言葉の意味の類似性判別に関するシソーラスと概念ベースの性能評価

An Evaluation of a Knowledge Base of Words and Thesauruses on Measuring the Semantic Similarity between Words

川島 貴広

Takahiro Kawashima

拓殖大学工学部情報工学科

Department of Computer Science, Takushoku University

hrnt@rivo.mediatti.net

石川 勉

Tsutomu Ishikawa

(同上)

ben@cs.takushoku-u.ac.jp

keywords: semantic similarity, knowledge base of words, thesaurus, similarity calculation method

Summary

We have developed a knowledge base of words as a tool to measure the semantic similarity between words. In this paper, we evaluate the knowledge base of words comparing with thesauruses, which are commonly used for measuring similarity. Thesauruses of NIHONGO-GOI-TAIKEI(NGT) and Japan Electronic Dictionary(EDR) are selected for the evaluation. For similarity calculation methods using thesauruses, we adopt a newly proposed method, in which each word is represented with vector using the structural feature of thesauruses and the degree of similarity between words is calculated by the inner product of their vectors, in addition to traditional methods based on the path length between categories or the depth of the subsumer. Evaluation is carried out through the two methods, that is, a traditional method based on human rating and the method we have already proposed, feasible for evaluating automatically without human judgment. Evaluation result shows that the knowledge base of word is superior to the both thesauruses(NGT outperforming EDR) as measurement tools, and the proposed calculation method outperforms the traditional ones. The result also shows that our evaluation method is a practical one, by investigating the correlation of both methods.

1. はじめに

現在の情報化社会において、インターネット上に氾濫する情報の中から、必要な情報を如何に効率的かつ的確に抽出するかは極めて重要な課題である。このため、これまで情報検索技術や自動要約技術に関するさまざまな研究がなされてきている [奥村 04, 奥村 02]。情報検索における基本的アプローチは、端的に言えば検索キーワードを含む文書の抽出といえるが、最近は検索精度の向上をめざし、ユーザの評価をフィードバックした段階的な検索 [中島 01, Rocchio 71]、単純なキーワード照合でなくその意味的な類似性をも考慮した検索 [飯田 98] 等について研究されている。特に後者、すなわち、意味的な類似性の利用は、必ずしもユーザが検索に最適なキーワードを選択できないことを考慮すると必須の要件になってくる。一方、自動要約では、テキストからの重要文抽出が核となるが [平尾 03]、この抽出においても文間の意味的な類似性を判定し、その結果を利用することが抽出精

度の向上につながると言える。このような意味的な類似性の判定は、一般的にキーワードあるいは文の構成要素である単語 (概念) 間の意味的な類似性を統合することによりなされる。

意味的な類似性の利用は、以上のような自然言語処理分野の応用技術だけでなく、コンピュータにより人間らしい処理をさせようとするあらゆる分野で必要になる。例えば、人工知能分野の核である推論技術では、類似性に着目した推論法として類推 [原口 86] や事例ベース推論 [松原 92] さらには不足知識を類似知識で補完し概略解を得る概略推論法 [Nguyen 01] 等が研究されている。これら推論法では、それぞれ、事物間の関係の類似性、事例間の類似性、知識間の類似性と、利用する類似性の形態は異なるものの、その類似性判定においてはすべて単語 (概念) 間の意味的な類似性が基本となる。

このような単語の意味の類似性判定には、これまでは専用のツールがないため、広く自然言語処理全般での利用を想定して構築されたシソーラスが用いられていた [長

尾 96]。シソーラスは、全単語を意味的に類似する単語の集合であるカテゴリに分類し、各カテゴリを is-a 関係や part-of 関係により階層的に関連づけたものであり、日本語単語については日本語語彙体系 [池原 97] や EDR 電子化辞書 [EDR 93] が、英単語については Word-Net [Miler 95] が代表的である。このシソーラスを単語の意味の類似性判定に用いる場合、一般的に類似の程度（以下、類似度と呼ぶ）は各単語が属するカテゴリ間のシソーラス上での距離やそれらカテゴリの共通上位カテゴリの位置等を基に算出されている [長尾 96, 平川 03]。

これに対し、我々は単語の意味の類似性判定を主目的とした概念ベースを構築してきた [笠原 97, 石川 98, Nguyen 02]。概念ベースは、国語辞書の語義文の情報をもとに構築されたもので、各単語（概念）は日本語語彙体系のカテゴリを基底とするベクトル（以下、概念ベクトルと呼ぶ）で表現される。前述のように is-a 等の固定的な視点により分類されているシソーラスでは、一定の視点からの類似性しか判定できないが、概念ベースではこのベクトル表現により、原理的には視点の情報を基に概念ベクトルを変調することによりあらゆる視点からみた類似度の算出が可能となり得る [笠原 97, 北川 93]。例えば、“馬”、“豚”、“車”という概念を考えたとき、“動物”という視点からみれば“馬”と“豚”が類似するが、“移動”という視点からみれば“馬”と“車”が類似することになる。このような判定がある程度可能なことが概念ベースの大きな特徴といえる。

