科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２１　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１５６

研究課題名（和文）語句の分布情報を利用する形式言語学習理論に基づく実用的アルゴリズムの研究

研究課題名（英文）Designing practical algorithms for learning formal languages based on distribution of strings in contexts

研究代表者
吉仲　亮（Yoshinaka, Ryo）

京都大学・情報学研究科・助教

研究者番号：８０４６６４２４

交付決定額（研究期間全体）：（直接経費）　　2,800,000円、（間接経費）　　840,000円

研究成果の概要（和文）：文脈自由言語の学習に関して，近年「分布学習」と総称される方法論が目覚ましい成果を挙げていた．本研究では，分布学習アプローチを２種類に大別し，その数学的な対称性を明らかにするとともに，既存のアルゴリズムに統一的な記述を与え，もってそれらを統合する強力なアルゴリズムを設計した．文脈自由言語を超える複雑な文法形式にも分布学習が一般的な方法で拡張できることを示し，さらに正例のみから高精度高確度の確率的アルゴリズムを提案した．

研究成果の概要（英文）：Recently the approaches generically called "distributional learning" have been making a great success in the learning of context-free languages. This research project revealed the exact mathematical symmetry of the two types of distributional learning approaches. Based on this observation, we gave a uniform view of the existing distributional learning algorithms and designed an algorithm which integrates existing distributional learning algorithms in it. The obtained algorithm is stronger than other distributional learners for context-free languages. Moreover, we showed that the techniques can be applied to different grammar formalisms that are more powerful than context-free grammars in a uniform way. We also proposed an algorithm that learns certain context-free languages from positive examples only with high probability and accuracy.

研究分野：　総合領域

科研費の分科・細目：　情報学　知能情報学

キーワード：　文法推論　文脈自由言語　弱文脈依存言語　分布学習　計算論的学習

１．研究開始当初の背景

さまざまな応用分野，たとえばパターン認識，生命情報科学，XML および関連技術，そして音声認識等の自然言語処理の諸課題等においては，正規言語を超える文脈自由な構造が極めて重要である．帰納的文法推論の研究史において，正則言語の学習については豊かな知識の蓄積があり，応用分野へも影響を与えてきたのに対して，それを超える文脈自由言語等の学習に関しては肯定的な理論的結果が少なく，応用分野へのインパクトは限定的であった．さらに，自然言語や生物配列には，交差依存と呼ばれる，文脈自由文法でも扱えないより複雑な現象があることも知られており，このような現象を記述可能な弱文脈依存文法と総称される文法形式の学習はさらに困難なものであった．これに対して，近年 Clark を中心とした研究者らは，文中での語句の分布情報を利用した学習アプローチを提唱し，文脈自由言語の興味深い部分クラスが効率的に学習できることを証明した．分布学習と呼ばれるこの方法論では，各語句とその語句が部分文字列として出現する文脈の関係 – すなわち文 $w=xyz$ における部分文字列 y と前後の語列 (x,z) – に注目し，文法規則を構築していく．嚆矢となった可代入文脈自由言語（substitutable context-free languages, Clark & Eyraud, JMLR 2007）の学習アルゴリズムの発表以降，この方法論に基づいた多様な文脈自由言語学習に関する肯定的な結果がいくつも提出されている．学習モデルは，正例のみからのものや質問を用いるもの，確率的な学習など多岐に渡っている．その中でも本研究申請者は，科研費（課題番号 20700124，平成 20～22 年度）の支援を受けながら，先行研究の議論を一般化し，アルゴリズムの適用範囲をより広い文脈自由言語に拡張し，さらに，この方法論を，弱文脈依存文法のひとつである多重文脈自由文法の学習へと飛躍的に発展させることに成功していた．

２．研究の目的

本研究では，以下の３つを目標とした．

⑴ 語句の分布情報に基づく学習の研究はこの数年で急速に多くの成果をもたらしており，極めて有望なアプローチである反面，まだ新しく，幾つものアルゴリズムがばらばらに提案されている状態であった．これまでに提案されている既存のアルゴリズム間の関係を整理し，言語構造的制約による学習のアプローチに関する理論を深化させる．

