語彙制限の無い発話の頑健な学習と理解を行う家庭用ロボット

船越孝太郎† 中野幹生† 鳥井豊隆† 長谷川雄二† 辻野広司†
木村法幸†† 岩橋直人††

† (株) ホンダ・リサーチ・インスティチュート・ジャパン 〒351–0188 埼玉県和光市本町 8–1
†† (独) 情報通信研究機構 〒619–0288 京都府相楽郡精華町光台 2–2–2
E-mail: †{funakoshi,nakano,tory,yuji.hasegawa,tsujino}@jp.honda-ri.com,
††{noriyuki.kimura,naoto.iwahashi}@nict.go.jp

あらまし 本稿では，家庭用ロボットに対話によって場所や人の名前を学習・理解させる方法を提案する． 名前の学習及び理解には，bag-of-words をベースにしたトピック認識技術を利用する．すなわち，ロボットは単語の出現頻度パターンとして名前を学習する． この方法により，音声認識誤りや音声認識の辞書に登録されていない単語に対しても頑健な学習と理解が可能となる．

キーワード 家庭用ロボット, 音声認識, 語彙外単語, 未知語，トピック認識

A domestic robot that acquires and recognizes spoken location names without vocabulary limitations

Kotaro FUNAKOSHI†, Mikio NAKANO†, Toyotaka TORII†, Yuji HASEGAWA†, Hiroshi TSUJINO†, Noriyuki KIMURA††, and Naoto IWAHASHI††

† Honda Research Institute Japan Co., Ltd. 8–1 Honcho, Wako-shi, Saitama, 351–0188 Japan
†† National Institute of Information and Communications Technology 2–2–2 Hikaridai Seikacho, Soraku-gun, Kyoto, 619–0288 Japan
E-mail: †{funakoshi,nakano,tory,yuji.hasegawa,tsujino}@jp.honda-ri.com,
††{noriyuki.kimura,naoto.iwahashi}@nict.go.jp

Abstract This paper presents a method that enables a conversational domestic robot to learn location names through speech interaction. Each acquired name is associated with a point on the map coordinate system of the robot. Both for acquisition and recognition of location names, a bag-of-words-based categorization technique is used. Namely, the robot acquires a location name as a frequency pattern of words, and recognizes a spoken location name by computing similarity between the patterns. This makes the robot robust not only against speech recognition errors but also against out-of-vocabulary names.

Key words Domestic Robot, Speech Recognition, Out-of-Vocabulary, Topic Recognition

1. はじめに

家庭用の移動ロボットに音声によって指示を与えるには，ロボットが家屋内の場所や物の名前を知らなければならない．家ごとに造りや配置，物の呼称が違う事ため，あらかじめロボットに全ての情報を与えておく事は出来ない．ユーザが各家庭でロボットにそれらの情報を教える必要がある．ユーザがロボットに場所の名前を教える最も簡単な方法は，ロボットを教えたい場所に連れて行き，その場所の名前を発話し覚えさせるという方法である．この際，自然な言語インタラクションを実現するために，ユーザに発話の語彙や発話方法に関する制約を課さないことが重要である．

従来，このようなインタラクションを可能とするための関連技術として，ユーザの発話から予め決められたいくつかのトピックを認識する手法 [1], [2] が提案されている．この手法では，大語彙音声認識によって認識された単語情報がトピック認識に利用される．トピックを学習時と認識時の両方で，ユーザによって自由に発声された発話が用いられることが特徴である．

しかしながら，大語彙音声認識であっても辞書に登録されていない単語（未知語）は正しく認識することができない．固有名詞など日常的に用いられるすべての単語をあらかじめ辞書に登録しておくことは現実的に不可能である．また，大きな背景雑音が存在したり，発話者とマイクの間の距離が長くなったりすることにより，マイクに入力される音声信号が歪む．この場合，たとえ既知語の音声が入力されたとしても認識誤りが生じやすくなる．実際に現状の大語彙音声認識器を用いて自由発話音声の認識を行った場合の単語認識率は，背景雑音がかなり低い場合でも 80%程度であり [3]，十分な性能とは言えない．

