拡張部分解析木を用いた漸進的構文解析

4F−8

秋葉友良　田中穂積

(東京工業大学)

自然言語の漸進的解析は、曖昧性解消のために有効な手段である。本稿では、漸進的解析のために構文解析器がどうあるべきかについて検討する。第1に、意味解析などを早期に開始するために、構文解析器は早期に文の各部分の間に関係をつけなければならない。第2に、早期に意味解析結果を評価した結果は、フィードバックして、構文解析器の動作に影響を与えるべきである。以上の考察から、新たな構文解析アルゴリズムを示す。

1　漸進的解析と中間構造の早期生成

漸進的解析は、自然言語処理における計算モデルとして有望である。文献[3]では、名詞句の指示対象決定問題について、漸進的モデルが有効であると主張している。また、文献[4]では、意味の曖昧性解消を漸進的に行なうことを提案している。

漸進的解析を行なうために、部分的な解析木を早期に生成することが構文解析に要求される。例えば、"John met Mary at the station." という英文の解析を考える。この文で、John と met の間の意味的関係(例えば、イベント meet の行為主体が John であること)は、両者の統語的関係(すなわち動詞 met の主語が John であること)が明らかになってから解析される。ボトムアップ構文解析器で、この統語的関係が得られるのは、met 以下の動詞句が全て解析された後となる。左隅構文解析法によって得られる構造を以下に示す。

[S [NP John] [VP ?]]
⇒ [S [NP John] [VP ?]] & [VP [V met] [NP ?] [PP ?]]
⇒ [S [NP John] [VP ?]] & [VP [V met] [NP Mary] [PP at ...]]
⇒ [S [NP John] [VP [V met] [NP Mary] [PP at ...]]

この事実は、漸進的解析にとって都合の良いものとはいえない。Mary 以下の解析を待つことなしに、John と met の意味的関係を解析するためには、次のような構造が構文解析中に得られればよい。

[S [NP John] [VP [V met] [NP ?] [PP ?]]]

同等の木表現を図1に示す。本稿の手法では、構文解析の中間構造として図1のような部分的構造を許すように、従来型の構文解析器を拡張する。ここで我々が必要とする構造の集合 P を以下に定義する。以下、P の要素を拡張部分解析木と呼ぶ。

1. 未決定項を含まない完全な構造の集合 C の要素は、P の要素である。
2. X をある範疇記号、$c_k(k = 1...n)$ を C の要素、$i_k(k = 1...m)$ を未決定項の集合 I の要素、p を P の要素とする時、構造 $[X\, c_1...c_n\, p\, i_1...i_m](n, m \geq 0)$ は P の要素である。

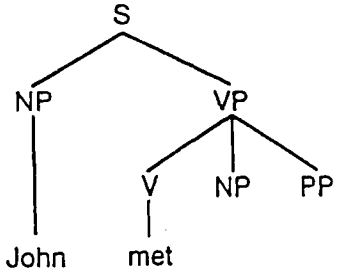

図1: 拡張部分解析木

2　フォーマリズム

構文解析の枠組の基盤としてボトムアップ・チャート法[2]を利用する。チャート法をベースにするのは、データ構造としてのチャートが構文解析手法を記述する、一般的枠組を提供すると考えられるからである。本アルゴリズムの具体化に関しては、チャート法に制約されるものではない。

2.1　複合活性弧

チャート法における活性弧を次のように拡張する。

⟨ from, to, dotted_rule ⟩

ここで、from と to は、弧の張られる、それぞれ左右側の接点を表す。また、ドットルール(dotted_rule)を次のように定義する。

- CFG 規則 $A \rightarrow A_1...A_k$ に対し、構造 $(A \rightarrow A_1...A_{i-1} \cdot A_i...A_k)$ はドットルールである(これを単純ドットルールと呼ぶ)。このとき、A を頂点範疇、A_i を最左未決定範疇と呼ぶ。
- α が A_i を頂点範疇とするドットルールである時、CFG 規則 $A \rightarrow A_1...A_i...A_k$ に対し、構造 $(A \rightarrow A_1 \ldots A_{i-1}\ (\alpha)\ A_{i+1}...A_k)$ はドットルールである。このとき、頂点範疇は A、最左未決定範疇は α の最左未決定範疇である。

