科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１０　日現在

機関番号：　１７５０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００２４５

研究課題名（和文）時間軸を考慮したニューロベース強化学習によるシンボル処理創発への突破口の模索

研究課題名（英文）Exploration of a Breakthrough Technology for Emergence of Symbol Processing by Neuro-based Reinforcement Learning Considering Time Axis

研究代表者
柴田　克成（Shibata, Katsunari）

大分大学・工学部・准教授

研究者番号：１０２６０５２２

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：膨大な時空間情報の中での効率的な学習は，実世界での高次機能創発の鍵を握る。本研究では，特に時間軸の扱いに注目した。事象の重要度を「主観的」に判断し，過去の事象に対する学習に利用する「因果トレース」の考え方を提唱し，状態価値の学習で飛躍的に学習性能が向上した。また，「概念」は行うべき行動の差に基づいて形成されるとの考えの下，強化学習によってリカレントネット内で離散的・抽象的な状態表現が形成された。さらに，自律的なコミュニケーション学習によって物体の動きの情報を伝達できることを示した。時間軸の処理に対して新たな展望をもたらしたが，シンボル処理創発にはダイナミクスの学習能力の更なる向上が必要である。

研究成果の概要（英文）：Efficient learning in a huge amount of spatio-temporal information holds the key to the emergence of higher functions in the real world. In this research, handling of time axis was especially focused on. A novel idea named "Causality traces" is propounded which judge the importance of events "subjectively" and are used for retrospective learning. Its learning performance exceeds that with the conventional method in value learning. Next, based on the idea that "concept" is formed from the difference of necessary motions, it is confirmed that discrete and abstract internal state representations are autonomously formed in a recurrent neural network through reinforcement learning. Furthermore, in autonomous communication learning, it was shown that information about target movement could be transmitted after learning. A novel perspective could be introduced to the handling of the time axis, but for the emergence of symbol processing, learning of dynamics should be further improved.

研究分野：　総合領域

科研費の分科・細目：　情報学，知覚情報処理・知能ロボティクス

キーワード：　知能ロボット　強化学習　ニューラルネット　高次機能　シンボル処理創発　因果トレース　概念形成　コミュニケーション学習

１．研究開始当初の背景

「高次機能」の脳における解明とロボットでの実現は，長年研究されて来たにもかかわらずなかなか進展して来なかった。特に,「ロボットがどうしたら論理的に物事を考えられるようになるか？」は非常に難しい問題であり，その糸口さえつかめない状態であった。また，その前提となるシンボル処理にはシンボルとパターンの乖離であるシンボルグラウンディング問題が立ちはだかって来た。シンボルとパターンをつなごうという研究はなされてきたが，いずれもシンボルとパターンを別々のものとして扱って，人間がいかに両者をつなぐかを設計して与えており，時空間情報であふれる実世界において，人間の手でシンボルとパターンを柔軟かつ有機的に融合することは困難なのではないかとの危惧があった。

一方，筆者らは，並列処理の学習が可能なニューラルネットを強化学習で自律的に学習させることで，ニューラルネット内部に，行動だけでなく，認識，記憶など目的達成に必要となるさまざまな機能が経験を通して創発することを確認してきた。そして，実世界におけるロボットの真の知能化のためには，柔軟な超並列処理に対する人間の持つプログラミング能力の限界から，人間が知識を与える形ではなく，ロボット自らが学習によって賢くなる方法を取ることが必要であると提唱して来た。

２．研究の目的

本研究では，シンボルグラウンディング問題を根本的に解決するために，パターンとシンボルを分けて考えず，シンボル処理も他の機能と同様に，センサからモータまでの並列で柔軟なシステム全体での学習の中で創発することを模索すべきだとの考えに基づく。そして，ニューラルネットと強化学習の組み合わせにより，少しでも「シンボル処理の創発」「論理的思考の実現」に近づき，その突破口を探し出すことを大きな目的とした。そして，従来，ロボットの処理を状態から行動の静的マッピングと捉え，時間を受動的に受け止めて，すべての時間をフラットに扱い，時間軸に注目して来なかったことに問題があると考え，本研究では特に時間軸の扱いに注目した。そして，(1)膨大な情報であふれる空間×時間（時空間）からの重要な情報の切り出しと(2)動詞コミュニケーションの獲得の２つを目標として掲げ，そこからシンボル処理，論理的思考への関わりを考えることを目指した。

