2D2-10

言葉の意味の類似性判別能力に関するシソーラスと概念ベースの性能比較

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

川島 貴広[*1] 石川 勉[*1]
Takahiro Kawashima Tsutomu Ishikawa

[*1]拓殖大学工学部情報工学科
Department of Computer Science, Takushoku University

We have developed a knowledge-base of words to measure the degree of semantic similarity between words. This paper describes an evaluation result about its capability comparing to the thesauruses of the EDR electrical dictionary and the ALT system. We also propose a new measuring method by thesaurus, in which vectorized data for representing each word are generated based on the structure of the thesaurus and then the degree of similarity are calculated using the data. Our evaluation shows that a knowledge-base of words is superior to both thesauruses. It is also shown that the proposed method is superior to conventional measuring methods using the distance between categories and so on, if thesauruses are used.

1. はじめに

我々は，単語 (概念) 間の意味的な類似性を判定するための概念ベースについて研究し，現在 25 万語規模の概念ベースを構築してきている [1].

本報告では，概念ベースとこれまで一般的に単語間の類似性を判定するために用いられていたシソーラスとの性能比較を行う．シソーラスにおける類似性判別の方法としては，シソーラス上のカテゴリ間の距離とカテゴリの段数による従来の 2 つの類似度計算法以外に，新たにシソーラス情報をベクトル化する方法を提案し，それぞれの類似性判別能力を比較評価する．

2. 評価対象

2.1 シソーラス

シソーラスとは，単語の上位／下位関係, 部分／全体関係, 同義関係, 類義関係などによって単語を分類し，体系づけた辞書である．今回使用するシソーラスは，日本語語彙体系 [2] と EDR 電子化辞書 [3](以下，前者のシソーラスを ALT，後者のシソーラスを EDR と略す) を使用する．

ALT は 2,715 のカテゴリからなる最大 12 段の，EDR は 202,797 のカテゴリからなる最大 16 段のシソーラスである．また，両シソーラスは構造的に異なり，ALT が完全な木構造であるのに対し，EDR は 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) ことがあるグラフ構造である．

2.2 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 [1]．具体的には，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属性，その出現頻度を属性値とし，基本的には tf・idf の考え方に基づいて構築されている．各概念は日本語語彙体系のカテゴリを基底とした 2,715 次元のベクトルで表現されている．

3. 類似度計算法

3.1 シソーラスでの計算法

3.1.1 距離，段数による類似度計算法

シソーラスを利用した類似度計算法としては，概念 A，B 間の類似度を求める場合，一般に以下の 2 つの方法が用いられている．

連絡先: 拓殖大学 工学部 情報工学科
〒 193-0985 東京都八王子市館町 815-1
E-mail: y3m307@st.takushoku-u.ac.jp

$$\text{[方法 1] : 類似度} = \frac{1}{\text{距離} + 1}$$

$$\text{[方法 2] : 類似度} = \frac{A, B \text{ の共通段数} \times 2}{A \text{ の段数} + B \text{ の段数}}$$

ここで，距離とはカテゴリ間の枝の数であり，段数とは根カテゴリを 1 段とし，それよりカテゴリが 1 つ下位になるごとに 1 つずつ加算したものである．これら詳細については，文献 [4] を参照されたい．

3.1.2 ベクトル化による類似度計算法

ベクトル化とは，シソーラスのカテゴリに属する各概念を，抽象化したカテゴリ数次元のベクトルで表現するものである．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 1，図 2 を使って説明する．図 1 は多重継承と 3 段に抽象化した際の例を，図 2 は基本的な重みの付与の考え方を示している．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，C_i はカテゴリの識別子である．また，R_1，R_2 はカテゴリ a から T へのルートであり，R_1 は C_1 を通るルート，R_2 は C_4 を通るルートを示している．

