データベース検索タスクにおける対話文脈を利用した音声言語理解

神田直之[†] 駒谷和範[†]
尾形哲也[†] 奥乃博[†]

データベース検索タスクにおける音声対話システムにおいて，音声認識誤りの棄却や解釈の曖昧性の解消のために対話文脈的な制約を取り入れる手法について述べる．まず，データベース検索タスクの対話は「検索条件の指定」と「情報の提示要求」の遷移からなるとモデル化する．さらに，検索条件をその入力の順序に従って木構造状に管理する．これらのモデルに基づく特徴を決定木としてまとめ，言語理解部に利用する．提案手法をレストランデータベース検索システムとして実装し，20 名の被験者による評価実験を行った．実験の結果，提案手法に基づく対話文脈的な特徴を加えることで，19.8%の意味理解誤り削減が得られた．さらに，ホテルデータベースを用いて同様の実験を行い，提案手法により得られた制約が対象データベースに依存せず適用できることを確認した．

Spoken Language Understanding Using Dialogue Context in Database Search Task

Naoyuki Kanda,[†] Kazunori Komatani,[†] Tetsuya Ogata[†]
and Hiroshi G. Okuno[†]

We describe how to introduce contextual information in spoken dialogue systems for the database search task. In this paper, we model dialogues in the database search task as consisting of two modes: “specifying retrieval conditions” and “requesting detailed information about specific entries”. Furthermore, we manage retrieval conditions as a tree structure considering their order. Based on those models, we introduce decision tree learning using features reflecting situations in the task as well as those derived from current utterances. By using the output of the decision tree, the system can appropriately select words from a speech recognition result even when it contains some errors. The experimental result with restaurant database showed that our method identified users’ intentions 19.8% better than that without the contextual information. Moreover, the experiment with hotel database showed that obtained constraints could be applied to another domain without any more training.

1. はじめに

近年，バス運行情報案内やレストラン検索など様々なタスクドメインにおいて音声対話システムが作成されている[8),17)]．音声対話においては音声認識誤りが不可避であるので，その言語理解処理において，誤りを含んだ音声認識結果からも正確にユーザの意図を抽出できる頑健さが要求される．これまで，音声認識誤り部分を識別するための研究として，音声認識結果の事後確率に基づく単語ごとの信頼度計算[16)]や，音響尤度比を用いた認識誤り検出[11)]，また GMM を用いた非音声入力の棄却[9)]などが行われてきた．これらの研究では，受理/棄却の判定をその発話から得られる情報のみから行っており，対話レベルの情報は用いていない．本研究では，対話文脈レベルの情報もあわせて内容語の取捨選択を行う．

また，音声言語の特徴として，書き言葉に比べ省略された表現が多く，解釈に曖昧性が生じやすい．たとえば我々のレストラン検索システムにおいても，「カード」とだけ発話があった場合，それだけでは「カードが利用できるレストランを検索したい」のか「レストランで利用できるカードが知りたい」のか曖昧性が生じる．言語理解処理では，それまでの対話情報を利用して，これらを正しく解釈できなければならない．本研究では対話から得られる情報を用いて決定木を学習し，言語理解に組み込むことで曖昧性の解消を行う．

さらに音声対話システムを構築する際には，熟練した技術者の作成したルールによる場合が多く，容易にシステム開発が行えないという問題がある．このため，

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

必要な情報を対話データから自動的に学習する，様々な研究が行われてきた[4),6),13),15)]．しかしながら，大量の対話データを収集するには多くの労力が必要である．もし特定のシステムで学習された情報が他のシステムでも利用できれば，非常に有用である．本研究では，関係データベースを検索するタスク一般に成り立つ対話モデルを設計する．これに基づいた特徴を用いて決定木学習を行うことにより☆，対象とするデータベースに依存せずに利用できる対話文脈的な制約を構成する．

