計算機はどこまで言葉を学習できるのか？

鳥澤　健太郎

■ 研究のねらい

言語の学習の問題は、自然言語処理あるいは言語学において大きな問題とされてきた。日本語や英語などの自然言語を計算機に処理させる方法論を研究する自然言語処理の分野においては、自然言語の文法や、単語の性質を自動的に学習する研究が行われてきた。これは文法や単語の性質が非常に多岐にわたり、それらを処理するプログラムを人手では書ききれないという見通しがあったからである。この言語の多様性に関わる問題、特に単語に関する問題は、インターネット上で自然言語で書かれたドキュメントが大量に利用可能になった昨今の状況において、よりクローズアップされる傾向にある。また、現代言語学の祖とされる Chomsky は人間の言語習得を可能たらしめる機構として「普遍文法」の概念を提案し、普遍文法の研究こそが言語学の本来の目的であると主張した。

本プロジェクトでは、このような状況をふまえ、言語の学習を大量のテキストを用いて計算機に行わせることを目標とした。この狙いを実現するにあたって、言語の学習が図１にあるような二つのステップに分割できるという仮定をおき、そのおのおのについて計算機にそれぞれの学習を行わせるという計画をたてた。

ステップA．語彙学習　英語、日本語といった個別言語に関して、ある程度の文法や品詞レベルでの単語の性質が与えられたという仮定から出発し、単語、特に名詞の持つ意味的な情報、知識をテキストから学習するプロセス。

ステップB．文法学習　第一のステップで前提となった文法を、より仮定の少ない状況から学習するプロセスである。より具体的には、単語の品詞（例えば、名詞や動詞）が与えられていない状況で、英語や日本語などの様々な言語に共通の性質を表現している「普遍文法」から出発し、与えられた単語列から、単語の品詞を推定しつつ、文法の学習を行う過程である。

実際に人間が言語を学習するときには、これらのプロセスが同時並行してすすむ可能性ももちろんある。場合によっては順序が逆転するであろう。しかしながら、そのようなことを考慮に入れたとたん、言語の学習が手に負えないほどむつかしい問題になってしまうのも事実である。学習プロセスのこのような分割は、研究を進める上での作業仮説であると考えてもよい。また、本プロジェクトでは、使用する文法の表現方法として主辞駆動句構造文法（Head-driven Phrase Structure Grammar, HPSG）とよばれる形式を採用して研究を進めた。

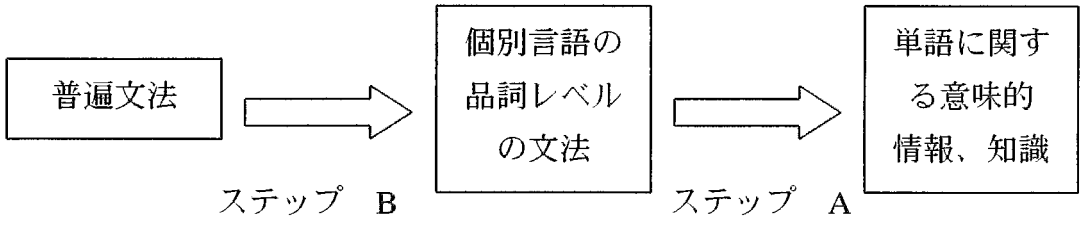

図１．言語学習における二つのステップ

■ 研究成果

本プロジェクトの主要な成果は、前述の言語学習の二つのステップに対応する、単語に関する意味的情報・知識の学習手法、普遍文法から出発した文法の自動学習の方法、並びに、本プロジェクトで利用した文法表現形式であるＨＰＳＧの構文解析器の性能向上である。以下ではこれらの成果各々について具体的にみていくこととする。

