社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
TL99-41 (2000-03)

帰納的学習を用いた自然言語処理の有効性について

荒木 健治　　棚内香次
北海道大学大学院工学研究科電子情報工学専攻

〒060-8628 札幌市北区北13条西8丁目
E-mail : araki@media.eng.hokudai.ac.jp

あらまし これまで我々は実例よりそこに内在している規則を獲得する帰納的学習を用いた自然言語処理手法を提案し，種々の応用を行ない各応用段階における有効性を確認してきた．しかし，研究の最終目的に沿った考察を行なっていなかった．そこで，本稿では，最初に本研究の目的について述べ，次に本手法を形態素解析，かな漢字変換，機械翻訳，対話処理に応用した結果について簡単に述べる．さらに，各応用を比較・検討し，統括することにより本研究の最終目的にとって，どこまでが明らかになり，現在何が課題であり，今後どのような方向に研究を進めるべきかを統一的に考察し，帰納的学習の自然言語処理における有効性と今後の研究課題について述べる．
キーワード 帰納的学習，自然言語処理，言語獲得，生得的能力，ヒューリスティクス

Evaluation of Effectiveness for Natural Language Processing Using Inductive Learning

Kenji ARAKI　　Koji Tochinai
Graduate School of Engineering, Hokkaido University

N13-W8, Kita-ku, Sapporo 060-8628, JAPAN
E-mail : araki@media.eng.hokudai.ac.jp

Abstract

We have proposed Natural Language Processing Using Inductive Learning Method, and applied it to a various kinds of tasks. We have confirmed its effectiveness each application from the results of their applications. However we do not consider them from point of view of the goal of our research. Therefore, in this paper, we describe the goal of our research, and the outline of their applications. Moreover, we compare their results and consider them from point of view of the goal of our research. As a result, we acquire what our proposed method makes clear, and the future problems and directions of our research. We describe these considerations in this paper.

key words inductive learning, natural language processing, language acquisition, innate capability, heuristics

1 はじめに

ここ数年間の自然言語を取り巻く環境の変化は劇的である．高度情報化社会の進展が，一般人が母国語以外の言語と遭遇する機会を増大させ，膨大な量の文書が日々電子化され流通している．このような状況から，人間の主要なコミュニケーション手段である自然言語を計算機で高度に利用するという課題は，緊急のものとなっている．このような試みは計算機が世に出現した初期の頃から存在したが，現在の状況から十分に成功したとは言い難い．

1980年代に，解析的なアプローチが提案された[1]．このアプローチでは，予め与えられた知識に基づいて処理が行なわれる．したがって，予め想定された曖昧さの無い状況にしか対処できない．このようなアプローチは頑健性が低いので，開発者が想定した言語現象に対しては良好に動作するが，想定し得ない状況には対処できない．このような言語現象は，自然言語処理システムが動作する実在の世界では頻繁に出現するので，結果としてユーザの要求を満足できない．また，このようなアプローチでは，未知の言語現象が出現した場合，人手により新たなルールを追加する必要がある．1990年代に入ってこの問題を解決するために，用例に基づくアプローチ[2]や統計的モデルに基づくアプローチ[3]が提案された．用例に基づくアプローチでは，人間が例を用いて問題を解決する過程を模倣し，大量の例より与えられた入力文と類似性の高い例の一部を組み合わせることにより自然言語を解析する手法であり，主に機械翻訳で多くの応用が行なわれている[4]．また，用例より統計的に格情報などを獲得することにより解析を行なうという手法も提案されている[5]．これらのアプローチは頑健性が高く，精度の向上のためには新しい実例を追加するだけで良く保守，更新が容易であるという利点がある．しかし，精度良く処理を行なうためには，非常に大量のコーパスを必要とする．また，構文解析結果などを利用する場合には構文解析ツールの精度や構文タグ付きコーパス作成の労力などに問題がある．したがって，このようなアプローチで自然言語処理システムを作成する場合，はじめに非常に大量のコーパスを収集する必要がある．近年，電子化された文書を収集するのは比較的容易となっているが，開発するシステムが対象とするタスクと整合性を有する構文タグ付きなどの良質なコーパスを収集することは容易ではない．