以下，各章の構成を説明する．2 章では関連研究の中での本研究の位置付けを述べる．3 章では我々が提案するデータベース検索タスクの対話モデルについて説明する．4 章で，この対話モデルに基づいた情報を反映させた言語理解部について説明し，5 章でその実験的評価を行う．最後に本稿のまとめを行う．

2. 関 連 研 究

対話文脈情報として，対話行為の連接確率をコーパスから学習して利用する研究がある[13),15)]．ここでは有用な制約を得るために，対話行為を「SET-START-TIME（会議室の使用開始時間の設定）」や「駐車場の検索の依頼」のように，ドメイン固有のレベルで設計している．このように，対象データベースの項目ごとに対話行為を設計する場合，対話行為数はデータベースの項目数に従って増加する．このため項目数の多いデータベースでは，上記の手法は必要な学習データ量が膨大になる．我々はデータベース検索タスクの概略をモデル化することで，対象とするデータベースに依存せずに成り立つ対話文脈的な制約を取得し利用する．

文献 14) では，文脈を反映させた言語理解を行うために，人手で記述したルールを用いている．本研究ではこのルールを包括する，より多くの特徴を定義し決定木学習を行うことで，多数のルール間の関係を自動的に学習する．

また，文献 5) では，言語的・音響的な特徴に加えて，直前のシステム発話タイプなどの対話文脈的な特徴を用いて機械学習を行い，言語理解結果の受理/棄却を行っている．同様に文献 4) や 6) では，対話文脈的な特徴を用いて，発話の受理/棄却を行っている．我々は，より多様な対話文脈的制約を得ることを目指し，データベース検索タスクにおける 2 つの対話モデル「対話の進行モデル」「履歴の構造モデル」を新たに考案し導入する．対話の進行モデルは，対話が「検索条件の指定」と「情報の提示要求」の 2 つの状態間の遷移からなるとモデル化したものである．また，履歴の構造モデルでは，検索条件を入力順に従って木構造状に管理することにより，各条件間の重要さを表現する．さらに，文献 4)〜6) の手法では，各々が対象とするドメインにおいて受理/棄却の判定精度の向上が報告されているが，得られた制約を対象ドメイン外に適用することは考慮されていない．我々が新たに導入した対話モデルはデータベースの内容に依存しないため，これらから得られる特徴量は，他ドメインにも適用可能である．

3. データベース検索タスクの対話モデル

本稿では，**表 1** のような関係データベース（各エントリは，属性名とその値の組で表され，各データは 1 つのキー属性を持つ）に対し検索するタスクを扱う．以下対象ドメインをレストランデータベースとして議論を進める☆☆．このタスクは，必要なスロットが事前に決定できるスロットフィリング型[1)] タスクに対し，以下の点で異なる．

- タスク達成に必要なスロットがユーザごとに異なる：ユーザによって，予算が重要であったり，場所が重要であったりする．
- ユーザの意図が対話中に変化する：システムからの返答を聞くまで満足できるエントリが存在するかどうかが分からないため，ユーザはしばしば検索された結果を見て要求を変更する．

対象とするデータベース検索タスクでの対話の例を**図 1** に示す．

表 1 関係データベースの例（レストランドメイン）
Table 1 Example of relational database (restaurant domain).

属性	値
店名（キー属性）	クスノキ食堂
フードタイプ	和食
説明	学生の町，京都らしく安くて ...
住所	京都市左京区吉田 ...
電話	555-5555
営業時間	18:30–23:30
休業日	木
交通	京阪出町柳駅より徒歩 10 分
カード	JCB，VISA，アメックス
駐車場	2 台
下限予算	400 円
上限予算	1,000 円

