自 然 言 語 処 理 98－7
（1993．11．19）

一般化された実例と確率を用いた曖昧性解消

李 航

NEC C&C 研究所

〒216 川崎市宮前区宮崎 4-1-1

E-mail: lihang@sbl.cl.nec.co.jp

本稿では、一般化された実例と確率を用いた、自然言語解析における曖昧性解消法を提案する。解析する時に、得られる各解釈の意味主観確率、意味確率、構文確率を計算する。意味主観確率は解釈の意味的なもっともらしさを表す。意味主観確率を計算する時にハイパーシソーラスと呼ばれる意味知識表現系を参照する。ハイパーシソーラスでは一般化された実例が記述されている。意味確率は解釈における単語の共起頻度を表し、構文確率は解釈の構文的な出現頻度を表す。最後に、意味主観確率、意味確率、構文確率を基に解釈の優先順位を決める。本曖昧性解消法を利用して 80% 以上の文を正しく解析することができた。

A Method for Sentence Disambiguation -Using Generalized Examples and Probabilities

Hang LI

C&C Technology Research Laboratories, NEC

Miyazaki 4-1-1, Miyamae-ku, Kawasaki 216, Japan

E-mail: lihang@sbl.cl.nec.co.jp

This paper proposes a method of computing preference for disambiguation in natural language analysis. Preferece is modeled as a vector of semantic subjective probability, semantic probability and syntactic probability. Semantic subjective probability represents the plausibility of an interpretation and is computed by refering to Hyper Theasaurus where generalized examples are stored as semantic knowledge. Semantic probability and Syntactic probability represent semantic frequency and syntactic frequency of an interpretaion, respectively. Using this method, sentences are analysed with an accuracy of more than 80%.

1 はじめに

自然言語を解析する時に生じる曖昧性を如何に解消するかがまだ完全に解決されていない問題である。本稿では、一般化された実例と確率を用いた曖昧性解消法を提案する。

解析は構文、意味解析からなるとし、文脈解析がないとする。というのは、文脈解析の実現は現時点では困難だからである。人間は、文脈知識がなくても自然言語の複数の解釈からある種の優先度の順に解釈を読みとることができる。従って、構文、意味解析の際、あらゆる可能な解釈を求め、人間と同じ或は似た方法で優先度を計算し、優先度の高い順に解釈を出力することができれば、曖昧性を「解消」できることになる。

このような前提で自然言語を解析する時に解釈の意味的なもっともらしさ (解釈が意味的に可能であるかどうか) の判断がもっとも重要である。例えば、人間が「I saw a girl with a scarf.」から一つの解釈をしか読みとれないのは、意味的なもっともらしさの角度からみて解釈が一意的に決まるからである。本研究では、意味的なもっとらしさを判断するための知識を意味知識という。

従来の自然言語解析では、主に二つの方法で意味的なもっともらしさを計算している。素性による意味的なもっともらしさの計算と実例による意味的なもっともらしさの計算である。素性による意味的なもっともらしさの計算では、素性を制約として用いているため、常に制約が強すぎて可能な解釈を排除してしまうか制約が弱すぎて多くの解釈を残してしまうかのいずれに陥る。また、この方法では、新しい素性を追加すると、素性の体系が変わるので、意味知識の追加が困難である。これらの問題点を解決するために、実例による意味的なもっともらしさの計算が提案された (例えば、[浦本 91])。具体的には、解釈と実例のシソーラスにおける距離を意味的なもっともらしさとする。実例の追加という簡単な操作で意味知識を追加することができるので、実例による意味的なもっともらしさの計算が優れた方法である。しかし、この方法では、各実例の適用範囲が直接表現されていないので、次の二つの問題点がある。まず、解析の時解釈とすべての実例との距離計算を行なうため、計算量が大きい。次に、慣用表現の実例がその表現にしか適用できないにもかかわらず、実際、過大適用されることがある。

実例による意味的なもっともらしさの計算の問題点を解決するために、本研究では、一般化された実例による意味的なもっともらしさの計算を考える。予め、実例 (具体的には、格フレーム) を集め、一般化し、一般化された実例 (一般化された格フレームともいう) をハイパーシソーラスと呼ばれる意味知識の表現系で表現する。解析の際、ハイパーシソーラスを参照することによって解釈の意味的なもっともらしさを計算する。一般化された実例による意味的なもっともらしさの計算は、実例による意味的なもっともらしさの計算のメリットを保つ一方、その問題点をうまく解決している。まず、実例が一般化されているため、その適用範囲がはっきりしており、実例の過大適用がない。また、処理を主に学習のフェーズで行なうため、解析時の計算コストが小さい。

