概念の意味属性と共起情報を用いた関連度計算方式

渡部 広一[†] 奥村 紀之[†] 河岡 司[†]

我々人間は曖昧な情報を受け取り適宜に解釈することで、会話を進めたり適切な行動を取ることができる。これは、長年の経験により蓄積された知識から築き上げられた言葉に関する「常識」を持っているからである。人間と自然に会話できる知的なコンピュータの実現には、単語の意味を理解するシステムの構築が必要であると考えられる。この実現には、ある概念から他の類似の概念ばかりでなく常識的に関連の強い概念を連想する連想メカニズムが不可欠である。そこで本稿では、単語の意味を定義している概念ベースを利用し、概念間の関連の強さをより一般的に評価する関連度計算方式について述べる。これまでの概念ベースの属性集合の一致度合いから概念間の関連性(類似度)を評価する手法を拡張し、概念空間における概念の共起情報を用いる関連度計算で補正する方式を提案する。

キーワード: 関連度, 概念ベース, 常識, 概念連鎖, 共起情報

The Method of Measuring the Degree of Association between Concepts using Attributes of the Concepts and Coincidence Information

HIROKAZU WATABE[†], NORIYUKI OKUMURA[†] and TSUKASA KAWAOKA[†]

Though we receive a word with ambiguous information, we humans can interpret it properly, so we can hold a conversation, and take proper actions. This is possible because we have “common sense” concerning the word, which comes from knowledge accumulated from long time experiences. In order to realize an intelligent computer which can talk with human beings, we think that a construction of the system which understands word meaning is necessary. An association mechanism, which associates one word (concept) with other similar concepts, is indispensable to this construction. This paper describes a method of measuring the degree of association, which evaluates the relevance between concepts based on the Concept-Base, which defines the meaning of words. There is a problem in the conventional method, which evaluates the relevance between concepts using the degree of overlapping of the attribute sets in the Concept-Base. This paper aims to solve the problem of this conventional method and proposes a method of measuring the degree of association using the coincidence information between concepts.

KeyWords: *degree of association, concept-base, commonsense, chain of concept, coincidence information*

[†] 同志社大学大学院工学研究科知識工学専攻, Department of Knowledge Engineering & Computer Sciences, Graduate School of Engineering, Doshisha University Graduate School

1 はじめに

人間はあいまいな情報を受け取り適宜に解釈して適切に会話を進めたり適切な行動を取ることができる。これは、人間が長年にわたって蓄積してきた、言語やその基本となる語概念に関する「常識」を持っているからである。すなわち、ある単語から概念を想起し、さらに、その概念に関連のある様々な概念を連想できる能力が重要な役割を果たしていると考えられる。ここで、ある単語に関連のある様々な単語を連想できるためには、単語間の意味的類似性だけでなく、単語間に存在する常識的な関係も含めた単語間の距離を評価できる必要がある。単語間の意味的な類似度の計算や距離計算は、自然言語処理における基本要素技術である。

本稿では、単語間の距離計算法を提案している。従来、単語間の距離は、単語同士が意味的にどの程度似ているかを表すものであるとして、「類似度」と呼ばれている。単語の意味的類似性には直接的類似性や間接的類似性があり、また、間接的類似性はさらに細かく分類される(内海 2002)。直接的類似性は辞書のカテゴリの類似であるのに対し、間接的類似性は辞書のカテゴリ以外の類似である。たとえば、「大人」と「子供」は同じ「人」に分類されるため意味的に似ており、類似度は高いはずであるが、「子供」と「おもちゃ」は意味的には似ていないし、同じ分類には含まれないであろうから類似度は低くなるであろう。しかし、実際には「子供」から「おもちゃ」を連想できることから、両者の距離はある程度近いものと思われる。「子供とおもちゃ」のような何らかの関連があるもの同士にも距離を定義できるようにするため、本研究では、単語間の距離のことを「関連度」と呼んでいる。もちろん、関連度には類似度の性質も含まれている。すなわち、直接的類似性が高いものも関連度は大きいと考えられる。本研究では、直接的類似性や間接的類似性を問わず、人間が常識的にイメージする単語間の距離に近いほど、その関連度計算法は優れていると判断する。

このような関連度を計算するには、従来用いられてきた単語間の意味的（あるいは分類的）上位下位関係を記述したシソーラス(NTT コミュニケーション科学研究所 1997) などでは困難である。また、ある文書空間内での共起情報を用いれば関連度を計算可能と思われるが、どのような文書空間を用いるべきかが問題となる。本研究では、文書空間として概念ベースを用いる。概念ベースは（後述するが）、国語辞書や大量の新聞記事を用いて構築したものであり、仮想的な文書空間と捉えることができる。

以下、2章では本研究で用いる概念ベースの構造について述べる。3章では、概念間の関連性の評価法に対する既存研究についてふれ、関連度計算法自体の評価の方法を述べる。4章では、本稿の主題である関連度計算法について従来法を述べ、評価考察を行った後、5、6章で新しい計算法についての提案と評価考察を行う。なお以下では、「単語」を「概念」あるいは「概念表記」と呼ぶ。これは、「単語」と言う言葉はその表記をさす場合とその単語の意味、すなわち、その単語が指し示す概念を表す場合があるため、それらを区別するために、表記を表す場合は「概念表記」、意味を表す場合は「概念」と呼ぶ。ただし、厳密な区別が困難な場合も多い

電車	電車,0.36	汽車,0.21	線路,0.10	...	一次属性
	電車,0.36	汽車,0.42	線路,0.38	...	
	汽車,0.21	電車,0.23	鉄道,0.25	...	二次属性
	線路,0.10	列車,0.11	電車,0.18	...	
	...	...	...	...	

図 1 概念「電車」(二次属性まで展開)

ので, その場合は「単語」と呼ぶこととする.