☆ ここでは分類器の一例として決定木を利用したが，決定木以外の分類器を利用することも可能である．

☆☆ 便宜上対象ドメインを設定しているが，以下の議論は関係データベースであれば対象ドメインによらず適用できる．

S1：こちらはレストラン案内システムです．どのような店をお探しですか？
U1：**祇園にある店を探しているのですが．**
S2：祇園という条件ですと 259 件あります．
U2：**和食の店で何かないですか？**
S3：祇園，和食という条件ですと 51 件あります．
U3：**3000 円以下**
S4：祇園，和食，3000 円以下という条件ですと 15 件あります．
U4：**1000 円以下だったら？**
S5：祇園，和食，1000 円以下という条件ですと 2 件見つかりました．吉田屋，クスノキ食堂です．
U5：**クスノキ食堂の住所を教えてください．**
S6：クスノキ食堂の住所は京都市左京区吉田 . . . です．

図 1 データベース検索タスクにおける対話の例
Fig. 1 Example dialogue in database search task.

対話の各時点での状態が表現されていれば，それを対話文脈的な制約として言語理解に利用することができる．スロットフィリング型タスク（天気案内，バス運行情報案内など）を扱うシステムでは，対話の流れを事前にすべて記述できるため，有限オートマトンによって対話の流れをモデル化し，対話の各時点で明示的に状態が定義できる[7),8)]．しかしながら，データベース検索タスクでは上記の特性より事前に対話の流れを記述しておくことが不可能であるため，これらとは異なるモデル化を行う必要がある．

また各システムごとに対話状態を設計すると，それに基づく制約はそのシステム以外では利用できないものとなる．本研究では対象とするデータベースに依存しない制約を得ることを目的としている．そのため発話の棄却や解釈に有用な制約を得ることができ，かつデータベースに依存しない対話状態設計を行った．

以下では，本稿で提案するデータベース検索タスクのモデルに関する説明を行う．

3.1 対話の進行モデル

本研究では，上で述べた条件を満たすモデルの 1 つとして，データベース検索タスクにおける典型的な対話の進行を，まず希望する検索条件を指定して数件までエントリを絞り込み，その後絞り込んだエントリに関するその他の属性の値を取得するものと想定する．これはレストランデータベースにおいては，はじめに希望する店（エントリ）を絞り込み，その後絞り込んだ店の住所や電話番号など（キー属性以外の属性）を取得することに相当する．ここでの前者を「検索条件の指定」モード，後者を「情報の提示要求」モードとする．本研究ではデータベース検索タスクにおける状態は 2 つのモードのいずれかに属し，対話はモード間

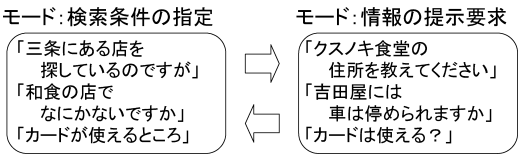

図 2 データベース検索タスクでの 2 つのモード
Fig. 2 Two modes in model for dialogue flow.

を遷移することで行われるとする（**図 2**）．以下，これを**対話の進行モデル**と呼ぶ．言語理解部では，各時点で対話がどちらのモードであるかや，その時点でのエントリ数などが，対話状態の表現として利用される．

たとえば図 1 では，U1〜U4 が「検索条件の指定」モードにおける発話であり，U5 によって対話の状態は「情報の提示要求」モードに移行している．また，得られた検索結果を吟味した結果ユーザが満足しなかった場合，再び「検索条件の指定」モードに戻ることもある．

3.2 履歴の構造モデル

データベース検索タスクではユーザごとに検索要求が異なる．もし，どの検索条件がユーザにとって重要かが分かれば，重要な検索条件が音声認識誤りによって上書きされてしまうことなどを避けることができる．本研究では以下の仮定を置き議論を進める．

- 長い間変更されない検索条件は重要である

これを表現するために本研究では，検索条件（属性・値ペア）の履歴を木構造状に管理する．この際，長い間変更されていない検索条件（すなわち，先の仮定のもとでの重要な検索条件）ほど木構造の上位に現れ，子ノードの数が多くなるよう木構造の管理を行う（**履歴の構造モデル**）．具体的には，以下の制約に従うようノードを管理する．

