日本語解析システムＭＳＬＲの効率化に関する研究

植木 正裕　　徳永 健伸　　田中 穂積
東京工業大学 大学院情報理工学研究科

1　はじめに

自然言語処理では、形態素解析・構文解析・意味解析などの解析が行なわれる。近年、構文解析と意味解析を統合する研究が行なわれているが、形態素解析と構文解析の統合に関しては研究が十分とはいえない。本研究で用いるＭＳＬＲシステムは、形態素解析と構文解析を完全に統合したシステムである[3, 4]。複数の解析を統合することの利点は、それぞれのレベルでの制約を同時に利用できることである。形態素解析と構文解析を統合すれば、形態素候補の選択の際に構文的な制約を利用できる。

ＭＳＬＲシステムでは、ＥＤＲ日本語単語辞書[5]を用いて入力文中の単語の候補を抽出し、冨田によるＧＬＲ法[2]を用いて形態素解析と構文解析を行なう。ＧＬＲ法では、文法からあらかじめＬＲ表を作成しておき、解析時にはＬＲ表を参照することで解析動作を決定する。ＭＳＬＲシステムでは、ＬＲ表を作成する際に、文法中の終端記号間の接続可能性に関する制約を用いて、ＬＲ表中の不必要な解析動作を削除しておく。これにより、構文解析を行なう過程で正しい形態素を選択することができる。

冨田によるＧＬＲ法では、グラフ構造化スタックと圧縮共有統語森という２つのデータ構造を用いることで、解析の効率化を図っている。ところが、冨田のアルゴリズム[2]では効率化が最大限には行なわれないことがある。本研究では、その問題点の指摘と解決法を述べる。

2　冨田によるＧＬＲ法

ＧＬＲ法では、解析動作に曖昧性がある場合、解析スタックを分岐させることにより、すべての解析動作を並行して行なう。解析が進むにつれ、分岐の数は組合せ的に増大するため、曖昧性の多い文の解析は難しい。冨田法では、解析の効率を上げるために、グラフ構造化スタックと圧縮共有統語森(packed shared forest)という２つのデータ構造を用いる。

グラフ構造化スタックは、複数のスタックトップに同じ記号が現れる場合に、スタックをマージすることで、スタックをツリー構造ではなくラティス構造にする。スタックをマージすることで、それ以降の重複した解析を回避できる。圧縮共有統語森は、２つの部分構文木が同じ非終端記号を親ノードに持ち、その部分木がカバーする終端記号列が同じ場合に、親ノードを共有する。

冨田法では、この２つのデータ構造を利用するために、merge, pack, share という３つの操作を用いている。ここでは、本稿の対象とする pack についてのみ簡単に説明する。

ＧＬＲ法では、解析が進む過程で曖昧性によりスタックを分岐するが、reduce によって複数のスタックに同じ非終端記号が現れることがある。

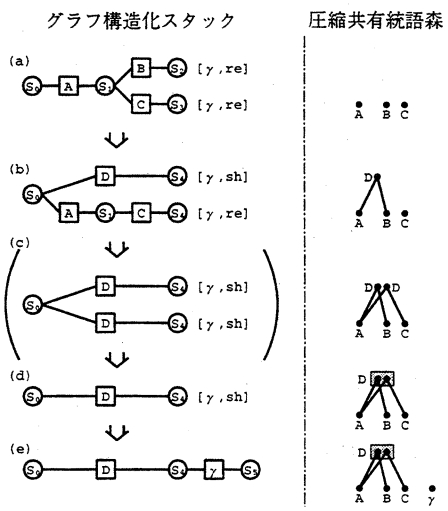


Fig. 1 スタックの pack

Fig. 1 の例ではスタックが２つに分岐しているが、図の上側の分岐では D → A, B、下側の分岐では D → A, C によりともに D に reduce されている (c)。そこで、スタック上には S_0 — D — S_4 を１本だけ生成して残し、２つのスタックをマージする。このような操作を pack とよぶ。スタックは解析動作を制御するための中間状態を保持するだけ

