知能と複雑系 117−16
(1999. 7. 23)

確率文脈自由文法を用いたエージェント間での言語共有

武田 稔也　東条 敏
北陸先端科学技術大学院大学
石川県能美郡辰口町旭台１－１
0761-51-1699 (1367)
ttakeda@jaist.ac.jp　tojo@jaist.ac.jp

言語獲得のモデル化は，認知科学における大きなテーマであり，原型言語から，進化，適応により，普段我々が使用する真の言語を獲得する過程においては自動化というスキルが存在する．本稿ではマルチエージェント環境での確率文脈自由文法 (ProbabilisticContext-Free Grammar,PCFG) の文法規則の共有化を自動化のプロセスの一部と考える．ここでは，各文法規則に確率を割り振り，会話における有用な規則を重視することにより，エージェント間での文法の共有化を目指す．実験には大規模並列計算機を使用し，各プロセッサ間での発話，認識という経験的学習によりチューニングし自己組織化を行うと同時に，仮説推論により文法規則を新たに生成し，共有文法を組織化するモデルの定式化を行う．

キーワード　マルチエージェント，帰納論理プログラミング，確率文脈自由文法

Large-Scale experimentation of autonomous grammar sharing

Toshiya Takeda　Satoshi Tojo
Japan Advanced Institute of Science and Technology , Hokuriku

1-1, Asahidai, Tatsunokuchi, Ishikawa 923-1292, JAPAN
0761-51-1699 (1367)
ttakeda@jaist.ac.jp　tojo@jaist.ac.jp

Abstract

Modeling language acquisition is one of important theme in cognitive science, where we can find the skill called 'automatization' in the process of language evolution from primitive form to mature language. In this paper, we propose a model of grammar sharing using PCFG as well as inductive logic programming. In that process, each processor exchanges utterances and recognizes the other, and the probability for the common grammar rules are tuned to be shared. We design an experimentation for a large-scale parallel computer.

key words　multi-agent , Inductive Logic Programming, Probabilistic Context-Free Grammar

1 はじめに

言語獲得のモデル化は, 認知科学における大きなテーマである. 大人のピジン語, 2 才以下の子供の言葉にみられる原型言語から, 進化, 適応により, 普段我々が使用する真の言語を獲得する過程において, 自動化 [1] というスキルが存在する. ここで自動化とは, 発話における文法規則の選択など, 試行錯誤的な注意を必要としていたパートを経験的学習により省略するプロセスを指す.

本稿ではマルチエージェント環境での確率文脈自由文法 (Probabilistic Context-Free Grammar,PCFG) [2] の共有化を自動化のプロセスの一部と考える. そして, 各エージェントの持つ PCFG の文法規則の木構造に確率を与え, これをエージェント間での発話, 認識という経験的学習によりチューニングし自己組織化する. また, 仮説推論 [3] により文法規則を新たに生成し, 共有文法を組織化することも同時に行い, 例えば慣用句や複合語の生成といった言語獲得過程のモデル化の実現を目標とする.

Lieberman によれば調音能力は複雑な調音器官のすばやい流れるようなコントロールを必要とし, そのためには自動化による注意の省略が必要だとしている. 彼は同様に人間の持つ統語装置の起源においても自動化と結び付いた仮説を提唱している [4].

エージェントアプローチによる言語に関する他の研究として, 橋本 [5] らはチョムスキー階層をもとに生成される文字列の共有を行っている. しかし, これは文法規則をもとに生成された文字列を, 文字列認識の段階での形態素解析を行っていないため, 語彙の共有が目的と言える. 本研究は語彙セットを先験的に与えた上での文法セットの共有化を目指しており, また, PCFG を用いている点で [5] と異なる. 小野 [6] らは仮説推論を用いた共有文法の組織化についての研究を行っている. 本研究でも仮説推論を用いるが, 本研究の立場はあくまでも PCFG の共有化であり, その上での必要な共有文法の獲得を目的としている. また, この研究は言語学的な正当性に基づくものであると同時に, 異なるプロトコルを持つ人工エージェント同士が柔軟に新たなプロトコルを自己組織化する

