E-062

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

Robust Acquisition and Recognition of Utterance Meanings Without Vocabulary Limitation for Domestic Robots

木村法幸　岩橋直人
(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
{noriyuki.kimura,naoto.iwahashi}@nict.go.jp

中野幹生　船越孝太郎
(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
{nakano, funakoshi}@jp.honda-ri.com

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]が Gorin らによって提案されている．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[2]，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[3][4]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[5]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[6][7]．

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. Bag-of-words モデルに基づいた発話のトピック認識

2.1. 提案手法

提案する音声によるトピック認識手法（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．

1．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．

2．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．

2.1.1. ワードグラフによる音声認識

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法が提案されている．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．

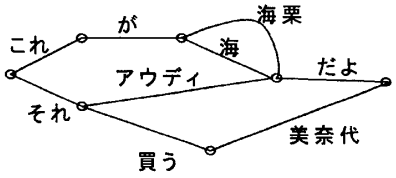

図 1　「これが海だよ」と発話したワードグラフの例

2.2. 文書トピック認識技術の適用

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[8]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検索語が 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．

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

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，$P(c|d)$ が最も大きくなるトピック c を認識結果とする．検索語は，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．

$$I(T_i,c)=H(c)-H(c|T_i)$$

ここで，T_i は，検索語 t_i が文書の中に存在する/存在しないの二値を取る．$H(\cdot)$はエントロピーを表す．

2.3. 評価実験

２種類の評価実験を行った．評価実験１では，トピックを決定付ける単語が１つという簡単な課題で評価を行い，評価実験２では，２単語がトピックを決定づける課題で評価を行った．

音声認識器には，（株）国際電気通信基礎技術研究所で開発した HMM モデルによる大語彙音声認識ソフトウェアを用いた．言語モデルには，旅行対話１００万文で学習を行った物を用いた（語彙数は１０万語）。マイクロホンは携帯型パソコンに内蔵のものを使用した．

発話者は男性話者２名とした．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．

2.3.1. 評価実験 1

トピックを決定付ける単語が１つの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 個の人名とした．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 の部分に人名が挿入される．

学習用文章
　彼は X さんです．
　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？
評価用文章
　X さんを呼んで下さい．
　X さんの席はどこですか．
　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．

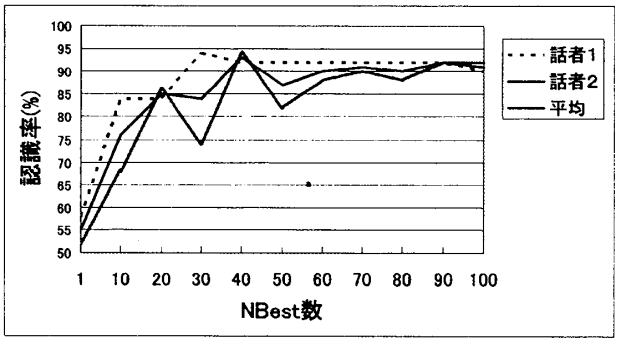

図 2　使用する文候補の数とトピック認識率の関係

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価を行った．使用するワードグラフの大きさを決める NBest 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[5]のように 1Best 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．

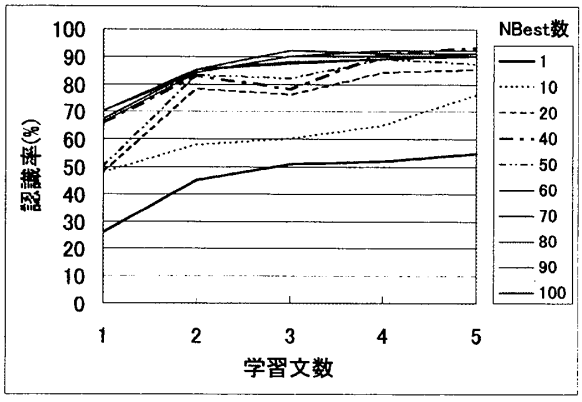

図 3　学習文数を変化させた時のトピック認識率
（２話者平均）

2.3.2. 評価実験2

トピックを決定付けるキーワードを人名＋兄弟名の2単語とした．兄弟の名称は「兄」，「弟」，「妹」の3つとした．人名は，既知語の人名2名だけの場合と，既知語の人名2名に未知語の人名2名を加えた場合で実験を行った．