⑵ 文脈自由文法以外の応用上重要なフォーマリズム，特に多重文脈自由文法や木接合文法などの弱文脈依存文法のための分布学習理論を構築する．

⑶ 従来の研究では，学習者からの質問に答える教師の存在の仮定のもとで豊かな言語族が学習可能であることを示していた．このような強力な道具なしに，大規模実データからの学習に直結する確率的な学習手法を提案する．

３．研究の方法

３つの目標⑴⑵⑶は順番に遂行されるものでも独立に達成されるものでもない．⑴によって得られる知見が⑵や⑶の目標に貢献し，⑵や⑶で考案される具体的なアルゴリズムが⑴の議論に示唆を与える．⑴のために，近年の語句の分布情報に基づく学習手法が成功している言語クラスの他，正則言語や NTS 言語の言語的性質と文法表現の関係について追究する．⑵について応用分野で使われている文法形式を対象にした学習アルゴリズムの提案を行ない，

⑶では，より現実的な学習環境での確率的精度保証のあるアルゴリズムを設計する．

⑵と⑶はそれぞれ対象と手法において実用を志向し，最終的に両者を兼ねたアルゴリズムとして融合される．

４．研究成果

上記の目的（1）に関して，本研究は期待以上の成果をあげた．従来の文脈自由言語に対する分布学習アルゴリズムは，主に部分文字列に注目して文脈自由文法の非終端記号を構成する「部分文字列駆動型」のアプローチと，文脈に注目する「文脈駆動型」アプローチとに分類されていたが，それらの関係については整然とした理論が与えられていなかった．本研究では，これらのアルゴリズムに潜む数学的な対称性を明確に示し，また，それにより，従来提案されていた文脈駆動型のアルゴリズムに対して，その対称形である部分文字列駆動型のアルゴリズムを設計した（学会発表⑦）．さらに，対称性的な性質を持つこれらのアルゴリズムに統一的な記述を与え，もって融合することに成功し，従来のどのアルゴリズムよりも強力な学習アルゴリズムを提案した（学会発表⑤）．これらにより，従来提案されていたそれぞれの文脈自由言語の分布学習アルゴリズムを統一的な視点で理解できるようになった．このいわば統一理論については，文法推論の国際会議でチュートリアルとして講演するに至っている．

上記の目的(2)に関しては，当初，自然言語処理等で用いられる木接合文法を目標として計画していたが，さらなる一般化である単純文脈自由木文法の分布学習アルゴリズムを提案するに至った（学会発表⑥）．通常の文脈自由文法が文字列を生成するのに対して，文脈自由木文法は，ラベル付き木を生成する．この発表では，従来の文脈自由文法を対象とするアルゴリズムを，単純文脈自由木の学習アルゴリズムに変換する一般的な手法を提案した．さらに，ラムダ計算に基づく抽象的範疇文法のうち，文脈自由な導出構造を持つ部分族に関する分布学習アルゴリズムも設計した（学会発表⑧）．抽象的範疇文法はラムダ項の集合を言語として生成する．ラムダ項は，文字列や木，その他のデータ構造を柔軟に表現することができるため，多くの弱文脈自由文法を抽象的範疇文法の枠組みで記述できる．そのため，この成果は弱文脈依存文法に関する一般的な分布学習の方法を与えているといえる．

さらに，当初の計画を超え，弱文脈依存文法を超える強力な文法形式である並列多重文脈自由文法についても，分布学習が有効であることを示した．この文法形式では，文字列の導出過程において，中間生成物を複製する規則が許される．このため，部分文字列と文脈との関係が非常に複雑になる．興味深いことに，この文法形式においては，文脈自由言語及び弱文脈依存言語で成り立っていた対称性が成り立たず，「文脈駆動型」アルゴリズムのみが有効であることも示されている（雑誌論文①，学会発表④）．

