国語辞書の意味解析情報を用いた概念ベースの多義性解消

寺田 宗紘　　森 亮介　　渡部 広一　　河岡 司
同志社大学大学院　工学研究科
dtd0743@mail4.doshisha.ac.jp, {watabe, kawaoka}@indy.doshisha.ac.jp

1　はじめに

コンピュータに人間に近い知的な判断を行わせるためには，コンピュータに高度な連想機能を持たせる必要がある．これは後述する概念ベースのように，辞書やWebのような媒体から語の知識を自動的に取得することで解決する．しかしながら，単純な語の羅列を与えても，コンピュータに知的な判断をさせるのは不可能である．なぜなら語は多義性を有しているからである．現在の概念ベースは，語の多義性を考慮して製作されていないため，多義性を有する語に関して知的な連想を行うことが出来ない．本稿では，国語辞書の意味解析情報を用いて多義語知識ベースと呼ぶ知識ベースを作成し，それを用いて概念ベース上の概念および属性の多義性を解消する方法について提案する．

2　概念ベースとその問題点

2.1　概念ベースと関連度

常識判断メカニズムにおいて概念ベースは中核をなすものである（図 1）．

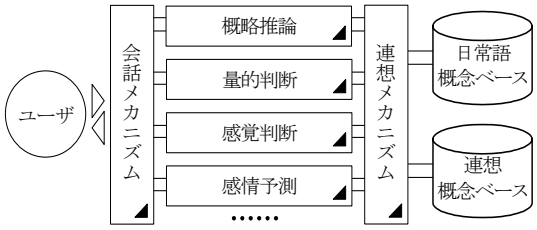

図 1　常識判断メカニズムの全体像

概念ベースは，電子化された複数の辞書から機械的に構築された知識ベースである[1,3]．概念ベースでは，概念を属性（関係のある概念）の集合で定義している．概念ベース内の概念数は約 9 万個，1 概念あたりの平均属性数は約 29 個である．概念 A は属性 a_i と，概念に対する属性の重要度を表す重み w_i の対で表されている．概念 A は（1）式のように表せる．

$$A=\{(a_1,w_1),(a_2,w_2),\ \cdots,(a_i,w_i)\} \qquad (1)$$

ここで，属性 a_i を概念 A の一次属性と呼ぶ．また，属性 a_i も概念であるので，a_i からも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．このように概念から高次の属性を展開することができる．

関連度とは概念間の関係の深さを定量化した 0 から 1 までの値で，関係が深いほど大きな値となる[2]．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導かれる一致度を用いて計算される．二次属性までの属性の一致度が高ければ関連度が高いと言える．よって関連度を用いて，概念ベースの精度を測ることが出来る．

2.2　概念ベースの問題点

概念ベースには不適切な属性（雑音と呼ぶ）が多く存在する．これは機械構築のための雑音と多義のための雑音がある．機械構築のための雑音は各種精錬処理によって精錬されている[3]が，多義のための雑音は従来の精錬によっては解決しない．図 2 に概念“バス”の概念ベース上での例を示した．この例では一つの概念“バス”の中に，複数の意味での“バス”の属性が含まれている．これが，表記から属性を構築した概念ベースの問題点である．また属性に雑音が存在すると関連度の精度も悪くなってしまう．本研究では，多義語知識ベースを利用して，概念の多義性判断を行った．

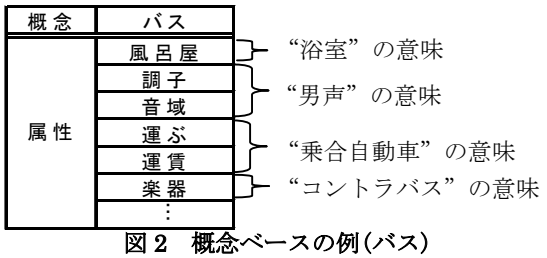

図 2　概念ベースの例（バス）

3　語義について

3.1　多義語の意味特定

語義は文脈に応じて決定される．人間ならば文脈より，語義を推定することが可能である．「バスが走る」という文章が与えられた場合，“走る”と言う動詞を用いて多義語である“バス”の意味を特定することが可能である．この場合，“走る”という動詞が“バス”の意味を特定するための手がかりになっていると言える．この“走る”のように多義語を特定するための手がかりとなる語を手がかり語と定義する．

