国語辞書と新聞を用いた概念ベースの構築と精錬

5X−4

広瀬　幹規　　渡部　広一　　河岡　司
同志社大学大学院　工学研究科

1. はじめに

人間らしい常識的な判断を行う知的判断メカニズムの中核として単語の意味を知識ベース化した概念ベースが存在する. 現在, 機械的に概念ベースを構築する方法として, 電子化された辞書や新聞記事を用いた自動構築が提案されている. 各概念ベースについて, 辞書より構築された概念ベースは概念の定義を表す属性が比較的多く獲得でき, 新聞より構築された概念ベースは日常的概念や属性が獲得できるという性質がある. 知的判断メカニズムをより向上させるためには偏った観点からではなく, 幅広い観点からの属性が構成されるべきである.

本稿では, 辞書から構築された概念ベースと新聞から構築された概念ベースを統合し, 属性の選別を行うことによって意味的な属性に偏らない日常的な概念ベースの構築手法を提案する.

2. 概念ベースの構造

2.1 概念の定義

概念Aはその属性a_iと重みw_iの対の集合で定義される.

$$A=\{(a_1, w_1), (a_2, w_2), \cdots, (a_n, w_n)\}$$

2.2 概念ベース

基本概念ベース(基本CB)：約4万の概念とその属性および重みを複数の国語辞書などの語義文から自動的に獲得した概念ベース[1].

連想概念ベース(連想CB)：基本概念ベースの重みを全て1にし, 属性に概念自身を必ず含むようにした概念ベース.

新聞概念ベース(新聞CB)：約3万件の記事から概念とその属性を自動的に獲得した概念ベース. 属性は共起情報を元に30個選択し, 重みは全て1としている[2].

Construction and refinement of the concept-base using a language dictionary and a newspaper.
Tomonori HIROSE, Hirokazu WATABE, Tsukasa KAWAOKA
Faculty of Engineering, Doshisha University

3. 概念ベースの構築手法

方針 1 ： 連想概念ベースと新聞概念ベースの統合

連想概念ベースと新聞概念ベースにおいて, 共通する概念に関しては属性を単純和することで概念の属性を生成する. 片方の概念ベースにしかない概念は構成されている属性をそのまま概念の属性とする. (構築された連想概念ベースをACB1とする. 属性の重みは全て1)

方針 2 ： 概念ベースの精錬

機械的に構築された概念ベースは非常に粗い情報を持ったものであるため, 概念ベースの質の向上を目的とした機械的な操作が必要となってくる.

本稿では外部から属性の修正が発生した場合でも継続的に精錬処理が可能である精錬手法を提案する. 精錬処理においては, 概念間の関連性を定量化した値(関連度)やシソーラスなどの外部知識ベースを利用する.

4. ルールを用いた自動精錬手法の提案

表1に示した精錬用ルールと概念間の関連の深さを定量化した値である関連度を用いた精錬手法を提案する. なお本稿では関連度計算方式として「重み付き概念連鎖関連度計算方式」[3]を用いた.

まず, ACB1を精錬用ルールの何れかに適合するかどうかで属性を選別し, 連想概念ベースACB2を生成する. (精錬 *RefA*, 属性の重みはすべて1)

次に, 精錬用ルールに概念Aとその属性である概念Bの関連度がある一定値以上でなければならないという条件を加える. さらに属性の中で概念との関連度が突出して高いというルールも追加する. 関連度計算には ACB2を利用する(図1).

