4S-2

マルチモーダル情報を利用した 未知語を含む発話のドメイン選択精度の向上

高橋　裕己[1]　中野　幹生[2,3]　岩橋　直人[4]　左　祥[5]　船越　孝太郎[3]
岡　夏樹[5]　菅野　重樹[6]

[1]早稲田大学 創造理工学研究科　[2]早稲田大学 理工学術院
[3]（株）ホンダ・リサーチ・インスティテュート・ジャパン
[4]京都大学　[5]京都工芸繊維大学　[6]早稲田大学 創造理工学部

１．はじめに

近年ロボットの技術の発展は著しく，既に様々なところで様々なタスクをこなすロボットが実用化されてきている．そこで重要になるのが，ロボットに行わせたい仕事を命令するためのインタフェースである．その際に音声で命令できると便利である．

現在音声対話ができるロボットの研究が進められている．家庭用ロボットなどの多くのタスクを行う音声対話ロボットには，複数のタスクドメインを扱うことのできるマルチドメイン対話システムが必須である．

マルチドメイン対話システムでは，ドメインの選択を行う必要がある．ドメイン選択をすることで，聞き返しができ，スムーズに対話を行うことができる．しかし，発話に未知語が含まれている場合，ドメイン選択が困難になる．本稿では，文型マッチングと画像認識結果からの特徴量を用いて，未知語を含む発話のドメイン選択を扱う．

２．関連研究

２．１ドメイン選択

ドメイン選択には大きく分けて二つの手法が存在する．分類問題として扱う方法（n 分類法）[1]と入力された発話がどのくらい各ドメインの発話である可能性が高いかを表すスコアをドメイン毎に算出し，最も高いスコアを出したドメインを選ぶスコアリング法 [2]がある．

前者の方法に比べて，後者の方法は，ドメイン毎に異なる特徴量を用いることができることや，ドメインが増えたときに訓練用のデータセットを用意し直さなくても，新しいドメインの訓練データのみを用意すればよいなどの利点があり，拡張性が高い．

Improving Domain Selection Accuracy of Utterances Including Unknown Words Using Multimodal Information : Yuki TAKAHASHI (Waseda Univ.), Mikio NAKANO (Waseda Univ./HRI-JP), Naoto IWAHASHI (Kyoto Univ.), Xiang ZUO (Kyoto Institute of Technology), Kotaro FUNAKOSHI (HRI-JP), Natsuki OKA (Kyoto Institute of Technology), Shigeki SUGANO (Waseda Univ.)

２．２ 未知語の扱い

未知語は，音声認識・言語理解の語彙に登録されていない単語のことである．従来のドメイン選択手法では未知語が発話されることを考慮していない．未知語が存在した場合，音声認識・理解が失敗してしまうために，ドメインを選択することも難しくなってしまう．

ドメイン選択において未知語を扱った研究に，自然な対話の中で物の名前を覚える研究がある[3]．しかし，これは統計的手法を用いていなかったために精度が良くないという問題があった．

３．提案手法

本稿では，統計モデルを用いて未知語を含んだ発話のドメイン選択を行う手法を提案する．そのためにドメイン選択に文型マッチング特徴量と画像認識結果からの特徴量を用いる．

３．１ 文型マッチング特徴量

本研究の目的はドメイン選択であるので，従来の未知語検出手法をそのままは用いず，発話がドメインに適合するかを調べる．これは，音声認識結果をフィラーモデルを含む文型とマッチングさせることにより行う．未知語の存在により，音声認識・理解の結果が誤っていても，文型とマッチングする際に未知語の存在を認めることで，未知語周辺の単語がドメイン特有の言い回しであった場合にそのドメインである確率を上げる．

３．２ 画像認識結果からの特徴量

場所の指示を行う際など，人間は命令に即したジェスチャを行うことがある．つまりドメインとジェスチャには相関があると考えられる．そして，ユーザの画像からはそれ以外にも様々な情報を得ることができる．そこで，ユーザのジェスチャと物体を持っているかどうかを特徴量として用いることでユーザの意図が推定でき，結果としてドメイン選択の精度が向上できると考えた．これによって文型とのマッチングでは判定しにくいドメインの判定ができるようになることが期待される．