3.2　語の置換

多義語は一義の同義語に置換することが可能である．例えば，「バスに乗る」という文章があるとする．この文章の“バス”を“バス”の同義語である“乗合自動車”に置換すると，「乗合自動車に乗る」という文章になる．置換前の文章と置換後の文章の意味は変化しない．しかし，“バス”には同義語として“浴室”も存在する．前述の文章の“バス”を“浴室”と置換すると「浴室に乗る」という文章になる．この文章は明らかに間違いである．このことから多義語の同義語は，その多義語の一部の意味分類に関してのみ，同義関係を持つと言うことが出来る．本研究では多義語のこの性質を利用し，多義の概念を一義の同義の概念（代表語）へと置換することにより語の多義性を解消した．図 3 に概念の多義性解消の流れを示す．詳細については 4.1.2 節で述べる．

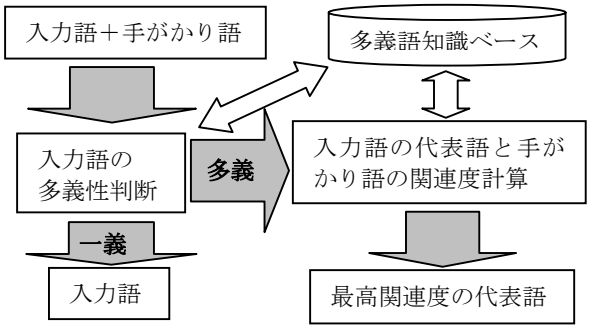

図 3　概念の多義性解消の流れ

4　概念ベースの多義性解消

概念ベースの多義性解消の手順は，大きく二段階に分かれる．第一段階では概念の多義性を解消する．この段階では概念構造情報[4]を用いて多義概念の意味を定義し，意味毎に適切な一義の概念を割り当てる．この操作によって作られる知識ベースを多義語知識ベースと呼ぶ．概念の多義性解消はこの多義語知識ベースを参照することによって行う（図 3）．第二段階では属性の多義性を解消する．この段階では，多義語知識ベースを用いて一義概念の多義属性を一義の概念に置換することによって，属性の曖昧性を解消する．

4.1　概念の多義性解消

4.1.1　多義語知識ベースの作成

4.1.1.1　概念構造情報

本研究では，多義語知識ベースの作成に概念構造情報を利用した．概念構造情報は，概念ベースと同じく辞書から自動構築された知識ベースであるが，概念ベースとは異なり，意味毎の詳細な分類情報を保存している[4]（図 4）．概念構造情報の見出し語数は約 16 万，1 見出し語あたりの平均関連語数は約 6 となっている．

意味番号	（関連語，論理関係）
1	（男声，不明）
2	（コントラバス，同義）
3	（チューバ，不明）（バスクラリネット，不明）

意味番号	（関連語，論理関係）
1	（風呂，同義）（浴室，同義）

意味番号	（関連語，論理関係）
1	（乗合自動車，同義）（自動車，上位）（大型，不明）

図 4　元の概念構造情報の例　“バス”（見出し語）

4.1.1.2　概念ベースと概念構造情報の対応

まず，概念構造情報の見出し語および関連語が，概念ベース上の概念として存在するかを調べる．概念ベースに収録されていない見出し語および関連語は，多義語知識ベースへの格納対象から除外する．次に，概念ベースの概念と概念構造情報の見出し語が 1 対 1 対応になるようにする．概念構造情報では同音異義語が別の見出し語として登録されている場合がある．例えば“バス”は概念ベースには概念として一つしか登録されていない（図 2）が，概念構造情報には見出し語として三つ登録されている（図 4）．よって三つの見出し語“バス”を一つに統合することによってこの問題を解決する（表 1）．

表 1　概念構造情報の例　“バス”（見出し語）

意味番号	（関連語，論理関係）
1	（男声，不明）
2	（コントラバス，同義）
3	（チューバ，不明）（バスクラリネット，不明）
4	（風呂，同義）（浴室，同義）
5	（乗合自動車，同義）（自動車，上位）（大型，不明）

4.1.1.3　多義語知識ベースの生成

多義語知識ベースは，概念構造情報から意味番号が一つしかない一義の見出し語を除いた上で，残った多義の見出し語について，意味分類ごとにその意味を代表する語を割り当てる．この語を代表語と定義する．代表語は以下の三つの手順で選定した．

