左隅型解析を利用した無情報からの教師なし係り受け解析

能地 宏　宮尾 祐介
総合研究大学院大学 情報学専攻
国立情報学研究所
{noji,yusuke}@nii.ac.jp

Mark Johnson　John Pate
Department of Computing
Macquarie University
{mark.johnson,john.pate}@mq.edu.au

1 はじめに

教師なし構文解析の歴史は古いが、ある程度の性能が得られ、研究が活発化したのはこの 10 年のうちである。木構造に対する確率モデルのパラメータは、それを文脈自由文法 (CFG) で表現しさえすれば、内側外側アルゴリズムによる期待値の計算と、EM アルゴリズムによって機械的に推定を行うことができる。しかしながら、初期の試みはことごとく失敗し、EM アルゴリズムでは教師なし構文解析は行えないと信じられていた (Manning and Schütze, 1999)。この状況を変化させたのが Klein and Manning (2004) であり、彼らは従来の句構造の代わりに係り受け構造を表現する CFG を定義し、この文法のパラメータは EM アルゴリズムである程度学習が可能であることを示した。従来の教師なし句構造の推定と異なり、係り受け構造に対する CFG では非終端記号が文中の単語を指すため（図 1)、意味のある表現が学習しやすいのだと考えられる。

しかしながら、このような文法のパラメータを教師なしで学習するときの探索範囲は依然大きく、単に EM アルゴリズムを実行するのでは学習がうまくいかないことが知られている。依存文法自体は projectivity などの制約を除けば特に構造に制限を設けないため、何の手がかりもなしにパラメータを学習しようとすると、モデルは簡単に悪い局所解にはまってしまうのである。これに対する従来の解決策は、EM アルゴリズムの初期値に工夫を施すというもので、実際 Klein and Manning (2004) で用いられた初期化なしでは彼らのモデルの精度は単純なベースラインを下回ることが報告されている (Gimpel and Smith, 2012)。しかしこのような英語をもとに設計された初期化が全ての言語で有効だとは限らず、Gimpel and Smith (2012) はまた、言語によっては一様分布による初期化がアドホックな初期化を上回る例を報告している。

本稿では、学習の困難さを緩和する別の方向性として、学習に使う木の構造に言語学的な制約を設ける手法を提案する。単純な一様初期化からの EM アルゴリズムが難しいのは、無数にある全ての木が同一に扱わ

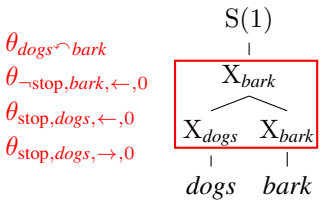

図 1: *dogs*⌢*bark* に対する句構造木と、DMV のためのパラメータ化の一部。赤で囲ったルールの確率は、その左側に示した DMV のいくつかのパラメータの積で表される。通常の PCFG のようなパラメータと CFG ルールの一対一の対応は、ルールを複数を unary ルールの集合に分解することで得られる。$\theta_{dogs \curvearrowright bark}$ は *dogs* を子に選ぶ確率、$\theta_{\neg\text{stop},bark,\leftarrow,0}$ は、*bark* が左側で停止しない（子を取る）確率で、0 は最初の子の生成であることを表す。これらは CKY チャート上のスパンから判定される。

れてしまい、その中から学習を行うのが困難なためである。提案法では、言語普遍的に有用と考えられる一部の木の集合のみを学習に用いることで、言語普遍的な構造の偏りを自動で抽出する。本稿では特に、EM アルゴリズムの最中に深い中央埋め込みを持つ木を期待値計算から取り除く手法を定式化し、その性能を評価する。英語を対象とした実験で、本手法は無情報の初期化からでも場合によっては Klein and Manning (2004) の初期化を上回る性能を示すという結果が得られた。

2 背景と関連研究

2.1 Dependency Model with Valence

Klein and Manning (2004) で導入された生成モデルは dependency model with valence (DMV) と呼ばれ、非常に多くの研究がなされている。これは文に対する projective な係り関係を生成するシンプルな生成モデルであり、拡張性も高いため、本研究でもこのモデルに焦点を当てる。DMV は、文の head から始まり、左右の外向きに各単語が dependent を生成する。各 head は

