社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－185 (14)
2008－SLP－ 71 (14)
2008／5／23

文書分類手法を応用したインタラクティブプレゼンテーションにおける視聴者発話の理解

簑津 真一郎 †, 中野 幹生 ††, 船越 孝太郎 ††, 木村 法幸 ‡,
岩橋 直人 ‡, 石塚 満 †, 辻野 広司 ††

† 東京大学大学院 情報理工学系研究科
†† (株) ホンダ・リサーチ・インスティチュート・ジャパン
‡ (独) 情報通信研究機構

本稿では，インタラクティブプレゼンテーションシステムにおける視聴者発話の理解のための発話分類法を提案する．提案法は文書分類法を視聴者発話の n-best 音声認識結果に適用する．分類時だけではなく，分類モデルの学習時にも音声認識結果を用いる．レストラン案内システムのデータを用いた評価実験により提案法の有効が示された．

Understanding Audience Utterances in Interactive Presentations using a Text Categorization Method

Shinichiro Minotsu †, Mikio Nakano †, Kotaro Funakoshi ††, Noriyuki Kimura ‡,
Naoto Iwahashi ‡, Mitsuru Ishizuka †, Hiroshi Tsujino ††

† Graduate School of Information Science and Technology, University of Tokyo
†† Honda Research Institute Japan Co., Ltd.
‡ National Institute of Information and Communications Technology

This paper presents a novel method for classifying audience utterances to recognize intentions in interactive presentation systems. It applies a text classification technique to the set of words that appeared in the n-best recognition outputs of audience utterances. Not only utterance classification but also training the classification model uses speech recognition outputs. The results of an experiment using the audience utterances collected using a restaurant explanation system showed that the effectiveness of the proposed method.

1 はじめに

ロボットやアニメーションエージェントの重要なアプリケーションの一つに自動プレゼンテーションがある．我々はプロジェクタスクリーンを用いてプレゼンテーションを行うヒューマノイドロボットシステムを構築している [10]．プレゼンテーションのコンテンツは簡易な記述言語 MPML-HR (Multimodal Presentation Markup Language for Humanoid Robots) で記述できる．さらに我々はシステムと言語を拡張し，視聴者の発話に応じてプレゼンテーションの内容を動的に変更できるようにした [11]．そのためには，視聴者の発話を正しく理解する必要がある．我々のシステムでは，一般的なタスク指向対話と異なり，視聴者の発話は限定されていて少数のカテゴリに分類できると仮定している．したがって，視聴者の発話を理解することは，発話をそれらのカテゴリに分類することと考えられる．

視聴者発話を理解する単純な方法は，ネットワーク文法駆動の音声認識を用いることである．しかしながら，バラエティに富む発話を理解するためには，それらをカバーする文法を構築する必要があり，そのような文法構築は容易ではない．

もう一つの方法は，実際の視聴者発話を使って構築された統計言語モデルを用いて音声認識を行い，その認識結果をキーワードスポッティングやベクトル空間モデルに基づく方法によって分類することである．しかしながら，この方法は発話の書き起こしが必要となるため，コストがかかる．これは，簡単にプレゼンテーションコンテンツが構

築できるという我々のシステムの利点を損なう．

本稿では上記とは異なるアプローチを提案する．正しい音声認識結果を得ることは目的とせず，分類しようとしている発話が過去に得られた訓練データとどのくらい近いかを推定する．本アプローチでは，視聴者発話を大語彙音声認識にかけた結果に現れる単語の集合に文書分類手法を適用する．従来の発話分類手法と異なり，分類モデルの学習時にも音声認識結果を用いる．本方法は書き起こしを必要とせず，学習用発話のカテゴリラベルのみを用いる．これにより，認識用文法や統計言語モデルを構築するより容易に発話分類器を構築することができる．

本方法は，家庭内ロボットのための場所の名前の獲得法 [3, 9] に基づいている．この方法は，ロボットに場所の名前を教える人と，その教えた名前を用いて指示を行う人が同一であることを仮定している．しかし，インタラクティブプレゼンテーションは不特定の視聴者を想定しているため，上記の仮定を用いることができない．そこで，理解しようとしている発話の話者とは異なる話者のデータを用いて学習を行う．その代わりに，学習データを増やす．数百発話を学習データとして用いることにより，異なる人の発話に対しても高い分類精度を達成することができた．

