様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月１１日現在

機関番号：62615
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21500025
研究課題名（和文）　型つきラムダ計算とＤａｔａｌｏｇに基づく構文解析・文生成の研究

研究課題名（英文）　Parsing and Generation Based on Typed Lambda Calculus and Datalog

研究代表者
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

研究成果の概要（和文）：本研究代表者は、構文解析と文生成の問題を統一的に扱うため、「ほとんど線形な」ラムダ項の集合を生成する「文脈自由な」文法フォーマリズムに対して、与えられたラムダ項が与えられた文法によって生成されるかどうかを判定する問題をデータベース問い合わせ言語 Datalog の問い合わせに帰着する手法をすでに開発していたが、本研究では、これを、与えられたラムダ項の集合と与えられた文法の生成するラムダ項の集合が空でない交わりを持つかどうかという問題に拡張することを試みた。ラムダ項の集合がある意味で「決定的な」データベースで表される場合に、Datalog への帰着が成り立つことを示すことができた。これにより、入力として意味表現の集合を用いる generation as intersection の問題を Datalog に帰着させることができた。

研究成果の概要（英文）：As a uniform approach to parsing and generation, the principal investigator had previously developed a method of reducing the recognition problem for grammars generating lambda-terms to query evaluation in Datalog. This applies to “context-free” grammars on lambda-terms restricted to “almost linear” lambda-terms. This method has been extended to the case where the input is not just a single lambda-term, but a set of lambda-terms represented by a “deterministic” database. This allows the reduction of the problem of “generation as intersection” to Datalog.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	700,000	210,000	910,000
２０１０年度	600,000	180,000	780,000
２０１１年度	600,000	180,000	780,000
年度			
年度			
総　計	1,900,000	570,000	2,470,000

研究分野：総合領域
科研費の分科・細目：情報学・情報学基礎
キーワード：構文解析、文生成、Ｄａｔａｌｏｇ、型つきラムダ計算、ほとんど線形なラムダ項、ACG、文脈自由ラムダ項文法

１．研究開始当初の背景

(1) 形式言語の研究において、文脈自由文法よりも高い記述力を持つが多項式時間で解析できるような文法形式が多く考案されて来たが、それらのうちの多くが持つ共通の特徴は文脈自由文法と同様の導出木の概念を

持つことである。10 年ほど前、de Groote (2001)によって多くの文法形式を統一的に捉える枠組みとして、抽象的範疇文法(ACG)が提案された。ACG は、通常の文法の導出や規則にあたるものを線形ラムダ項によって表現し、線形ラムダ項の集合を定義する。文字列や木は線形ラムダ項によって自然に表現できるので、ACG は文字列言語を定義する文法や木言語を定義する文法の一般化になっている。導出の形がラムダ抽象を含まない項（すなわち木）に制限された２階 ACG は、文脈自由文法(CFG)、木接合文法(TAG)、単純文脈木文法(simple CFTG)、多重文脈自由文法(MCFG)、多成分木接合文法(MCTAG)などを自然に表現できることがわかった(de Groote 2002, de Groote and Pogodalla 2004)。

(2) De Groote (2001)が定義した ACG においては、文法で扱うことのできるラムダ項はラムダ抽象が常にちょうど一つの変数出現を束縛する線形ラムダ項に限られていた。この制限のもとで、semilinear な言語を定義する CFG, TAG, MCTAG などの文法形式が ACG によって自然に表現できるが、一方、一般に semilinear でない言語を定義する IO-CFTG や並列多重文脈自由文法(PMCFG)を表現するには線形でないラムダ項を許すような ACG の拡張が必要になる。また、モンタギュー意味論を備えた文法に対する生成の問題は、適格な意味表現の集合を定義する文法に対する構文解析の問題と理解することができるが、後者は任意の型つきラムダ項を許す２階非線形 ACG と等価になる。本研究の研究代表者である金沢は、2007 年に発表した論文(Kanazawa 2007)で、２階 ACG のラムダ項に対する制限を線形ラムダ項からほとんど線形なラムダ項にゆるめた「ほとんど線形な文脈自由ラムダ項文法(CFLTG)」（２階疑似線形 ACG）を取り上げた。ほとんど線形なラムダ項は、原子型の変数に限ってラムダ抽象が２つ以上の変数出現を束縛することを許す。ほとんど線形な CFLTG は、IO-CFTG や PMCFG、さらにモンタギュー意味論のかなり広い範囲を表現できるので、構文解析と生成を統一的に捉える枠組みを提供する。Kanazawa (2007)は、線形ラムダ項についてよく知られている性質のいくつかがほとんど線形なラムダ項についても成り立つことを用いて、ほとんど線形な CFLTG が定義するラムダ項の集合の所属問題が、文法によって定まる Datalog プログラムと入力ラムダ項によって定まるデータベースに対する問い合わせによって表現できることを証明した。これは、よく知られている文脈自由文法の definite clause grammar による表現の一般化である。Datalog の問い合わせの評価は、データベースのサイズに関して多項式時間で計算できるが、ほとんど線形な CFLTG を表現する Datalog プログラムに限定した場合、導出木が多項式サイズになるため、所属問題を表す問い合わせが真と評価されるようなデータベースの集合が P の部分クラスである LOGCFL に属することがわかった。入力ラムダ項から対応するデータベースへの変換が logspace で計算可能であることも示すことができるため、結果として、ほとんど線形な CFLTG の定義するラムダ項の集合の所属問題が LOGCFL に属する（すなわち、この文法で表現できる範囲で、構文解析と文生成の問題がいずれも LOGCFL に属する）ことが帰結した。