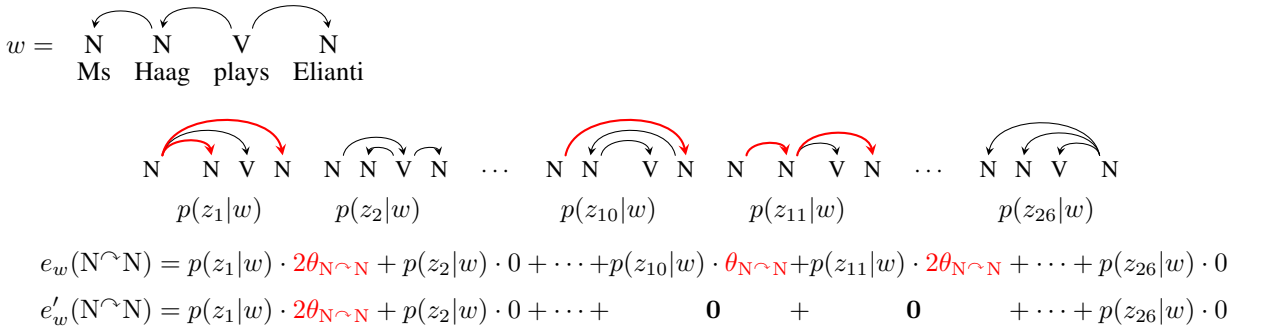

図 2: 中央埋め込みの上限を 0 とした場合（線形の構造しか許さない場合）の、提案法での期待値計算の変化。入力は品詞を用い、N から N への右向きの係り関係 N⌢N に対する期待値を考える。$e_w(\mathrm{N}^\frown\mathrm{N})$ は通常の EM アルゴリズムで計算される量である。これはある導出 z に対して割り当てられた条件付き確率を $p(z|w)$ として、z 中での N⌢N の出現回数 $c_w(\mathrm{N}^\frown\mathrm{N})$ を用いて $\sum_z p(z|w)c_w(\mathrm{N}^\frown\mathrm{N})\theta_{\mathrm{N}^\frown\mathrm{N}}$ と計算される。$e'_w(\mathrm{N}^\frown\mathrm{N})$ は提案法で計算する期待値であり、w に対する導出の候補から中央埋め込みを持つ z_{10} および z_{11} を取り除いて計算を行う。

単語の選択とは別に、それ以上 dependent を生成するか否かという二値の値に対する確率分布を持ち、この分布は更に、head がその方向に既に dependent を持っているかどうかによって条件付けがなされる。重要なのはこのモデルが CFG によって表現が可能なことであり、それによって、学習の問題はよく知られる内側外側アルゴリズムを利用した EM アルゴリズムの計算に帰着される。図 1 に DMV で計算される木の確率と、その CFG 表現に関する概略を示す。

2.2 良いモデルを学習するための工夫

DMV が初期値に敏感であることもよく研究されている。最もよく使われるヒューリスティックスは Klein and Manning (2004) が導入した harmonic initializer であり、これは学習の最初の E ステップの際に、文中の各単語同士に係り関係が生じる期待値を単語同士の距離の反比例により計算する。係り関係は一般的に短いものが好まれる性質があるが、この初期化はそのような情報を取り入れることができる。

より一般的な解決策、もしくはより高い精度を求めて、様々な研究がなされている。例えば、大量のランダムな初期値から尤度の高いものを選択する方法 (Headden et al., 2009)、機械学習によりモデルの非凸性を緩和する方法 (Gimpel and Smith, 2012; Gormley and Eisner, 2013)、目的関数を様々変化させることで局所解を脱する方法 (Spitkovsky et al., 2013) などが存在する。本研究では別の可能性として、言語普遍的な構造上の制約から探索範囲を狭めることに焦点を当てる。非凸緩和など既存手法の多くはモデル毎に理論及び実装を改良する労力を必要とするが、本稿の手法は、それが左隅型解析の CFG の上でパラメータ化できれば、木構造の生成モデルの推定に一般に適用可能である。本稿では既存モデルである DMV に制約を課すが、これがうまく働けば、教師なし木構造の推定問題に対する一般的な解決策になりうることを示唆する。事後分布正則化の枠組みで言語学的制約をモデルに加える研究も存在する (Naseem et al., 2010; Gillenwater et al., 2011)。これらは局所的なパラメータに対して制約を加えるのに対し、本稿の制約はより大域的な情報を捉えられる点が異なる。またこれらの手法では例えば動詞は名詞の head になりやすいなどの弱い教師情報を組み込むことに使われるが、我々はそのような基本的な構造の獲得がどのような制約によって可能になるかという点が興味の中心であるため、目的が若干異なる。技術的には、事後分布正則化と本稿での構造的な制約は直交しており組み合わせることが可能である。