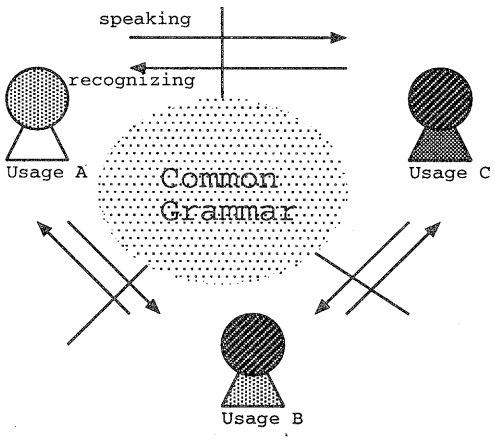

図 1: 本研究の自然言語におけるモデル

複雑系のシミュレーションの側面を持つ. 実装には, これまでの同様な研究とは異なり並列の大型計算機を用い, 各々の計算機に各エージェントを割り当てた大規模環境でのシミュレーションを行う.

本稿では以下の構成をとる. まず 2 章では, 本稿で提案するモデルの文法構造について述べる. 3 章ではエージェントの有する推論機構について述べる. そして 4 章でモデルの定式化を行い, 最後に 5 章では本モデルを計算機上で実装するにあたり, 超並列計算機での大規模並列化についての考察を行う.

2 文法の枠組

各エージェントは異なる文法規則を持ち, 他のエージェントとのコミニュケーションから, 帰納推論により新しい文法規則を生成する. ここでは, エージェントの持つ文法について考察する.

- 仮定 1
  エージェントの有する文法規則は文脈自由文法の形態をとる.

仮定 1 は, 自然言語の文法理論の多くが句構造を基にした文脈自由文法を基礎にしているためであ

る．また，文脈自由文法の終端記号である語については，最初に設定しておく必要がある．ただし，ここでは自然言語で用いられている語を取り扱う必要はない．便宜上，本システムでは2進数で表された数字列の並びを一つの語と見なす．

- 仮定2
  語彙は異なる文法規則を有するエージェント間においても共有されているものとする．

本モデルを異言語間でのコミニュケーションのモデルとして考えると，仮定2は奇妙に感じるかも知れない．しかし，現実のコミニュケーションにおいては，単語の意味はボディランゲージや指し示しなどを用いればある程度伝達可能であるし，各エージェントが辞書を保有していたと考えれば問題ないであろう．また，本稿での目的は文法の共有化であるため，簡単のためこのような仮定を行う．

次に，相互理解について考察した場合，エージェント間で互いの言語が理解されたことを主張するためには，各エージェントの持っていた意志と目的達成のための語用論的考察が本来必要である．しかし，意味の表示や意思・目的の表現を十分に準備することは現実的に困難である．しかし，仮定2にあるように文法の共有化を目標とするなら，その枠内で最低限コミニュケーションが成立する条件を定義することは，十分な意義を持つものである．本稿では以下の立場をとる．

- 仮定3
  エージェントの発話は，エージェントの持つ文法規則から確率的に選択されるものとする．

- 仮定4
  エージェントの発話は他のエージェントの持つ文法規則により構文解析できたとき，理解されたとする．

本モデルのエージェントが初期状態において持つ文法は，文に相当するトップカテゴリーをS，日常言語から類推可能なカテゴリー名，NP, VP, AdvなどをX_iで表す．各エージェントからは他のエージェントが持っている文法セットを直接見ることはできず，それによって生成される文を交換するのみである．この過程において，エージェントは他のエージェントの発話を理解できるように新しいカテゴリー・新しい文法規則を推論し，それを自分の文法セットに追加する．

本モデルでは，エージェントが持つことのできる文法の最大数は，パラメータとして設定される．仮定1，仮定3にあるようにモデルは文脈自由文法の文法規則に重みをつけた確率文脈自由文法である．使用頻度の低い文法規則や，他のエージェントに認識されなかった文法規則に対する重みは減少していき，閾値以下になると不要な規則として捨てられる．このようにエージェント内部では，生得的な機構により発現した文法が変化して，後で出来る部分に順応していく．