１．単語に関する意味的情報・知識の学習

１．１ 研究の動機

日本語には３０万語以上の単語があるといわれている。究極の自然言語処理ではこれらの多数の単語を意味的に処理しなければならない。３０万語の単語に関する辞書を作成するというのは並大抵のコストで可能となるものではない。また、さらに問題なのは、どのように辞書をつくれば、自然言語を使いこなすプログラムで利用可能になるのかがはっきりとはわかっていないということである。現在存在している辞書、シソーラスなどでは３０万語以上の見出しを持つものも存在するが、プログラムで利用できるような有用な情報を十分には提供してはくれない。例えば、次のような文が新聞の見出しに現れたと考えてみよう。

「磐田、４ゴール」

スポーツに関心のある人であるならば、「磐田」というのは都市としての磐田ではなく、磐田にあるサッカーチームであるところのジュビロ磐田であるということに気づくはずである。しかしながら、手元にある辞書をみても、磐田がジュビロ磐田というサッカーチームを指すとはかかれていない。また、そもそも、先の見出しの意味がわからないからといって辞書を引く人もあまりいないであろう。普通は意味がわかるようになるまで、続く記事を読むはずである。逆の言い方をすると、この見出しの意味するところがわからなかったとしても、見出しに続く記事を読むことで、磐田というのはサッカーチームであると学習するのである。単語の持つ意味的情報、知識に関する自動学習の研究を行う動機というのは以上の例でおわかりいただけるかと思う。もう少し正確に言い直すと、次の２点である。

工学的動機：現在ある辞書などでカバーされていないような単語の意味的性質を多数の語に関して明らかにしたい。

理論的動機：人間は意味のわからない単語があったとしても、前後の文脈からその意味を推測する能力をもつ。その過程に対する計算論的なモデルを構築したい。

１．２ 手法

具体的な研究内容は、大量のテキストが与えられた時に、そこに現れる単語の意味的性質を自動的に明らかにするプログラムを作成するということになる。この「意味的性質」というのはよくわからない概念であるが、暫定的に単語の「意味的な分類」のことだと考えてほしい。本プロジェクトで開発した学習手法が一体どのような意味的性質を学習できたのかについては後ほどより詳細に議論する。とりあえず、作成すべきプログラムというのは、磐田という単語が、大阪や名古屋と同じクラスに属すると同時に、マリナーズや巨人などと同じクラスに属する単語であることを判定するプログラムである。

本プロジェクトで開発した手法は、ＨＰＳＧとその統計的な構文解析器を用いて大量のテキストを

$$P(x,y) = \sum_{c} P(x \mid c)P(y \mid c)P(c) \qquad \text{(A)}$$

$$P(x,y,z) \stackrel{c}{=} \sum_{a,b} P(x \mid a)P(y \mid b)P(z \mid a,b),P(a,b) \qquad \text{(B)}$$

図２．単語分類のための確率モデル

解析して単語間の係り受け関係を抽出し、その結果に Expectation Maximization (EM) 法と呼ばれる統計的手法を適用することで単語の意味的分類を計算するというものである。例えば、ＨＰＳＧとその統計的な構文解析器が次のような文を与えられたとする。

「磐田は鹿島に連敗した。」

構文解析の結果としては次のようなデータが生成される。

〈磐田,〈は,連敗する〉〉,〈鹿島,に連敗する〉,〈磐田,鹿島,〈は,に,連敗する〉〉

このようなデータを大量の文に関して集めて学習データとし、それに統計的手法を適用するわけである。ちなみに、〈は,連敗する〉などのように助詞と述語の組を「補語位置」とよび、〈は,に,連敗する〉〉などのように二つの助詞と一つの述語の組を「述語テンプレート」とよぶ。ここで重要なことは、この手法では、学習に先立って意味に関する知識を天下りに与えるということは一切していないことである。与えられたのは解析時に必要となる文法だけである。しかも、与えられた文法も、各々の単語固有の性質についての記述はごく一部の機能語をのぞいて持っておらず、意味的な情報も一切含んでいない。これらを考えるに、本手法は教師なし学習であるとみなすことができる。

