F-021

観点を考慮した概念間の類似性判別法

Calculating Methods of the Degree of Similarity between Words Considering Viewpoint

野口　洋平† 森　麻美‡ 石川　勉‡
Yohei Noguchi Mami Mori Tsutomu Ishikawa

1　まえがき

我々は, 単語 (概念) 間の意味的な類似性判別を主な目的とした概念ベース (以下 GB) の構築を進めてきた [1]. GB の単純な概念間での類似性判別能力に関しては, これまでの評価でシソーラスを上回ることを確認している [2]. しかし, 概念間の類似性は, どのような観点で対象概念を比較するかにより異なってくる. 本報告では, このような観点を考慮した類似性判別法について述べる. 具体的には, 現 GB の構成を前提とした各種類似度計算法および, 新たな考えに基づき再構築した GB での計算法について比較評価する.

2　現 GB を用いた計算方法

現 GB 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 [3] 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 GB を用いた観点を考慮しない概念間の類似度計算は,2 概念のベクトルの内積をとることで行ってきた. 一方, 観点を考慮する場合は, その観点からみて重要となる属性値を強調した後, 内積をとることにより求める方法が基本となる. ここでは, 強調の方法として, 観点となる概念の属性値を利用して間接的に強調する方法 (間接法), 同じくその概念の属するカテゴリから直接的に強調する方法 (直接法), 両方法を組合せた方法 (組合せ法) の 3 つの方法を考えた. 図 1 にこれら強調法の考え方を示す.

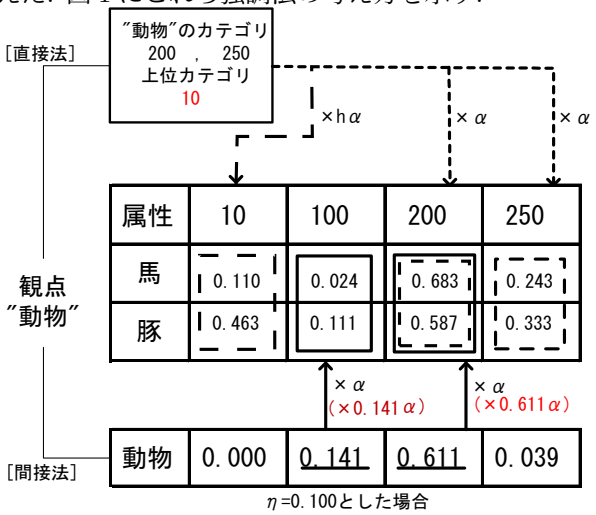

図 1: 観点による属性の強調方法

i) 間接法

観点の属性値があるしきい値 η を超える場合, その属性を重要と考え, 対象概念に対して, その属性に対応した属性値を強調 (α 倍) する方法 (一定強調) である. さらに, 観点の属性値 (w) を考慮することで ($w\alpha$ 倍) する方法 (可変強調) についても考えた. また, 観点の属性値がしきい値を超えなかった場合は, 対応する属性を強調しないで計算に用いる方法と属性値を 0 として類似度計算から除く方法の両方で評価した.

ii) 直接法

対象概念に対して, 観点のカテゴリに対応する属性値を強調 (α 倍) する方法 (上位無) である. さらに, 観点のカテゴリだけでなく, その上位カテゴリに情報量の比によって重み (h) をつけ, それらに対応する概念の属性値を強調 ($h\alpha$ 倍) する方法 (上位有) についても考えた.

iii) 組合せ法

間接法と直接法を組合せて強調する方法である.

3　再構築概念ベースでの計算方法

概念が以下のように, 観点となり得る属性と単なる重みではない属性値の組で表されていれば, ある観点から見た概念間の類似度は, その観点に対応する属性の属性値を比較することで得ることができる.

“ 概念 ”(属性, 属性値)(属性, 属性値)…(属性, 属性値)
“ 馬 ” (色, 茶色) (大きさ, 大型)…(足, 四本)

例えば, 概念“ 馬 ”と他の概念とを“ 色 ”を観点として比較する場合,“ 色 ”の属性値“ 茶色 ”を用い, 他の概念の同カテゴリの属性値との比較を行えばよい. この構成では, 観点となり得る属性を全て定めることが必要となるが, これは現実的には不可能に近い. 従って, ここでは近似的に現 GB の 2715 のカテゴリを観点となり得る基準の属性とみなし, 属性値はそのカテゴリに代表させる前の独立語とする. 具体的には, GB の原データとして“ ある概念 (見出し語) に他の概念 (独立語) がどの程度関連しているか ”というデータがある. この独立語の概念ベクトルを属性値として利用する. また, その重みとして, その出現頻度を与える. 図 2 に新概念ベース (以下新 GB) の構築方法を示す.

このような新 GB での類似度計算法として以下の方法を考えた.

新 GB 法

比較対象の概念に対して, 観点が属するカテゴリ内に存在する全ての属性値を重みを考慮して統合し, これをそのカテゴリの代表ベクトルとする. 類似度は対象概念の代表ベクトル同士の内積により求める.