(1) 最近に入力された検索条件ほど木の下位に置く．
(2) その時点での検索条件を木の最も右側に置く．
(3) その時点での検索条件に，新たに追加するノードと同一属性のノードが存在した場合，そのノードの弟の位置に新たなノードを追加した後，上記の制約 (2)，(1) の順にノードの移動を行う．

制約 (3) は検索条件が上書きされたときのノード管理制約を表す．上記の制約に従うことで，長い間変更されていない検索条件ほど木構造の上位に現れ，子ノードの数が多くなる．このことから，履歴の構造モデル上で検索条件がどの位置にあるかや，子ノードがどれだけ存在するかを，その検索条件が重要と見なせるかの情報として利用できる．

図 3 の (a) は【地名：祇園】，【フードタイプ：和

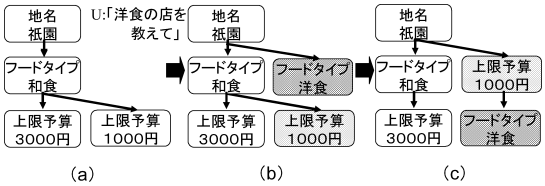

図 3 対話履歴の木構造上の管理
Fig. 3 Example of tree-structured dialogue history.

食】,【上限予算：3,000 円】,【上限予算：1,000 円】の順に入力が行われたときの履歴の構造モデルの状態を表しており，現在の検索条件は木の最も右側の子を順にたどることで得られる．図 3 の (a) の状態で「洋食の店を教えて」といった入力があると，【フードタイプ：和食】の弟ノードの位置にノード【フードタイプ：洋食】が生成される（図 3 (b)）．さらに，【上限予算：1,000 円】を現在の検索条件として木の最も右側に保持しておくためにノードの移動を行う．この際最近に入力されたものほど木の下位にくるという制約により，【上限予算：1,000 円】は【フードタイプ：洋食】よりも上位に移動される．この結果，履歴の構造は図 3 の (c) となる．この状態の場合，ユーザにとって【地名：祇園】という検索条件は，その他の検索条件よりも重要であると推定される．

4. 対話文脈を利用した言語理解

本章では，3 章で述べた対話モデルに基づく対話文脈的な制約を用いた言語理解処理について説明する．

まず，提案するシステムの全体像を**図 4** に示す．システムは大きく分けて音声認識部，言語理解部，対話管理部からなり，以下のように処理が行われる．

(1) 音声認識部によってユーザ発話が単語列に変換される．
(2) 言語理解部では音声認識部の出力と対話文脈的な制約から，ユーザの意図（対話行為と内容語）を推定する．具体的には，まず音声認識結果と想定発話パターンとの類似度計算を行う．その後決定木を用いて 各内容語ごとに 対話行為の信頼度や棄却すべき信頼度を求め，最後にそれらを統合して 発話ごとの 対話行為の決定や内容語の取捨選択を行う．
(3) 対話管理部では，言語理解部の出力を用いて対話の各状態を更新し，それに基づいてデータベースにアクセスし応答を行う．

従来の 1 発話のみから言語理解を行うシステム[17]は，決定木の部分がなく，想定発話パターンとの類似度計算の結果をそのまま言語理解結果として利用するものに相当する．

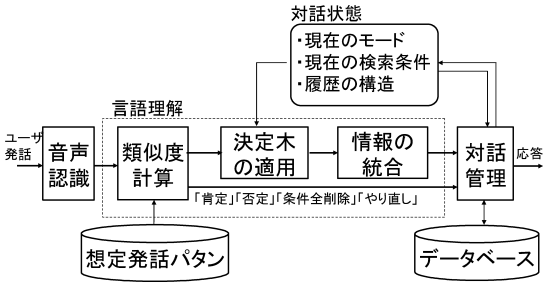

図 4 データベース検索システムの処理の概略
Fig. 4 Overview of our system for database search task.