従来のチャート法では、ドットルールが単一の CFG 規則から構成されるのに比べ、この定義では、ルールの構造を許している点が異なる。例えば、$(S \rightarrow NP(VP \rightarrow V \cdot NP))$ のような記述が許される。

一般に、ドットルール構造は、最も内側のルールのドットより右側にあるカテゴリ全てが未決定項である部分解析木を表す。このような不活性弧を中間構造として許すことで、拡張部分解析木を生成可能とする。本稿では、このような活性弧を複合活性弧と呼ぶ。一方、単純ドットルールを持つ従来の活性弧を単純活性弧、両者を合わせて活性弧と呼ぶことにする。

2.2　弧生成規則

ボトムアップチャート法では、2つの弧生成規則を使用する。以下にそれを示す。ただし、活性弧に関しては上記の拡張を行なっている。

A Method for Incremental Parsing using Extended Partial Trees
AKIBA Tomoyosi,　TANAKA Hozumi
Tokyo Institute of Technology

1. 不活性弧 $\langle x, y, A_1 \rangle$ に対し、CFG 規則 $A \rightarrow A_1 ... A_k$ があれば、活性弧 $\langle x, y, (A \rightarrow A_1 \cdot A_2 ... A_k) \rangle$ を張る。ただし、$k = 1$ の場合、不活性弧 $\langle x, y, A \rangle$ を張る。
2. 不活性弧 $\langle x, y, A_i \rangle$ に対し、最左未決定範疇が A_i である活性弧 $\langle w, x, (B \rightarrow ...(A \rightarrow A_1 ... A_{i-1} \cdot A_i ... A_k)...) \rangle$ があれば、活性弧 $\langle w, y, (B \rightarrow ...(A \rightarrow A_1 ... A_i \cdot A_{i+1} ... A_k)...) \rangle$ を張る。ただし、ドットがルールの右端に達した場合は、そのルールを消去し、(一つ上のレベルのルール中の) ルールを消去した部分にドットを挿入する。全てのルールが消去された場合は、不活性弧 $\langle w, y, B \rangle$ を張る。

本手法では、前節のように拡張した活性弧を生成するために、新たに次のような規則を加える。

3. 活性弧 $\langle x, y, (A_1 \rightarrow \beta) \rangle$ に対し、CFG 規則 $A \rightarrow A_1 ... A_k$ があれば、複合活性弧 $\langle x, y, (A \rightarrow (A_1 \rightarrow \beta) A_2 ... A_k) \rangle$ を張る。
4. 活性弧 $\langle x, y, (A \rightarrow \beta) \rangle$ に対し、最左未決定範疇が A_i である活性弧 $\langle w, x, (B \rightarrow ... \cdot A ...) \rangle$ があれば、複合活性弧 $\langle w, y, (B \rightarrow ...(A \rightarrow \beta)...) \rangle$ を張る。

我々のアルゴリズムでは、以上の 4 つの規則を用いる。これで、拡張部分解析木を利用する準備が出来たわけである。次に、これらの規則を、適切な場所、適切な時期に選択し、利用するためのアルゴリズムが必要となる。

2.3 基本アルゴリズム

左隅構文解析法 (left-to-right のボトムアップチャート法) では、チャートの接点に対し左から右に解析を進め、現在解析中の接点を to とする (以下、「最右の... 弧」と呼ぶ) 不活性弧に対して、必ず弧生成規則 1,2 を適用するといった戦略により、完全性と重複計算の回避を達成している。以下に、この戦略とのアナロジーから得られる、本稿の枠組での構文解析アルゴリズムを示す。

