様式 C-19

科学研究費補助金研究成果報告書

平成 23 年　6 月 17 日現在

機関番号：25301
研究種目：若手研究（B）
研究期間：2009～2010
課題番号：21700007
研究課題名（和文）　文法推論における教示可能性と自然言語処理への応用

研究課題名（英文）　Teachability on Grammatical Inference and its Applications for Natural Language Processing

研究代表者
但馬　康宏（TAJIMA Yasuhiro）
岡山県立大学・情報工学部・准教授
研究者番号：00334467

研究成果の概要（和文）：
文脈自由言語の部分言語族のひとつである単純決定性言語について，そのある部分言語族は多項式教示可能であることを示した．この結果は，多項式時間での教示可能性と多項式時間での質問による学習可能性の本質的な違いの一例となっている．
さらに，本研究により得られたアルゴリズムを文書の段落分割アルゴリズムに応用し，テキストデータを話題に応じた段落に分割する手法を開発した．この手法は，分割精度において従前の良く知られた手法よりも高性能であることが実験的に示された．
またこれらを思考ゲームの着手決定アルゴリズムに応用し得ることを示し，実験的にその有効性を示した．

研究成果の概要（英文）：
In this study, we developed a new learning algorithm for a subclass of context-free languages. The main result during the supported term is that we showed polynomial time teachability of a subclass of simple deterministic languages. This language class is polynomial time teachable but it would be hard to learn from example based queries in polynomial time. Thus, our result shows a difference between teaching and learning via queries. Next, we applied our algorithm to natural language processing, then we developed a text segmentation algorithm for conversations and showed the performance improvement. In addition, we applied our algorithm to game tree search. Then we showed an improvement of reinforcement learning algorithm for making the evaluation function. We also confirmed the learning performance is improved by our algorithm.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	900,000	270,000	1,170,000
2010 年度	900,000	270,000	1,170,000
年度			
年度			
年度			
総　計	1,800,000	540,000	2,340,000

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：教示可能性，質問による学習，テキストセグメンテーション，単純決定性言語

１．研究開始当初の背景

機械学習の研究において教示可能性とい

う考え方の研究が，いくつかの独立した定義に基づいて近年盛んに研究が行われている．本研究では，形式言語に対する教示可能性を研究目的とし，形式言語にふさわしい特徴集合および時間計算量を明らかにする．その結果，新たな学習アルゴリズムを提案し，自然言語の大規模データに対する応用手法を開発する．

２．研究の目的

一般的な機械学習アルゴリズムの研究では，サンプルの提示を受け，そのサンプル集合をよりよく説明する仮説を出力するアルゴリズムを考察する．したがって，提示されるサンプルに対する仮定はあまり多くなく，確率的近似(PAC)学習などにおいては任意の分布のもとで考えることが一般的である．しかし，教示可能性の研究においては，その枠組みを一般化し，どのようなサンプル集合が与えられれば，効率よく学習をおこなえるかを学習アルゴリズムとペアで考察するパラダイムである．形式言語に対する学習アルゴリズムの研究では，正則言語の完全提示による多項式時間学習不可能性(Angluin, Machine Learning, 1991)，文脈自由言語の所属性質問と等価性質問による多項式時間学習の不可能性(Angluin&Kharitonov, J.of Comp.& Sys.Sci.,1995)など，その困難さを示す結果が多く知られている．しかし，教示可能性の研究となると，現在知られている結果は少ない．教示可能性の定義は過去にいくつか独立した定義のもとで行われており，サンプルに無矛盾な仮説を生成するアルゴリズムならばどんなアルゴリズムでも正解を提示するような特別なサンプル集合(Shinohara&Miyano,NewGene.Comp.,1991)や，敵対的な教師がどのような妨害的なサンプルを加えても学習が成功するような特別なサンプル集合(Goldman&Mathias, J.of Comp.&Sys.Sci.,1995)などに対して研究が行われている．以上を踏まえ，形式言語に対する教示可能性を研究目的とし，形式言語にふさわしい特徴集合および時間計算量を明らかにする．その結果，新たな学習アルゴリズムを提案し，自然言語の大規模データに対する応用手法を開発することを目的とする．