3 帰納推論システム

概念学習をコンピュータのプログラムに行わせるためには，その形式化が必要である．概念学習の枠組みとして，例の一般化に基づく推論を行う帰納推論 [5] の枠組みを用いる．直感的にいえば，帰納推論とは演繹推論を逆転させることによって可能となる．すなわち，演繹推論の原型を二つの前提から（例えば三段論法によって）帰結を導く行為であるとするならば，帰納推論は前提の中の一つと帰結を与えて，もう一つの前提を導く行為に基づいていると考えられている．以下で帰納推論システムについて紹介する．

3.1 枚挙法による帰納推論アルゴリズム

枚挙法による帰納推論アルゴリズムは，可能なすべての仮説 (推測) $T_0, T_1, T_2, \ldots$を一つずつ提示し，全ての事例を説明できる仮説Tを探索するものである．つまり，新たな事例が観測されるたびに，それを説明できる一つの仮説を生成しようというものである．しかし，枚挙法では仮説を一つしか生成できないため，学習できるモデルは単純なものに限られてしまう．これに対して被覆集合アルゴリズムは一つの仮説で全ての例を説明しようとするのではなく，複数の仮説 (節) で例を説明しようというアルゴリズムである．

3.2 被覆集合アルゴリズム

被覆集合アルゴリズムでは, 与えられた背景知識 (仮説空間) と正例から例の一般化を試みる. つまり, ある正例から得られた仮説候補について, その無矛盾性, すなわちそれが負例を被覆していないことを調べ, さらにその仮説候補が弱過ぎないか検討を行う. ここで, 推測が弱過ぎるとは, 特殊化により, 本来説明されなくてはいけない例も, 説明されない状態をいう. もし, 新たな仮説候補を含む仮説の記述長に対して, 被覆する正例が少ない場合には, 学習の効果が得られているとはいえないので, このような仮説候補は採用されない. 以上の条件を満たした仮説候補, つまり説明力の強い仮説が見つかると, これを解 (仮説集合) に加え, それによって被覆される正例を元の正例集合から取り除き, 残された正例集合に対して, 同様の処理を繰り返す. これは正例集合が空になるまで繰り返される.

3.3 言語獲得への応用

機械学習の実現においては, その学習対象となる例や抽象概念の表現が必要となる. 一般に帰納推論システムにおいては, この表現方法として一階述語論理を用いられる. しかし, 本稿での学習対象は文脈自由文法の文法規則であるので, 表現方法として一階述語論理は使用しない.

また, 推論の方法については代表的なアルゴリズムを上で 2 つ挙げたが, 本稿では, 被覆集合アルゴリズム的立場をとる. 得られた仮説が過度に一般化されることを防ぐ方法として, 先に述べたように負例集合を被覆しないという条件がある. 本モデルでは, この他に 4.1 で述べる確率文脈自由文法を用いた方法の 2 つのアプローチにより一般化を行う. 具体的な仮説の生成方法については次章で述べる.

例の与え方については予め正例集合と負例集合を与えるのではなく, エージェント間のコミニュケーションにより逐次正例および負例集合に加えていく. ここで, 文脈自由文法の文法規則で与えられた例をどのように正例と負例に振り分けるかについては, 考察の必要がある. 例の与え方と仮説の提示方法, すなわちユーザインタフェースについては, 以下の 2 種類がある.

対話的システム システムが利用者に積極的に例を要求したり, 仮説を提示して正しさの検証を求めたりしながら仮説を洗練していく

経験的システム システムは利用者に積極的に例や仮説の検証を要求しない

本モデルは, システムが利用者に例や仮説の検証を要求しないという点では, 経験的システムと呼ぶことができるが, 実際に例を与えるのはエージェント間でのコミニュケーションであり, 仮説の検証を行うのもまた, エージェント間でのコミニュケーションによるものである. このため, 本モデルは「仮想的な対話的システム」と呼ぶことが出来るだろう.