統計的手法の適用に当たっては、図２にあるような確率モデルを仮定する。(A)式は、単語 x と補語位置 y が文中に組となって出現する確率を表している。また、右辺にある変数 c は単語の分類(クラス)を表す ID である。例えば、1000 個のクラスからなる単語の分類を考えたければ、c の値として、1 から 1000 までの整数を考えておけばよい。また、同様に（B）の式は単語 x およびｙが述語テンプレート z とともに出現する確率である。変数 a,b は単語 x,y に対応する単語クラスである。EM 法による単語分類では、以上のような確率モデルを考えたときに、学習データが出現する確率が最大になるようにパラメータ、つまり、(A),(B)式の右辺にある確率、例えば P(x|c)の値を自動的に調整する。

ちなみに(A)式のみを用いた EM 法による単語分類手法については先行研究がなされている。本プロジェクトの単語分類手法は、式（B）を加えることによって述語テンプレートも単語分類で考慮されるように拡張されている。この拡張は一見それほど大きなものとはみえないが、応用に関する節で述べるように、これによって、本手法は単なる単語分類手法ではなくなり、より単語の持つ意味に接近した知識を学習する手法と見なせるようになった。

以上の方法により、実際に新聞 14 年分から学習された単語の意味分類の一部を図 3 に示す。各クラスは、その ID 例えば、43 が与えられたときに、確率 P(x | 43)が大きい単語 x をリストアップすることによって得られている。この例から見るかぎり直感に近い意味的な分類がなされているが、この図にない他のクラスにおいても、意味的に妥当な分類が行われている。例えば、前述した「磐田」に関しても、都市のクラスとスポーツチームのクラスにそれぞれ大きな確率で属している。また、図には現れていないが、単語分類と同時に、ある補語位置、あるいは述語テンプレートがどのようなクラスの単語と組になって現れやすいかが確率として推定されている。これらは通常、文法、あるいは辞書

クラス 43	確率	クラス 869	確率	クラス 1730	確率
日立製作所	0.597	エアーニッポン	0.562	酒	0.486
日本ビクター	0.558	エアシステム	0.448	お茶	0.444
シャープ	0.513	ANK	0.423	コーヒー	0.331
三菱電機	0.502	航空	0.412	日本酒	0.317
東芝	0.493	空輸	0.370	ビール	0.288

図３．得られた単語分類の例

の一部とされる「格フレーム」と呼ばれる記述の確率での表現となっている。

しかしながら、上で言っている「直感にあっている」ということを客観的に評価できるかという疑問、あるいは、単語の意味的性質を明らかにする学習といっておきながら、結果は単なる分類にすぎないではないかという疑問も生じるであろう。これらの点について次のセクションで議論をしたい。

１．３ 応用ならびに議論

さて、これまでに説明した手法によって単語の分類が得られた訳だが、これらが何らかの意味で現実世界での単語の持つ意味をどのように反映しているかについての検討を行わなくてはならない。そこで、実験によって得られた単語の分類と格フレームを使い、自然言語処理において必要とされる次のようなタスクを行った。そのタスクとは、二つの名詞Ａ、Ｂを含む名詞句「ＡのＢ」を考え、ＡとＢの間に生じる意味的な関係を推測するというものである。典型的な関係としては、所有関係（ＡがＢを所有する）位置関係（ＡはＢにある）といったものがそれである。これまでの自然言語処理ではこのような関係を少数のプリミティブで分類をし、処理を行ってきた。しかしながら、次のような名詞句を考えると名詞間の関係を少数のプリミティブで置き換えるという前提がそもそも間違っていることに気づく。

英語の先生

アサヒビールのスーパードライ

レストランのビール

これらはそれぞれ、「英語を教える先生」「アサヒビールの作っている（販売している）スーパードライ」「レストランで飲むビール」といった意味を表していると考えることができ、また、これらの関係を所有、位置関係といった少数のプリミティブで表現するということが難しいということがわかる。可能性としては、すべての動詞の数だけのプリミティブを用意しなければならないかもしれない。また、もうひとつ重要なことは、このタスクを行うためには、単語の意味的性質が分かっている必要があり、これをある程度の精度でこなせるとすれば、我々の手法で計算された単語の分類と格フレームが意味的に妥当だとみなせる可能性があるということである。