このような問題を解決するために，我々は帰納的学習を用いた自然言語処理手法を提案し，種々の応用システムを開発してきた[6]．本手法は頑健性が高く，しかも膨大な量のコーパスを必要としない．また，字面レベルで学習を行なっているので，構文解析，意味解析などの解析ツールや構文タグ，意味タグなどを付加する必要もない．また，帰納的学習を用いた手法では，コーパスに基づくアプローチのように用例の部分部分を組み合わせたり統計的なアプローチのように確率を利用するのではなく用例中に内在しているルールを帰納的に獲得し利用する．本手法では，獲得されたルール中には正しいルールとともに多くの誤ったルールも存在する．したがって，正解を与えられた際にどれだけ多くの誤ったルールを削除できるかにその成否がかかってくる．また，獲得されたルールは，より抽象化される方向へ処理されるが，抽象化される各段階でのルールも記憶されるので，結果として抽象度の異なるルールが多段階に獲得され，これらのルールを抽象度の低いものより順に適用することにより処理が行なわれる．これは，抽象度の高いルールほど適用範囲は広いが精度は低く，逆に抽象度の低いルールほど適用範囲は狭いが精度は高いためである．本手法では，比較的少量の用例からでもルールを獲得することができるので，その時点で対象としているデータに適応したルールを獲得することができる．すなわち，文脈に依存した解を得ることができる．これは，学習する用例が限定されるのでその文脈に依存した解しか存在しないためである．この結果，その対象においては高い精度で処理をすることができるが，対象が変わると全く処理できないという事態に陥る．しかし，本手法ではこの問題を帰納的学習の学習機能による動的な適応により解決している．すなわち，本手法では対象が変化すると変化した対象の用例を自動的に学習することにより動的に対象に適応することができる．この結果，本手法に基づくシステムは高い精度で処理を行ない，かつ汎用性を有する．

これまで我々は本研究を進める上での各応用ごとの発表は行ってきたが，各応用での研究を比較し，研究の最終目的を達成するためにどこまでが明らかになり，現在何が課題であるのかを統一的に述べたことはなかった．そこで，本稿では最初に本研究の目的について述べ，次に本手法を形態素解析，かな漢字変換，機械翻訳，対話処理に応用した例について，そのアルゴリズムの概要と評価実験の結果について述べる．さらに各応用を比較，検討し，本研究の最終目的を達成するために本手法の自然言語処理における有効性と今後の研究課題を考察する．

2 研究の目的

本稿で述べる帰納的学習を用いた自然言語処理は以下のような研究目的から開発されたものである．

我々は人間が言語を獲得し，知識を学習していくメカニズムに興味を持ち，このメカニズムを解明して工学的に実現することを最終目的として研究を進めている．こ

のような観点に基づくシステムは，環境より情報を学習し，次第に成長することができる．したがって，本研究により，複雑で使いにくいシステムやプログラム作成の重労働といった高度情報化社会の種々の問題を解決し，人にやさしい柔軟なシステムを実現できるものと考えている．人間の幼児は，言語も知識も持たない状態から周囲の様々な環境より学習を行ない，言葉を話し，種々の知識を持つ大人に成長する．したがって，このような能力は確かに存在する．また，人間はこのような能力を生まれながらに有している．このような観点から，我々はこの研究の目的を「人間の言語及び知識獲得能力の工学的実現」としている．

この生得的な能力に関する研究は，いくつか存在するが，工学的に有効なシステムが完成するまでには至っていない [7] また，心理学の立場からの研究も多数行われているが [8]，これらの研究は，どのようにして子供が言語を獲得するのかを解明するという点では大きな意義があるが，そのような機構をどのようにして工学的に実現するかという問題は，当然ながら対象としていない．

3 基本的考え方

我々は「人間の言語及び知識獲得能力の工学的実現」の研究を行う際に，生得的能力を「二つの事物が同じか異なるかを判断する能力」と仮定した．この仮定の基で手法を開発し，その正当性を確認するための実験を行った．したがって，この能力から２種類の記号列の対応関係を獲得できなくては研究の目的に合致しない．ところで，人間は対応関係を有する未知の二種類の記号列に対してどのような方法を用いてその対応関係を決定しているのであろうか．表１に未知記号列からの対応関係抽出の例を示す．

表 1: 未知記号列からの対応関係抽出の例（１）

入力１	$\alpha\,\theta\,\sigma\,\psi\,\delta\,\lambda\,\vartheta$	
入力２	$\Xi\,\Sigma\,\psi\,\delta\,\Upsilon\,\Phi\,\Theta$	
セグメント１	$\alpha\,\theta\,\sigma$	$\Xi\,\Sigma$
セグメント２	$\psi\,\delta$	$\psi\,\delta$
セグメント３	$\lambda\,\vartheta$	$\Upsilon\,\Phi\,\Theta$