この問題に対して，従来，新しい単語を学習する方法 [4], [5] や，音声認識誤りを含んだままの情報を用いて処理を行う発話分類手法 [6] が提案されている．しかしながら，これらの手法では，多くの事前学習が必要であったり，発話方法に制約があったりするという問題があった．また，音声文書をキーワードにより検索する手法においては，音声認識誤りや未知語に対しての頑健性を高めるために，音声認識の複数認識候補を効率的に表現したワードラティスを用いる手法が提案されている [7], [8]．

音声によってロボットに場所や物の名前を学習／指示させる場合，音声認識結果が必ずしも完全に正しい必要は無く，認識結果が学習と指示の時に一貫していれば良い．そこで，本稿では，認識結果として得られるワードグラフを用いて，発話に未知の単語・認識誤りが含まれていても，内容を学習／指示できる，発話のトピック認識手法を提案する．提案手法は，少ない発話からの学習でも頑健な認識性能を示した．

最初に音声によるトピック認識手法を定義し，認識手法の評価実験を行う．次に，実環境下で評価するためにロボットへ実装し評価実験を行い，提案手法の有効性を示す．

2. 名前学習タスク

本稿で想定するタスクについて説明する．以後，このタスクを名前学習タスクとよぶ．名前学習タスクでは，ロボットはユーザの後について移動し，ユーザから場所の名前について教示を受ける．ロボットは，ユーザが発話した場所の名前と，その時のロボットの位置座標を連関させて記憶する．すなわち，ロボットが位置 P にいるときにユーザが発話 U を行った場合，ロボットは発話 U から場所の名前情報（すなわちある座標を指し示す情報）を抽出し，それを P に結びつけて記憶する．

名前情報を抽出する標準的な方法は，[9] のように，発話 U に対して音声認識を行い，認識された単語列から名詞句を取り出すというものである．しかし我々はある特定の名詞句の代わりに，単語の頻度情報を抽出して場所の名前情報として用いる．この抽出方法は 3. 節で説明する．

学習が完了したら，ロボットはユーザが音声で指示する場所へと移動する．ロボットはユーザの発話から名前情報を抽出し，それを用いて座標情報を読み出し，移動する．名前情報の同一性識別，すなわち認識が，座標情報を読み出す鍵となる．認識手法についても 3. 節で説明する．

本稿ではインタラクションに以下の仮定をおく．

- ロボットと人の間のインタラクションは，明確に分かれた二つのモード，学習モードと実行モードからなる．
- ユーザは学習か実行を開始する前に，次のインタラクションのモードを宣言する．
- 一度モードが宣言されれば，ロボットとユーザはそのモードに，別のモードが宣言されるまで従事する．
- ユーザの 1 発話は，一つの場所に関する名前情報か，モードを切り替えるコマンド一つだけを含む．
- 学習モードでは，提示された名前情報はロボットの現在地を指示する．
- 実行モードでは，提示された名前情報はロボットの目的地を指示する．

3. 発話トピック認識

3.1 問 題 点

大語彙連続音声認識は辞書に登録された単語（既知語）をつなぎ合わせて文として入力音声を認識するものである．音声によってロボットに場所を表す発話を学習・理解させるために，この大語彙連続音声認識ソフトウェアを用いることが考えられる．しかしながら，大語彙とはいえ，固有名詞など日常的に用いられるすべての単語をあらかじめ辞書に登録しておくことは現実的に不可能である．辞書に登録されていない単語（未知語）は正しく認識することができない．また，大きな背景雑音が存在したり，発話者とマイクの間の距離が長くなったりすることにより，マイクに入力される音声信号が歪む．この場合，たとえ既知語の音声が入力されたとしても未知語の場合と同じように認識誤りが生じやすくなる．一般に，ユーザは，辞書にどの単語が登録されているかを知ることができない．ユーザに負担をかけない自然な言語インタラクションのためには，ユーザが未知語を発話した場合や，認識誤りが起こっても，発話の意味を正しく理解できる手法が求められる．