なので、このようにスタックをマージすることで、以降の同じ解析動作を回避できるというメリットがある。

しかし、解析結果としては、D → A, B と D → A, C の2つの reduce によりできた2つの異なる構文木がいずれも必要である。そこで、圧縮共有統語森上には両方の構文木を生成しておく。ただし、この部分の違いだけで構文木全体を2つ別個に生成するのは空間的効率が悪い。そこで、2つの部分構文木で親ノード D を共有することにより、残りの構文木全体を共有する ((d) の右図参照)。これによって、D 以下の局所的に曖昧な部分だけ両方の部分構文木を持つことで、両方の解析結果を保持できる。

3 冨田法の問題点

ＧＬＲ法では、解析動作に曖昧性がある場合に、スタックを分岐させてすべての解析動作を並行して行なうことはすでに述べた。入力文を読み進んだ段階ですでにスタックに分岐がある場合には、すべての分岐について解析を行なうことになる。冨田のアルゴリズムでは、まず各分岐ごとに深さ優先で reduce を行なう。すべての分岐での reduce が終わったところで、先読みの shift を行なう。ところが、この方法では、pack のタイミングが遅れたり、pack に失敗する場合があることが指摘されている [1]。

例として Fig. 2 の場合を考えてみたい。

この例では、最初スタックが2つに分岐している。まず上側の分岐について解析を行なう。F → B, D と G → A, F の2つのルールによる reduce が行なわれて S_0 — G — S_7 が生成される。上側の分岐での reduce はこれですべて終わったので、次に下側の分岐の解析を行なう。F → C, E と G → A, F による reduce が行なわれたところで、下の分岐にも S_0 — G — S_6 が生成されるが、上の分岐と同一のスタックになるのでこの時点で pack を行なう。

ところが、この例では F → B, D と F → C, E によって S_1 — F — S_6 という枝が2回生成され、そのため、G → A, F による reduce も2回行なわれてしまうことに注意したい。つまり、本当に曖昧なのは F を根とする構文木であるのにも関わらず、下側の分岐で reduce により F が得られたときには、上の分岐では G → A, F の reduce により F がすでに pop されているため、F での pack のタイミングを失する。

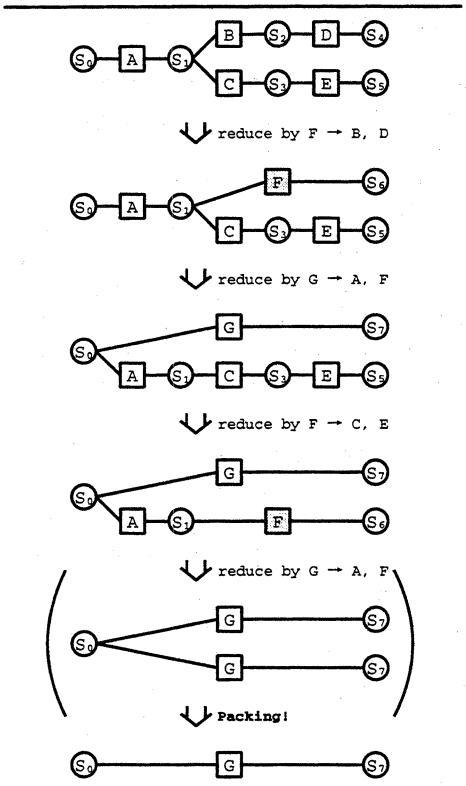


Fig. 2 pack のタイミングの遅れ

このように、冨田のアルゴリズムでは、分岐ごとの解析のタイミングにずれがあるために、pack のタイミングが遅れたり、文法によっては pack が起こらないことがあり得る。