4 モデルの定式化

4.1 発話

エージェントの発話は基本的に他のすべてのエージェントに対して行われる. 本来, 任意のエージェントが他の任意のエージェントにランダムに話しかけられるよう設定することが理想であるが, 本モデルにおいても十分長い時間をとれば, この本来の意図と同等の効果を期待できるため, 便宜上このように設定する. エージェントの発する文法規則は確率文脈自由文法 (以下 PCFG) により決定される.

PCFG を次の 4 つの組$< W, N, N^1, R >$で定式化する. ここで, W は終端記号の集合$\{w_1, ..., w_\omega\}$, N は非終端記号の集合$\{N^1, ..., w^\nu\}$, N^1は開始記号である. R は生成規則で, 本モデルではζ^jを任意の終端記号と非終端記号の列とすると$N^i \rightarrow \zeta^j$と定義する. これらの生成規則にはそれぞれ確率$P(N^i \rightarrow \zeta^j)$が存在する. これは非終端記号 N^iを展開する際, 全てのN^iに対する文法規則の中からζ^jを選択する確率である. ここで, 終端記号 w_kから w_lまでが N^iの下位カテゴリーとして存在するとき, N^iを $N^i_{k,l}$と書く. 図 2 の木構造文法について考察する. この文法規則を選択する確率は以下のように計算される.

$$P(figure2) = P(A \rightarrow BC)P(B \rightarrow w_1w_2)$$
$$P(C \rightarrow w_3w_4)$$

4.2 認識

エージェントは, 発話された文字列に対し形態素解析を行う. この段階では, 各エージェントの持つ文法規則の確率をもとにした形態素解析は行わず, 試行錯誤的に繰り返される. そして, 各エージェントの持つ文法規則に従って構文解析を行う. これに成功する, つまりコミニュケーションに成功した場合, この文法規則に対する確率を増やす. また, コミニュケーションに失敗した場合, 発話したエージェントの使用した文法規則の確率を減らし, 聞き手側のエージェントは, それを満たす文法規則を ILP の枠組みで新たに生成する.

例えば, 言語< 01111 >を受理する以下の文法規則を持つエージェント A がいたとする.

$$X_1 \rightarrow X_2X_3$$

$$X_2 = \{011, 010\}, X_3 = \{11, 00\}, X_4 = \{100\}$$

ここで, 言語< 01110011 >がエージェント B から発話されたとしよう. エージェント A が自分のもつ文法規則をもとに, この言語が非終端記号の並び X_2, X_4, X_3であることを認識する. しかし, エージェント A は X_4に関する文法規則は, $X_4 = \{100\}$ しか持っていないため, 構文解析に失敗し, 新しい文法規則を見つけようとする. そして, エージェント A の持つ背景知識と, 与えられた入力から以下の文法規則を新たに生成する.

$$X_1 \rightarrow X_2X_5$$

$$X_5 \rightarrow X_4X_3$$

エージェントの保有できる文法規則の最大数は PCFG に依存する. つまり, コミニュケーションに失敗し選択される確率の低くなった文法規則は淘汰され, 使用頻度の高い文法規則のみが保持される. また, 自らが保持していない全ての文法規則をインダクションによって生成すると過度に一般化されてしまい, 文法の共有化とは言えない. そこで, 負例によって過度の一般化を防ぐ必要がある. しかし, 言語を取り扱うというモデルの性質上, 正例と負例の判断基準を定めるのが困難である. そこで本稿では, 十分学習を行ったエージェントを教師とした, 一般の教師つき学習モデルとして取り扱う.

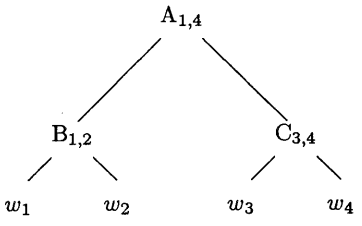

図 2: 木構造文法

5 大規模並列化の構想

本モデルを実装する計算機について検討を行う. 並列計算機を使用することにより, 得られる利益として処理速度の向上を考えるのは, ごく当然のことである. しかし, 本モデルで並列計算機を使用する目的は, 計算速度の向上よりも, 各プロセッサを一つのエージェントに見立てた大規模なシミュレーションにある. 純粋に計算速度の向上を目指す立場から見れば, 資源の無駄遣いに思えるだろう. しかし, マルチエージェントモデルを用いた言語共有において大規模なシミュレーションを行うことは, 実際の言語活動により近づくという意味では, 非常に意義のあるものである. また, 帰納推論自体非常に計算量の多いものであり, 各エージェントを各々のプロセッサで監視させた上で, 帰納推論などの処理を並列計算機にうまく割り振れば, 処理速度の向上も十分可能である. 以下に本モデルのシミュレーションを行う上で並列計算機上で実現したい事柄は以下の通りである (図 3).