本研究では、意味的なもっともらしさを主観確率で表し、それを意味主観確率と呼ぶことにする。解釈の優先順位を決める時に、解釈の意味主観確率の計算の他に、解釈の意味的な出現頻度と構文的な出現頻度の計算も必要であることが本研究でわかった。それぞれ意味確率と構文確率と呼ばれる確率で表すことにした。解析の時、意味主観確率と意味確率と構文確率を用いて解釈の優先順位を決める。その決め方は言語学の理論に基づいている。

また、従来では、確率による曖昧性解消法も提案されている [Su88][Fujisaki89][Jelinek90][Chitrao90][Magerman91][Hindle91]。本研究の曖昧性解消法は意味主観確率を用いる点と言語学の理論を実現している点等で従来と異なる。

本曖昧性解消法を用いて英語解析の実験を行なった。トレーニング文に対して、85% の文を正しく解析することができた。未トレーニング文に対しても、80% の文を正しく解析することができた。但し、本研究でいう正しく解析することは人間がもっともらしいと感じる解釈を一番最初に出力することである。実験を通じて、本曖昧性解消法が有効であることがわかった。

本稿の構成は以下の通りである。第 2 節ではハイパーシソーラスについて述べる。第 3 節では意味確率について述べる。第 4 節では構文確率について述べる。第 5 節では優先順位の決め方について述べる。第 6 節では曖昧性解消の実験について述べる。第 7 節でまとめる。

2 ハイパーシソーラス

この節ではハイパーシソーラスについて述べる。

2.1 構成

ハイパーシソーラスは、言語概念 (具体的には、名詞概念、動詞概念、形容詞概念、副詞概念) の階層 (木構造) からなる。言語概念は語義と対応する。階層は言語概念間の上位下位関係を表す。さらに、ハイパーシソーラスでは、意味知識を表層の格フレームの形で表現する。

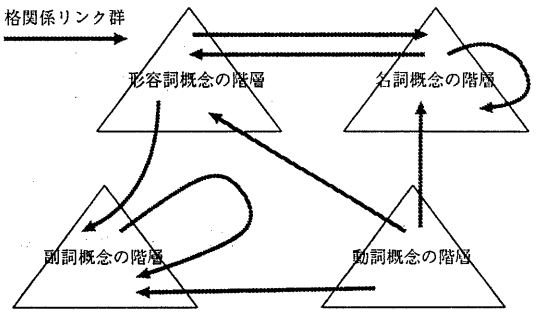

図 1: ハイパーシソーラスの構成

ヘッドになる概念が格フレームをもつ。格フレームの格スロットからフィラーになる概念 (格スロットに属する概念) へリンクが張られる。このようなリンクを格関係リンクという。例えば、動詞概念と名詞概念、名詞概念と名詞概念の間に格関係リンクが張られる (図 1を参照)。格関係リンクには Y(Yes)

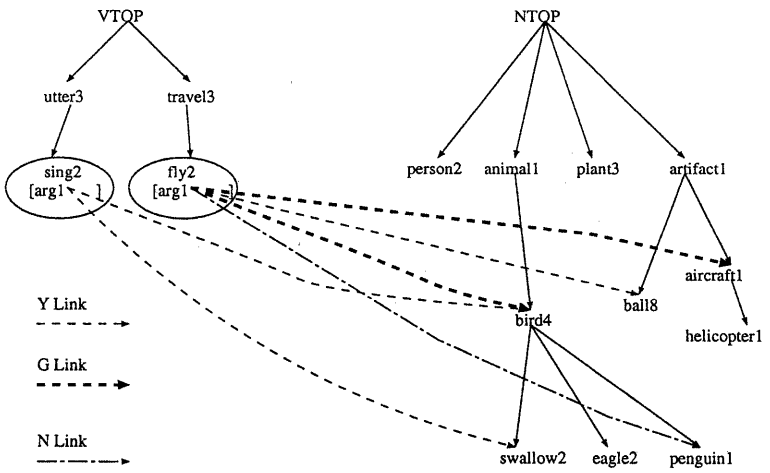

図 2: ハイパーシソーラスの例

リンク、N(No) リンク、U(Unknown) リンク、G(Group) リンクという四種類のものがある。ある概念がある格フレームのスロットに属することができる時、その概念とその格スロットの間に Y リンクを張る。ある概念とその下位概念がある格フレームのスロットに属することができる時、その概念とその格スロットの間に G リンクを張る。ある概念がある格フレームのスロットに属することができない時、その概念とその格スロットの間に N リンクを張る。ある概念がある格フレームのスロットに属するかどうかが判断しにくい時、その概念とその格スロットの間に U リンクを張る。デフォルトに N リンクを張らない。以上の四種類のリンクによって格フレームを柔軟かつ効率的に記述することができる。図 2のハイパーシソーラスでは、動詞概念 sing と fly のもつ格フレームが記述されている。図 3のハイパーシソーラスでは、実例「a girl with a telescope」を一般化したものが格フレームで記述されている。

