4N-5

シソーラスの分類情報を利用した概念ベースの属性追加手法

北川　晋也†　　奥村　紀之‡　　渡部　広一‡　　河岡　司‡

†同志社大学工学部知識工学科　‡同志社大学院工学研究科知識工学専攻

1. はじめに

本研究では，自然言語の意味理解をするために，類似度も含めた，一般的な概念間の関連度を利用し，柔軟な連想メカニズムの実現を目指している．連想メカニズムは概念ベースと概念間の強さを定量化する関連度計算により構成している．概念ベースは機械的に構築しているため，不適切な属性を含み，正しく特徴付けられていない概念が存在する．そのため本稿では，概念との関連性が明確である属性をシソーラスより取得し，より適切に概念を特徴付け，概念ベースの質の向上を図る方式を提案する．

2. 概念ベース

2.1. 概念ベースの定義

概念ベースは，電子辞書等から機械的に構築された約 9 万語の大規模知識ベースである．概念 A は，その語と関連の強いと考えられる語(属性) a_i と重み w_i の対の集合として定義する．

$$\text{概念 } A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (1)$$

概念自身の属性を 1 次属性と呼び，任意の 1 次属性 a_i は，必ず概念ベース内に含まれる語で定義する．つまり，属性を表す語もまた概念として定義されている．従って，1 次属性それぞれの属性を導くことができる．これを，2 次属性と呼ぶ．同様に導くことで，概念は任意の次元までの属性連鎖集合により定義されている．

2.2. 概念ベース *idf* [2]

概念ベース上での情報価値を表すものとして，概念ベースを仮想文章群として新たな *idf* を算出し，これを次式で定義する．

$$idf_{cb(m)}=\log\frac{V_{ALL}}{V(a_i)} \quad (2)$$

V_{ALL} は概念ベースに定義される全概念数，$V(a_i)$ は概念 a_i を m 次属性内に持つ概念数である．

3. 関連度計算方式[2]

概念ベースを利用した概念と概念の間にある関連性を定量的に評価する手法として，関連度計算方式を利用している．

4. 属性数による影響

関連度計算方式では，演算に使用する属性数を重み上位 30 個に限定してあり，概念の持つ属性が 30 個以上あるとき，より正確に関連度が算出される．しかし，30 個以下の属性しか持たない概念の割合は，65%存在する．そのため，概念の持つ属性数を 30 個以上にする必要がある．

5. 概念ベースの評価方法

4 つの概念の組を用意する．ここで，概念 X は任意の概念(基準概念)であり，概念 A は概念 X と極めて密な概念，概念 B は概念 X に密な概念，概念 C は概念 X に疎な概念である．例えば概念 X が「海」であるとき，概念 A は「海洋」，概念 B は「塩」，概念 C は「車」のようになる．

概念 X と概念 A との関連度を $\mathrm{Re}l(X,A)$ としたとき，各評価用データの組に関して

$$\mathrm{Re}l(X,A)-\mathrm{Re}l(X,B)>Ave\,\mathrm{Re}l(X,C) \quad (3)$$

$$\mathrm{Re}l(X,B)-\mathrm{Re}l(X,C)>Ave\,\mathrm{Re}l(X,C) \quad (4)$$

$$Ave\,\mathrm{Re}l(X,C)=\sum_{i=1}^{n}\mathrm{Re}l(X_i,C_i)/n \quad (5)$$

を満たせば，その組を正解とする． n は評価セット数である．このような評価用データ 1780 組で，何組のデータが正解したかによって概念ベースの精度評価を行う．

6. シソーラス

本稿では，日本語語彙体系[1]から作成された一般名詞の意味的用法を表す 2710 個のノードと約 13 万語のリーフからなるものを使用する．概念ベースに定義されている概念の内，約 5 万個が

Attribute Addition Method of The Concept Base using Classification Information on Thesaurus

†Shinya Kitagawa
Knowledge Engineering and Computer Sciences, Doshisha University

‡Noriyuki Okumura, Hirokazu Watabe, Tsukasa Kawaoka
Graduate school of Engineering, Doshisha University

