様式 C-19

科学研究費補助金研究成果報告書

平成　23　年　5　月　16　日現在

機関番号：１０１０１
研究種目：若手研究（Ｂ）
研究期間：２００８～２０１０
課題番号：２０７００１２４
研究課題名（和文）：
統語的方法論による文脈自由言語および弱文脈依存言語の正例からの効率的極限同定
研究課題名（英文）：Syntactic Approach for Efficient Identification in the Limit from Positive Data of Context-Free and Mildly Context-Sensitive Languages
研究代表者
吉仲　亮（YOSHINAKA RYO）
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

研究成果の概要（和文）：
自然言語との関わりの深い特殊な性質を満たす文脈自由言語が正例のみから効率的に学習可能であるという最近の先行研究の結果を一般化し，より豊かな文脈自由言語の族，および文脈自由言語では捉えきれないより複雑な言語現象を表現できる弱文脈依存言語の族について，これらを正例から学習する効率の良いアルゴリズムを提案した．さらに，正例に加えてある限定された種類の質問に答える教師を付けることで，より一層表現能力の高い弱文脈依存言語の族を効率的に学習する手法を提案した．

研究成果の概要（英文）：
Recently the literature showed that context-free languages with a special property reflecting an aspect of natural language phenomena are efficiently learnable from positive data. Generalizing the preceding research, our project has presented efficient algorithms that learn from positive data even richer classes of context-free languages as well as those of mildly context-sensitive languages, which handle some non-context-free phenomena observed in natural languages. Moreover, our research project has proposed techniques that learn even more expressive classes of languages from positive data with the aid of a teacher who answers limited questions from the learner.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,100,000	330,000	1,430,000
２００９年度	1,100,000	330,000	1,430,000
２０１０年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：計算論的学習理論
科研費の分科・細目：情報学・知能情報学
キーワード：学習と知識獲得，計算論的学習，文法推論，形式言語理論，弱文脈依存言語

１．研究開始当初の背景

文法推論は，形式言語をアルゴリズム的に学習することにかかわる研究分野である．形式言語は概念や自然言語の抽象的な表現であり，したがって文法推論は様々な応用を持つ機械学習の理論基盤を与える分野である．当該分野では，正規言語の学習に関する理論は，本研究課題開始期までにはすでに十分成熟していた．一方で，自然言語を初めとする応用分野において，文脈自由な構造を扱うことは必須であることは認識されていたものの，実際には文脈自由言語の効率的な学習に関する研究成果はほとんど知られていなかった．そのような状況下で，正例のみから効率的に学習可能な稀有な文脈自由言語族の例として，Clark and Eyraud (ALT2005, JMLR2007) による substitutable 文脈自由言語の学習に関する結果があった．substitutability は，２つの文字列が同じ前後の語列中に生起して文法的な文字列を構成するならば，これらの文字列は同じ統語範疇に属するとみなし，常に互いに代入可能であるとする性質である．すなわち任意の文字列 u,v,w,x,y,z について次を満たす言語 L である．

$uvw, uyw, xvz \in L$ ならば $xyz \in L$

これは単に効率的に学習可能な非自明な文脈自由言語の例というだけではなく，自然言語現象の一側面を単純化，抽象化したものとみなすことができるという意味で注目に価する結果であった．また，従来の文法学習の理論との大きな違いは，substitutable 文脈自由言語が文法の言葉による特徴付けではなく，その言語が満たすべき性質として定義されていることである．このような性質に注目して学習する戦略を統語的方法論と呼ぶことにする．

また，ドイツ語のスイス方言やオランダ語といった自然言語や，シュードノットと呼ばれる生物配列上の構造は，文脈自由言語では記述できない複雑さを持つが知られている．このような複雑な構造を扱うことが可能であり，かつ構文解析が容易な言語は弱文脈依存言語と呼ばれる。したがって弱文脈依存言語を正例から効率的に学習するアルゴリズムを設計することは重要な課題であるが、研究開始時点では肯定的な成果はほとんど知られていなかった。

２．研究の目的

本研究課題では，１で述べた Clark and Eyraud (ALT2005, JMLR2007) の研究結果を発展させ，統語的方法論によって正例のみから効率的に学習可能な文脈自由言語族に関するより一般的な理論を導くことを目標とした．さらに先行研究，および上記の統語的方法論による文脈自由文法学習に関する研究によって得られた知見をさらに拡張し，より複雑な構造を扱う弱文脈依存言語の学習理論へと一般化することを目標とした．

