Vol. 35 No. 3 情報処理学会論文誌 Mar. 1994

ショートノート

観点に基づく概念間の類似性判別

笠原 要[†] 松澤和光[†]
石川 勉[†] 河岡 司[††]

言葉によって表される「概念」間の類似性は，その概念が扱われる問題や状況等の「観点」によって変化する．本稿では，概念に関する知識（「概念ベース」）を収集することによって，「観点」に基づく概念間の類似性を判別しうる方式（「観点変調方式」）を提案する．この方式は，類似概念の判定を行う前に観点中の重要な情報に応じて概念中の情報の一部を強調することを特徴とする．また，約千個の概念に関する概念ベースを実際に構築し，情報検索における再現率-適合率の考え方を利用して類似判別方式の評価を行った．この結果，提案方式がシソーラスを用いる従来の方式に比べて有効となる見通しを得た．

Viewpoint-Based Measurement of Semantic Similarity between Words

Kaname Kasahara,[†] Kazumitsu Matsuzawa,[†] Tsutomu Ishikawa[†]
and Tsukasa Kawaoka[††]

This paper proposes a method for measuring semantic similarity between words by using a knowledge base of words consisting of attributes. Because the degree of similarity changes according to viewpoint, feature of this method allows some part of the attributes to be chosen by a word representing the viewpoint of judgement. To test the effectiveness of this method, we made an experimental knowledge base including a thousand words. The similarity measurements with the proposed method were closer to those decided by human judges than were similarity measurements made using the conventional way of using a thesaurus.

1. はじめに

近年，知識処理や言語処理において，人間の保有する「常識」を計算機上に表現し，利用しようとする試みが行われている．その中でも「概念」，すなわち言葉の持つ意味についての大規模な知識を獲得し，概念間の類似性を判別することが注目されている．

概念間の類似関係を表現するものとしては，シソーラス，類語辞典等が存在する．しかし，これらは固定的に概念間の関係を表現しているので，状況や問題に応じた類似判別を行うことはできない．例えば概念「馬」に対する類似概念を「豚」と「自動車」の内から選択するとき，動物に関する問題においては「豚」が，反対に乗物に関する問題では「自動車」が類似している判別が望まれる．しかし，現状のシソーラス等ではこうした選択はできない．

これに対し，問題や状況による類似性の違いに注目したいくつかの研究がある．例えば，概念の多重的な階層関係や比喩を「視点」を用いて表現する研究[1]や，複数のキー事例に共通する特徴を「観点」として抽出し，事例検索を行う研究[2]等である．ただし，これらの研究では，意味素のような概念を構成する根本的な情報があらかじめ厳密に獲得できることを前提としている．したがって「常識」を形成しうる膨大な量の概念に関する類似性を取り扱う場合には，これらの情報をいかにして獲得していくかが大きな問題となる．

そこでわれわれは，既存の知識源から獲得しうる，概念の情報データベース（概念ベース）をもとにした

† NTT 情報通信網研究所
NTT Network Information Systems Laboratories
†† NTT コミュニケーション科学研究所
NTT Communication Science Laboratories

類似判別の可能性の検討を行っている[3]．本稿では，その第一歩として，概念に関して辞書等の文書情報から機械的に獲得できる情報による概念ベースを用いて類似判別を行う類似判別方式（観点変調方式）を提案する．この方式は，問題や状況を表す「観点」を概念で表し，「観点」の保有する重要な情報に応じて比較概念の情報を強調した後で概念間の類似度を算出して判別を行うことを特徴とする．

また，情報検索における適合率-再現率を類似判別方式の評価に利用し，実際に千個程度の概念に関する概念ベースを構築して，観点変調方式の評価を行う．

2. 観点変調方式

観点に基づく概念の類似判別法である観点変調方式について提案する．この方式の特徴は，概念間の類似判別を行う前に，判別の観点に基づいて概念の内容を変調させる点にある．これは，人間が概念を用いて思考するときに，概念に含まれる内容すべてを用いるのではなく，観点によって概念の一部分だけに注目して思考していることをモデル化したものである．