表 1[1] の入力１，入力２のような対応関係を有する二つの未知記号列を見た場合に人間は，まず二つの記号列に共通な部分を検出する．表１の例では，下線部 ($\psi\,\delta$) に注目すると考えられる．そして，その両側の差異部分をその出現順に対応付ける．すなわち，表１のセグメント１，２，３のような対応関係を考える．このような抽出過程は，我々が生得的な能力として仮定している「二つの事物が同じか異なるかを判断する能力」を用いて行なうことができる．しかし，二つの記号列に共通な部分は別として異なる部分は，表１以外に，対応関係が出現順ではなく出現順と逆順になる対応関係を考えることができる．ここで出現順の対応関係を取るか出現順と逆順の対応関係を取るかは，対象とする問題に依存したヒューリスティックスを用いて行われると考えられる．

[1] 表１で無意味な記号列を用いているのは，計算機上のシステムから自然言語を見た場合，解析するための背景知識や文法知識の存在しない状態では，システムが直面する状況はちょうど人間が表１のような記号列を見たときと同じ状態になるのではないかということを想定したためである．

表 2: セグメントからのプリミティブ抽出の例

セグメント１	$\alpha\,\theta\,\sigma$	$\Upsilon\,\Phi\,\Theta$
セグメント２	$\theta\,\sigma\,\gamma\,\mu$	$\Phi\,\Theta\,\Sigma$
プリミティブ１	α	Υ
プリミティブ２	$\theta\,\sigma$	$\Phi\,\Theta$
プリミティブ３	$\gamma\,\mu$	Σ

さらに，このようにして抽出されたセグメントから共通部分を抽出することにより，プリミティブな単位に分解することができる．表２にセグメントからの共通部分の抽出によるプリミティブ抽出の例を示す．表２に示したように，セグメント１，２の対応関係にある二つの記号列の各々の共通部分 ($\theta\,\sigma$, $\Phi\,\Theta$) をプリミティブ２として抽出し，その両側の差異部分 (α, Υ)，($\gamma\,\mu$,Σ) をそれぞれプリミティブ１，プリミティブ３として抽出する．このように，共通部分と差異部分を分離することにより三つのプリミティブを得ることができる．この三つのプリミティブを合成することにより元の二つのセグメントを得ることができるので，この三つのプリミティブがあれば，二つのセグメントは不要となる．このような抽出方法は，基本的には我々が生得的能力として仮定している「二つの事物が同じか異なるかを判断する能力」を用いて行なうことができるものであるが，そのほかにやはり上述したような出現順に対応関係があるというヒューリスティクスが使われている．このように実例より共通部分と差異部分を多段階に抽出することにより知識を獲得する手法は帰納的学習の一つである．なお，本研究におけるヒューリスティクスの導入の意味などについては，5.3 で述べる．

4　帰納的学習を用いた自然言語処理

我々は「人間の言語及び知識獲得能力の工学的実現」の研究を行う際に，対象データを段階的に高度化しながらその対象に耐え得る手法を提案するという手順で研究を進めてきた．具体的には，漢字かな混じり文という一種類の記号列，べた書き文とその漢字かな混じり文という記述体系が同じだが表層表現が異なる二種類の記号列，原文とその訳文といった意味的には非常に近いが記述体系が全く異なる二種類の記号列，記述体系は同じだが意味的に異なり因果関係が存在する二種類の記号列という４つの段階を対象とし，それぞれの段階に対して「帰納的学習を用いた形態素解析手法 [9]」，「帰納的学習を用いたかな漢字変換手法 [10]」，「遺伝的アルゴリズムを用いた帰納的学習による機械翻訳手法 [11]」，「遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法 [12]」を提案し，さらに実験によりその有効性を確認している．本章では，これらのアルゴリズムの概要とその性能評価実験の結果を簡単に述べる．