これらシソーラスや概念ベースの単語の意味の類似性に関する判別能力（以下、類似性判別能力）の評価は、これまで単独にかつ別々の尺度でなされてきており [平川 03, 笠原 97, 石川 98]、同一の尺度での比較評価はなされてこなかった。本論文の目的はこの比較評価にある。具体的には本論文では、各種類似度計算法に基づくシソーラスの類似性判別能力と概念ベースのそれを、既に提案している機械化が可能な評価法と一般的に用いられている人間の感覚による評価法により比較評価する（以下、前者を機械的評価法、後者を感覚評価法と呼ぶ）。比較対象とするシソーラスとしては、日本語単語については最も大規模かつ代表的といえる日本語語彙体系と EDR 電子化辞書のシソーラスを選ぶ。さらにシソーラスについては、従来の距離や共通カテゴリ位置をベースとした計算法に対し、新たに単語が属するカテゴリとシソーラスの構造から得られる情報をもとに各単語をベクトル表現し類似度を算出する方法（以下、ベクトル化法と呼ぶ）を提案する。この方法はベクトル表現を用いているため、概念ベースと同様に視点を考慮した類似性判別の可能性を有する。また、感覚評価法による評価結果との相関をとることにより、機械的評価法自体の評価を行う。

以下、2章では、比較対象である日本語語彙体系と EDR 電子化辞書のシソーラスおよび概念ベースの構成について述べた後、それらにおける一般的な類似度計算法と提

案するベクトル化法について示す。また、人間の感覚が指数的事であることから、類似度を指数関数を用いて算出する方法を示す。3章では、機械的評価法と感覚評価法について述べる。4章では、各種の類似度計算法と前記2種の評価法のあらゆる組み合わせについての評価を行う。また、ベクトル化法についてはベクトル表現する場合の最適条件を示す。5章では、機械的評価法自体の評価を行うとともに、シソーラスと概念ベースの類似性判別能力について、得られた評価結果を総括する。

2. 評価対象と類似度計算法

2.1 評価対象

§1 シソーラス

評価対象として選んだ、日本語語彙体系 [池原 97] と EDR 電子化辞書 [EDR 93] (以下、前者を NGT, 後者を EDR と略す) について簡単に紹介する。

(1) NGT

日本電信電話株式会社 (NTT) が、日英機械翻訳システムの実現を目指して開発したシソーラスであり、2,715 のカテゴリからなる最大 12 段の木構造を持つ。NGT シソーラスの一部を図 1 に示す。同図に示すように頂点のノードは“名詞”という概念であり、その下位になるごとに、より具体的な物事を表わす概念となっている。日本語語彙体系では約 30 万個の単語（概念）がこれら 2,715 のカテゴリのいずれに属するかが示されている。

(2) EDR

EDR 電子化辞書は、コンピュータの先進的な言語処理のために開発され、そのシソーラスは、202,797 のカテゴリからなる最大 17 段のグラフ構造を持つ。EDR シソーラスの一部を図 2 に示す。各ノードは単語（概念）に対応し、概念識別子と呼ばれる 16 進の数字で表現される。グラフ構造のため、同図に示すように 1 つのカテゴリが複数の上位カテゴリを持つ（以下、多重継承と呼ぶ）ことがある。

図 1 NGT シソーラスの一部

図 2 EDR シソーラスの一部

§ 2 概念ベース

概念ベースは、国語辞書の語義文を用いて構築されている [笠原 97, Nguyen 02]。図 3 に概念ベースの構成イメージを示す。概念ベースでは、見出し語を概念とし、各概念について語義文中の独立語を抽出し、それを各概念の元の属性としている。さらにこの元となる属性を用いて孫引き、逆引きを行い、新たな属性を付加することで属性数の増加を図っている。その後、属性は日本語語彙体系の 2,715 のカテゴリにまとめられる。属性の重み（属性値）は、語義文中での属性の出現頻度および概念間での出現頻度から、基本的には $tf \cdot idf$ [Salton 94] の考え方に基いて設定している。また、明らかにノイズと考えられる属性は人手で除く等により属性値の精練を行っている。

こうして構築された概念ベースでは、各概念は 2,715 次元のベクトルで表現されることになる。なお、属性値はその 2 乗和が 1 になるように正規化されている。概念ベースの詳細については文献 [笠原 97], [Nguyen 02] を参照されたい。

図 3 概念ベースの構成イメージ

2.2 類似度計算法

§ 1 シソーラスでの計算法

これまでシソーラスを利用した類似度計算法としては、冒頭にも述べたように、比較対象の概念が属するカテゴリ間の距離やその共通の上位カテゴリの段数をもとに計算されていた [長尾 96, Li 03] (以下、前者を距離法、後者を段数法と呼ぶ)。ここでは、従来法について述べた後、新たにシソーラスの構造的な特徴を利用したベクトル化法を提案する。

(1) 従来法

距離法、段数法について、図 4 を用いて説明する。同図で A, B は対象概念であり、a, b は対象概念のカテゴリである。また、対象概念を A, B としたときの距離法、段数法の類似度を $R_D(A, B)$, $R_L(A, B)$ とする。なお、対象概念は、そのカテゴリの下位に存在すると考え、その分も 1 段とし計算を行う。

(a) 距離法

距離法では、下式により類似度を算出する。

$$R_D(A, B) = \frac{1}{D_{AB} + 1} \quad (1)$$

ここで、 D_{AB} は図中の矢印のように概念 A, B 間のシソーラス上の距離（カテゴリ間の枝の数）である。図中の例では、距離が 5 であるので $R_D(A, B)$ は $1/6$ となる。