３．研究の方法

(1) 文脈自由文法の学習について

Clark and Eraud の提案した substitutability という制約は，正規言語における reversiblity というよく知られた制約のアナロジーとして捉えられる． Reversible 正規言語は各自然数 k に対して k-reversibility が定義され，正例のみから効率的に学習可能な言語族の階層を成すことが知られていた．言語 L が k-reversible であるとは次のような制約を満たすことをいう．

s の長さが k で $usv, usy, xsv \in L$ ならば $xsy \in L$

そこでまず，この reversible 正規言語の階層を参考にして，類似の階層構造を substitutable 文脈自由言語に導入することで，統語的方法論による正例からの学習の一般化を行うことを目指す．

(2) 応用上も重要な弱文脈依存文法という概念に対しては様々な具体的な形式化が提案されているが，中でも最も自然な文脈自由言語の拡張である多重文脈自由文法を対象に研究を進める．正規文法，文脈自由文法，多重文脈自由文法の拡張過程に沿って，reversibility, substitutability の自然な一般化概念を多重文脈自由文法に対して考案する．

多重文脈自由文法	m−次元 substitutable (研究成果 (2)①)
文脈自由文法	k,l-substitutable (研究成果 (1)①)
	substituable (Clark and Eyraud 2005)
正規文法	k-reversible (Angluin 1982)
	(zero-)reversible (Angluin 1982)

４．研究成果

(1) 文脈自由文法の学習について．
①本研究代表者は，２つの自然数 k,l に関して k,l-Substitutability という制約を次のように提案した．

s の長さが k, t の長さが l でありかつ
$usvtw, usytw, xsvtz \in L$ なら $xsytz \in L$

すなわち Clark and Eyraud のオリジナルの substitutability は 0,0-substitutable にあたる．そして，それぞれの k,l に対して k,l-substitutable 文脈自由言語が正例から効率的に学習可能であることを証明した．これは，古典的な k-reversible 正規言語の効率的学習の，文脈自由言語における相似を成す成果と言える．（学会発表[6]）

② また，この①の証明において重要な役割を担ったのが，文脈自由文法の特殊な標準形である．文脈自由文法の Greibach 標準形の一般化である k,l-Greibach 標準形を定義し，それが標準形であるとの簡潔な証明を与えた．k,l-Greibach 標準形の文法の非終端規則は，右辺が長さ k の終端記号の列で始まり，長さ l の終端記号の列で終わらなければならない．この標準形の証明手法はそれ自体が形式言語理論上の成果として意義のあるものであり，文法推論上の成果とは独立に発表している．（雑誌論文 [2]）

(2) 多重文脈自由文法の学習について
① 多重文脈自由文法は弱文脈依存文法の一種であり，文脈自由文法の自然な一般化である．正規文法における reversibility が，文脈自由文法における substitutability に対応したように，多重文脈自由文法に於いても類似の制約を定式化し，自然数 m に関して m-次元 substitutabilty を提案，かかる制約を満たす文法も，先行研究と同様に正例のみから効率的に学習可能であることを示した．m=1 の場合が先行研究の文脈自由言語における substitutabilty に相当し，真の一般化になっている．多重文脈自由言語は，最も代表的な弱文脈依存言語であり，本研究成果はこのクラスに関する効率学習の極めて稀有な例となった．（雑誌論文 [1], 学会発表 [5]）

(3) 正例と質問による学習
① (1),(2) の成果をさらに発展させ，学習の基本的な資源を正例としながら，さらに教師に対する質問を許した場合の学習可能性についても追究した．本研究で仮定したのは，任意の文字列が学習対象言語に含まれるか否かに答える教師である．その結果，この学習モデルにおいては，正規言語や従来手法で学習可能であった文脈自由言語を真に含む，非常に豊かな多重文脈自由言語の族が効率的に学習可能になることを示した．（学会発表 [3]）．