格フレームは、まとめて記述すべきである時もあるし、各格スロットが独立であるとし、別々に記述したほうがよい時もある。本研究では、前者の格フレームを I タイプの格フレームといい、後者の場合、ヘッド、格スロット名、格スロット値の組を P タイプの格フレームという。I タイプの格フレームで慣用表現を記述することができ、いくつかの P タイプの格フレームを合わせてプロダクティブに言語表現を記述することができる。例えば、「A car eats up money.」を I タイプの格フレームで表現すべきである。「They ate supper at a restaurant.」を P タイプの格フレームで表現したほうがよい。ハイパーシソーラスにおいても、格フレームが必ず I タイプと P タイプのいずれである。

また、ハイパーシソーラスでは、格フレームは別の観点からも分類される。下位概念に継承できる格フレームと下位概念に継承できない格フレームである。

さて、格フレームを一般化し、一般化された格フレームで意味知識を完全に表現できるであろうか。本研究では、人間 (著者) がコーパスから 10 単語のもつ格フレームをランダムに集め、機械が人間に質問しながらそれらの格フレームを一般化した[1]。ある程度格フレームを集めて、一般化すれば、一般化された格フレームの累積数がほぼ収束することがわかった (図 4)。つまり、一般化された格フレームを意味知識とするのはほとんど十分であることがわかった。

ハイパーシソーラスは機械学習のための知識表現系として提案されているものである。今後、機械で自動的、或は半自動的にハイパーシソーラスを構築する研究が必要である。

2.2 メリット

ハイパーシソーラスによる意味知識表現には以下のメリットがある。

- 知識の追加が簡単である。機械が学習で獲得した知識をハイパーシソーラスの中に簡単に入れることができるし、人間も知識をその中に簡単に入れることができる。いわば、ハイパーシソーラスはオプーンな意味知識の表現系である。ハイパーシソーラスはこの点で従来の意味ネットワークと異なる。

- 正確に意味知識を表現することができる。普通のシソーラスはその形が変われば、表現する意味知識の内容も変わる [荻野 87]。しかし、ハイパーシソーラスでは、普通のシソーラスと対応する言語概念の階層の形が変わっても、意味知識の記述効率が変わるだけで、意味知識の内容が変わらない。

- 異なる精度で意味知識を表現することができる。完全正確に意味知識を表現する必要のない時、ほどほど正しく意味知識を表現することができる。その時、G リンクを多く用いて近似的に記述すればよい。

[1]その方法については後に述べる。

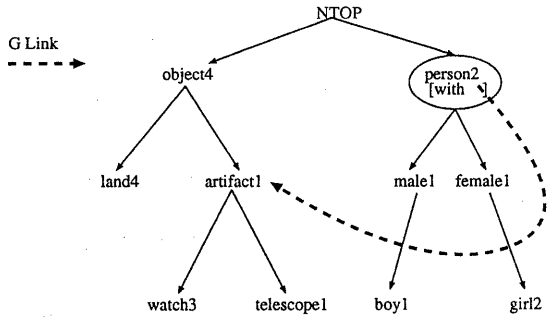

図 3: ハイパーシソーラスの例

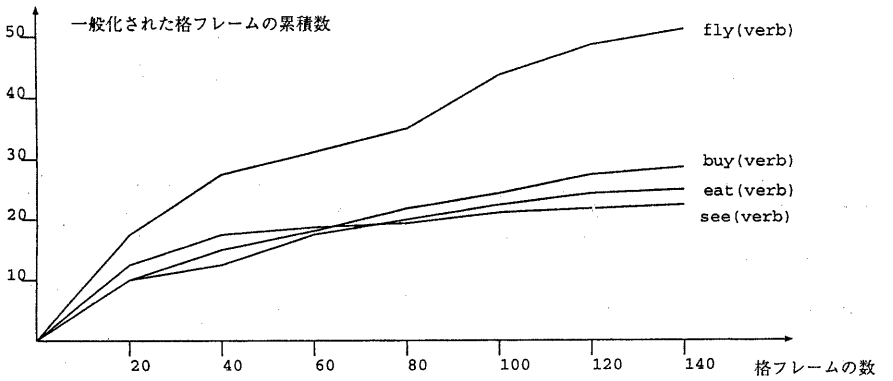

図 4: 一般化された格フレームの累積数

- 処理の効率がよい。ハイパーシソーラスでは、意味知識が格フレームパターンで表現されているため、それを利用した処理はほとんど格フレームパターンの検索で済むので、処理の効率がよい。