(1)に関しては，時空間にあふれる莫大な情報の中から重要な事象を切り出して分節化した状態表現ができるからこそ，それに対応するシンボルが創発すると考えた。そこで，従来の強化学習における Eligibility Trace やリカレントネットの教師あり学習における BPTT 法がいずれも時間をフラットに扱って来たことに対し，ここでは，重要な状態では時間をゆっくり進めて十分に記憶し，そうでない場合は時間を早く進めることを考えた。その際，重要な状態の判断を人間が与えるのではなく，自分で判断させるため，ニューロンの出力の時間変化で個々のニューロンが個別に重要度を判断する。従来，リカレントネットの学習に対するアイデアとして用いて来たこの考え方を，過去の事象の学習に対して一般化するとともに，強化学習に導入し，実際に効率的な学習が実現できるかどうかを検証する。

また，シンボル処理と密接に結びつく「概念」の形成については，センサ信号空間上の距離による分類だけで説明するのは不可能であり，必要な行動の違いを生成するための離散的かつ抽象的な内部状態表現として形成されると考えた。そして，リカレントネットを用いた強化学習を行うことで，連続値センサ信号を入力として，行動生成の必要性から離散的，抽象的な内部表現が獲得できるかどうかを確認する。

さらに，論理的思考は，リカレントニューラルネット内の単なる固定点収束のダイナミクスで実現できるものではなく，複数の状態間遷移などの複雑なダイナミクスが求められる。そこで，強化学習における複雑なダイナミクスの利用，および，固定点収束以外のダイナミクスが強化学習を通して実現できるかどうかを検証する。

(2)に関しては，論理的思考が言語的であることから，コミュニケーションが自分から自分への内在的なものになったものが論理的思考であるという可能性に注目した。そして，他人とのコミュニケーションの発達がシンボル創発，論理的思考につながる可能性があると考えた。すでに，現在の視覚センサ信号から，必要な情報を抽出し，相手に伝えることを学習によって獲得できることは示していた。コミュニケーションにおいては，見ている現在の状況を必要に応じて相手に伝えるだけでなく，状態の時間的変化を伝えることも必要である。そこで，本研究では，それを動詞的表現と呼び，リカレントネットと強化学習で，物体の動きを相手に伝えるコミュニケーションの自律学習が可能であると示すことを目指した。

３．研究の方法

本研究では，ニューラルネットを用いた強化学習において，研究目的で述べた最初の観点「時空間からの重要な事象の切り出し」をさらに３つに分け，(1)過去の事象に対する効率的な学習，(2)離散的かつ抽象的内部表現の学習による獲得，(3)固定点収束以外のダイナミクスの学習による獲得，(4) 動詞コミュニケーションの学習による獲得 の４つのサブテーマで進めた。

(1) 過去の事象に対する効率的な学習

状態の時間変化が大きいときの情報を重点的に記憶して学習で用いるため，各ニューロンで過去の入力の値を保持するメモリの更新時定数(実際には定数ではなくなるが)を出力の時間微分の絶対値に反比例させる方法を一般化し，ここではこれを強化学習に適用する方法を定式化した。そして，簡単なタスクで，一定の時定数で過去の情報の取り込みと保持を行う Eligibility Trace の場合と学習速度などを比較し，その効果を観察した。

まず最初に，簡単な一次元の空間を移動するタスクの学習を行った。本手法の動作の確認と問題点の洗い出しを目的とし，状態評価と行動を同時に学習する強化学習のうち，行動は理想的なものを与え，状態評価の学習に本手法を適用した。続いて，本手法の効果が大きく発揮されると考えられる各状態の存在時間が状態ごとに大きく異なるタスクに適用して，Eligibility Trace の場合と比較して学習性能が大幅に向上するかどうかを観察し，さらに，各中間層ニューロン間で，異なる過去の事象を表現するように役割分担が進むかどうかを観察した。