黒橋らは「ＡのＢ」という名詞句から、名詞同士の関係を子供用の辞書に書かれた単語の定義と人手で作成された単語分類から推測するという研究を行っている。出力される関係は前述のように述語でしか表現できないものが多数含まれている。これとほぼ同じタスクを辞書、あるいは人手で書かれ

イチローの本塁打	＝イチローが放つ本塁打	着物の女性	＝着物を着る女性
数学の教授	＝数学を教える教授	ローソンのビール	＝ローソンで販売するビール
数学の学生	＝数学を学ぶ学生	パブのビール	＝パブで飲むビール
テレビのオペラ	＝テレビで見るオペラ	玄関の車	＝玄関に止まる車

図４．得られた単語分類を元にした意味的関係の推定の例

た単語分類ぬきで行うというのが、本プロジェクトの単語分類手法の応用として設定したタスクである。具体的には、「英語の先生」という名詞句から「英語を教える先生」という名詞句を生成するタスクとなる。手法の詳細については本報告ではふれないが、最終的に得られた成功例を図４に示す。精度に関して言えば、本手法では新聞からとられた「ＡのＢ」の形を持つ名詞句のうち、およそ３／４に対してなんらかの意味的な関係を推定することができたが、それらのうち妥当であったものは約60％程度であった。この値はそれほど高いとは言えないが、意味的関係の候補として５個までの出力を許すと、そのうちには妥当な関係が８５％以上で含まれていた。

最後にこのタスクの持つ意味合いについてより詳細に検討してみたい。問題は本プロジェクトで得られた単語意味分類ならびに格フレームの持つ意義である。本研究での主張は、得られた単語意味分類ならびに格フレームが十分意味の一部を切り取っているというものである。先ほど、同じタスクを子供用辞書を用いて行う研究について言及したが、子供用辞書が単語の持つ意味の少なくとも一部を表しているということについては、それほど議論の余地はないであろう。本研究での主張の端的な根拠のひとつは、そのような辞書を使用して行うべきタスクを、本手法では自動学習された単語分類ならびに格フレームの確率的表現のみで、ある程度の精度で達成したということである。

より具体的な例に則して言えば、本手法では「アサヒビールのスーパードライ」から「アサヒビールが生産するスーパードライ」という名詞句を生成できたが、これは「アサヒビールはなにかを生産するもの」であるという意味的な事実をふまえたものと捉えることができる。本手法では「アサヒビール」や、「スーパードライ」という単語自体が学習データ中でどのような出現の仕方をしているかを直接参照することはしていない。「アサヒビール」がクラスαに属する確率、「スーパードライ」がクラスβに属する確率、ならびにクラスαとクラスβの間に「生産する」という関係が成立している確率を用いているだけである。にもかかわらず、このような「言い換え」に成功するのは、クラスαがすくなくとも「生産するなにか」を含む単語クラス、クラスβが「生産されるなにか」を含む単語クラスになっているということを示している。これは例えば「キリンビールのスーパードライ」といった学習データには現れなかった表現が、単語分類中のクラスによる一般化によって同様に得られていることからもわかる。また人のために書かれた辞書との関係に関して言えば、仮に理想的な辞書が存在したとすると「アサヒビール」は「ビールを製造する会社」であり、「スーパードライ」は「ビールの商標」であるといったことが定義としてかかれているべきであろう。本手法ではそこまで詳細な知識ではないが、それに近いものを学習できたことになる。

さらに、我々の手法では辞書だけではカバーできないような名詞句も取り扱うことができる。例えば、「アメリカの車はよく故障する。」という文を考えてみよう。ここで「アメリカの車」は「アメリカで作られている車」という風に解釈すべきものである。実際に我々の手法ではこれに近い結果を得