2 インタラクティブプレゼンテーションにおける視聴者発話理解

我々が構築しているインタラクティブプレゼンテーションシステムは，プロジェクタスクリーンを用いるヒューマノイドロボット，スクリーン上を動くアニメーションエージェント，または，音声のみを用いて，与えられたシナリオにしたがってプレゼンテーションを行う．そして，視聴者からの発話にしたがって，プレゼンテーションの内容や実行順序を変更する [11]．

プレンゼンテーションコンテンツとインタラクション機能は MPML-HR の拡張版である MPML-HR ver. 3(以後 MPML-HR3 と呼ぶ) と呼ぶスクリプト言語で記述することが可能である [11]．これにより，音声言語処理の専門知識がなくても，インタラクティブプレゼンテーションのコンテンツを作成することが可能である．

MPML-HR3 で記述されたプレゼンテーションコンテンツは複数の部分に分かれており，その各々が概ね一つのプレゼンテーションスライドに相当する．システムは，視聴者発話が「現在のスライドの説明の繰り返し」,「前のスライドに戻る」,「最初に戻る」,「プレゼンテーションを終了する」,「特定のトピックの詳細な説明をする」などのいくつかのカテゴリに分類されることを仮定している．図 1 にレストラン説明システムと人間のユーザとの対話の例を示す．このシステムは 4 節で記述するデータ収集で用いている．

システム	これから中華料理店香港楼を紹介します．どうぞよろしくお願いします．(中略) これで香港楼の紹介を終わります．みなさまのお越しをお待ちしております．
ユーザ:	営業時間は何時から何時まで？
システム:	はい，営業時間は午前 11 時から夜 10 時 30 分です．ラストオーダーは 9 時半になります．

図 1: インタラクション例

MPML-HR3 では，プレゼンテーション記述者は各カテゴリの発話の認識のためのネットワーク文法を単純な記法で書くことができる．以下に例を示す．

1. 次 の スライド [に (行って | 移って) [ください]]
2. (初め | 最初) の スライド [に 戻って [ください]]

この記法において，“[ ]” はあってもなくてもよい部分を表し “(|)” はどちらでもよいことを表す．「次のスライドに移って」という発話は 1 番目の文法で受理できる．現在のシステムはこれらの文法を全部一度に用いる音声認識を用いている．そして，どの文法を用いて認識したかという情報を用いて視聴者発話を分類する．

しかしながら，プレゼンテーション記述者にとって，様々な視聴者発話をカバーする文法を書くことは容易ではない．文法でカバーされない発話は誤認識されやすい．そのような文法外発話は音声認識結果を棄却する発話検証技術により無視することができるが，文法カバー率が低いと多くの発話が無視されてしまうという問題が生じる．

3 文書分類法を用いた発話の分類

視聴者発話は正確に認識する必要はなく，分類できればよい．本稿で提案する方法は，以下に詳述する BWG (Bag of Words in a Graph) と呼ぶ場所の名前を覚えるロボットのために開発された

方法 [3] をベースにしている．このロボットに，現在ロボットがいる位置の名前を「そこはキッチンの前」などの発話で教えると，その後「キッチンの前に行って」などの発話で場所を指示することができるようになる．これは，場所の名前を教える発話を場所ごとにカテゴライズして分類モデルを学習し，指示する発話をそのモデルに基づいて分類することで可能になる．ここではロボットに名前を教える人と，その教えた名前を用いて指示を行う人が同一であることを仮定しており，1 カテゴリ（場所）あたり非常に少ない数（1-4）の発話で学習が可能である．

我々は，この方法を視聴者発話の分類に応用した．実際のプレゼンテーションでは，あらかじめ視聴者の発話を収録することができないので，学習データの発話者は分類しようとしている発話の話者と異なっていることを仮定する．そのかわり，学習データは数百発話存在すると仮定する．さらに，学習データは書き起こされていないがカテゴリラベルが与えられているとする．カテゴリラベルをつけるのは，書き起こすよりずっと少ない労力ですむ．

