自然な対話の中で物体の名前を覚えるロボット

A Robot That Can Acquire Object Names in Natural Spoken Dialogues

中野 幹生[*1] Mikio Nakano　能勢 隆[*2*3] Takashi Nose　田口 亮[*2*4] Ryo Taguchi　水谷 了[*5] Akira Mizutani　中村 友昭[*5] Tomoaki Nakamura　船越 孝太郎[*1] Kotaro Funakoshi
長谷川 雄二[*1] Yuji Hasegawa　鳥井 豊隆[*1] Toyotaka Torii　岩橋 直人[*2*6] Naoto Iwahashi　長井 隆行[*5] Takayuki Nagai

[*1](株) ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

[*2](株) 国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International

[*3]東京工業大学
Toko Institute of Technology

[*4]名古屋工業大学
Nagoya Institute of Technology

[*5]電気通信大学
University of Electro-Communications

[*6](独) 情報通信研究機構
National Institute of Information and Communications Technology

This paper presents architecture for a robot that can acquire the names of objects in natural, multi-domain spoken dialogues. It is based on a multi-expert model of the dialogue and behavior controller of robots. It employs an expert for word acquisition dialogue, an expert for understanding commands using the acquired words, and an expert for replying to questions on the name of objects. These experts communicate with a word acquisition module and an image learning and recognition module. In addition, adding other experts for various task domains enables building a robot that can perform word acquisition in multi-domain dialogues. We have built a robot based on the proposed architecture, and it showed the effectiveness of the architecture.

1. はじめに

音声対話機能を持つロボットやアニメーションエージェントなどの研究が盛んに行われている．音声対話を行うには音声認識が必要であるが，現在の音声認識技術は，あらかじめ登録した単語やフレーズしか認識ができない．これは，実際にロボットが家庭やオフィスなどで使われるようになる際に問題となる．家庭やオフィスなどでは，その家庭・オフィス特有の言葉が用いられ，しかも，日々新しい言葉が誕生する．たとえば，ある家庭で新しい物品を購入すると，それにその家庭特有の名前をつけることがよく行われる．そのような物品を家庭内で探すタスクを言葉で指示しようとするとき，ロボットがその言葉を知らないと指示が理解できない．キーボード入力により新しい言葉を登録することが考えられるが，煩雑である．したがって，ロボットは音声コミュニケーションによって新しい名前を覚える能力を持つことが望ましい．

今までに物体の名前を覚えるシステムの研究が数多く行われてきたが，そのほとんどは，名前を覚えるというタスクに特化しており，複数の発話のセットから統計的に獲得するものである [Roy 00, Yu 94]．しかしながら，実際の家庭用ロボットの使用を考えると，自然な対話の中で，名前を教示する発話を検出し，その発話の中にある物体の名前を抽出して物体と結びつけて覚える必要がある．

本稿では，対話の中で新しい名前を覚えるロボットのアーキテクチャを提案する．提案するアーキテクチャは，マルチドメイン対話行動モデル RIME (Robot Intelligence with Multiple Experts)[Nakano 08b] に基づいている．RIME は特定のドメインのインタラクションを行うエキスパートというモジュールを複数組み合わせることによって複雑なインタラクションを行うことができる．RIME のエキスパートの一つとして語彙獲得対話を行うエキスパートを用い，獲得した名前を他のエキスパートでも利用できるようにすることにより，マルチドメイン対話の中で語彙獲得を行うことができるようになる．

提案したアーキテクチャの有効性の実証のために対話ロボットを構築した．このロボットは対話により物体の名前を覚えるとともに，その名前で物体の探索を指示されると，移動して物体を探しに行くことができる．

連絡先: 〒 351-0188 埼玉県和光市本町 8-1 (株) ホンダ・リサーチ・インスティチュート・ジャパン，中野 幹生，E-mail: nakano@jp.honda-ri.com