ることができる。(「アメリカで生産する車」) しかしながら、この解釈はアメリカ、車というそれぞれの単語の辞書的定義からはでてこない。アメリカの定義が「何かを生産する場所」である、あるいは車の定義が「生産されるもの」であるとは考えにくい。これは辞書にある定義が、単語が指す概念を十分に「くくる」記述でなければならないからであり、アメリカを「何かを生産する場所」という表現でくくるのは無理があるからである。したがって、この例は辞書でカバーできる意味解釈の限界をしめすものと考えることができる。辞書の有効性を否定するものではないが、実際に自然言語の文を解釈する際には、このような単語の意味の「定義」からだけで得られない、「常識」とでもよぶべき意味的知識が必要であり、今回提案した単語の意味分類を自動学習するアルゴリズムはそのような知識のすくなくとも一部を学習できたことになると考えている。

1.4 単語の意味分類の自動学習に関するまとめ

従来より、自然言語処理において統計的学習手法は一定の役割をはたしてきた。しかしながら、従来研究においては、人手で注釈を加えられたテキストからの「教師あり学習」として利用されることがほとんどであり、「教師なし学習」としての利用はごく少数の例にとどまってきた。意味に関する知識を取り扱う「教師なし学習」と見なせるものはさらに少数であったといえる。これらを考えるに、本プロジェクトにおける単語の意味に関する学習は今後の研究の方向性を考える上でも重要な成果であると考えている。また、関連する研究として WWW 上の表など、文以外のメディアから単語の分類を自動学習する研究、さらには、今回学習によって求めた単語分類あるいは関係する意味的な知識と、人手で書かれた対応物との比較などを現在進めている。

2. 普遍文法から出発した文法の自動学習

本プロジェクトにおけるもう一つの主要な研究は、日本語、英語などの個別の言語に特有の性質を捨象し、自然言語一般の性質を記述した「普遍文法」から出発して、日本語、英語などの個別言語に関する文法を自動学習する方法に関するものである。普遍文法は本報告の冒頭でも触れたように、言語学者によって提案された概念であり、言語学の目指す目標とされるものである。人間の幼児は個別の言語を習得する以前にこの普遍文法をもっており、これをもとに個別言語の学習をおこなうとされている。この普遍文法の存在を仮定した言語習得の計算機上でのモデルに関しては先行研究が存在するが、それらの研究では学習データが品詞列である、あるいは、品詞レベルの単語に関する辞書が存在するなどの現実的ではない仮定をおいている。一方、言語習得の本来のモデルとしては、音声言語・信号から出発するべきであるが、現状では困難すぎる研究課題である。本プロジェクトでは、これらの両極端の中間の仮定すなわち、単語列からの文法学習を問題と
して設定した。結果として、本研究では以下のものを出発点として学習モデルを構築することになる。
入力1：普遍文法の記述　**入力2**：学習データとなる単語列
また学習モデルで学習すべきものとして考えているのは次の二つである。
出力1：普遍文法内にあるパラメータの具体的な値　**出力2**：各単語に対する品詞の割り当て
普遍文法にあるパラメータとは、言語共通の性質ではない個別言語の性質、例えば語順などを指定するものである。普遍文法ではこのパラメータの値が具体的には定まっていない。したがって、普遍文法をそのまま使うと、英語、日本語といった個別言語を分け隔てなく生成できることになる。学習においては入力である個別言語の文をもとに、これらのパラメータの値がきまっていき、これらのパ

