1Q-3

概念ベースにおける属性傾向と属性集合の評価

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

コンピュータが人間と円滑にコミュニケーションを行うためには, コンピュータに人間と同じような連想体系(概念ベース)を持たせる必要がある. 概念ベースは, 概念とその概念を構成する概念(属性)から構成されている知識ベースである. 概念ベースを n 次元の連鎖集合として定義することができる. 本論文では, 概念ベースの連鎖集合の深さを機械的に解析して, どの程度の次元数で概念ベースが収束していくのかを検討し, 各次元における属性集合の特徴の評価について述べる.

2.概念ベース

コンピュータと人間が円滑にコミュニケーションを行うためには,人間と同じような連想体系を持たせる必要がある.概念ベース[1]は,電子化された国語辞書などを機械的に解析して構築する.辞書の見出し語を概念表記とし,各見出し語に付与されている説明文を形態素解析することによって得られる自立語をその概念を表す属性 a_i とする. さらにそれら属性がその概念にとってどの程度価値のあるものかを表す重み w_i の対の集合として定義する(式 1).

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

本研究では, 特に属性の収集手法に関して報告を行う,

3.属性連鎖

概念ベースにおいて, 概念に付与されている属性を 1 次属性とすると, 1 次属性からさらに 2 次属性を導くことができる. 概念ベースに登録されている概念は有限であるので, これを繰り返していくことで属性の連鎖は n 次元で収束すると思われる.

本研究では, 概念ベースにおける属性の収束性と各次元において抽出できる連鎖属性の有用性についてそれぞれ評価していく.

属性連鎖の評価を行うにあたり, 概念ベースを EDR 概念辞書[2]と MeCab[3]から構築する. EDR 概念辞書は概念見出しと概念説明文より構成されている. 日本語概念見出しを概念とし, 概念説明文を MeCab で用いて形態素解析行うことで, 1 次属性を抽出し概念ベースを構築した. この際に, 概念表記として定義されているもののみを属性とした. また同じ概念見出しが複数存在するものに関しては, 別の概念として辞書登録していることに意図があると判断し, それぞれ別の概念として登録した.

このようにして, 総概念数 221025, 概念あたりの平均属性数が約 3.32 の概念ベースを構築した.

4.評価実験

概念ベースに登録されている 221025 の概念のうち, 評価セットとして任意で 100 個の概念を選出し, 属性連鎖の評価を行う. 選出した評価リスト例を表 1 に示す. 属性連鎖の収束と属性の精度を評価し, 連鎖集合によって適切な属性が取得されるかについて考察する.

表 1 評価セット例

働く	トマト	椅子	うどん
魚	遅れる	スプーン	教える
読書する	雪	桜	訪れる
泣く	滑る	噛む	踊る
夏休み	スコップ	商業	タバコ

各次元における属性は, 連鎖属性と前の次元の属性が付与されるため, 連鎖が深くなる程, 属性数は増加していく. 連鎖属性の取得を属性数が収束するまで繰り返し, 収束性について調査する. また, 取得された属性の内, 関連属性として正しい属性の割合を属性精度と定義し, 属性精度の評価を行う. 属性精度の評価の際には, 概念の持つ属性が概念のみの場合には, 常に精度が 1 となるため, 93 の概念に着目し評価を行った.

評価セットに対し, 属性数が収束するまでの各次元の平均属性数と, 各次元の属性に対して属性精度の推移を図 1 に示す. 評価セットにおける平均属性数は, 連鎖が深くなるにつれ増加していき, 33 次属性において収束している. 属性精度について見てみると, 次元が深くなる程減少しており, 17, 18 次元程度で収束することがわかる.平均属性数も同様の次元で収束しはじめているの

「An Evaluation of Attributes Direction and Set for Each Concept in Concept-base」
†「Akihiro TOYOSHIMA」
†「Noriyuki OKUMURA」
Kagawa National College of Technology, Department of Information Engineering

で，収束するまで展開した場合と,この時点まで展開した場合の取得される属性の中で,正しいと思われる属性はたいして変化しない.