BWG 法は，文書分類法 SVMV (Single Random Variable with Multiple Values) 法 [7] に基づいている．SVMV 法は索引語を用いて文書を特徴付ける方法で，比較的少量の学習データでも高い分類精度が得られる．BWG 法は，音声認識結果に SVMV 法を適用する．大語彙統計言語モデルを用いた音声認識によって視聴者発話を認識した際の n-best 認識結果に現れる単語の集合を用いる（本来はワードグラフを用いるが，実装上の都合から n-best 認識結果を用いる）．たとえば，もし n-best 認識結果が以下の (a) であれば，単語リスト (b) を用いる．

(a) 最初 の スライド に 戻って ください
最後 の スライド に 行って
最初 の スライド に 戻って

(b) 最初, の, スライド, に, 戻って, ください,
最後, 行って

BWG 法では，発話 u がカテゴリ c である確率 $P(c|u)$ を，索引語と呼ぶ単語の集合を用いて，以下のように分解する．

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

ここで $T = t_i$ は，ある単語集合の中に存在する索引語からランダムに選んだ索引語が t_i である確率事象を表す．$P(T = t_i|u)$ は，u の n-best 音声認識結果に現れる索引語からランダムにひとつを選んだ時にそれが t_i である確率であり，u の n-best 音声認識結果中の索引語における t_i の相対頻度で推定できる．$P(c|u, T = t_i)$ は，u の n-best 音声認識結果に現れる索引語からランダムに一つ選んだものが t_i であった時，その発話がカテゴリ c である確率である．$T = t_i$ が与えられたときに u と c が条件付独立と仮定すると以下のようになる．

$$P(c|u) \quad \approx \quad \sum_{t_i} P(c|T = t_i) P(T = t_i|u) \quad (2)$$

学習フェーズでは，カテゴリラベルのついた発話から $P(c|T = t_i)$ を得る．$P(c|T = t_i)$ はベイズ則を用いて以下のように変形できる．

$$P(c|T = t_i) = \frac{P(T = t_i|c) P(c)}{P(T = t_i)} \quad (3)$$

ここで，$P(T = t_i|c)$ は，c に属する学習用発話データの n-best 音声認識結果に現れる索引語における t_i の相対頻度，$P(T = t_i)$ はすべての学習用発話データの n-best 音声認識結果に現れる索引語における t_i の相対頻度で推定できる．n-best 認識結果内に現れた索引語の頻度を求める際には，同じ単語が何度現れても 1 回と数える．たとえば，上記の (a) の場合,「スライド」という単語は 3 回現れているが，1 と数える．すなわち (b) の各語はすべて 1 回の出現と数える．これは，回数を反映させた場合，予備実験で性能が低下したからである [3].

理解フェーズ，すなわち，分類フェーズでは，$P(T = t_i|u)$ が入力発話から得られる．そして，各カテゴリ c に関して $P(c|u)$ を求め，これが最大になる c を選ぶ．

索引語の集合は，学習用発話の n-best 音声認識結果に現れる単語の中から，カテゴリとの相互情報量に基づいて選択する．相互情報量 $I(T_i; C)$ は以下のように定義される．

$$I(T = t_i; C) = H(C) - H(C|T = t_i) \quad (4)$$

ここで，$H(C)$ は発話カテゴリを値にとる確率変数 C のエントロピーで，$H(C|T = t_i)$ は $T = t_i$ が与えられたときの C の条件つきエントロピーである．もし C と $T = t_i$ が独立ならば，$I(T = t_i; C)$ は 0 になる．相互情報量を用いることにより，カテゴリの特徴付けに寄与しない一般的な単語を除くことができる．BWG 法では，学習データの音声認識結果に現れるすべての単語の中での検索語の割合が定めた値になるように，検索語の数を制限する．

