5P-2

語の連鎖構造と相関に基づく概念ベースの構築

豊嶋 章宏 † 奥村 紀之 ‡
†,‡ 香川高等専門学校 情報工学科

1 はじめに

近年の情報化社会の発達に伴い, コンピュータと人間の円滑なコミュニケーションが1つの課題となっている．コンピュータが人間と同様の振る舞いを行うためには連想体系を持たせる必要がある．これを実現するための中核機構として概念ベースが用いられる．本論文では，概念ベースの機械的構築手法の提案し，関連度計算を用いた概念ベースの評価を行う．

2 概念ベース

コンピュータに柔和な連想体系を持たせるために概念ベースは用いられる．概念ベース [1] は，電子国語辞書等より機械的に構築される．辞書の見出し語を概念表記とし，各見出し語に対応した説明文に形態素解析を行い，自立語を属性 a_i として抽出する．さらにそれぞれの属性に概念に対してどの程度の価値があるかを示した重み w_i を付与する．

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

本研究では，基本概念ベースを構築する際にテキストコーパスとしてEDR電子化辞書を，形態素解析器としてMeCabをそれぞれ使用した．このとき，EDR電子化辞書の単語をユーザ辞書として登録した．このようにして構築した基本概念ベースの規模は総概念数が170,499，平均属性数が約3.89となった．

また，本研究では概念ベースにおける重み付けの手法として，*tf·idf* を用いる．*tf·idf* は，ある文書における単語の出現頻度と，文書集合における単語の特定性に基づいた重み付け手法である．概念 A における単語 t の重みは以下の式で定義される．

$$w_t^A = tf_A(t) \cdot idf(t) \quad (2)$$

概念 A における $tf_A(t)$ は，辞書コーパスにおける見出し語Aに対応した説明文中に存在する t の出現頻度である．また，$idf(t)$ は，概念ベース全体における概念あたりの逆出現頻度である. これらの値の積を重みとして属性にそれぞれ付与する．

「A Construction of Concept-base based on Concept-Chain Model and Correlation」
† 「Akihiro TOYOSHIMA」
†† 「Noriyuki OkKUMIRA」
Kagawa National College of Technology, Department of Information Engineering (‡‡)

3 概念ベースの構築手法

現在，概念ベースにおける様々な構築手法が提案されている．しかし，従来手法は概念ベースの構築を行う際に人手による評価が必要であるため，再現性の観点で問題がある. 本節では，概念ベースにおける語の連鎖構造や相関を用いた再現性の高い機械的構築手法を提案する．

3.1 連鎖構造を用いた概念ベースの構築

電子化辞書等を基に機械的に構築した基本概念ベースにおいて，見出し語に対応した説明文中に存在する自立語群を1次属性とする．基本概念ベースを構築する際に，見出し語に存在する語のみを属性として使用したため，1次属性を概念として参照することで2次属性を連鎖的に抽出することができる．そのため概念ベースはN次の連鎖集合として定義される．連鎖構造より抽出できる属性を本研究では連鎖属性と定義する．本項では，連鎖属性の抽出手法について説明する．次数 α における概念 A_α が次式であるとする．

$$A_\alpha = \{(a_{\alpha 1}, w_{\alpha 1}), (a_{\alpha 2}, w_{\alpha 2}), \cdots, (a_{\alpha i}, w_{\alpha i})\} \quad (3)$$

このとき，次数 $\alpha+1$ における概念 $A_{\alpha+1}$ は次式で表される．

$$\begin{aligned} A_{\alpha+1} &= \{(a_{(\alpha+1)1}, w_{(\alpha+1)1}), (a_{(\alpha+1)2}, w_{(\alpha+1)2}) \\ &\quad , \cdots, (a_{(\alpha+1)j}, w_{(\alpha+1)j})\} \end{aligned} \quad (4)$$

連鎖属性を構築する際に，属性 $a_{(\alpha+1)j}$ がn個抽出されたとき, 重み $w_{(\alpha+1)j}$ は

$$w_{(\alpha+1)j} = \sum_{k=1}^{n} w_k \cdot tf(\alpha) \quad (5)$$

となる．$tf(\alpha)$ は連鎖を構築する際に元となった属性の重みとなる．

3.2 相関を用いた概念ベースの構築

連鎖構造を用いた概念ベースの構築では抽出できない属性を，テキストマイニングを用いた語の相関情報より抽出する． 本研究では，テキストマイニングシステムとしてIBM Content Analytics(ICA) を利用する．ICAを用いてEDR電子化辞書を解析し，見出し語をファセットとして与えることで，文書間で共起している語を取得し概念ベースを構築する．