４．実験

４．１ 概要

提案手法の有効性を確かめるために実験を行った．本実験のタスクは発話のドメインを判定することである．ドメイン選択法は，n 分類法とスコアリング法の両方を用いた．

４．２ データ

物体の名前を覚える，物体移動，電話番号検索，掃除の指示，人探しの 5 つのドメインについて各ドメイン 30 発話ずつ 20 人（男女 10 人ずつ）の被験者から合計で 3000 発話を収集した．

４．３ 音声認識結果からの特徴量抽出

音声認識には，Julius ver. 4.2.1*を用いた．音響モデルは，Julius ディクテーション実行キット 4.2 付属のものを用いた．言語モデルは，自作したドメイン用言語モデルと，ディクテーションキット付属の大語彙言語モデルの両方を用いて認識を行った．これは，発話検証スコアを求めるためである．自作した言語モデルを使って行った音声認識では，集めた発話を二つのグループに分け，片方のグループから作った言語モデルを使い．もう片方の音声認識を行った．この結果から，発話時間や単語数，発話検証スコアなど 12 個の特徴量を抽出した．さらに 3.1 で述べた文型マッチング特徴量 4 つを抽出した．

４．４ 画像認識結果からの特徴量抽出

発話中にユーザがどのようなポーズを取っているか，物体を持っているか等の特徴量を抽出するため画像認識を行った．Microsoft 社の Kinect for windows と OpenNI を使用し，骨格追跡からポーズ認識を行った．物体を持っているかどうかは Kinect と OpenCV を用いて，判別を行った．画像認識結果から，ポーズ 3 種類の判定とオブジェクト有無で計 4 つの特徴量を抽出した．

４．５ 実験条件

下記の Table 1 の通り，All は組み合わせた特徴量を含め，すべての特徴量を，Pattern は文型マッチング特徴量のみを，Image は画像認識結果からの特徴量のみを使ってドメイン選択を行った．

Table 1 Experimental conditions

条件名	条件
All	全ての特徴量を使用
Pattern	Base + 文型マッチングの特徴量を使用
Image	Base + 画像認識結果からの特徴量を使用
Base	文型マッチング，画像認識結果からの特徴量以外を使用

４．６ 分類結果

使用した分類器は，n 分類法では Weka ver.3.6.8[4] の Logistic(ロジスティック回帰)，スコアリングでは SMO (サポートベクタマシン)である．

また，特徴選択を Weka の ClassifierSubsetEval で Greedy Stepwise の forward を用いて行い，特徴選択を行わない場合と結果を比較した．

結果を下記の Table 2 に示す．5 分類法では特徴選択をした結果の条件 Image が最も高い精度となっている．スコアリングの結果で一番高い精度となったのは，同じく特徴選択をした，条件 Pattern である．

Table 2 Accuracy of domain selection (%)

条件	5 分類法		スコアリング法	
	特徴選択なし	特徴選択あり	特徴選択なし	特徴選択あり
All	90.65	93.67	85.50	86.07
Pattern	91.67	94.00	85.93	88.93
Image	92.23	97.33	84.74	87.84
Base	92.67	95.33	87.40	87.60

４．７ 考察

文型マッチング・画像認識結果からの特徴量をそれぞれ用いた場合は精度の向上を確認できた．しかし，両方を加えた場合の精度向上は確認できなかった．これは，物体の認識精度や二つに分けたグループの偏りにより特徴選択が上手くいかなかったなどの原因が考えられる．これらは，物体認識精度を向上すること，グループ間の偏りが少なくなるように分けることで改善できる可能性がある．

５．おわりに

本稿では，文型マッチングと画像認識結果からの特徴量を用いて未知語が存在する発話のドメイン選択を行う手法を提案した．そして実験の結果．それぞれの特徴量の有効性が示唆された．今後は，提案手法をロボットへ実装することを目指していく．

謝辞

本研究成果の一部は，科学研究費補助金基盤研究(S)25220005 の助成を受けたものであり，また，早稲田大学理工研プロジェクト研究「自然と共生する知能情報機械系に関する基盤研究」の一環として行われたものである．ここに謝意を表す．

参考文献

[1] R. Lane et al.: Topic classification and verification modeling for out-of-domain utterance detection. *Proc. Interspeech.* (2004)

[2] M. Nakano et al.: A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems, *Proc. SIGDIAL*. pp. 18-29 (2011)

[3] 中野他: 自然な対話の中で物体の名前を覚えるロボット, 第 23 回人工知能学会全国大会 (2009)

[4] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition. Morgan Kaufmann (2005).

* http://julius.sourceforge.jp/