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

令和　５ 年　 ６ 月　 ８ 日現在

機関番号：３２６７５
研究種目：基盤研究(C)（一般）
研究期間：2017～2022
課題番号：１７Ｋ００００２６

研究課題名（和文）形式言語の分布学習の理論の深化

研究課題名（英文）Advances in the Theory of Distributional Learning of Formal Languages

研究代表者
金沢　誠（Kanazawa, Makoto）

法政大学・理工学部・教授

研究者番号：２０２６１８８６

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：文脈自由言語の分布学習アルゴリズムでは，文法の非終端記号に対して，所属性質問によって判定できるような文字列集合を割り当てる。従来の研究では，有限個の所属性質問の論理積によって表せるような文字列集合を割り当てていた。これに対して2つの一般化が可能であることを示した。1つ目の一般化では，論理積のかわりに任意のブール結合を許す。2つ目の一般化では，ブール演算に加えて正規演算も許し，文法の各非終端記号を所属性質問に対応する原子式を含む一種の拡張正規表現で表す。これらの一般化により，学習の対象とできる文脈自由言語のクラスが飛躍的に広がった。

研究成果の学術的意義や社会的意義
いまだに謎に包まれている人間の母語習得のメカニズムの解明のためには，研究の指針となるような学習の数理モデルの確立が欠かせない。この観点から，母語習得のモデルとして一定の説得力を持つ学習の枠組みのもとで，どれだけ広い文脈自由言語の部分クラスが学習可能になるのかを調べることは，非常に重要な課題である。本研究は，正例と所属性質問からの極限同定の枠組みのもとで，従来の分布学習のアルゴリズムで目標言語とすることができる文脈自由言語のクラスを飛躍的に拡大することに成功した。

研究成果の概要（英文）：Distributional learning algorithms for context-free languages work by assigning to each nonterminal of the hypothesized grammar a string set that can be decided by making queries to the membership oracle for the target language. In previous works, these string sets were limited to those that can be represented by finite conjunctions of membership queries. The present study presented two generalizations. The first generalization allows arbitrary Boolean combinations of membership queries in place of finite conjunctions. The second generalization allows regular operations in addition Boolean operations, and represents each nonterminal by an extended regular expression containing atoms for membership queries. These generalizations greatly extend the class of context-free languages that can be targeted by distributional learning algorithms.

研究分野：形式言語理論，数理論理学

キーワード：文脈自由文法　文法推論　正例と所属性質問からの極限同定　分布学習　拡張正規閉包

１．研究開始当初の背景

文法推論（形式言語のアルゴリズム的学習）の研究の主要な動機の一つに人間の母語獲得の数理モデル化があり，このため，**正例からの学習可能性**が重要視されて来た。しかし，多項式時間のような計算資源に関する制限のもとで，正例からの効率的な学習可能性を示した結果は，正規言語やパターン言語のような非常に単純な文法形式を対象にしたものを除いて，極めて限定的なものにとどまっている。近年，Clark と Yoshinaka は，**正例と所属性質問からの多項式時間極限同定**の枠組のもとで，**分布学習**と呼ばれる手法が，通常の文脈自由文法のみならず，自然言語の記述に必要とされる多重文脈自由文法や並列多重文脈自由文法などのより複雑な文法形式に対しても，効率的な学習アルゴリズムをもたらすことを示した。この枠組で，**所属性質問**は，ある対象が学習の目標である未知の言語の要素であるかどうかを質問するもので，母語獲得における**間接的否定証拠**のモデルであるとされる (Clark and Lappin 2011)。

分布学習は，文法の非終端記号から導出される対象の**分布**に注目する。非終端記号から導出される対象は，文法が生成する言語に属する文の中に**部分構造**として出現する。文からある部分構造を取り除いたものが**文脈構造**である。たとえば，通常の文脈自由文法の場合，部分構造は文字列で，文脈構造は文字列の順序対であり，言語 L における文字列xの分布は，$D_L(x) = \{(u,v) \mid uxv \in L\}$と定義される。（より複雑な文法形式では，部分構造と文脈構造はより複雑な対象になる。）