2 概念ベース

2.1 概念ベースの構造

概念ベースにおいて, 任意の概念 A は, 概念の意味特徴を表す属性 a_i と, この属性 a_i が概念 A を表す上でどれだけ重要かをあらわす重み w_i の対で表現される. 概念 A の属性数を N 個とすると, 概念 A は以下のように表せる. ここで, 属性 a_i を概念 A の一次属性と呼ぶ.

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_N, w_N)\} \quad (1)$$

また, 概念 A の一次属性 a_i は概念ベースに定義されている概念としているため, a_i から同様に属性を導くことができる. a_i の属性 a_{ij} を概念 A の二次属性と呼ぶ. さらに三次属性, 四次属性, ... と属性展開が可能であるため, 任意の概念 A は属性 (概念) の無限連鎖により定義されていることになる.

概念「電車」を二次属性まで展開した様子を図 1 に示す.

2.2 概念ベースの構築方法

当初の概念ベースは, 複数の電子化国語辞書を用いて機械的に自動構築されたものである. この概念ベース (基本 CB) (笠原, 松澤, 湯川, 石川, 河岡 1993; 笠原, 松澤, 石川 1997) では, 約 3 万 4 千の概念表記 A とその属性 a_i および重み w_i を複数の国語辞書の語義文から自動的に獲得している. 辞書の見出し部の単語を概念表記とし, 語義文に含まれる自立語を属性として抽出し, それらの重みは属性の出現頻度を基に付与している. さらに, 属性の自己参照による新たな属性の追加, 及び不要な属性の統計的な除去からなる精錬を行うことによって概念ベースを機械構築している.

しかし, 基本 CB は国語辞書の語義文から機械的に構築されているため, 人間の感覚では必要な属性が抜け落ちていたり, 明らかにおかしな属性が雑音として含まれているといったよう

表 1 関連度の例

基準概念	対象概念	関連度
電車	汽車	0.36
	自動車	0.25
	馬	0.09

に、必ずしも適切なデータのみで構成されているわけではない。また、直接的類似性（辞書的カテゴリの類似）を評価するには適しているが、間接的類似性（辞書のカテゴリ以外の類似、関連）を評価するには必ずしも十分ではないと思われる。そこで、本研究では、不適切なデータを削除し、必要なデータを追加する自動精練処理を行った概念ベース（概念数約 9 万）（広瀬，渡部，河岡 2002）を構築し利用する。この概念ベースは、基本 CB の概念に加えて、新聞記事における単語（概念表記）の共起情報を元に新たな概念表記と属性を追加し、属性信頼度（小島，渡部，河岡 2002, 2004）と使用した新聞記事全体における *idf* 値を元に属性の重みを付与し、重みの小さい属性は削除するといった処理を施したものである。ただし、この概念ベースもあくまで機械構築であるため、人間の感覚と一致する属性の率は（サンプル評価によると）約 6 割である。

3 概念間の関連性評価法

3.1 概念間の関連度

概念間の関連度とは、2つの概念 *A* と概念 *B* の関連の強さを定量的に評価するものである。例えば、概念「電車」に対して、「汽車」、「自動車」、「馬」の関連の強さがどれほどのものかを知りたいとき、表 1 のように関連の強さを定量化（数値化）できれば、コンピュータにも判断できるようになる。この場合、概念「電車」に対しては、「汽車」が最も関連が強いということがわかり、また「汽車」、「自動車」、「馬」の順に関連が強いこともわかる。

概念と概念の関連性（類似度または関連度）を評価する方法としては、シソーラスを用いる方法（黒橋，長尾 1992; 長尾 1996; Sumita and Iida 1992; Fujii, Inui, Tokunaga, and Tanaka 1998; Uramoto 1994; 大井，隅田，飯田 1997）やベクトル空間モデルによる方法（Salton and McGill 1993; Schutze 1998; Fujii et al. 1998; 稲子，笠原，松澤 2000; 笠原他 1997; 小嶋，伊藤 1997）等がある。

ベクトル空間モデルによる方法では、単語ベクトルの表現方法が問題となる。たとえば、概念数 100 個の概念ベースを用いて単語ベクトルを構成する場合、もっとも単純には、100 次元の単語ベクトル空間を用いることとなる。しかし、ベクトル空間モデルでは各基底ベクトルは互いに独立している必要があるのに対し、このような単純な方法では、各基底ベクトルはまっ

たく独立していない。そこで通常は、概念数より少ない次元に次元圧縮を行い各基底ベクトル間の独立性を高める方策が採られる。(笠原他 1997) では辞書の語義文を元に作成された概念ベース (基本 CB) を利用し、各単語 (概念) をシソーラスを用いて次元圧縮を行っている。すなわち、約 2700 のシソーラスのノードを基底ベクトルとみなすことで、約 3 万 4 千個の基本概念ベースの概念を約 2700 次元のベクトル空間で扱っている。

また、同じく概念ベースを利用する方法として、概念の属性の一致度と重みを利用する意味関連度計算方式 (井筒, 渡部, 河岡 2002; 渡部, 河岡 2001) がある。この方法では、概念をベクトルとはみなさず、重み付き属性の集合として扱う点が特徴である。詳細は 4 章で述べる。

シソーラスを用いる方法よりも概念ベースを用いたベクトル空間モデルによる方法が優れているという報告 (川島, 石川 2005) があり、また、ベクトル空間モデルによる方法よりも意味関連度計算方式の方が良い評価結果が得られている (渡部, 河岡 2001) ため、本稿では意味関連度計算方式の拡張方式について報告する。すなわち、概念ベースにおける概念表記と概念表記の共起情報を用いた共起関連度計算方式を提案し、また、共起関連度計算方式と意味関連度計算方式を複合利用する関連性評価方式について提案する。

なお、本論文では、既存論文ではほぼ同じ状況で比較実験が行われている中で最も優れていると考えられる手法との比較実験のみを行うが、より広い意味での概念間関連度計算手法同士の比較実験は必要であり、今後の課題とする。特に、(岡本, 石崎 2001) では、大勢の人間を使った連想実験により概念間距離の定式化を試みており、人間の感覚により近い概念間距離を算出できる可能性が高いが、概念数が少なく網羅性が低いいため、本論文では比較対象にはしていない。

