形態素情報と係り先範囲の制約に基づく日本語長文の骨格構造解析

1G－7

兵藤安昭　池田尚志
岐阜大学工学部

1 はじめに

日本語の構文解析では，格構造を用い，意味を考慮して係り受けの曖昧性の問題を解決しようする方法がよく行なわれている．しかし，意味的な面から係り受けの制約を記述するには，かなり精密な意味情報が必要であり，格構造による処理で行なわれるような意味素性やシソーラスを用いた方法では十分な記述は困難である．

本論文では，意味理解に入り込まない範囲で，すなわち形態素情報と係り受けに関するいくつかの表層上の制約規則のみを用いて，日本語長文の骨格構造を解析する方法について述べる．骨格構造とは，完全な係り受けの木構造をなすものではなく，並列構造など意味に立ち入らなければ解析できない部分は曖昧なブロックとしてそのまま残し，文の全体的な構造を把握しようとするものである．つまり，意味の問題に立ち入らない範囲で可能な最大限度の構文解析を追求した．

2 骨格構造解析

2.1 文節カテゴリと係り受け関係の表示

まず初めに，入力文に形態素解析処理を施し，各文節に文節カテゴリを付与する．文節カテゴリとは，文節自身のタイプと係りうる文節のタイプによりカテゴライズしたもので，文節自身のタイプを，体 (名詞)，用 (動詞・形容詞・形容動詞)，副 (副詞・連体詞)，接 (接続詞) の４つに分類し，これらの組合せにより１０種の基本的文節カテゴリを設けた (表 1)．

その他に「体並，用並，時用，は用」の４つの文節カテゴリを設けた．「体並，用並」は，その文節が並列構造の可能性があることを示している [1]．

表 1: 文節カテゴリ

体用	用言に係る体言文節．
体体	体言に係る体言文節．
体終	係り先のない体言文節．
用体	体言に係る用言文節．
用用	用言に係る用言文節．
用終	係り先のない用言文節．
副体	体言に係る副言文節 (連体詞)．
副用	用言に係る副言文節 (副詞等)．
副副	副言に係る副詞文節．
接用	用言に係る接言 (接続詞) 文節．

また，文節中に時を表す名詞や，主題を表す機能語 (は，では，...) が含まれる時は，各々「時用」「は用」として扱う．

次に，文節カテゴリより，すべての文節の可能な係り先を求め三角表上 [2] に表示する (図 1)．

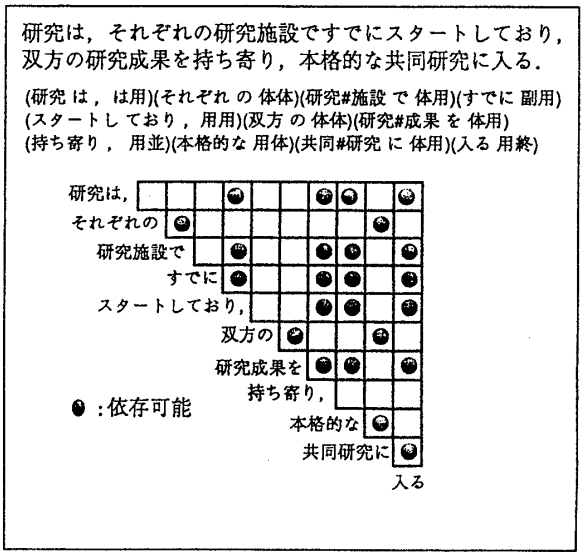

図 1: 文節カテゴリと係り受けの三角表

2.2 係り先範囲の制約に基づく係り受けブロック化

係り受けブロックという用語を，その範囲内で係り受けが行なわれる文節ないし係り受けブロックの列として定義する．本手法で述べる骨格構造解析では，意味解析を用いずに，できるだけ小さなブロックからなるように与えられた文節列のブロック化を行なう．

Skelton Analysis of Long Japanese Sentence based on Surface Information and Restriction of Dependency Area
Yasuaki Hyodo, Takashi Ikeda
Faculty of Enginerring, Gifu University
Gifu-shi,501-11,Japan