2. 課題

本稿で扱う課題は，様々なドメインの対話を行うロボットが人間との対話の中で物体の名前を覚えることである．つまり，いわゆるマルチドメイン対話の一つのドメインとして，語彙獲得対話ドメインがあるとする．語彙獲得対話では，人間は物体を見せながら，自然な発話でその名前をロボットに教える．ここで自然な発話とは，定型的な発話ではなく,「これは … だよ」や「… を覚えて」など，様々な表現を用いた発話のことである．

語彙獲得対話以外のドメインとして，獲得した語彙を用いた人間の指示発話を理解して実行するドメインと，獲得した語彙を尋ねる発話に答えるドメインを扱う．獲得した語彙を用いた指示とは，例えば,「… はどこにある？」などの物体の探索の指示である．獲得した語彙の質問は「これは何ですか？」などである．これらのタスクを遂行するためには，語彙獲得の際

に，語彙の音韻列を正しく獲得している必要がある．

さらに，ロボットはマルチドメイン対話機能を持つので，他の全く異なるドメインの対話も行える必要がある．例えば，天気情報に関する質問応答などの一般的に扱われている対話ドメインを扱うことができなければならない．

類似の課題を扱った研究として，Holzapfel ら [Holzapfel 08] の研究がある．ロボットは対話の中で未知語 (out-of-vocabulary word) を発見すると語彙を学習する対話を行う．Holzapfel ら研究と本研究の違いは，Holzapfel らが定型的なパターンの中で未知語が現れる場合のみを扱っているのに対し，我々は自然な発話で物体の名前を教える発話を扱っていることである．さらに，我々は対話ドメインの追加が容易であることを目指しており，マルチドメイン対話アーキテクチャをベースにしている．

3. 語彙獲得対話ロボットアーキテクチャ

2. 節で述べたようなロボットは，発話を検出すると，その発話が，どのドメインの発話なのかを判定しなくてはならない．これは，マルチドメイン対話システムにおけるドメイン選択の問題として捉えることができる．我々も一般的なマルチドメイン対話システムのアーキテクチャをベースにするが，その中でも，RIME[Nakano 08b] を用いる．RIME は，後述するように，その他のマルチドメイン対話システムアーキテクチャに比べて，対話中のドメインの変更が柔軟であること，および，ドメインの追加に関して制限が少ないという利点を持っている．本節では，まず RIME の概略を述べてから，提案アーキテクチャを説明する．

3.1 マルチエキスパートモデル RIME

RIME では，特定の種類のサブタスクに特化した知識と内部状態を持つエキスパートと呼ぶモジュールを用いる（これは，マルチドメイン音声対話システムで用いられているドメインエキスパート [O'Neill 01] の概念を拡張したものである [Nakano 08b]）．たとえば，天気予報に関する質問に答えられるロボットであれば，「天気予報に関する質問を理解する」というサブタスクのためのエキスパートや「天気予報を人に伝える」というサブタスクのためのエキスパートを持つ．また，「特定の場所に移動する」という物理行動を行うサブタスクのためのエキスパートなども用いることができる．これらのエキスパートを順次利用することにより，複雑なタスクを遂行することができる．たとえば，ある物を説明するタスクは，その物のところに人を案内して，言葉で説明するという 2 つのサブタスクを順次遂行することによって行うことができる．

RIME では，このようなエキスパートを利用して全体のシステムを動作させるためのプロセス群（調整プロセス群と呼ぶ）が走っている．RIME のモジュール構成を図 1 に示す．調整プロセスは 3 つあり，並行動作する．理解プロセスは音声認識結果をエキスパートに送信し，最適なエキスパートを選択し，タスク計画プロセスにその情報を送る．行動選択プロセスは，選択されたエキスパートに対し，次の動作の決定を要求する．タスク計画プロセスは，タスクを遂行したり，音声認識結果に反応したりするために，どのエキスパートをアクティベートし，どのエキスパートをディアクティベートするかを決定する．これら 3 つのプロセスは発話割り込みを扱うために並列で動作する．