3.2 関連度計算法の評価方法

本研究では、直接的類似性や間接的類似性を問わず、人間が常識的にイメージする概念間の距離に近いほど、その関連度計算法は優れていると判断する。そこで、関連度計算法の評価をするために、人手によって作成した評価用データを用いて行う。

評価用データ

人間が任意の概念 X (基準概念) に対して、高い関連を示す概念 A 、関連のある概念 B 、ほとんど関連のない概念 C と判断した 4 つの概念 ($X-A, B, C$) を 1 セットとして大量に用意した。そして、さらにそのデータに対して作成者以外の 3 人に判断してもらい、3 人中 3 人が正しいと判断したデータのみを使用する。また、評価用データの概念はすべて概念ベースで定義されている概念のみで構成されている。本稿では、このように作成した合計 2370 組の評価用データ (表 2) を用いて関連度計算法の評価を行う。

表 2 評価用データ (一部)

X	A	B	C
椅子	腰掛け	机	像
医師	医者	看護婦	山
海	海洋	塩	車
病	病氣	医者	勇気
...	...	...	...

評価方法

評価は概念 X と概念 A との関連度 $Rel(X, A)$, 概念 X と概念 B との関連度 $Rel(X, B)$, 概念 X と概念 C との関連度 $Rel(X, C)$ が,

$$Rel(X, A) - Rel(X, B) > AveRel(X, C) \quad (2)$$

$$Rel(X, B) - Rel(X, C) > AveRel(X, C) \quad (3)$$

$$AveRel(X, C) = (1/2370) \sum_{i=1}^{2370} Rel(X_i, C_i) \quad (4)$$

を満たすときを正解とする. この評価法を C 平均評価法と呼ぶことにする. 概念 X と無関連である概念 C との関連度 $Rel(X, C)$ は 0.0 になるのが理想である. しかし実際はある程度の数値が雑音として算出される. そこで, ただ単に概念 X と概念 A, B, C との関連度が A, B, C の順になるだけではなく, 誤差である $AveRel(X, C)$ の値より優位な差がありなおかつ概念 X と概念 A, B, C との関連度が A, B, C の順になるものだけを正解とした. 関連度計算法で評価用データ全体の正解率が高ければ人間の感覚に近い判断ができていくことになる.

なお, 評価方法についても, 各論文ごとに異なっているのが現状である. 本論文では, 既存研究の評価方法を参考にしながら, 人間の感覚により近いものをよしとし, なるべく機械的に客観的に評価したいという観点から, 上記の評価方法を取っている.

4 意味関連度計算方式

概念ベースでは, 各概念はその概念に何らかの意味で関連する概念 (属性) と重みの対の集合として定義されている. この属性と重みの対の集合は, その概念の意味を近似的に表しているものととらえられる. したがって, 二つの概念間の意味的な距離は, それらの属性と重みの対の集合がどの程度似ているかを評価することで得られると考える. ところで, 各概念の属性はまた概念でもあるため, 属性同士が完全に同じでなくとも似ている度合いを評価可能である. すなわち, 二つの概念間の意味的な距離は, 再帰的に, 属性と重みの対の似ている度合いを評価することにより得られる. この再帰は, 厳密には, 無限に深いものであるが, 計算効率と精度を考慮して, 二次属性までを使用する.

すなわち, 二つの概念 A, B 間の意味関連度は, 各概念を二次属性まで展開し, 一致する属性と重みによって求める一致度を使い計算する. 一致度は $0 \sim 1$ の実数値をとる. 具体的には, 一致する二次属性を調べ, その重みを使って計算する一致度の和が最大になる一次属性の組み合わせを作る. 一次属性の組み合わせを作る場合には, 遺伝的アルゴリズムなどを使うことによって最適な組み合わせを作ることが考えられるが, 最大値を取る組み合わせを順に取ることによって比較的良い関連度が得られることがわかっている (浮田, 渡部, 河岡 1998). 以下, 二つの概念 A, B の意味関連度を $mathit{MR}(A, B)$ と定義し, 意味関連度の計算方法を示す.

4.1 計算方法

一致度の計算方法

概念 A, B をその一次属性を a_i, b_j , 重みを u_i, v_j とし, 属性がそれぞれ L 個, M 個 ($L \leq M$) とすると,

$$A = \{(a_1, u_1), (a_2, u_2), \dots, (a_L, u_L)\} \quad (5)$$

$$B = \{(b_1, v_1), (b_2, v_2), \dots, (b_M, v_M)\} \quad (6)$$

と表現できる. このとき, 概念 A と B の一致度 $MatchWR(A, B)$ は,

$$MatchWR(A, B) = \sum_{a_i=b_j} \min(u_i, v_j) \quad (7)$$

(各概念の重みの総和は 1 に正規化する) と定義する. ただし, $a_i = b_j$ は属性 (概念表記) a_i と b_j とが同じであることを表す. このとき, 一致度は一致する属性のうち小さい方の重みとなるが, これは両方の属性に共通して存在する重み分は有効だと考えるためである.

意味関連度の計算方法

意味関連度は, 対象となる全ての一次属性の組み合わせについて一致度を計算し, 一次属性どうしの対応を決定することにより計算する. 具体的には, 一致する一次属性どうし ($a_i = b_j$) については優先的にその対応を決定する. 一致しない一次属性については, その一致度の合計が最大になるように一次属性どうしの対応を決定する. 一致度を利用することによって, 一致しない (概念表記が異なる) 一次属性についても関連の度合いを考慮に入れることができる.

