概念ベースにおける概念の表現方式の提案

佐藤 浩史　　笠原 要　　松澤 和光
ＮＴＴ(株)コミュニケーション科学研究所
e-mail: {hiroshi, kaname, matuzawa}@cslab.kecl.ntt.co.jp

1 はじめに

今日、一般にコンピュータは正確な解を得ることを目的に使用されている。その為に、ユーザに正確かつ適切な入力を求め、融通が利かない頭の固いものとなりがちである。そこで我々は、完全解が得られない場合でも、「常識」を用いて適切な概略解を得る柔軟な推論システム、「アバウト推論」の研究を進めている[1]。

その常識を構成する要素の一つとして、これまでに約4万語の単語の意味知識を保有する知識ベース、「概念ベース」を辞書から自動構築し、単語の類似性判別方式を提案した[2][3]。この方式は、単純に単語間の類似性を判別するのではなく、その場に応じた「観点」を考慮した上で計算を行うのが最大の特徴である。

我々は、単語の「概念」を属性(特徴)の重みからなるベクトルで表現し、空間上での距離を使って類似性を判別している。しかし、属性として単純に単語を選択した場合、それらの中には類義語も存在し、独立であるべき空間基底としてはふさわしくない。同様な手法で単語の意味表現を行っている研究はいくつかあるが（例えば[4][5]）、多くはこの点を考慮していない。そこで我々は、各属性をシソーラスのカテゴリで一般化し、それぞれが独立とみなすことでベクトル空間を張って概念を表現している。その空間上で、ベクトルの内積を使って類似度計算を行う。

ところが、シソーラスのカテゴリ間には一般に従属関係が存在し、完全に独立ではないため、本来のベクトル空間の基底としてはなお問題がある。

そこで本稿では、カテゴリ間の従属関係を考慮した概念の表現方式を提案する。カテゴリ同士の持つ従属関係を、概念の属性の重みに反映させることによって独立性を高める。この概念の表現法を類似性判別に適用して評価する。

2 背景

2.1 従来技術の概要

ここでは、我々が構築した「概念ベース」および「概念の類似性判別方式」について解説する。

「概念ベース」とは、複数の辞書より取得した日常語約4万語の概念のデータベースであり、おのおのの概念を、特徴を表す属性と、その属性が概念においてどれだけ重要であるかを表す重要度の対の集合により表現したものである。ここで属性は、その概念の語義文中に現れる自立語、特に日常語をとり、また重要度はその語義文中の出現頻度より算出する。

さらに、このままでは属性同士が独立でない為、シソーラスを使ってこれらの属性を一般化する。具体的には、30万語の単語を約3,000種のカテゴリに分類したALT シソーラス[6]によって、属性をそれぞれが属するカテゴリにグループ化する。これにより、各属性を独立とみなすことができ、概念 g を多次元ベクトル空間のベクトルとして表現することが可能となる（図1,2）。その際、ベクトルの大きさが1になるように正規化を行っている。

$$\mathbf{g} = (q_1, q_2, \cdots, q_n),\ \sqrt{{q_1}^2 + {q_2}^2 + \cdots + {q_n}^2} = 1$$

概念「馬」

属性 p_i	動物	乗り物	娯楽	…
重み q_i	0.7	0.4	0.2	…

図1:概念の例

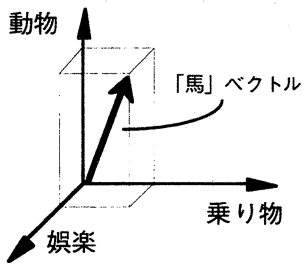


図2:概念の幾何イメージ

そして、概念間の類似度をそれらベクトルの内積をもって定義するが、我々の方式で特徴的なのは、状況に応じた類似判別を行なう点である。例えば「馬」は「豚」と「自動車」どちらにより似ているかと考えた場合、「動物」と言う点では「豚」に近く、「乗る」ではむしろ「自動車」に近いと考えられる(図3)。このように、対象を比べる際の指標として「観点」を定義し、観点 k における概念 g_1, g_2 の類似度を $Sim(k, g_1, g_2)$ ($0 \leqq Sim \leqq 1$) で表す。数値上は、観点のもつ属性を類似度計算の際に強調する。言い換えれば、ベクトル空間の該当する軸を調整することで実現している。

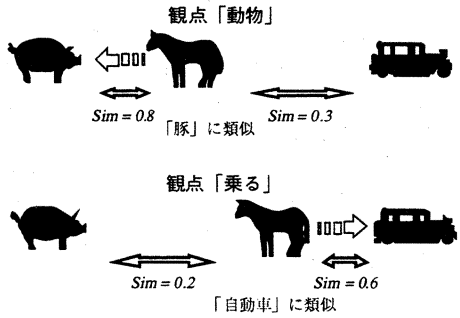


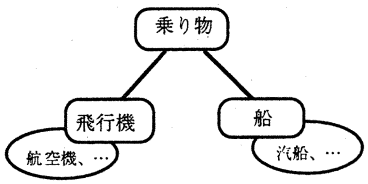
図3:観点作用のイメージ

2.2 問題点

シソーラスのカテゴリは、意味の近い単語をまとめた集合であり、単語同士に比べてその独立性は高い。それを根拠として、現在はこのカテゴリをベクトル空間の基底として採用しているが、一般にシソーラスはカテゴリ同士が完全に独立ではない。その多くは、カテゴリ間の関連性を上下関係で表現し、木構造をとっている。従って、各空間軸が完全に等価ではなく、ベクトルとしての概念を正確に表現するには至っていない。その為、ベクトルの内積に偏りが生じ、類似性判別の際に不自然な結果を出している可能性がある。