2.1 概念の表現

概念は，理想的には属性値やその関係を表す属性名等種々の表現法が考えられるが，多量の概念について表現することは難しい．そこで，ここでは機械的に獲得しうる概念の情報に着目し，概念 G を，概念の特徴を表す属性 p_i と，その属性 p_i が概念 G においてどれだけ重要であるかを表す重要度 q_i $(0 \leq q_i \leq 1)$ の対の集合により表現する．

$$G = \{(p_1, q_1), (p_2, q_2), \cdots, (p_n, q_n)\} = \{(p_i, q_i)\} \quad (i = 1, \cdots, n, \sum_{i=1}^{n} q_i = 1). \quad (1)$$

ここでは，重要度の和が一定とする．また，属性や重要度を厳密に決定するのは不可能であり，辞書等における説明文中の自立語を属性，属性と概念の共起頻度を重要度とみなす．これにより，多数の概念の獲得，追加が機械的に行える．例えば，概念「馬」は以下のように表現される．

$G_{馬}$ = {(家畜, 0.2), (たてがみ, 0.1), (蹄, 0.1), (草食, 0.05), …}.

2.2 概念の規格化

類似判別は，概念中の属性同士の比較を基本とするが，表記の比較のみでは異なる表記で意味の近い属性を類似していると判定できない．そこで，概略的な比較を行うために，属性を分類する分類体系 C を導入する．

$$C = \{c_1, c_2, \cdots, c_j, \cdots, c_k\}. \quad (2)$$

ここで，分類 c_j は，同様な意味を持つものとして分類された属性の集合を表す．

$$c_j = \{p_{j1}, p_{j2}, \cdots, p_{jl}\} \quad (3)$$

例えば，分類「体」ならば，

$c_{体}$ = {手, 足, たてがみ, 蹄…}.

となる．分類体系には，既存のシソーラス等を用いることができる．

この分類体系 C を用いて，概念の規格化表現 $\hat{G}$ を以下のとおり定義する．

$$\hat{G} = \{(c_j, Q_j)\} \quad (j = 1, \cdots, l, \quad Q_j = \sum_{p_i \in c_j} q_i / \sum_{i=1}^{n} q_i). \quad (4)$$

例えば，先に示した概念 $G_{馬}$ を規格化表現すると，属性「たてがみ」と「蹄」は同じ分類「体」に入るので統合される．

$\hat{G}_{馬}$ = {(生物, 0.5), (体, 0.2), (捕食, 0.1), …}.

ここで，「生物」，「体」，「捕食」は分類体系に含まれる分類名である．

2.3 観点による概念の変調

類似性を判別する際の観点としては種々考えられるが，ここでは，概念ベースに含まれる任意の概念が観点となりうると考える．前述したように，観点中の重要度の高い分類と関連のある概念中の分類の強調を行う．観点の規格化表現 $\hat{K} = \{(c_j, Q'_j)\}$ に基づいて，対応する分類ごとに規格化された概念 $\hat{G} = \{(c_j, Q_j)\}$ 中の重要度を変調させることにより，変調概念 G^K を得る．

$$G^K = \{(c_j, Q_j^K / \sum_{i=1}^{l} Q_i^K)\} \quad (Q_j^K = Q_j \cdot M(Q'_j)). \quad (5)$$

式(5)における M は，変調の度合を決定する変調関数であり，例えば**図1**のような閾値関数が考えられる．この関数は，観点中の一定の基準値以上の重要度

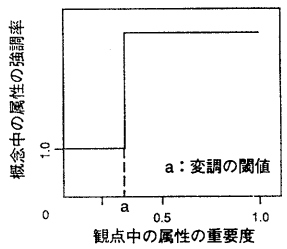


図 1 変調関数の一例
Fig. 1 An example of Modulation Function.

を持つ分類と同じ概念中の分類の重要度を強調する作用がある．また，変調の閾値には，概念ベースに含まれる属性の重要度の平均を用いる．

例えば，観点「動物」の規格化表現

$$\hat{K}_{動物}=\{(生物,0.4),(体,0.1),(捕食,0.1),\cdots\}.$$

に基づいた概念「馬」の変調概念は，変調関数の閾値を 0.2，強調率を2とすると，分類「生物」に対する重要度だけが強調されて，

$$G^{動物}_{馬}=\frac{1}{1.0+0.2+0.1+\cdots}\{(生物,1.0),(体,0.2),(捕食,0.1),\cdots\}.$$

となる．変調した概念の重要度の和は変動するので，再び正規化（和を1とする）を行ったものが最終的な変調概念となる．

2.4 類似度の算出

二つの概念 G_A, G_B 間の観点 K に基づく類似度 $R(G_A, G_B, K)$ は，概念の規格化と観点による変調を行った変調概念 $G^K_A=\{(c_j, Q^K_{Aj})\}$，$G^K_B=\{(c_j, Q^K_{Bj})\}$ をもとにして以下のように定義する．

