1D−6　ルール抽出手法のテキスト解析への適用

倉島 顕尚[1]　高須 淳宏[2]　安達 淳[2]
東京大学工学部[1]　学術情報センター研究開発部[2]

1　はじめに

データ処理システムを構築する際、処理の対象となっている事例の集合の性質をルールの形で取り出すことができれば好都合である。筆者らは、個々の事例はトークンの列によって構成されているとした上で、トークンの並びの集合から、その中に潜んでいる規則を抽出するための記号列からのルール抽出の問題について研究を行ってきた[1)]。この問題を、形式言語の獲得という視点から考えた場合、理論的には正例の集合から規則を同定するのは殆んど不可能である[2)]。それはルール抽出過程において、例の一般化の処理の限度を定める指標が存在しないことに原因がある。そこで、本発表ではルールの一般化を抑えるための評価法について考慮しながら、ルール抽出手法のテキスト解析処理への適用についての検討を行う。

2　記号列からのルール抽出のモデル

ルール抽出の基本モデルとして、図1に示すような記号列からのルール抽出という処理モデルを前提とする。記号列からのルール抽出とは、記号列の構成要素であるトークンが任意個並んだものを入力とし、トークンを終端記号とする文法記述を出力とするものである。文法記述を導出式の形で表した場合、部分列を表すための中間的な要素として変数が用いられるが、この変数はすべて計算機側が生成するものとする。その他、ここで仮定する制約条件は次の通りである。

- 入力として与えられる記号列は、正例のみとする。
- 文法の記述レベルは文脈自由文法クラスとする。
- ルール抽出に際して与えられた記号列が、すべて受理できるルールが正しいルールである。
- 記号列の集合から正しいルールを少なくとも一つ抽出する。
- ルールの抽出を補助する教師は存在しない。

Application of a Rule Extraction Method to Text Processing
Akihisa KURASHIMA[1], Atsuhiro TAKASU[2], Jun ADACHI[2]
[1] University of Tokyo
[2] National Center for Science Information Systems

3　応用モデル

3.1　応用モデルの定式化

基本モデルは、そのままではテキスト処理などの応用問題に利用できない。そこで、図2に示すような、応用モデルの定式化を行う。基本モデルでは、単に入力として与えられた記号列の集合を受理できるルールを抽出すれば良かったが、応用モデルでは目的に応じたルールの抽出を目指すことになる。ここでは、テキストの構造解析を行うためのルールの抽出を目標においている。構造解析を行うためのルールを抽出するには、与える入力の中に、構造解析のための情報を付加しなければならない。そのために、応用モデルにおける入力は、単なる記号列ではなく、記号列の部分に対応した属性情報を付加したものとなる。図3に示す例では、ある文献の著者名リストにおいて、各著者を表す部分列に author という属性が与えられている。この部分列に対応する属性のことをシンボルと呼ぶ。ルール抽出過程では、この部分列とシンボルとの対応をも加味したルールを抽出する。

3.2　応用モデルの実現手法

応用モデルの定式化に基づき、これを実現するルール抽出手法の処理過程を図4に示す。その内容は、まず、入力を基本モデルに合うように分解し、分けられた各問題毎に基本モデルの手法を用いてルールの抽出

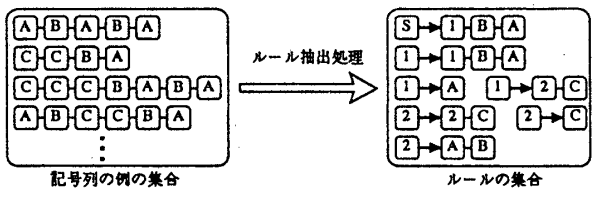

図 1: 記号列からのルール抽出のモデル

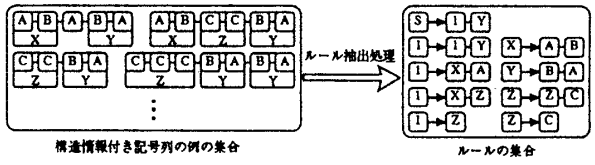

図 2: ルール抽出の応用モデル

Afsarmanesh, Hamideh / McLeod, Dennis / Knapp, David / Parker, Alice
└author┘ └author┘ └author┘ └author┘

