自然言語処理 96－4
（1993．7．9）

Dempster-Shafer 理論を応用した コーパスからの言語知識の獲得手法

関根 聡
東京情報システム研究所
松下電器産業株式会社

現在、電子化された言語データの増加にともなって、計算機上で電子化された言語データを利用した言語知識獲得の研究が非常に盛んになっている本論文では、コーパスから抽出した単語間の係り受け関係のデータを基に、単語間距離を計算し、コーパス中に出現しなかった単語間の係り受け関係を推測する方法について述べる。特徴としては、可能な解をすべて出力するパーザーを利用し、Dempster-Shafer 理論における「無知」という概念を導入することでコーパス中に現れていないかかり受け関係を扱う。また、そのデータに基づいて計算される単語間距離の情報や効率的な人間の介入により、かかり受け関係の有無を計算する。そして、提案するアルゴリズムに基づいて日本語復号名詞に対する実験を行ない、その結果を考察する。

Linguistic Knowledge Acquisition from Corpora Using Dempster-Shafer theory and Human Intervention

SEKINE, Satoshi
Matsushita Electric Industrial Co.,Lts
Tokyo Information System Laboratory

Currently, several attempts are made on research of linguistic knowledge acquisition from corpus. The purpose of this paper is to propose a linguistic knowledge acquisition based on data of word dependency. The system uses a parser which produces all possible analyses to derive word dependency data. Also a smoothing technique and a human intervention method are introduced. The Dempster-Shafer's theory which can handle disbelief data is applied to the system. An experiment for Japanese compound nouns is conducted and evaluation of the result is presented.

1 はじめに

現在、自然言語処理研究において、枠組や計算効率の話題にも増して自然言語の知識の重要性が認識されている。これまでの自然言語知識に関する研究では知識の型やその知識の操作手法に力点が置かれていた。しかし、最近は実際に埋め込むべき知識そのものへの関心が強まっている。過去に開発された自然言語処理システムでは、自然言語処理に利用される知識は主に言語学者やシステム開発者の内省や考察から作成されることが多かったが、現在、電子化された言語データの増加にともなって、計算機上でそれらのデータを利用した言語知識獲得の研究が非常に盛んになっている [1] [2] [3] [4]。

本論文では、コーパスから抽出した単語間の係り受けのデータを基に、単語間距離を計算し、コーパス中に出現しなかった単語間の係り受け関係を推測する方法について述べる。単語間の類似度の計算方法については、白井ら [5] の方法や、藤原ら [6] の方法がある。これらの方法は基本的に、テキストデータから抽出された係り受け関係の出現頻度を基に単語間の距離を、2つの単語が同じ環境で出現した頻度を2つの単語のそれぞれの出現頻度で正規化した形で求めている。また、n個の係り受けのデータをn次元のベクトルと考え、そのベクトル間の距離や角度を用いて、単語間の距離とするという方法もある [1]。

しかし、これらの方法には次のような欠点がある。それは、データには現れなかった係り受け関係の対処の方法である。例えばある係り受け関係が、対象としているコーパス中に現れなかったとしても、それだけを根拠に、その係り受けは不可能なものであると断定することはできない。その係り受け関係は本来成立するものであるが、たまたまそのコーパスに存在しなかっただけであるという可能性もある。つまり問題は、コーパス中に現れなかった係り受け関係は、その係り受けが成立不可能なものなのか、たまたまそのコーパスに現れなかっただけのものなのか判断することができないということである。

この問題を解決するために、本論文では「無知量」を扱うことのできる Dempster-Shafer 理論を応用し、また効率的な人間の介在を導入することでより確からしい係り受け関係の知識を生成する手法を提案する。特に考察では、効率的な人間の介在の手法について力点をおいて考える。

2 係り受けデータ

本手法の入力データとなるのは、コーパスから抽出した係り受けデータであるが、このデータを自動的に獲得することは非常に難しい。そのためには、完全な文法だけでなく、ここで獲得しようと考えている単語の意味的な知識までも必要となってくる。この問題は、関根ら [1] において「鶏と卵の問題 (Egg and Chicken Problem)」として提起されている。

そこで、この手法では係り受けデータを、関根ら [1] において提案された手法と同様の手法を用いて獲得することにする。まず、可能な解をすべて出力するパーザーを使ってコーパス中の文を解析する。つまり、このパーザーは曖昧な部分がある場合には、すべての可能な係り受け関係を出力する。次に、ここで出力された関係に信頼値と呼ばれる値を付与する。この値は、曖昧な解析の数に反比例する。例えば、ある範囲において曖昧な解析が3つある場合にはそれぞれの係り受け関係に1／3の信頼値を与え、ある部分において、曖昧性がなく1つしか解析結果がない場合にはその係り受け関係には1の信頼値を与える。そのようにして、コーパス中の文をすべて解析した後に、それぞれの係り受け関係の信頼値をある式にしたがって積算し、それぞれの関係のコーパス全体における真偽値という値を計算する。この真偽値、それぞれの係り受け関係が成立する度合を表している。