3.2 提 案 手 法

提案する音声トピック認識手法（Bag of Words in Graph: BWG）は，語彙や文法を制限されること無しに自由に発声された音声のトピックをロボットが理解できるようにするものである．手法は，学習と認識の二つのフェーズからなる．まず学習フェーズでは，場所や人物などの個々のトピックに関して，ユーザによって話された音声を一つまたは複数用いることにより，音声とトピックの対応付けを学習する．次に認識フェーズでは，入力された音声に対して，学習フェーズで学習された複数のトピックのうちから適切なトピックを一つ選択する．提案する BWG 法の特徴は次の二つである．

- 入力音声を一つの文として認識するのではなく，複数の文の候補を含む，単語をエッジとした非循環グラフ（ワードグラフ）として認識する．
- 認識されたワードグラフを文書とみなし，bag-of-words モデルに基づいた文書トピック認識技術を適用する．

以下，これらの特徴について説明する．

3.2.1 ワードグラフを用いた音声認識

大語彙連続音声認識において，認識文の探索過程で文仮説をワードグラフで生成することにより効率的な探索を行う手

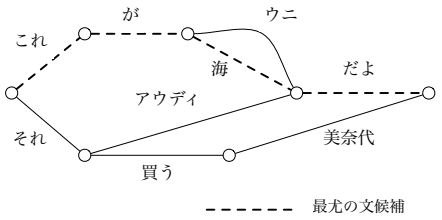

図 1　発話「これが海だよ」のワードグラフの例

法 [10] が提案されている．ワードグラフは，認識単語を表す edge の集合と，単語の始まりと終わりの時刻が付与されている vertex の集合からなる．ワードグラフの例を図 1 に示す．

ワードグラフの中から音響的および言語的な尤度を基準にして第 1 位から N 位までの認識文候補を選択することができる．未知語を含む音声が入力された場合，未知語の音声部分は，辞書中で音素系列が類似した単語や複数の単語の組み合わせとして表現される．音声認識結果として第 1 位に選択された文ではなく，ワードグラフそのものを用いることで，情報の消失を少なくして，未知語入力や誤認識に対して，後に続く処理の頑健性を高めることができる．

3. 2. 2　文書トピック認識の適用

BWG 法はワードグラフを文書とみなし，これに統計的な文書トピック認識の手法を適用するものである．文書トピック認識の手法として，Single random Variable with Multiple Value 法 [11] を用いた．この手法は，トピックが文法や単語の出現位置，順序に関係なく単語の出現頻度のパターンで定義される bag-of-words モデルに基づいたものである．テキストからランダムに選択された索引語が t_i である事象を表す確率変数 $T = t_i$ を与え，テキスト d がトピック c である確率 $P(c|d)$ を以下のように表す．

$$P(c|d) = \sum_{t_i} P(c|d, T = t_i)P(T = t_i|d)$$
$$\approx \sum_{t_i} P(c|T = t_i)P(T = t_i|d) \quad (1)$$

$$topic(d) = \underset{c \in C}{\operatorname{argmax}} P(c|d) \quad (2)$$

学習フェーズでは，学習用音声サンプル集合を用いて $P(c|T = t_i)$ を計算する．認識フェーズでは，入力音声から $P(T = t_i|d)$ を求めて，$P(c|d)$ を計算し，式 (2) のように $P(c|d)$ が最も大きくなるトピック c を認識結果とする．