3 提案手法：左隅型 DMV

提案法の基本方針は、EM アルゴリズムの際に言語学的に可能性が低い構造を取り除きながら学習を行うというものである。本節ではまず直感的に、提案法によって EM アルゴリズムの挙動がどう変わるかを説明し、その後それを実現するための文法について述べる。

3.1 中央埋め込みを避ける EM アルゴリズム

言語学的に可能性が低い構造として、本稿では中央埋め込み構造に着目する。我々はこれまで係り受け構造に対する左隅型構文解析の研究を行ってきた (Noji and Miyao, 2014)。そこで示したように、係り受け構造の中での中央埋め込み度合いは左隅型の構文解析中に捉えることが可能であり、また我々は中央埋め込み構造が言語普遍的に稀な構造であることを示した。

図 2 は、ある文に対する EM アルゴリズムの挙動が、提案法によってどのように変化するかを説明した

ものである。PCFG に対する EM アルゴリズムでは、各ステップで訓練コーパス中でのルール $A \to \alpha$ の期待値 $e(A \to \alpha)$ を計算し、それを正規化することでパラメータを更新していく。

$$\theta_{A\to\alpha} = \frac{e(A \to \alpha)}{\sum_{\alpha'} e(A \to \alpha')}. \quad (1)$$

この期待値計算は各文毎に行われ、$e_w(r)$ で文 w における r の期待値を表せば、$e(r) = \sum_{w\in C} e_w(r)$ となる。C は訓練文の集合を表す。これらの各文に対する期待値は、内側外側アルゴリズムを用いることで効率的に計算できる。図において、中央の二つの導出 z_{10}, z_{11} は深さ 1 の中央埋め込み構造を持つ (Noji and Miyao, 2014) [1] が、提案法では内側外側アルゴリズムでの期待値計算の際に、これらの木を取り除いた不完全な分布の上で期待値計算を行う。どの深さまで中央埋め込みを許すかは事前に設定するパラメータである。

3.2 チャート上での左隅遷移とスタック深さの表現

ではどのように期待値計算中に中央埋め込み構造を取り除けるのだろうか。ここでは遷移型の左隅型解析器がある CFG によって表現可能であり、それによって CKY チャート上で非終端記号に解析中の部分木の中央埋め込み度合いが表現可能となることを説明する。

Noji and Miyao (2014) の解析器は、常にスタック上の部分木の右側をダミーノードによって抽象化しながら解析を行う。提案法では、この解析器の各動作を CFG ルールに対応付ける。例えば一語 *the* がスタック上にあるとき、LEFTPRED 動作は *the* がその右側の語の dependent となることを予測し $\overset{the}{\frown}{}^{x}$ という部分木をつくる。遷移型解析器では逐次的に処理を行うため、ダミーノードにより現時点で観測されない head を抽象化することが必要となるが、チャート型の解析器では動的計画法を用いるため、対応する CFG ルールはこの x に入る具体的な語を含める。例えば現在の文が *the big dog is* $\cdots$ と続く場合、 *the* の head は *dog* なので、対応するルールは次のようになる。

$$\mathrm{X}_{dog/dog} \to \mathrm{X}_{the}$$

X_{the} は *the* が head となる部分木を表す。$\mathrm{X}_{dog/dog}$ は少し複雑で、 head が *dog* となる木において、 *dog* はまだ観測されていない、つまり予測しているという状態を表す。提案法において一般に $\mathrm{X}_{a/b}$ は、 head が a となる不完全な部分木を表し、それが部分木の外（右側）にある語 b を待っている状態を抽象化する。