4.1　形態素解析手法

最初に対象としたのは，一種類の記号列である漢字かな混じり文である．この漢字かな交じり文から単語を認識し，獲得することが仮定した生得的能力により可能であることを確認した．これは，形態素解析手法に相当する．この形態素解析に対しては，従来の既知語の処理の中でその例外として未知語を処理する手法とは異なり未知語の処理の中でその例外として未知語を処理するという手法を提案した．その結果，従来の自然言語処理の問題である未知語を高い精度で処理することが可能となった．この「帰納的学習を用いた形態素解析手法 [9]」では，未知語の処理が基本なので辞書が全く空の状態から単語そのものを帰納的学習により獲得することができる．すなわち，本手法における帰納的学習ではテキスト中の漢字かな混じり文より共通部分と差異部分を多段階に抽出することにより語を獲得する．本手法の処理過程は二つの学習過程，二つの認識過程より構成される．処理過程を図１に示す．

辞書が全く空の状態から情報処理，機械工学，応用化学，人文科学の各分野に関する論文，合計４４編，総文字数 302,703 文字を用いて評価実験を行った結果，各分野の変化点で分野の変動による多少の精度の低下が見られたが，しばらく学習すると迅速に回復し，どの分野もおおよそ９０％程度まで認識できることが示され，仮定した生得的能力に基づく帰納的学習により単語の認識という段階まで言語獲得能力を工学的に実現できることが明らかになった．

図 1: 帰納的学習を用いた形態素解析手法の処理過程

4.2　かな漢字変換手法

次に，対象を高度化し，記述体系が同じだが表層表現が異なる二種類の記号列に対する適用を行なった．このような記号列としてべた書き文とその漢字かな混じり文を用いた．この「帰納的学習を用いたかな漢字変換手法 [10]」では，べた書き文とその漢字かな混じり文から帰納的学習によりかな漢字変換に必要な語の表記と読みを獲得し，次に獲得状況及び変換精度に基づく確実性の高い順に多段階に変換を行なう．処理の流れを図２に示す．

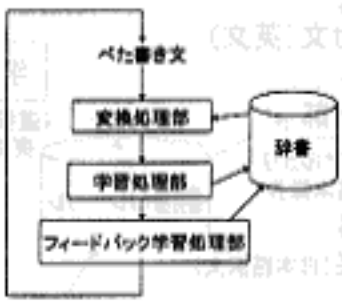

図 2: 帰納的学習を用いたかな漢字変換手法の処理過程

4.1 で用いた資料と同様の実験データを用いて本手法の性能評価実験を行った．実験条件も同様に辞書が全く空の状態から行っている．その結果各分野の変化点で一時的に正変換率が未知語の出現のために低下するものの学習能力により迅速に回復し，各分野で９０％以上の精度を示し，本手法のこの応用における有効性を確認した．

4.3　機械翻訳手法

次に，本研究を進める第三段階として原文とその訳文といった意味的には非常に近いが記述体系が全く異なる二種類の記号列を対象とした帰納的学習手法の開発を行った．対象としたデータは翻訳例，すなわち原文

とその訳文の組である．このレベルでの応用を行うには，対象としている二類の記号列の相違が大きく，これまでの帰納的学習手法の学習能力では不十分であった．そこで，遺伝的アルゴリズム [13] を帰納的学習に適用することにより帰納的学習手法の学習能力の強化を図った．このような手法を「遺伝的アルゴリズムを用いた帰納的学習による機械翻訳手法 [11]」と呼ぶ．また，この手法を GA-ILMT(Machine Translation method using Inductive Learning with Genetic Algorithms) と表記する．

本手法は，翻訳実例から翻訳ルールを帰納的に学習し，翻訳を行なうものである．学習型の機械翻訳手法では，一般に非常に大量の実例を必要とするという問題がある．しかし，本手法では遺伝的アルゴリズムの交叉処理を応用することにより少数の実例より多くの翻訳例を自動的に生成し，それらを用いて多くの翻訳ルールを得ることによりこの問題を解決している．また，システム全体としても遺伝的アルゴリズムを実現しているので，使用につれてその翻訳情報をフィードバックし，誤った翻訳ルールが淘汰され最適なシステムに進化することができる．本手法の処理過程を図 3 に示す．