それぞれのエキスパートは内部状態にアクセスするためのメソッドを持っていなければならない．initialize メソッドはエキスパートが作られたときに呼ばれ，内部状態を初期化する．understand メソッドは音声認識結果を受け取った際に理解プロセスから呼び出され，音声認識結果に基づいて情報を更新する．select-action メソッドは，行動選択プロセスから継続的に呼び出され，発話待ちの状態でなければ，内部状態に基づき，行動を 1 つ出力する．その他に割り込み発話を扱うためのメソッドなどを持っていなくてはならない．

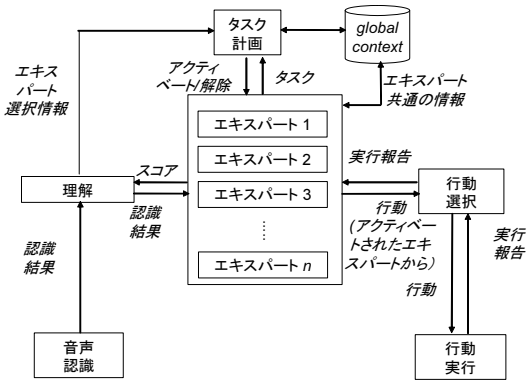

図 1: RIME のモジュール構成

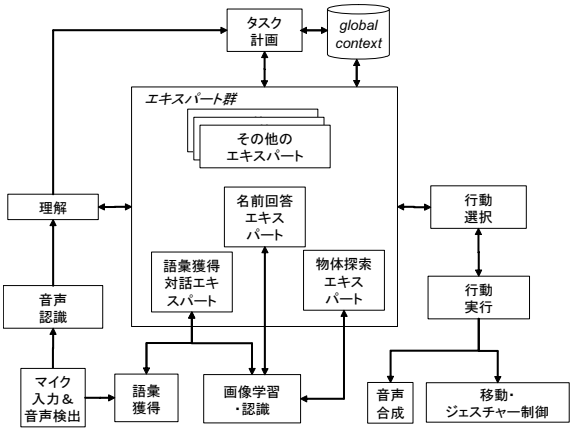

図 2: 語彙獲得対話ロボットアーキテクチャ

understand メソッドの返り値は，その音声認識結果がどのくらいそのエキスパートで処理されるべきかを表す 0 以上 1 以下のスコアである．理解プロセスは，音声認識結果を，現在アクティベートされているエキスパートおよび，新規にアクティベートされる可能性のあるエキスパートに，この understand メソッドを用いて送り，最も高いスコアを返したエキスパートを選択して，その情報をタスク計画部に送る．これは，マルチドメイン音声対話システムにおけるドメイン選択の機能にあたる．

これらのインタフェースを実装しさえすれば，内部で知識や状態をどのような形で保持しているか，また，どのようなアルゴリズムで理解や行動選択を行うかに関わらず，どのようなエキスパートでも導入することができる．

各エキスパートは，global context と呼ばれるデータ格納部を介して，共通に使える情報（例えば，話題になった事物，ユーザの興味，周囲の状況など）を授受できる．

3.2 アーキテクチャの概要

語彙獲得対話ロボットのアーキテクチャを図 2 に示す．各タスクドメインごとにエキスパートを用意する．語彙獲得エキス

パートは，語彙獲得モジュールおよび画像学習・認識モジュールと通信することにより，語彙獲得対話を行う．獲得した語彙の情報は global context に蓄えられ，他のエキスパートも用いることができる．語彙獲得対話エキスパートは獲得した語彙の情報を global context 経由で他のエキスパートに送る．

物体認識が必要なエキスパート，すなわち，物体探索エキスパートや物体の名前を返答するエキスパートは，必要に応じて画像学習・認識モジュールと通信する．

語彙獲得モジュールが独自に音声認識を行う必要があるため，音声区間検出後の音声を分離して RIME の音声認識と語彙獲得モジュールの両方に送る．

3.3 エキスパート選択

RIME ではエキスパート選択は，各エキスパートの understand メソッドが返すスコアを用いて行われる．本語彙獲得ロボットでは，音声認識は，各エキスパートが持っている文パターンに基づく有限状態文法と，大語彙統計言語モデルの 2 つを併用する．大語彙統計言語モデルを用いた認識結果は，BWG(Bag-of-Words in Graph) 法 [Funakoshi 07] のような発話分類手法で用いる．この発話分類の結果と，有限状態文法を用いた認識結果が言語理解文法でカバーされているかどうか，有限状態文法を用いた認識結果の信頼度，および文脈などの情報などを利用してスコアを算出する．

