人間からの指示を含めたルール抽出過程の検討

1H−1

倉島 顕尚　安達 淳
東京大学工学部　学術情報センター研究開発部

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。本発表では、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処理システムについて紹介する。ここでは前提として、テキストによる入力から、文脈自由文法により記述された規則を抽出することを仮定した。この問題を、形式言語の獲得という視点から考えた場合、理論的には例の集合から規則を同定するのは殆んど不可能である。しかし、元の規則に一致していなくとも、与えられた例の集合を受理できる規則を抽出することは、意味のあることである。なお、計算機処理のみで意味のある規則を抽出することも困難であるので、筆者らは人間からの指示を含めた形でルール抽出過程を捉えることによって、この点を解決しようとしている。

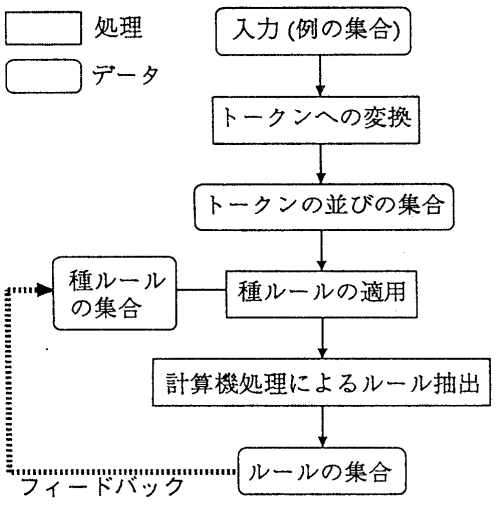

図 1: 処理モデル

A Rule Extraction Method with Human Instructions
Akihisa KURASHIMA[1], Jun ADACHI[2]
[1]Faculty of Engineering, University of Tokyo
[2]Research and Development Department, National Center for Science Information Systems

2　処理モデル

ルール抽出処理過程の全体を図1に示す。各処理の働きは、次のようになっている。

トークンへの変換　トークンとは、システム内で取り扱う要素の単位である。外部からの入力は、システムの入口で、それを構成する要素に分解される。テキストを入力とする場合、テキストを構成する各文字がトークンとなり、テキストはトークンの並びに置き換えられる。

種ルールの適用　ルールとは、トークンの並び方を表現するものである。ここでは、トークンの並びの持つ規則を文法表現の形で取得することを最終目的としているので、ルールはすべて導出式の形で表現される。ルールを導出式で表現したときの非終端記号は、変数と呼ぶ。さて、種ルールとは、結果として得られるルール表現を構成するための簡単なルールであり、外部から明示的に与えられるものである。トークンの並びの集合に置き換えられた入力は、この段階において種ルールと照合され、種ルールによって変数に置き換えられる部分は、すべて変数にしてしまう。種ルールの簡単な例としては、各文字に対応したトークンのうち、アルファベットに対応したものを一つの変数に変換するなど、複数のトークンをクラスタ化するものや、アルファベットに対応した一続きのトークンの並びを一つの変数に変換するものが挙げられる。

計算機処理によるルール抽出　この部分では、前段階までに処理されたトークンの並びの集合に対して、これを受理できるルールを計算機処理によって自動的に抽出する[2), 3)]。このために、例の集合からルール表現を得るための抽出アルゴリズムを複数用意する。複数のアルゴリズムを用意するのは、それぞれ異なった視点からルールの抽出を行わせるためである。よって、この段階の結果は複数得られることがある。外部から与える種ルールを、生成されるルール表現の核の部分とすると、計算機処理によるルール抽出は、与えられた全ての例のが受理できるように核の部分を補うように働くものと見ることができる。

フィードバック　得られた結果に利用者が満足できない場合、再度処理を行うことになる。このとき、一連の処理を行うための条件を変更しなければならない。この処理条件の変更は、種ルールを変更することによって行われる。

