3R－4

テキスト文からの多変量分析による語彙概念の分類

倉本 英剛　上原 悟　牧野 武則

東邦大学理学部情報科学科

1 はじめに

語彙概念を分類して階層化することは、計算機に単語や文の意味理解をさせるのに必須である。しかし、単語の意味（概念）を分類するには、多様な意味上の見方を考慮する必要があり、人手に頼らない自動的な分類手法が望まれる。「語の意味は語と語の関係から決定される。」という観点から、テキスト文のみから得られる情報をもとに、概念を分類する手法を提案する。まず、テキスト文を入力して、形態素解析を行ない、単語に品詞を付与する。そして、単語のペアの集合を、多変量分析のクラスター分析を用いて語彙概念を分類する。＜形容詞＞-＜名詞＞、＜副詞＞-＜動詞＞、＜動詞＞-＜名詞＞の関係について実験を試みた。

2 品詞間の依存関係と語彙概念

上の3つを実験した理由は以下にある。

直接の依存関係にある形容詞と名詞は、形容詞が名詞の概念の属性を指示していると見ることができ、動詞と名詞では、格関係子について可能な名詞概念の属性を指示している。このことは、依存関係がある形容詞と名詞のペアの集合から形容詞と名詞の属性を分類できることである。つまりそれらの意味分類ができることを示している。

また、動詞と名詞の共起関係から、その格関係子に対して名詞、動詞を分類できることが予想される。

副詞と名詞については、副詞は動詞の意味だけでなく、時制、時相（アスペクト）、様相と関連を持つと考えられる。

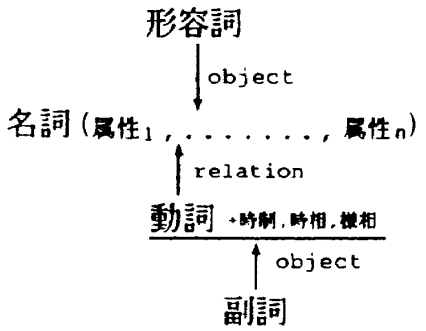

図 1: 品詞間の依存関係

3 多変量分析法

名詞を分類するには、その名詞と係り受ける形容詞のセットの動詞および関係子のセットが手掛かりとなる。つまり、概念の類似度はそれらのセットの類似度に対応している。また、形容詞、動詞も同様に分類される。

3.1 クラスター分析法

クラスター分析は、異質なものの混ざりあっている対象を、それらの間に何らかの意味で定義された類似度を手掛かりにして、似たものを集めいくつかの均質なものの集落（クラスター）に分類する分析法である。

3.2 群平均法

群平均法とはクラスターの類似度を、それに含まれる対象間の平均的な値で定義する方法である。いまｐ－クラスター、ｑ－クラスター、ｒ－クラスターの大きさ（クラスターに含まれる対象の数）をそれぞれ n_p, n_q, n_r とするとき、

$$S_{tr} = \frac{n_p S_{pr} + n_q S_{qr}}{n_p + n_q}$$

と表される [5]。

4 実験

概念の意味は、他の概念との関係、つまり「属性」によって決定される。したがって、その概念が持っている属性の束によって、概念の分類が出来る。

Lexical Concept Classification with Cluster Analysis
Eigo Kuramoto, Satoru Uehara, Takenori Makino
Department of Information Science, Faculty of Science, Toho University
2-2-1 Miyama,Funabashi,Chiba 274

そして、「属性」のデータ（ある概念が他の概念とどのように関係しているかという情報）を、実際の文章から得る。これは、概念の意味が事実のみによって決定されることを表しているので、その分類も事実に即したものになり、人手によらない分類が行なわれることになる。

分類に使用するデータとして、ＥＤＲ電子化辞書を用いた。形態素解析には、藤川 [3] のシステムを用いて、単語に品詞を付与した。これで得たデータをもとに、クラスター分析を行ない、語彙概念の分類を行なう。