ただし, 観点が複数のカテゴリに属していた場合は, 対象概念の各カテゴリ間で合成を行いその合成ベクト

†拓殖大学大学院工学研究科電子情報工学専攻
‡拓殖大学工学部情報工学科

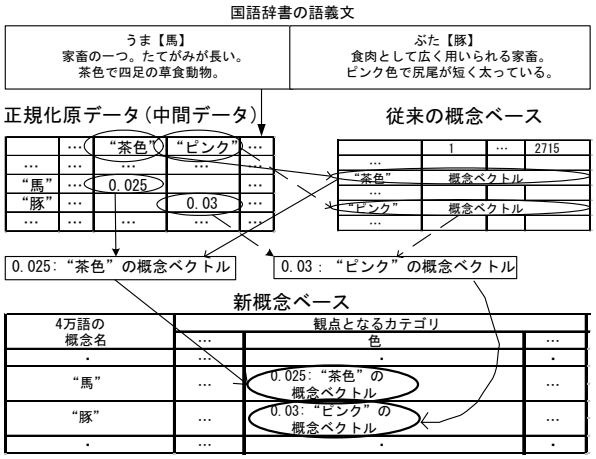

図 2: 新 GB の構築方法

ルを代表ベクトルとする方法 (合成法) とはじめに各カテゴリ間の代表ベクトルで内積を求め, その最大 (最大法), もしくは平均を類似度とする方法 (平均法) について評価する.

4 評価

評価には, 以下の 2 式が成立する概念 3 つと観点 2 つからなる評価データの組を用意する.

R(概念 1, 概念 2| 観点 A) > R(概念 1, 概念 3| 観点 A)
R(概念 1, 概念 3| 観点 B) > R(概念 1, 概念 2| 観点 B)

評価では, このような評価データ 100 組を作成し, これを用い, 以下の方法で評価した. 図 3 に評価データの組の例を示す.

データ組の構成 [概念1 概念2 観点A / 概念3 観点B]

[博物館 美術館 観光 / 図書館 調査] [音楽 映画 鑑賞 / 数学 授業]

[手紙 小包 郵便 / 電話 連絡] [蜂 注射 刺す / 蝶 昆虫]

[鼻 肺 呼吸 / 目 顔] [新聞 雑誌 記事 / 郵便 配達]

図 3: 評価データ組の例

4.1 評価方法

厳しい評価としては, 前述の 2 つの式がともに成り立ったときを正解 (◎) とする. また, ゆるめの評価として, 式単体で評価し, 右辺と左辺の大小関係が成り立った場合を正解 (○), 反転してしまった場合を不正解 (×), 一致した場合は判別不能 (△) とする. この両方法で,2,3 章のすべての計算法について評価した. なお,α については最適値を選んだ.

4.2 評価結果

各計算法の評価結果を表 1 に示す. ◎は 100 組中の正解数, ○, ×, △は 200 式中の値である. また, 条件は各計算法で最適となった条件あるいは方法を示している. なお, 観点の属性値がしきい値を超えなかった場合は, いずれの方法でも対応する属性を 0 とする方法が優れていた.

表 1: 各強調法における評価結果

強調法	◎	○	△	×	条件
間接法	33	122	42	36	可変強調
直接法	18	87	111	2	上位有
組合せ法	32	122	43	35	可変強調, 上位有
新 GB 法	18	83	91	26	最大法

同表からわかるように, 厳しい評価では, 間接法と組合せ法が優れていた. ゆるめの評価でも, 間接法と組合せ法が正解数からは優れていると言える. なお, △は類似度が 0 となる場合がほとんどであった.

4.3 考察

新 GB 法は観点を考慮した類似性判別性能の向上を目指した方法であるが, 表 1 で示したように, ◎と○の数で評価をすると, 現 GB の性能以下となった. この原因を新 GB の属性数の不足と考え, 概念が持つ属性数と○, △, ×の各判定に含まれる概念数の関係を測定した. 測定の結果を図 4 に示す.

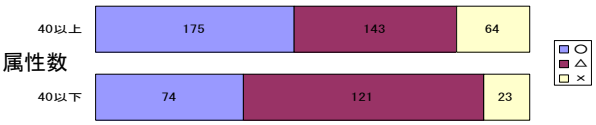

図 4: 属性数と各判定における概念数の関係

同図から, 属性数が多い場合は, 判別不能 (△) が減少することがわかる. また, △を除いた正解率 ($\frac{○}{○+×}$) は, 属性数にあまり依存していない. 従って, 新 GB において属性数の不足を解消できれば, 表 1 における正解率 (76.1 %) が得られることが期待できる. この値は GB の値 (△の半数を○と×にそれぞれ加える) を上回るものである.

5 まとめ

観点を考慮した概念間の類似性判別法について, 現 GB を前提とした方法と再構築した GB を用いる方法について検討を行った. 評価の結果, 現 GB を用いて間接的に強調を行う方法が最も正解率が高かった. 新 GB を用いた方法については, 判別不能となる場合が多く, それが原因で高い性能が得られなかった. これについては, 属性数の不足を解消することにより, 現 GB を用いた計算法を上回る可能性があると考えられる. これについての具体的な方法については, 今後の課題である.

参考文献

[1] Nguyen Viet Ha, 帆苅譲, 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] 川島貴広, 石川勉: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, 情報処理学会 第 65 回全国大会 (2003)

[3] 池原悟ほか, 日本語語彙体系, 岩波書店 (1997)