３．研究の方法

(1)既存のアルゴリズムの発展による教示可能性および評価パラメータの検討

我々は既に，単純決定性言語に対して，特殊なサンプル集合と所属性質問による効率的な学習アルゴリズムを提案している．この結果は，そのまま単純決定性言語の教示可能性につながるが，我々のアルゴリズムで用いている特殊なサンプルのサイズは，非終端記号から生成される最短の記号列長を利用しているため，(Goldman&Mathias, J.of Comp.&Sys.Sci.,1995)の教示可能性の定義では多項式で抑えられない．しかしこれは，ある種の単純決定性文法が生成できる記号列の長さは，非終端記号の種類の超多項式下界を持つという文法の性質から示されたものである．このパラメータを学習において必要なパラメータとみなすことは不自然ではなく，評価のパラメータとして妥当性がある．

このように，形式言語に本質的な量は，学習パラメータに取り込むことが妥当な場合があり，その結果新たな言語の教示可能性を示すことができる．また，我々のアルゴリズムは，線形言語の部分言語族に対しても有効であることが示されている(Tajima et al.,IPSJ-SIG,2005)．この結果を応用し，単純決定性言語の場合と同様に形式言語における教示可能性にふさわしい線形言語の評価パラメータを特定する．線形言語に関しては，多項式サイズでの教示が不可能であることが示されているが，その部分言語族に対する研究はいくつか行われており，有用性がある．

さらに，他の言語族に対して応用が可能か否かを調べる．

(2)自然言語処理への応用

自然言語処理において，対話などのコーパスを段落に分割することは基本的な問題である．従来の研究における段落分けでは，与えられた文書の単語出現に関する特徴量を文章ごとに算出し，その特徴量の変化点を観察する方法，および音声認識技術に端を発するHMMなどが多く用いられている．本研究では我々が以前提案した，文章を分類器によりひとつの記号に対応させ，文の数と同じ数の記号をもつ記号列として文書全体をとらえる手法を用いる．本研究の成果を用いて対話およびニュース記事などの一般的な文書の段落分割を行う．この手法は従来よりも効率的かつ高性能であり，HMMを文法推論アルゴリズムに置き換えることによりさらに性能向上が見込まれる．

(3)本研究による学習アルゴリズムの新たな適用領域の開拓

本研究による学習アルゴリズムの構築法をゲームのアルゴリズムに適用する．思考ゲームにおけるアルゴリズムに欠かせない局面の評価関数について，強化学習などを用いて自動作成することは広く行われているが，その構成方法に本研究による成果を反映させる．具体的には，強化学習における報酬設定を本研究による知見からより適切なものに設定する．

４．研究成果

(1) 既存のアルゴリズムの発展による教示可能性および評価パラメータの検討

以上の課題に対しては，文脈自由言語の部分言語族のひとつである単純決定性言語について，そのある部分言語族は多項式教示可能であることを示した。この部分言語族は，所属性質問と等価性質問を用いて，アルファベットの要素数を定数とみなせば多項式時間で学習可能であるが，これを変数とみなす場合は超多項式下界であると見込まれる．すなわち，本研究により示した結果は，多項式時間での教示可能性と多項式時間での質問による学習可能性の本質的な違いの一例となっている．本研究内容は電子情報通信学会論文誌に Teachability on a subclass of simple deterministic languages として投稿中である．以下本結果について詳細を述べる．

単純決定性言語とは，単純決定性文法であらわされる言語族であり，単純決定性文法とは，グライバッハ標準形の文脈自由文法において生成規則の右辺が，左辺の非終端記号と右辺の先頭の終端記号を見るだけで文法内で一意に決定できる形式文法である．本研究ではこの単純決定性文法に対してさらに，右辺の非終端記号列の長さが先頭の終端記号によって文法内で一意に決定できるという制限を課し，スタックユニフォーム単純決定性文法と定義した．このスタックユニフォーム単純決定性文法で表現される言語族に対する，以下の質問を用いた学習アルゴリズムを示した．