(1) 概念構造情報から意味分類ごとに一義の同義または類義の語（既存概念と定義する）を抜き出す．

(2)（1）の条件に合う語が見つからない場合は，その意味分類の関連語を属性として持つ新しい概念（新概念と定義する）を作成する．

(3) 関連度計算を用いて意味的に近い分類を統合する．

まず（1）と（2）の操作を行う．（1）の操作において，一義か多義かの判定は概念構造情報に存在するかどうかで決める．また，代表語候補が複数ある場合は，見出し語と最高関連度の語を代表語に選定する．（2）の操作によって得られた新概念の概念名として，概念構造情報の見出し語の後に半角数字を昇順に割り当てた．なお，現行概念ベースの仕様に沿って，新概念の属性には新概念自身も追加した．新概念の属性の重みは，現行概念ベース内の属性の *idf* を算出し，それを重みとして使用した[5]．*idf* が算出できない新概念自身については，その新概念内で最も重みの大きい属性の値を新概念自身の重みとしている．（1）と（2）の操作によって得られた多義語知識ベースの例を表 2 に示す．“バス 1”は“男声”を属性として持つ新概念であり，“バス 2”は“チューバ”と“バスクラリネット”を属性として持つ新概念である（表 1）．

表 2　多義語知識ベースの例　“バス”（見出し語）

意味番号	代表語
1	バス1
2	コントラバス
3	バス2
4	浴室
5	乗合自動車

（1）と（2）の操作終了後に，（3）の操作を行う．この操作では高関連度の代表語同士を統合することによって意味の重複を解消する．ここで関連度の閾値を設定し，閾値以上の代表語同士を統合することとする．閾値の設定は評価の段階で行う．

統合先に関するルールについて述べる．統合する代表語集合に既存概念が存在する場合，既存概念を統合した意味における代表語に選定する．既存概念が複数ある場合，見出し語と最高関連度の既存概念を代表語に選定する．統合する代表語が全て新概念の場合，全

ての新概念の属性を統合したものを属性として持つ新しい概念（新概念）を代表語とする．表 3 は代表語“コントラバス”と“バス 2”が統合され，統合された意味の代表語が“コントラバス”となる場合の例である．

表 3　多義語知識ベースの例　“バス”（見出し語）

意味番号	代表語
1	バス1
2	コントラバス
3	浴室
4	乗合自動車

4.1.2　概念の多義性解消の流れ

概念の多義性解消は，多義語知識ベースを参照して行う．多義の概念（見出し語）を代表語（一義の概念）に置換することによって，概念の多義性を解消する．複数の代表語（意味）から 1 つの代表語（意味）を選ぶ方法は，多義概念の意味を特定するための手がかり語と，多義概念の全ての代表語との関連度を取り，最高関連度の代表語に置換することとする(図 3)．なお，一義か多義かの判定は多義語知識ベースに見出し語として存在するかどうかで判定する．

4.2　属性の多義性解消

概念の多義性は解消されたが，属性の多義性はまだ解消されていない．そこで多義性解消の第二段階として，一義概念の多義属性を一義概念（代表語）に置換することによって，属性の曖昧性を解消する．図 5 は一義の概念“タクシー”の多義属性“バス”が一義の代表語“乗合自動車”に置換されたときの例である．

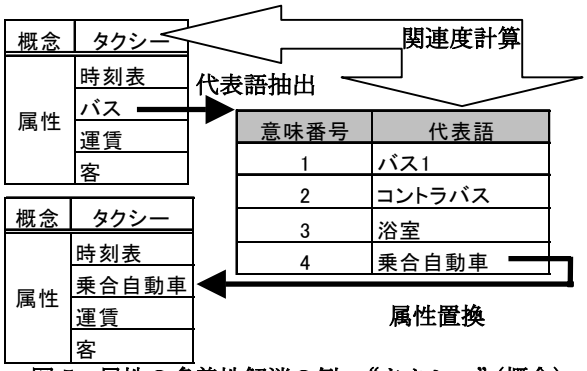

図 5　属性の多義性解消の例　“タクシー”（概念）

5　評価と考察

5.1　評価方法

提案手法を用いて多義性解消を行った概念ベースを評価する．表 4 に示すような評価データを人手により 200 セット作成し，基準概念と 4 つの対象概念との関連度を取り，それぞれについて 200 語の平均関連度を求め，改良前後の概念ベースで比較する方法により評価を行う．なお，この評価セットでは，基準概念の意味特定を行うための手がかり語は関連名詞とする．