分布学習のアルゴリズムは，入力として与えられた正のデータの中に観察される部分構造または文脈構造を要素に持つさまざまな有限集合を仮説文法の非終端記号として用いる。部分構造の有限集合を用いる方法を primal approach，文脈構造の有限集合を用いる方法を dual approach と呼ぶ。dual approach では，文脈構造の有限集合Cを非終端記号と見なし，$C^{\triangleleft} = \left\{ x \,\middle|\, \text{すべての } (u,v) \in C \text{ に対して } uxv \in L_* \right\}$を$C$の解釈とする。ここで，$L_*$は，学習の目標言語である。primal approach では，部分構造の有限集合Kを非終端記号と見なし，$K^{\triangleright\triangleleft}$を$K$の解釈とする。ただし，$K^{\triangleright} = \left\{ (u,v) \,\middle|\, \text{すべての } x \in K \text{ に対して } uxv \in L_* \right\} = \bigcap \{ D_{L_*}(x) \mid x \in K \}$である。このように非終端記号の意味を定めると，自然に文法規則の正しさが定義できる。例えば，dual approach で 通常の文字列上の文脈自由文法を扱う場合，規則$C_0 \to C_1\, C_2$が正しいのは$C_0^{\triangleleft} \supseteq C_1^{\triangleleft}\, C_2^{\triangleleft}$が成り立つときである。文法規則の正しさは，所属性質問によって判定できるような条件に全称量化が施された形をしているので，極限において判定でき，このことから，学習アルゴリズムは正しい文法規則のみからなる文法に収束することができる。Clark と Yoshinaka は，学習の目標文法の各非終端記号Xについて，Xから導出される対象の集合 $L(X)$ が文脈構造の有限集合C_Xまたは部分構造の有限集合K_Xによって**特徴付けられる**（つまり，$L(X)^{\triangleright\triangleleft} = C_X^{\triangleleft}$または$L(X)^{\triangleright\triangleleft} = S_X^{\triangleright\triangleleft}$が成り立つ）という前提が成立するとき，目標文法の各規則と同じ形の規則が学習アルゴリズムの出力する文法に含まれることになり，学習が成功することを示した。目標文法に仮定されるこの性質は，finite context property（C_Xが存在する場合）または finite kernel property（K_Xが存在する場合)と呼ばれる。（それぞれ，FCP, FKP と略す。）

研究代表者は，Yoshinaka, Clark らとの共同研究で分布学習の理論の発展に寄与して来たが，その中で，従前の分布学習の理論が非本質的な要素を多く含み，不必要に制限されたものになっていることに気付くことができた。特に， dual approach の場合，従来の FCP は，目標文法が満たすべき条件として不必要に強い条件であり，実際には，$X = C_X^{\triangleleft}$という解釈のもとで目標文法の各規則が正しいという仮定で十分であることを示した(Kanazawa and Yoshinaka 2017)。この条件は，従来の FCP よりも真に弱い条件であり，非終端記号を集合を表す変数とみなし，文脈自由文法を連立代数方程式として定義する定式化のもとでは，$X = C_X^{\triangleleft}$が連立方程式の **pre-fixed point** を構成するという条件に対応する。

２．研究の目的

次の 3 つの課題に取り組むことを目的とした。

(1)　**分布学習の学習対象の拡張**

分布学習が有効であるような文法形式は，広い意味での「文脈自由文法」に限られない。対象の集合に対する単調な演算の least fixed point により言語が定義されるような文法形式に対して広く有効なはずである。そのような文法形式の一つとして，すでに Yoshinaka (2015)が conjunctive grammar の分布学習が可能であることを strong finite context property の前提のもとで示していたが，この結果をさらに拡張して，他行式時間認識可能な文字列言語を特徴づける hereditary elementary formal system (Ikeda and Arimura 1997)などのより表現力の強い文法形式に対しても同様

の結果を得ることを試みる。

(2) **目標文法に対する前提条件の緩和**
分布学習において目標文法に対する前提条件を，従来の FCP/FKP （を pre-fixed point の概念を使って弱めたもの）から「有限回の所属性質問によって要素であるかどうかが判定できるような pre-fixed point が存在する」という形の条件に緩和できることを示す。pre-fixed point は，$C^{\triangleleft}$または$K^{\triangleright\triangleleft}$の形をしている必要はない。例えば，$C_1^{\triangleleft} \cup C_2^{\triangleleft}$や $C_1^{\triangleleft} C_2^{\triangleleft}$，あるいは$C_1^{\triangleleft} - C_2^{\triangleleft}$のような形を許すことも可能なはずである。

(3) **学習モデルの一般的な性質の解明**
分布学習は，特定の形をした学習アルゴリズムが幅広い形式言語のクラスを効率的に学習可能であることを示したものだが，分布学習の成果を適切に評価するために，正例と所属性質問からの極限同定という学習モデルに関する一般的理論を展開することを試みる。

３．研究の方法

東北大学の吉仲亮を連携研究者として研究を進めた。2017 年の国際会議 LATA での発表に対する聴衆の反応に触発され，当初の予定を変更して，平成 29 年度～30 年度は，まず次の 2 つの課題に取り組んだ。

