数式表現を含む自然言語文の解析

2Q－1

谷 幹也　宮部 隆夫　市山 俊治
NEC 関西C&C研究所

1 はじめに

自然言語インタフェースにおける日本語入力文では、数式記号と自然言語が混在した入力が行なわれることが多い.このような入力文を解析する場合、単一の文法記述で行なおうとすると、数式記号も他の自然言語の構成要素と同等に扱う必要がある．そのため、数式記号に関する文法が複雑になるとともに、解析途中に作成する予測エッジが増大していた．本報告では、入力文を数式記号を区切り記号として分割し、自然言語表現、数式記号表現それぞれの文法を利用した解析部を用いることで、数式記号を含む表現の解析を、他の自然言語処理の解析と分離する構文解析方式を提案する．本方式では、数式記号に関する文法を簡略化することができるとともに、解析途中で余分に発生していた予測を減少させることができる.

2 数式表現の入った入力文

データベースの自然言語インタフェースなどにおいては、検索結果の演算結果を再度検索するなど数式記号を含んだ入力文の解析を行なうことが多い.入力文に数式記号の記述を許した場合、数式記号及び括弧にはさまれる部分に関しては、「名詞」「数詞」「未登録語」「名詞句」「文」など様々な句構造をとり得る．従って、解析を多段階にしている場合、最も広い係受けを行なう解析部分で数式表現を記述する必要があり、このレベルに記述することによって、解析途中の予測エッジを増加させる結果となっている．また、数式演算のルールも複雑になるため、ルールの開発／メンテナンスの立場からも、問題であった．数式記号及び数式に関する文法記述は、正規表現の範囲にあり、自然言語解析用の文法とはレベルが異なる．このように文法記述のレベルが異なるものが混在する表記に対する解析を行なうために、数式部分については数式文法及び正規文法を解析できるパーザで、自然言語部分については、自然言語処理文法及びそのためのパーザで解析することによって解析の途中結果を減少させるとともに、ルール記述(特に数式文法の記述)量を減らすことができる.

A parsing method for nautral language with numerical formulas.
Mikiya Tani , Takao Miyabe and Shunji Ichiyama
Kansai C&C Research Lab., NEC Corp.

3 アルゴリズム

数式記号の混在した自然言語文を、数式表現、自然言語表現のそれぞれのパーザで処理するために、次のアルゴリズムを導入する．なお、入力文は形態素解析部によって、形態素列 ($\{D_i\}_{i=0}^{n}$) に分解されているものとする.

1. 部分文構造 (M_j^k) への分解

入力形態素列 $\{D_i\}_{i=0}^{n}$ の i=0 から順番に着目し、範囲表現開始記号であれば、部分文構造を一つ作成し、対応する範囲終了記号までの間の形態素列を部分文構造 M_j^k の形態素列に登録し、現在着目している形態素 D_i から範囲表現終了記号までを、M_j^k で置き換える．数式記号であれば、範囲開始形態素から、着目形態素の一つ前までを M_j^k として登録し、その部分を M_j^k で置き換える．部分文構造に登録されている形態素列に数式記号が存在する部分文構造の種別には「数式(NU)」を、そうでないものには「自然言語(NL)」を登録する。図1は部分文構造の分割例である。

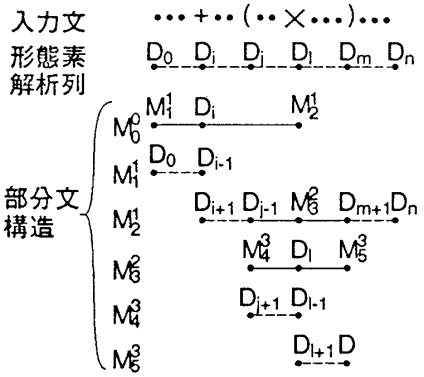

図 1: 部分文構造

2. 部分文構造のボトムアップ解析

レベルkをその最大値から0までの間、M_j^k の形態素列を構文解析する．種別が「NL」の時は、自然言語パーザで、「NU」の時は数式パーザで解析し、解析した部分解析木を部部文構造 M_j^k の部分解析木として登録する．レベルの大きいものから順番に処理を行なっているため、形態素列の中に部分文構造を示すマーカ M_j^k が存在する場合にも、それが示す品詞、意味構造、意味分類などの属性は既に決定されているため、自然言語の解析も一般の自然言語解析と同様に行なうことが可能である.

