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　元 年　 ６ 月 ２８ 日現在

機関番号：３２６９２
研究種目：若手研究(B)
研究期間：2014～2018
課題番号：２６７３０１２３

研究課題名（和文）Distributional学習に対するノンパラメトリックベイズの適用と応用

研究課題名（英文）Application of Nonparametric Bayesian Methods to Distributional Learning

研究代表者
柴田　千尋（SHIBATA, Chihiro）

東京工科大学・コンピュータサイエンス学部・講師

研究者番号：００６３３２９９

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：Distributional Learning の基本的な考えに基づき，Hierarchical Pitman-Yor Processesと呼ばれる階層化ノンパラメトリックベイズのモデルを用い，(k,l)-文脈依存確率をもつ文法構造を提案した．また，同時に，推定のための高速なMCMC手法を提案した．実際に実験を通して，高い予測精度を得ることができることを示した．最終的な成果として，教師なし構文解析の高精度化のためのアルゴリズムを開発することができた．

研究成果の学術的意義や社会的意義
近年の深層学習の発展により，自然言語処理の分野においても，様々なことが加速度的に可能となってきている．例えば，機械翻訳がその代表的な例である．一方で，現在の手法では，大量の教師ありのデータを必要とする．構文解析であれば，構文木を含む構文情報が付与された文のセットが必要となる．本研究成果は，構文木など教師となるものがまったくなくても，構文解析を統計的な学習手法を用いて行うものであり，特に，確率文脈自由文法の学習の観点から見て，より制度の高い確率構造と学習アルゴリズムを新規に提案している．

研究成果の概要（英文）：Based on the basic idea of distributed learning, we propose a grammatical structure with (k, l)-context-dependent probabilities using a hierarchical nonparametric Bayesian model called Hierarchical Pitman-Yor Processes. At the same time, a fast MCMC method for the estimation is proposed. In experiments, it was shown that high prediction accuracy could be obtained. As a final result, we developed the algorithms for improving the accuracy of unsupervised parsing.

研究分野：機械学習

キーワード：機械学習　自然言語処理　構文解析　ノンパラメトリックベイズ

１．研究開始当初の背景

ノンパラメトリックベイズ推定により，形式言語を学習する手法は，教師なし学習の手法として，最も精度の高い有望な手法の一つである．とくに，文脈自由文法や範疇文法など，記号列に潜む木構造の生成規則に対する教師なし学習については，未だ多くの課題が残されていた．一方，Distributional 学習とは，記号列から，その木構造をなす生成規則の教師なし学習手法として，有望なものであったが，精度の点で，他の統計的な手法と比べ，まだまだ改善の余地がある状況であった．

２．研究の目的

本研究課題の目的は，Distributional 学習の手法に対し，ノンパラメトリックベイズの手法を取り入れた手法を開発することで，生成された文法規則の，言語モデルとしての精度や，予測された構文の正確さを向上させることである．ここでいう言語モデルとしての精度とは，テストデータにおける，次の単語の予測確率を意味する．また，構文の推定を行う際，MCMC などのサンプリングにより，推定を行うが，一般的に言って，サンプリングを行うには非常に計算コストがかかる．したがって，本研究課題では，計算量の算出および，計算量を削減できるアルゴリズムを開発することも，もうひとつの目的となる．