一次属性どうしが一致するものがない場合, 概念 A, B のうち属性数の少ない概念を $A(L \leq M)$ とし, 概念 A の一次属性の並びを固定する.

$$A = ((a_1, u_1), (a_2, u_2), \dots, (a_L, u_L)) \quad (8)$$

概念 B の各一次属性を対応する概念 A の各一次属性との一致度 ($MatchWR$) の合計が最大に

なるように並べ替える.

$$B_x = ((b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \dots, (b_{xL}, v_{xL})) \quad (9)$$

($\{b_{xL+1}, \dots, b_{xM}\}$ は無視する.) このように対応を決めると, 概念 A と B の意味関連度 $MR(A, B)$ は,

$$MR(A, B) = \sum_{i=1}^L MatchWR(a_i, b_{xi})(u_i + v_{xi})(\min(u_i, v_{xi})/(\max(u_i, v_{xi}))/2) \quad (10)$$

となる. すなわち, 意味関連度は対応する一次属性の一致度と, それらの属性の重みの平均および重みの比に比例すると考える.

一次属性どうしが一致する (概念表記が同じ) ものがある場合 ($a_i = b_j$) は, 別扱いにする. これは概念ベースには約 9 万の概念が存在し, 属性が一致することは稀である. 従って, 属性の一致の扱いを別にすることにより, 属性が一致した場合を大きく評価するためである. 具体的には, 対応する属性の重み u_i, v_j の大きさを重みの小さい方にそろえる. このとき, 重みの大きい方はその値から小さい方の重みを引き, もう一度, 他の属性と対応をとることにする. 例えば, $a_i = b_j$ で $u_i = v_j + \alpha$ とすれば, 対応が決定するのは (a_i, v_j) と (b_j, v_j) であり, (a_i, α) はもう一度他の属性と対応させる. このように対応を決め, 対応の取れた属性の組み合わせが T 個の場合,

$$A' = ((a'_1, u'_1), (a'_2, u'_2), \dots, (a'_T, u'_T)) \quad (11)$$

$$B' = ((b'_1, v'_1), (b'_2, v'_2), \dots, (b'_T, v'_T)) \quad (12)$$

となる.

概念 A と B の意味関連度 $MR(A, B)$ は,

$$MR(A, B) = \sum_{i=1}^T MatchWR(a'_i, b'_i)(u'_i + v'_i)(\min(u'_i, v'_i)/(\max(u'_i, v'_i))/2) \quad (13)$$

となる.

一致度と意味関連度の計算例を概念「机」と「椅子」を例に示す. 二つの概念の一次, 二次属性を表 3 に示す.

一次属性「勉強」と「本棚」の一致度は, 属性「本」のみが一致するため, その重みの小さい方を取り, 次式のように計算できる.

$$MatchWR(\text{勉強}, \text{本棚}) = \min(0.2, 0.1) = 0.1 \quad (14)$$

同様に全ての一次属性の組み合わせについて一致度を計算した結果が表 4 である.

意味関連度の計算は, まず一致部分 (概念表記が一致するもの) から行う. 次に残りの部分

表 3 一次属性と二次属性

概念	一次属性
机	(学校, 0.6), (勉強, 0.3), (本棚, 0.1)
椅子	(勉強, 0.5), (教室, 0.3), (木, 0.2)

(a) 一次属性

一次属性	二次属性
学校	(大学, 0.4), (校舎, 0.4), (木造, 0.2)
勉強	(予習, 0.5), (試験, 0.3), (本, 0.2)
本棚	(図書, 0.6), (書物, 0.3), (本, 0.1)
教室	(教師, 0.4), (校舎, 0.4), (生徒, 0.2)
木	(森林, 0.5), (木造, 0.4), (葉, 0.1)

(b) 二次属性

表 4 一致度マトリックス

	学校	勉強	本棚
勉強	0	1	0.1
教室	0.4	0	0
木	0.2	0	0

について一致度の大きいところから順に対応を決めると意味関連度は次式のように計算できる.

$$\begin{aligned}
 MR(\text{机}, \text{椅子}) = & 1 \times (0.3 + 0.3)(0.3/0.3)/2 \\
 & + 0.4 \times (0.6 + 0.3)(0.3/0.6)/2 \\
 & + 0.1 \times (0.1 + 0.2)(0.1/0.2)/2
 \end{aligned} \tag{15}$$

表 3(a) より, 一次属性「勉強」が一致している. そこで, 重みの大きいほうの「勉強」を二つに分解し, 一致度が大きい順に対応を決めると (一致度は表 4 より, $MatchWR(\text{勉強}, \text{勉強})=1, MatchWR(\text{学校}, \text{教室})=0.4, MatchWR(\text{本棚}, \text{勉強})=0.1$) 以下のようになる.

$$\begin{aligned}
 \text{机} &= ((\text{勉強}, 0.3), (\text{学校}, 0.6), (\text{本棚}, 0.1)) \\
 \text{椅子} &= ((\text{勉強}, 0.3), (\text{教室}, 0.3), (\text{勉強}, 0.2)) \\
 \text{一致度} &= \quad \quad \quad 1, \quad \quad \quad 0.4, \quad \quad \quad 0.1 \\
 &\text{そして, 式 13 に代入すると (この場合 } T = 3 \text{) 式 15 となる.}
 \end{aligned}$$

4.2 評価実験

評価用データを用いて, 意味関連度計算方式の評価実験を行った. 評価実験の結果は C 平均評価法で正解率は 71.1%であった. 詳細を表 5 に示す. ただし, $a >^* b$ は, $Rel(X, a) - Rel(X, b) > AveRel(X, C)$ を表す.

表 5 の結果から概念 X に対して中程度の関連のある概念 B と関連のない概念 C の判定が上

表 5 意味関連度の評価実験結果

判定	正解率
$A >^* B >^* C$	71.1 %
$A >^* B$	88.4 %
$B >^* C$	80.8 %
$A >^* C$	96.4 %

表 6 意味関連度の誤り (一部)

X	A	B	C
道	道路	(車)	(点)
買う	購入	(金)	(雨)
話す	会話	(人)	(煉瓦)
歩く	歩行	(動物)	(黒板)
フェリー	船	(海)	(地藏顔)
でぶ	肥満	(脂肪)	(穴居)

手くいつていないものが多いことがわかる. 表 6 にその誤りであったものの一部を示す. この表のうち, 括弧付で表している概念 (B, C) の判定が誤っていた.