本研究では，内容語を関係データベース中の属性名と値とした．また，対話行為は以下の 7 種類を設定した．まず，データベースに検索条件を指定するための対話行為として，「検索条件の追加」および「検索条件の削除」を，また特定のエントリに関する情報を取得するための対話行為として「情報の提示要求」を設定した．これに加えて「肯定」「否定」および「条件全削除」「やり直し」を設定した．以下では言語理解部での処理を詳細に述べる．

4.1 想定発話パターンとの類似度計算

各対話行為ごとに複数用意した想定発話パターンと音声認識結果の類似度を計算する．類似度の計算は文献 17) に従った☆．これは，音声認識結果が文としてどれだけその対話行為らしいかを表す☆☆．レストランドメインにおける想定発話パターンには「*FOODTYPE* がおいしい店を教えてください」といったものを 590 文用意した．なお想定発話パターン中の *FOODTYPE* はレストランデータベース中のフードタイプ属性の値に対応している．

ここで，「肯定」「否定」「条件全削除」「やり直し」は内容語をともなわない．これらの想定発話パターンとの類似度が最も高かった場合，4.2 節以降の処理は不要であるため，それを発話の対話行為として決定する．それ以外の場合，さらに以下の処理を行って対話行為の決定と内容語の取捨選択を行う．なお，本節で得られた類似度計算の結果は以下で利用する決定木の特徴量としても用いる．

4.2 各内容語に対する決定木の適用

決定木を用いて，各内容語 w_i の対話行為の信頼度

☆ ただし，内容語 w_i に対する重みとして，(構文上の重み $weight_i$) × (音声認識器が出力する信頼度 cm_i [10]) を用いた．
☆☆ 言語的な情報から各対話行為とその尤度を出力する手法であれば，ベクトル空間モデルに基づく類似度計算[3] などを利用することもできる．

S1：類似度計算で最尤となった対話行為
S2：類似度計算で最尤となった対話行為の尤度
S3：類似度計算で尤度が 2 番目の対話行為の尤度
S4：(S2) / (S3)
S5：内容語の種類（属性，値，キー属性，キー属性値）
S6：単語信頼度
S7：対になる内容語の存在の有無（「上限予算」と「1000 円」など）

図 5 1 発話のみから得られる特徴（7 個）
Fig. 5 Features obtained from single utterance.

G1：ユーザの前発話での対話行為
G2：直前のシステム発話が質問かどうか
G3：当該内容語が確認されたことがあるか
G4：当該内容語が否定されたことがあるか
G5：当該内容語が削除されたことがあるか

図 6 一般的に用いられる文脈的特徴（5 個）
Fig. 6 Contextual features used generally.

$CF(s|F_i, w_i)$ および内容語 w_i を棄却すべき信頼度 $CF(reject|F_i, w_i)$ を得る．ここで，F_i は各内容語ごとに得られる特徴量，$s \in$ {“検索条件の追加”，“検索条件の削除”，“情報の提示要求”} である☆．

特徴量 F_i には，各内容語ごとに得られる 33 種類を利用する．ここでは，**1 発話のみから得られる特徴（図 5）**や**一般的に用いられる文脈的特徴（図 6）**に加えて，**データベース検索タスクのモデルに基づく文脈的特徴（図 7）**を新たに定義し，使用する．1 発話のみから得られる特徴としては，想定発話パターンとの類似度計算の結果（S1〜S4）や単語信頼度[16]（S6）などを用いる．なお S2，S3 で類似度が最も高い対話行為が複数存在した場合，「曖昧」というラベルを付与しておく．一般的に用いられる文脈的特徴としては，直前のユーザ発話の対話行為（G1）や，各内容語が否定されたことがあるか（G4）などを用いる．また，対話の進行モデルに基づく特徴として，現在の発話がどちらのモードで行われているか（C1）や，現在の検索条件で絞り込まれたエントリ数（C2）などを設定する．ここで，C7 は現在の検索条件に合致するエントリのうち，情報の提示要求モードに入ってから言及されたものの割合である．また，C8 は現在の検索条件に合致するエントリのうち，現在までに言及されたことのあるものの割合を表す．さらに，履歴の構造モデルから変更対象ノードの深さ（C14），子ノードの数（C16）などを設定する．なお，C18，C19，C20，C21 は変更対象ノードの深さを現在の木の深さとの差および比で正規化したものである．