2.3 構築

次に、ハイパーシソーラスの現在の構築方法について述べる。WordNet[Miller92] の名詞概念と動詞概念の階層をそのまま名詞概念と動詞概念の階層として利用した。また、形容詞概念と副詞概念の階層を自分で定義した。

ハイパーシソーラスの構築を以下の手順で行なう。

1. 四つの言語概念の階層が予め与えられたものであるとする。
2. 統計的な手法を用いてコーパスから格フレームを抽出する。現在のところ、実際、自動的に抽出できたのはPタイプの格フレームだけであり、Iタイプの格フレームの抽出は人間がやる。
3. 機械が言語概念の階層を利用して人間に質問しながら、抽出できた格フレームを一般化する。さらに、ハイパーシソーラスに格フレームをもたせ、格関係リンクを張る。

まず、(Pタイプの) 格フレームの自動抽出について述べる。最初は格スロット名を抽出する。例えば、動詞概念 eat のもつ格フレームの格スロット名を抽出する時に、コーパスに現れる eat(その活用系も含めて) の後に一定サイズ (12 単語) の「窓」をあげる。その「窓」に現れる前置詞の出現頻度を統計し、頻度の高い前置詞を eat のもつ格フレームの格スロット名とみなす。この方法で格スロット名を抽出する時の誤り率が 18% であった。表 1 に自動的に抽出できた格スロット名の例を示す。間違って抽出したものをアンダーラインで示している。

次に格スロット値を抽出する。例えば、動詞概念 (名詞概念でもある)demand のもつ格フレームの格スロット for の格スロット値を抽出する際、コーパスに demand が現れ、更にその後に for が現れる時、for の後に一定サイズ (5 単語) の「窓」をあげる。「窓」に現れる単語の中で品詞が名詞である単語の出現頻度を統計し、出現頻度の高い単語を格スロット値とみなす。名詞概念と動詞概念のもつ格フレームの格スロット値の自動抽出結果の例を表 2 に示す。間違って抽出したもの

表 1: 格スロット名の自動抽出結果の例

eat の格スロット名	of	in	to	for	at	up	with	by	into	on	from
出現頻度	227	192	141	108	107	78	74	63	54	50	37
fly の格スロット名	to	in	of	from	on	for	with	over	at	into	by
出現頻度	712	337	194	175	136	117	100	86	69	67	59

表 2: 格スロット値 X の自動抽出結果の例

X gain	a	million	one-time	exordinary	net	will	capital	special	sales
出現頻度	573	296	107	90	79	57	53	41	41
demand for X	in	products	a	oil	stocks	goods	services	dollars	bonds
出現頻度	124	109	99	66	38	34	33	28	28
eat at X	a	home	time	in	restaurants	times	house		
出現頻度	12	7	6	6	4	3	3		
buy X	a	in	shares	back	stock	million	selling	more	stake
出現頻度	3181	2767	2437	1433	1252	1052	767	731	618

をアンダーラインで示している[2]。この方法で名詞概念のもつ格フレームの格スロット値を抽出する時の誤り率が 26% で、動詞概念のもつ格フレームの格スロット値を抽出する時の誤り率が 44% であった。

格フレームを抽出した後、機械が言語概念の階層を利用して人間に質問しながら、格フレームを一般化していく。例えば、「a swallow flies.」の格フレームをコーパスから抽出できたとする。機械が名詞概念の階層を参照し、swallow の上位概念 bird が fly の主語 (格スロット arg1) になれるかどうかを人間に質問する。具体的には、「a bird flies.」が意味的に正常であるかどうかを質問する。yes が返ってきたとする。より正確に意味知識を記述するためには bird の下位概念の penguin と eagle が fly の主語になれるかどうかについても人間に質問する必要がある。記述する意味知識がほどほどに正しければよいのであれば、その時点で、bird の下位概念がすべて fly の主語になれるとし、fly の格スロット arg1 から bird へ G リンクを張る。機械がさらに上位概念 animal が fly の主語になれるかどうかを人間に質問する。no が返ってきたとする。そこで、その格フレームの格スロット arg1 の一般化を終了する。次にヘッドの一般化を行う。fly の上位概念 travel がその格フレームをもつことができるかどうかについて人間に質問する。no が返ってきた時、一般化を終了する[3]。

