

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

川島貴広[†] 石川勉[‡]
拓殖大学 工学部 情報工学科[‡]

1 はじめに

単語 (概念) 間の意味的な類似性を判定するには、これまで一般的にシソーラスが利用されていた。しかし、シソーラスは、is-a や kind-of の関係で概念群を木構造化したものであり、観点を考慮した類似性判定は困難である。これに対し、我々はあらゆる観点から見た類似性判定が可能な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]。これまで、これらの類似性判別能力については、観点を考慮しない場合についても同一尺度での比較評価はなされていない。

本報告では、概念ベースとシソーラスの類似性判別能力を定量的に比較する。具体的には、シソーラスとして日本語語彙体系 [2] と EDR 電子化辞書 [3] のものを選び (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)、観点を考慮しない基本的な類似性判別能力に関して比較評価する。

2 概念ベースの構成と類似度計算法

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。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ている。さらに、その属性を日本語語彙体系の 2715 のカテゴリで代表させ、各概念を 2715 次元のベクトルで表現している。この概念ベースでは、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。

3 シソーラスを利用した類似度計算法

シソーラスを利用した類似度計算法は、一般に図 1 の 2 つの方法 (方法 1、2) が用いられている。

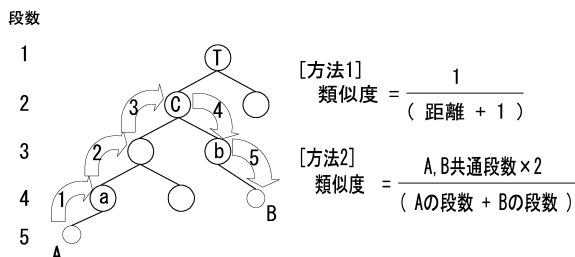


図 1: シソーラスによる類似度計算法

A、B は、類似度を求める対象概念であり、a、b はそれらの概念のカテゴリである。方法 1 の距離とは、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。方法 2 の段数とは、

根のカテゴリ (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ものである。また、共通段数とは、対象概念の共通の上位カテゴリの段数である。すなわち同図の例では、方法 1、方法 2 の類似度は、それぞれ、1/6、4/9 となる。

なお、以上の方法では、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、平均の 3 通りの場合について比較評価する。

4 評価方法

文献 [4] で提案した評価法を用いる。以下、これについて簡単に説明する。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、及び、②2 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ということ、が重要である。従って、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 2 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。

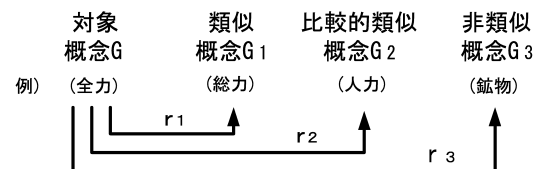


図 2: 対象概念とその評価に用いる概念の関係

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。

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

[†]Takahiro Kawashima, Tutomu Ishikawa

[‡]Department of Computer Science, Takkushoku University

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

なお、 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 [5] 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 3 に示す。

対象 概念G	類似 概念G1	比較的類似 概念G2	非類似 概念G3
全力 視野 悪臭 絶食 歓声 跳躍 熟睡 近道 失踪 行為 ...	総力 視界 臭い 断食 欢呼 飛躍 熟眠 早退 失踪 行動 ...	人力 視線 臭い 試食 声色 快速 仮眠 通行 尾行 言動 ...	鉱物 戦死 船舶 貯蓄 唾 番号 学友 球橋 ...

図 3: 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” をつけて評価する。

5.2 シソーラスの類似性判別能力

図 4 に評価結果を示す。

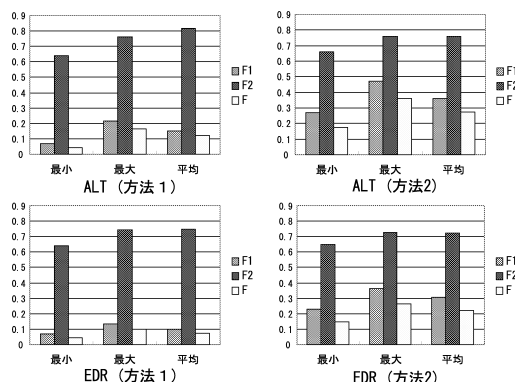


図 4: シソーラス評価結果

同図より、①類似度の計算法としては両シソーラスとも方法 2 の方が優れている、②類似度の求め方としてはそれが最大となる計算法が優れている、③いずれの計算法でも ALT の方が EDR より優れていることが分かる。

5.3 概念ベースとの性能比較

同様に測定した概念ベースの類似性判別能力と前節の最良の場合のシソーラスのそれらを図 5 に比較して示す。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圧倒的に優れていることが分かる。

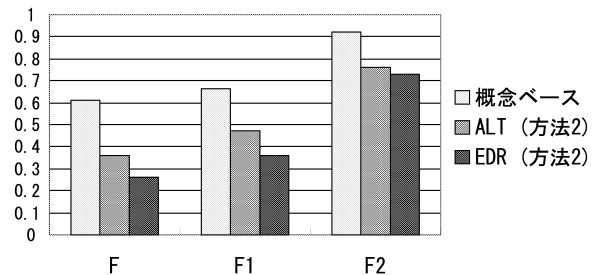


図 5: 概念ベースとシソーラスの性能比較

5.4 信頼係数をパラメーターとした評価

5.2、5.3 節の結果は、信頼係数 90% で評価したものだが、これを变化させた場合の影響について調べた。図 6 にその結果を示す。同図より、信頼係数を変化させても、評価値の変化はわずかであることが分かる。

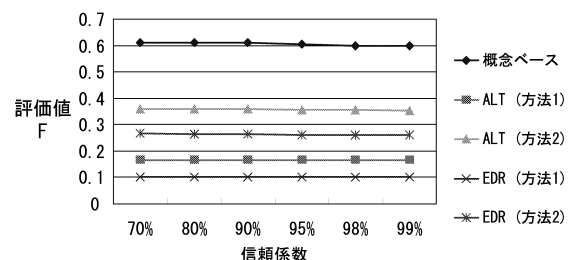


図 6: 信頼係数の評価値への影響

6 まとめ

これまで開発してきた概念ベースとシソーラスの類似性判別能力について比較評価した。概念ベースは、観点を考慮した類似性判別が可能という特徴を有するが、それを考慮しない単純な類似性判別能力についても従来のシソーラスに対し、圧倒的に優れている (これまで提案してきている評価値において 2 倍程度) ことが分かった。

参考文献

- [1] Nguyen Viet Ha、穂刈謙、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
- [2] 池原悟、他: “日本語語彙体系 1 意味体系”, 岩波書店
- [3] “EDR 電子化辞書”, http://www.ijnet.or.jp/edr/J_index.html
- [4] 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)
- [5] 遠藤織枝、他 (編): “類語例解辞典”, 小学館 (1994)