この手法の特徴は、完全に正確な答を出すものではないが、現在の技術では完成された技術になっていないパーザーや文法の不完全さを考えに入れて、その弱点をカバーするものであるといえる。つまり、1つしか解を出さないパーザーは、そのパーザーのもつ文法やそこに含まれているヒューリスティックを反映した1つの解を出すものであり、その文法を矛盾なく完全に作り上げることは現在のところ不可能であると考えられているため、求められた1つの解は完全に誤りであることがしばしばあり得る。しかし、そのような完全性を追求せずに、ある程度大雑把ではあるが、そのパーザーの作り出す解の中に正解がかなりの確率で含まれる文法を書くことは、比較的難しくないことであると考えられる。したがって、後者の可能な解をすべて出力するパーザーをうまく利用してより確からしいデータを作ることを目標として、上記の手法は開発された。ただ、本来1つの解しかないところに、可能な複数の解を出力するということは、確実に誤った情報を取り込むことであり、その危険性は避けられないものである。しかし、上記の手法は、コーパス全体を見通した際に、誤った解の出現する回数より正解の出現する回数の方が多いであろうという仮定により成立している。したがって、上記の簡単なアルゴリズムの説明で述べように、真偽値は信頼値の積算により求まるという点が本手法の重要な部分である。

3 Dempster-Shafer 理論

このように計算された真偽値の値域は、[0, 1] であり、1であるということは、その係り受け関係は、真であるということを意味し、値が低くなればなるほどその係り受け関係は成立しないであろうと捉えることができる。しかし、この値を求めた方法はコーパスにおける係り受け関係の出現頻度が主なソースとなっているため、0に近いからといって、その関係が成立しないものである決めつけることはできない。この問題は、単語や文中の関係の出現頻度に基づいて何かを求める方法においては避けられない問題である。そこで、この問題を解決す

るために無知量を効率的に取り扱うことのできるDempster-Shafer理論を導入する[7]。一般的に確率を取り扱うBayse確率では、

$$P(A)+P(\bar{A})=1 \qquad (1)$$

という関係が要求されるため、信用の欠如(lack of belief)と不信用(disbelief)の区別ができず、ともに$P(\bar{A})$となってしまう。それに対し、Dempster-Shafer理論では、それぞれの要素を部分集合で表し、無知な状態は、そのすべての集合を包含する全体集合として表すことにより無知量を取り扱える枠組になっている。この理論は本手法において実際に以下のように応用している。

A_T : 係り受け関係が真である集合
A_F : 係り受け関係が偽である集合
A_0 : 全体集合

$m(A_T)$: 係り受け関係が真である基本確率
$m(A_F)$: 係り受け関係が偽である基本確率

$$\sum_{A_i \subseteq A_0} m(A_i)=1 \quad (i=0,T,F) \qquad (2)$$

このように定義した場合、必ず要素がある集合に含まれるという下界確率は基本確率と等しくなり、これをもって、それぞれの要素の確率とする。また、無知であるという確率は式.2から

$$m(A_0)=1-m(A_T)-m(A_F) \qquad (3)$$

として求めることができる。アルゴリズムの説明部分で詳しく説明するが、本手法では、この「係り受け関係が真である確率」、「係り受け関係が偽である確率」、「係り受け関係が無知である確率」を使い分けることによりより確からしい知識を獲得している。

4 Gradual Approximation

本論文で提案する手法に含まれる他の特徴あるテクニックはGradual Approximationである。これに似たアルゴリズムは、藤原ら[6]によって「反復クラスタリング」として提案されているが、この手法はクラスタリング結果を次のクラスタリングに利用するという手法であるのに対し、関根ら[1]において提案されたGradual Approximationは、文に含まれる解析データと知識として抽出されたデータの間の相補完的な知識の収束を目的としており、異なった種類のデータを利用し知識を収束させるという点で本質的に異なっている。本手法におけるGradual Approximationを説明する。