ここで例として、「航空機」と「汽船」の類似度を考えてみる。概念ベースにおける両者の属性の一部とその重みは、次の図4の通りである。

概念＼属性	乗り物	飛行機	船
航空機	0.073	0.960	0
汽船	0	0	0.742



(シソーラス構造図)

図4:「航空機」「汽船」の概念表現

それぞれの属性「飛行機」「船」の重みは高い値を持っているが、シソーラス上で上位に位置する属性「乗り物」の重みが低い。特に「汽船」に関しては0である。これは、「汽船」の語義文に「船」は出てくるが、「乗り物」という単語が出てこなかったことに起因している。従って、観点「乗り物」で両者の類似度を計算しても、属性「乗り物」での強調が生かされず、期待より類似度が小さくなる。

この点を改良することで、類似性判別の精度の向上、すなわち、より人間の感覚にあった結果が期待される。

3 提案方式

我々が属性の一般化に用いているシソーラスは、木構造をとっている為、カテゴリ間の従属関係が「上位カテゴリ」「下位カテゴリ」として表現されている。この関係に着目し、シソーラスに存在する上下関係を、具体的数値として概念の表現に反映させ、カテゴリ間の独立性を高める。つまり、ある属性に重みが与えられているときは、その上位カテゴリにあたる属性にも、重みがいくらかあるのが当然だと考え、操作を行う。この操作により、計算上はカテゴリ同士が独立だとみなすことができ、より自然な類似度が期待される。

そこで、§2.2 で挙げた例の属性「飛行機」「船」の重みを、一定の割合で、その上位カテゴリである属性「乗り物」に伝搬することを考える（図5）。

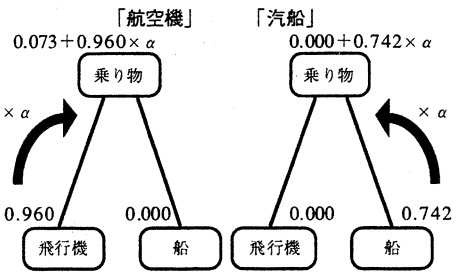


図5:重みの伝搬のイメージ

これにより、下位の属性から上位へ自然に重みが付与され、概念「航空機」「汽船」両者の共通属性「乗り物」の重みが増し、必然的に類似度が上がると予想される。この例では重みの伝搬は1つ上への属性だけだったが、さらに上位への伝搬も同様である。その際には、伝搬の次数によって適切な係数を掛けることが必要となる。これらの考えを元に、次式での新しい概念の表現を提案する。

$$\mathbf{g}' = \alpha_0 \cdot \mathbf{g} + \alpha_1 \cdot up(\mathbf{g}) + \alpha_2 \cdot up^2(\mathbf{g}) + \cdots \quad (1)$$

$$\alpha_i \in \mathbf{R}_{\geq 0}$$
$$up : \mathbf{g} \mapsto \mathbf{g}$$

$up(g)$ は、概念 g が持つ属性を全てそれらの上位カテゴリにシフトしたものであり、n ($\in \mathbf{N}$) 階上までシフトした $up^n(g)$ の線形和をもって、新しい概念 g' の表現とする。類似度計算を行う際には、g' を正規化、つまり $g'/\|g'\|$ を新たに g と置き換える。なお、複数のカテゴリから同一の上位カテゴリへシフトしてきた場合は、単純に重みの和をとる。また、上位カテゴリを持たない属性、すなわち木構造の最上位カテゴリに対しては操作を行わない。

4 実験

4.1 予備実験

方式の有効性を評価する為、予備実験を行った。ここでは、提案方式 (1) に関するパラメータをもっとも単純な、$\alpha_0 = \alpha_1 = 1$, $\alpha_n = 0$ ($n \geqq 2$) とする。

$$\mathbf{g}' = \mathbf{g} + up(\mathbf{g})$$

まず、§3 で挙げた例に対しては次のような類似度が得られた。

概念の表現	無観点	観点「乗り物」
従来方式 (g)	0.047	0.031
提案方式 (g')	0.396	0.586

図6: *Sim* (乗り物, 航空機, 汽船)

類似度が、無観点・有観点ともに従来方式では不自然に低かったが、提案方式では増大している。また従来方式では、観点「乗り物」を与えたにも関わらず、類似度が無観点時よりも下がっている。しかし、提案方式ではこの点も改善されており、人間の感覚と適合する。

そこで、類似度が上がることが期待される観点を与えた際に、逆に下がる例を10組選び、同様に評価を行った。その結果、10組中7組が上の例と同様に改善された。このことから、提案方式は有効であると予想される。