索引語は，学習データの中に含まれる単語のうち，トピックとの相互情報量が大きいものを選択する．相互情報量 $I(T_i; c)$ は次式で計算される．

$$I(T_i; c) = H(c) - H(c|T_i) \quad (3)$$

ここで，T_i は索引後 t_i が文書の中に存在する/存在しないの二値を取る．$H(c)$ は確率変数 c のエントロピーを，$H(c|T_i)$ は事象 T_i のもとでの c の条件付きエントロピーを表す．もし，c と T_i が互いに独立ならば，$I(T_i; c)$ の値は 0 になる．相互情報量を用いることで，トピックの識別に貢献しない索引語を排斥できる．

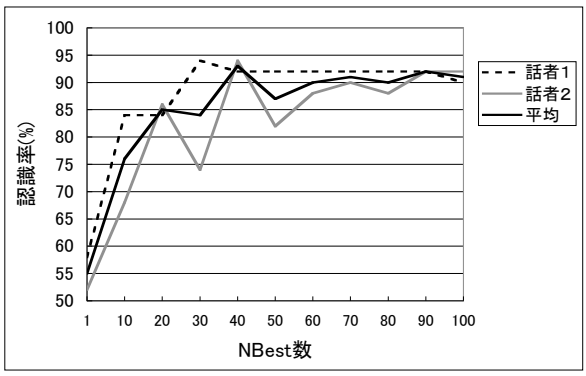

図 2　使用する文候補の数とトピック認識率の関係

3. 3　評　　価

音声からワードグラフを生成するために，Advanced Telecommunication Research Institute で開発した HMM モデルによる大語彙音声認識ソフトウエア [12] を用いた．マイクロホンは携帯型パソコンに内臓のものを使用した．

トピックを決定付ける単語が未知語である場合の BWG 法の評価を行うため，トピックは辞書に登録されていない名前を持つ 10 名の人物とした．男性話者 2 名の発話音声を用いて評価した．学習フェーズと認識フェーズではともに，各トピックに対して 5 回ずつ発話した音声を用いた．学習フェーズと認識フェーズで発話された文は次の通りである．X の部分に人物名が挿入される．

学習フェーズ

- 彼は X さんです．
- X さんは有名です．
- これは X さんのものです．
- 部長の X です．
- X さんをご存知ですか？

認識フェーズ

- X さんの席はどこですか？
- X さんを呼んでください．
- X さんを探しています．
- ここからは X さんに任せます．
- それは X さんの責任です．

まず，ワードグラフを使用することの効果を評価するために，出力されたワードグラフの中で，SVMV 法による文書トピック認識で使用する部分ワードグラフの大きさを変化させ，これがトピック認識率にどう影響するかを調べた．部分ワードグラフは，第 1 位から n 位までの文候補からなるものとした．使用するワードグラフのサイズが大きくなるに従って高い認識率が得られた（図 2）．このことから，ワードグラフを用いて情報量の欠落を少なくすることで，トピック認識率を向上させることに成功したと言える．