ある段階で獲得された単語間距離のデータは単語間の類似度を示す指標であり、ある2つの単語間の距離は、その2つの単語を含みそれ以外の要素が同じである係り受け関係に反映することができる。つまり、係り受け関係A，Bがそれぞれ単語W_i，W_jを同じ種類の要素として含み、それ以外の要素が全く同じであるとするならば、2つの係り受け関係A，Bの関係は、単語W_i，W_j間の関係を反映することができると考えられる。そして、この情報を基に係り受け関係の真偽値を再計算させることは無知な情報を埋める(Smoothing)ために有効である。具体的には、2つの係り受け関係A，Bの関係は、その距離として単語W_i，W_j間の距離を利用し、それぞれの係り受け関係の真偽値を利用してお互いに影響を与えるという方法をとる。そのようにして再計算された係り受け関係の真偽値をもとに、次の単語間距離のデータを計算し直し、新たな単語間距離の知識を得ている。そしてまた、その単語間距離とその時求まっている係り受け関係の真偽値を基に、新しい係り受け関係の真偽値を再計算するというように、2つの種類の情報の間でGradual Approximationを行ないながらより確からしい知識へと収束させていく。以上が、本手法で用いられるGradual Approximationの手法である。

5 効率的な人間の判断の介在

しかし、このように統計的な手法を活用しても、コーパスから抽出できる知識には限度がある。したがって、本手法では人間の持っている自然言語知識を利用することにより、単語間の距離や係り受け関係の知識の獲得を効率的に獲得することを考える。人間の介入のやり方には、いくつかの方法が考えられる。例えば、無知の係り受け関係の真偽をすべて人間に尋ね、情報を完全に作り上げてしまうという方法である。しかし、これではコーパスが大きくなると、非常に大きな手間となってしまい、現実的ではない。すべてを尋ねることが困難であるのなら、人間に尋ねる無知の係り受け関係の数をなるべく少ない状態でより高精度な情報を獲得することが望ましいと考えられる。本研究では、人間の介在の効率化も1つの研究課題に取り上げている。

例えば、単純に無知な係り受け情報についてランダムに人間に尋ねるという方法もある。しかし、その時点で得られているかかり受け関係の情報を利用して効率的に質の高い情報を得ることを考える。また、前述したGradual Approximationの手法を組合せ、1回のサイクルに少しずつ人間の判断を仰ぎ、その情報の追加に基づき再計算を行ない、Gradual Approximationの中でより確からしい情報に収束させていくという戦略を取る。

本手法では、係り受け関係に「真」「偽」「無知」という3つの値を導入したことを利用し、以下の3つの基準に従い人間の介在の対象とするかかり受け関係を決定し、それを人間に判断させるという方法を検討する。

1. 「真」の値と「偽」の値の差が小さく、かつ「無知」の値の小さいものから順番に人間に判断を仰ぐ。

2. 「無知」の値を持ち、「真」の値の高いものから順番に人間に判断を仰ぐ。

3. 「無知」の値を持ち、「偽」の値の高いものから順番に人間に判断を仰ぐ。

最初の手法では、「無知」の値が低いもののうち「真」か「偽」かどちらになるかの境界上にあるものから決定していけば、データの収束に効果的であるという推測に基づいている。次の手法では、ほとんど「真」であると思われるものを完全に「真」であると決定し、また、本来「偽」であるのに、「真」に近い値を与えられ、全体に誤った影響を与えることを防ぐ効果も期待できる。3番目の手法も2番目の手法と同様の期待に基づいている。

6 アルゴリズム

次に、本手法のアルゴリズムを説明する。まず、概要において具体的な例に基づいてアルゴリズムを簡略的に説明した後、用語の説明およびアルゴリズムの定義を行なう。

6.1 概要

図.1に示した全体の構成図に基づいて説明を行なう。

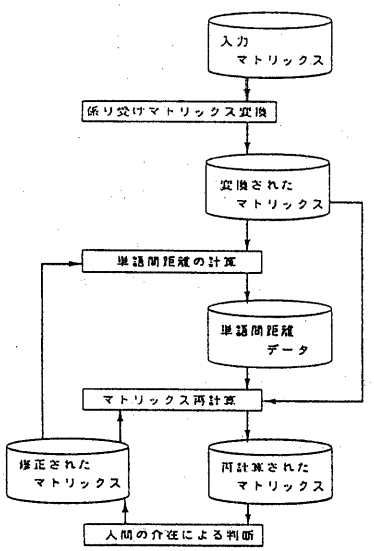

図 1: 全体構成図

	木	花	草	手首
育つ, 動作主	0.1	0.0	1.0	0.0
咲く, 動作主	0.0	0.3	0.0	0.3
枯れる, 動作主	1.0	0.0	0.8	0.0
切る, 目的格	1.0	1.0	0.0	1.0