(b) 段数法

段数法では、下式により類似度を算出する。

$$R_L(A, B) = \frac{C_{AB} \times 2}{L_A + L_B} \quad (2)$$

ここで、 L_A, L_B は根カテゴリ（図中のカテゴリ T）を 1 段とし、そこから、1 つ下位カテゴリになるごとに 1 を加算してカウントした概念 A, B のシソーラス上の段数である。また、 C_{AB} は対象概念 A, B の共通の上位ノード（図中の C）の段数である。同図の例では、 L_A は 5, L_B は 4, C_{AB} が 2 なので、 $R_L(A, B)$ は $4/9$ となる。

なお、これらの方法では、概念に多義が存在しその概念が複数のカテゴリに属する場合、類似度が一意に定まらない。従ってここでは、計算可能なすべての組み合わせで類似度を計算し、その中で類似度が最大、最小、平均となる 3 通りの場合について比較評価する。

(2) ベクトル化法

図 4 距離法、段数法による類似度計算法

ベクトル化法とは、シソーラスのカテゴリに属する各概念を、後述する手順に基づいて（ステップ1～3）抽象化 [平川 03] したカテゴリ数次元のベクトルで表現する方法である。ベクトルの値としては、シソーラス構造に基づいて各カテゴリに適切な重みを付与することにより決定する。類似度はこのベクトルの内積で計算する。以下、この方法について図5、図6を使って説明する。図5は多重継承と3段に抽象化した際の例を示しており、図6は重みの付与についての説明図である。ここで両図中のaは類似度を求める対象概念のカテゴリ、bは重みを付与するカテゴリ、Tは上位カテゴリが存在しない根カテゴリである。

ステップ1：ベクトルの次元数の削減

各概念をシソーラスの全カテゴリ数次元のベクトルで表現した場合、カテゴリ数が多いシソーラスではベクトルのサイズが大きくなりすぎることになる。従って、カテゴリの抽象化によりベクトルの次元数の削減を図る。抽象化は元のカテゴリをより上位のカテゴリにマップする操作であり、ベクトルの次元数は元のカテゴリ群を適切な粒度のカテゴリ群で代表させることにより削減される。この抽象化の方法としては、文献 [平川 03] における均等深度法を採用する。均等深度法とは段数がN以上であるカテゴリを、段数Nに位置するカテゴリに抽象化する方法である。なお、多重継承がある場合には、複数の上位カテゴリにマップされることがある。例えば、シソーラスを3段に抽象化する図5では、カテゴリaは、ルートR₁ではC₂、ルートR₂ではC₄にマップされることとなる。

ステップ2：初期重みの付与

対象概念のカテゴリ（図6中のa）に対して初期重み“1”を与える。また、多義や抽象化する際の多重継承により複数のカテゴリが対象概念のカテゴリとして存在する場合は、それらすべてのカテゴリに対して初期重みを与える。例えば、図5では、aに対しC₂、C₄のそれぞれに初期重み“1”を与える。

ステップ3：上位カテゴリへの重みの付与

対象概念のカテゴリの上位カテゴリは、その概念に意味的に関連するので所定の重みを付与すべきといえる。この場合、上位カテゴリほど意味的に幅広い概念であり、情報量は少ないため、初期重みと同じ重みを与えるべきではない。従って、ここでは情報量の比により重みを減らし、下式により重みを付与することとする。

$$W_b = W_a \times \frac{\log\{(C_b + 1)/N_c\}}{\log(1/N_c)} \quad (3)$$

ここで、W_bは重み付与の対象となるノードbの重み、W_aはノードaの重み（初期重み）、N_cはカテゴリ総数、C_bはノードbの下位カテゴリの総数である。ただし、重み付与の対象となるカテゴリに既に重みが存在する場合、または、多数の重みが下位カテゴリから上がってくる場合には、最大の重みをそのカテゴリの重みとする。例え

図5 多重継承と抽象化の考え方

図6 ベクトル化法による重みの付与の考え方

ば、対象概念が多義により図6のカテゴリaだけでなくカテゴリsにも属していたとすると、これらカテゴリから同図の矢印の2つのルートで上がってくる重みをそれぞれ上式を用いて計算し、大きい方の値をカテゴリbの重みとする。あるいは、対象概念が同じく多義によりカテゴリaとカテゴリb自体に属していたときは、カテゴリaから上がってくる重みとカテゴリb自体にある初期重みの大きい方をカテゴリbの重みとする（ここでは初期重みは“1”としているのでカテゴリbの重みは“1”となる）。また、多重継承の場合にも全ての上位カテゴリに対し重みを付与する。例えば、図5のように多重継承しているシソーラスを3段で抽象化する場合、同右図のように木構造に展開し、カテゴリC₂、C₄の上位カテゴリC₁、C₃に対しそれぞれ重みを付与する。