1）抽象化

抽象化とは，カテゴリをより上位のカテゴリにマップする操作である．これにより，与えられたカテゴリを適切な粒度のカテゴリに抽象化しベクトルの次元を減少する．また，抽象化の方法としては，均等深度法 [5] を採用する．均等深度法とは根のカテゴリから段数が一定 N 以上であるカテゴリを，その上位カテゴリで段数 N に位置するものに抽象化する方法である．また，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．例として図 1 では，シソーラスを 3 段に抽象化している．この場合には，ルート R_1 では C_2，ルート R_2 では C_4 に，それぞれその下位に属するカテゴリが，マップされることとなる．

2) 初期重みの付与

対象概念のカテゴリ (図 2 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 1 のよう 3 段に抽象化した場合，a に対し C_2，C_4 のそれぞれに初期重み “1” を与える．

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に対しても，関連するカテゴリとして重みを付与する．この場合，シソーラスは上位カテゴリになるほど情報量は少ないため，上位カテゴリに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 (図 2 中の式) により重みを減らして付与することとする．ただし，すでに上位カテゴリ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，最大の重みをそのカテゴリの重みとする．また，多重継承の場合にも全てのカテゴリに対し重みを付与する．例として，図 1 のように多重継承しているシソーラスを 3 段で抽象化した場合，右図のように Tree を展開し，R_1,R_2 上のそれぞれのカテゴリに対し，上述したように情報量の比により重みを付与する．

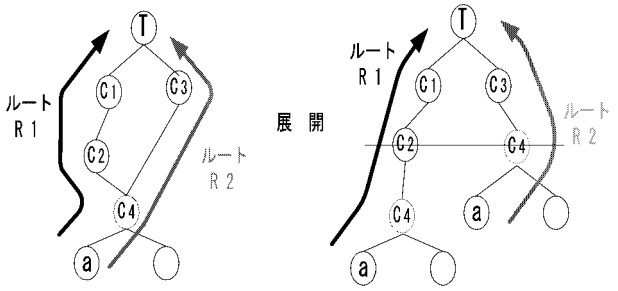

図 1: 多重継承と抽象化の考え方

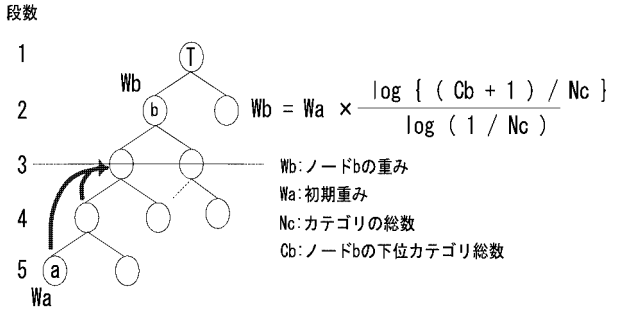

図 2: シソーラスのベクトル化法による重みの付与

以上，ベクトル化の方法を示したが，この処理を類似度計算時に行うのではなく，あらかじめこのベクトル化を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．

3.2 概念ベースでの計算法

概念ベースでは，概念間の類似度は比較する概念同士のベクトルの内積で算出され，0〜1 の実数で表される．これら詳細については，文献 [1] を参照されたい．

4. 評価法

文献 [6] で提案した評価法を用いる．以下，これについて説明する．

4.1 考え方

類似度計算ツールの特性としては，①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こと，② 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識別可能であること，が重要である．従って，これら特性を考慮した評価指数を設定する．

4.2 評価指数

対象概念 G，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 1 組とし，それを N 組つくり評価データとする．ここで，$G-G_1$ 間，$G-G_2$ 間，$G-G_3$ 間の類似度をそれぞれ r_1，r_2，r_3 とし (図 3 参照)，前述の特性に対し，それぞれ以下のような評価指数を設定する．

