日本語文の経験則による２分木解析

4H－1

酒井 克英　河田 浩二　牧野 武則

東邦大学理学部情報科学科

1 はじめに

従来の構文解析では規則の集合を用いたルールベースドシステムが用いられてきたが、ルールベースドシステムには、全ての規則がない場合解析に失敗すること、局所的な構造を把握することはできるが文の大局的な構造が把握しにくいこと、膨大な規則が必要になること、などの問題点が挙げられる。

本論文では、日本語言語の本来有している構文上の特徴である左下がり構造を生かし構文解析に積極的に利用する、２分木解析方法を提案する。この２分木構造は、文の大局的な係受けのスコープを示し、特定の文法による構文解析に利用される。これに関して、英語においては２分木解析の自動学習 [1] が研究されている。

2 日本語の構造

日本語の構造は、基本的に左下がりの構造になる。ただし、文要素の種類や品詞によっては、左下がり構造の中でも、句の構造が変化し、左下がり構造にはならない場合がある。そのため、例外となる場合については、木構造の変形を行なう。

3 ２分木解析

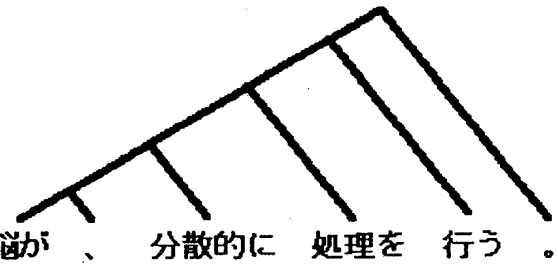
3.1 前処理

ここでは、形態素処理され品詞付けされたデータについて、解析を考える。文を扱う際の最小単位は文節とする。文節に関しては、形態素情報から、語順、品詞、機能、活用などの情報を付加されたワードリスト [2] [3] が得られる。

3.2 括弧付け操作

3.2.1 初期状態

最初に文に対し、左下がりの２文木に対応する括弧付けを行なう（図１）。

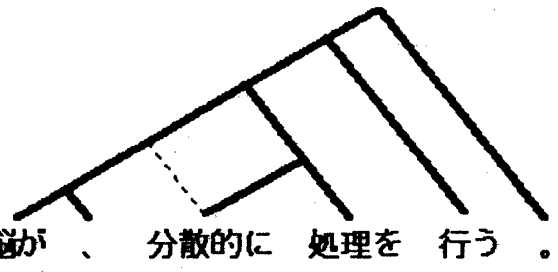


（（（（（脳が 、）分散的に）処理を）行なう）。）

図 1: 左下がり２文木と括弧付け

3.2.2 変形操作

次に、経験的に見つけだした例外、つまり左下がり２分木にならない場合に対して、変形を行う（図２）。変形の規則の条件部は、語の品詞、活用、機能



（（（（脳が 、）（分散的に 処理を））行なう）。）

図 2: 変形された２文木と括弧付け

語といった構文上の素性で記述される。条件が満たされれば、アクション部に記述された括弧オペレータによって括弧付けの変形が実行される。変形には、

- 語の（左 | 右）に、（左括弧 | 右括弧）を、（付加 | 削除）する。
- 語と語の間に、（左括弧 | 右括弧）を（付加 | 削除）する。

Analysing Binary Structure with Heuristic rules for Japanese Sentences
Katsuhide Sakai, Kouji Kawada, Takenori Makino
Faculty of Science, Toho University
2-2-1 Miyama,Funabashi,Chiba 274

という１２種類のテーブル [1] が考えられる。また、２分木構造を保つため、係り受けは非交叉である。

変形は、左括弧、右括弧をある位置に挿入する事によって行われる。この時変形は、非交叉条件と２分木構造を満たすための付随操作を行う。つまり、ある位置に"("を付加しようとした場合、

- 付加した"("に対応する")"を削除。
- ")"に対応していた"("を削除。
- 付加した"("に対応する")"を付加。

という一連の動作を行う。

解析は、以下の手順で行われる。

1. 初期状態へ、２文木をつくる。
2. 変形規則に対し、文頭から文末までで、条件が満たされれば変形を行う。
3. 文末まで見た後、次の規則に対して、チェックを行う。
4. 最後の規則について確認したら、解析終了。

3.2.3 変形の種類

変形の条件は、以下のようなものを用意した。

1. 「、」がでてきた時、それ以前の文の構造を区切り、小構造を作る。
2. 提題の「は」が出て来た時、外側に出す。

2'. 提題の「は」のあとに「、」がある時は、その「、」を含め外側に出す。

3. 文頭に出てくる時制、場所名詞で、機能のないものを外側に出す。

3'. 文頭に出て来る時制、場所名詞で、機能がなく、そのあとに「、」がある場合、その「、」を含め、外側に出す。

ここで言う「外側に出す」という操作は、２文木において上位の枝に移動することに対応する（図３）。また変形は条件１から順に見ていかれる上、非交差を満たすので、１度区切られた小構造から外へ、要素が出ていくことはない。

4 実験

実験では、あらかじめ単語単位に分割され、品詞づけされた文について、解析を行なった。単語への分割と品詞情報の付加には、juman [4] を用い、その中から最もらしいものを、人手によって選択した。

用意した日本語文、約 120 文に対し、２文木解析を行なった。その結果、そのうちの約 75%については、期待された結果が得られた。解析に失敗したのは約 25%である。そのうち、形態素の処理上問題があるもの約 10%（10 文）、並列句の発見ができていないため解析がうまくいっていないもの約 10%（13 文）、それ以外約 5%であった。

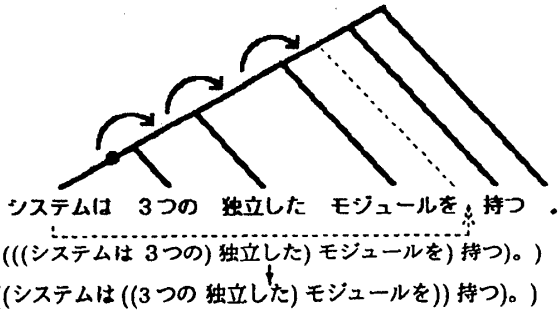


図 3: 変形操作

このように、3+2 個の規則で、75%の精度の解析が行なうことができた。

並列句の発見の問題は、前後の語の品詞を見るだけでは、うまく行かない。前後の文節の品詞情報のみで解析すると、格情報に連体を持つ連続した名詞に関し、誤った解析をしてしまう。

5 終わりに

本論文では、日本語の構造を表す手法として、２分木を用いた解析方法を提案した。この方法は、あらかじめ正解に近い構造を与えて解析を行うため、解析不能になる事はなく、ロバストである。そして、解析を行うのに、初期状態の括弧付けと、たった５つの変形規則の適用しか行っていない。また、この２分木構造は、特定の文法に基づかない構造であり、文全体の大局的なスコープを持っている。こうして得た構造は、単語情報に付加する事で、依存文法による解析 [3] にも、句構造文法による解析にも用いることができる。

参考文献

[1] Eric Brill "Automatic Grammer Induction and Parsing Free Text: A Transformation-Based Approach" ACL 93(1993)

[2] 河田、牧野「語彙依存文法における語彙項目の記述について」情報処理学会第 49 回全国大会 3G-8(1994)

[3] 河田、酒井、牧野「日本語文における２文木構造から依存構造の導出」情報処理学会第 51 回全国大会 4H-2(1995)

[4] 松本、黒橋、宇津呂、妙木、長尾「日本語形態素解析システム juman 使用説明書 ver1.0」(1993)