4 評価

4.1 データ

提案法を評価するため，レストラン紹介システムとユーザとの対話データを用いて実験を行った．システムは音声のみを出力モダリティとして用いた．システムは名前，場所，お勧め料理などの情報を提供することによりレストランを紹介した．もし，ユーザからの割り込みの発話がなければ，約1分で説明を終える．システムは文法ベースの音声認識を用いて発話の分類を行った．音声認識はJulianとその付属の音響モデルを用いた[8]．音声合成はNTT-IT社のFinceVoiceを用いた．システムはユーザ発話の認識結果の信頼度に応じて，棄却，確認要求などを行った[11]．

各被験者は以下の課題にしたがって質問をした．

> 説明を聞いて以下の問いに答えてください．
>
> 1. 営業時間とラストオーダーの時間
> 2. パーティーは開催できますか？できるとすると何人くらいまでのパーティーができますか？
> 3. 全体的に高めだと思いますか？安めだと思いますか？

被験者には，次のカテゴリの割り込み（質問・要求）ができることを教示した．

- 一つ前にもどる　・次に行く
- 最初にもどる　・おわる
- もう一度繰り返す　・予算
- 営業時間　・ラストオーダー
- お勧め料理　・パーティー
- 場所

ただし，「ここに書いてある通りに発音せず，命令したり依頼したりする口調で話してください．たとえば，『ひとつ前に戻って』や，『ひとつ前に戻ってください』のように言ってください．」と教示した．課題が終了した場合，「おわり」と言ってプレゼンテーションを終わらせるように教示した．対話の制限時間は5分とした．

被験者は21歳から47歳までの男性9名，女性4名の計13名である．各々の被験者に4つのレストラン案内のシステムそれぞれと一度づつ対話を行ってもらい，計52の対話を収録した．マイクの入力を録音し，発話の開始・終了時間のタグ付けと発話内容の書き起こしを手作業で行った．

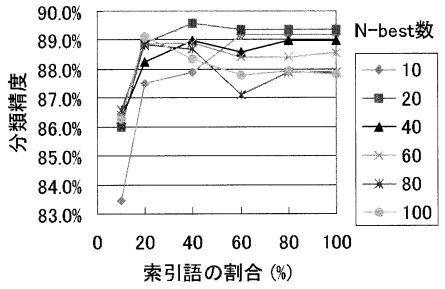

図 2: 提案法での分類精度

カテゴリ	発話数	異なり発話の数
一つ前にもどる	3	3
次に行く	8	2
最初にもどる	15	5
おわる	71	24
もう一度繰り返す	13	9
予算	135	63
営業時間	138	41
ラストオーダー	33	9
お勧め料理	77	44
パーティー	108	31
場所	71	39
違う	2	2

表 1: 各カテゴリの発話の頻度

全部で977のユーザ発話が収集され，各々の発話を手作業で15のカテゴリに分類した．カテゴリは，教示した11のカテゴリ，および「はい」,「いいえ」,「違う」,「その他」である．複数のカテゴリに分類された発話もある．実験では，複数のカテゴリに分類されたものと「その他」のものを除いた．さらに,「はい」「いいえ」の発話は人間でも分類しにくいものが多かったため，除外した．結果として12カテゴリの674発話を用いた．表1に，各カテゴリの発話数とモーラレベルでの異なり発話数を示す．

カテゴリ毎の発話数はかなり異なる．しかし，この実験設定はカテゴリ毎に同じ量の発話を用意するのに比べて，より現実的である．一人のユーザの平均発話数は51.8($s.d.$ 16.8)であった．