この時の，学習に用いる発話数と認識性能の関係について評

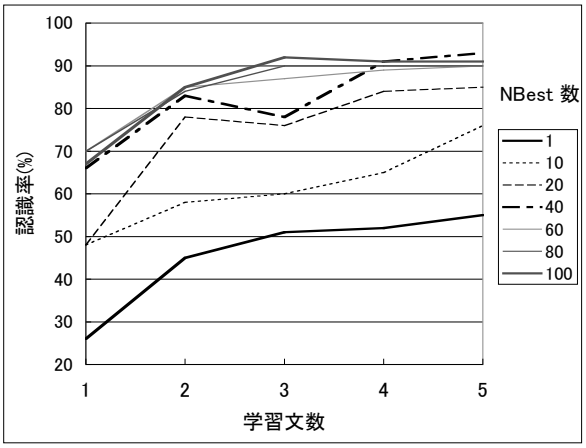

図 3　学習分数を変化させた時のトピック認識率（2 話者平均）

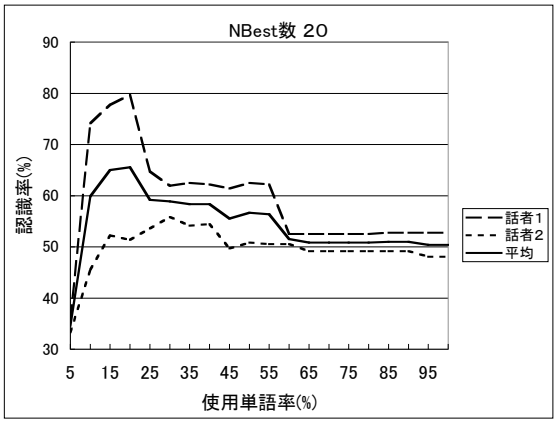

図 4　全体の入力単語に対する索引語の割合を変化させた時のトピック認識率

価を行った．使用するワードグラフの大きさを決める N-Best 数を変えた場合の結果を図 3 に示す．学習文数を増やすことによって認識率を向上させることが可能であることがわかる．また，従来手法 [6] のように 1-Best だけで学習を行うよりも，ワードグラフを用いて学習することにより，少ない発話数で高い認識率が実現できた．

次に，相互情報量を基準にして索引語の数を制限することの効果を調べた．入力単語全体に対する検索語の数の割合を変化させたときの認識率を図 4 に示す．個人差もあるが，使用する検索語の数が全体の 20%〜30%とした場合，平均して高い認識率を得られた．しかし，60%を超えると話者 1，話者 2 共に認識率が低下している．検索語の数が多すぎても少なすぎても認識率を悪化させてしまうことは，学習データの量とモデルの複雑さとの関係から理解できる．よって，検索語の選択に相互情報量を用いることの有効性が示された．(注1)

4.　音声対話ロボット

4.1　システムアーキテクチャ

図 5 に対話ロボットのアーキテクチャを示す．図 5 中の対話

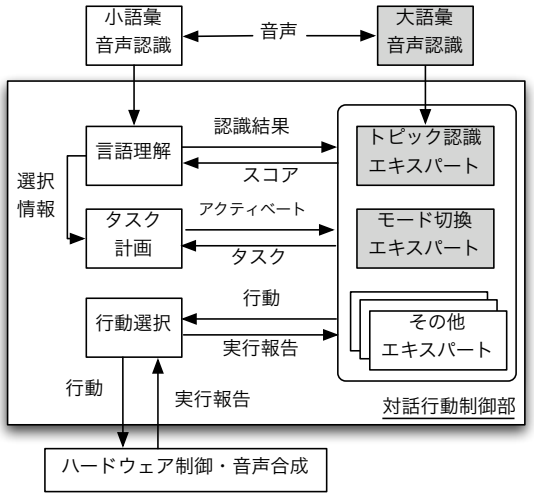

図 5　対話ロボットアーキテクチャ

行動制御部は，特定のタスクに 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 ルを複数用いるモデル RIME (Robot Intelligence based on Multiple Experts)(注2)に基づいている．各エキスパートは，特定ドメインの対話に従事したり，移動などの物理タスクを遂行したりする．対話行動制御部の出力は MADL (Multi-modal Action Description Language) の形式をとり，実際にロボットを物理レベルで制御する行動制御部に送られる．

名前学習のために，モード切換エキスパートとトピック認識エキスパートを用意した．エキスパートは文法ベースの小語彙音声認識器を互いに共有するが，トピック認識エキスパートだけはそれ自身の統計ベースの大語彙音声認識器を用いる．上記の 2 エキスパートと大語彙音声認識器は，図 5 中に灰色で示してある．

4.1.1　エキスパート及び対話行動制御部

前述のように対話行動制御部はその内部に複数のエキスパートを持つ．各エキスパートはある特定のサブタスクの遂行を受け持つ．ロボットがあるサブタスクを実行しようとするとき，そのサブタスクに対応するエキスパートが対話行動制御部の中で活動状態にあり，ロボットの次の行動を選択する．ユーザからの発話が認識されると，言語理解結果と文脈から，次に活動状態になるエキスパートが決まる．エキスパートはオブジェクト指向言語におけるオブジェクトに相当する．各エキスパートはそれ自身の内部状態を持ち，言語理解結果や基盤化状態，局所的な行動計画などを保持する．