冨田法では、pack は空間的・時間的効率の両方に影響するので、pack を最大限行なうことが望ましい。そこで、本稿では次のようなアルゴリズムを提案する。

4 アルゴリズムの改良

3 で指摘した問題点を解決するために本稿で提案するアルゴリズムの基本となる考え方は次の2点である。

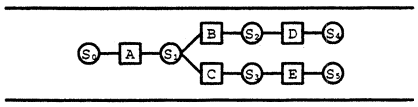
1. **スタックの分岐間での解析のタイミングを揃える**

   冨田のアルゴリズムの問題は、分岐によって解析の進み具合にずれができることであった。提案するアルゴリズムでは、各分岐ごとに reduce を深さ優先で行なうのではなく、reduce の範囲を制限することで、分岐間での解析のタイミングを制御する。

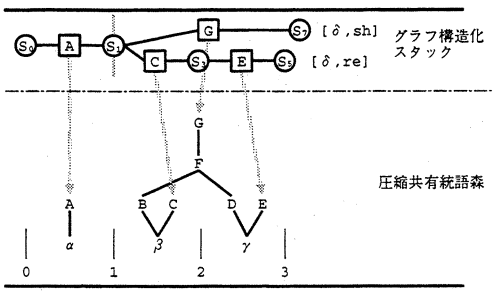
2. **圧縮共有統語森を過去の reduce 動作の履歴として参照する**

   後述するが、1 により Fig. 2 に示した pack のタイミングのずれを避けることができるが、文法中にユニットルールが含まれる場合、似たような失敗が生じることがある。これは、スタック上には過去の解析の履歴が残らないことに起因する。一方、圧縮共有統語森は過去の reduce 動作の履歴そのものであるので、これを参照することで、すでに reduce により pop され、スタック上には存在しない非終端記号の中に、pack 可能な非終端記号があるかどうかを調べる。

以下では、例により解析の様子を具体的に説明する。



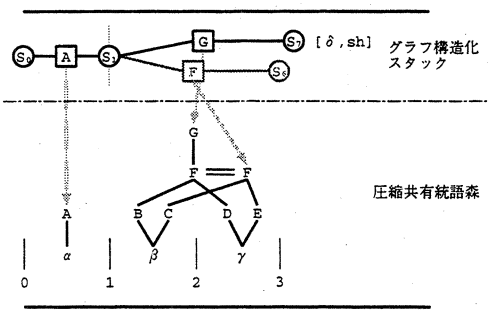
上の状態から、まず上側の分岐を用いた解析で F → B, D と G → F による reduce が行なわれると下図の状態になる。



図の上半分がグラフ構造化スタック、下半分が圧縮共有統語森を表している。矢印は、スタックと統語森上の非終端記号の対応関係を表している。なお、簡単のため、例では終端記号 α, β, γ はそれぞれ 1 形態素であるとする。

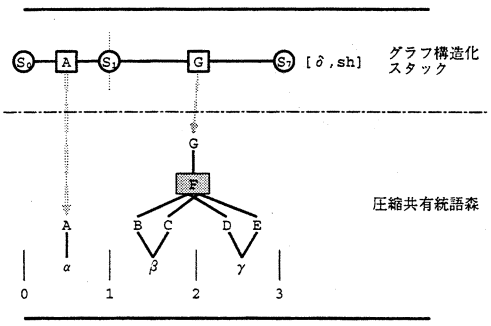
このとき、reduce 範囲に制限がなければ、上の分岐での解析をそのまま続けることになるが、例では S_1 を越える reduce は禁止しているため、上の分岐の解析をここで一時中断する。

次に下側の分岐を用いた解析に移ると、F → D, E による reduce が行なわれる。



冨田のアルゴリズムでは、スタック上にほかの F がないため、ここで F をスタック上に push するが、本アルゴリズムでは**圧縮共有統語森を参照**して、過去に F という非終端記号が存在したかどうかを調べる。

この例では、上側の分岐で F → B, D によってできた部分構文木が見つかる。そこで、2 つの部分木で親ノード F を共有させ、スタック上には F は push しない。