評価用データの概念 A は概念 X と同義・類義といったように意味的に近い単語であるのに対して, 「道—車」, 「買う—金」, 「でぶ—脂肪」のように概念 B は概念 X と意味的に近い単語というよりは概念 X から連想により導き出せる単語が多い.

意味関連度計算方式は概念の属性の一致度合いから関連性を判断しているため, 連想により導き出せるような単語の間の関連性判断は上手くできないと考えられる. (ただし, 概念ベースには新聞記事における共起情報によるデータも多く含まれているため, ある程度高い正解率にはなっている.) 人が「買う—金」の方が「買う—雨」よりも関連が強いと思うのは, 「買う—金」が共に使われる頻度が「買う—雨」が共に使われる頻度よりも高いためであると考えられる. このように概念表記の対がよく出現する概念の間には関連があることを考慮に入れた関連度計算方式を次章で提案する.

5 共起関連度計算方式

人が概念間の関連性を判断する場合, 「自動車—車」, 「電車—汽車」といったように意味的な近さ (直接的類似性) が重要な判断基準となると考えられる. そこでこれまで概念の意味属性の一致度と重みを利用する意味関連度計算方式により, 意味的にどれだけ近いかを判断することで概念間の関連性を評価してきた. ところが意味関連度計算方式では「道—車」, 「買う—金」のように連想により導き出せるような単語の間の関連性の判断が十分に行えないことが評価実

図 2 表記的共起

験によりわかった．人がこれらの単語の間に関連があると判断するのは，共に使われる頻度が他の単語と使われる頻度よりも高いためであると考えられる．そのため意味的な近さを判断するだけの意味関連度計算方式ではこのような単語の間の関連性の判断は難しい．そこで概念表記の対でよく出現する概念間には関連があることを考慮に入れた関連性評価法について考える．

5.1 表記的共起関連度計算方式

ある概念表記の対が複数の文書で共出現する頻度が高ければ高いほどその概念の間には関連があると考えられる．本稿では，これらの共起情報を求めるために概念ベースを利用し，概念を構成する属性集合の中でどれだけ共出現するかを調べる．（概念ベースは，国語辞書や新聞記事から単語を切り出して作成しているため仮想的な文書空間ととらえることができる．）

計算方式

概念表記の対が共出現する頻度から関連性を判断する表記的共起関連度を以下の式で定義する（図 2）．

$$Co(A, B) = \left(\frac{df(A \& B)}{df(A)} + \frac{df(A \& B)}{df(B)} \right) / 2 \quad (16)$$

ただし， $df(A \& B)$ は概念表記 A ， B が共に一次属性に出現する概念数， $df(A)$ は概念表記 A が一次属性に出現する概念数， $df(B)$ は概念表記 B が一次属性に出現する概念数とする．

例えば， $df(\text{自動車} \& \text{運転})$ は，「自動車」と「運転」を共に属性に持つ概念数を数えるわけであるが，「車」と「運転」を共に持つ概念は対象にならない．このように，意味的に近い単語との共起は考慮せず，「自動車－運転」という『表記』の対がどれだけ出現するかを評価するものであるので，表記的共起関連度計算方式と呼ぶことにする．

評価実験

評価用データを用いて，表記的共起関連度計算方式の評価実験を行った．評価実験の結果は C 平均評価法で正解率は 49.7% であった．表 7 に評価実験で誤ったものの一部を示す．

表 7 表記的共起関連度の誤り (一部)

X	A	B	C	$Co(X,A)$	$Co(X,B)$	$Co(X,C)$
火	炎	火事	海	0.455	0.481	0.007
警察	警官	拳銃	青空	0.205	0.000	0.000
ランナー	走者	マラソン	脊養	0.315	0.000	0.000
コーナー	曲り角	道	名作	0.000	0.026	0.000
マスタード	辛子	香辛料	拳世	0.000	0.000	0.000
空オケ	カラオケ	音楽	疎略	0.000	0.000	0.000

表 7 の $Co(\text{コーナー}, \text{曲がり角}) = 0.000$, $Co(\text{マスタード}, \text{辛子}) = 0.000$, のように $Co(X, A) = 0.000$ となるものが 2370 組中 354 組 (14.9%), $Co(\text{警察}, \text{拳銃}) = 0.000$, $Co(\text{マスタード}, \text{香辛料}) = 0.000$ のように $Co(X, B) = 0.000$ となるものが 838 組 (35.4%) と表記的共起関連度 0 が頻出した. 表記的共起関連度計算方式で 5 割程度の正解率を得ることしかできなかったのは, 概念 X に対して高関連である概念 A や中関連である概念 B と共起せず表記的共起関連度が 0 になる評価セットが多数存在したためである.

5.2 意味的共起関連度計算方式

表記的共起関連度計算方式では, 概念ベースの一次属性を用いて, 概念表記が共起する率を調べている. しかし, 評価結果からもわかるように, 関連が強いと思われる概念同士でもその概念表記がまったく共起しない場合が多々現れる. そこで, 一次属性のみではなく, 二次, 三次と使用する属性の範囲を広げていけば, 関連の強い概念同士では共起率が高くなっていくものと思われる. (このことは, 「自動車」と「運転」の共起を調べるときに, 「車」と「運転」の共起も同時に調べることに相当する.) ところが, 二次, 三次と使用する属性の範囲を広げると, 扱うべき概念表記の数が指数関数的に増大していくため計算効率が急激に悪くなる. さらに, 全概念空間を探索することになるが, ほとんどは無関係の部分であり非常に無駄が多い. そこで, 全概念空間を探索するのではなく, 可能性の高そうな部分のみに限定して意味的な共起を調べる手法を提案する.