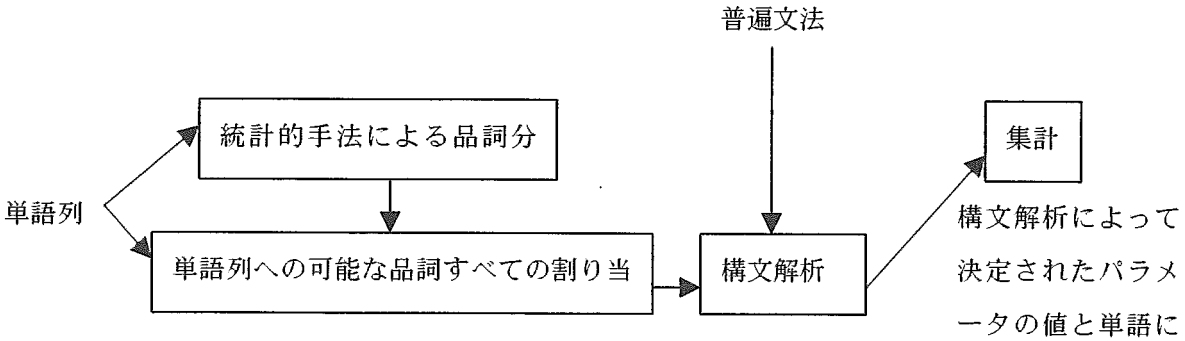

図５．学習プロセスの流

ラメータの値が具体的に定まった後の「普遍文法」は個別言語の文法と見なすことができる。本研究では、使用した文法記述であるＨＰＳＧの中にこのパラメータを埋め込み、普遍文法として機能するような文法を記述した。この文法で構文解析木を生成すると、埋め込まれたパラメータの値がその木を生成するのに必要な値にセットされる。この値の設定は、ＨＰＳＧの構文解析で使われる単一化という操作によって自然な形で自動的になされる。また、異なるパラメータ間には複雑な相互作用があり、この相互作用を考慮して具体的なパラメータの値を決定しなければならないが、このような方式によってパラメータ間の相互作用が統一的に扱えるようになった。

本プロジェクトで提案する学習の枠組みでは、制約を解くプロセスとして学習をとらえる。まず、普遍文法は単語列に対する一種の制約と考えることができる。とりあえず、語順は未定であるが、動詞と名詞が存在し、動詞はその主語、補語として、名詞句をその周囲にとる。さらには、一定の性質を満たす単語として、冠詞、助動詞、助詞といった機能語も存在し得るが、言語によってはそれらの品詞を含まないかもしれないといった制約がそれである。先に、本研究で使用した普遍文法の記述を使って構文解析木を生成すると、その木を生成するのに必要なパラメータが定まると書いたが、このパラメータの値は、もともと普遍文法が持っていた制約と、与えられた単語列に関する制約の両者を満足する値と見なすことができる。一方、具体的な単語の性質である品詞の割り当てに関しても、統計的方法によって単なる単語列だけから緩い制約を求めることができる。例えば、ある単語は名詞である可能性が高いが、別の単語は冠詞、助詞、助動詞などの機能語である可能性が高いといった制約である。この統計的手法においては、各単語の前後の文脈や、一般的に動詞の出現頻度は名詞よりも低いということ、機能語はどのようなテキストであれ、高頻度で一様に出現するといった言語一般の性質を利用している。これらの性質を利用して求まった単語列に関する緩い制約と、普遍文法に書かれた制約とをあわせて解くことで、機能語の品詞を一意に定めることができる。まとめると、本研究における文法学習は、普遍文法に内在する制約と、単語列に現れる制約の両者を解き、それらの制約を満足するパラメータ、品詞割り当てを計算することで進んでいく。

本研究で提案する学習のプロセスは図５にある。まず、入力された単語列には統計的品詞分類手法が適用されるが、これは各々の単語をI)動詞、II)名詞、III)冠詞、助詞などの機能語の３つのクラスに分類する。前述した単語列だけから求まる、品詞割り当てに関する制約というのはこの分類のこと