なお、以上の処理を類似度計算時に行うのではなく、あらかじめ全ての概念について行っておけば、概念ベースと同様に扱えることは言うまでも無い。

§2 概念ベースでの計算法

概念ベースでは、比較する概念同士のベクトルの内積で類似度を算出する。値は0～1（同一概念間で1）の実数で表わされる。本来、内積を用いる場合はベクトルの基底が直交している必要があるが、本概念ベースにおけるベクトルの基底は前述したように日本語語彙体系の2,715のカテゴリであり、必ずしも意味的に直交するとは言えない。しかし、これに対してはその構築時に各種の対処策を講じており（これら詳細については、文献 [Nguyen 02] を参照されたい）、ここでは単純に内積で算出することとする。

§3 指数関数による類似度

人間の感覚は指数的であるため、類似度も指数的に扱うことで性能向上が期待できる。Li 等は、Word-net を評価対象とした各種の類似度計算法について評価し、指数的な扱いの優位性を示している [Li 03]。本論文では、2・2・1 節と 2・2・2 節で得た類似度による直接的な評価に加え、同文献で使用されている以下の指数関数 (式 (4)) とこれを簡略化した指数関数 (式 (5)) により変換した類似度 R を用いた評価も行う。

$$R = \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}} \quad (4)$$

$$R = \frac{1 - e^{-\beta x}}{1 + e^{-\beta x}} \quad (5)$$

ここで、 x は前節までの方法により得た変換前の類似度、 β は定数である。 β はサンプル概念を用いた評価を通して最適な値に設定する。すなわち、これは各評価対象と類似度計算法の組み合わせ毎に設定することになる。

3. 評価法

3.1 機械的評価法

(1) 考え方

シソーラスや概念ベース（以下、類似度判定ツールと呼ぶ）を単語（概念）の意味の類似性判別に利用する場合、これらの特性としては、

- イ) 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。
- ロ) 2 つの類似する概念が存在する場合、どちらが類似しているかを正確に識別可能。

が重要である。前者は類似度の定義域（ここでは 0 ~ 1）の有効な活用を、後者は微妙な判定の高精度な実現を意味し、これら 2 つの特性を同時に満足するツールほど高性能な類似性判別ツールであると言うことができよう。我々は、これらを考慮した評価法を既に文献 [石川 98] で提案しているが、ここではこの方法を若干改良して用いることとする。以下、この評価法について簡単に説明する。

本評価法は、まず、対象概念 G に対し、それに類似する概念 G_1 、比較的類似する概念 G_2 、非類似概念 G_3 をそれぞれ用意しこれらを 1 組とし、これを N 組つくり評価データとする (図 7 参照)。次に、前述のイ)、ロ) の特性を判定可能な評価指数をそれぞれ設定し、用意した評価データに対してこれら評価指数を計算しその結果を統合して類似性判別能力の評価とするものである。評価指数は以下のように設定する。なお、 $G - G_1$ 間、 $G - G_2$ 間、 $G - G_3$ 間の類似度をそれぞれ r_1, r_2, r_3 とする。

(2) 評価指数

イ) に対する評価指数:

基本的には、 $r_1 - r_3$ が大きい程良いが、類似度判定ツールあるいは類似度計算法により、類似度の値に特定の傾向が出ることが予想される。例えば、シソーラスの

距離法では、同一概念間の類似度は 1 であるが、次に類似する概念間の類似度は式 (1) より $1/3$ となってしまう。すなわち、類似度の値が小さくなる傾向があると言える。よって、これらを考慮し、評価指数として以下を設定する。

$$F_1 = \frac{(\bar{r}_1 - \bar{r}_3)}{(1 + \sigma_1 + \sigma_3)} \quad (6)$$

ここで、 $\bar{r}_1, \bar{r}_3$ と σ_1, σ_3 は、それぞれ r_1, r_3 の平均と標準偏差である。上式の分母は前述の傾向を考慮したもので、類似度の値が小さくなる場合はそのバラツキも小さくなる傾向があり、逆の場合は逆の傾向がでるとの仮定に基づくものである。なお、バラツキを考える際、重要となるのは、 r_1 については $\bar{r}_1$ より小さい方のバラツキであり、 r_3 については $\bar{r}_3$ より大きい方のバラツキである。従って、標準偏差の算出にはこの方向のデータのみを用いる。

ロ) に対する評価指数:

r_1 と r_2 の関係としては基本的には $r_1 > r_2$ でなければならない。従って、評価指数として以下を設定する。

$$F_2 = \frac{m}{N} \quad (7)$$

ここで、 m は前述の関係が成立した組の数である。ただし、この判定は有意差を考慮して行われるべきと考え、以下の基準で行う。その結果を評価データ N 組について総計し、 m の値とする。

$$\begin{aligned} r_1 > r_2 + \alpha &: 1 \quad (\text{正解}) \\ r_2 + \alpha \geq r_1 > r_2 - \alpha &: 0.5 \\ r_2 - \alpha \geq r_1 &: 0 \quad (\text{誤り}) \end{aligned}$$

ここで、 α は判定のための有意差であり、これについては、統計的考えを取り入れ、区間推定を用いて設定することとした。具体的には、 r_1, r_2 について普遍分散 s^2 を求め、それぞれについて以下の式により、 α_1, α_2 を算出する。

$$\alpha_i = t_{\delta/2} \times \frac{s_i}{\sqrt{N}} \quad (8)$$

ここで、 $t_{\delta/2}$ は信頼度係数 $1 - \delta$ における自由度 $N - 1$ の t 分布表より求めた係数である。こうして得た α_1, α_2 の和を α とする。この有意差を考慮するようにしたこと が文献 [石川 98] からの改良点ある。