その結果、上図に示すスタックと統語森ができあがる。pack による変更は、F 以下の部分構文木の

追加のみで、統語森のほかの部分やスタックには変更は必要ない。

5 評価実験

提案したアルゴリズムをMSLRシステム上に実装し、冨田法との比較実験を行なった。

評価用の文としては、EDRコーパスより 5,000 文をランダムに抽出した。そのうち、固有名詞や異表記などにより解析に失敗するもの、冨田法ではメモリがたりないなどの理由で解析できないものを除いた 1,226 文を用いて比較を行なった。

比較材料としては、グラフ構造化スタックのノード数、圧縮共有統語森の深さ 1 部分構文木数を用いた。pack が行なわれることにより、スタックのマージが行なわれるため、pack の回数が多く、タイミングが早いほどスタックのノード数は少なくなる。また、pack により重複した解析動作を回避でき、無駄な部分構文木の生成も抑制される。

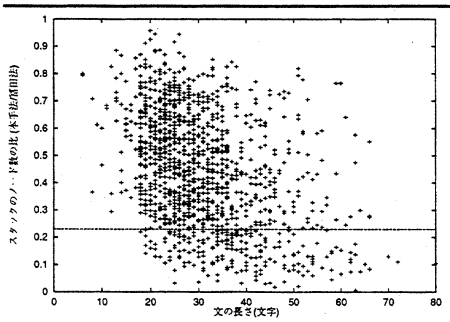


Fig. 3 グラフ構造化スタックのノード数

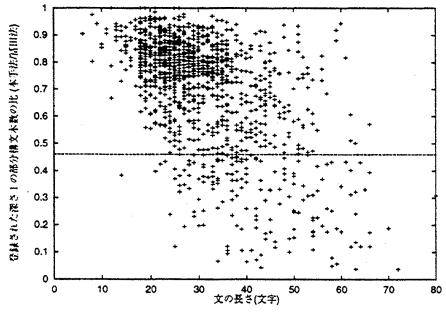


Fig. 4 深さ 1 部分構文木数

Fig. 3 と Fig. 4 は、冨田法でのノード数・構文木数を 1 としたときの本手法での結果を各文ごとにプロットしたものである。文が長くなるほど改良の効果が高いことがわかる。入力全体の平均では、スタックのノード数で 0.23、深さ 1 部分構文木数で 0.46 まで削減できた。

6 結論

本稿では、冨田によるGLR法の問題を解消し、解析をより効率的に行なうための pack に関する新しい手法を提案した。実験により、本手法は空間的・時間的効率において、冨田の方法よりも改善されることが確認できた。また、冨田の方法ではメモリ不足により解析できなかった長文でも解析が可能になった。

今回の実験では文法は 1 種類しか用いなかったが、pack の効果は文法に依存する部分があるので、他の文法を用いての実験が必要と思われる。

また、実験に用いた文の中には、固有名詞などにより解析が失敗するものも多かった。今後は、未定義語処理など、解析精度向上についても改良を行ないたい。

References

[1] 伊東秀夫. LR 表を用いたチャートパージングアルゴリズム. 情報処理学会 自然言語処理研究会, Vol. 99, No. 7, pp. 49–56, 1 1994.

[2] Masaru Tomita and See-Kiong Ng. The generalized LR parsing algorithm. In *Generalized LR Parsing*, pp. 1–16. Kluwer Academic Publishers, 1991.

[3] 伴光昇, 福田譲, 白井清昭, 田中穂積. 圧縮統語森上での形態素解析候補の絞り込み -品詞列統計情報の利用-. 1994 年度 人工知能学会全国大会 (第 8 回) 論文集, pp. 527–530, 6 1994.

[4] 植木正裕, 徳永健伸, 田中穂積. EDR 辞書を用いて日本語文の形態素解析と統語解析を行なうシステム. EDR 電子化辞書利用シンポジウム論文集, pp. 33–39, 7 1995.

[5] 日本電子化辞書研究所. EDR 電子化辞書利用マニュアル, 第 2.1 版, 1994.