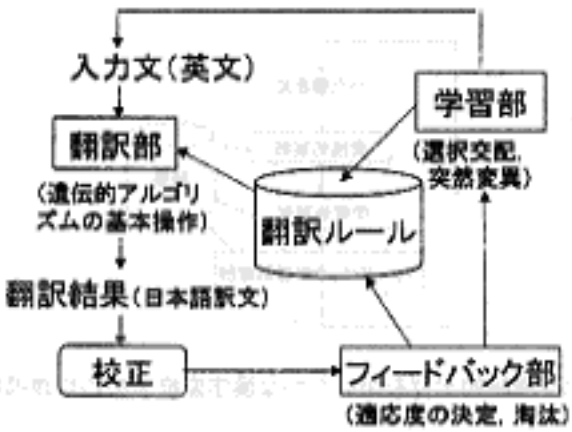

図 3: GA-ILMT の処理過程

また，本手法において染色体は翻訳例，遺伝子は翻訳例中の単語に割り当てられている．すなわち，染色体は英文とその日本語訳文を組とした翻訳例を，そして染色体を構成している遺伝子には翻訳例を構成している単語を割り当てている．また，交叉の方法としては，1 点交叉を用い，突然変異率を 2% としている．中学 1 年生用の教科書ガイドに出現する翻訳例 (1,010 組，11,479 文字) を用いて性能評価実験を行った結果，有効な翻訳率が 61.9% となり本手法のこの応用における有効性が確認された．

4.4 対話処理

次の段階として記述体系は同じで意味的には異なるが因果関係は存在する記号列に対する適用を行なった．このような記号列として対話例を用いた．本研究で対象としたのは，対話の中でも雑談である．雑談は，目的が達成されることよりは自然に対話すること，対話をし続けること自体に意味がある．ここで，雑談を対象としたのは本研究の最終目的である「人間の言語及び知識獲得能力の工学的実現」の最終段階である対話処理において，まず実現すべきなのは幼児の段階だからである．すなわち，幼児の段階ではまだそれほど十分に言語を獲得していないので，質問応答システムのような特定の情報を得る目的の対話は行なえず，まずは雑談から言語獲得のためのデータを収集すると考えたからである．

雑談を行なうシステムとしては，ELIZA[14] がある．ELIZA は精神科医が患者に行なうインタビューを代行するシステムで雑談を行なうことができる．ELIZA は複数のキーワードの組み合わせることにより文脈に依存したように見える応答文を生成する．また，アドホックな方法のみで動作し，意味を理解してはいないので，非常に頑健性の高い応答を行なうことができる．ELIZA では，ユーザが自分の興味のあることを深く話し出すとそれに追従して会話を継続することができず，うまく話題をそらして会話を継続しようとする．そのためユーザの満足度は低くなってしまうという問題がある．そこで，本研究では，この問題を解決するために「遺伝的アルゴリズムを用いた帰納的学習手法 (GA-IL)[11]」を雑談に応用した，すなわち，GA-IL によりシステムとユーザの対話例より応答文生成のためのルールを獲得し，それらを用いて応答を行なうことによりユーザの興味のある話題についても次第に深く話すことができる．このような手法を我々は，「遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法 [12]」と呼ぶ．また，この手法を GA-ILSD(Spoken Dialogue method using Inductive Learning with Genetic Algorithms) と表記する．本手法においては，対話例より応答文生成ルールを獲得するが，応答文生成ルールが十分に存在しない状態では，ELIZA が対話を継続し対話例を収集する．学習が進むにつれて獲得したルールを用いて応答を行なうことができるようになるので，ELIZA による応答が減少し GA-IL を用いた応答が増大する．本手法の処理過程を図 4 に示す．

実験システムを開発し実験を行なった結果，ELIZA のみのシステムと GA-IL に ELIZA を付加したシステムでは，正応答，準応答[3] の合計が，66.4% から 76.1% まで 9.7 ポイント上昇し，本手法のこの応用における有効性を確認することができた．