(3) 構文解析・文生成の問題が Datalog の問い合わせで表現できれば、Datalog の問い合わせに対する効率的な評価アルゴリズムを構文解析・文生成に応用することが可能になる。そのような効率的なアルゴリズムのひとつにマジックセット書き換えというプログラム書き換えを用いたボトムアップ式評価がある。この書き換えを文脈自由文法を表す Datalog プログラムに適用すると、Earley の構文解析アルゴリズムと実質的に同じものが得られることが知られている。他の文法形式に対する構文解析やモンタギュー意味論に対する生成の問題を表す Datalog プログラムに対して同じ方法を用いることによってこれらの広いクラスの形式に対して Earley 型のアルゴリズムを自動的に得ることができる。ただし、文脈自由文法より複雑な文法形式に対する構文解析にこの方法を適用して得られる Earley 型アルゴリズムは、一般に、誤りをもっとも早い時点で検出するという correct prefix property を満たさない。金沢は、2008 年に発表した論文(Kanazawa 2008)で、MCFG を表す Datalog プログラムの場合、マジックセット書き換えの前にもうひとつの簡単な書き換えを適用することによって prefix-correct な Earley 型アルゴリズムが得られることを示した。この手法は、MCFG や TAG に対してこれまで提案された Earley 型アルゴリズムと比較してきわめて単純明快であり、アドホックな論法を用いることなく、アルゴリズムの正しさを簡単に示すことができる。

(4) 本研究の海外共同研究者である Salvati は、論文 Kanazawa 2007 にヒントを得て、任意の CFLTG（２階非線形 ACG）について、その認識問題が決定可能であることを証明した。（後に Salvati 2010 として発表。）これは、Montague 意味論に基づく文生成の問題が、決定可能であることを意味する。

２．研究の目的

本研究の目的は、上で示した研究代表者のこれまでの研究成果をさらに拡張・発展させることであった。次の３つの方向への拡張を目指した。

(1) 扱えるラムダ項に対する制限の緩和 Salvati の研究によって、任意の型つきラムダ項を許す CFLTG については、認識問題は決定可能だが計算量は non-elementary であることがわかっていた。Datalog の問い合わせはプログラムを固定した場合 P 完全であり、LOGCFL とは開きがあるため、「ほとんど線形」という制限をさらに緩めても、何らかの形で CFLTG から Datalog への帰着が成り立つ可能性がある。

(2) 単独のラムダ項からラムダ項の認識可能集合への拡張　CFG や、TAG などの CFG の緩い拡張に対する構文解析においては、入力として１つの文字列でなく文字列の認識可能集合を用いた場合も同様の手法が有効であることが知られている。このような問題の拡張は、parsing as intersection と呼ばれる。文生成においても、意味表現の集合を入力として用いることは、underspecified representation の表現として有効であると考えられる。Kanazawa (2007)で、すでに入力として木の認識可能集合を用いた場合について、Datalog による表現が可能であることが指摘されていた。これを、ほとんど線形可能な認識可能集合に拡張する。

(3) Prefix-correct な Earley 型アルゴリズムの一般化　マジックセット書き換えともうひとつの単純な書き換えを合わせて用いて prefix-correct な Earley 型アルゴリズムを得る手法を MCFG から PMCFG、さらにはほとんど線形な CFLTG が生成する文字列集合に拡張することを試みる。

３．研究の方法

上の「研究の目的」で述べた(2)の問題に最初に取り組み、その結果を応用して(1)と(3)の問題に取り組むことを計画した。これと同時に、すでに 2008 年度までの研究で得られている成果も含めて、これまでの成果を詳細に記述した論文の執筆を進めた。

４．研究成果

(1)「研究の目的」の(2)の問題について、入力を１つのラムダ項から、ある自然な意味で「決定性」のデータベースによって表されるラムダ項の集合に一般化した場合について、Datalog 問い合わせへの帰着が成り立つことを示すことができた。この成果について以下に詳しく述べる。