図 2: 入力となる係り受け関係

図.2は本手法の入力となる係り受け関係の例である。この例では、縦軸に単語間の意味的な関係におけるかかり元の単語と意味的な関係が示され、横軸には単語間の意味的な関係におけるかかり先の単語が示されている。また、値は0から1に正規化されており、1に近付くほどその単語間の意味的関係が真である程度が高いことを意味する。本来はパーザーを使った解析により、その曖昧性の程度から値を決めるべきであるが、説明のための例として上記のような例を使う。

このデータから、本手法の特徴である「真」「偽」「無知」の3つの値を決定する。まず、係り受けマトリックスのそれぞれの値はそのまま「真」である確率に割り当てる。それ以外の部分、つまり1から「真」である確率を引いた部分は、9：1の割合で「無知」と「偽」に割り当てることとする。この「偽」の確率の決定方法は、真偽値の計算が「真」の確率を中心に行なわれており、「真」である確率が低い場合には一般的に「偽」である確率は高いと考えられるため、このような方法をとった。また、「無知」と「偽」に振り分ける比率については、予備的な実験を行なった上、この割合に決定した。この係り受けマトリックスの変換により、図.2の係り受け関係のデータは、図.3のように変換される。図において、各要素は上から、「真」「偽」「無知」の値を示す。つ

	木	花	草	森
育つ, 動作主	0.1	0.0	1.0	0.0
	0.09	0.1	0.0	0.1
	0.81	0.9	0.0	0.9
咲く, 動作主	0.0	0.3	0.0	0.3
	0.1	0.07	0.1	0.07
	0.9	0.63	0.9	0.63
枯れる, 動作主	1.0	0.0	0.8	0.0
	0.0	0.1	0.02	0.1
	0.0	0.9	0.18	0.9
切る, 目的格	1.0	1.0	0.0	1.0
	0.0	0.0	0.1	0.0
	0.0	0.0	0.9	0.0

図 3: 変換された係り受け関係

ぎに、このデータに基づいて単語間距離を求める。それぞれ、係り受け先と係り受け元の単語について、「真」「偽」「無知」の値に配慮した式（式.7）を利用し単語間の距離を計算する。式は両方が同じ要素を持つ値から違う要素を持つ値を引いたものを全体の数で正規化したものであり、この値の値域は［－1，1］となる。値が大きいほど、類似性が高く、低いほど類似性はないということになる。図.3のデータからこの計算方法にしたがって求めた単語間距離を図.4に示す。ここで計算された単語間距離と、図.3の係り受け関係の真偽値をもとに、真偽値を再計算する。この再計算の公式は、式.8 - 式.12である。これによって、単語の類似性に基づいて、似ているかかり受け関係間においてそれぞれの関係の値を高めたり、似ていないかかり受け関係の間においてそれぞれの関係の値を低くするといった影響を与えあう。具体的

	木	花	草	手首
木	1.0	0.081	0.065	0.081
花	0.081	1.0	-0.029	0.097
草	0.065	-0.029	1.0	-0.029
森	0.081	0.097	-0.029	1.0

図 4: 計算された単語間距離

には、この再計算において、ある係り受け関係が他の係り受け関係に与える影響は、それぞれの係り受け関係において、1つの単語だけが異なり、他の要素は全く同じ場合に、その異なる単語間の距離と影響を与える係り受け関係の真偽値に比例し、その影響の中でその関係に与える「真」および「偽」のいずれかの影響がもっとも大きなもの1つだけの影響を繰り入れることとしている。「真」及び「偽」の影響は、影響を受ける関係の「無知」の値によって場合わけが複雑になっているが、直観的には、大きな真偽値を持つ似ている係り受け関係があれば、それから大きな影響を受け、それと同じ方向の値となるように設計されている。この影響により、図.3の係り受け関係の値は図.5のようになる。ここで、簡単に計算結

	木	花	草	手首
育つ, 動作主	0.181	0.007	1.000	0.007
	0.163	0.125	0.000	0.125
	0.656	0.868	0.000	0.868
咲く, 動作主	0.032	0.489	0.001	0.489
	0.190	0.114	0.103	0.114
	0.778	0.397	0.896	0.397
枯れる, 動作主	1.000	0.073	0.944	0.073
	0.000	0.190	0.024	0.190
	0.000	0.737	0.032	0.737
切る, 目的格	1.000	1.000	0.059	1.000
	0.000	0.000	0.190	0.000
	0.000	0.000	0.751	0.000

図 5: 変換された係り受け関係

果の評価を行なってみる。「真」「偽」「無知」のうちもっとも大きな値がついている要素がその関係を代表するものであるとすると、上記のデータの内、「真」「偽」のいずれかと答えが一致しているものが7つ、そうでないものは9つであるので、この状態での正解率は43.8%である。