その後，行動の学習も含めた強化学習に適用し，特に，試行錯誤(探索)のために加える乱数成分の影響を観察した。

(2) 離散的かつ抽象的内部表現の学習による獲得

時間とともに変化する連続値のセンサ信号の中で，行うべき行動生成のために離散的，抽象的な状態識別が役立つタスクとして，複数の部屋からなる環境をコンピュータ上に構築した。ロボットが各部屋の真ん中に配置されたスイッチを押すと四方のうちの一つの扉が開き，そちらの部屋に移動できるようにする。それを繰り返して，4 つ目の部屋に到達したら報酬が得られるタスクを設定する。そして，8 方向の壁までの距離とスイッチまでの距離と方角に関する全部で 11 個の連続値信号を入力とし，移動のための連続値信号 2 個とスイッチを押すか移動するかを選択するための 2 個の Q 値の信号を出力とする。そして，Actor-Q 学習という強化学習の方法でリカレントネットを学習させる。そして，学習後，中間層ニューロンを観察し，状態変化を離散的に表現し，かつ，センサ信号によらない抽象的な状態表現をしたニューロンが創発するかどうかを観察する。また，距離センサ等の 11 個の信号からの状態表現では，「抽象的」と呼ぶのに不十分であると考え，冗長な入力信号を付加して，より多くのセンサ信号から「離散的」かつ「抽象的」な状態表現が獲得できるかを観察する。

当初の計画では，その後，実ロボットを用いて実験する予定であったが，利便性から，ロボットシミュレータ Webots 上で部屋とロボットよりなる環境を構築し，ロボットへの入力信号を画像データとして学習を行った。

(3) 固定点収束以外のダイナミクスの学習による獲得

Tani らが提案し，階層的かつ相互に引き込みを持つダイナミクスの生成が確認されている MTRNN (Multiple Time Scales Recurrent Neural Network)において，ダイナミクスの階層性が強化学習の効率的探索実現につながる可能性を検討した。さらに，強化学習に基づいてリカレントニューラルネットを学習させることで，固定点収束以外のダイナミクスを形成できるかを調べるため，ホイールを一定周期で回転させるという周期運動が必要なタスクを学習させ，それにあった振動子が形成されるかどうかを確認した。さらに，これも Tani らが提案している RNNPB (Recurrent Neural Network with Parametric Bias)と組み合わせて，バイアス入力による周期の制御を試みた。

(4) 動詞コミュニケーションの学習による獲得

送信者と受信者の 2 つのエージェントを用意し，送信者は簡単な視覚センサを有し，動く物体を捉え，その視覚センサ信号をリカレントネットへの入力とし，受信者に送る信号を決定するとともに，視覚センサの動きも自ら決定する。受信者は送られて来た信号を元に，物体の動きがどのようなものかを出力する。そして，受信者が認識結果を出力した際に正解か不正解かによって与えられる報酬や罰を元に両者が個別に強化学習を行った。一つの画像だけからでは物体の動きの情報を伝えることができないため，リカレントネットを用いて，送信者は適切に視覚センサを動かしながら，複数ステップかけて物体の動作の情報を伝えられるようになるかどうかを確かめる。

最初は，簡単のため，マス目状に区切られた 5x5 の離散環境において，送信者は環境より小さい 3x3 のセルよりなる視覚センサを有し，受信者は物体の場所によらず，上下左右のどの運動を行ったかを識別する。視覚センサの動作も上下左右の離散動作に限定し，送信者は，4 つの信号のどれかを受信者に送るか，センサを 4 方向のどこかに動かすかを 8 個の Q 値出力をもとに選択した。

次に，連続空間でも学習できるかどうかを確認するため，10x10 の 100 個のセンサセルを持つ視覚センサを用いて，上，下，右に回転運動を加えることで，物体の動作を識別するために視覚センサで 3 回観察する必要がある場合について学習させた。この場合，視覚センサを動かすための 2 つの連続値およびどの信号を相手に伝えるかを決めるための信号 4 個の計 6 個を出力とした。一方，受信者は，送信者からの信号をリカレントネットに入力し，4 つの動作か認識を先送りするかのどれかを決めるための 5 個の Q 値出力を持つ。そして，連続値動作出力を有する送信者は Actor-Q 学習，Q 値出力のみの受信者は Q 学習によって学習した。

４．研究成果

(1) 過去の事象に対する効率的な学習

まず，強化学習において，従来の時間をフラットに扱って過去の入力を記憶する Eligibility Trace を参考にし，時間をさかのぼることなく，割引率を考慮した過去の状態価値を学習できるように，因果トレース(Causality Trace)の定式化を行った。また，簡単な環境でのタスクのシミュレーションを通して，学習初期の学習速度が遅いことおよびトレースの初期値の与え方をどうするべきかという 2 つの問題点が明らかになったのに対し，過去の出力変化と現在の出力変化の相対的な関係に基づいて入力信号を取り込む割合を変化させる方法，および，試行開始時に，入力の総和が 0 の場合の出力からの変化量に基づいて入力信号を取り込む方法を導入した。また，Eligibility Trace では，入力を取り込む割合が常に一定であるだけでなく，その値を予め適切な値に決める必要があり，学習するタスクに関する知識が必要になるが，因果トレースはその必要がないにもかかわらず適切に働くことが確認できた。