3　ルールの変更処理の検討

前章で説明した通りにシステムを動作させようとする場合、計算機処理によるルールの抽出の段階において、種ルールを補うようなルール抽出を行う必要がある。例えば、文献データベースの著者フィールドに登録されている著者名群のデータからルールを抽出するとする。ここで種ルールが次のように設定されていたとする。

$$S_{family} \rightarrow S_{word} \quad (1)$$
$$S_{first} \rightarrow S_{word} \quad (2)$$
$$S_{name} \rightarrow S_{first}\ T_{comma}\ T_{space}\ S_{family} \quad (3)$$
$$S_{authors} \rightarrow S_{authors}\ T_{space}\ T_{slash}\ T_{space}\ S_{name} \quad (4)$$
$$S_{authors} \rightarrow S_{name} \quad (5)$$
$$S \rightarrow S_{authors} \quad (6)$$

ここでS_{word}は文字の並びを表している。これに対して、次のような入力は種ルールでは受理できない。

America, Pierre / De Bakker, Jaco

そこで、計算機処理によるルール抽出部において、種ルールで受理できない部分をルールとするために、無条件照合法 (Voluntary Token Matching Method: 以下 VTM 法と記す) を考案した。

VTM 法　この手法は、ある特定の変数について、無条件にトークンの並びと一致するような仮定を置いた上で、入力のトークンの並びとルールとの照合を行うものである。先の式 (1) から (6) のうち、変数S_{first}とS_{family}に対して無条件にトークンの並びに照合して良いと仮定した場合の動作は、図 2のようになる。先の種ルールで受理できない例については、S_{family}に対するルールを変更すれば良いという結果が得られる。

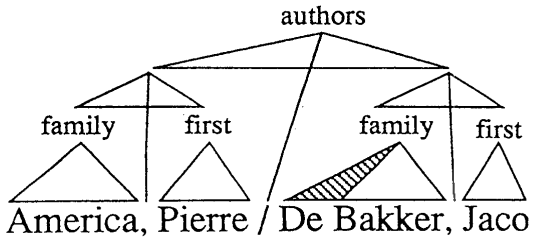

図 2: VTM 法による動作

4　利用者とのやりとり

本モデルにおいて、利用者とのやりとりは、次の点に集約することができる。同時に、ユーザインタフェースを向上するために検討している手法を示す。

種ルールの準備　まず、計算機に処理を行わせる前に、利用者は例の集合に潜むルールを大まかに記述して、種ルールとして与える。この際、VTM 法で無条件照合の対象となる変数も併せて指定する。

得られたルールの表示と評価　与えられた種ルールにより計算機は処理を行い、その結果を利用者に表示する。結果は複数得られることがある。利用者が、最善の結果を判断するため、あるいはフィードバックにより再度処理を行うかどうか決めるにはルールの評価を行う必要があるので、そのための指標を用意する。指標としては、ルール表現における導出式の数や変数の数などルール表現の表面的な性質を表すものや、ルールによって受理できるトークンの並びの範囲を計算したもの、あるいはルール表現をダイアグラムの形でグラフ化したときのパラメータの値などを用意する。

種ルールの改良　再度計算機に処理を行わせるにあたり、利用者は種ルールを変更する必要がある。その変更の際に参考にできるようなルールが得られる抽出アルゴリズムを用意することで、利用者の負担を軽くする。

5　むすび

ここでは、人間とのやりとりを含めた例の集合からのルール抽出処理モデルの提案とその実現について述べた。本アプローチは、実際に与えられた例を受理できるルールを理論的に、あるいは机上で定めることが困難な問題に対して有効である。今後、抽出されたルールの評価方法についての検討を進める。さらにその応用として、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に利用することを考えている。

参考文献

[1] T.G. Dietterich and R.S. Michalski. イベント系列の予測学習. 概念と規則の学習—例からの学習. 共立出版, 1988.

[2] 倉島顕尚, 安達淳. モデルの照合によるトークン列からの規則抽出手法の検討. 知識のリフォーメーションシンポジウム論文集, pp. 167–176, 1991.

[3] 倉島顕尚, 安達淳. トークン列からの文法規則抽出手法の検討. 1992 年電子情報通信学会春季大会講演論文集, 1992. D-204.