$$R(G_A, G_B, K)=R'(G^K_A, G^K_B). \quad (6)$$

R' は，観点が介在しない場合の単純な類似度を表し，以下の条件を満たす関数である．

- $0\leq R'\leq 1$.
- $R'(G_A, G_A)\equiv 1$.
- $R'(G_A, G_B)\leq R'(G_C, G_B)$ ならば，概念 G_B に対して G_A よりも G_C の方が類似している．

上記条件を満たす関数 R' は種々考えられるが，ここではその一つとして，

$$R'(G_A, G_B)=\sum_{j=1}^{l}\sqrt{Q^K_{Aj}\cdot Q^K_{Bj}}. \quad (7)$$

を考えた．例えば，観点「動物」に基づいた概念「豚」の変調概念を

$$G^{動物}_{豚}=\{(生物,0.4),(体,0.3),(捕食,0.1),\cdots\}.$$

とすると，観点「動物」に基づく概念「馬」と「豚」の類似度は，

$$R(G_{馬}, G_{豚}, K_{動物})=R'(G^{動物}_{馬}, G^{動物}_{豚})=\sqrt{0.5*0.4}+\sqrt{0.1*0.3}+\cdots.$$

と求められる．以上のように類似度 R を計算して観点に基づく類似判別を行う．

3. 評価法

ここでは，類似判別方式を評価する基準として人間の行う類似判別結果との“近さ”を採用する．提案した方式は，変調関数 M や類似関数 R 等におけるパラメータが種々存在するので，これらを最適化した上で評価する必要がある．しかし，種々のパラメータに対する判別結果をその都度人間によって評価することは困難である．そこで，一度人間が選択した類似概念を基準データとし，情報検索の検索効率を求める手法を用いて評価する．

情報検索では，人間が期待する情報 A，機械が検索した情報 B に対し，適合率 $T=(A\cap B)/B$，再現率 $S=(A\cap B)/A$ の値より検索効率が評価され，$T=1$，$S=1$ に近い検索方式が望ましいとされている[4]．この再現率と適合率を方式評価に利用する．

与えられる比較概念と観点の組に対する類似概念の判別において，人間の選択した類似概念の集合を A，類似度に基づいて選択した類似概念の集合を B とすれば，情報検索と同様に再現率 S と適合率 T が定義される．ここで，類似度の高い順で第何位までの概念を類似概念の集合 B に含めるかに応じて再現率と適合率の値は変化し，図2のように再現率-適合率曲線が描かれる．再現率 [0, 1] 区間の適合率の積分値を平均適合率 $\overline{T}$ と定義する．この値が1に近いほど理想的な判別法と考えられるので，この $\overline{T}$ により方式を評価する．

4. 方式の評価実験

4.1 実験条件

方式評価のため，文献 5) に基づいて出現頻度が高く基本的と思われる 1,049 語の名詞を選び，概念ベースを実際に構築した．各概念に対する属性は，一般の国語辞典等の概念を示す語の説明文から抽出した自立語を用い（平均 12 属性/概念），重要度は説明文中の属性の出現頻度により決定した．人間による評価データは，42個の比較概念と観点の組に対し，被験者（6名）が選択した類似した概念（計 426 概念）を用いた．

観点変調方式と比較するための従来方式としては，シソーラス上での概念間の距離より類似度を算出する

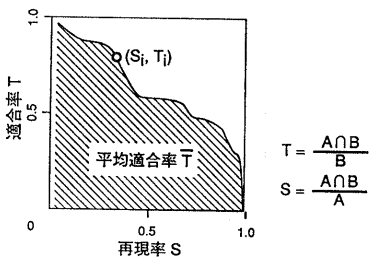


図 2 再現率・適合率曲線
Fig. 2 Recall and precision curve.

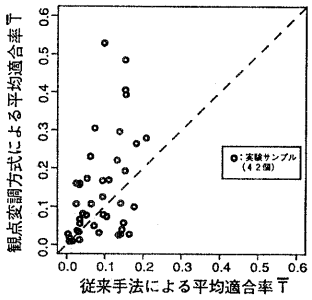


図 3 提案方式の評価
Fig. 3 Estimation of performance of the proposed method.