F_1, F_2 をどのように統合し類似性判別能力を判定するための最終的な評価指数 F とするかについてはいろいろな方法が考えられるが、ここでは以下のように単純に

図 7 対象概念とその評価に用いる概念の関係

表 1 人間の与える類似度の基準

類似度	判断基準
4	ほとんど同じ (同義)
3	よく似ている
2	少し似ている
1	似ていないが関連性はある
0	全く無関係

それらの積とすることとした．この評価指数 F の値は，理想的な類似度判定ツールで 1 となる．

$$F = F_1 \times F_2 \quad (9)$$

本評価法は，いずれかの方法により評価データが与えられれば人手を介さない機械的な評価が可能となるもので，大量のデータによる評価を行う場合には極めて有効な方法と言えよう．

3.2 感覚評価法

これまで類似度判定ツールやそれを用いた類似度計算法の評価は，被験者に評価データとなる 2 概念間の類似性の程度を判定してもらい，その結果と比較する等，人間の感覚に基づいて行われるのが一般的であった [平川 03, Li 03]．ここでは，文献 [Li 03] で用いられている方法を利用する．

この方法（感覚評価法）は，複数の概念組に対して人間の感覚により 0～4 の 5 段階で類似度を与え，その値と評価対象となる類似度計算法で得られた類似度との相関をとり評価するものである．具体的には，まず評価概念組に対して，あらかじめ決めておいた指標に基づいて複数の被験者に 5 段階のいずれかで類似度を与えてもらう．それにより得た各類似度の平均値をその概念組に対する人間の感覚による類似度とする．こうして得た類似度と評価対象となる類似度計算法により得られた類似度との相関係数 M を求め，この値により類似性判別能力の評価を行う．ここでは，5 段階の評価基準として表 1 を用いることとした．

4. 評価結果

評価対象である類似度判定ツール“概念ベース”，“NGT”，“EDR”について，以下の項目全ての組み合わせについて評価を行った．ただし，概念ベースでは計算法が 1 種なので“類似度”と“評価法”の組み合わせについて評価した．

類似度計算法	： 距離法 (最大, 最小, 平均), 段数法 (最大, 最小, 平均), ベクトル化法 (抽象化段数をパラメータ)
類似度	： 変換なし, 指数変換 (式 (4), 式 (5))
評価方法	： 機械的評価法, 感覚評価法

対象 概念G	類似 概念G1	比較的類似 概念G2	非類似 概念G3
200組 全力 視野 悪臭 絶食 歓声 跳躍 熟睡 近道 失踪 行為 : :	総力 視界 異臭 断食 歓呼 飛躍 仮眠 早道 失跡 行動 : :	人力 視線 臭い 試食 声色 快速 仮眠 通行 尾行 言動 : :	鉦物 戦死 船本 貯蓄 睡号 学友 球橋 : :

図 8 評価データの一部

4.1 評価データ

評価データをどのように設定するかは評価の信頼性を左右する重要な事項と言える．すなわち，機械的評価法では対象概念 G と，それに類似する概念 G_1 ，比較的類似する概念 G_2 ，非類似概念 G_3 の組をどのように設定するかである．評価対象である“NGT”，“EDR”のカテゴリを利用し，例えば G と G_1 を同一カテゴリに属する概念から選ぶ等々も考えられるが，この場合評価の公平性が保たれないことになる．従って，ここではこれらと作成者が異なる情報源であること，意味的に類似する概念が明確に示されていること等の視点から，類語辞典（これも段数の少ないシソーラスであるが）を利用しこれをもとに評価データを設定することとした．

具体的には，類語辞典として類語例解辞典 [遠藤 94] を選び，まずこの中より対象概念 200 個をランダムに選んだ．次に，これら 200 個の各概念について，類似概念は同辞書で類語グループを構成する概念の中から，比較的類似概念は中分類が同一の概念の中から，非類似概念は大分類が異なる概念の中からランダムに選んで評価データとした．図 8 に評価データの一部を示す．

以上のように抽出した 200 組の評価概念により評価を行うが，類似度を指数変換する場合には定数 β の値を決める必要がある．従って，ここでは，200 組の評価概念を 100 組ずつに分け，片方の 100 組で最適な β を求め，残りの 100 組に対して評価を行う．さらに，これらの評価概念を入れ替えた場合についても評価を行い，最終的な評価値はそれらの平均値とすることとした．

また，感覚評価法でも同じ評価データを利用することとした．具体的には，200 組の各評価概念に対して， G 、 G_1 、 G_2 、 G_3 をそれぞれを 1 つの概念組として計 600 組を構成し，これらを前述の理由から半分の 300 組ずつに分け使用した．被験者はこの概念組みに対して表 1 の基準により類似度を与える．また，この評価は 14 人の被験者で行った．この感覚評価法で機械的評価法と同じ評価データを利用したのは，前者での結果により，機械的評価法自体を評価するためでもある．なお，EDR 電子化辞書に関しては，サ変名詞が動詞で登録されているので，評価概念に“する”をつけて評価した．

図 9 ベクトル化法評価結果 (機械的評価)