学習／評価で発話された文の例は次の通りである．Xの部分に人物名，Yに兄弟の名称が挿入される．

学習用文章

定型文1：(キーワードのみ)
　XさんのY
定型文2：(キーワードの前後に文章あり)
　彼はXさんのYです．
　この方が，XさんのYです．
　Xさんの Y は、有名です．
　これは、X さんの Y のものです．
　X さんの Y をご存知ですか？

評価用文章

定型文1：(キーワードのみ)
　XさんのY
定型文2：(キーワードの前後に文章あり)
　XさんのYを呼んでください．
　XさんのYを探しています．
　XさんのYの席はどこですか？
　ここからは、XさんのYに任せます．
　それは、XさんのYの責任です．

自由文：(人名，兄弟名称の間に他の単語が挿入)
　それはXさんのだから，彼のYに渡しておいて．
　彼のYは、Xさんだよ。
　Xさんが来たら、Yの部屋で待ってもらって。
　XさんはYです。
　Xさん．その方はYですか．

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，キーワード1とキーワード2の組み合わせに対して上記文章を発話した音声を用いた．定型文1（キーワードのみ）は、同じ文章を5回発声した．

この時の学習，評価に用いた音声認識器の性能を確認するために，既知語の人名で行った発話音声の認識結果を表1に示す．

表1　音声認識性能（既知語のみ）

	評価値	話者1	話者2
既知語	単語正解精度(1Best)	92.77%	85.65%
	単語正解精度(NBest)	96.24%	94.61%
	文正解精度(1Best)	72.00%	52.00%
	文正解精度(NBest)	78.67%	74.67%

人名を既知語のみとし，トピック数を6にした場合の実験結果を表2に示す．2人の被験者の平均結果が最も高かったNBest数30の認識率である．

表2　既知語のトピック認識率

学習	評価	話者1	話者2
定型文1	定型文1	90.00%	70.00%
	定型文2	76.67%	66.67%
	自由文	83.33%	76.67%
定型文2	定型文1	100.00%	80.00%
	定型文2	96.67%	76.67%
	自由文	90.00%	86.67%
平均		89.44%	76.11%

次に，未知語を加えトピック数を12にした場合の結果を表3に示す．2人の被験者の平均結果が最も高かったNBest数20の認識率である．

表3　未知語を含むトピック認識

学習	評価	話者1	話者2
定型文1	定型文1	96.67%	50.00%
	定型文2	66.67%	41.67%
	自由文	83.33%	41.67%
定型文2	定型文1	76.67%	55.00%
	定型文2	76.67%	58.30%
	自由文	78.33%	61.67%
平均		79.72%	51.38%

評価実験1に比べて認識率が低くなった。これはトピック間でキーワードが重複しているためであると考えられる．また，学習，評価方法，個人差によって認識率に大きな差があった．特に個人差が大きく，定型文1で学習し定型文1で評価した場合，話者1では96.67%という高い認識率であるが，話者2では50%以下となった．話者2の場合，表1に示した文正解精度がNBestに対して1Bestで大きく低下している．これは，大語彙音声認識のデコーディングの際にワードグラフが大きく広がっていることを意味し，このようにワードグラフが広がったことにより不要な単語が増えトピック認識結果に影響していると考えられる．

次に，不要な単語の悪影響を軽減するために相互情報量を基準にして検索語の数を制限することの効果を示す．表3の結果の場合（NBest数20）に，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complexity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．