☆ このとき CF は，w_i が決定木のある葉に分類されたときに，その葉の持つ要素数を N，そのうち s（または reject）であるものの要素数を M，分類ラベルの数を P として，$(M+1)/(N+P)$ で与えられる[2]．

C1：現在のモード（初期状態：検索条件の指定）
C2：現在の検索条件に合致するエントリ数
C3：現在の検索条件に合致するエントリで，そのキー属性が「情報の提示要求」モードに入ってから言及されたものの数
C4：対話開始時から現在までに発話されたキー属性数
C5：現在の検索条件に合致するエントリで，そのキー属性が現在までの対話で言及されたものの数
C6：「情報の提示要求」モードに入ってから言及されたキー属性が存在するか
C7：(C3) / (C2)
C8：(C5) / (C2)
C9：当該内容語がキー属性の場合，すでに言及されているか
C10：当該内容語がキー属性の場合，現在の検索条件と合致するか
C11：現在の検索条件に合致するエントリ数が 0 か
C12：現在の検索条件に合致するエントリ数が 1 か
C13：履歴の構造モデルにおける現在の木の深さ
C14：履歴の構造モデルで，当該内容語が上書きする対象のノードの深さ（同一属性のノードが存在しない場合，現在の木の深さ+1）
C15：履歴の構造モデルで，当該内容語と同一属性を持つノードの深さの平均（同一属性のノードが存在しない場合，木の深さの平均+1）
C16：履歴の構造モデルで，当該内容語が上書きする対象のノードの子の数
C17：当該内容語が属性・値の場合，
現在の検索条件に同一のものがあるか
C18：(C14) / (木の深さ+1)
C19：(現在の木の深さ+1) − (C14)
C20：(C15) / (木の深さ+1)
C21：(木の深さ+1) − (C15)

図 7 データベース検索タスクのモデルに基づく文脈的な特徴（21 個）
Fig. 7 Features based on proposed models.

この決定木の学習は，正解として各内容語に対話行為を付与しておいた対話データを用いて行う．また音声認識誤りである内容語に対しては「棄却」というラベルを与えておく．これらのラベルに基づき決定木の学習を行うことで，出力として各内容語ごとに対話行為の信頼度と棄却すべき信頼度を得る．

4.3 対話行為の決定と内容語の取捨

発話中の各内容語に対する決定木の出力を統合して，発話としての対話行為の決定と，各内容語の取捨選択を行う．具体的には以下の手順に従う．

(1) 決定木により得られた対話行為の信頼度 $CF(s|F_i, w_i)$ を，対話行為ごとにすべての内容語に関して足し合わせ，その総和が最大となる対話行為を発話の対話行為 S とする．具体

的には次式で与える．

$$S = \arg\max_{s} \sum_{i} CF(s|F_i, w_i)$$

(2) 各内容語 w_i を棄却すべき信頼度 $CF(reject|F_i, w_i)$（以下これを R_i とする）を用い，以下により内容語の取捨選択を行う．

- $CF(S|F_i, w_i) \geq R_i$ ならば w_i を受理
- それ以外の場合，棄却

ただし，当該内容語を棄却すべき信頼度が低く（$R_i < \alpha$），かつ $CF(S|F_i, w_i) \neq 0$ の場合は，当該内容語を受理するか否かをユーザに確認するシステム発話を行い，本来受理すべき内容語の誤棄却を抑える．実験時 $\alpha = 0.9$ とした．

5. 評価実験

5.1 評価用データの収集

提案手法を評価するために，レストランデータベースを検索する音声対話システムを実装した．データベースのキー属性は「店名」，キーでない属性は「フードタイプ」「住所」など図 1 に示す 11 種類である．また，データベースのエントリ数は 1,217 である．