4.2 各評価法における評価結果

§ 1 機械的評価法による評価結果

(1) ベクトル化法の最適抽象化段数

抽象化段数に対する評価値 F の変化を、それぞれ図 9 に示す。同図では合わせて抽象化によるカテゴリ数の変化についても示している。同図 (a) より、NGT のベクトル化では、指数変換なしの場合が 11 段、他の 2 つの指数変換法では 12 段が最良となり、あまり抽象化しない方が良いという傾向があることが分かる。一方、EDR のベクトル化では、同図 (b) より、指数変換なしの場合が 7 段、他の 2 つの指数変換法では 17 段で最良となっている。変換なしでは最適なベクトルの次元数があるが、変換する場合には抽象化しない方が良いと言える。

また、ベクトル化した場合の 1 概念ベクトルあたりの非零の要素数について調べた。EDR について評価に用いた 800 概念についての分布を図 10 に示す (NGT もほぼ同様の傾向であった)。非零の要素数は 5~9 個以上になることが分かる。なお、非零の要素数の平均は、NGT の 11 段、12 段で 10.4 個、EDR の 7 段で 11.1 個、17 段で 14.1 個であった。

(2) シソーラスの類似性判別能

図 11 に評価結果を示す。同図より、シソーラスとしては、EDR より NGT の方が全体的に高い評価値が得られることが分かる。また、類似度計算法としては、ベクトル化法、段数法、距離法の順に優れ、距離法、段数法で

図 10 EDR ベクトル化での各概念の非零要素数分布

は類似度が最大の場合の条件が最良であると言える。さらに、ベクトル化法、距離法では、類似度を指数的に扱った方が良い結果が得られている。

(3) 概念ベースとシソーラスの比較

概念ベースでの評価結果を、シソーラスによる各種計算法と比較しながら図 12 に示す。シソーラスにおいては、最適の条件の場合 (図 11 で評価値が最大になる場合) での評価値を示している。同図から、概念ベースがいずれのシソーラスより高い評価値を得ていることが分かる。また、概念ベースでも類似度を指数的に扱った方が好ましいことが分かる。なお参考までに、指数変換なしの場合の単体の評価指数 F_1 , F_2 における比較を図 13 に示す。いずれの評価指数においても概念ベースの方が両シソーラスより高い評価値を得ている。なお、指数変

図 11 シソーラスによる方法の類似性判別能力比較 (機械的評価)

図 12 機械的評価法による類似性判別能力比較

図 13 機械的評価法による評価指数 F1, F2 の比較

換を行った場合にも同様の結果が得られている。

(4) 信頼係数の変化

以上は、3・1 節 (2) の信頼係数を 90% として評価した結果であるが、これを変化した場合の影響についても調べた。ここでは、類似度を指数変換しない場合の評価値 F を例として図 14 にその影響を示す。同図より、信頼係数をこの程度の範囲で変化させても、評価値はあまり変わらないことが分かる。また、指数変換した 2 つの方法に関しても同様な結果が得られている。

§ 2 感覚評価法による評価結果

(1) ベクトル化法の最適抽象化段数

抽象化段数に対する評価値 M (相関係数) の変化を、図 15 に示す。同図より、NGT のベクトル化では、指数変換の有無に関わらず 12 段で最良となり、抽象化しない方が良い傾向があると言える。一方、EDR のベクトル

図 14 信頼係数の変化による評価値への影響

(a) NGT

(b) EDR

図 15 ベクトル化法評価結果 (感覚評価)

化では、指数変換なしの場合が 6 段、他の 2 つの指数変換式では 9 段で最良となっている。全体的に抽象化した方が良いという傾向がみられ、最適なベクトルの次元数が存在すると言える。ここで注目すべき点は、全体的にみると機械的評価とほぼ同様な傾向 (特に EDR で変換なしの場合) を示しているということである。これは機械的評価法が感覚的評価法に近いことを示しているとも言えよう。

(2) シソーラスの類似性判別能力

図 16 に評価結果を示す。この評価法でもシソーラスとしては、EDR より NGT の方が全体的に高い評価値が得られている。また、類似度計算法としては、最大値でみると若干ベクトル化法と距離法が段数法に比べて優れているが、ほとんど差は出ていない。距離法、段数法では、機械的評価と同様に、類似度が最大の場合の条件が最良であると言える。なお、類似度の指数関数的な扱いは、ベクトル化法および EDR の距離法で若干有利となっている。

(3) 概念ベースとシソーラスの比較

概念ベースでの評価結果を、シソーラスによる各種計算法と比較しながら図 17 に示す。機械的評価と同様にシソーラスにおいては、最適の条件の場合の値を示している。同図より、感覚評価法でも、概念ベースがいずれのシソーラスより高い評価値を得ていることが分かる。また、類似度は指数的に扱った方が若干高い性能が得られ

図 16 シソーラスによる方法の類似性判別能力比較 (感覚評価)