概念 A と概念 B の関連が強いほど, それぞれの概念の意味特徴を表す属性集合内に対象とする概念表記が数多く出現すると考えられる. すなわち, 概念 A の n 次属性に概念表記 B が多数現れ, 概念 B の n 次属性に概念表記 A が多数出現するものと考えられる. このような性質から関連性を判断する方法を意味的共起関連度計算方式と呼ぶことにする (図 3).

図 3 意味的共起

計算方式

対象とする属性集合をどこまで広げるか, すなわち何次属性までを対象とするか, という問題がある. そこで一次, 二次, 三次属性まで展開した場合の意味的共起関連度を以下のように定義する.

$$CoM1(A, B) = \left(\frac{Nb_1}{NA_1} + \frac{Na_1}{NB_1} \right) / 2 \quad (17)$$

$$CoM2(A, B) = \left(\frac{Nb_2}{NA_2} + \frac{Na_2}{NB_2} \right) / 2 \quad (18)$$

$$CoM3(A, B) = \left(\frac{Nb_3}{NA_3} + \frac{Na_3}{NB_3} \right) / 2 \quad (19)$$

ただし, NA_n は概念 A の n 次属性数, NB_n は概念 B の n 次属性数, Na_n は概念表記 A が概念 B の n 次属性内に出現する回数, Nb_n は概念表記 B が概念 A の n 次属性内に出現する回数とする.

しかし, 上記の計算方法では多少問題があるように思われる. 例えば概念 A の属性集合内に概念表記 B が 10 回出現する場合と概念表記 C が 10 回出現する場合を考える. 出現回数が同じだからといって概念 A に対して概念 B と概念 C は同じくらい関連があるとはいえないであろう. なぜなら, 概念ベースの全概念の属性内に 10 回しか出現しない概念表記 B が, 概念 A の属性集合内に 10 回出現する場合と, 概念ベースの全概念の属性内に 10000 回出現する概念表記 C が, 概念 A の属性集合内に 10 回出現する場合とでは, 明らかに概念 A にとって概念 B の方が概念 C よりも関連があると考えられるからである. そこで, 情報検索の分野でよく用いられる稀に出現する語を重要とする idf 値を利用して評価する方法を, 以下の式で定義する.

$$CoM1_{idf}(A, B) = \left(\frac{Nb_1 \times idf(B)}{NA_1} + \frac{Na_1 \times idf(A)}{NB_1} \right) / 2 \quad (20)$$

$$CoM2_{idf}(A, B) = \left(\frac{Nb_2 \times idf(B)}{NA_2} + \frac{Na_2 \times idf(A)}{NB_2} \right) / 2 \quad (21)$$

$$CoM3_{idf}(A, B) = \left(\frac{Nb_3 \times idf(B)}{NA_3} + \frac{Na_3 \times idf(A)}{NB_3} \right) / 2 \quad (22)$$

図 4 パラメータ実験結果

$$idf(t) = \log \frac{N_{All}}{df(t)} + 1 \quad (23)$$

ただし, N_{All} は概念総数 (87242), $df(t)$ は概念表記 t を三次属性内に持つ概念の数である. ここで, idf 値を求める際の出現頻度 df 値は三次属性を用いているが, これは, 実験的に検証した (坂田, 渡部, 河岡 2004) ものである.

評価実験

出現回数のみで評価する方法と idf 値を用いて評価する方法で精度比較を行う. ここで属性を何次まで展開すればいいか, また属性をいくつまで使うのがいいのかを評価する必要がある. そこで, 評価用データに対し, 打ち切り属性数を 10 から 100 まで 10 間隔で変化させたものと打ち切りなし, つまりすべての属性を使う場合で, 評価用データの正解率にどのような変化があるのか調査した. その結果を図 4 に示す. 図 5 にはそれぞれの計算方式で最も良かった正解率のグラフを示す.

一次属性, 二次属性まで展開する場合は, 属性はすべて使用した場合がもっとも精度がよかった. 三次属性まで展開する場合は, 打ち切り属性数 30 のとき最も精度がよかった. また, すべての場合において, ただ単に出現回数だけで評価する方法よりも idf 値を利用した方法のほうが精度がよかった.

今回提案した意味的共起関連度計算方式のうち最も精度がよかったのは, それぞれ使用する属性数を 30 にして三次属性まで展開して idf 値を用いる計算方式であり, 正解率は 68.1% であった.

図 5 意味的共起関連度評価結果

5.3 共起関連度に関する考察

概念表記の共起情報から関連性を判断する計算方法として表記的共起関連度計算方式と意味的共起関連度計算方式を提案した。評価実験の結果、表記的共起関連度計算方式より意味的共起関連度計算方式のほうが精度が良かった。

意味的共起関連度計算方式で最も精度の良かった方式 *CoM3_idf* の正解率は 68.1% であった。これは意味関連度計算方式の正解率 71.1% より低い値である。しかし、意味関連度計算方式で誤った $(X-A, B)$ の 274 組, $(X-B, C)$ の 455 組, $(X-A, C)$ の 85 組に対して *CoM3_idf* で評価したところ, $(X-A, B)$ で 48 組 (正解率 17.5%), $(X-B, C)$ で 243 組 (正解率 53.4%), $(X-A, C)$ で 31 組 (正解率 36.5%) が正解になった。注目すべきところは、意味関連度計算方式で判断が上手くできなかった $(X-B, C)$ の判断が共起関連度計算方式を用いれば適切な判断ができるようになることである。このことから意味関連度計算方式と共起関連度計算方式を複合利用することで、よりよい関連性の判断ができると考えられる。

6 意味共起関連度計算方式

4章で述べた意味関連度計算方式と5章で述べた共起関連度計算方式を複合利用することで、よりよい関連性の判断の実現を目指す。

表 8 パラメータ実験結果

計算方式	C_w	正解率
MR	0.000	71.1%
MR+Co	0.255	75.1%
MR+CoM 1	1.130	72.3%
MR+CoM 1_idf	0.540	72.3%
MR+CoM 2	9.620	76.0%
MR+CoM 2_idf	8.970	76.5%
MR+CoM 3	9.695	75.6%
MR+CoM 3_idf	9.815	75.8%