表 4　評価データの例

基準概念	対象概念			
	関連名詞	関連動詞	低関連名詞	低関連動詞
バス	自動車	乗る	温泉	洗う
バス	温泉	洗う	自動車	乗る

5.2　意味統合の際に用いる閾値の設定

現段階では，まだ代表語統合の際に使用する関連度閾値を決定していないため，ここで最適な閾値を決定する．閾値の候補を 0.2～1.0 の範囲に限定し，その範囲で刻み幅を 0.05 として，評価結果が最良の閾値を最終的な閾値に決定する．閾値の候補の範囲を 0.2 以上に限定したのは，同じような意味の語は属性が類似していると言えるので関連度が高いと考えるからである．関連度 0.2 以上が高関連度というのは過去の研究で実験的に分かっている．なお，閾値を決める段階では属性の曖昧性解消は行わないものとする．

評価尺度は関連名詞と低関連名詞の平均関連度の差とする．この差が大きいほど，概念ベースの精度が良いと言える．関連名詞と低関連名詞の平均関連度の差を評価尺度としたときの各閾値での結果を図 6 に示す．

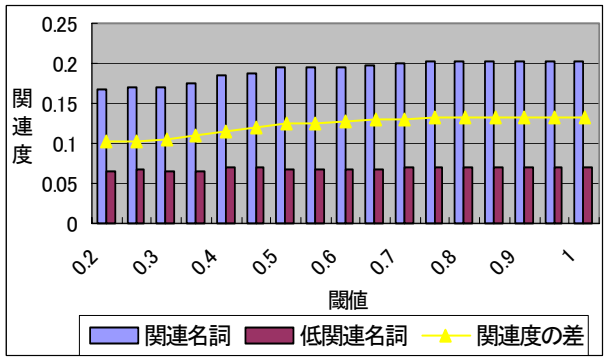

図 6　各閾値での関連名詞と低関連名詞の平均関連度の差

図 6 に示した結果より，関連度閾値 0.5～1.0 の範囲が良好であるが，際立った差が見られないため，他の評価尺度も考慮する．

代表語には既存概念と新概念が存在する．新概念は既存概念と比較して属性の精度が悪い．このことに着目すると，代表語に既存概念が多く含まれているものほど良いと言える．そこで，全代表語に占める既存概念の割合（既存概念率）を，閾値を決める上での 2 つ目の評価尺度とする．この結果を図 7 に示す．

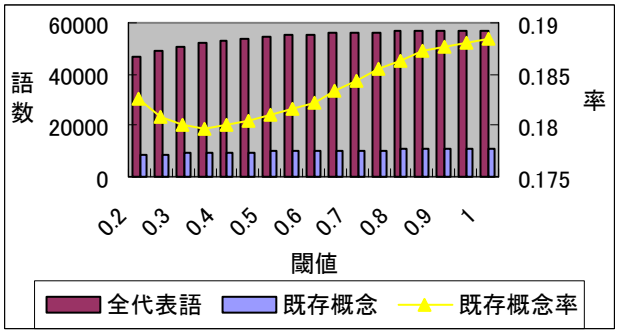

図 7　各閾値での既存概念率

図 7 に示した結果より，閾値 0.35 未満では，閾値が低いほど既存概念率が高く，逆に閾値 0.35 以上では，閾値が高いほど既存概念率が高いことが分かる．閾値が低いほど多くの代表語が統合されるため，既存概念率が高いように思えるが，閾値が 0.35～1.0 の範囲では逆に，閾値が低いほど既存概念率が低いという結果になっている．これは，閾値が低いほど統合する代表

語集合に既存概念が数多く存在すること，つまり，代表語に選ばれなかった既存概念が増加することに起因すると考えられる．さて，問題の閾値だが，最も既存概念率の低い閾値 0.35 と最も既存概念率の高い閾値 1.0 で既存概念率を比較すると，差は 0.01 もなく大きいとは言えない．よって，この評価尺度でも閾値を決めるのは困難である．

代表語統合によって新概念の属性数が増加するものもあるので，新概念の平均属性数を，閾値を決める上での3つ目の評価尺度とする．この結果を図8に示す．

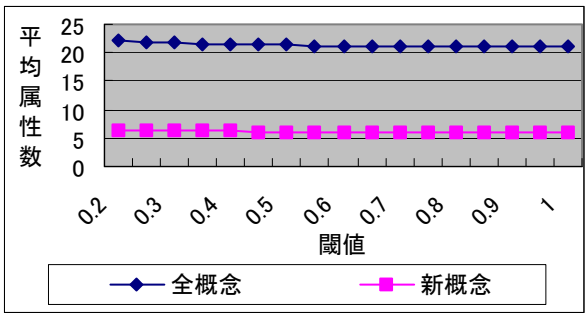

図 8　各閾値での平均属性数

新概念の平均属性数は新概念同士の統合の場合に増加するが，統合する代表語集合に既存概念が含まれている場合，もしその中に比較的属性数の多い新概念が含まれていたら，その新概念がなくなってしまうことによって平均属性数が減ってしまうことも考えられる．図 8 の結果において，あまり属性数に大きな差がないのはこのような理由からだろう．