図 17 感覚評価法による類似性判別能力比較

と言える。

5. 考 察

(1) 機械的評価法の評価

機械的評価と感覚評価の結果 (図 12, 図 17) を比べると, 前者では, 概念ベースおよびシソーラスのベクトル化法が距離法や段数法に対して明確な差をもって優位性を示している。これに対し, 後者ではその差は少なくあるいはほとんどなくなっている。従って, 機械的評価法は概念をベクトル表現する方式に有利になる傾向があると考えられ, この方法による評価結果の信頼性が懸念されることになる。そこで, 機械的評価と感覚評価の結果の相関の有無について調べた。図 18 に, 4 章の冒頭に示したすべての組み合わせでの評価結果 (計 45 通り) についてプロットした結果を示す。また, 表 2 に, 全体の相関係数 (データ数: 45) と類似度の指数変換の有無に

図 18 機械的評価法と感覚評価法の相関

表 2 機械的評価法と感覚評価法の評価値の相関係数

指数変換	相関係数
変換なし	0.699694
変換式 (4)	0.774212
変換式 (5)	0.793211
全体	0.756717

対する相関係数 (データ数: 各 15) を示す。これから, 両評価法にはっきりとした相関があることがわかり, 機械的評価法は前述のような傾向はあるものの全体的にはその評価結果は信頼できると言える。また, 表 2 から類似度は指数的に扱った方がより人間の感覚に近い結果が得られると言える。

(2) 類似性判別に関する人間の感覚のバラツキ

人間の感覚は, 万人共通ではない。そのズレは, 個人の感覚の差によるものであり, 人間が与える類似度にもバラツキが生じている。このバラツキを調べることにより, 類似度判定ツールの性能がどの程度なら人間に近いと言えるのかについての目安を推定した。具体的には, 感覚評価法において被験者 14 人の与えた類似度の平均と各被験者の与えた類似度との相関を求めた。その結果, 14 人の平均の相関係数は 0.90 であった。すなわち, 概念間の類似性判定は人間が行ってもこの程度のずれが生じることが分かった。従って, 各種の類似度判定ツールも, 感覚評価において 0.90 に近い評価値 M (相関係数) が得られれば, ほぼ人間に近い類似性判別を実現できたと考えよう。

(3) 評価結果の総括

類似度判定ツールについては, 機械的評価, 感覚評価ともに概念ベース, NGT, EDR の順で高い評価値を示していることから, ここで用いた類似度計算法を想定した場合, この順で類似性判別能力が高いと言えよう。特に, 概念ベースの感覚評価における 0.83 ~ 0.88 (指数変換の有無) という相関係数は, 前述の人間のバラツキを

示した 0.90 という値にきわめて近いと言える。また、機械的評価では指数変換しなくても $F=0.615$ という評価値が得られているが、これは例えば、類似概念との類似度が 0.8、非類似概念との類似度が 0.1、類似概念間で約 90%の精度で類似の程度を正確に判定できることを意味しており、実用性の高さを示していると言える。

シソーラスの類似度計算法については、機械的評価では NGT, EDR とともに、ベクトル化法、段数法、距離法の順で良いこと、感覚評価ではほとんど差がないこと、および両評価法に相関があることから、ほぼこの順で類似性判別能力が高いと言えよう。

また、類似度を指数関数的に扱うことは、全体的な評価結果から判断すると性能向上に有効と言える。ただし、あらかじめ評価サンプルを用いて適切な定数（ここでの評価では最適な β は 0.1 ~ 35.0 程度であった。）を設定する必要があるため、その採否はこれを考慮して決定すべきである。

なお、以上は類似性判別に関する性能を平均的にとらえた場合の結果であり、概念ベースが類似性判別に関し常に NGT, EDR より優れているというものではない。そもそも概念ベースは国語辞典の語義文から機械的に構築しているため、語義文が短かったり、類似の意味を持つ単語（概念）を全く異なる視点から説明している場合、それら概念間の類似度が人間の感覚による類似度と大きくずれることがある。例えば、“研究”と“勉強”の類似度（指数変換なし）を求めると 0.05 と小さく、逆に“視野”と“磁場”の類似度は 0.87 と大きくなる。紙面の都合で詳細は示さないが、概念ベースでの類似度と感覚評価法での類似度の関係は、実験データを平均化すると、前者の類似度が約 0.7 のとき後者の 3（よく似ている）に対応し、約 0.43 のとき 2（少し似ている）、約 0.17 のとき 1（関連性はある）にそれぞれ対応する。これより、先の例の類似度が人間の感覚と大きくずれていることが分かる。この点が国語辞典の語義文のみを情報源としている現在の概念ベースの本質的な欠点である。従って、具体的な応用においては、多くの概念間の類似度を統合する等、平均的に高い精度を必要とする場合には概念ベースが適すると言える。一方、概念間の類似度を単独で評価し、その結果が人間の感覚と大きくずれることが禁止的な場合は、人手で構築されているシソーラスの方が適していると言える。あるいは、ここで提案したベクトル化法では各概念がベクトル表現されているため、現在の概念ベースと容易に結合可能であり、これによりシソーラスの特性をも併せ持つ改良型の概念ベースを構築するという方向（今後検討の予定）も考えられる。

6. む す び

概念ベースとシソーラスの言葉の意味に関する類似性判別能力を、いくつかの類似度計算法を想定し比較評価

した。シソーラスとしては、国内では代表的な日本語語彙体系と EDR 電子化辞書のもの（それぞれ NGT, EDR）を選んだ。また、シソーラスでの類似度計算法として、その構造的特徴を利用して各概念（単語）をベクトル表現し、その内積で類似度を算出する方法（ベクトル化法）を提案した。