まず、ある高階シグネチャー Σ を固定したとき、Σ に含まれる定数記号を述語記号に使ったデータベース D と型 α に対して、ラムダ項の集合 $\Lambda(D, \alpha)$ を定義した。CFTG G から定まる Datalog プログラム P とデータベース D から $S(\vec{\alpha})$ が導出可能であることが、G の導出木に結びつけられたラムダ項で $\Lambda(D, \alpha)$ に属するものが存在することと同値になる。（ただし、α は S と結びつけられた型であり、$\vec{\alpha}$ は α から→を取り除いて得られる原子型の列である。）さらに、文法 G がほとんど線形のとき、$\Lambda(D, \alpha)$ が non-erasing かつ almost non-duplicating な β 簡約の逆について閉じていれば、同じ条件が G が生成する集合と $\Lambda(D, \alpha)$ が空でない交わりを持つことと同値になることがわかる。したがって、$\Lambda(D, \alpha)$ が non-erasing かつ almost non-duplicating な β 簡約の逆について閉じているような D を認識問題の入力に用いればよいことになる。

このための十分条件として、「決定性のデータベース」という概念を定義した。決定性の文字列オートマトンや決定性のボトムアップ木オートマトンをデータベースで自然に表現すると、決定性のデータベースが得られる。また、与えられたデータベースが決定性であるかどうかという問題は、アルゴリズムで判定可能である。

決定性のデータベースを使って、入力として意味表現の集合を用いる generation as intersection の問題を Datalog 問い合わせに帰着できることを示した。例えば、文法を固定した上で、与えられた文が持つ解釈の集合が決定性のデータベースで表現できる場合、その文と少なくとも１つの解釈を共有する文をすべて求めることが、Datalog 問い合わせに対するアルゴリズムを使ってできることになる。しかし、与えられた文の持つ導出木の集合から、対応する解釈の集合を表す決定性のデータベースを求めるアルゴリズムは簡単ではなく、今後の課題である。

(2)　「研究の目的」の(1)の問題について、上の(2)の問題の解決を通してアプローチすることが可能である。たとえば、チャーチ自然数にように、ほとんど線形でないラムダ項を木の認識可能集合で表現することが可能な場合がある。与えられた文法がほとんど線形でないラムダ項を生成する場合に、これをほとんど線形なラムダ項を生成する文法とほとんど線形なラムダ項からほとんど線形でないラムダ項への準同形写像とに分解することができれば、(2)の問題の解決を応用して、ほとんど線形でないラムダ項を生成する文法に対しても Datalog 問い合わせへの帰着を行うことが可能になる。この手法は、並列多重文脈自由文法や、IO マクロ文法に対して確かめることができた。しかし、入力が文字列

でないより一般の場合について、興味深い応用例を見つけることはできなかった。
「研究の目的」の(3)の問題については、十分な時間を割くことができず、成果は上がらなかった。

(3) 以上の研究成果と、すでに 2008 年度までに得られていた成果を合わせて、74 ページの詳細な論文を完成させ、専門誌に投稿すると同時にウェブサイトで公開した。Datalog への帰着の証明を始めとして、厳密な記述は、2007 年の論文(Kanazawa 2007)では記載するスペースがなかったため、数学的な裏付けとなる理論を公表したのは初めてである。細かい点で 2007 年の時点で解決していなかった部分が見つかったところも修正して完全な証明を提示した。また、今回得られた新たな成果である generation as intersection に関わる結果も、簡潔に記述した。この論文は 2011 年 8 月に完成し投稿したが、現時点でまだ査読結果は返って来ていない。
さらに、この論文に書ききれなかった結果として、ほとんど線形なラムダ項と各原子型の負の出現の数が 1 つ以下に限られるような型づけがちょうど対応することを証明した論文を執筆し、NII Technical Report として発表した。

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 1 件）
① Makoto Kanazawa, Almost Affine Lambda Terms, NII Technical Reports, 査読なし, NII-2012-03E. May,2012. http://www.nii.ac.jp/TechReports/12-003E.html

〔学会発表〕（計 2 件）
① Makoto Kanazawa, Datalog as a Uniform Framework for Parsing and Generation, Workshop: Parsing with Categorial Grammars, 21st European Summer School in Logic, Language and Information, Bordeaux, France, July 24, 2009.
② Makoto Kanazawa, Generation as Intersection and Datalog, ACG@10: Ten Years of Abstract Categorial Grammars, Bordeaux, France, December 7, 2011.

〔その他〕
ホームページ等
これまでの研究成果について詳細に記述した論文が次の URL よりダウンロード可能。
http://research.nii.ac.jp/~kanazawa/publications/pagadqe.pdf

６．研究組織
(1)研究代表者
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

(2)研究分担者
なし

(3)連携研究者
なし