また, 上記のルールにおいて, 関連度の条件が厳しい場合の判定を *J1_High*, 条件が比較的緩い場合の判定を *J1_Low* とし, 各判定において属性が適合したルール

数を Nとする．さらに，概念の属性内で概念との関連度順 n 番目以内かつ関連度が w 以上という判定を $J2$ とする．以上の判定を利用して図 2 に示すような精錬処理 *Ref B* で連想概念ベースACB3を生成する．

表 1．概念 A 内のある属性(概念 B)に対する精錬用ルール

ルール名	ルール内容
相互リンク	概念 Bの属性に概念 Aを含んでいる
表記特徴	概念 Aの表記と概念 Bの表記に同一の漢字が含まれる
シソーラス	概念 Aと概念 Bが上位・下位・仲間の関係である
関係データ	概念 Aと概念 Bが辞書解析結果(文献[4])より関係が明確である

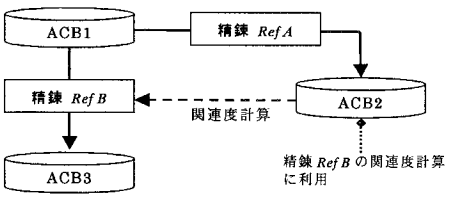

図1．精錬 *Ref B* の各概念ベースの関係

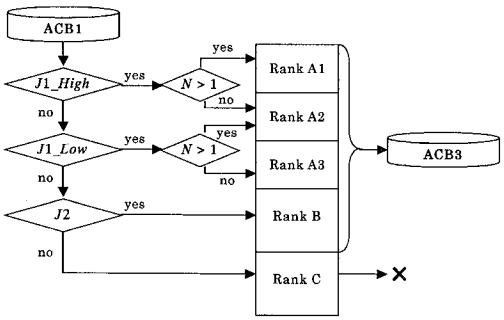

図2．精錬 *Ref B*

5．　選別された属性の重みの検証

ルールによる精錬手法により属性が A1,A2, A3, B の 4 つのランクに選別されるが，4 つにランクされた属性が属性内でどれだけの比重を持てば良いかは検証されていない．ここでは，属性には概念自身も含まれているため，4 つのランクと概念自身を含んだ 5 つのランクがどれだけの比重を持てばよいか検証する．検証には評価尺度を用いる．(評価尺度とは基準概念 Mx, 同義・類義の概念 Ma, 関係のある概念 Mb, 関係のない概念 Mc で構成され，Mx との関連度が Ma が最大で Mc が最小になる時，正しく解釈されたとする)．また，適切な重みの傾向が得られても，得られた結果が評価尺度によって変動する可能性は否めない．よって傾向の信頼性を確認するために Mda1(200 セット)，Mda2(200 セット)，Mda3(190 セット)の 3 セットの評価尺度において評価を行った．検証によって，次のような重みを得た(概念自身，A1,A2, A3, B の重みをそれぞれ w_5, w_4, w_3, w_2, w_1とする)．

(w_5, w_4, w_3, w_2, w_1)=(1.0, 0.68, 0.57, 0.31, 0.07)

6．　評価

今回の処理により構築された ACB3は各概念ベースよりも評価尺度(590 セット)の解釈成功率において最もよい結果を得ることができた．

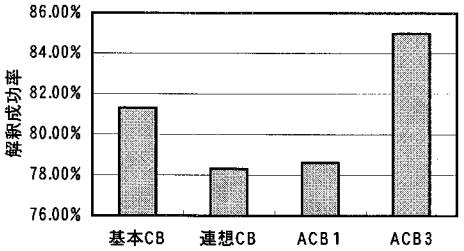

図 3．各概念ベースの評価尺度での解釈成功率

7．　おわりに

本稿では国語辞書と新聞を用いて概念ベースを構築し，精錬することによって，意味的な属性に偏らない日常的な概念ベースを構築することができた(図 3)．

今後は構築された概念ベースの質をより高める精錬手法，新規属性の追加手法について検討していきたい．

謝辞

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]橋本　隆志，渡部　広一，河岡　司:新聞記事による概念ベースの自動構築，情報処理学会第 61 回全国大会講演論文集，分冊 2，pp.87-88(2000)

[3]渡部　広一，河岡　司:常識判断のための概念間の関連度評価モデル，自然言語処理，Vol.8，No.2，pp.39-54(2001)

[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，電子情報通信学会信学技報，Vol.99，No.99，pp.45-52(2000)