様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３１　日現在

機関番号：　１１３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３００１３

研究課題名（和文）分布学習に基づく自然言語文とその意味表現の対からなる形式言語の学習に関する研究

研究課題名（英文）Study on the learning of formal languages consisting of natural language sentences and their semantic expressions based on distributional learning

研究代表者

吉仲　亮（Yoshinaka, Ryo）

東北大学・情報科学研究科・准教授

研究者番号：８０４６６４２４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：近年，弱文脈依存文法の学習に対して強力なアプローチとして発展していた分布学習を，より複雑な文法形式に対して展開した．まず，生成規則が文脈自由文法の拡張型になっているような任意の文法形式について，分布学習が適応可能となる抽象的一般的な条件を導いた．また，連言文法と呼ばれる文法形式に対する多項式時間学習アルゴリズムを設計した．非線形ラムダ項を用いた文法形式について，分布学習が可能となる条件に関する事例的成果をあげた．さらに，分布学習可能にする文脈自由文法に関する最も代表的な条件について，これを弱め，より広いクラスの文法が学習可能であることを証明した．

研究成果の概要（英文）：In recent years, approaches genericallly called ``distirbutional learning'' towards learning mildly context-sensitive languages have been making many positive results. Our project developped the theory of ``distributional learning'' further and tackled learning even more complex grammar formalisms. We gave a uniform view on those existing learning algorithms for mildly context-sensitive languages and derived a general condition with which a grammar formalism shall be distributionally learnable. We targeted grammar formalisms more complex than mildly context-sensitive grammars, including conjunctive grammars, which may define the intersection of languages, and non-linear lambda grammars, which have copying production rules. Furthermore, we have succeeded in weakening the two representative conditions that make grammars distributionally learnable and showed that even richer classes of languages than those used to be defined are distributionally learnable.

研究分野：　文法推論，アルゴリズム論，形式言語理論

キーワード：　文法推論

１．研究開始当初の背景

形式文法のアルゴリズム的学習（文法推論）の研究の主要な基本的動機の一つとして，人間の母語獲得メカニズムの数理モデル化がある．文法推論では，文脈自由言語の効率的な学習は非常に重要かつ困難な課題であったが，近年の「分布学習」とよばれるアプローチがブレークスルーをもたらした．ここでいう分布とは確率分布ではなく，単語列と単語列が出現しうる文脈の間の共起関係を指し，この関係の分析に基づいて文法を構築する技法を総称して分布学習という．これまでに，様々な制約の下で．文脈自由言語の豊かな部分族を，効率的に学習する，多様な分布学習アルゴリズムが提案されてきた．分布学習の技法は，自然言語文を表現するために妥当な形式言語とされている弱文脈依存言語の学習にも応用可能であることがわかっている．しかし，幼児は単なる文の集合として母語を学ぶのではなく，文が発せられた環境を観察し，文や単語の意味理解と共に統語規則を学んでいく．このような意味表現の学習を意識した分布学習の研究は従来ほとんど行われてこなかった．

２．研究の目的

本課題研究では，母語獲得過程を，自然言語文とその表現する意味の対の学習としてモデル化し，分布学習の技法を発展させ，文とその意味表現を対として導出する文法形式を学習するアルゴリズムを設計することを目的とした．

本課題研究では，自然言語文は単純に文字列とし，意味表現としては，変数束縛を含む論理式を使って意味を表現するという仮定のもとで，特定の論理に依拠せずに一般性を保つためにラムダ項を用いる．

このような設定は，自然言語文とその意味表現の導出をモデル化した，範疇文法と総称される文法形式などにおいて見られる一般的なものである．

本課題研究で対象とする文法形式は，これまでの分布学習研究との親和性・一貫性から，自然言語文を生成する文脈自由文法と意味表現を担う文脈自由ラムダ文法を，導出木を共有するように結合したものを考える．この本提案課題の対象とする文法形式と，従来の分布学習が対象としていた典型的な文法形式との違いとして，以下が挙げられる．

(a) [構造分解の探索空間の爆発]
従来の分布学習が対象としていた文法形式は，導出過程で中間生成物が複製されることのない「線形型文法形式」が主であった．しかし自然言語の意味表現を生成する文法としては，再帰代名詞等の意味表現を実現するために導出過程で中間生成物が複製されうるような生成規則を用いる「非線形型文法形式」を考えることが一般的である．一般に分布学習では，与えられた正例に対して，可能な部分構造と文脈構造への分解}を列挙・分析するが，線形型文法形式においてはこのような分解は多項式個の可能性しかないことが通常である一方，非線形型文法形式では指数個にのぼる可能性がある．少なくともナイーブな解釈では指数爆発は避けられない．