音声認識エンジンは Julius☆を用いた．言語モデルとして，想定発話パターンから作成した言語モデル（語彙サイズ 2,185）と，大量のテキストコーパスから学習した一般的な言語モデル（グルメレシピドメイン；語彙サイズ 19,447[12]）を言語モデル融合ツール[12]を用いて混合した．混合比は 9:1 とした．得られた言語モデルの語彙サイズは 21,565 である．

提案する言語理解処理は，事前に学習した決定木を必要とする．この学習データには，本研究室の学生 6 名による予備実験での対話データ（内容語数：748）を使用した．決定木学習には C5.0[2] を利用した．

システムからの応答はコンソール上の文字出力と音声出力の両方で行った．検索した結果が 8 件より多い場合は検索された件数だけを示し，8 件以下の場合は検索された件数と検索された店名を出力した．

上記のシステムを用いて，音声対話システムを利用したことのない 20 名から対話データを収集した．ユーザはまず，システムの説明とシステムが理解できる発話例を読み，音声入力のタイミングに慣れるため 5 分ほど練習を行う．その後「和食が食べたくなりましたが，あいにく VISA カードしか持っていません.」といった状況シナリオに基づき対話を行う．同様の条件で 3 対話を行った後，ユーザが自由に状況を設定したうえで，さらに 1 対話を行った．どの対話においても，ユーザが自分の判断で満足できる店が見つかった時点で対話を終了してもらった．

5.2 評価用データの詳細

実験により得られた発話は総計 3,015 発話（151 発話/人，38 発話/対話），認識された内容語は 2,803 内容語であった．また，全体の単語正解精度は 78.9%であった．得られたデータの対話行為ごとの発話数と内容語数を**表 2** に示す．このうち「検索条件の追加」と「検索条件の削除」が同時に発話されたものが 8 発話存在し，表中ではいずれも 1 発話として二重に計数している．またその他に分類された発話のうち，「肯定」「否定」「条件全削除」「やり直し」の発話が計 342 発話，タスク外発話が 161 発話であった．この表より，「検索条件の追加」と「情報の提示要求」の発話がほぼ同数で，かつどちらも内容語に約 20%の音声認識誤りを含んでいることが分かる．これらから正確な言語理解を行うには音声認識誤りの棄却や意味曖昧性の解消が必須である．

すべての評価用データにより，すべての特徴を用いて学習した決定木の一部を**図 8** に示す．ここでは，対話の進行モデルに基づく特徴である「現在のモード」や，履歴の構造モデルに基づく文脈的な特徴（木構造

表 2 対話行為ごとの発話数と内容語数
Table 2 Number of utterances and content words per dialogue act.

発話・内容語の内訳	検索条件の追加	情報の提示要求	検索条件の削除	その他	計
発話数	1,220	1,013	279	503	3,015
実際に発話された内容語数	1,388	1,253	307	-	2,948
正しく認識された内容語数	1,133	1,037	244	-	2,414
認識された内容語数	1,279	1,177	287	60	2,803

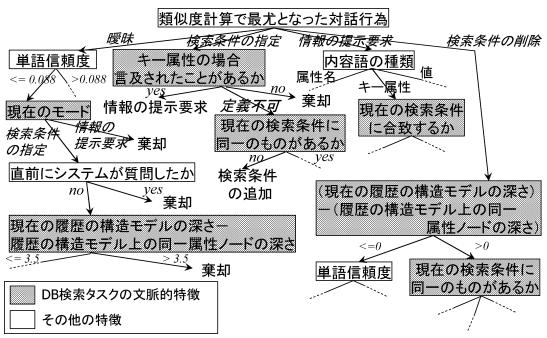

図 8 生成された決定木の一部
Fig. 8 Part of obtained decision tree.