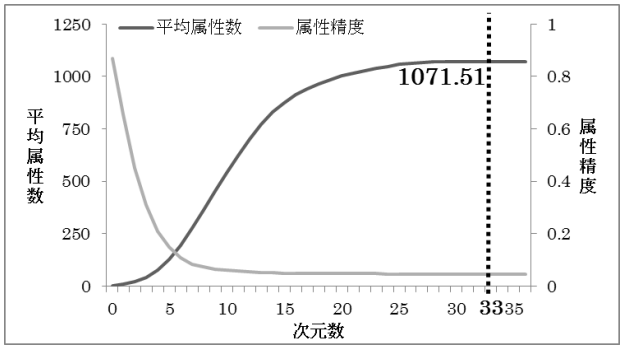

図 1 平均属性数の推移

次に,解属性数の推移を図 2 に示す,属性精度や平均属性数と同様に 17, 18 次元程で収束していることがわかる.このとき,正しいと思われる属性の数は 33.76 であるが,概念ベースにおける属性数の検討の研究[4]により,概念ベースに持たせる属性数として 30 が適切であることが実証されている.故に,連鎖属性として 17, 18 次程度が適切であることが属性数の観点からもわかる.

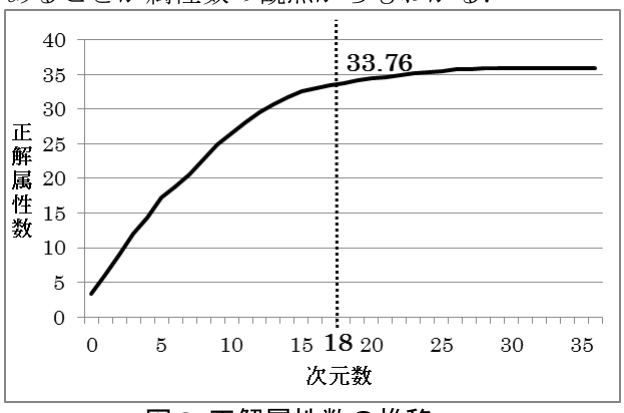

図 2 正解属性数の推移

5.相関属性

属性の精度概念ベースの連鎖属性を展開することで, 概念ベースにとって有用な属性を精錬することができた. しかし連鎖を辿るだけでは, 概念が取得する属性は意味的なものが多くなり, 連想的な単語が少なくなってしまう, 概念 1 つ 1 つを特徴付けるために十分な属性を取得するのは不可能である. そこでテキストマイニングシステムである IBM Content Analytics[3]に EDR 概念辞書のデータを入力し, 単語間の相関関係より属性を取得させた. ここではトマトという概念に対する 33 次の連鎖属性と相関関係から取得した属性の 1 例を表 2 と表 3 に示す.

表 2 属性連鎖によるトマトの属性

果実	食べる	実	物	料理
味	植物	食用	トマト	食べる

表 3 相関関係によるトマトの属性

ペースト	レタス	ソース
トマト	さくらんぼ	はさむ
チェリー	赤	鶏肉
プチトマト	チリソース	ポテト

属性連鎖により取得した属性に比べ, 相関関係により抽出した属性の方が, 連想的なものであることがわかる. 連鎖属性に相関関係による属性も加えれば, 概念ベースとしての質が向上するはずである.

6.おわりに

本論文では, 概念ベースにおける属性を概念として参照することで抽出できる連鎖属性について評価を行った. 総概念数 221025 個, 平均属性数 3.32 個の概念ベースの連鎖属性を展開すると 33 次元で収束し, 17, 18 次元における連鎖属性が概念ベースにおいて有用な属性が取得できると目視実験により確かめた. また, 概念に相関している単語を属性として付与し, 精錬する手法について提案した. これらの手法は, いずれも人の手による精錬を行っておらず, 理論に基づいて概念ベースの精錬を行っている. 精度の高い概念ベースを, 機械的に構築することが今後の課題となる.

謝辞

本研究の一部は研究費(23720222)の助成を受けたものである。

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] EDR 概念辞書.(2007).情報通信機構
[3] Taku Kudo , Kaoru Yamamoto , Yuji Matsumoto.(2004).Applying Conditional Random Fields to Japanese Morphological Analysis.EMNLP
[4] IBM Content Analytics.2011.日本アイ・ビー・エム株式会社
[5] 入江毅,渡部広一,河岡司:概念ベースにおける属性数の検討と概念間の関連度計算方式,電子情報通信学会技術研究報告書,99 巻,37-44