3.4 語彙獲得対話エキスパート

物体の名前を教示する発話が入力されて，このエキスパートがアクティベートされると，このエキスパートは画像学習認識モジュールに画像学習を要求する．画像学習認識モジュールは，見えている物体が過去に覚えたものと同じかどうかを判断し，同じならその物体の ID を，そうでなければ，物体の画像の特徴を記憶するとともに物体の ID を送り返す．学習に失敗した場合は，失敗フラグを送る．語彙獲得対話エキスパートは，学習が失敗であれば，その旨を音声でユーザに伝える．また，物体の ID が得られた場合は，語彙獲得モジュールに語彙獲得を要求する．語彙獲得モジュールは，あらかじめ学習してある教示発話の言語モデルを用いて名前を獲得して送り返す．語彙獲得対話エキスパートは，獲得した語彙と物体 ID との関係を global context に書き込むとともに，音声認識の有限状態文法に獲得した語彙を追加する．

3.5 物体探索エキスパート

物体探索エキスパートは，物体探索要求が認識されると，認識結果から物体 ID を得て，画像学習・認識モジュールに物体探索要求を送るとともにあらかじめ指定したルートでロボットを移動させる．物体探索は，高速だが精度の低い方法で物体を探索する．物体が存在する可能性があれば，ロボットを停止させる．そして画像学習・認識モジュールに物体認識要求を送るとともに，物体の存在する可能性の高い方向にロボットを移動させる．物体が認識されれば，探索は終了する．この探索の過程でロボットは，獲得した物体の名前を用いて「… を探しに行きます」「… を発見しました」等の発話を行う．

3.6 物体の名前を返答するエキスパート

物体の名前を尋ねる発話が認識されると，画像学習・認識モジュールに画像学習要求を送る．返ってきた結果がすでに名前を覚えた物体の ID であれば，その物体の名前を答える．そうでなければ「わかりません」と答える．

図 3: 語彙獲得対話ロボットの外観

4. 実装

上記アーキテクチャを用いて語彙獲得対話ロボットを構築した．語彙獲得，物体探索，名前の回答，その他の対話が行えることを確認している．以下に実装に用いたハードウエアおよびソフトウエアモジュールの詳細を述べる．

4.1 ロボット

ロボットは図 3 に示すような全方向移動台車ロボットである．ロボットには指向性マイクロホン (三研マイクロホン (株) CS-3e) とステレオカメラ (東京エレクトロンデバイス (株) TD-BD-SCAMv2) が取り付けられており，音声対話処理と画像処理はロボットに搭載された 2 台の PC で行っている．ロボットの移動は別のサーバコンピュータで制御されていて，ロボットおよびロボット上の PC とは無線 LAN で通信する．モジュール間通信は MMI-IF[鳥井 06] を用いることにより容易に実現している．

4.2 対話行動制御

対話行動制御は RIME をベースにしたツールキット RIME-TK[中野 08a] を用いて構築した．音声認識には，複数の言語モデルを用いてデコードできる Julius Ver.4[*1] を用いている．また，音響モデルおよび大語彙言語モデルは Julius 付属のものを用いている [Kawahara 04]．音声合成は NTT アイティ (株) の FineVoice を用いた．

現在は，実装の都合上，語彙獲得対話エキスパートと物体探索エキスパートは一つのエキスパート (語彙獲得対話・物体探索エキスパート) になっている．その他，物体の名前を返答するエキスパート，天気情報の要求を理解するエキスパート，天気情報を提供するエキスパート，内線番号の質問を理解するエキスパート，内線番号を教えるエキスパートなどを用いている．