例えば"概念"が3個（S_1,S_2,S_3）で、"属性"が7個（$K_1,K_2,..,K_7$）あったとする。ここで、$\alpha_i(j)$という変数を次のように定義する。

$\alpha_i(j)=1$ … S_iとK_jとの関係が文章データにあるとき。

$\alpha_i(j)=0$ … そうでないとき

ここで表1のような結果が与えられたものとする。

	K_1	K_2	K_3	K_4	K_5	K_6	K_7
S_1	0	1	1	0	1	1	1
S_2	0	1	0	0	1	1	0
S_3	1	0	1	1	1	0	0

表 1: 概念の属性度

そこで、"概念"におけるお互いの類似度を表す数値としてe_{ij}を用意すると、以下のように考えればよい。2つの対象S_i,S_jに対して、属性1と0の関係が(1,1)または(0,0)のように同じであれば、S_iとS_jは似ていると考えられる。非類似度を用いるので、(1,0)という組があれば+1とする。これによりS_1とS_2を調べてみると、非類似度$e_{12}=2$となる。これを全ての組み合わせで調べると、

$$e_{ij}=\sum|\alpha_i(j)-\alpha_k(j)|$$

から求めることができる。よって以下の表2を得る。ただし$e_{ij}=e_{ji}$とし、$e_{ii}=0$とする。

この概念間の非類似度を基に、群平均法により樹形図を構成する。

	S_1	S_2	S_3
S_1	0	2	4
S_2	2	0	6
S_3	4	6	0

表 2: 概念間の非類似度

5 結果

EDR電子化辞書より5000文の中から単語のペアを抜き出し、今回の実験を試みた結果、形容詞に関して一部分ではあるが、図2のような結果を得た。少ない情報量からでも、ある程度の精度を持った分類が行なわれた。

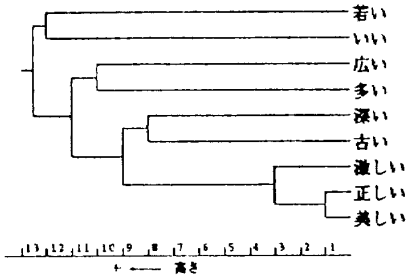

図 2: クラスター分析の結果

図2は名詞との係わりで形容詞を分類した例であるが、問題点もみられる。例えば「若い」と「美しい」については、一番距離が遠く分類されており、一番似ていない概念となった。しかし「若い人」、「美しい人」のように同じ属性を持つ事もあり、他の概念よりも短い距離で分類されるべきである。しかしこのような結果が見られるのは、今回用いたデータの中にそういったデータが存在しなかったためである。

今回の結果より、本方式が有効と見られ現在10万文に対して、実験を続けている。

6 おわりに

本論では、「語の意味は語と語の関係から決定される。」という観点から語の分類と、階層化するための方法を提案した。属性を用いることにより、分類に対する明確な基準がないという従来の概念分類の問題点を解消した。

だが、今回予備的な実験において、分類がうまくいかない部分もあった。この原因は以下の2点である。

まず、データ量の不足である。これを解決すれば、データの増加に伴い単語のペアも増加し、概念の属性が増すので、より細かなクラスター分析が可能となる。

もう一つは、形態素解析の精度の問題である。今回の予備的実験で用意した単語のペアの集合は、品詞で抜き出すために、形態素解析に依存してしまう。自然言語の曖昧さから本来違う品詞でも、"形容詞"や"動詞"という情報があるとその単語も抜き出して解析してしまうので、形態素解析の精度の向上が必須である。

参考文献

[1] 牧野「自然言語処理」オーム社 (1991)
[2] 前川、伊藤、古郡「係り受け関係と相互情報量を用いた単語の意味獲得」情報処理学会第55回全国大会 2R-1(1997)
[3] 藤川「bigram 情報と文法情報による形態素解析」情報処理学会第55回全国大会 4AE-2(1997)
[4] 牧野「機械翻訳」オーム社 (1989)
[5] 田中、脇本「多変量統計解析法」現代数学社 (1983)