次に人間の介在を導入する。この例では、データの数が少ないので分かりにくいが、一般に大きなデータを扱うと、「無知」の値を持つデータの数が多くなり、人間の介在によって、効率的に係り受け関係のデータを求めることが有効になってくる。この例では、「真」と「偽」の値の差が小さくかつ「無知」の値の小さいものから順に人間によって係り受け関係の真偽を判断し、効率的に値を収束させようとしている。1回のサイクルにおいて判断を行なうかかり受け関係の個数は2個とする。つまり、「真」か「偽」かに値が偏っている関係は、ほぼその係り受け関係の成否は分かっており、ちょうど境界にあり「無知」ではない、つまりどちらとも決定し難い関係から決定させることにより、効率的に値が収束するという考えである。この順序づけの公式は、式.13であり、その式に基づくと図.6の順に関係を人間に係り受け関係の成立不成立を判断してもらうことになる。

関係	真	偽	無知	正解
育つ, 動作主, 木	0.181	0.163	0.656	真
咲く, 動作主, 手首	0.489	0.114	0.397	偽

図 6: 真偽の境界上にある係り受け関係

このようにして人間の介在を行なった結果を基に、単語間距離を計算し直し、類似かかり受け関係の影響を計算し、また人間の介在を行なう というプロセスを繰り返す。この例では、1回の人間の介在につき2つの関係を訂正し、3回の繰り返しを行なった結果、係り受け関係の値は図.7のようになり、ここでの正解率は93.75%となった。

	木	花	草	手首
育つ, 動作主	1.000	0.237	1.000	0.255
	0.000	0.572	0.000	0.563
	0.000	0.191	0.000	0.182
咲く, 動作主	0.159	1.000	0.133	0.000
	0.620	0.000	0.555	1.000
	0.221	0.000	0.312	0.000
枯れる, 動作主	1.000	1.000	0.976	0.000
	0.000	0.000	0.024	1.000
	0.000	0.000	0.000	0.000
切る, 動作主	1.000	1.000	1.000	1.000
	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	0.000

図 7: 3サイクル後の係り受け関係

6.2 用語

アルゴリズムで使用される用語の定義を行なう。

V_{ij}	要素 (i,j) の入力時の値
T_{ij}	要素 (i,j) の「真」の値
F_{ij}	要素 (i,j) の「偽」の値
U_{ij}	要素 (i,j) の「無知」の値
$D(x,y)$	x,y の単語間距離
T'_{ij}	再計算時の基になる「真」の値
F'_{ij}	再計算時の基になる「偽」の値
U'_{ij}	再計算時の基になる「無知」の値

6.3 アルゴリズム

1. 係り受け関係マトリックスの変換
入力された係り受けマトリックスの値 V_{ij} に従い、「真」「偽」「無知」の値 T_{ij}, F_{ij}, U_{ij} を式.4 - 式.6にしたがって計算する。

$$T_{ij} = V_{ij} \quad (4)$$
$$F_{ij} = 0.1 * (1 - V_{ij}) \quad (5)$$
$$U_{ij} = 0.9 * (1 - V_{ij}) \quad (6)$$

2. 単語間距離の計算
上記において作成された係り受けマトリックスを基に、以下の公式にしたがって単語間距離を計算する。

$$D(W_a, W_b) = \frac{d(W_a, W_b)}{(d(W_a, W_b) + \alpha)} \quad (7)$$

$$d(W_a, W_b) = \sum_i \{T_{ia} * T_{ib} + F_{ia} * F_{ib} - T_{ia} * F_{ib} - F_{ia} * T_{ib}\}$$

この単語間距離は、係り受け側、係り先側ともに計算を行なう。α は単語間距離の相対的な大きさを決定する定数であり、次の節で述べる実験では１０としてある。

3. 類似係り受け関係による影響
類似係り受け関係からの影響の計算を行なう。影響は、単語間距離と影響を与える関係の真偽値に依存しており、もっとも影響を与える関係のみから影響を受けるものとする。計算式では場合分けが複雑であるが、基本的には大きな値を持つ似ている関係があればその関係から大きな影響を受けるように設計してある。

$$T_{ij} = T'_{ij} + (U'_{ij} * f(i,j)) \quad (U'_{ij} > 0) \quad (8)$$
$$F_{ij} = F'_{ij} + (U'_{ij} * g(i,j)) \quad (U'_{ij} > 0) \quad (9)$$
$$T_{ij} = \frac{fval}{fval + gval} \quad (U'_{ij} = 0) \quad (10)$$
$$F_{ij} = \frac{gval}{fval + gval} \quad (U'_{ij} = 0) \quad (11)$$
$$U_{ij} = 1 - T_{ij} - F_{ij} \quad (12)$$