2.4 意味主観確率

本研究では、解釈の意味的なもっともらしさを主観確率で表現し、意味主観確率と呼ぶことにする。ハイパーシソーラスにおける一般化された格フレームを参照することによって解釈の意味主観確率を計算する。

解釈 (格フレーム) の意味主観確率を以下のように計算する。

1. まず、格フレームが一般化された格フレームに属するかどうかをチェックする。属する時に、以下のケースが考えられる。

   (a) 格フレームが一つの I タイプの一般化された格フレームに属する。或は、格フレームは、一つの I タイプの一般化された格フレームと上位から継承されるいくつかの P タイプの一般化された格フレームを合わせたものに属する。

   (b) 格フレームがいくつかの P タイプの一般化された格フレーム (上位から継承されるものも含めて) を合わせたものに属する。

   このいずれのケースにおいても、格フレームの各格スロットの意味主観確率を求めることができる。対応する一般化された格フレームの格スロットの格関係リンクが Y リンク、或は G リンクの時、そのスロットの意味主観確率が 1 であり、格関係リンクが U リンクの時、そのスロットの意味主観確率が 0.5 であり、格関係リンクが N リンクの時、そのスロットの意味主観確率が 0 である。

2. 一般化された格フレームに属さない格スロットの意味主観確率が 0 であるとする。

3. 格フレームの意味主観確率がその格フレームを構成する格スロットの意味主観確率の積であるとする。

意味主観確率の値の大きさで解釈の優先順位をつけることができる。

図 4からもわかるように、一般化された格フレームで自然言語の表現を完全にカーバーするのは不可能である。ハイパーシソーラスに蓄えていなければ、解釈の意味主観確率が 0 になる。以上の意味主観確率の決め方だと、一つの格スロットが一般化された格フレームに属しない時でも、その解釈の意味主観確率が 0 になるし、すべての格スロットが一般化された格フレームに属しない時でも、その解釈の意味主観確率はやはり 0 になる。意味知識が不完全な時、前者の 0 が後者の 0 より「まし」であることを表現したい。そういうことを表現できれば、不完全な意味知識を利用しても、意味主観確率で解釈の優先順位を決めることができる。本研究では以下のような工夫でそのことを実現している。実際、格フレームの一般化された格フレームに属さない格スロットの意味主観確

[2] a、in、will は辞書の中では名詞にもなっている。

[3] このように構築できたハイパーシソーラスは図 2のハイパーシソーラスと少し異なる。

率が0でなく、0.01(十分小さい値)であるとする。例えば、「The local elite ate chocolate for dessert.」を解析する時に、

$$[elite, [mod, local]] \quad (1)$$

の意味知識が定義されていないとする。また、

$$[eat, [arg1, elite], [arg2, chocolate], [for, dessert]] \quad (2)$$

の意味知識が定義されているとする。よって、文全体では、前置詞句「for dessert」が「ate」に係る解釈の意味主観確率が0.01になり、前置詞句「for dessert」が「chocolate」に係る解釈の意味主観確率が0.0001になる。前者の解釈が優先される。

3 意味確率

意味主観確率を計算し、優先度とすれば、多くの場合解釈の優先順位を正しくつけることができる。しかし、意味主観確率だけを計算して優先度とするのは不十分である。例えば、

He killed a girl with a knife. (3)

という文には、「彼はナイフで女の子を殺した」と「彼はナイフをもっている女の子を殺した」という二つの解釈がある。意味主観確率だけをみると、前者の解釈の意味主観確率と後者の解釈の意味主観確率が共に1である。人間が意味主観確率だけを優先度としているのであれば、人間が両者の解釈を同じ度合いで解釈するはずである。しかし、英語の母国語話者に聞けば、ほとんどの人が前者の解釈を好む。これはkillとknifeの間に強い共起関係があることによると思われる。意味主観確率だけを優先度とするのでは人間が前者を好むことを表現することができない。