4　評価実験

本項では，EDR 辞書を MeCab により解析して構築した基本概念ベース (UCB) と ICA による概念ベース (ICACB) の評価を行う．概念ベースの評価手法として関連度計算方式に基づく評価手法を用いる．

4.1　連鎖属性の収束性の評価

UCB と ICACB に対し，連鎖属性の構築をそれぞれ行う．図 1 に，形態素解析により構築した基本概念ベースを連鎖構造を用いて展開したときの 1 概念あたりの平均属性数の推移を示す．図 1 より UCB は 20 次，ICACB は 22 次で収束していることがわかる．

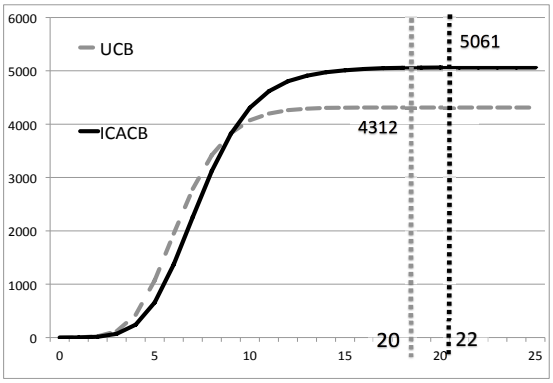

図 1: 形態素解析による概念ベースの平均属性数の推移

4.2　関連度計算方式を用いた評価

関連度計算は，概念ベースを利用して概念間の関連の強さを定量化する手法である．本研究では，重み比率付き関連度計算アルゴリズム [2] を用いて評価を行う．関連度計算を行うための評価用データとして，奥村ら [1] が使用した評価セットの中から EDR 電子化辞書の単語して定義されているものを抽出し，224 セット用意した．評価用データの例を表 1 に示す．

表 1: 評価用データ例

X	A	B	C
自転車	二輪車	ペダル	日本国憲法
迷い	惑う	迷子	父兄

概念 X は任意の対象概念となり，概念 A は概念 X に類似または高関連の概念，概念 B は概念 X にある程度関連が認められる概念，概念 C は概念 X に対して無関連な概念となっている．r_A を概念 X と概念 A の関連度，r_b を概念 X と概念 B の関連度，r_c を概念 X と概念 C の関連度とし，このとき

$$(r_A > r_B) \wedge (r_B > r_C) \tag{6}$$

を満たす場合，関連度評価は成功，それ以外は失敗として関連度を求める．以下に形態素解析による概念ベースの 1 次，2 次，3 次の連鎖属性，ICA による概念ベースの 1 次属性，それぞれの 1 次属性を合わせた複合概念ベースの 1 次属性，奥村ら [1] が構築した概念ベースの関連度評価結果を示す．

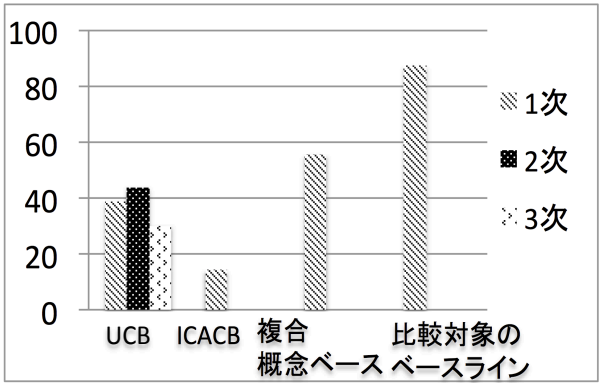

図 2: 関連度評価結果

図 2 より，UCB は 2 次まで展開すると精度が良くなるが，3 次まで展開すると精度が悪いことが分かる．連鎖属性の展開によりノイズとなる属性が増加することが要因であると考えられる．また，それぞれ個別の概念ベースの精度に比べ，UCB と ICACB を組み合わせた複合概念ベースの精度が高いことがわかる．連鎖属性を構築する際に，概念にとって重要な属性を抽出し精度が向上するよう重み付けを考える必要がある．

5　おわりに

本稿では，概念ベースの機械的な構築手法として語の連鎖構造に基づく構築手法と，テキストマイニングによる語の相関を用いた構築手法について述べた．評価実験により，二つの手法を組み合わせることでより質の良い概念ベースが構築できることが分かった．今後の課題としてより規模の大きい概念ベースの評価を行うことや連鎖属性を構築する際の重み付け手法について考える必要がある．

参考文献

[1] 奥村紀之, 渡部広一, 河岡司,(2007). 概念間の関連度計算のための大規模概念ベースの構築. 自然言語処理,Vol.4,No.5,pp41-64

[2] 井筒大志, 東村貴裕, 渡部広一, 河岡司,(2002). 概念ベースを用いた関連度計算方式の精度評価, 信学技報,PRMU2001-259,pp117-112