デュアルチャートを用いた依存文法の構文解析について

1R－5

山端 潔
NEC C&C メディア研究所
e-mail: yamabana@ccm.cl.nec.co.jp

1 はじめに

依存文法は、二単語間の係受けとその種類により統語構造を記述する文法の枠組みである。自由語順の扱いが容易なため日本語文法の記述にしばしば用いられてきたが、文法の語彙化や統計モデルとの相性の良さから英語などの文法記述に用いられるケースも増えている[1, 2]。

依存文法の構文解析法には種々あるが、基本的にチャート法をはじめとする文脈自由文法の解析法がほぼ同様な形で適用できる。しかし、例えばチャート法を単純に適用した場合、入力単語数に対する計算量のオーダーは文脈自由文法の3乗に対し依存文法では5乗となることが知られている。これを3乗とするアルゴリズムもあるが、部分解析結果が一般に非連結構造になり、他の言語処理との相性がよくないという問題点がある。

本稿では、これら従来の問題を解決し、入力単語数に対し3乗のオーダーで解析が可能でありかつ途中結果として連結な依存構造だけが現れるアルゴリズムを示す。

2 依存文法とその解析アルゴリズム

2.1 依存文法とその構文解析

依存文法では、単語間の依存関係により統語構造を記述する。依存関係は種類と方向を持ち、一方の単語がヘッド、他方がディペンダントとなる。本稿では、全体の構造が単純な rooted tree で記述できるもの、すなわち依存関係のグラフ中に閉路がなく各単語がヘッドを一つだけ持つものだけを扱うことにする。

文脈自由文法のチャート法は、ほとんどそのまま依存文法の解析に流用できる。構造下部のノードへの文法適用が終了した依存構造をチャートへの登録単位とし、同じ区間を張り以後の文法適用の可能性が同じものをパックして解析をすすめればよい。

2.2 計算量

文脈自由文法において入力単語数の3乗のオーダーで可能な構文解析結果がすべて得られる理由のポイントは、解析結果に対する文法適用の可能性が非終端記号により完全に表現されることにある。チャート法の基本演算は、隣接する区間の二つのエッジを一つのエッジにまとめあげる演算である。各区間において区別しなければならない(パックできない)エッジの数は非終端記号の数という定数で抑えられる。隣接区間の組み合わせの数は単語数の3乗のオーダーなので、全体の計算量が単語数の3乗のオーダーとなるのである。

一方、依存文法には文脈自由文法の非終端記号に相当する概念がなく、依存構造の統語的性質はヘッドワード(一般には文法適用可能な語)とその状態により決まる。区間中のどの語もヘッドワードになる可能性があるので、ある区間に対して統語的に異なりパックできない部分解析結果の数は最悪で区間長のオーダーとなる。こうして隣接エッジの組み合わせの部分で単語数の2乗の因子が導入されるので、全体の計算量は単語数の5乗のオーダーとなる。

2.3 Eisner の方法

Eisner は、チャート法を変形して単語数の3乗のオーダーで依存文法の構文解析が行えることを示した[1]。ポイントは、チャートに登録する単位(エッジ)として両端の単語だけが新たに依存関係を結べるような部分構造を採用したことにある。区間中パックできないエッジの数が区間長によらないので、計算量は文脈自由文法と同様単語数の3乗で抑えられる。ただし、このように適格なエッジを制限する代償として、非連結な依存構造の組をエッジとして許容している。

3 デュアルチャートによる依存文法の解析

Eisner のアルゴリズムでは、解析途中でできるエッジが一般に非連結な依存構造となり言語的な意味をつけにくい。そのため、部分解析結果を各種の言語処理、例えば優先度付与や枝刈りの対象とすることが困難である。この問題を解決するため、計算量が単語数の3乗のオーダーであり、かつエッジがすべて連結な依存構造だけからなる依存文法の解析アルゴリズムを以下に提示する。

このアルゴリズムでも、Eisner と同様計算量を単語数の3乗のオーダーに抑えるために、エッジの依存構造を、両端の単語にのみ文法適用が行われ得るものに限定する。異なるのは、代償として非連結な依存構造を許容する代わりに、デュアル(双対)チャート上での解析処理を行うことにある。ここで、デュアルチャートとは、単語間の位置をノードとする通常のチャートに対し単語をノードとするチャートのことをいうものとする。

以下、本手法の道具立てを説明する。まず、チャートを二つ用意する。一つは単語間の位置をノードとする通常のチャート(第1のチャート)で、他方は単語をノードとするデュアルチャート(第2のチャート)であ

Parsing dependency grammars using dual chart
Kiyoshi YAMABANA
C&C Media Research Laboratories, NEC Corporation