各カテゴリの発話には多くのバリエーションがあった．以下に「おわる」カテゴリの発話例を示す．

> 「あーじゃーおわります」,「ありがとおわっていーよ」,「おしまい」

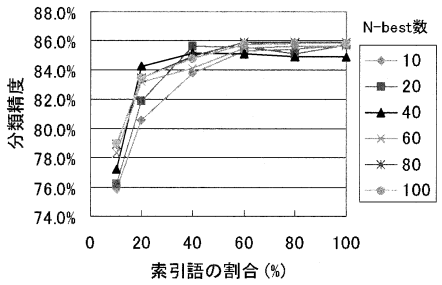

図 3: 6 人分の発話を用いて学習したときの分類精度

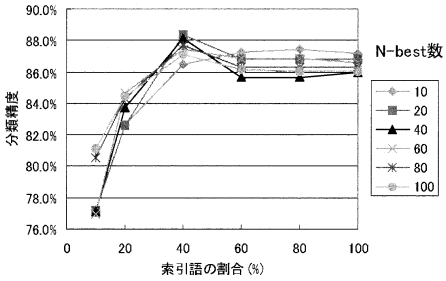

図 4: 半分の学習データ量を用いた場合の分類精度

4.2 発話分類結果

上記の発話を分類することによって，提案手法を評価した．大語彙音声認識は ATR で開発された ATRASR[6] を用いた．言語モデルと辞書は旅行ドメインの 100 万発話から構築されたものを用いている．辞書のサイズは 10 万語である．

一人のユーザの発話を理解するのに残りの 12 人分の発話を用いて学習を行った．各話者の発話の分類精度の平均を計算した．

ある 1 人のユーザの発話をテストデータとして除外した場合，学習データの 10-best, 20-best, 40-best, 100-best の音声認識結果に現れた単語数はそれぞれ 3,939，4,541，5,168，5,705 であった．このとき学習発話数は 651 であった．100%の単語が索引語として用いられた場合には，これらの数の単語が用いられたことになる．

図 2 に，学習時と分類時に用いた音声認識結果の数（n-best 数）と，索引語の割合を変化させた時の分類精度を示す．20-best の音声認識結果を用い，索引語の割合を 40%にした時に 89.6%の精度を得た．これらの結果は，学習データの話者が入力発話の話者と異なっていても，BWG 法が効果的に動作することから，提案アプローチの有効性を示しているといえる．20-best の音声認識結果を用いた時の方がより多くの音声認識結果を用いた時より分類精度が高かったのは，もしより多くの音声認識結果を用いると，音響的に遠い単語が含まれる可能性があるからだと考えられる．また，学習データに現れた認識結果のすべての単語を用いた場合（索引語の割合が 100%）よりも，索引語を制限した場合の方が良い精度が得られている場合があることから，この実験結果は，索引語を相互情報量を用いて制限することの有効性を示唆している．

逆に，データ収録時の文法ベース音声認識を用いて分類した時の精度は 66.3% であった．これは文法外発話の割合が非常に高かった (79.7%) からだと考えられる．書き起こしに基づいて文法を拡張し文法外発話の割合を 33.0%にした結果，83.2%の精度を得た．しかしながら，文法の拡張は，非常にコストがかかり，MPML-HR3 を用いたプレゼンテーションシステム開発の良さを損なう．さらに，ATRASR と Julian は別の音響モデルを用いているので単純に比較はできないが，提案法の分類精度がまだ高い．これらのことから，提案法は文法ベースの音声認識に基づく方法に対して精度と開発労力の両方で利点があることが示唆された．

4.3 学習データの量とバリエーションの影響

学習データの量とバリエーションの影響を推定するため，二つの追加実験を行った．一つは学習データを 12 人分から半分の 6 人分にしたものである．各々のユーザの発話の分類の精度を求める時に，学習データとして，残り 12 人のうちランダムに選んだ 6 人の発話データのみを用いた．結果を図 3 に示す．もう一つは，学習の際に，各ユーザの発話の半分を用いたものである．図 4 に結果を示す．6 人分の発話のみを用いた場合の方が，用いた発話の人数は同じで半分のデータを用いた場合よりも精度が低下した．これは，人数が多い方が発話のバリエーションが多いからではないかと考えられる．

5 関連研究

発話分類技術が用いられている別のドメインにコールルーティングがある [1, 2]．多くの場合，コールルーティングでは，ユーザ発話の書き起こしか

ら得られた統計言語モデルを用いた音声認識の結果に対し，ベクトル空間モデルなどを用いた文書分類法を適用する．

このような方法は，統計言語モデルと分類モデルの学習のために，多くの発話の書き起こしが必要になる．書き起こしは労力がかかるため，プレゼンテーションシステムの構築を簡単にするという我々の目的に合わない．能動学習や半自動学習なども提案されているが [12]，ブートストラップのためにある程度の書き起こしは必要となる．