評価は、既に提案している機械的評価法と、これまで一般的に用いられている人間の感覚による方法により行った。また、シソーラスの類似度計算法としては、提案したベクトル化法と従来から広く用いられているカテゴリ間の距離に基づく計算法（距離法）およびそれらの共通段数に基づく計算法（段数法）を採り上げた。両評価法による評価の結果、類似度判定ツールとしては前述の類似度計算法の範囲で、概念ベース、NGT, EDR の順で高い性能が得られた。特に概念ベースでは人間による類似性判定に近い性能が得られた。また、類似度計算法としては、提案したベクトル化法は従来法と同等以上の性能を示し、従来法では段数法が距離法より若干高い性能を示した。ベクトル化法は概念ベースと同様に視点を考慮した類似性判定の可能性もあり、従来法と同等以上の性能が確認されたことから有望な方法と言える。

さらに、提案した機械的評価法と人間の感覚に基づく評価法には明確な相関があり、前者の実用性が確認された。

なお、本論文での評価結果は前章でも述べたように、類似性判別に関する性能を平均的にみた結果であり、現在の概念ベースでは概念によっては人間の感覚と大きくずれる類似度が得られることがあるという欠点もある。このため、実際の応用では用途に応じ概念ベースとシソーラスを使い分けることが現実的である。また、概念ベースの視点を考慮した類似性判別については、現在、それ自体の構成法と絡めた各種計算法を検討中であり、別途報告する予定である。

謝 辞

本研究の一部は、文部省科学研究費補助金（課題番号 15560351）および同志社大学学術フロンティア「知能情報科学とその応用」研究プロジェクトの一環として実施された。

◇ 参 考 文 献 ◇

- [EDR 93] Japan Electronic Dictionary Research Institute, Ltd. : EDR Electronic Dictionary Technical Guide, tr-042 edition (1993).
- [遠藤 94] 遠藤織枝ほか（編）：類語例解辞典，小学館（1994）.
- [原口 86] 原口誠，有川節夫：類推の定式化とその実現，人工知能学会誌．Vol.1, No.1, pp.132-139 (1986).
- [平川 03] 平川秀樹，木村和広：概念体系を用いた概念抽象化手法と語義判定におけるその有効性の評価，情報処理学会論文誌．Vol.44, No.2, pp.421-432 (2003).
- [平尾 03] 平尾努，磯崎秀樹，前田英作，松本祐治：Support Vector Machine を用いた重要文抽出法，情報処理学会論文誌．Vol.44, No.8, pp.2230-2243 (2003).

- [飯田 98] 飯田敏幸, 松澤和光, 池上徹彦, 石野福弥, 今井賢一: 想起型情報検索システムについて, 情報処理学会研究報告 98-OS-77/98-DPS-87, pp.19-24 (1998).
- [池原 97] 池原悟ほか: 日本語語彙体系, 岩波書店 (1997).
- [石川 98] 石川勉, 井澤潤次郎, 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, Vol.13, No.3, pp.470-479 (1998).
- [笠原 97] 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).
- [北川 93] 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).
- [Li 03] Li, Y., Bandar, Z. A., McLean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources, IEEE Trans. Knowledge and Data Engineering, vol.15, No.4, pp.871-881 (2003).
- [松原 92] 松原仁: 推論技術の観点からみた事例に基づく推論, 人工知能学会誌, Vol.7, No.4, pp.567-575 (1992).
- [Miler 95] G. A. Miler: WordNet: A Lexical Database for English, Comm. ACM, Vol.38, No.11, pp.39-41 (1995).
- [長尾 96] 長尾真: 自然言語処理, 岩波書店 (1996).
- [中島 01] 中島浩之, 木谷強, 岩城修: レレバンスフィードバックにおける検索語の共起関係推定処理の高速化, 情報処理学会論文誌. Vol.42, No.10, pp.2429-2438 (2001).
- [Nguyen 01] Nguyen Viet Ha, 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).
- [Nguyen 02] Nguyen Viet Ha, 帆効譲, 石川勉, 笠原要: 単語の意味の類似性判別のための大規模概念ベース, 情報処理学会論文誌. Vol.43, No.10, pp.3127-3136 (2002).
- [奥村 02] 奥村学ほか: 特集「テキスト自動要約」, 情報処理学会誌. Vol.43, No.12, pp.1285-1316 (2002).
- [奥村 04] 奥村学 (編): 特集「WWW 上の情報の知的アクセスのためのテキスト処理」, 人工知能学会誌 Vol.19, No.3, pp.295-323 (2004).
- [Rocchio 71] Rocchio, J.J.: Relevance Feedback in information retrieval, The SMART Retrieval System, pp.313-323, Prentice-Hall (1971).
- [Salton 94] G. Salton and J. Allen: Text Retrieval Using the Vector Processing Model, 3rd Annual Symposium on Document Analysis and Information Retrieval (1994).

〔担当委員: 西田 豊明〕

2005 年 4 月 25 日 受理

著者紹介

川島 貴広

2005 年 3 月拓殖大学大学院工学研究科電子情報工学専攻修士課程終了。在学中は言葉の意味の類似性判別に関する研究に従事。現在、ヴァル研究所勤務。

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。1955 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。電子情報通信学会, IEEE 各会員。