> 開始記号 X に対し、CFG 規則 $SS \rightarrow X$ を CFG 規則集合に追加する。まず左隅構文解析法と同様、最右の非活性弧に対しては必ず規則 1,2 を適用する。これによって完全性が保証される。そして、任意に選択した最右の活性弧に対して、規則 3,4 を可能な限り全て適用する。適用後、最初の活性弧は以降の解析の対象から取り除く。このようなアルゴリズムによって、左隅構文解析法と同等の重複計算の回避が可能となる。

このアルゴリズムのポイントは、規則を適用する最右の活性弧を任意に決めることが出来る点にある。どの最右活性弧に対しても規則を適用しない場合は、左隅構文解析法と同等である。一方、全ての最右活性弧を展開するのは、計算量的に明らかに問題がある。したがって、解析を進める最右活性弧をどのように選択するかが問題となる。以下、規則を適用する最右活性弧の選択方法を、活性弧選択基準と呼ぶ。

2.4 活性弧選択基準

漸進的解析にとって意味のある最右活性弧を選択する基準が必要となる。適切な選択が行なわれれば、必要な解析だけを早期に行なったり、意味・文脈解析から不適切と判断された弧を早期に取り除くことで、構文解析 (あるいは、自然言語処理全体) の効率を上げることが期待できる。このような判断基準を、構文的情報 (文法) からだけではなく、意味的情報、文脈的情報から得ることが、本稿の第 2 の目的である。以下に、活性弧選択基準の例を示す。

- 構文的情報のみを使い、結び付きやすい範疇を指定する。例えば、主節の動詞句と主語位置の名詞句、関係節と先行詞など。
- 意味解析や文脈解析に成功した部分解析木を優先的に選択する。
- 複合活性弧を生成していくと、中央埋め込み構造の深くなるものが生成される場合がある。そのようなものを選択しない (あるいは枝刈する) といったヒューリスティクスを用いる。

自然言語の構文解析器は、構文的知識を用いて入力文を処理する問題解決プログラムと見ることができる。しかし、実際の実用的な自然言語処理では、意味解析・文脈解析も同時に行なう統合的なシステムとなり、利用できる知識は構文的知識に留まらない。したがって、意味・文脈知識を同時に利用して構文解析の効率を上げることが期待できる。しかし、従来型の構文解析器では、上位レベルの知識 (意味解析) は、構文解析器が提案した構造に対する可否の判定 (意味チェック) を下すに過ぎない。我々は、真に統合的な自然言語処理を求めるなら、構文解析器の制御に関してより柔軟な機構が必要であると考える。その機構では、構文的知識と同様に、他のレベルの知識も同等に制御に対する影響力を持っている。本稿のアルゴリズムにおける活性弧選択基準は、このような機構を部分的に実現するものである。

3 関連研究

漸進的解析に関する研究において、本稿のように構文要素の間の関係を早期に得るために Categorial Grammar(以下 CG) を用いるアプローチがある [1]。本稿の手法は、一般の句構造文法を用いて、漸進的解釈を実現するものである。また、CG が部分構造を得るための情報を語彙項目にすべて記述している (構文知識だけを用いている) のに対し、我々は意味・文脈的知識が影響を与え得るより柔軟で動的なモデルを示している。

また我々は文献 [5] において、依存構造の漸進的解析手法を示しているが、本稿の手法は句構造への一般化と考えることが出来る。

参考文献

[1] Nicholas J. Haddock. Incremental interpretation and combinatory categorial grammar. In *IJCAI*, 1987.

[2] M. Kay. Algorithm schemata and data structures in syntactic processing. Technical Report CSL-80-12, Xerox PARC, 1980.

[3] C. S. Mellish. *Computer Interpretation of Natural Language Descriptions.* Ellis Horwood, Chichester, 1985.

[4] 奥村学, 田中穂積. 自然言語解析における意味的曖昧性を増進的に解消する計算モデル. 人工知能学会学会誌, Vol. 4, No. 6, 1989.

[5] 秋葉友良, 伊藤克亘, 奥村学, 田中穂積. 増進的曖昧性解消モデルに基づいた日本語解析. コンピュータソフトウェア, Vol. 9, No. 5, 1992.