図 3: 属性情報のついた記号列の例

を行い、最後にこれらをまとめ、全体として与えられた入力を満たすルールが抽出できているかの確認を行うというものである。検討の結果、この手法を用いれば、ルール抽出の際に与えられた正例を、すべて正しく処理できるルールが得られることが確認されている。

3.3 得られるルールについての検討

ルール抽出によって得られるルールの性質について考察する。応用モデルで得られるルールは、次の二つの条件によって、その性質が変わる。

- **与えられる入力データの構造** 入力データは、記号列と、その部分列に対応するシンボルの組合せという構造情報から成っているが、この構造がルールの性質に影響する。
- **基本モデルによって得られるルールの種類** 応用モデルの実現において用いている、記号列からのルール抽出処理で得られるルールの文法クラスが、最終的に得られるルールの性質に影響する。例えば、基本モデルで得られるルールが正則文法の範囲か、それを超えた文脈自由文法の範囲かなど。

3.4 一意性

一般に、文脈自由文法による解析結果は曖昧となる可能性があることが分かっており、これは応用モデルによって得られたルールについても同じことが言える。しかし、記号列に対してシンボルが対応づけられる部分を特定するという当初の目的に限定した場合、シンボルが対応付けられる部分が一意に決まるのであれば、ルール全体が曖昧であっても構わない。そこで、得られたルールに対して、一意性という性質を定義する。一意性とは、各シンボルの受理のルールを含んだ、入力全体を受理するルールに任意の入力を与えた時、常に入力の部分列とシンボルとの対応が一意に定まるというルールの性質を指す。

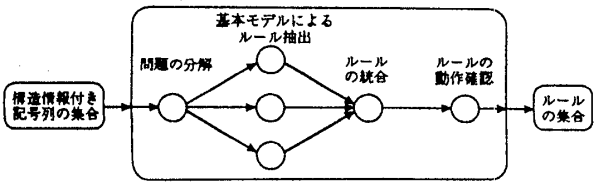

図 4: 応用モデルの実現

さて実際に、あるルールが与えられた時、そのルールに一意性が存在するかどうかを確かめる手法について検討した結果、得られるルールのクラスが正則文法の範疇にある場合には、そのルールについて一意性を決定的に求められるということが明らかになった。また、ルールが正則文法の範囲を超え、文脈自由文法クラスになる場合には、任意のルールに関して決定的に求めることは不可能である。

3.5 応用モデルの実証試験

ここで取り上げた応用モデルを実現するルール抽出手法のインプリメントと実験を行うことで、提案したモデルの検証を行った。この実験では、実際に文献データベースから 1,037 例を取り出して用いた。そして、その用意したテキストの処理例を数種類作成し、各例の集合から、データベーステキストの解析を行うためのルールをそれぞれ抽出させてみた。その結果、ここで提案した処理モデルがうまく動作することと、一意性を持つという制約条件によって候補が絞り込めるということが確認された。

4 むすび

以上のように、筆者らは例からの学習の一端として、与えられた記号列、あるいは構造情報を持った記号列からのルール抽出に着目し、問題の定式化と理論的な検討、および実験を行った。その結果として、問題の枠組としてのルール抽出のモデル、また実際問題に適用可能なルール抽出手法が得られた。この研究によって、記号列の構造解析に関するルールは、理論的には同定不可能であるものの、ある条件の下では取得可能であるという知見が得られた。また実証試験により、応用モデルを用いれば、データベース上のデータ処理システム構築の支援ツールなどが構成可能であることが、その実証試験により確かめられた。

参考文献

[1] 倉島顕尚, 安達淳. 人間からの指示を含めたルール抽出過程の検討. 第 45 回情報処理学会全国大会講演論文集, 1992. 1H-1.

[2] 横森貴, 篠原武. 形式言語の学習. 情報処理, Vol. 32, No. 3, pp. 226–235, 1991.