方式を用いた．すなわち，木構造のシソーラスにおいて，同じ分類に入る概念間の意味的近さを表す距離を0，異なる分類に入る概念間の距離については，二つの分類を隔てる分類のノードの数を距離とする．そして，距離0のときに類似度1，距離が l のときには最長距離を L としたときに類似度を $1-l/L$（L：最長距離）とした．平均適合率 $\overline{T}$ を求めるためには類似度に基づく類似概念間の順位自体を必要とするので，類似度に基づく概念間の距離は評価に影響しない．なお，シソーラスは観点変調方式に利用したシソーラスと同じものを使用した．

4.2 実験結果

図3は，42の評価サンプルについての平均適合率 $\overline{T}$ を求めたものである．縦軸は観点変調方式による $\overline{T}$，横軸は従来手法方式による $\overline{T}$ の値である．

42サンプル中で提案方式の評価が従来手法より高くなるのは29サンプル（約7割），平均適合率 $\overline{T}$ の比は従来手法の2.0倍（42サンプル平均）となる．このことは，辞書から機械的に構築した概念ベースであっても，観点変調方式によって従来手法以上の類似判別が行えることを示している．

5. おわりに

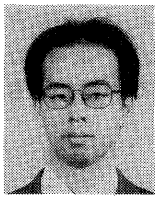
本稿では，概念に関する知識を収めた概念ベースを用い，観点に基づく概念間の類似判別を行う観点変調方式の提案を行った．また，情報検索における評価手法を類似判別に応用し，実際に千個の概念を含む概念ベースを作成して実験を行い，本方式の有効性を示した．

今後はさらに，数万規模の概念を内蔵した概念ベースを構築し，観点変調方式の検討／改善を行う予定である．

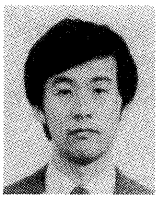
参考文献

1) 岩山 真，徳永健伸，田中穂積：比喩を含む言語理解における視点の役割，情報処理学会自然言語処理研究会報告，NL 73-7, pp. 51-58 (1989).
2) 沢田裕司，大川剛直，馬場口登，手塚慶一：観点を考慮した連想機構の一モデル化，情報処理学会情報学基礎研究会報告，28-2 (1992).
3) 笠原 要，松澤和光，石川 勉：アバウト推論における多観点概念ベース，第45回情報処理学会全国大会論文集，3-27 (1992).
4) 日本ユニバック総合研究所編：共立コンピュータ辞典，共立出版，p. 635 (1976).
5) 国立国語研究所：高校教科書の語彙調査Ⅱ，秀英出版 (1984).

（平成5年 4 月 19 日受付）
（平成5年12月 9 日採録）



笠原 要（正会員）
1964 年生．1989 年慶応大学理工学部化学科卒業．1991年東京工業大学総合理工学研究科電子化学専攻修士課程修了．同年日本電信電話(株)入社．現在，NTT 情報通信網研究所知識処理研究部にて知識処理技術の研究に従事．人工知能学会会員．



松澤 和光（正会員）
1953 年生．1975 年東京工業大学工学部電子工学科卒業．1977年同大学院修士課程修了．同年日本電信電話公社武蔵野電気通信研究所入所．以来，フルウェーハシステム，大規模 ROM，ヒューマンインタフェース，知識処理技術の研究に従事．現在，NTT 情報通信網研究所知識処理研究部主幹研究員．IEEE，電子情報通信学会，人工知能学会各会員．

石川 勉
1949 年生．1967 年電気通信大学電気通信学部応用電子科卒業．同年日本電信電話公社武蔵野電気通信研究所入所．以来，主記憶装置，高信頼化技術，フルウェーハシステム，並列プロセッサ，知識処理技術の研究に従事．工学博士．現在，NTT 情報通信網研究所知識処理研究部グループリーダ．IEEE，人工知能学会，電子情報通信学会各会員．

河岡　司（正会員）

1943 年生．1966 年大阪大学工学部通信科卒業．1968年同大学院修士課程修了．同年日本電信電話公社入社．オペレーティング・システムの開発，ネットワークアーキテクチャの研究，人工知能の研究に従事．工学博士．現在，NTT コミュニケーション科学研究所所長．電子情報通信学会，人工知能学会，IEEE 各会員．