(4) **閉包性の成立・不成立の解明**
分布学習のアルゴリズムが有効であるような文脈自由言語のクラスについて，よく知られた閉包性が成り立つかどうかを調べる。

(5) **congruential な文脈自由文法の研究**
FKP を持つ文法クラスの部分クラスであり，MAT (minimally adequate teacher)のもとで学習可能であることが知られている Clark の意味での congruential な文法の性質について調べる。

この 2 つの課題については，平成 29 年度に当時研究代表者が在籍していた国立情報学研究所にインターンとして訪問した Tobias Kappé（当時 University College London の博士課程学生）との共同研究を行った。

令和元年度～4 年度は，連携研究者の吉仲と密接に連携しながら，研究の目的の項で掲げた（2）の課題に取り組んだ。この課題の射程は，当初の想定をはるかに超えて大きく広がることになった。そのせいもあり，研究の目的の項で掲げた（1）と（3）の課題には取り組むことができなかった。

４．研究成果

次のような成果を上げた。番号は研究の目的と研究の方法の項で使った番号に対応している。

(4) **閉包性の成立・不成立の解明**

分布学習が有効なクラスには，大きく分けて FKP を満たす文法を持つ言語からなるクラスと FCP を満たす文法を持つ言語からなるクラス の 2 種類があるが，FKP には強いバージョンと弱いバージョンがあり，FCP には強いバージョン，弱いバージョン，さらに弱いバージョンの 3 つの定義がある。正規演算，正規言語との共通部分，準同型写像，準同型写像の逆像については，強いバージョンの FKP/FCP を満たす文法の両方を持つ言語からスタートしても，これらの演算を施した結果できる言語に対しては，もっとも弱いバージョンの FKP/ FCP が成り立たなくなることがあることがわかった。次に，正規言語との共通部分をとる演算について，演算の結果が FKP/FCP を満たすことを保証するようなもとの文法に対する条件をいくつか考案した。

(5) **congruential な文法の研究**

Clark の意味での congruential な文法の部分クラスである pre-NTS 文法は正規言語との共通部分をとる演算について閉じており，2 つの pre-NTS 文法が与えられたときにそれぞれが生成する言語が等しいかどうかを判定するアルゴリズムが存在することが知られている。congruential な文法に対しても同様に正規言語との共通部分をとる演算について閉じていることが示せた。そして，pre-NTS に対するアルゴリズムを修正して，congruential な 2 つの文法が与えられたときにそれぞれが生成する言語が等しいかどうかを判定するアルゴリズムを与えることができた。また，与えられた 1 つの文脈自由文法 G が決定性である（決定性プッシュダウンオートマトンに対応する）という仮定のもとで，G が congruential であるかどうかをアルゴリズムにより判定することができることを証明した。

(2) **目標文法に対する前提条件の緩和**

FCP は，文法が生成する言語 L に対して，

$$C^{\triangleleft} = \left\{ x \,\middle|\, \text{すべての } (u,v) \in C \text{ に対して } uxv \in L \right\}$$

の形をした集合を非終端記号に当てはめて文法規則の正しさを規定するものである。（ここで，C は文字列の順序対の有限集合である。）このかわりに，

$$\left\{ x \,\middle|\, \text{すべての } (u,v) \in C \text{ に対して } uxv \in L \text{ であり，すべての } (u,v) \in D \text{ に対して } uxv \notin L \right\}$$

の形をした集合（C, D は文字列の順序対の有限集合）を用いると，より広い言語のクラスを捉えることができ，この言語のクラスは，FCP の場合とまったく同様に正例と所属性質問を用いた学習アルゴリズムによって学習可能である。

有限集合 C と D の要素の数をそれぞれ高々 k と l に限定して得られる言語のクラスを FCP(k, l)とする。FCP(k, 0)が従来 k-FCP と呼ばれて来たクラスに相当する。k-FCP が k に関して階層 をなすように，FCP(k, l)は，l に関して階層をなすこと，具体的には，FCP(1, l + 1)に属し，どんな k についても FCP(k, l)に属さない言語が存在するという定理を証明することができた。同様に，FCP(k + 1, 0)に属し，どんな l についても FCP(k, l)に属さない言語が存在すると予想できるが，これについてはまだ部分的な結果（k が 5 以上の場合）しか得られていない。

学習可能性の観点からは，

$$\left\{ x \,\middle|\, \text{すべての } (u,v) \in C \text{ に対して } uxv \in L \text{ であり，すべての } (u,v) \in D \text{ に対して } uxv \notin L \right\}$$