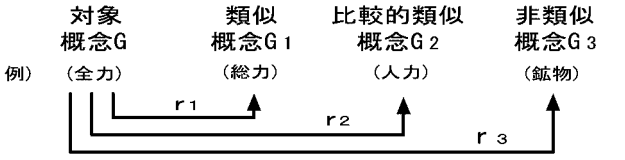

図 3: 対象概念とその評価に用いる概念の関係

①に対する評価指数:

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

ここで，$\bar{r}_1$，$\bar{r}_3$ はそれぞれ，r_1，r_3 の平均値，σ_1，σ_3 はそれぞれ，それらの標準偏差である．ただし，σ_1 は $\bar{r}_1$ より小さい方のデータ，σ_3 は $\bar{r}_3$ より大きい方のデータを用いて算出する．

②に対する評価指数:

r_1 と r_2 の関係としては $r_1 > r_2$ 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．

$$F_2 = \frac{m}{N}$$

ここで，m は前述の関係が成立した組の数である．ただし，m の成り立つ条件として以下のように有意差 α を考慮する．

$r_1 > r_2 + \alpha$	:	1	(正解)
$r_2 + \alpha > r_1 > r_2 - \alpha$	:	0.5	
$r_2 - \alpha \geq r_1$	:	0	(誤り)

この，α については，区間推定を用いて設定する．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．

$$\alpha_i = t \times \frac{s}{\sqrt{N}}$$

ここで，t は信頼係数を与えて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．

最終的な評価指数 F は，これらの積とし，以下のように設定する．この評価指数は，理想的な類似度計算ツールで 1 となる．

$$F = F_1 \times F_2$$

5. 評価結果

5.1 評価データ

類語例解辞典 [7] より 200 組を抽出し評価した．ここで，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いる．評価データの一例を図 4 に示す．

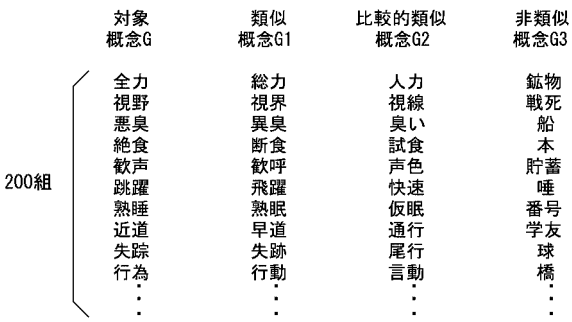

図 4: 評価データサンプル

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．

5.2 評価結果

5.2.1 シソーラスの類似性判別能力

図 5，図 6 にそれぞれ，ALT，EDR のベクトル化による評価結果を示す．方法 1，2 の評価結果の詳細については，文献 [4] を参照されたい．

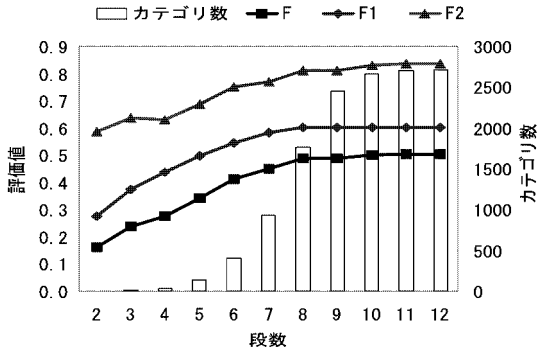

図 5: ベクトル化法 (ALT) の評価結果

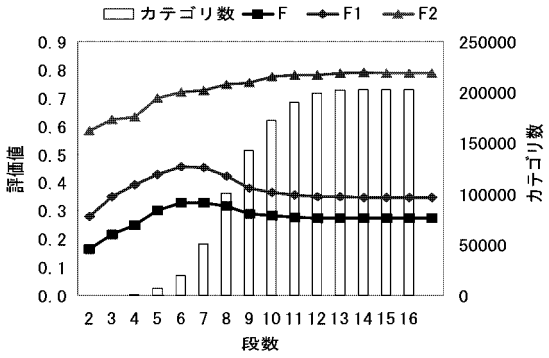

図 6: ベクトル化法 (EDR) の評価結果

以上の結果より，ALT では 11 段で，EDR では 7 段で最良となることがわかる．EDR の結果については，文献 [5] で紹介されている均等深度法とほぼ同様な傾向がみられる．なお，4 章における信頼係数は 90%とした．