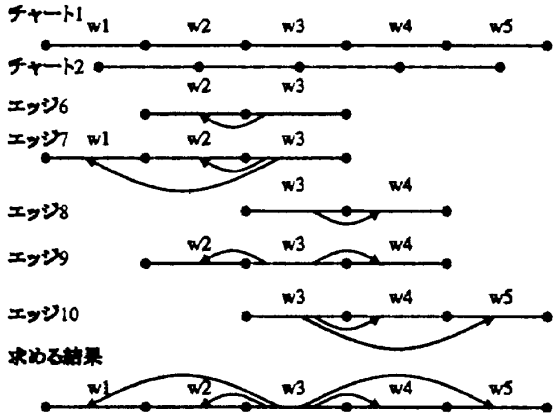

図 1: 解析例 1

る。次に、チャートへの登録単位として適格なエッジとして、連結な依存構造であって両端の単語以外への文法適用 (依存関係での結合) が終わっているもののみを使う。エッジは、単語エッジを除き、双方のチャートに共通に登録される。

次に、それぞれのチャートの上で隣接する二つのエッジをまとめあげる演算を定義する。第 1 のチャートでは、二つのエッジを新たな依存関係で連結することによりまとめあげを行う。その際、適格性を保つため、両端の単語以外への文法適用が終了したと見なしてかまわないかどうかをチェックする。これは、左側のエッジの右端の単語と右側のエッジの左端の単語に対して行えばよい。第 2 のチャートでは、隣接エッジは必ず端点の単語を共有しているので、この単語を仲立ちに二つのエッジを一つの依存構造にマージする演算をまとめあげ演算とする。この際、共有される単語上で、それぞれのエッジに属する依存関係に矛盾がないことをチェックしてエッジの適格性を保つ。

構文解析は、基本的には第 1 のチャート上でのボトムアップチャートパーシングである。注目エッジに対して第 1 のチャート上で左に隣接するエッジとのまとめあげを行っていくが、特に、第 1 のチャートでのまとめあげに引き続き、同じ注目エッジに対し第 2 のチャート上で左に隣接するエッジとのまとめあげをも行うのが特徴である。新規に作成されたエッジは双方のチャートに登録する。第 2 のチャート上でのまとめあげにより、適格エッジの制約のため第 1 のエッジでの処理だけではまとめあがらなかった構造がまとまるようになる。

図 1 に解析例を示す。一番上が第 1 のチャート、次が第 2 のチャートである。エッジ 1 から 5 までは w1 から w5 までの語彙エッジ、エッジ 6 から 10 までが解析の途中結果としてあらわれるエッジである。中央の w3 をヘッドとし、左右の単語二つずつをディペンダントとする深さ 1 の依存構造が求める構造だとする。解析をすすめると、第 1 のチャート上で、6 から 10 までのエッジがまとめあがる。内部のエッジ (点線) は文法適用が終了しており、外部と依存関係をこれ以上持てない。これらのエッジを第 1 のチャート上でどう組み合わせても求める構造は得られないが、エッジ 7 と 10 が第 2 のチャート上で隣接しており、これらを第 2 のチャート上でまとめあげることで正解が得られる。

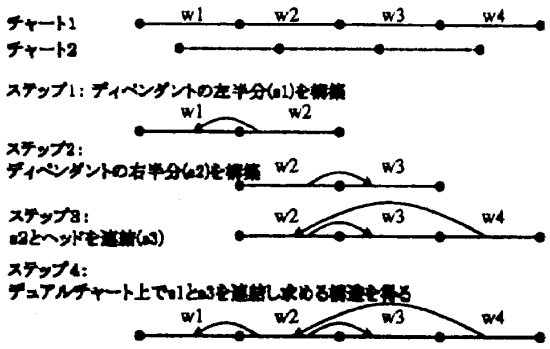

図 2: 解析例 2

4 考察

このアルゴリズムにより文法に従う依存構造がすべて正しく求められることは、単語数に関する数学的帰納法により示せる。以下概略を示す。まず単語数が小さい時にこれが成り立つことはすぐにわかる。ヘッドが両端以外にある場合は、例 1 に示したように、ヘッドワードを含みその左側の部分依存構造と右側の部分依存構造を第 2 のチャート上でまとめればよい。ヘッドワードが左端または右端にある場合だが、図 2 に典型例を示す。この場合、$w2$ を右端とする左半分の依存構造と、$w2$ を左端としヘッド $w4$ を右端とする右半分の依存構造がまとめあげられれば、これらを第 2 のチャート上でまとめることにより求める構造が得られる。ところが、これら部分構造は帰納法の仮定により正しくまとめあげできるのである。

このアルゴリズムでは、計算量のうち単語数に関わる部分は、二つの区間の組合せの数だけである。チャートを二つ使うことのオーバーヘッドはファクター 2 であり、単語数にはよらない。結局、最悪時の計算量は単語数の 3 乗のオーダーであることがわかる。また、基本演算を検討すると、処理の過程であらわれるのが連結な依存構造だけであることも明らかである。

5 まとめ

入力単語数に対して 3 乗のオーダーで全解を求めることができ、かつ、連結な依存構造のみが途中解析結果としてあらわれる依存文法の構文解析アルゴリズムを示した。今後実装と評価を行う予定である。

参考文献

[1] Eisner., J. "Bilexical grammars and a cubic-time probabilistic parser". Proceedings of IWPT97 (1997).

[2] Sleator., D. et al. "Parsing English with a Link Grammar". CMU Technical Report CMU-CS-91-196 (1991).