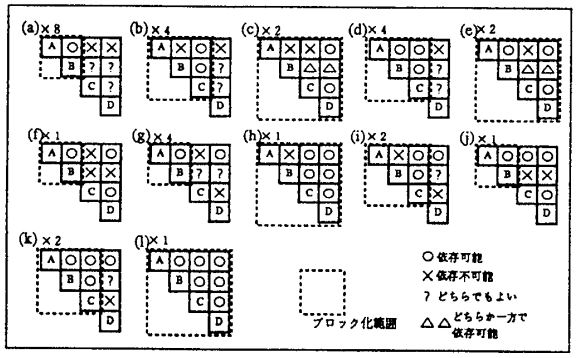

図 2: ブロック化可能なパターン

さて，本手法では (1) 文頭側から順次ブロック化し，(2) Nブロック先までをブロック化の範囲として調べる．$N = 3$ とした時，ブロック内の依存可能性のパターンは $2^6 = 64$ 通りが考えられる．これらの組合せの中で，図 2に示す 32 通りがブロック化可能なパターンである．その他のパターンは，非交差条件によりブロック化不可能なパターン (26 通り)，出現しえないパターン (6 通り) である．

2.3 骨格構造解析

骨格構造解析の手順は以下の通りである．今回の実験では $N = 3$ ブロック先までをブロック化の範囲とした．

(1) 入力文を形態素処理し，文頭のブロックを IBとする．

(2) IBに対して，前節で述べた方法により係り受け解析を行ないブロック化処理を可能な限り遂行する．ただし，係り先が表 2に示すブロック化停止文節 I，あるいは II となった場合には，そこでブロック化を停止する．

(3) ブロック化が停止した次のブロックを IBとする (次のブロックが無ければ終了)．

(4) IBより $N = 3$ ブロック前の係り受けブロックを PBとする．PBに対して，(2) と同様の処理を行なう．ただし，ブロック化を停止するのは，ブロック化停止文節 I の場合のみとする．ブロック化が停止した時，次のブロックが IBより文頭側のブロックであったら，それを PBとして (4) を繰り返す．そうでなければ，それを IBとして (2) に戻る (次のブロックが無ければ終了)．

図 1の例文の解析例を図 3に示す．実験は，朝日新聞記事より，文字数が５０文字以上８０文字未満，８０文字以上，の各５０文，合計１００文に対して行なった．骨格構造解析が正しく行なわれたのは，５０文字以上８０文字未満の文で４８文，８０文字以上の文で４７文であった．

表 2: ブロック化停止文節

I. (a) 「体言＋は＋，(読点)」
   (b) 「体言＋ (では，にとっては，としては，によると) ＋，(読点)」
   (c) (a),(b) が出現しない文で，「体言＋は」
   (d) 「接続詞＋，」

II. (a) 文節カテゴリ「体並，用並」
   (b) 読点が含まれる文節．

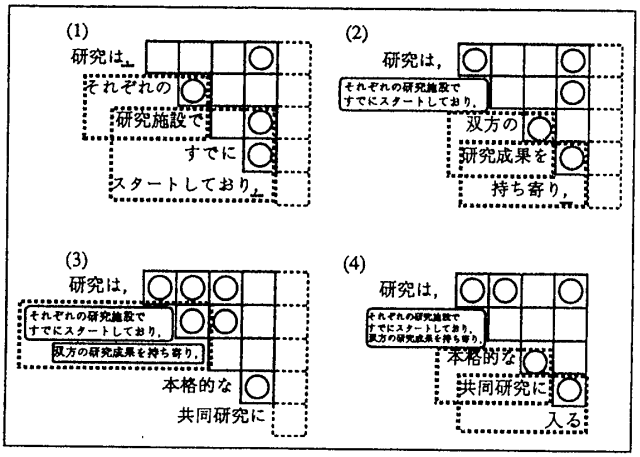

図 3: 解析例

3 おわりに

本稿では，形態素情報と係り受けに関するいくつかの表層上の制約規則のみを用いて，日本語長文の骨格構造を解析する方法について述べた．今後は，本手法により解析された骨格構造データを用いて，類似文検索など構造を考慮した高度なテキスト検索 [3] のための大規模データベースの構築に役立てたいと考えている．

参考文献

[1] 黒橋，長尾：長い日本語文における並列構造の推定，情報処理学会論文誌，Vol.33, No.8，1992

[2] 黒橋，長尾：格構造解析への評価関数の導入による統語的曖昧性の解消，情報処理学会ＮＬ研，92-9，1992

[3] 兵藤，河田，青山，浅井，池田：構文テキストベースと意味分類コードを用いた類似例文検索への応用，情報処理学会ＮＬ研，100-13，1994