１．所属性質問

２．等価性質問

この学習アルゴリズムは，スタックユニフォーム単純決定性文法ならばある与えられた記号列に対する導出木のスケルトンが多項式時間で一定数に絞り込めるという性質を利用している．まず，学習アルゴリズムの内部で，我々が以前に提案した構造反例付き等価性質問を用いた単純決定性言語に対する多項式時間学習アルゴリズムを動作させる．その内部アルゴリズムが構想反例付き等価性質問を行うたびに，外側の学習アルゴリズムでは，得られた仮説をそのまま用いて一般の等価性質問を行う．反例として終端記号列が与えられたら，その記号列を生成できる導出木のスケルトンを学習アルゴリズム内で絞り込み，その種類の数だけ内部学習アルゴリズムを並列化する．それぞれの内部学習アルゴリズムに対して，導出木のスケルトンという形で反例を渡す．この動作によりスタックユニフォーム単純決定性言語に対する上記１．２．の質問を用いた学習アルゴリズムが構成できる．この学習アルゴリズムの時間計算量は，学習対象となる文法の非終端記号集合のサイズ，反例として与えられる記号列のサイズに関する多項式であるが，終端記号の集合のサイズに関しては指数関数的である．したがって，一般の多項式時間学習アルゴリズムの枠組みからは外れたものとなる．この時間計算量は，与えられた反例に対する導出木のスケルトンが超多項式個存在することにより得られる計算量であるため，これ以上改善することが難しいと見込まれる．

一方，教示可能性の観点からは，スタックユニフォーム単純決定性文法について，導出木のスケルトンの種類を最小にするためには，高々，終端記号の種類の数だけサンプルがあればよいことが示せる．したがって，この結果より，スタックユニフォーム単純決定性言語は，多項式時間での質問による学習は難しいが，準多項式計算量で教示可能であると言える．これは形式言語に対する教示可能性に関する新たな事実である．

(2)自然言語処理への応用

この課題に対しては，本研究により得られたアルゴリズムを文書の段落分割アルゴリズムに応用し，テキストデータを話題に応じた段落に分割する手法を開発した．本手法は，以前に我々が提案した対話文の文書を話されている話題の切れ目を見つけ出す手法に，本研究による成果を統合したものであり，分割精度において従前の良く知られた手法よりも高性能であることを実験的に示した．

この手法は，従来研究では，単語自体を出力記号とした HMM により段落分割を行うことが主であったものを，文章をクラスタリングアルゴリズムによりひとつの記号に集約し，その記号列を学習することにより実現される点に新規性がある．この操作により文書中での単語出現分布が少々乱れても頑強な分割精度を出すことができる．したがって，本研究による成果は，対話やチャットなどの話題の脱線が多いテキスト，マイクロブログなど短い文章の連なりで構成されたテキストに対する話題分割に効果を発揮する．

(3)本研究による学習アルゴリズムの新たな適用領域の開拓

強化学習によるゲームの途中局面を評価する評価関数の獲得を行うアルゴリズムに対して，ランダムシミュレーションに基づく途中報酬を設定し，学習が効率的に行われることを実験的に示した．

従来の研究では，強化学習における報酬はゲームの決着，すなわち勝敗のみを用いることが多かったが，本手法により途中経過を適切に評価することができる．また実験的にそのことを示した．

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３件）
① 但馬康宏，分割位置を教師値としたテキストの段落分割，情報処理学会研究報告，査読無，2011-MPS-82，pp.1-3，2011.
② 但馬康宏，強化学習による評価関数の獲得における報酬設定について，情報処理学会研究報告，査読無，2010-GI-24,pp1-7，2010.
③ 但馬康宏，強化学習によるゲームの評価関数の獲得，電子情報通信学会技術研究報告，査読無，COMP2009-28，pp.21--26，2009.

〔学会発表〕（計２件）
①大多悠介，但馬康宏，多段 UCB1 アルゴリズムによるオセロの実装と評価，第 61 回 電気・情報関連学会中国支部連合大会，発表番号:24-14，pp.474，Oct. 2010.
②福永直起，但馬康宏，オセロのヒューリスティックな評価要素に対する重み付け，第 61 回 電気・情報関連学会中国支部連合大会，発表番号:24-15，pp.475，Oct. 2010.

６．研究組織
(1)研究代表者
但馬　康宏（TAJIMA Yasuhiro）
岡山県立大学・情報工学部・准教授
研究者番号：00334467