[3] 意味的に正しい応答ではあるが，表現が不自然なもの．

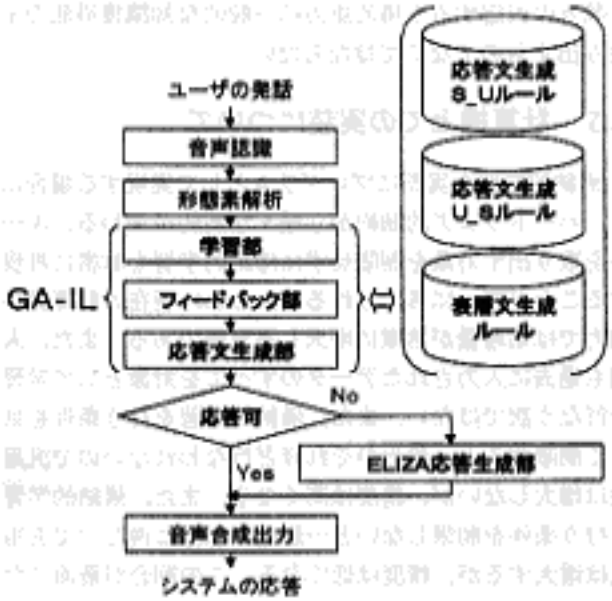

図 4: GA-ILSD の処理過程

表 3: 応用結果

応用分野	形態素解析
データの性質	一種類の記号列，記述体系同じ 意味的に同じ
精度	90%
ヒューリスティクス	複数回出現する文字列は語 漢字１字の差異部分は接尾辞 尤度評価関数 段落ごとに学習
応用分野	かな漢字変換
データの性質	二種類の記号列，記述体系同じ 意味的に同じ，表記が異なる
精度	90%
ヒューリスティクス	双方向解析，一字一音語の処理 尤度評価関数，語の階層化の条件
応用分野	機械翻訳
データの性質	二種類の記号列，記述体系異なる 意味的に非常に近い
精度	62% (GA 無し：52%)
ヒューリスティクス	尤度評価関数，交叉位置，突然変異率 淘汰の条件，70 文ごとに学習
応用分野	対話処理
データの性質	二種類の記号列，記述体系同じ 意味的に異なる (因果関係は存在)
精度	76%
ヒューリスティクス	突然変異率，交叉位置，選択の条件 応答文生成ルールの一致率

5 考察

生得的能力を「二つの事物が同じか異なるかを判断する能力」と仮定し，「人間の言語及び知識獲得能力の工学的実現」を目的として研究を進めてきた．対象データを段階的に高度化しながらその対象に耐え得る手法を開発するという手順で研究を進め，各応用で仮定した生得的能力から開発された帰納的学習手法が有効であることを確認した．本章では，これら各段階での応用を比較・検討することにより本研究の意義を明らかにし，今後の研究の方向について考察する．表 3 に各応用の結果を示す．

5.1 精度について

各段階での精度であるが，表 3 に示すように特に機械翻訳での応用においてかなり低下している．これは，ここで対象とした日英翻訳では言語間の相違が大きく，GA-ILMT で用いたような表層レベルの情報ではうまく対応が取れないこと原因であると考えられる．その結果，変換ルールが抽出できなかったものと考えられる．したがって，日英翻訳のような対象では，対応関係が明確になるレベルまで各文を解析しなければならない．また，対話処理においても 75% 程度の正当率なので，それほど高くはなっていない．これも帰納的学習が本来対応関係が存在することを前提としてルールを抽出するものであるが，対話例では表層レベルの情報にはその対応関係が明確に存在しないので，精度良くルールを獲得することができないためである．したがって，精度を向上させるためには，構文解析，意味解析を行い，その構文構造，意味構造を明らかにした上で，構文構造間，意味構造間で帰納的学習を行なうことにより機械翻訳，対話処理を行なう必要がある．ここで，構文解析，意味解析を行なうといっても本研究の目的を達成するためには，帰納的学習を用いて構文解析，意味解析を行なえなければ意味がない．そこで，現在帰納的学習を用いた構文解析手法及び意味解析手法を開発中である．

5.2 GA 導入の意味について

GA 導入の目的は，交叉処理による実例の自動生成とフィードバック処理の淘汰処理による実現である．すなわち，システム全体として GA を構成することにより学習により進化し，対象に最適化するシステムを実現している．実例の自動生成は，言語獲得の観点から考えると幼児が意味もわからず大人の表現の真似をし，いろいろな文を繋ぎ合わせて新たな文を生成する過程に相当すると考えることができる．すなわち，試行錯誤の過程である．また，幼児が誤った文を生成し使用した場合には，相手の大人に意味が通じない，おかしな反応をされたという経験をする．その経験からその文を生成する際に用いられたルールがおかしいということに気づき，以後そのようなルールを用いないようにする．このような過程が

淘汰処理に相当すると考えることができる．したがって，GA 導入は本研究の目的に合致している．また，その有効性についても機械翻訳，対話処理においてGA の導入によりそれぞれ10% 程度の向上が見られ，言語獲得能力の工学的実現という観点からもその有効性を確認することができた．