対話行動制御部はその内部に言語理解，タスク計画，行動選択の三つのモジュールを持ち，これらのモジュールがエキスパートの活動を調整する．言語理解部は小語彙音声認識器の認識結果を各エキスパートに配布し，各エキスパートが返すスコア（理解の確信度）に基づいてタスク計画部に最も適切なエキスパートを提示する．それに基づき，タスク計画部は実際にどのエキスパートをアクティベートするかを決める．行動選択部

(注1)：相互情報量の有効性検証に用いたデータについては [13] を参照されたい．

(注2)：[14] では MEBDP と呼ばれていた．

はアクティベートされたエキスパートから次の行動を受け取り，ハードウェア制御・音声合成部に渡す．

各エキスパートはその内部状態に対話行動制御部の各モジュールがアクセスするためのインタフェース（メソッド）を備える必要がある．ここでは，名前学習タスクに関係するメソッドのみを説明する．*understand* メソッドが音声認識結果が得られたときに言語理解部によって呼び出されると，エキスパートはドメイン依存の発話パターンを用いて認識結果の解釈を行い，0 から 1 の間でスコアを返す．このスコアはその発話が当該エキスパートによって扱われるべきであるかどうかについての確信度を表す．*select-action* メソッドは行動選択部によって呼び出され，内部状態に基づいて次にとるべき行動を返す．このメソッドは，直前の行動の遂行が完了すると呼ばれる．しかし，直前の行動を表す MADL において発話待ちフラグがセットされていた場合は，このメソッドの呼び出しはユーザ発話があるまで保留される．

4.1.2 名前学習タスク用エキスパート

図 6 中のシナリオに沿って，前述のモード切換エキスパートとトピック認識エキスパートの振る舞いについて説明する．以下，それぞれのエキスパートを MSE と TRE とよぶ．

a） 学習モード

ロボットへの教示を始めるために，まずユーザはモード切換コマンドを発話する（U0）．この発話は小語彙音声認識器によって認識され言語理解部を通じて全てのエキスパートに送られる．認識結果が十分によければ，MSE が全てのエキスパートの中で最も高いスコアを返し，次にアクティベートされるエキスパートとなる．

アクティベートされた MSE はコマンドに従い TRE を次にアクティベートして学習モードに入るようにタスク計画部に指示する．TRE はまず，学習モードに入ったことをユーザに知らせる（R1）．これ以降，他のエキスパートがアクティベートさ れるまでの間，TRE は言語理解の都度，中間的なスコア (0.5) を返す．(アクティベートされていない TRE は常に 0 を返す.）これにより，より高い確信度を返すことによって他のエキスパートがアクティベート状態を奪わない限り，TRE がユーザ発話に応答する．また TRE がアクティベートされていて学習モードにいる時，常にロボット がユーザの近くにいるように制御する．

TRE がユーザ発話（U2）に反応するときは，小語彙音声認識器からの認識結果は無視して大語彙音声認識器からの認識結果を使用する．ユーザ発話を受け取った TRE はワードグラフから得られた単語の頻度情報とロボットの現在位置を関連づけて記憶し，名前を学習した旨をユーザに伝える（R3）．ユーザはその場でさらに発話を追加することで，名前認識の認識精度を向上させることができる．追加発話から得られた頻度情報によって，記憶されている頻度情報は更新される．