さらに，短い滞在時間の状態と長い滞在時間の状態が混在するタスクでは，図 1 のように，環境全体にわたるグローバルな学習(a)と隣接する状態間のローカルな学習(b)が両立でき，Eligibility Trace に対して圧倒的な学習性能を示すことを確認した。また，因果トレースは，個々のニューロンの出力の変化によって入力を取り込むため，ニューロンごとに取り込む入力が違ってくることから，個々の中間層ニューロンが表現する過去の事象がニューロンごとに異なる傾向が Eligibility Trace の場合より大きいことも示した。

また，動作の学習も含めた強化学習へ因果トレースを適用したが，状態価値のみの学習時のような圧倒的な優位性を示すには至っていない。潜在能力が非常に大きいだけに，乱数を用いない試行錯誤の導入という根本的な方法も含めて今後の大きな課題である。

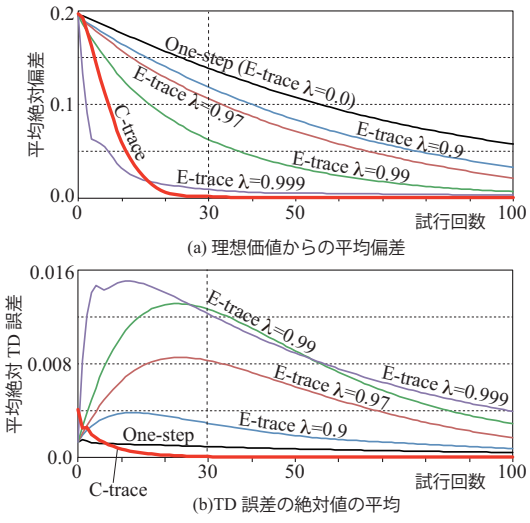

図 1　状態価値学習時の学習曲線

(2) 離散的かつ抽象的内部表現の学習による獲得

研究方法で述べた複数の部屋の移動タスクにおいて，最初は２つ目の部屋に移動できたら報酬をもらえるようにし，それができたら，３つ目の部屋，４つ目の部屋と徐々にタスクを難しくすることで，扉の開き方がどのような場合でも，壁までの距離情報などの 11 個の時系列の連続値センサ入力から，必要なときにスイッチを押し，扉が開いたらそちらに移動し，４つ目の部屋まで到達することができるようになった。学習後に中間層ニューロンを観察したところ，図２のように，開くドアの方向によらず,ドアが開いて新しい部屋が出現する前と後を離散的に区別したり，違う部屋に移動すると離散的に出力が変化したりするニューロンが発現していることを確認した。ここでは，センサ信号が異なっていても同様な状態表現をすることを「抽象的」と呼んだが，入力として毎ステップランダムに変化する 39 個の入力を与えることで，より多くの入力信号の中から必要な情報を抽出して状態表現ができるかどうか学習させたところ，同様に離散的な情報表現ができることを確認した。また，４部屋で学習させたロボットをそのままさらに部屋数を増やした環境に置いたところ，追加学習をすることなく 20 個以上の部屋を通過してゴールにたどり着くことができ，状態表現の汎化能力を示していると言える。

また，ロボットシミュレータ Webots 上で部屋とロボットよりなる環境を構築し，ロボットが画像データを取り込んで学習を行ったが，まだ最初の部屋でボタンを押すことまでしか学習ができなかった。

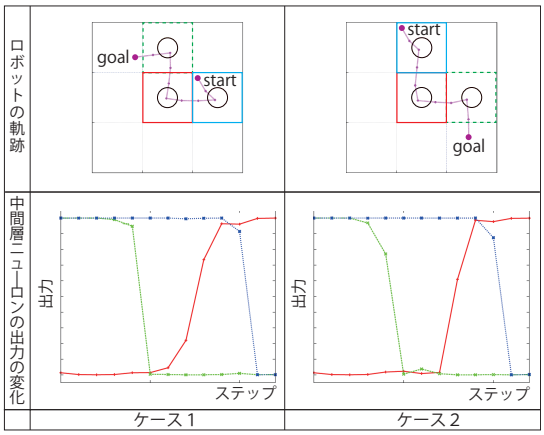

図 2 部屋移動タスクでの学習後のロボットの動作軌跡と中間層ニューロン出力の離散的・抽象的変化

(3) 固定点収束以外のダイナミクスの学習による獲得

Tani らが提案した MTRNN の階層的ダイナミクスを強化学習で利用することを考え，特に時定数が大きい抽象空間上で外乱を与えることによる高次のダイナミックな探索の可能性を探った。そこでまず，MTRNN 内

