科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１８　日現在

機関番号：　１２１０２
研究種目：　研究活動スタート支援
研究期間：　2012～2013
課題番号：　２４８０００００４

研究課題名（和文）構文解析モデルの階層的確率オートマトンへの等価変換

研究課題名（英文）Establishing Equivalent Transformation from Syntactic Parsing Models to Hierarchical Probabilistic Automaton

研究代表者
若林　啓（Wakabayashi, Kei）

筑波大学・図書館情報メディア系・助教

研究者番号：４０６３１９０８

交付決定額（研究期間全体）：（直接経費）　　2,300,000円、（間接経費）　　690,000円

研究成果の概要（和文）：構文解析は，言語などの系列データから「自然なまとまり」を抽出し構造化する系列解析技術である．従来の構文解析モデルは，対象の系列が長くなると極端に計算時間が増加する問題があり，ビッグデータへの適用は困難であった．本研究では，従来の構文解析モデルを階層的確率オートマトンと呼ばれる，系列長に対して線形な計算時間で解析可能なモデルに等価変換する手法を確立した．この変換によって長い系列の近似構文解析を高速に実行できることを示し，文章の名詞句抽出やフレーズ抽出の応用において高速かつ効果的な解析を実現する手法を確立した．

研究成果の概要（英文）：Syntactic parsing is a data analysis technique for estimating structures of sequence data. Existing syntactic parsing models have a problem that the computation time extremely increase for longer sequences, therefore they can hardly be applied to the big data analysis. In this project, we established the equivalent transformation from the existing parsing models to the hierarchical probabilistic automaton which can parse in the linear computation time with respect to the length of sequence. I demonstrated that the proposed transformation enable us to execute approximate parsing of longer sequences very faster, and established the fast and effective sequence data analysis applications of noun phrase extraction and topical phrase extraction.

研究分野：　工学

科研費の分科・細目：　知能情報学

キーワード：　教師なし構文解析　階層的確率オートマトン　階層型隠れマルコフモデル　確率文脈自由文法　依存構造解析　チャンキング　フレーズ抽出

１．研究開始当初の背景

自然言語文章の構文解析は、自然言語の記述内容を解析する上で重要な基盤技術である。計算機による構文解析では、構文解析モデルとして適切な確率モデルを獲得する方法が本質的に構文解析の精度を決定する。特に近年では、人手で構文情報を付与したコーパスを全く利用せず、単語の系列または品詞の系列のみを学習データとして用いる教師なし学習に基づく構文解析の必要性が強く指摘されている。

しかし、教師なし構文解析では、確率モデルの学習に要する計算量が大きく、大規模なコーパスを利用した学習が難しいという問題がある。教師なし学習では、全ての可能な構文木の確率を推論し、それに基づいてパラメータ学習を行う必要がある。構文木の推論には、基本的に Inside-Outside アルゴリズムの拡張が用いられており、文章長 T に対して、T の 3 乗のオーダの計算量を要する。教師なし構文解析では大規模なコーパスを用いた学習が必要であるが、計算量の問題のため短い文章のみで構成されたコーパスを利用せざるを得ず、長い文章で適切な解析結果を得ることは難しい。このため、文章長について計算量の小さい構文解析モデルの推論アルゴリズムが求められている。

２．研究の目的

本研究では、既存の構文解析モデルから階層的確率オートマトンへの等価変換手法を確立することで、階層的確率オートマトンの効率的な推論手法を適用し、従来よりも小さい計算量での構文解析モデルの推論及び学習の実現を目的とする。

階層的確率オートマトンは、確率オートマトンが別の確率オートマトンを階層的に生成する動作をモデル化している。研究代表者は階層的確率オートマトンの理論的な研究を進めてきており、系列長 T に対して線形の計算量の効率的な推論アルゴリズムが存在することが明らかになっている。

階層的確率オートマトンが生成する状態系列は木構造を成すが、生成可能な木構造は、与えた状態空間とパラメータ制約に依存して決まる。本研究では、生成可能な木構造を構文解析モデルが生成する構文木の構造と一致させることで、等価な階層的確率オートマトンを獲得する手法を確立する。