また一方で，単純文脈自由木文法を拡張し，導出過程での中間生成物の複製を許すような文脈自由木文法を考えることができる．このような文法形式は，自然言語の単純な意味表現のために用いることができる．しかしこの文法形式においては，部分文字列駆動型にせよ文脈駆動型にせよ，従来の分布学習手法を当てはめることができず，学習のためには相当強い条件が必要になることがわかった（学会発表②）．

上記の目的（3）に関しては，正データのみから高確率高精度で，いくつかの文脈自由言語が学習可能な条件を示した（学会発表①）．従来のアルゴリズムは教師への質問により各文字列が学習対象言語に含まれるか否か決定していたが，文字列の分布と学習対象分布に関するある仮定のもとで，正例のみから十分に文字列と文脈の関係を推測できることを示した．これにより，高確率高精度の確率的学習が可能になった．

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 Alexander Clark, Ryo Yoshinaka.

Distributional learning of Parallel Multiple Context-free Grammars. *Machine Learning*. Oct. 2013. **査読有**
DOI: 10.1007/s10994-013-5403-2.

② Chihiro Shibata, Ryo Yoshinaka**.** A

Comparison of Collapsed Bayesian Methods for PFAs. *Machine Learning*. Oct. 2013. **査読有**
DOI: http://10.1007/s10994-013-5410-3

〔学会発表〕（計　8　件）

① Shibata Chihiro and Ryo Yoshinaka.

PAC Learning of Some Subclasses of Context-Free Grammars with Basic Distributional Properties. In *proceedings of the 24th International Conference on Algorithmic Learning Theory*. Springer-Verlag, LNCS 8139, pp.143-157, 2013. **査読有**

② Ryo Yoshinaka. An Attempt Towards

Learning Semantics: Distributional Learning of IO Context-Free Tree Grammars. In *proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms*. Paris, France. pp. 90-98. 2012. **査読有**

③ Chihiro Shibata and Ryo Yoshinaka.

Marginalizing Out Transition Probabilities for Several Subclasses of PFAs. *ICGI 2012, JMLR Workshop and Conference Proceedings,* Vol. 21, pp. 259-263. 2012. **査読有**

④ Alexander Clark and Ryo Yoshinaka.

Beyond Semilinearity: Distributional Learning of Parallel Multiple Context-free Grammars. *ICGI 2012, JMLR Workshop and Conference Proceedings,* Vol. 21, pp. 84-96. 2012. **査読有**

⑤ Ryo Yoshinaka. Integration of the

Dual Approaches in the Distributional Learning of Context-Free Grammars. In *proceedings of the 6th International Conference on Language and Automata Theory and Applications*. A Coruña, Spain. Springer-Verlag, LNCS 7183, pp.538-550. 2012. **査読有**

⑥ Anna Kasprzik and Ryo Yoshinaka.

Distributional Learning of Simple Context-Free Tree Grammars. In *proceedings of the 22nd International Conference on Algorithmic Learning Theory*. Espoo, Finland. Springer-Verlag, LNAI 6925, pp.398-412. 2011. **査読有**

⑦ Ryo Yoshinaka. Towards Dual

Approaches for Learning Context-Free Grammars Based on Syntactic Concept Lattices. In *proceedings of the 15th International Conference on Developments in Language Theory*. Milan, Italy. Springer-Verlag, LNCS 6795, pp.429-440. 2011. **査読有**

⑧ Ryo Yoshinaka and Makoto Kanazawa.

Distributional Learning of Abstract Categorial Grammars. In *proceedings of the 6th International Conference on Logical Aspects of Computational Linguistics*. Montpellier, France. Springer-Verlag, LNCS 6736, pp.251-266. 2011. **査読有**

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
該当なし

６．研究組織
(1) 研究代表者
吉仲　亮（YOSHINAKA, Ryo）
京都大学・大学院情報学研究科・助教
（平成 23 年９月までは，北海道大学・大学院情報科学研究科・学術研究員）
研究者番号：80466424

(2) 研究分担者
なし

(3) 連携研究者
なし