発話が入力されたときのエキスパートの選択は，大語彙統計言語モデルの認識結果を用いた BWG 法 [Funakoshi 07] による発話分類の結果と，有限状態文法を用いた認識結果を Finite-State Transducer(FST) で言語理解した結果を用いて行っている．BWG 法による発話分類は，名前の教示か，探索要求か，その他の発話かに分類する．語彙獲得対話・物体探索エキスパートの understand メソッドは，発話分類の結果が名前の教示か探索要求の場合に一定のスコアを返す．

物体の名前を返答するエキスパート，天気情報の要求を理解するエキスパート，内線番号の質問を理解するエキスパートは，理解できる発話のパターンを FST の形で保持しており，それらの FST と同等の有限状態文法が音声認識用言語モデルとし

∗1 http://julius.sourceforge.jp/

て用いられている．その有限状態文法を用いた音声認識結果が自分の FST で理解できるかどうか，音声認識結果の信頼度，自分がすでにアクティベートされているかどうか，の 3 つの情報から，手書きのルールに基づきスコアを計算する．

天気情報を提供するエキスパートや内線番号を教えるエキスパートは，要求を理解するエキスパートによってタスクが設定された場合にのみアクティベートされる．

現在のエキスパート選択スコアの計算のための規則は開発者の試行錯誤に基づいたものである．データに基づいたスコアの最適化が今後の課題である．

4.3 語彙獲得

語彙獲得の方法には様々なものが考えられる．例えば，Holzapfel ら [Holzapfel 08] が行っているように文のパターンをあらかじめ与える方法や，山本ら [山本 04] のように，統計的言語モデルの中に音韻 ngram などで表された未知語の統計モデルを埋め込む方法がある．文のパターンをあらかじめ与えると，それ以外のパターンの文が認識できない．また，我々は，どのような語彙でも獲得できるように，未知語の統計モデルは用いない．

そこで，新規な語彙獲得の方法として以下の方法を用いている．名前を教える発話のパターンが，個人ごとにある程度限られていると仮定し，あらかじめ田口らの方法 [田口 09] を用いて，語彙を教示する発話の集合から言語知識を学習しておく．ここで，発話のうち物体の名前以外の部分のことを言い回しと呼ぶ．学習した言語知識には，言い回しのリスト，および言い回しと名前の bigram が含まれている．ここで名前の部分はクラス化し，クラス bigram とする．

発話が入力され，語彙獲得が要求されると，まず発話を音素認識（音素タイプライタ）により音素列に変換する．これには ATR で開発された音声認識システム ATRASR[伊藤 04] を用いている．次に，二段 DP マッチング [Sakoe 79] を用いて言い回しを音素列に当てはめる．このとき，bigram 確率を用いて，あり得ない単語列の当てはめが起こらないようにする．最後に当てはめた言い回しと当てはめられた部分の音素列との編集距離が閾値以下の場合，そこは言い回しではなく，名前であるとみなす．以上の方法で語彙獲得を行う．

4.4 画像学習・認識

画像学習・認識モジュールはステレオカメラの情報を用いて物体の画像の学習および物体の探索を行う．

物体を見せて学習させる際にまず問題となるのは，画像中のどの領域が学習すべき物体かという，物体の切り出しの問題である．この問題は，動きアテンションを用いることで解決する．これは，人が物体を持ちロボットに見せることで教示するため，その際に物体を動かすと仮定し，画像中の動いている塊が物体であるという事前知識を与える．つまり，画像中の動きを検出し，その領域の色や奥行きの情報を基に最終的な物体領域を確率的に推定するもので，ステレオの計算を含めても 10fps 程度で動作する．

物体探索では，シーン中のどこに認識すべき物体があるかを抽出する．但しこの際は必ずしも人が物体を持っている保証がないため，動きに注意を向けた抽出手法を用いることはできない．そこで，探索時の領域抽出には，色ヒストグラムと奥行き情報を併用した高速なアクティブ探索による領域抽出手法を利用する．