5.3 ヒューリスティクスの導入について

各段階において，仮定された生得的能力の他に表3に示すいくつかのヒューリスティクスを導入している．したがって，本研究の初期条件として仮定した生得的能力以外知識を何も与えないとしているが，ヒューリスティクスも知識の一つであるので，厳密には生得的な能力以外何も与えないということではない．しかし，これまで我々の行なった帰納的学習に関する他研究 [15, 16, 17] では，帰納的学習を用いた手法に各応用において必要な解析的な知識を直接与えることによりその精度を向上させているのに対し，4 章で述べた応用においては学習のためのヒューリスティクスを与えている[3] ので，知識を与えるといってもその意味が質的に異なる．このように与える知識のレベルは様々であるが，本研究の目的からするとこれらの知識は何らかの手段で獲得されなければならない．それは，他の人間から明示的に与えられるものかもしれないし，暗に与えられるものかもしれない．いづれにしてもシステムは，プログラムや辞書中に書き込むといった計算機特有の方法ではなく人間が日常生活で通常使用する手段を用いて与えられる情報より獲得しなければ本研究の目的に合致しない[4]．

このようなことを考えると本来ヒューリスティクスは生得的能力から派生するもので，生得的能力が個々の対象ごとに特化して発現したものでなければならない．このようなことが仮定した生得的能力から可能なのかということは非常に重要な問題であるが，同時に非常に大きなテーマでもあるので今後研究を進めることとし本研究においてはすでに獲得されたものとして予め与えている．

5.4 言語獲得能力の位置付け

Chomsky らは言語獲得のために必要な機能を人間が生得的に有していると考えている [18]．また，Anderson[19]，綿見 [7] らは一般的な学習の仕組みが存在し，それが特定の形で発現したものが言語獲得能力であると考えている．我々の立場は後者である．すなわち，言語もまた知識の一つであるので，一般的な知識獲得能力を用いて言語を獲得したという立場である．したがって，最終的には我々の仮定する生得的能力は一般的な知識獲得能力を生み出すものでなくてはならない．

5.5 計算機上での実装について

帰納的学習を実際にプログラムとして実現する場合には，ハードウェア的制約から様々な制約が加わる．ルールを取り出す対象を制限せずに帰納的学習を厳密に再現することが第一に考えられるが，これは現在の計算機の能力では処理量が急激に増大し不可能である．また，人間も過去に入力されたデータのすべてを対象として学習を行なう訳ではない．また，帰納的学習を行う条件を厳しく制限すると一般化がそれほど行なわれないので汎用性は増大しないが，精度は高くなる．また，帰納的学習を行う条件を制限しないと一般化が急激に進むので汎用性は増大するが，精度は低くなる．この割合が最適になるように帰納的学習を実現する必要がある．4 章で述べた各応用ではこの点を考慮して各応用に最適な種々の帰納的学習を実現している．

また，計算機の処理能力は一定なので，帰納的学習を計算機上に実現する際には帰納的学習が行われる条件を制限して学習の対象とするデータを多くする場合と逆に条件を制限せず学習の対象とするデータを少なくする場合が考えられる．どちらを行なうかは対象に依存する．例えば，機械翻訳では原文と訳文は意味的には非常に近いので，各単語の間にはある程度普遍的な対応関係が存在する．したがって，帰納的学習起動の条件を制限せずよりルールを一般化する方向へ学習を行ない，その代わり対象とする学習データを少なくする．この場合に帰納的学習起動の条件を制限せず抽象的なルールを多く持つことにより汎用性を保ち，学習対象データを制限して文脈に依存した解のみを持つことにより精度を高くしている．しかし，対話処理のように対話例の各発話が意味的に同じではなく因果関係が存在するという場合，各単語間に普遍的な対応関係は存在しない．その対応関係は文脈に依存して決定される．したがって，このような対象においては帰納的学習起動の条件を制限し，その代わり対象とするデータを多くする．この場合には多くのデータから厳しい制限を加えてルールを抽出するので確実性の高いルールしか抽出されず精度良く応答文を生成することができる．また，汎用性も対象とするデータを大量にすることにより具象的ルールを多く持つことで補っている．4.4 と 4.5 で述べた各応用においてはこのようにして精度を向上させている．

6 おわりに

本稿では，これまでの自然言語処理の研究の概要を述べ，その問題点を解決するために開発された帰納的学習

[3] 対話処理の GA-ILSD については ELIZA を含んでいるので，この部分には解析的な知識を直接与えているが，GA-IL の部分には学習のためのヒューリスティクスを与えている．
[4] 本研究においてはその手段を自然言語に限定している．