5.2.2 概念ベースとシソーラスの性能比較

概念ベースとシソーラスによる 3 種の計算法の類似性判別能力を表 1 に示す．

表 1. 概念ベースとシソーラスの類似性判別能力

類似度算出法		F	F1	F2
概念ベース		0.624	0.679	0.920
ALT	ベクトル (11 段)	0.504	0.602	0.837
	距離 (方法 1)	0.168	0.220	0.762
	段数 (方法 2)	0.364	0.478	0.762
EDR	ベクトル (7 段)	0.328	0.451	0.727
	距離 (方法 1)	0.100	0.135	0.742
	段数 (方法 2)	0.263	0.361	0.727

表 1 の結果より，①総合的な評価指数 F だけでなく，F_1，F_2 のいずれも概念ベースの方が両シソーラスより優れている，②シソーラスでは，いずれの計算法でも ALT の方が EDR より優れている，③シソーラスを用いた計算法では，ベクトル化法，段数による方法 (方法 2)，距離による方法 (方法 1) の順に優れている，ことが分かる．

6. 考察

文献 [8] では WORDNET を用いて各種の類似度計算法について人間の感覚に基づいて評価している．具体的には，28 組の評価概念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，その値と提案手法により得られた類似度との相関係数により評価している．表 2 に用いられている評価概念とそれらに対する人間の感覚による類似度の一例を示す．

表 2. 評価概念とその類似度の一例

評価概念	類似度 (0〜4)
moon-string	0.04
glass-magician	0.44
food-rooster	1.09
food-fruit	2.69
coast-shore	3.60
automobile-car	3.92

同文献では，各種計算法に対して評価が行われているが，その最大値として，段数をベースとする方法において類似度を指数関数的に表わすことで 0.8914 という値が得られている．

ここでは，概念ベースに対しても同じ評価概念に基づき同様な手法で評価した．その結果，類似度を以下の関数（x：もとの類似度，β：定数（=7.9））で指数関数的に表したとき，相関係数は 0.93 と同文献の値以上となった．

$$\text{変換後の類似度} = \frac{1 - e^{\beta x}}{1 - e^{\beta}}$$

すなわち，概念ベースは人間の感覚に基づく評価でも高い類似性判別能力が得られるといえる．なお，評価は，各概念を日本語に訳して使用した．

7. まとめ

概念ベースとシソーラスの類似性判別能力を比較評価した．また，シソーラスを用いる方法として，その構造を利用して各概念をベクトル表現する方法を提案した．評価の結果，類似性判別能力は，概念ベース，ALT シソーラス，EDR シソーラスの順で，シソーラスによる方法では，ベクトル化法，段数による方法，距離による方法の順で高いことが分かった．

参考文献

[1] Nguyen Viet Ha, 穂刈譲，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)

[2] 池原悟，他:“日本語語彙体系 1 意味体系”，岩波書店 (1997)

[3] “EDR 電子化辞書”，http://www.iijnet.or.jp/edr/J_index.html

[4] 川島貴広，石川勉:“言葉の意味に関する類似性判別能力における概念ベースとシソーラスとの性能比較”，情報処理学会第 65 回全国大会,2M-1,pp.2-135 - 2-136 (2004)

[5] 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)

[6] 川島貴広，室伏秀幸，石川勉:“単語のカテゴリ情報とシソーラス情報を利用した概念ベースの構築”，FIT2002 情報科学技術フォーラム,E-38,pp.157-158 (2002)

[7] 遠藤織枝，他（編）:“類語例解辞典”，小学館 (1994)

[8] 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, pp871-881 (2003)