にアトラクタ空間がどのように形成されるのかを調べた。しかし，形成されたアトラクタの大きさに大きな不均一があるなど，探索ダイナミクスの階層性を思うように発揮さることはできなかった。

次に，固定点収束以外のダイナミクスが，必要に応じて強化学習で獲得できるかという点に着目し，タスクに必要な周期に近い振動子が創発するかを通常のリカレントネットを用いて学習させたところ，学習成功率が低かったり，獲得できる周期の範囲が狭かったり，学習が途中で壊れると振動子の波形が不規則なものになったものの，振動子の形成が確認できた。また，ダイナミクスを別の信号で変化させるために RNNPB を使って学習したところ，不安定ではあるものの，PB 値によってホイールの回転周期を変化させられることを確認した。

(4) 動詞表現のコミュニケーションの学習による獲得

コミュニケーションにおける動詞表現の獲得については，研究方法で述べた２つのタスクについて，いずれも，最初からコミュニケーションの学習をさせてもできなかった。これに対し，送信者がまず単独で，センサの動かし方と物体の動きの認識を事前に学習した後に，受信者を含めて学習することで，コミュニケーションの学習ができるようになった。さらに，連続空間で認識のために３ステップ必要なコミュニケーションタスクの学習では，事前学習後のコミュニケーション学習時に，事前学習で獲得されたセンサ動作が壊れてしまうことが観察され，センサ動作と信号生成のネットワークを分ける，もしくは，中間層以下の重み値を固定することでそれを回避することができた。新たに難しいタスクの学習をさせる場合には，ある程度，前に学習したことが壊れないような工夫が必要であることを示唆していると考えられる。

以上の４つの項目における成果をまとめると，特に，「因果トレース(Causality Trace)」について，新たな時間軸の扱い方を提示し，その優れた潜在能力を明確に示すことができたことは，非常に大きな意味があった。本手法は，強化学習に適用する場合と使い方が少し異なるものの，今後その必要性がますます増していくと予想されるリカレントネットの学習においても，その実用的な計算量やメモリの量などから本手法が BPTT や RTRL などの従来の主要な方法に取って代わる可能性がかなり高いのではないかと考えている。近年注目されている Deep Learning が，筆者が以前から提唱して来た，人間による作り込みをできるだけ排除し，生のセンサ信号を入力し，並列で柔軟なニューラルネットで学習させることの重要性を後押ししており，空間的に膨大な情報の扱い方を示していると考えることができる。その一方で，本研究で示した時間の扱い方と合わせることで，膨大な時空間情報を効率的に扱うことができるようになり，実世界での高次機能への道を開くのではないかと期待される。そのためには，行動の学習も含めた強化学習でその有効性を示すこと，さらには，リカレントネットを用いて強化学習を行う場合では，因果トレースは両者に威力を発揮すると期待されるので，それを引き出すためにどのように適用すれば良いかを確立することが急務である。

その他の項目についても，連続値入力から離散的抽象的な状態表現を学習によって獲得できることを示したことは，「概念」形成への大きな一歩と考えられるし，リカレントネットを使って強化学習をさせることで，振動子が形成できたことも，固定点収束以外のダイナミクスの創発の第一歩ということで意味が大きい。さらに，動的な情報もコミュニケーションで表現できるということも，新たな可能性を示した大きな一歩であると考えている。