ある解釈が意味的にもっともらしいかどうかとその解釈の出現頻度が高いがどうかとは別問題である。曖昧性を解消するために、さらに、言語表現における単語の共起頻度を表現する必要がある。本研究では、意味確率を定義する。格フレームにおいて、ヘッドがXで、格スロット名がLで、格スロット値がYである時、その格スロットの意味確率を以下のように定義する。

$$P(X|L) \quad (4)$$

格スロットの意味確率を$P(X|L,Y)$にしないのは、そうすれば、膨大な量のデータからの推定が必要になり、現時点でその実現が困難だからである。

解釈の意味確率はその解釈の格フレームを構成する格スロットの意味確率の積であるとする。

格スロットの意味確率を以下の手順で推定する。まず、ハイパーシソーラスから、あらゆる可能なヘッドと格スロット名のペアを集める。次に、コーパスを利用してヘッドと格スロット名のペアの出現頻度を統計する。ヘッドと格スロット名のペアの出現頻度に基づいて格スロットの意味確率を推定する。ヘッドと格スロット名のペアの出現頻度を統計する時に、以前の「窓」をあげる方法をとる。この方法では、間違って統計することが多い(表3を参照)。というのは、係受けの曖昧性を解消しないまま統計を行なっているからである。しかし、本研究の曖昧性解消法では意味確率だけを頼りに優先順位を決めていないので、意味確率の推定の誤差が大きくても曖昧性解消にそれほど悪い影響を及ぼしていない。

表3: 格スロットの意味確率の統計結果

	正解	タイプ2エラー	タイプ1エラー
P(名詞|前置詞)	40%	60%	0%
P(動詞|前置詞)	54%	37%	9%

次に、格スロットの意味確率の推定の仕方について述べる。実際、同じヘッドと格スロット名のペアが膨大な量のコーパスの中でもそう頻繁に表れてこないことがある。本研究では、Lapalace推定量を用いて格スロットの意味確率を推定する。Lapalace推定量とは、標本がNで、この内xであるものの数がN_xである時、xの発生する確率を

$$\frac{N_x + 1}{N + s} \quad (5)$$

とするものである。但し、sは確率変数のレンジの大きさであるとする。Lapalace推定量が多くの優れた性質をもっており、特に、データが不十分な時にそれを用いれば、真の分布に近い推定ができることがわかっている[竹内, 安倍 93]。

4 構文確率

意味主観確率と意味確率を計算して、優先度とするのは依然不十分である。例えば、

John phoned a man in Chicago. (6)

という文には二つの解釈がある。一つは「ジョンはシカゴにいる人に電話をした」で、もう一つは「ジョンはシカゴで人に電話をした」である。人間は前者の解釈を好む。前者の解釈の意味主観確率と後者の解釈の意味主観確率が同じであるし、前者の解釈の意味確率と後者の解釈の意味確率も同じである[4]。意味主観確率と意味確率だけを優先度とするのでは人間が前者の解釈を好むことを表現することができない。この場合、人間が前者の解釈を優先的に解釈するのは、人間の書いた文章には前置詞句が近くにあるものを修飾しやすいという傾向があるからである。このような傾向は構文的なもので、しかも、確率によって表現することができる。本研究では構文確率をも定義する。

まず、CFG規則(或は、拡張されたCFG規則でもよい)の確率を定義する。非語彙規則のCFG規則が$\alpha \rightarrow \beta$である時、条件付き確率$P(\beta|\alpha)$がその確率であるとする。語彙規則のCFG規則が$A \rightarrow a$である時、条件付き確率$P(A|a)$がその確率であるとする。

解釈の構文確率がその解釈の句構造を導出する際に適用したCFG規則の確率の積であるとする。例えば、式6の例文の前者の解釈の構文確率が以下である。

[4]少なくともある範囲において同じである。

$$
\begin{aligned}
&P(\text{解釈 } 1) = \\
&P(NP, VP|S) \times P(N|NP) \\
&\times P(V, NP|VP) \times P(NP, PP|NP) \\
&\times P(DET, N|NP) \times P(P, NP|PP) \\
&\times P(N|NP) \\
&\hline
&\times P(N|John) \times P(V|phone) \\
&\times P(DET|a) \times P(N|man) \\
&\times P(P|in) \times P(N|Chicago)
\end{aligned}
\tag{7}
$$

右辺では、非語彙規則の確率が解釈の句構造の頻度を表しており、語彙規則の確率が各単語がどれだけ句構造に関与しているかを表している。語彙規則と非語彙規則の確率を別々に定義するのは、こうすれば、語彙規則の確率が推定しやすくなるからである。非語彙規則の確率と語彙規則の確率はそれぞれ独立した事象の確率として考えられている。

非語彙規則の確率の推定を以下のように行う。まず、意味主観確率と意味確率を優先度とする場合正しく優先順位を付ける[5]ことのできない例文を集める。次に、それらの例文を解析する。曖昧性が生じた場合、人間が正しい解釈を選ぶ。次に、機械が正しい解釈を導出する時に適用した各非語彙規則の適用回数を数える。各非語彙規則の適用頻度を基に各非語彙規則の確率を推定する。