6.1 意味関連度と共起関連度の合成

意味関連度計算方式と共起関連度計算方式を合成して評価する式を以下の式で定義し、この計算方式を意味共起関連度計算方式と呼ぶことにする。

$$MCR(A, B) = \frac{MR(A, B) + C_w \times CR(A, B)}{1 + C_w} \quad (24)$$

ただし、 $MR(A, B)$ は意味関連度、 $CR(A, B)$ は共起関連度、 C_w は重み定数とする。

意味共起関連度計算方式では、共起関連度に重み C_w を付与し、意味関連度に足し合わせ意味共起関連度を求める。

6.2 意味共起関連度計算方式の評価実験

評価用データを用いて、意味共起関連度計算方式の評価実験を行う。評価方法は、共起関連度に付与する重み C_w を 0.0 から 10.0 まで 0.005 間隔で変化させ、評価用データの正解率が最大になる C_w を調べた。その結果を表 8 と図 6 に示す。

表 8、図 6 から、最も精度が良かったのは、意味関連度計算方式と共起関連度計算方式 $CoM2_idf$ に重み $C_w=8.970$ を付与して組み合わせた意味共起関連度計算方式 (正解率 76.5%) で、意味関連度計算方式単独の正解率 71.1% と比較して 5.4% 精度向上したことがわかる。

評価用データ 1 組に対し 3 回 ($X-A, X-B, X-C$) の関連度計算を行っているので 2370 組では 7110 個の関連度の値が算出される。そこで意味関連度計算方式と共起関連度計算方式 $CoM2_idf$ の関連度分布 (降順) の全体を図 7、関連度上位 20 個の部分を図 8 に示す。図より意味関連度の値が 0 から 1 の範囲に分布しているのに対して、共起関連度が 0.1 より大きいのは 7110 個中わずか 9 個であり、共起関連度の値は 0 から 0.1 の範囲に分布していることがわかる。意味関連度に対して、共起関連度に付与する重みが $C_w=8.970$ と大きな値になっているのはこのためである。

意味関連度計算方式に対して共起関連度計算方式がどれだけ効果があるかを知るためには、意味関連度と共起関連度の範囲をあわせる必要がある。そこで以下の式で共起関連度 CR の値

図 6 意味共起関連度評価結果

図 7 関連度の分布 (全体)

を 0 から 1 の範囲に拡張した.

$$CR = \begin{cases} 1.0 & \text{if } CoM2_idf \geq 0.10 \\ 10 \times CoM2_idf & \text{otherwise} \end{cases} \quad (25)$$

意味関連度と拡張した共起関連度を組み合わせた意味共起関連度計算方式について同様の実

図 8 関連度の分布 (上位 20)

験を行ったところ図 9 のようになり、 $C_w = 0.90$ の場合に意味共起関連度計算方式の精度が最大の 76.5% になることがわかった。

6.3 意味共起関連度計算方式の考察

意味関連度計算方式のみで概念間の関連性を評価した場合の正解率は 71.1% であった。意味関連度と重みを付与した共起関連度を加える意味共起関連度計算方式では 76.5% と意味関連度計算方式よりも 5.4% 高い精度で関連性の判断を行うことができた。また、意味関連度計算方式 (正解率 71.1%) と共起関連度計算方式 ($CoM2_idf$, 正解率 62.9%) それぞれの正解率をみると、意味関連度計算方式の方が共起関連度計算方式より正解率が高いことから、意味関連度に対して、共起関連度に付与する重み、 $C_w = 0.90$ という値は意味関連度計算方式の誤差を共起関連度計算方式が補正するための値として適切であるといえる。

表 9 に意味関連度計算方式と意味共起関連度計算方式の評価結果の詳細を示す。注目すべきところは、意味関連度計算方式で ($X-B, C$) の正解率が 80.8% だったのが、意味共起関連度計算方式では 86.2% と 5.4% 精度向上しているところである。

表 10 に意味関連度計算方式で誤って、意味共起関連度計算方式で正解した評価結果の一部を示す。これは、表 6 で示した意味関連度の誤りの評価例であり、意味共起関連度計算方式により正しく判断ができるようになった。

ただし、同じく表 10 からわかるように、関連度の値そのものは人間が感覚的に抱く概念間

図 9 パラメータ評価実験結果 (一部)

表 9 意味関連度と意味共起関連度の評価結果

	意味	意味共起
判定	正解率	正解率
$A >^* B >^* C$	71.1%	76.5%
$A >^* B$	88.4%	88.8%
$B >^* C$	80.8%	86.2%
$A >^* C$	96.4%	97.0%

表 10 意味共起関連度評価結果 (一部)

X	A	B	C	MCR(X,A)	MCR(X,B)	MCR(X,C)
道	道路	車	点	0.30561	0.00918	0.00299
買う	購入	金	雨	0.38682	0.00677	0.00080
話す	会話	人	煉瓦	0.23099	0.01080	0.00083
歩く	歩行	動物	黒板	0.29655	0.00501	0.00024
フェリー	船	海	地藏顔	0.02995	0.01704	0.00002
でぶ	肥満	脂肪	穴居	0.24363	0.02628	0.00132

の距離に相当するところまでは行っていない。現状では、高関連・有関連・無関連をほぼ区別できる程度であり、更なる改良が必要である。

7 おわりに

本稿では、直接的類似性や間接的類似性を問わず、人間が常識的にイメージする概念間の距離を算出する関連度計算法を提案した。従来、主に用いられていた概念の意味属性の一致度合から関連性を判断する意味関連度計算方式では、「自動車一車」のように意味的に近い単語に対しては精度の良い関連性判断が行えた。特に、概念ベースが辞書をベースに作成されたものである場合は、辞書の意味が近いかどうかを評価する意味関連度計算方式が優れていると言える。しかし、「道一車」のように意味的に近い単語ではなく連想により導き出せるような単語の間の関連性の判断は十分な精度が得られていなかった。そこで、本稿では、概念ベースに新聞記事からの共起情報を元に抽出した概念表記と属性を加え、さらに、概念表記と概念表記の共起情報から関連性を判断する共起関連度計算方式を提案した。この共起関連度計算方式と意味関連度計算方式を複合利用する意味共起関連度計算方式により、従来、関連性評価として用いていた意味関連度計算方式よりも的確に関連性の判断ができるようになったと考えられる。もちろん、適用目的が類似度の場合には意味関連度計算方式のみを用いることになる。