このように，個々の項目としては一定の成果を上げることができ，本研究の大きな目的である「シンボル処理の創発」についても前進があったと考えられる。しかし，何分，簡単な振動子を学習で獲得されることも不安定でなかなか難しく，部屋移動タスクの学習も，コミュニケーションの学習も，タスクがまだまだトイプロブレムの領域を出ないにも関わらず，学習に非常に時間がかかったり，易しい問題から徐々に難しくするというスケジューリングが必要であったりと学習させるのに大変な苦労をした。そう考えると，人間の高次機能とのギャップはあまりにも大きいという点については，残念ながら本研究を通じて何か明確な突破口が見えたと言うことは難しい。Deep Learning に見られるように，量が質を変える可能性も期待できるが，少なくとも，リカレントネットの学習によって，より複雑なダイナミクスをより容易に安定して獲得できるようになることは必須であると考えられる。そのためには，本研究で当初予定していたが，ほとんど時間が割けなかったリカレントネットの構造面での工夫や，単なる乱数ではない，たとえばカオスなどを利用し，時間軸を考慮した新たな探索方法の開発などが必要ではないかと考えている。

また，本研究において，小型のロボットを用いた学習を当初予定していたが，途中で出て来た様々な問題の解決に時間がかかったことなどから，予定通り進めることができなかった。研究期間終了後も，できる範囲で引き続き研究を進めていきたい。

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８件）

① Katsunari Shibata and Kenta Goto, Emergence of Flexible Prediction-Based Discrete Decision Making and Continuous Motion Generation through Actor-Q-Learning, Proc. of Int'l Conf. on Development and Learning and on Epigenetic Robotics (ICDL-Epirob) 2013, ID 15 (CDROM) 2013
② Yoshito Sawatsubashi, Mohamad Faizal bin Samsudin and Katsunari Shibata, Emergence of Discrete and Abstract State Representation in Continuous Input Task through Reinforcement Learning, Advances in Intelligent Systems and Computing, Robot Intelligence Technology and Applications 2012, Proc. of RiTA 2012, M1C-2.pdf, pp. 13-22 2012
③ Katsunari Shibata and Shuji Enoki, Differential Trace in Learning of Value Function with a Neural Network, Advances in Intelligent Systems and Computing, Robot Intelligence Technology and Applications 2012, Proc. of RiTA 2012, M2C-4.pdf, pp. 55-64 2012
④ Mohamad Faizal bin Samsudin, Yoshito Sawatsubashi and Katsunari Shibata, Emergence of Multi-Step Discrete State Transition through Reinforcement Learning with a Recurrent Neural Network, LNCS(Lecture Notes in Computer Science), Neural Information Processing, Proc. of ICONIP (Int'l Conf. on Neural Information Processing) 2012, Vol. 7664, pp. 583-590, 2012
⑤ Katsunari Shibata and Shunsuke Kurizaki, Emergence of Color Constancy Illusion through Reinforcement Learning with a Neural Network, Proc. of ICDL-EpiRob (Int'l Conf. on Development and Learning - Epigenetic Robotics) 2012, PID2562951.pdf, 2012 DOI: 10.1109/DevLrn.2012.6400580
⑥ Katsunari Shibata and Kazuki Sasahara, Emergence of Purposive and Grounded Communication through Reinforcement Learning, LNCS(Lecture Notes in Computer Science), Vol. 7064, Proc. of ICONIP (Int'l Conf. on Neural Information Processing). 2011, pp.66-75 ,2011

〔学会発表〕（計１３件）

① 柴田克成, 因果トレース － 並列かつ主観的時間スケールの導入による過去の処理の効率的学習 －, 電子情報通信学会ニューロコンピューティング研究会, 2014.3.18, 東京
② 品矢 裕介, 強化学習によるリカレントニューラルネットワーク内部での振動子創発の可能性, 第 32 回計測自動制御学会九州支部学術講演会, 2013.12.1, 長崎
③ 朱 祺, RNN を用いた強化学習によるセンサ信号の時間変化を表すコミュニケーションの創発, 第 32 回計測自動制御学会九州支部学術講演会, 2013.12.1, 長崎
④ 榎修志, , ニューラルネットを用いた価値関数の学習における 微分型トレースの提案, 計測自動制御学会システム・情報部門学術講演会, 2012, 2012.11.23, 名古屋
⑤ 沢津橋由人, リカレントネットを用いた強化学習における 離散的かつ抽象的な状態表現の創発, 計測自動制御学会 システム・情報部門学術講演会 2012, 2012.11.23, 名古屋
⑥ 柴田克成, あめとむちで知能を作る？ －知能ロボットって本当に賢いの？－, SOFT 九州支部夏季ワークショップ (招待講演), 2011.9.1, 玉名(熊本県)

６．研究組織
(1)研究代表者
柴田 克成（Katsunari Shibata）
大分大学・工学部・准教授
研究者番号：10260522