(b) [合理的な関係の導入] 従来の分布学習の対象言語は，文字列や木など，単一種類の構造物の集合であったが，本問題設定では異なる構造物の対の集合となり，より複雑な概念を学習対象とする．そこで，新たに異なる構造物間の関係に関する制約を導入して学習可能性を議論することになる．

３．研究の方法

(a) に関して，我々の研究グループは，本研究開始前に，「並列正則木文法」について，分布学習が可能になる条件を発見しつつあった．並列正則木文法は，最も単純な部類の非線形型文法形式であり，木の集合を言語として導出する．我々は，複製された中間生成物が導出過程を経ても局所域に留まるような条件（局所複製条件）を与えた場合に，並列正則木文法の効率的な分布学習が可能になるとの見通しを得ていた．この知見を定式化し，これを手がかりにさらなる一般化を目指す．これらの制約を緩和し，あるいは参考にしながら，自然な意味表現と効率的分布学習を両立する制約も探っていく．また，ここでは，必ずしも課題を克服することのみを目標とするのではなく，課題の困難さの原因を詳細に明らかにすることも重要な目的である．この(a)については，本課題の問題設定において従来の分布学習の対象に比して学習がより困難になる要因である．一方で，文字列と意味表現の関係に関する(b)については，学習対象が組になって複雑になるためより洗練された議論が必要になるものの，しかし必ずしも学習を理論上より困難にする要素ではない．むしろ，文と意味の間に合理的な制約を考えることで，(a)や(b) の問題を克服できる可能性がある．文脈自由文法については，既存の分布学習手法が適用可能なクラスを設定し，文字列と意味表現の間の関係に強力な仮定を置くことで，これを梃子として，文字列の学習を経由して単独では学習困難な意味表現集合を同時に学習する方法を探る．

４．研究成果

(1) Conjunctive grammars と呼ばれる文脈自由文法の拡張にあたる文法形式に対する分布学習アルゴリズムの設計を行った．Conjunctive grammars は文脈自由文法に集合積演算を導入したものとみなすことができ，文脈自由文法では表現できない多様な言語を表現できる．分布学習には，部分文字列

を非終端記号の意味付けに用いる第１アプローチと，文脈を非終端記号の意味付けに用いる第２アプローチがあり，文脈自由言語の学習においてはこれらのアプローチは美しい対称形を持つことが知られているが，conjunective grammar においては，対称性が崩れ第２アプローチのみが有効であることがわかった（学会発表 [9]）.

(2) 分布学習が適応可能な文法形式一般に関する議論を行った．これまで文脈自由文法やその一般化とみなせるような多重文脈自由文法，文脈自由木文法といった各種の文法それぞれについて，分布学習概念が適用され議論されてきた．これに対して本研究代表者は，分布学習概念を適用可能な文法形式について一般的抽象的な枠組みを与えた．このような枠組みは，本課題で対象としている，文字列とラムダ項の対という特殊な数学オブジェクトに対しても有効であり，分布学習がいかに定義され議論されるべきか良い見通しを与えた．本成果は，国際研究会の招待講演（学会発表 [10])において発表された.

(3) 非線形ラムダ項を用いた文法形式について，分布学習が部分的に可能となる特殊ケースの条件に関する事例的成果をあげた．非線形性は分布学習を計算量的に困難にする要因であるが，文脈構造か部分構造の一方が非線形度が定数で抑えられる場合には，分布学習が部分的に可能である．そのような導出構造を与える文法規則の条件について考察した（学会発表 [8]）.

(4) さらに，非線形度が定数で抑えられない文法であっても，文脈構造と部分構造への分解が多項式時間で可能である場合を発見した．並列正則木文法がこのような性質を持っていること，さらに一般化した文脈自由木文法について，このような性質を持つ部分クラスが定性的に定義可能であることを証明した（雑誌論文 [3]）.