である。ついで、単語列中の各単語はこれらの分類に従って、可能な品詞を割り当てられる。ここで機能語のクラスに分類された単語には機能語に含まれるすべての品詞が単語に対して割り当てられる。つまり、一つの単語に複数の品詞が同時に割り当てられることになる。例えば、英語の場合であれば、単語”the”には助詞、冠詞、助動詞などのすべての機能語の品詞が割り当てられる。次いで、得られた品詞列をもとに普遍文法を用い構文解析をおこなう。前述したように構文解析時に、普遍文法内にあるパラメータの値が決定されていく。また、”the”などの機能語に関して言えば、theが助動詞であるような、誤った品詞割り当てを元にした構文解析木はその生成が失敗する可能性が高い。逆に言えば、構文解析の成功の前提となるような品詞割り当てを集めていけば、機能語の品詞が決定できることになる。

言語学習プロセスの最後は、構文解析時に成功裏に生成された構文解析木から、具体的な値となったパラメータ、およびその前提となった品詞割り当てを多数の単語列に関して集計することである。統計的手法による品詞分類には誤りも存在するし、仮に正しい品詞分類が与えられたとしても、その品詞列が、例えば英語の文法と日本語の文法の両方で成功裏に構文解析されてしまうというケースもある。言語学習プロセスの最後にあたる集計モジュールはこのようなノイズを取り除くために存在する。

これまでに、日本語、英語を例としてパラメータの設定、機能語の品詞の決定に関する実験をおこなった。日本語を約４千文入力した場合には、パラメータとして日本語型の語順が設定された構文解析木のみ生成された文が約 600 文、英語型の語順を設定する構文解析木のみ生成されたものが約 200 文であった。後者の構文木が生成されたのは一見奇妙であるが、これは統計的手法の品詞への分割で生じたエラーが原因になっているものが多い。英語の場合はこの割合がほぼ逆転した結果となっている。これらの結果を元に優勢な語順を正しいパラメータとすれば、この二つの言語については語順が正しく決定されたことになる。　ついで、この語順に関するパラメータが決定されたとして、機能語に関する品詞を決定する。実験では、英語の”the”,”in”、あるいは日本語の「が」「を」などの基本的機能語に対して、その品詞が正しく決定されることがわかった。

当然のことながら、普遍文法を巡る探求は言語学者によって現在も続いており、現在決定版とよべるものは存在しない。したがって、本研究で出発点とした普遍文法はいわば暫定版にすぎないということになる。実際に実験で使用した文法で、扱えていない言語現象は多々ある。従って、研究の主眼は、言語習得の正確なモデルを構築するというよりは、暫定的な普遍文法を与えたときに、その文法と単語列が相互作用することによって、個別言語の性質を学習すると同時に単語の品詞を推測する計算の枠組みを構築するという段階にとどまっている。しかしながら、このような計算の枠組みをつくることによって、仮説として生み出された普遍文法が、学習に対してどの程度寄与できるかをチェックすることが可能になった。本研究はこの点で意義があると考えている。

３．ＨＰＳＧパーザーの性能向上

以上述べてきた２つの研究と平行して、プロジェクトで使用したＨＰＳＧを用いる構文解析器など、ツールに関する改良、整備もおこなった。具体的には、統計的曖昧性解消機構の導入、構文解析の高速化などである。

■ 今後の展開

３年間にわたる本プロジェクトの期間においては、当初の目標であった言語習得に関する完結したモデルを構築するには至らなかった。現状では、語彙学習で前提とした文法と文法学習で学習された文法の間には大きなギャップが存在する。しかしながら、自然言語処理研究者であれば誰でもがぶつかり、また多くの場合に避けて通ってきた２つの大問題、「意味、あるいは常識はどのようにとらえたらよいか？」、「単語のもつ多様性を計算機は学習できるのか？」に関して、少なくとも「とっかかり」を作ることができたと考えている。大量のテキストからの単語分類の教師なし学習によって、構文解析や形態素解析といった比較的単純なタスクではなく、人のために人手で書かれた辞典、つまりは単語の意味的定義を必要とするタスクを、ある程度の精度で実際に行うことができたことは予想していなかった成果であった。別の言い方をすれば、純粋に客観的な概念である単語の出現頻度が、主観的とされる意味（あるいは少なくともその一部）との間に一定の相関をもっているということは、ある程度予想されていたこととは言え、驚きであった。これらの結果は、今後の研究にとって原動力となると考えている。より具体的にいえば、単語に関する学習での成果を、これまでに整備してきたＨＰＳＧを用いるためのツール群などと組み合わせ、対話システムにおけるユーザーの意図の認識、あるいは、テキストの自動要約といったような困難な課題、特にこれまで「計算機が常識を持たないが故にできないとされてきたタスク」の処理方法の実現に結びつけたいと考えている。