探索に成功した後，ロボットが物体に近づいてから最終的に認識を行う際には，SIFT(Scale Invariant Feature Transform) を用いた局所特徴のマッチングを利用する．この際，色情報を用いて候補を絞った上で，学習時に様々な方向から見て取得した物体の SIFT 情報とのマッチングを行い最終的な認識結果を得る．

5. おわりに

本稿では，マルチドメイン対話の中で物体の名前を獲得し，その名前を利用してコミュニケーションを行うことができるロボットのアーキテクチャを提案した．また，本アーキテクチャに基づいて実装したロボットについても述べた．

今後は，本ロボットの詳細な評価を行っていくとともに，語彙獲得・画像学習の性能向上を目指す．

参考文献

[Funakoshi 07] Funakoshi, K., Nakano, M., Torii, T., Hasegawa, Y., Tsujino, H., Kimura, N., and Iwahashi, N.: Robust acquisition and recognition of spoken location names by domestic robots, in *Proc. IROS-2007*, pp. 1435–1440 (2007)

[Holzapfel 08] Holzapfel, H., Neubig, D., and Waibel, A.: A dialogue approach to learning object descriptions and semantic categories, *Robotics and Autonomous Systems*, Vol. 56, No. 11, pp. 1004–1013 (2008)

[伊藤 04] 伊藤 玄, 葦苅 豊, 實廣 貴敏, 中村 哲：音声認識統合環境 ATRASR の概要と評価報告, 日本音響学会 2004 年秋季研究発表会講演論文集 Vol.1, pp. 221–222 (2004)

[Kawahara 04] Kawahara, T., Lee, A., Takeda, K., Itou, K., and Shikano, K.: Recent progress of open-source LVCSR engine Julius and Japanese model repository, in *Proc. Interspeech-2004 (ICSLP)*, pp. 3069–3072 (2004)

[中野 08a] 中野 幹生, 船越 孝太郎, 長谷川 雄二, 辻野 広司：オブジェクト指向に基づくロボット・エージェントのマルチドメイン対話行動制御モジュール構築ツール RIME-TK, 人工知能学会研究会資料 SIG-SLUD-54 (2008)

[Nakano 08b] Nakano, M., Funakoshi, K., Hasegawa, Y., and Tsujino, H.: A Framework for Building Conversational Agents Based on a Multi-Expert Model, in *Proc. 9th SIGdial Workshop*, pp. 88–91 (2008)

[O'Neill 01] O'Neill, I. M. and McTear, M. F.: Object-oriented modelling of spoken language dialogue systems, *Natural Language Engineering*, Vol. 6, No. 3&4, pp. 341–362 (2001)

[Roy 00] Roy, D.: Integration of Speech and Vision Using Mutual Information, in *Proc. ICASSP-2000*, pp. 2369–2372 (2000)

[Sakoe 79] Sakoe, H.: Two-level DP-matching–A dynamic programming-based pattern matching algorithm for connected word recognition, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 27, No. 6, pp. 588–595 (1979)

[田口 09] 田口 亮, 岩橋 直人, 能勢 隆, 船越 孝太郎, 中野 幹生：モデル選択による言語獲得手法とその評価, 第 23 回人工知能学会全国大会講演論文集 (2009)

[山本 04] 山本 博史, 小窪 浩明, 菊井 玄一郎, 小川 良彦, 匂坂 芳典：複数のマルコフモデルを用いた階層化言語モデルによる未登録語認識, 電子情報通信学会論文誌 D-II, Vol. J87-D-2, No. 12, pp. 2104–2111 (2004)

[Yu 94] Yu, C. and Ballard, D.: On the Integration of Grounding Language and Learning Objects, in *Proc. 19th AAAI* (488-494)

[鳥井 06] 鳥井 豊隆, 長谷川 雄二, 中野 幹生, 中臺 一博, 辻野 広司：人・ロボットインタラクションシステムの為のミドルウェアの開発, 計測自動制御学会第 7 回システムインテグレーション部門 講演会 (SI2006), pp. 2D2–1 (2006)