本研究の構文確率は [Fujisaki89][Jelinek90] 等の (構文) 確率とは異なる。

5 優先順位の決め方

この節では解析における解釈の優先順位の決め方について述べる。

解析は、構文解析と意味解析からなる。構文解析では解釈の句構造をつくる。と同時に、それぞれの解釈の構文確率を計算する。意味解析で句構造に対応する格フレームをつくる。と同時に、それぞれの解釈の意味主観確率と意味確率を計算する。構文、意味解析の後に、意味主観確率、意味確率、構文確率を基に各解釈の優先順位を決める。実際、意味主観確率と意味確率と構文確率をベクトルとしてまとめ、そのベクトルを優先度とする。解釈の優先順位を以下のように決める。

$$
\begin{aligned}
&[X1, Y1, Z1] > [X2, Y2, Z2] \\
&\text{if} \quad X1 > X2 \\
&\text{else if} \quad X1 = X2, Y1 > Y2 \\
&\text{else if} \quad X1 = X2, Y1 = Y2, Z1 > Z2
\end{aligned}
\tag{8}
$$

但し、X が意味主観確率を、Y が意味確率を、Z が構文確率を表す。

言語学では、Minimal Attachment Principle(以下では MAP) と Right Association Principle(以下では RAP) という原理が人間の英語解析における曖昧性解消 (特に PP Attachment の曖昧性の解消) の時に働くと言われる [Frazier79][Hobbs90]。MAP によれば、修飾語 (Modifier) が項 (Argument) になる解釈が修飾語が付加詞 (Adjunct) になる解釈より優先的である。RAP によれば、後位修飾語 (PostModifier) が一番近いヘッドを修飾する解釈が優先的である。さらに、MAP が RAP より優位であると言われる [Hobbs90]。

[5]その方法について後に述べる。

本研究の優先度の計算法は言語学における以上の原理を実現している。意味主観確率と意味確率の大きさで解釈の優先順位を決めるのは MAP を実現しており、構文確率の大きさで解釈の優先順位を決めるのは RAP を実現している。さらに、式 8の優先順位の決め方が MAP が RAP より優位であることを実現している。

この優先度計算法では、まず意味主観確率を基に優先順位を決めるので、慣用表現の解釈がより優先的になる。確率だけによって優先順位を決める場合、慣用表現の解釈が前に出る保証がない。

6 実験

以上の曖昧性解消法を利用して、英語の解析を行った。パーザーは SAX[松本 86] を利用している。コーパスは Wall Street Jornal[6]の記事 ($300MB$) を利用している。辞書は The Oxford Advanced Learner's Dictionary[7]を変換したものを利用している。また、拡張された CFG 規則 (非語彙規則) を 110 個定義した。

まず、動詞概念と名詞概念 (合わせて 180 概念) のもつ格フレームを約 1000 フレーム集め (機械が自動的に抽出したのもあるし、人間が抽出したのもある)、その一般化を行なった。それから、コーパスを利用して意味確率の推定を行なった。次に、コーパスからランダムに単文を 360 文抽出し、構文確率の推定を行なった。なお、構文確率の収束が観測されなかった。

格フレームの一般化、意味確率の推定、構文確率の推定のいずれにも利用した単文 (トレーニング文) を 40 文集め、その解析を行なった。また、新たに単文 (未トレーニング文) を 50 文集め、その解析も行なった。トレーニング文と未トレーニング文を解析する時の曖昧性解消の結果を表 4 に示す。解析の誤りは主に意味確率の統計の誤り、不十分な構文確率のトレーニングに起因する。なお、解析する文の平均長さが 8.9 単語であり、文平均曖昧性の数が 21.5 であった。

さらに、未トレーニング文を解析する時、意味主観確率だけによる優先度の計算と、意味主観確率と意味確率だけによる優先度の計算で曖昧性解消を行なった。それぞれの結果を表 5 に示す。意味主観確率の計算が曖昧性解消にとってもっとも重要であることがわかった。

実験を通じて、本研究で提案する曖昧性解消法が非常に有効であることがわかった。

7 おわりに

本稿では一般化された実例と確率を用いた曖昧性解消法を提案した。さらに、それによる曖昧性解消の実験についても述べた。

本曖昧性解消法には以下の特徴がある。

- 一般化された実例で意味知識を表現している。実例の追加とその一般化という単純な操作で意味知識を追加することができる。しかも、実例は一般化された形で蓄えら

[6]ACL/DCI CD-ROM 1。

[7]Oxford Text Archive。

表 4: 実験結果 1

	正解が一位になる率	正解が上位三位内に入る率
トレニーング文 (40 文)	85%	98%
未トレニーング文 (50 文)	80%	94%
未トレニーング文に対し意味知識を追加登録後	84%	94%