$$fval = T'_{ij} * (1 + f(i,j))$$
$$gval = F'_{ij} * (1 + g(i,j))$$
$$f(i,j) = f_0(i,j) \quad (f_0 + g_0 < 1)$$
$$f(i,j) = \frac{f_0(i,j)}{f_0(i,j) + g_0(i,j)} \quad (f_0 + g_0 \geq 1)$$
$$g(i,j) = g_0(i,j) \quad (f_0 + g_0 < 1)$$
$$g(i,j) = \frac{g_0(i,j)}{f_0(i,j) + g_0(i,j)} \quad (f_0 + g_0 \geq 1)$$
$$f_0(i,j) = \max_k f_1(i,j,k), f_2(i,j,k)$$
$$g_0(i,j) = \max_k g_1(i,j,k), g_2(i,j,k)$$

$$f_1(i,j,k) = D(W_j, W_k) * T'_i k \quad (\mathrm{D(W_j, W_k)} > 0)$$
$$f_1(i,j,k) = -D(W_j, W_k) * F'_i k \quad (\mathrm{D(W_j, W_k)} \leq 0)$$
$$g_1(i,j,k) = D(W_j, W_k) * F'_i k \quad (\mathrm{D(W_j, W_k)} > 0)$$
$$g_1(i,j,k) = -D(W_j, W_k) * T'_i k \quad (\mathrm{D(W_j, W_k)} \leq 0)$$
$$f_2(i,j,k) = D(W_i, W_k) * T'_k j \quad (\mathrm{D(W_i, W_k)} > 0)$$
$$f_2(i,j,k) = -D(W_i, W_k) * F'_k j \quad (\mathrm{D(W_i, W_k)} \leq 0)$$
$$g_2(i,j,k) = D(W_i, W_k) * F'_k j \quad (\mathrm{D(W_i, W_k)} > 0)$$
$$g_2(i,j,k) = -D(W_i, W_k) * T'_k j \quad (\mathrm{D(W_i, W_k)} \leq 0)$$

4. 人間の介在
係り受け関係の知識を効率的に収束させるために、人間の介在により、係り受け関係の成否を判断する。この効率化のために、人間に判断させる係り受け関係を以下のいずれかの基準で決定する。人間に判断させる係り受け関係は、基準１（式.13）の値が小さいものから、基準２（式.14），基準３（式.15）の場合は値の大きいものから順に決められた個数までのものを対象とする。

$$C = |T_{ij} - F_{ij}| + U_{ij} \quad (13)$$
$$C = T_{ij} \quad (14)$$
$$C = F_{ij} \quad (15)$$

これらの基準に基づいて提示された係り受け関係について、人間は「係り受け成立する」か「成立しない」かのどちらかを選ぶ。ここで、「成立する」と判断された関係については、$T_{ij} = 1.0$, $F_{ij} = 0.0$ とし、逆に「成立しない」と判断された関係については、$T_{ij} = 0.0$, $F_{ij} = 1.0$ という値を係り受けマトリックスに代入する。

5. 繰り返し
上記の人間の判断を加えた新しい係り受けマトリックスに対して、２．の単語距離の計算からのプロセスを繰り返し、Gradual Approximation を行なった後、値が収束したと判断されたら、プロセスを終了させる。

7 計算機マニュアルの複合名詞句を対象とした実験

これまで述べたアルゴリズムに従い実際のコーパスを対象に実験を行なった。

7.1 実験データ

実験データは、日本語計算機マニュアル（約８３００文）から抽出された複合名詞句（異なり名詞句数６１６，延べ名詞句数１８８１）で、このコーパスから抽出された係り受け関係のうち頻出の係り先、係り元の３２個による係り受けマトリックス（全１０２４要素）を対象に実験を行なった。人間の介在はそれぞれの基準に基づき１回のサイクルにおいて、２０個の係り受け関係に対して行なった。計算機によって判断すべき関係が画面上に

表示され、オペレータはその関係が成立するか否かを判断するだけであるので、この作業は非常に簡単なものであった。

7.2 評価

評価は、それぞれの係り受け関係において、最高の値を示す要素が正しいものと一致しているかどうかで評価を行なった。図.8には、基準1に基づいた人間の介在による実験におけるサイクルごとの「正解率」「誤答率」「無知率」の変化を示す。