シソーラス内にも存在する．

7. 属性の追加方法

シソーラスは人手によって作成されているため，機械的に構築した概念ベースとは異なり，語と語の関係に信頼性がある．従って，各概念に対して，シソーラスから属性を追加することは有効であると考える．以下，シソーラスを用いて属性を追加する手法について述べる．

7.1. 属性の重み付け手法[2]

シソーラスで関係付けられている語を概念ベースに追加する場合の重みは，関連度×$\sqrt{idf}_{cb(3)}$による重み付け手法を用いる．

7.2. 追加属性の選択方法

シソーラスを用いた属性の拡張では，各概念から見て，3 個上までの親にあたる語，子にあたる語，並びに，共通の親を持つ兄弟にあたる語の 3 通りを考え，属性に追加する．

まず，親にあたる語は，その概念をより抽象的に表したものであり，概念を特徴付けることができる．また，親にあたる語は最大 38 個，平均 4.0 個取得できる．平均が 4.0 個と少なく，目視評価により，3 個上までの親に関連があると見られたため，3 個上までの親全てを属性として追加する．

次に，子にあたる語は，その概念をより特徴として持つもの，兄弟にあたる語はその概念により近いものが存在する．子にあたる語は最大 639 個，平均 36.9 個，兄弟にあたる語は最大 1423 個，平均 152.9 個取得されてしまう．そのため，全て追加すると関連度計算に使用する 30 個の属性が全てシソーラスに存在する語に置き換えられるという問題が考えられる．そこで，子，兄弟にあたる語を属性追加候補語として取得し，前節の手法で重み付けを行った後，重みの上位より追加する個数に上限を設定し，属性を追加した．今回は，子，兄弟にあたる語の上限をそれぞれ 0 個から 30 個まで 5 個刻みにし，その全ての組み合わせを調べた．

また，複数のノードのリーフとなっている概念が多数存在する．その場合，共通の親を持つ語ごとに均等に概念ベースに追加を行った．

8. 評価結果

属性を追加した概念ベースに対して，5 章で述べた方法を用いて評価を行った．

図 1 に追加後の精度上位 5 位を示す．ただし，グラフの軸項目については，X_Yと表記し，子にあたる語をX個，兄弟にあたる語をY個，親にあたる語を全て追加したことを表している．

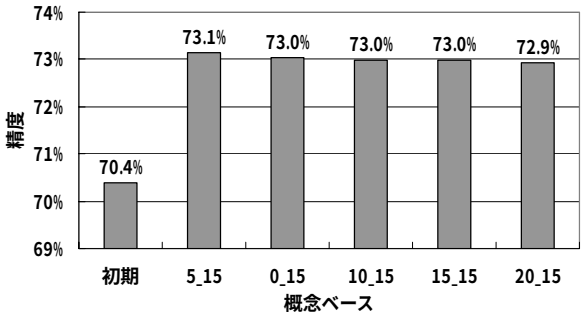

図 1：評価結果

図 1 より，3 個上までの親にあたる語に加え，子にあたる語上位 5 個，兄弟にあたる語上位 15 個を追加した場合，元の概念ベースの精度 70.4%に対して 73.1%となり 2.7%の精度向上が見られた．また，全属性数は約 328 万，平均属性数は約 38 個となり，関連度計算に使用される 30 個の属性を持たない概念の割合は，65%から 51%と改善した．

9. おわりに

人手により作成され，人間の観点から見て関係の確かなシソーラスの親，子，兄弟を属性として機械的に付加することにより，電子化辞書/新聞から構築した概念ベースの属性を拡張した．この手法により，概念をより正しく特徴付け，概念ベースの質向上を図ることができることを実験により示した．

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]池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦(編)．日本語語彙体系．岩波書店(1997).

[2]奥村紀之，小島一秀，渡部広一，河岡司．電子化新聞を用いた概念ベースの拡張と属性重み付与方式．情報処理学会研究報告，2005-nl-166，55-62(2005).