また、ここまでで言及してきた言葉の意味はあくまで言語の中で閉じた相対的なものであり、言語の一側面でしかないテキストのみを扱うものであった。言葉に関するポピュラーな捉え方である、世界内に存在する対象を指し示すものとしての言葉、あるいはコミュニティー内で共有され、生成されるものとしての言葉については最後までふれることがなかった。これはプロジェクト開始時にそのような作業仮説をおいたからであり、そのような見方を否定する考えはない。むしろ、本プロジェクトの結果をふまえ、言語内で閉じた学習における成果を、言葉に対するこのような見方と結びつけるのも興味ある研究課題であると考えている。

■ 成果リスト

- Kentaro Torisawa. A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases, In Proceeding of Workshop on Automatic Paraphrasing: Theories and Applications, to appear, 2001
- Kentaro Torisawa. An unsupervised method for canonicalization of Japanese postpositions, In Proceeding of 6th Natural Language Processing Pacific Rim Symposium, to appear, 2001.
- Minoru Yoshida, Kentaro Torisawa, Jun'ichi Tsujii. A method to integrate tables of the World Wide Web, In Proceeding of the International Workshop on Web Document Analysis (WDA 2001),2001.
- Minoru Yoshida, Kentaro Torisawa, Jun'ichi Tsujii. Extracting ontologies from World Wide Web via HTML tables, In Proceeding of Pacific Association for Computational Linguistics 2001, 2001.
- Kenji Nishida, Kentaro Torisawa, Jun'ichi Tsujii. Compiling an HPSG-based grammar into more than one CFG, In Proceeding of Pacific Association in Computational Linguistics 2001, 2001.
- Takashi Ninomiya, Kentaro Torisawa, and Jun'ichi Tsujii_An Agent-based Parallel HPSG Parser for

Shared-memory Parallel Machines, 言語処理学会学会誌, Vol 8(1), pp. 21・48, 2001.

- Kentaro Torisawa, Kenji Nishida, Yusuke Miyao, and Jun-ichi Tsujii . An HPSG Parser with CFG Filtering in Journal of Natural Language Engineering, Cambridge University Press, Vol 6(1), pp. 63--80, 2000.
- Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuisi, Jun'ichi Tsujii. A Hybrid Japanese Parser with Hand-crafted Grammar and Statistics, In Proceeding of the 18th International Conference on Computational Linguistics, pp. 411-417, 2000.
- 金山博、鳥澤健太郎、光石豊、辻井潤一。3つ以下の候補から係り先を選択する係り受け解析モデル、言語処理学会学会誌、 Vol 7(5), pp. 71・92, 2000.
- Minoru Yoshida, Takashi Ninomiya, Kentaro Torisawa, Jun'ichi Tsujii. Efficient FB-LTAG parser and its parallelization, In Proceeding of Pacific Association for Computational Linguistics '99, pp. 90-103, 1999.
- Kenji Nishida, Kentaro Torisawa, Jun'ichi Tsujii. Efficient HPSG parsing algorithm with array unification, In Proceeding of Natural Language Pacific Rim Symposium '99, pp. 144-149, 1999.
- Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuisi, Jun'ichi Tsujii. Statisitical dependency analysis with an HPSG-based Japanese grammar, pp. 138・143, 1999