しばしばロボットはユーザのコマンド発話を場所の名前として誤学習する．このような場合，ユーザは「取り消し」ということで直前の誤学習を取り消すことができる．

U0: 学習モードに入って．
R1: 場所の名前を覚えます．
(ロボットはユーザの側に移動)
U2: ここがキッチン．
R3: この場所の名前を覚えました．
(ユーザが移動し，ロボットは追随する)
U4: ここは正面ドアだよ．
R5: この場所の名前を覚えました．
U6: 実行モード．
R7: 実行モードに入ります．
U8: キッチンにいって．
R9: 向かいます．
(ロボットはキッチンに移動する)

図 6 Example Interaction

図 7 The Robot

b） 実行モード

学習が完了したら，ユーザはモード切換コマンド（U6）を発話する．このコマンドに対して TRE がアクティベートされ実行モードに入る．TRE は実行モードにはいったことをユーザに伝える（R7）．TRE はユーザ発話（U8）に対して，音声認識結果から頻度情報を取り出し，それと学習結果が最もよく一致するトピックを選択する．そして応答するとともにそのトピックに連関した座標に移動する（R9）．

ユーザは場所を連続して指定することで，ロボットに移動経路を指示することができる．経路を指示するためには，まず経路指定開始コマンドを「経路指定開始」と発話する．コマンドが正しく理解されればロボットがその旨を返答するので，ユーザは場所の名前を一つずつ発話する．最後に「経路指定終了」と経路指定終了コマンドを発話すれば，ロボットは指定された場所へ順番に移動する．これらのコマンドは MSE によって処理される．

4.2 実　　装

台車ロボット（図 7）を用いて提案手法を実装し，予備的な実験を行った．大語彙音声認識には Julius [15] を用い，小語彙音声認識には Julian を用いた．Julian は Julius が ngram 言語モデルを用いる代わりにネットワーク文法を用いるようにしたものである．

ユーザとロボットの位置情報の取得には超音波タグを用いた．

表 1 認 識 結 果

被験者 #	1	2	3	4	平均
正解率 (%)	73.3	86.7	86.7	86.7	83.3

ユーザとロボットは共に超音波タグを装備し，超音波センサーを備えた部屋の中で活動する．したがって，ロボットは自身とユーザの位置を，部屋の中の絶対座標系の上で認識する．

超音波タグによって検出される座標は，ノイズのために静止状態であっても揺らぐ．加えて，ユーザ発話を待って待機中のロボット自体もわずかにドリフトしてしまう．そのためロボットは二つの座標点間の距離がある閾値以下にある場合はそれらを同一地点と見なす．このような措置は．一カ所で複数の発話によって教示を行う場合に必要となる．上記の閾値は実験的に設定した．

4.3 評　　価

4.3.1 方　　法

実験は前述の部屋の中で行った．部屋の大きさは 7×4 メートル四方である．この部屋の中の五カ所を選択し，#1 から #5 と書かれたカードを置いた．

被験者は 5 枚のカードの場所に順に移動し，それぞれの場所の名前を説話マイクを用いて発話した．場所の名前は各被験者が任意に与えた．

4.3.2 結　　果

4 名の被験者が実験に参加した．学習モードでは 2 名の被験者（被験者 #1 と #3）が各地点において 1 回だけ教示したのに対して，残りの 2 名の被験者は各地点で 3 回教示を繰り返した．

表 1 に結果を示す．実験に使用した音声認識機（Julius）の辞書と言語モデルは Web テキストから構築されソフトウェアとともに配布されているものを使用した．辞書のサイズは 60248 語である．全教示発話の中に含まれた 26 単語（数え）のうち，2 単語だけが辞書に登録されていなかった．従って未知語率は 7.7%であった．既知語と未知語の間で，トピック認識正解率に違いはなかった．

5. おわりに

本稿では，オンラインの音声インタラクションを通じて場所の名前を学習するロボットを提示した．場所の名前は，文書分類において広く用いられている bag-of-words モデルに基づいた BWG (Bag of Words in a Graph) 法を用いて発話中のトピックとして獲得および認識される．