３．研究の方法

本研究では、既存の構文解析モデルとして平成 24 年度に PCFG モデルを、平成 25 年度に依存構造生成モデルを対象に、以下の課題を実施した。

(1) 既存の構文解析モデルから階層的確率オートマトンへの等価な変換の導出。階層的確率オートマトンが生成可能な状態系列が既存の構文解析モデルと対応するような、階層的確率オートマトンの状態空間とパラメータ制約を導出する。

(2) 等価な階層的確率オートマトンの推論及び学習アルゴリズムの導出。構文解析モデルは無限に深い木構造を生成可能であるため、等価な階層的確率オートマトンは無限の階層を持つ必要がある。このため、厳密推論は理論的に不可能である。しかし、階層的確率オートマトンでは、階層を降りれば降りるほど指数的に確率が減少していく性質があるため、近似的な手法によって十分効果的な推論が実現できることが予想される。このアイデアに基づいて、効果的な近似推論および学習アルゴリズムを導出する。

(3) 推論及び学習アルゴリズムの実装。計算機上で実データを用いた構文解析モデルの学習を行い、既存のアルゴリズムとの実行時間の比較実験を行う。また、導出した近似学習アルゴリズムについて実際の自然言語処理応用における有効性の実験を行い、提案手法の評価を行う。

４．研究成果

(1) PCFG モデルの階層的確率オートマトンへの等価変換を確立し、その推論アルゴリズムの実装および実行時間の検証を行った。導出した等価変換は、PCFG の Right-Corner 変換を用いたプッシュダウンオートマトンに基づいている。従来の確率的プッシュダウンオートマトンは、解析木を子ノードから親ノードに向かって生成する特性をもつことから、確率モデルとして表現する確率分布が元の PCFG と異なっている。本研究では子ノードが何代前の親から生成されたかを表す「左導出カウント」と呼ぶ補助潜在変数を導入することにより、ベイズの定理を利用して確率分布が一致する階層的確率オートマトンの構成法を確立した（**図１**）。

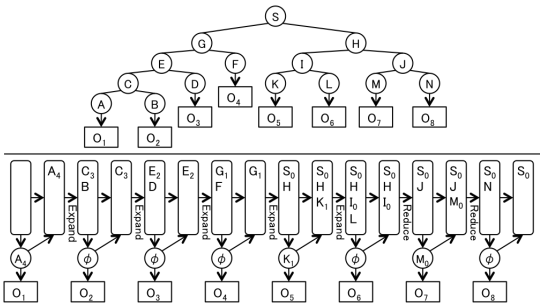

図１ 左導出カウントを含む PCFG の階層オートマトン表現

実行時間の検証として、PCFG モデルの代表的な推論手法である Inside-Outside アルゴリズムとの比較を行った。本手法による推論は、深い階層を必要とする構文解析では実行時間が不利であるという実験結果を得たが、一方で長い系列長のデータに対して極めて優位であるという結論を得た。近年の PCFG を用いた応用研究では、音素列の解釈や、ア

クセスログの解析などにおいて比較的浅い構文文法の有用性を示した結果が多く示されている。本研究成果は、このような浅い構文解析を長い系列に対して実行する際の有効な高速化手法として発展を期待できると考えられる。

(2) 代表的な構文解析モデルのひとつである依存構造生成モデルと等価な階層的確率オートマトンを導出し、これを応用した自然言語文章の教師なしチャンキング手法を提案した。チャンキングは、単語の系列から名詞句や前置詞句といった浅い統語構造を抽出する技術であり、固有表現抽出や機械翻訳などで重要な前処理であると考えられている。依存構造生成モデルの等価変換によって導出された階層的確率オートマトンは、研究成果（1）と同様、階層数が大きくなることによって計算量の観点から厳密推論が困難になるが、階層数を制限することによって得られる近似モデルがチャンクの推定を非常に効率よく高い精度で行うことができることを明らかにした（**図2**）。

