法では評価が下がることを報告した．つまり，ユーザが指定する「文脈や状況」を構成する単語は，その一部分を表現しているのみであり，それら単語のみで「文脈や状況」のすべてを表すことはできない．したがって，観点として与えられた単語に関わる属性の重みを強調し，それ以外の属性はそのまま保持する本手法の方が，より現実的な扱いであるといえよう．以上が文献11)，12)の手法と，本論文の違いである．

このほか，本論文は語彙知識を機械的に構築する際の「精錬」の手法を大きな特徴とする．語彙知識を獲得する辞書は，元々計算機処理のために作られたものではないため，その記述から意味を機械的に取り出すことには種々の困難がある．これは日本語の辞書の場合，語義文の形態素解析が難しい等の理由から特に顕著であった．従来のこの種の研究は，主に語義文の解析精度を上げる方向で，これに対処しようとしてきた．しかし，本来，単語の意味というものは，ある言語を表す単語全体で意味を持つはずである．この考えに基づいて，まず辞書からラフな知識を獲得し，次にこの知識全体を利用して知識の質を高める手法とした．これが本論文で述べた「精錬」であり，我々が独自に提案したものである．なお，文献12)にある「活性伝搬」の手法は，本論文における「再帰的参照」に似た手法であるが，これは人間の記憶構造をモデル化したものと考えられ，本論文の精練手法とは考え方も実際の操作も異なる．

最後に，本論文で採用した様々な仮説/パラメータ等について述べる．本論文では，概念の知識表現をはじめ，精練手法・観点変調・類似計算等々の様々な提案を行った．これらはその多くが従来にない新しい方式であるがゆえに，個々の方式における種々の前提やパラメータ，たとえば概念ベースどうしの加法性の仮定等については，その十分な根拠を示すことができなかった．しかし，ここで主張したかったのは，こうした個々の方式の最適性ではなく，辞書の語義文といういわば非工学的なデータから「工学的に有用なデータ」を構築し，単語間の類似性判別を行うことができるという全体の枠組みである．方式の詳細な最適化については，今回得られた概念ベースを第一ステップとして利用しつつ，段階的に研究を進めていく計画である．

6. おわりに

本稿では，概念の知識を辞書等より自動獲得して得られた概念ベースを用い，観点に応じた概念の類似性判別方式の提案を行った．また，実際に構築した4万規模の日常語の概念に関する概念ベースによる類似性判別をシソーラスを用いた類似性判別と比較評価し，機械的に構築した単純な構造の概念ベースによる提案手法であっても有効であることを示した．さらに，観点の効果を調べる実験を行い，多義性の高い日常語の概念の類似性判別に本方式が有効であることを明らかにした．

今後は，概念ベースの構築に統計手法を採り入れ，概念の質の向上を目指す．また，得られた方式を元に，数十万語規模の概念ベースの構築を目指す．そして，談話や文章より観点を求める方式の検討を行う予定である．

謝辞 概念ベースの作成および評価実験に協力していただいたNTT-AT（株）金杉友子さんに感謝いたします.

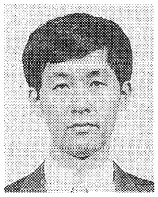
参考文献

1) 大津展之ほか：特集「リアルワールドコンピューティング研究計画」，情報処理学会誌，Vol.34, No.12, pp.1423–1448 (1993).
2) 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，Vol.SIG-J-9401, pp.103–110 (1994).
3) Tversky, A.: Features of Similarity, *Psychological Review*, Vol.84, pp.327–352 (1977).
4) Suzuki, H., Ohnishi, H. and Shigemasu, K.: Goal-directed Processes in Similarity Judgements, *Proc. 14th Annual Conference of the Cognitive Science Society*, pp.327–352 (1992).
5) 岩山 真，徳永健伸，田中穂積：比喩を含む言語理解における顕現性の役割，人工知能学会誌，Vol.6, No.5, pp.674–681 (1991).
6) 沢田裕司，大川剛直，馬場口登，手塚慶一：観点を考慮した連想機構の一モデル化，情報学基礎，Vol.28, No.2, pp.9–16 (1992).
7) 大野 晋，浜西正人：類語国語辞典，4th edition，角川書店 (1990).
8) Hindel, D.: Noun Classification from Predicate-Argument Structures, *Proc. ACL*, pp.268–275 (1990).
9) Pereira, F., Tishby, N. and Lee, L.: Distributional Clustering of English Words, *COLING-93*, pp.183–190 (1993).
10) Grishman, R. and Sterling, J.: Generalizing Automatically Generated Selectional Patterns, *COLING-94*, pp.742–747 (1994).
11) 北川高嗣，清木 康，人見洋一：意味の数学モデルとその実現方式について，信学技報，Vol.DE93-4, pp.25–31 (1993).
12) 小嶋秀樹，伊藤 昭：意味空間のスケール変換による動的シソーラスの実現，信学技報，Vol.NL95, No.19, pp.1–8 (1995).
13) 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, No.3, pp.505–509 (1994).
14) Kasahara, K., Ishikawa, T., Matsuzawa, K. and Kawaoka, T.: Viewpoint-based Measurement of Semantic Similarity between Words, *Proc. 5th International Workshop on Artificial Intelligence and Statistics*, pp.292–302 (1995).
15) 安西祐一郎：認識と学習，岩波書店 (1989).
16) 笠原 要，藤本和則，松澤和光，石川 勉：精錬に基づく概念ベース構成法，信学技報，Vol.DE95-7, pp.49–56 (1995).
17) Kasahara, K., Matsuzawa, K. and Ishikawa, T.: Refinement Method for a Large-scale Knowledge Base of Words, *Working Papers of the Third Symposium on Logical Formalizations of Commonsense Reasoning*, pp.73–82 (1996).
18) Salton, G. and McGill, M.: *Introduction to modern Information Retrieval*, McGraw-Hill (1983).
19) 日本ユニバック総合研究所（編）：共立コンピュータ辞典，共立出版 (1976).
20) 武部良明（編）：必携類語実用辞典，三省堂 (1977).
21) 三省堂編修所（編）：必携用事用語辞典，4th edition，三省堂 (1992).
22) 見坊豪紀（編）：三省堂現代国語辞典，2nd edition，三省堂 (1992).
23) 松村 明，三省堂編修所（編）：大辞林，三省堂 (1992).
24) 新村 出（編）：広辞苑，岩波書店 (1992).
25) 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，Vol.84-13, pp.95–102 (1991).
26) 笠原 要，松澤和光，湯川高志，石川 勉，河岡 司：アバウト推論のための多観点概念ベース：構築と評価，人工知能学会全国大会，pp.11–3 (1993).

(平成8年4月19日受付)
(平成9年5月 8 日採録)

笠原 要（正会員）

昭和39年生．平成3年東京工業大学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．



松澤 和光（正会員）

昭和28年生．昭和52年東京工業大学大学院工学研究科電子工学専攻修士課程修了．同年電電公社入社．以来，フルウェーハシステム，大規模ROM，ヒューマンインタフェイス，知識処理技術の研究に従事．現在，NTTコミュニケーション科学研究所グループリーダ．IEEE，電子情報通信学会，人工知能学会，言語処理学会，ファジー学会各会員．

石川　勉（正会員）

昭和22年生．昭和45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．以来，主記憶装置，フルウェーハシステム，並列プロセッサ，知識処理技術の研究に従事．平成7年より拓殖大学工学部教授．IEEE，電子情報通信学会，人工知能学会各会員．