３．研究の方法

本研究では，次の４つを順に行う．また，実証のために，アルゴリズムの実装を行い，自然言語などの比較的規模の大きい実データを用いて実証実験を行う．

(1) まず，ノンパラメトリックベイズを取り入れた弱文脈依存の確率的生成文法を考案することにより，推定すべき確率言語のクラスを決定する．

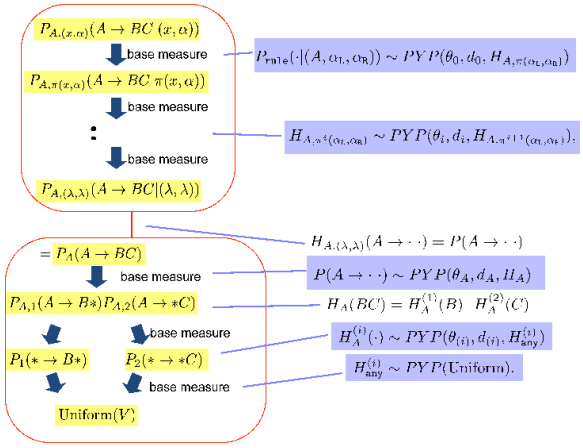
(2) 次に，定式化を行い，既存の MCMC サンプリングのアルゴリズムをナイーブに本研究の確率モデルに適用したときに，イテレーションあたり，どの程度の計算量になるのかを明らかにする．また，実装を行い，実際の計算速度および混合速度を計測する．

(3) その後，具体的にどのようなアルゴリズムを改良すれば，計算量を減少させることができるのかを追求する．また，改良したアルゴリズムを実装し，その速度を計測する．

(4) 最終的に，サンプリングが高速に行うことができるのか，および，その並列化の可能性について，検討を行う．

４．研究成果

(1) 右図に，本研究で対象とする 階層化 Pitman-Yor 過程(HPYP)をしめす．これは，研究代表者により提案されたものであり，(k,l)-文脈依存確率文法と呼ばれるものである．k, l はそれぞれ，左文脈，右文脈の長さを示す．k, l が共に 0 であるとき，無限確率文脈自由文法(I-PCFG ： Infinite Probabilistic Context-free Grammar)と等価となる．この PYP の階層は，意味的には，徐々に文脈を短くして，データがスパースであった場合のスムージングをしていると捉えることができる．階層を下ると文脈が徐々に少なくなり，最終的に無くなると I-PCFG へ帰着される．



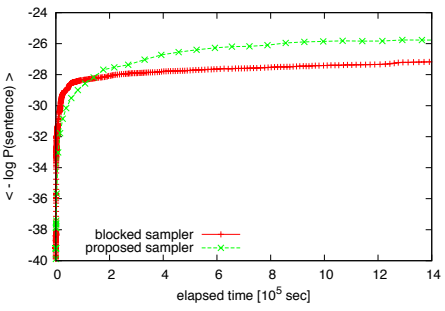
(2) 上記のような確率モデルに対し，構文木の推定をおこなう．推定には，MCMC のアルゴリズムを用いる．Blocked Gibbes Sampling と呼ばれる手法を，本確率モデルに適用した場合のアルゴリズムを Algorithm 1 に示す．文ごとに，Inside Outside アルゴリズムを用い，拡張された内側確率，および外側確率を計算し，それらの値に基づき，新しい木をサンプルする．計算量は，$O(|V|^{l+3}|w|^3)$ となる．$|V|$ は非終端記号の数，$|w|$ は文長を表す．右文脈の長さ 1 が $|V|$ の肩にかかっており，右文脈を考慮すると，非終端記号がごく少ない場合を除き，非常に計算が遅くなることがわかる．したがって，実際の実装においては，左文脈のみを考慮するほうが良いことがわかる．

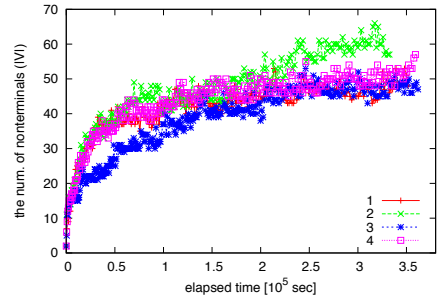
Algorithm 1: Blocked sampling; draw T according to $P(T|w)$.

Input: A sentence w, a CFG in CNF G, and a (k,l)-context-sensitive probability P.
Output: A derivation tree $T \in \mathcal{T}(w)$.