[1] 係り受けに対する中央埋め込み度合いは、係り受け木をラベルなしチョムスキー標準形の句構造に変換することで調べられる。

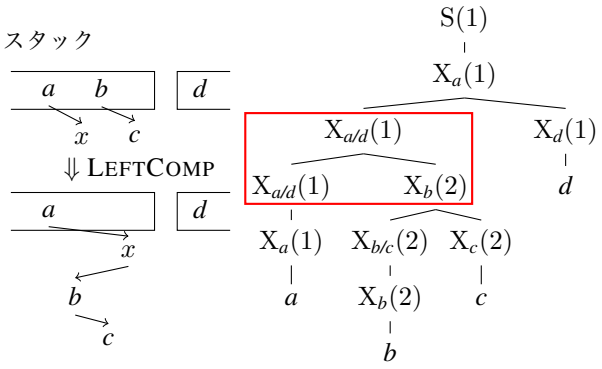

図 3: （右）: 図 2 の z_{10} に対応する中央埋め込み構造を認識する場合の CFG 木。非終端記号の数字は、その部分木を左隅型解析器で認識する際のスタック深さを表す。開始記号 S の深さは 1 でないといけないという制約と、各 CFG ルールで決められた深さの計算によって各ノードの深さが決定する。（左）: 赤で囲った CFG ルールに対応する遷移型解析器の動作。この動作を適応する際にスタック深さが 2 となることが、右側の $\mathrm{X}_b(2)$ によって表現されている。

現在解析中のスタックの深さは CFG 木の非終端記号にエンコードされる。図 3 は、図 2 の z_{10} に対応する中央埋め込みに対する導出の CFG 木を表す。図中の $\mathrm{X}_{a/d}(1) \to \mathrm{X}_{a/d}(1)\ \ \mathrm{X}_b(2)$ は二つの木の合成動作 LEFTCOMP に対応するルールである。左隅型解析でスタックの要素が二個以上になるのは、この LEFTCOMP もしくは RIGHTCOMP が使われるときのみであるため、これらの動作に制限を加えることによって、中央埋め込みを持つ木をチャート上から取り除くことができる。LEFTCOMP 動作は一般に次のルールで表現される[2]。

$$\mathrm{X}_{a/b}(\gamma) \to \mathrm{X}_{a/b}(\gamma)\ \ \mathrm{X}_c(\gamma+1)$$

このルールを適応する際に右側の子の深さが一つ増えてしまうが、最大深さが γ である場合はこのルールの確率は 0 となり、結果として内側外側アルゴリズムの中でこのルールを利用する全ての木が取り除かれる。

これまで各ルールのパラメータ化については議論を省略したが、DMV と同じように、各 CFG ルールを DMV で使われるパラメータの積として表現することが可能である。またこの CFG をそのまま CKY パーザとして実装した場合の計算量は $O(n^6)$ であるが、内部をよく観察することにより Eisner and Satta (1999) と似た手法を適用でき、$O(n^4)$ での実装が可能となる。

[2] 正確には、この深さの関係はノード X_c のスパンの長さが 1 より大きい場合にのみ成り立ち、そうでない場合 X_c の深さは γ となる。これは単語を shift してすぐに composition を行うことに対応するが、この動作は例えば右枝分かれの構造を処理する際に必要とされる。

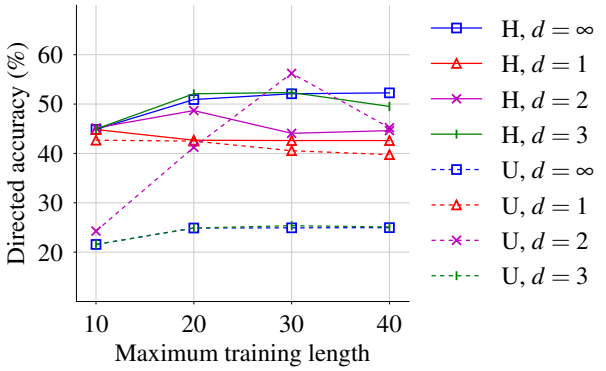

図 4: WSJ Section 23 （最大長さ 10）に対する精度の比較。 H, U は初期化の違いであり、H は harmonic initializer、 U は一様初期化を表す。d は最大深さで、$d = \infty$ は制約を加えない通常の DMV である。

4 実験

本研究の目標は、初期値の工夫なしで教師なし学習可能な手法の開発であるから、特に提案法を無情報（一様）の初期値から実行した場合の精度と、Klein and Manning (2004) の初期化を用いた場合の精度の比較に焦点を当てる。また本稿は英語のみの結果を報告する。先行研究に従い、ある程度カテゴリー化された列から意味のある構造を獲得できるかを確かめるために品詞を入力として用い、Penn Treebank WSJ の Section 2 – 21 を訓練文に、Section 23 をテストに用いる。なお深さの制約は訓練時のみに用いテスト時には用いない。