の形の m 個の集合の和を考えることも自然であるが，このようにして FCP(k, l)の真の拡張が得られるかまだわかっていない。

上で述べた拡張では，文法の非終端記号に対して，文法が生成する言語Lの商集合（$\{x \mid uxv \in L\}$の形の集合）のクラスの**ブール閉包**の要素を対応させる。これに対して，L の商集合に適用できる演算として，ブール演算だけでなく，正規演算も許すことが考えられる。つまり，非終端記号に対して，L の商集合の**拡張正規閉包**の要素を対応させるのである。このように学習の対象となる文法のクラスを広げても，従来の分布学習のアルゴリズムと同様のアルゴリズムが成立することを示すことができる。

文法 G が，G が生成する言語の商集合のクラスを Γ に属する演算で閉じたクラス（Γ 閉包）の要素からなる pre-fixed point を持つとき，G は **Γ 閉包性**を持つと言う。従来の（一番弱い意味での）FCP は，$\Gamma = \{\cap\}$の場合の Γ 閉包性に対応する。Γがブール演算の集合のときは Γ 閉包性を**ブール閉包性**と呼び，Γがブール演算と正規演算からなる集合のときは Γ 閉包性を**拡張正規閉包性**と呼ぶ。ブール閉包性を満たす文法は持たないが，拡張正規閉包性を満たす文法を持つ文脈自由言語の例として，

$$\overline{O_1} = \left\{ x \in \{a,b\}^* \,\middle|\, x\text{の中の}a\text{の出現回数} \neq x\text{の中の}b\text{の出現回数} \right\}$$

がある。拡張正規閉包性を満たす文法を持つ文脈自由言語のクラスがどのようなものであるか，まだほとんどわかっていない。本質的に曖昧な文脈自由言語で，拡張正規閉包性を満たす文法を持たないものの例を示すことができたが，本質的に曖昧でない文脈自由言語で拡張正規閉包性を満たす文法を持たないものはまだ見つかっていない。拡張正規閉包性は，非常に大きな文脈自由言語の部分クラスに対応する可能性がある。また，拡張正規閉包性と **star-free 閉包性**（Γ としてブール演算と連接の演算からなる集合をとった場合の Γ 閉包性）が実質的に異なるのかわかっていない。上記の$\overline{O_1}$は star-free 閉包性を満たす文法を持つことが示せる。拡張正規閉包性と star-free 閉包性について詳しく調べることは，今後の課題である。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　3件）

１．著者名 Makoto Kanazawa and Ryo Yoshinaka	４．巻 153
２．論文標題 A Hierarchy of Context-Free Languages Learnable from Positive Data and Membership Queries	５．発行年 2021年
３．雑誌名 Proceedings of Machine Learning Research	６．最初と最後の頁 18-31
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Makoto Kanazawa and Tobias Kappe	４．巻 93
２．論文標題 Decision problems for Clark-congruential languages	５．発行年 2019年
３．雑誌名 Proceedings of Machine Learning Research	６．最初と最後の頁 3-16
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Makoto Kanazawa and Ryo Yoshinaka	４．巻 -
２．論文標題 Extending Distributional Learning from Positive Data and Membership Queries	５．発行年 2023年
３．雑誌名 Proceeding sof Machine Learning Research	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Makoto Kanazawa	４．巻 -
２．論文標題 Learning Context-Free Grammas from Positive Data and Membership Queries	５．発行年 2023年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計5件（うち招待講演　1件／うち国際学会　5件）

1．発表者名
Makoto Kanazawa and Ryo Yoshinaka

2．発表標題
A Hierarchy of Context-Free Languages Learnable from Positive Data and Membership Queries

3．学会等名
The 15th International Conference on Grammatical Inference（国際学会）

4．発表年
2021年

1．発表者名
Makoto Kanazawa and Tobias Kappe

2．発表標題
Decision problems for Clark-congruential languages

3．学会等名
The 14th International Conference on Grammatical Inference（国際学会）

4．発表年
2018年

1．発表者名
Makoto Kanazawa and Tobias Kappe

2．発表標題
Decision problems for Clark-congruential languages

3．学会等名
LearnAut 2018（国際学会）

4．発表年
2018年

1．発表者名
Makoto Kanazawa and Ryo Yoshinaka

2．発表標題
Extending Distributional Learning from Positive Data and Membership Queries

3．学会等名
The 16th International Conference on Grammatical Inference（国際学会）

4．発表年
2023年

１．発表者名 Makoto Kanazawa
２．発表標題 Learning Context-Free Grammars from Positive Data and Membership Queries
３．学会等名 The 29th Workshop on Logic, Language, Information and Computation（招待講演）（国際学会）
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
連携研究者	吉仲　亮 (Yoshinaka Ryo) (80466424)	東北大学・情報科学研究科・准教授 (11301)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
英国	University College London			