② さらに，文脈自由文法と多重文脈自由文法の導出構造の類似点と相違点の分析を通じ，文脈自由言語族の学習アルゴリズムを多重文脈自由言語のそれへと翻訳する一般的な枠組みを提案した．さらに多重文脈自由言語に加え，他の弱文脈依存文法形式である文脈自由木言語などへの拡張を可能にするため，ラムダ計算を用いた抽象的かつ包括的な学習アルゴリズムを提案した．本研究は弱文脈依存言語族学習一般に関する極めて稀有かつ強力な肯定的結果をもたらした．（学会発表 [1,2]）

(4) 付随する研究成果
① 正例からの極限同定における効率的な学習の定義は学会においても定説はなく，如上の研究を進める前提として，異なる可能な定義の妥当性や関係性について整理する必要があった．そこで，正例からの極限同定に関する一般的な議論を試みるとともに，その議論を very simple languages と関連言語クラスの学習に適応して学習効率の議論を深化させた．（雑誌論文 [3]）
② 多重文脈自由文法の数理的性質，とくに文脈自由文法との類似性に関する理解を深めた．文脈自由文法において広く知られている Chomsky-Schützenberger の定理の類似の定理が多重文脈自由文法においても成立することを示した．（学会発表 [4]）

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 3 件）

[1] Ryo Yoshinaka: Efficient learning of multiple context-free languages with multidimensional substitutability from positive data. Theoretical Computer Science 412(19): 1821-1831 (2011)［査読有］

[2] Ryo Yoshinaka: An elementary proof of a generalization of double Greibach normal form. Information Processing Letters 109(10): 490-492 (2009)［査読有］

[3] Ryo Yoshinaka: Learning efficiency of very simple grammars from positive data. Theoretical Computer Science 410(19): 1807-1825 (2009)［査読有］

〔学会発表〕（計 8 件）

[1] Ryo Yoshinaka and Makoto Kanazawa: Distributional Learning of Abstract Categorial Grammars. The 6th Conference on Logical Aspects of Computational Linguistics. July 1st, 2011, Montpellier, France. [査読有]

[2] Ryo Yoshinaka: Distributional Learning of Extensions of Context-Free Grammars. The 5th International Workshop on Data-Mining and Statistical Science & the 7th Workshop on Learning with Logics and Logics for Learning. March 29th, 2011. Osaka, Japan. [合同招待講演]

[3] Ryo Yoshinaka: Polynomial-Time Identification of Multiple Context-Free Languages from Positive Data and Membership Queries. The 10th International Colloquium on Grammatical Inference. September 15th, 2010. Valencia, Spain. [査読有]

[4] Ryo Yoshinaka, Yuichi Kaji, and Hiroyuki Seki: Chomsky-Schützenberger-Type Characterization of Multiple Context-Free Languages. The 4th International Conference on Language and Automata Theory and Application. May 27th, 2010. Trier, Germany. [査読有]

[5] Ryo Yoshinaka: Learning Mildly Context-Sensitive Languages with Multidimensional Substitutability from Positive Data. The 20th International Conference on Algorithmic Learning Theory. October 4th, 2009. Porto, Portugal. [査読有]

[6] Ryo Yoshinaka: Identification in the Limit of *k,l*-Substitutable Context-Free Languages. The 9th International Colloquium on Grammatical Inference, September 22th, 2008. St-Malo, France. [査読有]

〔図書〕（計 4 件）

[1] Ryo Yoshinaka: Polynomial-Time Identification of Multiple Context-Free Languages from Positive Data and Membership Queries. In ICGI. LNCS 6339, pp. 230-244. Springer-Verlag, 2010.

[2] Ryo Yoshinaka, Yuichi Kaji, and Hiroyuki Seki: Chomsky-Schützenberger-Type Characterization of Multiple Context-Free Languages. In LATA. LNCS 6031, pp. 596-607. Springer-Verlag, 2010.

[3] Ryo Yoshinaka: Learning Mildly Context-Sensitive Languages with Multidimensional Substitutability from Positive Data. In ALT. LNCS 5809, pp. 278-292. Springer-Verlag, 2009.

[4] Ryo Yoshinaka: Identification in the Limit of *k,l*-Substitutable Context-Free Languages. In ICGI. LNCS 5278, pp. 266-279. Springer-Verlag, 2008.

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

６．研究組織
(1)研究代表者
吉仲　亮（YOSHINAKA RYO）
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