(1) 各プロセッサでそれぞれのエージェントの動作を模倣する

(2) エージェント間のコミニュティとしてプロセッサ間の通信を行う

(3) エージェント内部での計算についても並列処理を行う

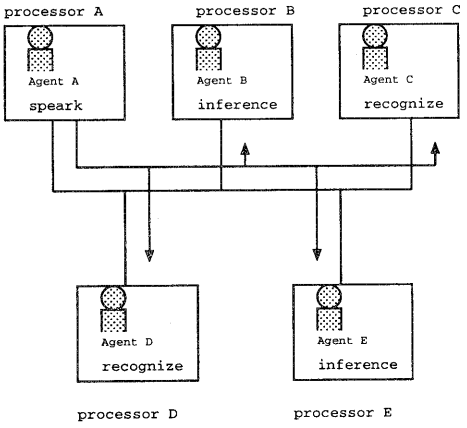

図 3: 並列化の概念

本学で利用可能な並列計算機のうち, メッセージ通信の可能な計算機として nCUBE3 がある. nCUBE3 は分散メモリとハイパーキューブネットワーク方式による MIMD 型アーキテクチャを採用した超並列システムで, 多数のプロセッサが各々プログラムとデータを持ち, 必要に応じて相互に通信を行いながら同時並列に処理を行うことができる. 本研究では, その機能性から nCUBE3 を採用する. ここで MIMD は複数命令複数データの略語である. また, 並列プロセッサを使用して実験を行う以上, その処理速度の検証も求められる. このため, 各エージェントを一つのプロセッサに割り当てる以外に, エージェントのもつ各パラメータについても並列化を試みる.

6 おわりに

本稿では, エージェント間での言語共有を目的とするモデルの定式化を行った. 実装に向けての展望として, 帰納推論における正例, 負例の判断基準がある. 文法共有の問題においては正例と負例の厳密に判断することはできない. 本モデルでは, 早期に多くの文法規則を学習したエージェントが, 他のエージェントに対して正例, 負例を与える教師的役割を果たすことにより解決されるが, この問題に対しては他の方法についても考察する必要があるだろう.

PCFG については, 文法規則に与えられた確率の評価方法が問題になる. 本モデルでは, エージェントの発話に対し, 他のエージェントが認識できるかどうかで, 確率の操作を提案したが, 例えば, GA などのような突然変異的なオペレータなどによる確率の操作も考えられる.

本稿で提案したモデルは幾つかの点で人工的な設定を行っているため, 自然言語のような複雑な現象のメカニズムを説明することは不可能である. 今後はこれらの人工的設定に対しても, より根拠のある (認知科学的に説明可能な) モデルに改善していくこと, そしてそれが良いモデルであることを主張するために実際の言語データから統計的検証を常に考えていく必要がある.

参考文献

[1] 大津由紀雄 他: 言語科学と関連領域, 岩波書店, 1998.

[2] Eugene Charniak: STATISTICAL LANGUAGE LEARNING, *TechBooks*,1993.

[3] 古川 康一: 帰納論理プログラミング— チュートリアル —, 人工知能学会誌 *vol.12, No.5*, pp655-664, 1997.

[4] Derek Bickerton 著, 筧 寿雄 監訳: ことばの進化論, 勁草書房, 1998.

[5] 小野哲雄, 東条 敏: 推論機能を有するエージェント群による共通文法の組織化, 人工知能学会誌 *vol.13, No.4*, pp546-559, 1998.

[6] Hashimoto, T. and Ikegami, T.: Emergence of net-grammar in communicating agents, *BioSystems 38*, pp1-14, 1996.