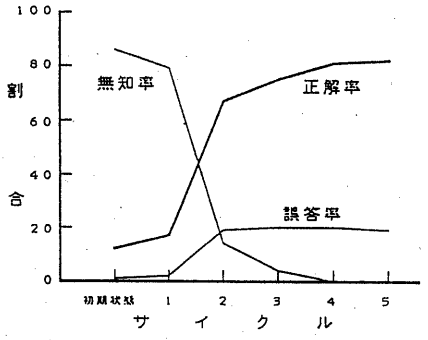

図 8: サイクルごとの正解率などの変化

図.8 から分かるように、本アルゴリズムを利用することにより5サイクル（１００の係り受け関係に対する人間の判断）だけで、全体の８０％余り（８３０個）の係り受け関係に対して正しい関係を付与することができた。また、人間の介在を行なうかかり受け関係の基準を変えた場合の結果は図.9 のようであった。図では、それぞれ人間の介在対象の基準を変えた時の、5サイクル後の正解率を示す。この件については考察において詳しく見ることにする。

介入対象の基準	基準1	基準2	基準3
正解率	81.05 %	80.76 %	77.34 %

図 9: 人間の介在基準と正解率

8 考察

この実験結果を観察すると、最終的に誤答となったものの中には本来係り受け関係が成立するはずなのに、計算結果からは成立しないと判断されたものが、多くあった（図.10）。これは、最初の係り受けマトリックス変換部で、「真」でない値を、どのように「偽」と「無知」に分配したかということに関係すると思われる。つまり、この分配において「偽」である割合を多くすればする程、全体的に不成立と計算されるかかり受け関係が多くなる傾向にあるためである。例えば、逆にこの割合を変化させ、「偽」である初期値を小さくすれば最後の誤答の中

実験結果	成立	不成立	無知
本来、成立する関係	173	166	0
本来、成立しない関係	28	657	0

図 10: 実験結果の分析

に含まれる「成立するはずだったのに駄目だったもの」の割合は減るかも知れない。しかし、同時に「成立しないはずだったのに成立してしまった」ものが増えてしまい、パラメータによるトレードオフという状態になろう。また、もう1つの距離を計算する際のパラメータ（式.7におけるα）についても同じ問題が考えられ、パラメータ決定の客観的な判断が必要である。

次に、人間の介在対象の基準に関する問題を考えてみる。コーパスという不完全な形のデータからの知識獲得において人間の介在が避けられないものであるならば、その介在を効率的に行なうべきであるということは明白である。そこで、アルゴリズムでは、人間の判断の対象とする曖昧な係り受け関係の選択に3つの判断基準を提案し、それぞれの実験結果を考察してみる。

まず、基準1（式.13）の場合を考えてみよう。この手法では、「無知」の値が低いもののうち「真」か「偽」かどちらになるかの境界上にあるものから判断していくという方法である。具体的にどのようなかかり受けが対象になったか例を示す。以下に示すのは5サイクル目に、人間の判断対象として提示されたかかり受け関係のうち上位の5つの関係である。このように、「真」か「偽」

かかり受け関係	真	偽	無知	正解
(専用, 番号)	0.494	0.495	0.010	真
(キー, オプション)	0.487	0.512	0.001	偽
(入力, コマンド)	0.482	0.518	0.000	真
(特殊, 位置)	0.482	0.518	0.000	偽
(ワード, 検索)	0.481	0.519	0.001	真

図 11: 基準1での判断対象関係の例

か境界線上にあり、「無知」の値が低いものが提示されていることがわかる。最終的な正解率を見てみると（図.9）、基準2とほぼ同じ値ではあるが3つの基準の中でもっとも良い結果を示していることがわかる。境界線上のものから決定することによって、より効率的に知識を獲得できていると言うことができる。

次に、基準2の場合を見てみる。この手法では、ほとんど「真」であると思われるものを完全に「真」であると決定し、また、本来「偽」であるのに「真」に近い値を与えられ全体に誤った影響を与えることを防ぐ効果を期待している。ここでは、具体例として、最初のサイクルと3回目のサイクルにおいて、人間の判断対象として提示されたかかり受け関係のうち上位の5つの関係を示す。図.13でわかるように、「真」の値が高い関係は、や

かかり受け関係	真	偽	無知	正解
1サイクル目				
(ファイル, 幅)	0.999	0.000	0.001	真
(非, 文字)	0.999	0.001	0.000	真
(コマンド, ファイル)	0.988	0.009	0.004	真
(名, リスト)	0.973	0.014	0.012	真
(標準, ファイル)	0.973	0.015	0.011	真
3サイクル目				
(終了, 入力)	0.906	0.092	0.002	偽
(全, 端末)	0.819	0.164	0.017	真
(ホスト, リスト)	0.819	0.164	0.017	真
(ファイル, リスト)	0.819	0.164	0.017	真
(ファイル, 機能)	0.767	0.233	0.000	偽