表 5: 実験結果 2

	正解が一位になる率	正解が上位三位内に入る率
意味主観確率による優先度	73%	88%
意味主観確率と意味確率による優先度	75%	90%

れているため、その適用範囲がはっきりするし、意味的なもっともらしさの計算コストも低い。

- 意味主観確率、意味確率、構文確率を定義し、言語学の理論に基づいて優先度を計算している。より人間に近い優先度計算法で曖昧性を解消する点で従来の確率だけによる曖昧性解消により優れている。
- 一般化された実例と確率を用いているため、知識を統一した枠組で表現することができ、優先度計算の整合性が保たれている。この点で普通のヒューリスティクスによる曖昧性解消より優れている。

従来の実例による曖昧性解消と確率による曖昧性解消のメリットを生かし、欠点をなくし、うまくそれらを統合したのが本研究の曖昧性解消法であるといえる。今後は、ハイパーシソーラスの学習を研究していく予定である。

謝辞

本研究の機会を与えてくださったNEC C&C研究所システム基礎研究部の中村勝洋部長に感謝いたします。本研究を進めるにあたって、有益なコメントを数多くいただいたNEC C&C研究所システム基礎研究部の安倍直樹主任、竹内純一氏等多くの方々に感謝します。さらに、SAXを提供してくださった奈良先端科学技術大学院大学の松本裕治教授、実験用データを提供してくださった方々にも感謝します。

参考文献

[Chitrao90] M. V. Chitrao, R. Grishman, *Statistical Parsing of Messages, Proc. of DARPA Speech and Natural Language Workshop, 1990.*

[Frazier79] L. Frazier, J. Fordor, *The Sausage Machine: A New Two-Stage Parsing Model, Cognition,6.191-325, 1979.*

[Fujisaki89] T. Fujisaki, F. Jelinek, J. Cocke, E. Black, T. Nishino, *A Probabilistic Parsing Method for Sentence Disambiguation, International Parsing Workshop'89, 1989.*

[Hindle91] D. Hindle, M. Rooth, *Structural Ambiguity and Lexical Relations. ACL91. 1991.*

[Hobbs90] J. R. Hobbs, *Two Principles of Parse Preference, COLING90, 1990.*

[Jelinek90] F. Jelinek, J.D.Lafferty, R.L.Mercer, *Basic Methods of Probabilistic Context Free Grammars, IBM Research Report, 1990.*

[Lakoff87] G. Lakoff, *Women,Fire,and Dangerous Things, What Categories Reveal about the Mind. The University of Chicago Press, Chicago and London, 1987.*

[李 92a] 李航, 曖昧性解消のための優先度計算法：優先度 = 意味ファジイ ∧ 意味確率 ∧ 構文確率, 情報処理学会自然言語処理研究会,*92-10,1992.*

[李 92b] 李航, ハイパーシソーラスとその学習, 情報処理学会自然言語処理研究会, *92-11,1992.*

[李 92c] 李航, ハイパーシソーラス: 意味知識の表現モデル, 自然言語処理シンポジウム, 電子情報通信学会, 日本ソフトウェア科学会, *1992.*

[Magerman91] D. Magerman, M. Marcus, *Pearl: A Probablistic Chart Parser, International Workshop on Parsing Technology, 1991.*

[松本 86] 松本裕治, 杉村領一, 論理型言語に基づく構文解析システム *SAX*, コンピュータソフトウェア, *Vol.3, No.4, 1986*

[Miller92] G.Miller, R.Beckwith, C. Fellbaum, D. Gross, K. Miller *Introduction to WordNet: An On-line Lexical Database, Anonymous FTP, internet:clarity.princeton.edu , 1992.*

[荻野 87] 荻野綱男, シソーラス作成の問題点, 言語, *Vol.6, No.5, pp.64-71, 1987.*

[Su88] K. Su, J. Chang, *Semantic and Syntactic Aspects of Score Function, COLING88, 1988.*

[竹内, 安倍 93] 竹内純一, 安倍直樹, *Laplace* 型推定量の確率的 *PAC* 学習モデルによる性能評価, 信学技報,*IT92-128, 1993.*

[浦本 91] 浦本直彦, 長尾確, 一般辞書から抽出した事例を用いた制約と選好, 人工知能学会全国大会, *1991.*