(5) 文脈自由言語を分布学習可能にする文脈自由文法に対する条件の最も代表的な２つとして，Finite kernel property (FCP) と Finite context property (FKP) が従来提案されていたが，より厳密にはこれらの条件にはそれぞれより強い条件のものとより弱い条件のものが提案されていた．学習のためには弱い意味のもので十分であることが知られていたが，果たして文法の言語表現能力として違いをもたらすのか否かについてはわかっていなかった．この未解決問題に対して，我々は，強い FKP と弱い FKP および強い FCP と弱い FCP の間で表現能力に違いがあることを明らかにした．また，典型的な分布学習アルゴリズムが学習を成功させる条件について，この弱い FKP がちょうど必要十分であり，一方で，弱い FCP については更に弱めた形がちょうど必要十分条件になることを明らかにした．これらは，分布学習可能という概念の本質に迫る研究となった （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 4 件）

[1] Seishi Ouchi, Tomohiko Okayama, Keisuke Otaki, Ryo Yoshinaka, Akihiro Yamamoto: Learning concepts and their unions from positive data with refinement operators. Ann. Math. Artif. Intell. (査読有) 79(1-3): 181-203 (2017)
DOI: 10.1007/s10472-015-9458-6

[2] Shuhei Denzumi, Ryo Yoshinaka, Hiroki Arimura, Shin-ichi Minato: Sequence binary decision diagram: Minimization, relationship to acyclic automata, and complexities of Boolean set operations. Discrete Applied Mathematics (査読有) 212: 61-80 (2016)
DOI: 10.1016/j.dam.2014.11.022

[3] Alexander Clark, Makoto Kanazawa, Gregory M. Kobele, Ryo Yoshinaka: Distributional Learning of Some Nonlinear Tree Grammars. Fundam. Inform. (査読有) 146(4): 339-377 (2016)
DOI: 10.3233/FI-2016-1391

[4] Chihiro Shibata, Ryo Yoshinaka: Probabilistic learnability of context-free grammars with basic distributional properties from positive examples. Theor. Comput. Sci. (査読有) 620: 46-72 (2016)
DOI: 10.1016/j.tcs.2015.10.037

〔学会発表〕（計 10 件）

[1] Yuki Igarashi, Diptarama, Ryo Yoshinaka, Ayumi Shinohara: New Variants of Pattern Matching with Constants and Variables. SOFSEM 2018 (査読有): 611-623

[2] Davaajav Jargalsaikhan, Diptarama, Yohei Ueki, Ryo Yoshinaka, Ayumi Shinohara: Duel and Sweep Algorithm for Order-Preserving Pattern Matching. SOFSEM 2018 (査読有): 624-635

[3] Hayato Mizumoto, Shota Todoroki, Diptarama, Ryo Yoshinaka, Ayumi Shinohara: An efficient query learning algorithm for zero-suppressed binary decision diagrams. ALT2017 (査読有): 360-371

[4] Makoto Kanazawa, Ryo Yoshinaka: The Strong, Weak, and Very Weak Finite Context and Kernel Properties. LATA 2017 (査読有): 77-88

[5] Yohei Ueki, Diptarama, Masatoshi Kurihara, Yoshiaki Matsuoka, Kazuyuki Narisawa, Ryo Yoshinaka, Hideo Bannai, Shunsuke Inenaga, Ayumi Shinohara: Longest Common Subsequence in at Least *k* Length Order-Isomorphic Substrings. SOFSEM2017 (査読有): 363-374

[6] Diptarama, Ryo Yoshinaka, Ayumi Shinohara: AC-Automaton Update Algorithm for Semi-dynamic Dictionary Matching. SPIRE 2016 (査読有): 110-121

[7] Diptarama, Ryo Yoshinaka, Ayumi Shinohara: Fast Full Permuted Pattern Matching Algorithms on Multi-track Strings. Stringology 2016 (査読有): 7-21

[8] Makoto Kanazawa, Ryo Yoshinaka: Distributional Learning and Context/Substructure Enumerability in Nonlinear Tree Grammars. FG 2015 (査読有): 94-111

[9] Ryo Yoshinaka: Learning Conjunctive Grammars and Contextual Binary Feature Grammars. LATA 2015 (査読有): 623-635

[10] Ryo Yoshinaka: General Perspective on Distributionally Learnable Classes. MOL 2015 (招待講演): 87-98

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
吉仲　亮（YOSHINAKA, Ryo）
東北大学・情報科学研究科・准教授
研究者番号：80466424

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
金沢　誠（KANAZAWA, Makoto）
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：20261886

(4)研究協力者
（　　）