書き起こしを用いずに分類モデルを学習する方法が提案されている [4, 5]．それらの方法では，音素 n-gram を用いた音素タイプライタによる音声認識を行い，各カテゴリを特徴づける音素列を相互情報量を用いて発見する．しかしながら，一般的に，音素タイプライタの認識精度は，大語彙言語モデルを用いた場合に比べて悪いため，結果として，発話分類の精度も高くない可能性が高いと考えられる．詳細な比較は今後の検討課題である．

6 おわりに

本稿では，発話分類手法をインタラクティブプレゼンテーションの視聴者発話の理解に適用する方法を提案した．この手法では，n-best 音声認識結果に現れた単語の集合を分類モデルの学習と実際の分類の両方で用いる．我々は実験的に提案手法の有効性を確認した．本方法は，学習発話の書き起こしを必要としない．これは，我々が構築しているインタラクティブプレゼンテーションシステムの特徴が容易にプレゼンテーションコンテンツを作成できることであることにマッチしている．

本稿では，理解の精度の向上のみを扱ったが，インタラクティブプレゼンテーションシステムでは，理解結果の信頼度が大きな役割を果たす [11]．今後は理解結果の信頼度を精度よく求める方法の研究を行っていく予定である．

参考文献

[1] Jennifer Chu-Carroll and Bob Carpenter. Vector-based natural language call routing. *Computational Linguistics*, Vol. 25, No. 3, pp. 361–388, 1999.

[2] Stephen Cox and Ben Shahshahani. A comparison of some different techniques for vector based call-routing. In *Proc. Eurospeech-2001*, pp. 2337–2340, 2001.

[3] Kotaro Funakoshi, Mikio Nakano, Toyotaka Torii, Yuji Hasegawa, Hirosh Tsujino, Noriyuki Kimura, and Naoto Iwahashi. Robust acquisition and recognition of spoken location names by domestic robots. In *Proc. 2007 IEEE/RSJ IROS*, pp. 1435–1440, 2007.

[4] A. L. Gorin, D. Petrovksa-Delacretaz, G. Riccardi, and J.H. Wright. Learning spoken language without transcriptions. In *Proc. ASRU-99*, 1999.

[5] Qiang Huang and Stephen J. Cox. Automatic call-routing without transcriptions. In *Proc. Eurospeech-2003*, pp. 1909–1912., 2003.

[6] 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.

[7] Makoto Iwayama and Takenobu Tokunaga. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proc. Applied Natural Language Processing Conference (ANLP-94)*, pp. 162–167, 1994.

[8] Tatsuya Kawahara, Akinobu Lee, Kazuya Takeda, Katsunobu Itou, and Kiyohiro Shikano. Recent progress of open-source LVCSR engine Julius and Japanese model repository. In *Proc. Interspeech-2004*, pp. 3069–3072, 2004.

[9] 木村法幸, 岩橋直人, 中野幹生, 船越孝太郎. 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 講演論文集 E-062, 2007.

[10] Yoshitaka Nishimura, Kazutaka Kushida, Hiroshi Dohi, Mitsuru Ishizuka, Johane Takeuchi, Mikio Nakano, and Hiroshi Tsujino. Development of multimodal presentation markup language MPML-HR for humanoid robots and psychological evaluation. *International Journal of Humanoid Robotics*, Vol. 4, No. 1, pp. 1–20, 2007.

[11] Yoshitaka Nishimura, Shinichiro Minotsu, Hiroshi Dohi, Mitsuru Ishizuka, Mikio Nakano, Kotaro Funakoshi, Johane Takeuchi, Yuji Hasegawa, and Hiroshi Tsujino. A markup language for describing interactive humanoid robot presentations. In *Proc. IUI-07*, 2007.

[12] Gokhan Tur, Dilek Hakkani-Tür, and Robert E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, Vol. 45, No. 2, pp. 171–186, 2005.