4.2 本実験

次に、提案方式のパラメータを検討する実験を行った。本稿では第一段階として、式 (1) について、

$$\mathbf{g}' = (1-\alpha) \cdot \mathbf{g} + \alpha \cdot up(\mathbf{g})$$

の形で評価を行うこととする。

実験は、類似性判別において概念の多義性を判定する方式を用いた。一般的なシソーラスにおいて、概念 g_0 が2つのカテゴリ c_1, c_2 に属する時、それぞれのカテゴリ名を観点 k_1, k_2 とする。そして、各カテゴリに属する単語の内からランダムに選んだ概念を、それぞれ g_1, g_2 とする。その場合、以下の様な判別が期待される。

$$Sim(\mathbf{k}_1, \mathbf{g}_0, \mathbf{g}_1) > Sim(\mathbf{k}_1, \mathbf{g}_0, \mathbf{g}_2)$$
$$Sim(\mathbf{k}_2, \mathbf{g}_0, \mathbf{g}_1) < Sim(\mathbf{k}_2, \mathbf{g}_0, \mathbf{g}_2)$$

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、類語辞典[7]を用いて2931組作成した。各 α (α=0.0,0.1,...,1.0) について、評価した結果を図7に示す。

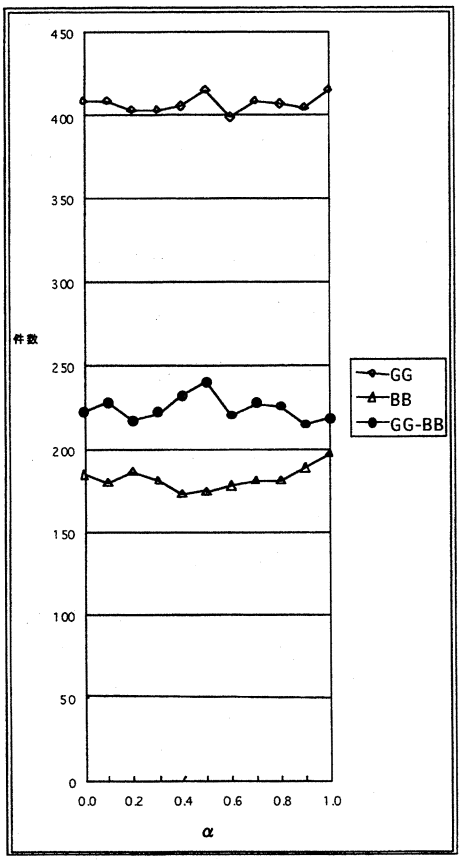


図7:実験結果

GG は両不等式ともに成立した評価データの件数、BB は両者ともに不成立だったものを表している。すると、GG がもっとも多くなるのは α=0.5 及び α=1.0 のときであり、BB は α=0.4 付近で極小となる。従って、多義性の判別を GG−BB で評価すると、α=0.5 が最適な値という結果となった。また、α=0.0 が表すものは従来の方式での評価結果であり、それと α=0.5 の値を比べることで、本方式は、従来の概念の表現を改善した方式であることが言える。

5 まとめ

本稿では、概念の表現においてカテゴリ同士の上下関係を考慮し、属性の独立性を高めた新しい概念の表現方式を提案した。この提案方式により、各概念に辞書からの自動生成だけでは得られなかった上位属性が加わり、より人間の感覚にあった類似性判別が可能になったことが、今回の実験で確かめられた。

なお、提案した概念の表現方式は 我々の概念ベースに限定された話ではなく、木構造のシソーラスを用いた言葉のベクトル表現一般に適用することができる。それら共通の問題である、空間基底の独立性の確保に1つのアプローチを示し、実証した。

今後は、高次への重みの伝搬の検討を行い、この効果をより高いものにするべく改良を加えていく予定である。また、属性を上位にシフトする際に、一律の割合ではなく、属性のカテゴリのシソーラス上での深度、概念表現内での重要度等を考慮することも考えている。

6 参考文献

[1] 松澤, 石川, 湯川, 河岡 (1993) : “アバウト推論方式の基本構想について”, 信学技報, AI93-77, pp.41-48.

[2] 笠原, 松澤, 石川, 河岡 (1994) : “観点に基づく概念間の類似性判別”, 情報処理学会論文誌, 35-3, pp.505-509.

[3] Kasahara, K., Matsuzawa, K. and Ishikawa, T. (1996) : ”Refinement method for a large-scale knowledge base of words.”, Common Sense ‘96, pp.73-82.

[4] Salton, G., Wong, A. and Yang, C.S. (1975) : “A vector space model for automatic indexing.”, Communications of the ACM, 18-11,pp.613-620.

[5] Waltz, D.L. and Pollack, J.B. (1985) : “Massively parallel parsing : A strongly interactive model of natural language interpretation.”, Cog. Sci., 9, pp.51-74.

[6] 池原, 宮崎, 横尾 (1991) : “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, 第84-13巻, pp.95-102.

[7] 大野, 浜西 (1990) : “類語国語辞典”, 第4版, 角川書店.