謝辞 本研究を進めるにあたって、評価実験などで適切な支援を頂いた青田正宏氏に感謝いたします。また、本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った。

参考文献

- Fujii, A., Inui, K., Tokunaga, T., and Tanaka, H. (1998). "Selective sampling for example-based word sense disambiguation." *Computational Linguistics*, **24** (4), pp. 573-597.
- 広瀬幹規, 渡部広一, 河岡司 (2002). "概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精練手法." 信学技報 TL2001-49, pp. 109-116.
- 井筒大志, 渡部広一, 河岡司 (2002). "概念ベースを用いた連想機能実現のための関連度計算方式." 情報科学技術フォーラム FIT2002, pp. 159-160.
- 稲子希望, 笠原要, 松澤和光 (2000). "複合語内単語共起による名詞の類似性判別." 情報処理学会論文誌, **41** (8), pp. 2291-2298.
- 笠原要, 松澤和光, 石川勉 (1997). "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, **38** (7), pp. 1272-1283.
- 笠原要, 松澤和光, 湯川高志, 石川勉, 河岡司 (1993). "アバウト推論のための多観点概念ベース構築と評価." 人工知能学会全国大会.
- 川島貴広, 石川勉 (2005). "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." 人工知能学会論文誌, **20** (5), pp. 326-336.
- 小島一秀, 渡部広一, 河岡司 (2002). "連想システムのための概念ベース構成法-属性信頼度の

- 考え方に基づく属性重みの決定.” 自然言語処理, **9** (5), pp. 93–110.
- 小島一秀, 渡部広一, 河岡司 (2004). “連想システムのための概念ベース構成法—語間の論理的関係を用いた属性拡張—.” 自然言語処理, **11** (3), pp. 21–38.
- 小嶋秀樹, 伊藤昭 (1997). “文脈依存的に単語間の意味距離を計算する一手法.” 情報処理学会論文誌, **38** (3), pp. 482–489.
- 黒橋禎夫, 長尾真 (1992). “長い日本語文における並列構造の推定.” 情報処理学会論文誌, **33** (8), pp. 1022–1031.
- 長尾真 (編) (1996). 自然言語処理. 岩波書店.
- NTT コミュニケーション科学研究所 (編) (1997). 日本語語彙体系. 岩波書店.
- 岡本潤, 石崎俊 (2001). “概念間距離の定式化と電子化辞書との比較.” 自然言語処理, **8** (4), pp. 37–54.
- 大井耕三, 隅田英一郎, 飯田仁 (1997). “意味的類似性と多義解消を用いた文書検索手法.” 自然言語処理, **4** (3), pp. 51–70.
- 坂田光広, 渡部広一, 河岡司 (2004). “関連度と属性の情報価値を考慮した概念ベースの自動精錬手法.” 同志社大学理工学研究報告, **45** (1), pp. 14–22.
- Salton, G. and McGill, M. (Eds.) (1993). *Introduction to modern Information Retrieval*. McGraw-Hill.
- Schutze, H. (1998). “Automatic word sense discrimination.” *Computational Linguistics*, **23** (1), pp. 97–123.
- Sumita, E. and Iida, H. (1992). “Example-based transfer of Japanese adnominal particles into English.” *IEICE Transactions on Information and Systems*, **E75-D** (4), pp. 585–594.
- 浮田知彦, 渡部広一, 河岡司 (1998). “概念間の関連度計算への遺伝的アルゴリズムの適用.” 情報処理学会春期全国大会 1U-2.
- Uramoto, N. (1994). “Example-based word sense disambiguation.” *IEICE Transactions on Information and Systems*, **E77-D** (2), pp. 240–246.
- 内海彰 (2002). “言語と類似性.” 人工知能学会誌, **17** (1), pp. 8–13.
- 渡部広一, 河岡司 (2001). “常識的判断のための概念間の関連度評価モデル.” 自然言語処理, **8** (2), pp. 39–54.

略歴

渡部 広一: 1983 年北海道大学工学部精密工学科卒業. 1985 年同大学院工学研究科情報工学専攻修士課程修了. 1987 年同精密工学専攻博士後期課程中途退学. 同年, 京都大学工学部助手. 1994 年同志社大学工学部専任講師. 1998 年同助教授. 工学博士. 主に, 進化的計算法, コンピュータビジョン, 概念処理などの研究に従事. 言語処理学会, 人工知能学会, 情報処理学会, 電子

情報通信学会, システム制御情報学会, 精密工学会, 日本知能情報ファジィ学会各会員.

奥村 紀之: 2003 年同志社大学工学部知識工学科卒業. 2005 年同大学院工学研究科知識工学専攻博士前期課程修了. 同年, 同志社大学大学院工学研究科知識工学専攻博士後期課程入学. 主に, 概念処理の研究に従事.

河岡 司: 1966 年大阪大学工学部通信工学科卒業. 1968 年同大学院修士課程修了. 同年, 日本電信電話公社入社, 情報通信網研究所知識処理研究部長, NTT コミュニケーション科学研究所所長を経て, 現在同志社大学工学部教授. 工学博士. 主にコンピュータネットワーク, 知識情報処理の研究に従事. 言語処理学会, 人工知能学会, 電子情報通信学会, 情報処理学会, IEEE(CS) 各会員.

(2005 年 4 月 4 日 受付)

(2005 年 6 月 6 日 再受付)

(2005 年 7 月 18 日 採録)