The inside probabilities P^{IN} is calculated recursively.;
Let $Z \in (V\mathbb{N}^2)^*$ be a sequence of pairs of nonterminals and intervals representing positions of substrings. ;
$Z := \langle S, 1, |w| \rangle$ and $D = \lambda$;
while $Z \neq \lambda$ **do**
  Let $\langle A, i, j \rangle$ be the first element in Z. ;
  Trim $\langle A, i, j \rangle$ from Z. ;
  if i=j **then**
    Add $A \to w(i, i)$ to the end of D. ;
  else
    Sample $B, C \in V$ and $k \in \mathbb{N}$ with a probability proportional to $f_{(w(1,i-1),\alpha)}(B, C, k)$, where α is the sequence of nonterminals in Z. ;
    Add $A \to BC$ to the end of D. ;
    $Z := \langle B, i, k \rangle \langle C, k+1, j \rangle Z$
  end
end
Make the left-most derivation d_T from D.;
Make the derivation tree T identified by d_T.;

(3)及び(4) 本研究では，上記のブロックサンプリングに対して，合成サンプリングと呼ばれる手法を提案した．これは，CYK テーブルにおいて，構文木の形(shape とよぶ)と，各ノードに割り当てる非終端記号とを，別々にサンプリングする手法である．これにより，計算量が $O(|V|^1|w|^3 + |V||w|^2)$ に減少する．一方で，混合速度自体は低下すると考えられる．右図上は，実際の自然言語のデータ(Brown Corpus)に対して本研究で構築した手法を適用した場合の，経過時間に対する，精度(テストデータにおける次単語の予測確率)の変化の様子である．また，右図下は，$|V|$ の変化の様子を表す．混合サンプリング(図中 proposed sampler) のほうが ブロックサンプリングに比べて，混合速度が劣るため，$|V|$ が小さい初期段階においては精度の上昇速度が遅いが，$|V|$ が大きくなるに連れてサンプリング速度の差が大きくなり，最終的には精度で大きく上回っていることがわかる．





また，並列計算の可能性については，CYK テーブルの構築の際に，行列演算を行うため，その部分での並列化は可能であるものの，最も計算コストがかかる，個々の HPYP の確率の計算においては，並列計算はそれほど有効ではない．したがって，現状では，並列化の際には，単純ではあるが，文レベルでの並列化が有効であるといえる．

５．主な発表論文等

〔雑誌論文〕（計　5　件）

(1) Chihiro Shibata and Jeffery Heinz, ″Subregular Complexity and Deep Learning″. In proceedings of Conference on Logic and Machine Learning in Natural Language (LaML). 2017/6(査読有)
(2) Chihiro Shibata and Jeffery Heinz, ″Predicting Sequential Data with LSTMs Augmented with Strictly 2-Piecewise Input Vectors″, JMLR: Workshop and Conference Proceedings : ICGI, 57:137-142. 2016/10(査読有)
(3) Chihiro Shibata and Ryo Yoshinaka, ″Probabilistic learnability of context-free grammars with basic distributional properties from positive examples″. Theoretical Computer Science 620:46-72, Elsevier. 2016/3(査読有)
(4) Chihiro Shibata, ″Inferring (k,l)-Context-Sensitive Probabilistic Context-Free Grammars using Hierarchical Pitman-Yor Processes″, JMLR Workshop and Conference Proceedings : ICGI, 34: 153-166, 2014/9 (査読有)
(5) Chihiro Shibata and Ryo Yoshinaka, ″A Comparison of Collapsed Bayesian Methods for Probabilistic Finite Automata″, Machine Learning, 96(1): 155-188, Springer. 2014/6 (査読有)

〔学会発表〕（計　3　件）

(1) 岡本(柴田)千尋, 内海慶, 持橋大地, 構文情報を陽に与えたときの LSTM-RNN による内部表現について, 第 237 回自然言語処理研究会 2018/9
(2) 加藤和樹, 柴田千尋, 田胡和哉, 文のテンプレートの学習および感情を考慮した会話文の生成情報処理学会,　第 77 回情報処理学会全国大会 2015/3
(3) 柴田千尋, 階層化 Pitman-Yor 過程を用いた文脈を考慮した確率文脈自由文法の推定-分布学習の実データへの適用にむけて-,　第 17 回情報論的学習理論ワークショップ(IBIS2014) 2014/11

６．研究組織