特定深さの中央埋め込み構造は、文が長くなるほど指数的に増えるため、提案法の精度は訓練文の長さに依存する可能性が高い。図 4 は様々な設定に対し、訓練に使う文の最大長さを変化させたときの精度の変化をプロットしたものである。ただし評価にはテスト文のうち長さ 10 以下のものしか使用しなかった。これは教師なし解析でよく評価される設定である。長さ 40 までの文で評価しても傾向は変わらなかった。Gimpel and Smith (2012) が報告するように、従来の何も制約を加えない一様初期化 (U, $d = \infty$) の精度は 25%程度と非常に低い。最大深さを 1 にした場合 (U, $d = 1$)、特に長さ 10 以下の設定では Klein and Manning (2004) の harmonic initializer とほぼ同等の性能を示すほど大きく精度の向上が見られる。最大深さが 1 とは、図 2 のように中央埋め込みを全く許さない場合に対応するが、英語の単純な構成の文はほとんどがこの制約のもと解析を行うことができ、また短い文ほどそのような単純な構造が多数を占めるため性能の向上が大きかったと考えられる。この結果から、本研究のアプローチである EM アルゴリズム中の木の候補の削減は、無情報の初期値からの学習に対し有効に働くといえそうである。最大深さを 2 にすると、一段の中央埋め込み構造が許容される。関係節なども大抵の構造であればこの深さで処理ができるため、カバーできる範囲は非常に大きいが、その分短い文の解候補の削減には有効に働かない。しかし $d = 2$ で長さ 30 以下の文から学習したときは驚くほど高い精度を達成し (56.22)、harmonic initializer を 4 ポイントほど上回った。深さ 2 以下の構造は Noji and Miyao (2014) においてほとんどの言語で 90%以上を占めることが観察されており、多くの自然言語の構造をカバーしながら解候補の削減を行える数値である可能性がある。しかしながら最大長さ 40 では精度が低下しており、$d = 2$ で任意の長い文から構造を抽出できるというわけではなさそうである。このように限定的ではあるが、主に英語に向けてチューニングされた Klein and Manning (2004) の手法を上回る性能を無情報から得ることができたという結果は興味深く、今後多言語の実験を通して本手法の可能性を詳細に探っていく予定である。

5 おわりに

教師なし構文解析の学習を向上させる手段として、従来の初期値の工夫とは別の、学習候補の削減という方向性を検討し、従来難しかった無情報の初期値からの学習が行える可能性を示した。本手法はまた、認知的制約を加えた教師なし文法獲得のモデルとも見なすことができ、文法の学習可能性を説明する道具となるかもしれない。特に $d = 1$ の設定は中央埋め込みを全く認識できない、つまり記憶容量が非常に制限された学習者と見なすことができ、この設定で実際に短い文からの学習が高精度を示したことは興味深い。ここでは従来の問題設定である品詞列からの学習に焦点を当てたが、これをより認知的に望ましい条件に変えていくことも今後の課題である。

参考文献

J. Eisner and G. Satta. Efficient parsing for bilexical context-free grammars and head automaton grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 457–464, 1999.

J. Gillenwater, K. Ganchev, J. Graça, F. Pereira, and B. Taskar. Posterior sparsity in unsupervised dependency parsing. *J. Mach. Learn. Res.*, 12:455–490, 2011.

K. Gimpel and N. A. Smith. Concavity and initialization for unsupervised dependency parsing. In *Proc. of NAACL-HLT*, pages 577–581, 2012.

M. R. Gormley and J. Eisner. Nonconvex global optimization for latent-variable models. In *Proc. of ACL*, pages 444–454, 2013.

W. P. Headden, M. Johnson, and D. McClosky. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proc. of HLT-NAACL*, pages 101–109, 2009.

D. Klein and C. D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. of ACL*, pages 478–485, 2004.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

T. Naseem, H. Chen, R. Barzilay, and M. Johnson. Using universal linguistic knowledge to guide grammar induction. In *Proc. of EMNLP*, pages 1234–1244, 2010.

H. Noji and Y. Miyao. Left-corner transitions on dependency parsing. In *Proc. of COLING*, pages 2140–2150, 2014.

V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. Breaking out of local optima with count transforms and model recombination: A study in grammar induction. In *Proc. of EMNLP*, pages 1983–1995, 2013.