図 12: 基準2での判断対象関係の例

はり正解が「真」である関係であるものが多い。判断のための提示された２０個の関係のうち正解が「真」であったものは、サイクルの順に、１６個、１６個、６個、６個、１０個となっている。特にサイクルの回数が浅いうちには、正解が「真」であるものが多く、サイクルが繰り返されるにつれ、少なくなっている。これは、より「真」であると思われるものは最初の方のサイクルに提示され後になるに従って、怪しいものが増えると考えられるため当然の結果であろう。また、正解率は、基準1の方式とともに高いが、これは期待していた通り、本来成立しない関係であるのに、「真」に近い値を与えられたものを矯正し、全体に誤った影響を与えることを防いだものと思われる。

最後に、基準3の場合を見てみる。この手法では、「偽」の値が高いものから人間に提示し、人間の判断を行なうという方法である。ここでは、具体例として、5回目のサイクルにおいて、人間の判断対象として提示されたかかり受け関係のうち上位の5つの関係を示す。この基準

かかり受け関係	真	偽	無知	正解
(フィールド, 幅)	0.191	0.796	0.014	偽
(ホスト, 可能)	0.192	0.794	0.013	偽
(端末, 可能)	0.193	0.794	0.013	偽
(c, 処理)	0.199	0.790	0.010	偽
(フィールド, 可能)	0.180	0.789	0.032	偽

図 13: 基準3での判断対象関係の例

の場合には、基準2の場合と全く逆に、判断対象として提示されたかかり受け関係には、正解が「偽」が多く含まれている。順にそれぞれのサイクルにおいて提示された２０個のうちの個数を見てみると、１５個，１８個，１６個，１５個，１８個となっている。この場合では、前の基準2で見られたような数の変化は見られない。つまり、「偽」の値が大きいものについては、サイクルの回数とは関係なく正解が「偽」であるものの個数はほぼ一定であるということが分かる。また、この基準を用いた時の最終的な正解率は3つの中で最低であった。これは、正解が「偽」であり「偽」の値も大きなものがほとんどであり、誤ったものを矯正し、人間の介入の影響がほとんど役に立っていないと考えられる。

また、本手法ではマトリックスの項目として、単語を使用したが単語には意味的な曖昧性が存在し、1つの単語であれ2つ以上の項目をあげて計算を行なった方がいいものがあると考えられる。このような単語の意味的曖昧性に対処する方法はこれからの課題である。

9 むすび

自然言語の大きな問題となっている、知識獲得方法において無知量を扱うことのできる Dempster-Shafer 理論を応用した手法と適切な人間の介在により効率的に単語の距離と係り受け関係の知識を獲得できる手法を提案した。中規模な実験を行なった結果その手法の有効性も実証できた。また、「無知」の概念の導入と人間の介入の親和性及び人間の介入方法により正解率が変わってくるということが確認された。今後は、大規模な実験を行ない、アルゴリズムをより堅固なものに仕上げていくとともに、他の統計的な手法と組み合わせて、色々な知識を効率的に獲得する手法を研究してゆきたい。最後に、本研究のヒントを下さった、ＵＭＩＳＴの辻井潤一教授に感謝の意を表します。

参考文献

[1] S.Sekine, J.J.Carroll, S. Ananiadou, J.Tsujii: "Automatic Learning for Semantic Collocation" *3rd Conference on A.N.L.P.* (1992)

[2] S.Sekine, A.Ananiadou, J.J.Carroll, J.Tsujii: "Linguistic Knowledge Genarator" *COLING '92 pp.560-566* (1992)

[3] Ralph Grishman: "Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments" *Comp. Linguistics Vol.12 No.3* (1986)

[4] Kenneth Ward Church: "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text" *2nd Conference on A.N.L.P* (1988)

[5] 白井 克彦, 林 良彦, 平田 裕一, 久保田 淳市: 「係り受け解析のための辞書の構築とその学習機能」 情報処理学会誌 Vol.26 No.4 pp706-714 (1985)

[6] 藤原 祥隆, 菊池 英夫: 「段階的スコアリングによる単文表現間の類似度計算方」 電子情報通信学会論文誌 D Vol.J70-D No.11 pp.2273-2279 (1987)

[7] 石塚 満: 「Dempster & Shafer の確率理論」 電子情報通信学会論文誌 66 No.9 pp.900-903 (1983)