3 つの評価尺度で閾値を決めようと試みたが，いずれにおいても大きな差がなかったため，最終的な評価と直結している 1 つ目の評価尺度を優先する．1 つ目の評価尺度による結果では，0.5～1.0 の範囲が良好であったので，この範囲で閾値を決定する．閾値が高すぎるとほとんど統合されず，統合する意味があまりないという点と，新概念の数は少ないほうが良いという点に着目して，この範囲で最も低い値の 0.5 を閾値に決定する．表 5 に閾値を 0.5 とした時の概念ベースの詳細な情報を示す．

表 5　閾値 0.5 時の多義性解消概念ベース

	概念数	平均属性数
元の概念	87242	29.1096
新概念	45034	6.12992
全体	132276	21.2861

5.3　評価結果と考察

閾値を 0.5 に設定して評価を行った．評価結果を図 9 に示す．また，関連名詞と低関連名詞の平均関連度の差を図 10 に示す．概念の多義性解消前では，関連名詞と低関連名詞の平均関連度の差はない．この理由は，作成した評価セットには同じ見出し語が 2 セットずつ，関連語と低関連語でクロスした形で収録しているためである（表 4）．一方，多義性解消後ではこの差は大きいという結果から，多義性解消後の概念ベースは多義性解消前と比較して精度が向上したと言える．このことは概念の多義性解消の精度が良いことを示している．しかし，さらに属性置換を行った概念ベースは，属性置換を行わなかった概念ベースと比較して，少し精度が落ちている．この原因は属性置換の際，多義の属性が新概念に置換されてしまったためだと考えられる．したがって属性置換を行う前に，新概念の属性，つまり概念構造情報の関連語の精度を上げる必要がある．

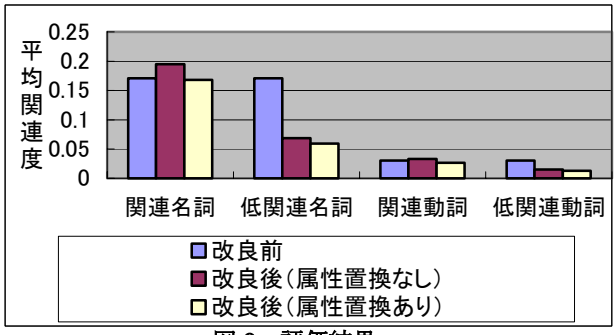

図 9　評価結果

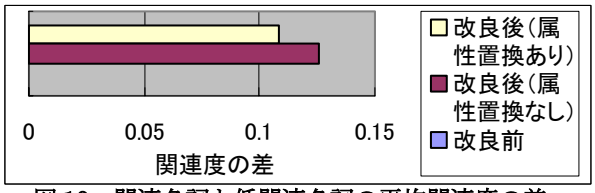

図 10　関連名詞と低関連名詞の平均関連度の差

関係動詞,低関連動詞は名詞との比較のために作成したが，動詞の平均関連度は改良前，改良後ともにかなり低い値となっている．この理由は名詞と比較して動詞は曖昧性の度合いが強いためだと考えられる．

6　おわりに

本稿では，コンピュータに，人間に近い高度な連想機能を持たせることを目的に，概念ベースの多義性を解消する方法について述べた．提案手法を用いて多義性解消を行った概念ベースは，多義性解消前の概念ベースと比べて精度が向上していることを示した．多義性解消概念ベースの精度をさらに向上させるためには，概念構造情報の精度を高める必要があると思われる．

なお，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

[1]笠原要，松澤和光，石川勉，“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，Vol.38，No.7，pp.1272-1283（1997）

[2]井筒大志，渡部広一，河岡司，“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（2002）

[3]広瀬幹規，渡部広一，河岡司，“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錬手法”，信学技報，TL2001-49，pp.109-116（2002）

[4]小島一秀，渡部広一，河岡司，“常識判断のための概念ベース構築法－国語辞書からの抽出した概念間の論理関係の利用”，同志社大学理工学研究報告 Vol.42 No.1，pp1-8，(2001)

[5]坂田光広，渡部広一，河岡司，“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（2004）（印刷中）