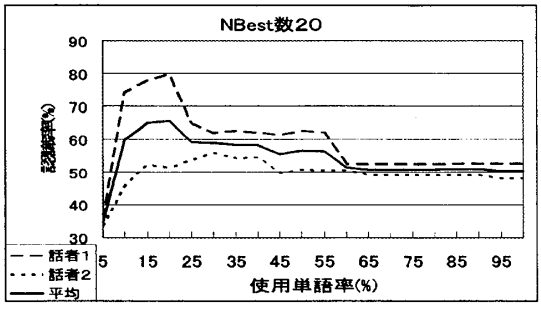

図4　全体の入力単語に対する検索語の割合を変化させた時のトピック認識率

3. ロボットへの適用

提案した手法を，ロボットへ適用し評価を行った．図5に示す台車ロボットを使用した．音声認識はJuliusを使用し，マイクには接話型のマイクを使用した．実験環境は，超音波タグによって絶対座標が取得できる4m×7mの部屋を使用した．ロボットと人がそれぞれ超音波タグを装着することによって，座標情報を取得する．このロボットシステムには学習モードと実行モードがある．学習モードでは，ロボットは常にユーザに近づくように移動し，ユーザが場所の名前を発話するとそれをトピックとして覚える．実行モードでは，ユーザの発話した場所に移動する．学習モードと実行モードの切り替えは音声による指示で可能である．この指示の認識はトピック認識とは別の小語彙音声認識器を用いて行う．

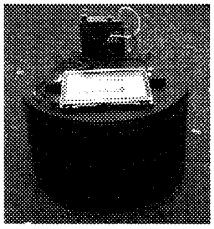

図 5 台車ロボット

3.1. 評価実験

部屋内に5つの位置を選択し，＃1～＃5と書かれたカードを置いておく．ロボットを学習モードにし，被験者は＃1～＃5の場所へ行き，ロボットに対して位置の名前を発声する．学習が完了したら，実行モードに切り替えそれぞれの場所の名前を発声する．4人の被験者が実験に参加した．話者1，3は，学習時に1つの位置につき1回の発話で学習を行い，話者2，4は1つの位置につき3回の発話で学習を行った．評価時は，1つの場所に対して3回の発話を行った．以下に，学習及び実行方法の例を示す．

話者　　：「学習モード．」
ロボット：「場所の名前を覚えます．」
　（ロボットはユーザの近くに移動する）
話者　　：「ここは台所．」
ロボット：「この場所の名前を覚えました．」
　（人は次の場所に移動し，ロボットは後をついてくる）
話者　　：「ここは入り口．」
ロボット：「この場所の名前を覚えました．」
話者　　：「実行モードに移って．」
ロボット：「実行モードに移りました．」
話者　　：「台所へ行って．」
ロボット：「行きます．」
　（ロボットは台所に移動する）

結果を表4に示す．

表 4　ロボットを用いた実験での場所認識の正解率

	話者1	話者2	話者3	話者4	平均
認識率(%)	73.3	86.7	86.7	86.7	83.3

位置の学習に使われた名前の単語のうち，26単語中2単語が辞書に登録されていない単語であった．未知の単語を含む場合と含まない場合で，場所の認識率にほとんど差は無かった．

4. まとめ

音声の認識結果をワードグラフで表現し，ワードグラフを文書とみなして文書トピック認識の手法を適用することを特徴にした，音声トピック学習・認識手法を開発した．評価実験により，トピックを決定付ける単語が未知語である場合でも高いトピック認識率が達成可能であることを示した．今後の課題としては，トピックを決定付けるキーワードが重複する場合の性能向上や，背景雑音が大きかったり，発話者とマイクの距離が遠かったりした場合の性能評価や，複数のトピックが順番に表現された発話からトピック列を認識する手法の開発などが挙げられる．

参考文献

[1] A. L. Gorin, et al. “Learning spoken language without transcriptions,” in In Proc. of IEEE Workshop Speech Recognition and Understanding, 1999.
[2] T. Hori, et al., “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*, 1 (11), 2005.
[3] N. Iwahashi, “Language acquisition through a human-robot interface by combining speech, visual, and behavioral information,” Information Sciences, 156, 109–121, 2003.
[4] D. Roy and A. Pentland, “Learning Words from Sights and Sounds: A Computational Model”, Cognitive Science, 26(1), 113-146, 2002.
[5] 浅見克志, et al., ”音声インタフェースのための発話を単位とした話題及び発話行為タイプ推定”, 電子情報通信学会論文誌, J87-D-2 (2), 436-446, 2004.
[6] M. Saraclar and R. Sproat, “Lattis-Based Search for Spoken Utteracne Retrieval”, Proc. HLT-NAACL, 2004.
[7] 西崎博光, 中川聖一, “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), 1369-1381, 2003.
[8] M. Iwayama and T. Tokunaga, “A probabilistic model for text categorization based on a single random variable with multiple values,” In Proc. 4th Applied natural language processing conference, 162–167, 1994.