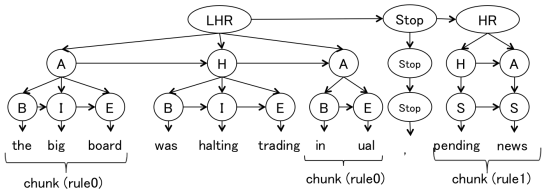

図2 平坦近似依存構造モデルによる文章のチャンキング

提案した「平坦近似依存構造モデル（Flat Approximated Dependency Grammar, FADG）」は、従来の教師なし依存構造生成モデルである DMV と比較して、実際の英語文章の学習で１万分の１以下の圧倒的に少ない計算時間で実行できることを実験により示した。同じオーダの計算量の教師なしチャンキング手法との比較では、これまで構文構造を考慮できないために抽出が難しかったチャンクの抽出や、誤って抽出されていたブロックの効果的な除去が可能になることを示し、FADG は現在の世界最高精度を達成した。本研究成果においては英語を対象にした実験を行ったが、今後の展望としては異なる言語への適用が考えられる。適切な平坦近似は言語によって異なると考えられるため、より一般的な近似手法の確立が今後の主な課題といえる。

また、依存構造生成モデルを用いた教師なし構文解析が近年盛んに研究されており、本研究成果の応用が期待される。教師なし構文解析は学習するコーパスの規模の大きさが本質的に精度に影響するため、計算量の小さい依存構造の推論アルゴリズムは単なる計算時間の短縮だけでなく、広い分野での高精度な構文解析の実現に貢献できる。提案手法は局所的な依存構造のみを推定するが、部分構文解析をカスケードさせることで完全な構文解析を行う手法が提案されており、本研究成果を用いた計算量の小さい近似的な完全構文解析の手法を導出できると考えられる。

(3) 階層的確率オートマトンによる浅い構文解析を応用した、自然言語文章の教師なしフレーズ抽出手法を提案した。従来のフレーズ抽出手法は頻出部分系列の探索に基づいた系列データマイニングを応用した手法が主流であるが、単純に単語系列に対して適用すると「名詞＋助詞」という日本語の普遍的なパターンをフレーズとして抽出してしまう。従来手法では、このような不適切な部分系列を除くためにあらかじめ助詞や動詞を取り除く前処理を行うことから、「核の傘」や「ティファニーで朝食を」といった有用なフレーズを原理的に抽出することができなかった。提案手法は、どの文書にもまんべんなく出現する「ストップフレーズ」と，特定の文書に偏って出現する「トピックフレーズ」を浅い構文解析の枠組みで統一的にモデル化することで、有用なフレーズを選択的に抽出することに成功している（**図3**）。新聞記事１年分を利用した実験により、提案手法が現実的な実行時間で有効なフレーズを抽出できることを確認した。提案手法に関する論文は「DEIM フォーラム 2014 優秀論文賞」を受賞するなど分野内からの高い評価を得ており、フレーズ抽出手法の新しい展開として位置づけられると言える。本研究成果は、階層的確率オートマトンへの変換を用いた構文解析アプローチの有効性および拡張性の高さを示しており、浅い構文解析をベースとした系列解析手法の新たな研究基盤として今後の発展を示唆するものであると考えられる。

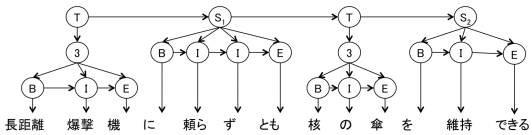

図3 ストップフレーズとトピックフレーズを識別する浅い構文解析に基づくフレーズ抽出

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 若林 啓：部分統語構造を考慮した階層的確率オートマトンに基づく教師なしチャンキング. 情報処理学会論文誌：データベース(TOD), Vol. 7, No. 62, 9p., 2014.6（印刷中）（査読有）

② 若林 啓：部分統語構造を考慮した確率オートマトンに基づく教師なしチャンキング. 電子情報通信学会他共催第６回 Web とデータベースに関するフォーラム論文集, II-3-2, 7p., 2013, 11.（査読有）

〔学会発表〕（計２件）

① 若林　啓：階層型 HMM に基づくフレーズ生成トピックモデルの提案. In 電子情報通信学会他共催　第６回データ工学と情報マネジメントに関するフォーラム (DEIM), A9-2, 6p., 2014 年 3 月 3 日～5 日, 淡路夢舞台＆ウェスティン淡路.（DEIM フォーラム 2014 優秀論文賞　受賞）

② 若林　啓：確率オートマトンに基づく確率文脈自由文法モデルの推論. 電子情報通信学会　総合大会, D-20-7, 1p., 2013 年 3 月 19 日～22 日, 岐阜大学.

６．研究組織

(1)研究代表者

若林　啓　（WAKABAYASHI, Kei）
筑波大学・図書館情報メディア系・助教
研究者番号：４０６３１９０８