を用いた自然言語処理の有効性について述べた．本研究の目的である「人間の言語及び知識獲得能力の工学的実現」に向けて生得的能力を「二つの事物が同じか異なるかを判断する能力」と仮定して対象データを高度化しながら応用システムを作成してその有効性を確認した．このことにより帰納的学習を用いた自然言語処理の有効性が実証された．さらに，本手法の意義を各対象での応用結果を比較・検討することにより明らかにした．精度については，対象とする記号列の間の対応関係が曖昧になるにつれて精度の低下がみられた．これについては，帰納的学習を用いて構文解析，意味解析を行なうことにより対応関係が明確になるところまで解析を行なう必要があると考えられる．また，帰納的学習における GA 導入の意味を試行錯誤の過程に相当することを述べた．さらに，ヒューリスティクス導入の意味，言語獲得能力の位置付け，計算機上での実装の仕方についても考察した．

今後の課題としては，構文解析，意味解析といった解析レベルを深くする方向へ帰納的学習を適用し，さらに解析された結果を用いて種々の応用システム作成し，その有効性を確認することが挙げられる．また，帰納的学習の結果と学習に用いる実例の乱雑さとの関係を統計的な手法を用いて定量的に評価することにより，帰納的学習により理論的に抽出可能なルールの上限を明らかにし，帰納的学習を用いた自然言語処理が実在するデータにどの程度有効なのかを理論的に評価するという研究も予定している．

参考文献

[1] 野村浩郷：自然言語処理の基礎技術，信学会 (1988).

[2] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S. : A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol.16, No.2, pp.79-85(1990).

[3] Church, K. W. and Mercer, R. L. : Introduction to the Special Issue on Computational Linguistics Using Large Corpora, *Computational Linguistics*, Vol.19, No.1, pp.1-24(1993).

[4] 佐藤理史：アナロジーによる機械翻訳，共立出版，東京 (1997).

[5] 宇津呂武仁，松本裕治，長尾真：二言語対訳コーパスからの動詞の格フレーム獲得，情処学論，Vol.34, No.5, pp.913-924(1994).

[6] 荒木健治：学習に基づく自然言語処理，信学技報，SS 99-14, pp.33-40(1999).

[7] 綿見美貴子：言語を獲得するコンピュータ，共立出版，東京 (1998).

[8] H. H. Clark and E. V. Clark : *Psychology and Language*, Harcourt Brace Jovanovich(1977).

[9] 荒木健治，桃内佳雄，栃内香次：帰納的学習による語の獲得および確実性を用いた語の認識，信学論 (D-II), Vol.J75-D-II, No.7, pp.1213-1221(1992).

[10] 荒木健治，高橋祐治，桃内佳雄，栃内香次：帰納的学習を用いたべた書き文のかな漢字変換，信学論 (D-II), Vol.J79-D-II, No.3, pp.391-402(1996).

[11] 越前谷博，荒木健治，桃内佳雄，栃内香次：実例に基づく帰納的学習による機械翻訳手法における遺伝的アルゴリズムの適用とその有効性，情処学論，Vol.37, No.8, pp.1565-1579(1996).

[12] 木村泰知，荒木健治，桃内佳雄，栃内香次：遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法の提案，第 59 回情処全大，Vol.2, pp.403-404(1999).

[13] Goldberg, D. E. : *Genetic Algorithms in Search, Optimisation, and Machine Learning*, Addison-Wesley(1998).

[14] Weizenbaum, J. : ELIZA - A Computer Program for the Study of Natural Language Communication Between Man And Machine, *Communications of the Association for Computing Machinery*, vol.9, No.1, pp.36-45(1966).

[15] 笹岡久行，荒木健治，桃内佳雄，栃内香次：帰納的学習を用いた訳語推定手法の派生語および複合語における有効性の評価，信学論 (D-II), Vol.J81-D-II, No.9, pp.2146-2158(1998).

[16] 工藤晃一，荒木健治，桃内佳雄，栃内香次：学習型機械翻訳手法に適用された遺伝的アルゴリズムにおける知識による制約の有効性について，信学論 (D-II), Vol.J82-D-II, No.11, pp.2035-2047(1999).

[17] 松原雅文，荒木健治，桃内佳雄，栃内香次：文字情報縮退方式を用いた帰納的学習によるべた書き文のかな漢字変換手法，情処学研報，98-NL-128, pp.1-7(1998).

[18] Chomsky, N. : *Reflections on language*, Pantheon Books(1975).

[19] Anderson, J. R. : *The architecture of cognition*, Harvard University Press(1983)