BWG 法では，ワードグラフの形で表現された音声認識結果に対して SVMV 文書トピック認識手法 [11] を適用する．3. 節で述べた評価により，BGW 法が支配的なトピックを表す言語表現中に未知語が含まれていても正確にトピックを認識することが示された．

ロボットのアーキテクチャは RIME モデル [14] に基づく．RIME は，特定のタスクに従事する複数のエキスパートを駆使することで会話ロボットの複雑で知的な振る舞いを実現する．本稿で提案した BMG 法は RIME の 1 エキスパートとしてロボットに実装された．RIME のおかげで，会話ロボットを簡単に実装することができた．

現在のところ，BMG 法は 1 発話に 1 トピックだけが含まれることを前提としているため，インタラクションのスタイルが大きく制限される．今後の課題として，1 発話中に複数のトピックが含まれる場合にも対応できるように BMG 法を拡張することが考えられる．

また，BMG 法はトピックを既知語の集合として非明示的な形で獲得する．そのためロボットが獲得したトピックを明示的に表現することが難しく，ロボットがどのような学習を行ったのかユーザが把握しにくい．この問題については現在 BWG 法と音素列獲得を組み合わせた手法の開発を進めている．

文　　献

[1] A. L. Gorin, G. Riccardi and J. H. Wright: “How may i help you?”, Speech Communication, **23**, pp. 113–127 (1997).

[2] J. Chu-Carroll and B. Carpenter: “Vector-based natural language call routing”, Computational Linguistics, **25**, 3, pp. 361–388 (1999).

[3] T. Hori, C. Hori, Y. Minami and A. Nakamura: “Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition”, IEEE Transaction on audio, speech and language processing, 11 (2005).

[4] N. Iwahashi: “Language acquisition through a human-robot interface by combining speech, visual, and behavioral information”, Information Sciences, **156**, pp. 109–121 (2003).

[5] D. Roy and N. Mukherjee: “Towards situated speech understanding: visual context priming of language models”, Computer Speech & Language, **19**, 2, pp. 227–248 (2005).

[6] 浅見, 竹澤, 菊井：“音声インターフェースのための発話を単位とした話題および発話行為タイプ推定”, 電子情報通信学会論文誌, **J87-D-II**, 2, pp. 436–446 (2003).

[7] 西崎, 中川：“音声認識誤りと未知語に頑健な音声文書検索手法”, 電子情報通信学会論文誌, **J86-D-II**, 10, pp. 1369–1381 (2003).

[8] M. Saraclar and R. Sproat: “Lattis-based search for spoken utterance retrieval”, HLT-NAACL 2004 (2004).

[9] C. Yu and D. H. Ballard: “On the integration of grounding language and learning objects”, In Proc. of the 19th National Conference on Artificial Intelligence (AAAI) (2004).

[10] M. Oerder and H. Ney: “Word graphs: an efficient interface between continuous speech recognition and language understanding”, Proc of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 119–122 (1993).

[11] M. Iwayama and T. Tokunaga: “A probabilistic model for text categorization: Based on a single random variable with multiple values”, In Proc. of the 4rth Applied Natural Language Processing Conference (ANLP), pp. 162–167 (1994).

[12] 伊藤, 葦苅, 實廣, 中村：“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, pp. 221–222 (2004).

[13] 木村, 岩橋, 中野, 船越：“家庭用ロボットのための語彙制限の無い発話の頑健な学習と理解”, FIT2007 第 6 回情報科学技術フォーラム予稿集 (2007).

[14] M. Nakano, Y. Hasegawa, K. Nakadai, T. Nakamura, J. Takeuchi, T. Torii, H. Tsujino, N. Kanda and H. G. Okuno: “A two-layer model for behavior and dialogue planning in conversational service robots”, Proc. of IROS 2005, pp. 1542–1548 (2005).

[15] A. Lee, T. Kawahara and K. Shikano: “Julius — an open source real-time large vocabulary recognition engine”, In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691–1694 (2001).