3. 部分文構造のトップダウンマージ

M_0^0 の部分解析木が定まった段階で、ネットワークをトップダウンにマージしていく．この際、解析木中に存在する「指示詞」及び「省略」は上位部分解析木と下位部分解析木において、従来用いている文間文脈処理［宮部 92］を行なうことで解消する．

解析部は図 2のような構造を持つ．入力文は、数式記号文分割部によって部分文構造 (M_j^k) によって分割され分割文データ記憶部によって管理される、それぞれの分割文の種別によって、数式処理解析部か自然言語解析部のどちらかを立ち上げて解析を行なう．

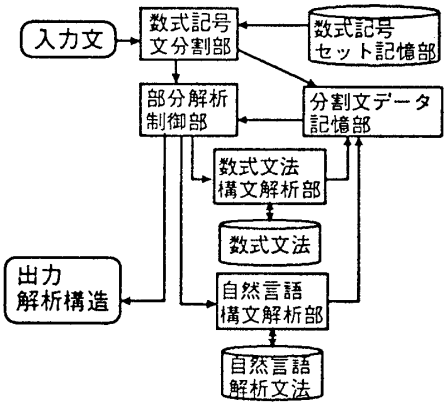

図 2: 解析部の流れ図

4 解析例

「日本電気の (1990 年以降の株価の最高値 - ここ 3 ヶ月の株価の平均× 1.4) は？」という入力文に対して、ここで提案したアルゴリズムの動作を検証する．入力文は形態素解析によって形態素列に分解されたものとする．形態素列を左から順番に解析し、部分文構造への分割を行なったものが図 3で表される $M_0^0 \sim M_4^2$ の部分文構造である．分割された段階では、解析木属性はいずれも空である．

マーカ	種別	形態素解析列	解析木
M_0^0	NL	日本電気,の, M_1^1, は,?	
M_1^1	NU	M_2^2 - M_3^2 × M_4^2	
M_2^2	NL	1990,年,から,現在,まで,の,株価,の,最高値	
M_3^2	NL	ここ,3,ヶ月,の,株価,の,平均	
M_4^2	NU	1.4	

図 3: 部分文構造への分解結果

レベル 2$(\{M_j^2\}_{j=2}^4)$ からボトムアップに解析し、レベル 0 の (M_0^0) の解析が終了したものが、図 4の (A) である．M_0^0 の解析が終了した段階で、M_0^0 からトップダウンに

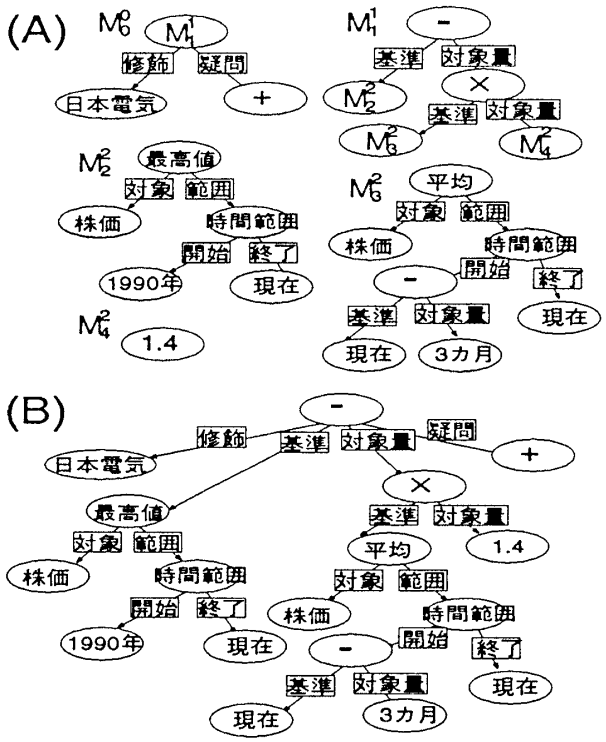

図 4: 複合解析の結果

マージを行なって解析結果図 4の (B) を作成する．

5 おわりに

数式表現を含む自然言語表現の入力文の解析を、それぞれの範囲に対して異なった解析モジュールで解析することで、途中結果を増大させることなく、また、数値表現ルールを複雑にすることなく、解析を行なえる解析方式の提案を行なった．本方式では、形態素解析の後処理時に、数式記号に着目して入力文を部分文構造に分解して、分解された部分文の種別に応じた解析文法、解析パーザを用いてボトムアップに解析し、トップダウンに結果をまとめ上げることで、レベルの異なる記述能力を持つ数式文法と自然言語文法の両方で記述された入力文を効率的に解析できる．現在、本方式を当社で開発中の自然言語インタフェース IF-Kit ［谷 91］の解析部として、自然言語処理部はボトムアップチャートパーザ、数式処理部を正規表現パーザで解析するようにインプリメントし、評価を行なっている．

［参考文献］

［谷 91］ 谷幹也, 飯野香, 山口智治, 市山俊治：自然言語インタフェース構築キット:IF-Kit, 信学技法 NLC91-62, 1991.
［宮部 92］ 宮部隆夫, 市山俊治: 日本語インタフェース文脈文法－解析手法－, 情処第 45 回全国大会予稿集,1992.