☆ http://julius.sourceforge.jp/

上での同一属性の深さなど）が上位に現れている．一例として，ユーザの「“居酒屋無門” の住所をお願いします」という発話の音声認識結果が「居酒屋を，もんの住所をお願いします」であった場合を取り上げる．ここで “居酒屋無門” は固有名詞で店名である．音声認識結果に対する想定発話パターンとの類似度計算では，「検索条件の追加」と「情報の提示要求」が同値となって一意に対話行為が定まらず，「曖昧」という結果が得られた．ここで，音声認識結果からは，データベースの値に相当する「居酒屋」と属性名に相当する「住所」という 2 つの内容語が抽出される☆．内容語「居酒屋」は音声認識誤りであり，棄却されるべき内容語であるが，単語信頼度は比較的高く 0.65 であった．システムの現在のモードは「検索条件の指定」で，直前のシステム発話はユーザへの質問ではなかった．さらに内容語「居酒屋」と同一属性のノードが対話の早い時点で入力されていたため，「現在の履歴の構造モデルの深さ–履歴の構造モデル上の同一属性ノードの深さ」が 4 と大きかった．このとき，内容語「居酒屋」に着目して図 8 の決定木をたどると，棄却という判別結果が得られる．同様にして内容語「住所」に着目して決定木をたどると「情報の提示要求」という判別結果が得られた．2 つの内容語「居酒屋」と「住所」に対する判別結果を総合することで，ユーザ発話の対話行為は「情報の提示要求」であり，内容語は「住所」のみである，という言語理解結果が得られる．この言語理解結果は，店名は音声認識誤りにより得られていないものの，ある店の住所に関して情報の提示要求をユーザが行った，という正しい解釈を表す．これより，「どの店の住所ですか？」といった適切な応答を行うことが可能となる．

5.3 言語理解精度の評価

提案手法の評価のために，以下の 3 種類の手法を試行した．

手法 1：音声認識結果との類似度が最も高かった想定発話パターンの対話行為を，その発話および発話に含まれる内容語の対話行為とする．また，各内容語ごとにその単語信頼度が閾値未満ならば棄却する（ベースライン）．

手法 2：データベース検索タスクのモデルに基づく文脈的な特徴（図 7）を用いずに学習した決定木を用いる．対話行為推定と内容語の受理/棄却は提案手法に従う．

☆ 内容語の抽出は，データベースの値と属性名をもとに，あらかじめ用意しておいた内容語リストにあるものを取り出す．

表 3 レストラン検索システムにおける内容語ごとの意味理解精度（F 値）
Table 3 Classification accuracy of language understanding in restaurant system.

	手法 1	手法 2	手法 3
検索条件の追加	0.926	0.907	0.903
情報の提示要求	0.945	0.953	0.949
検索条件の削除	0.815	0.730	0.857
棄却	0.100	0.368	0.550
計	0.809	0.834	0.867

表 4 言語理解部による内容語ごとの意味理解誤り数
Table 4 Number of language understanding errors in restaurant system.

	手法 1	手法 2	手法 3
意味理解誤り数	536	465	373

総内容語数 2,803 語

手法 3：提案手法に従い，すべての特徴を用いて，対話行為推定と内容語の受理/棄却を行う．

言語理解部での内容語ごとの意味理解の精度と誤りの総数を**表 3** と**表 4** に示す．内容語の棄却判定を含む意味理解の精度としては再現率と適合率がともに重要であるため，ここでは評価尺度として F 値☆☆を用いた．手法 1 では，棄却の閾値を 20 通り試行し，意味理解誤り数が最小となった 0.05 を選択した．また，想定発話パターンとの類似度計算では一意に対話行為が定まらなかった内容語の 78.2%が「検索条件の追加」だったため，その場合はすべて「検索条件の追加」と判定し計数している．手法 2，3 では，10-fold クロスバリデーション☆☆☆によって評価を行った．また，決定木の枝刈りのパラメータを 20 通り試行し，意味理解誤り数が最小となる値を選択している．

手法 2 で生成された決定木の上位には，「類似度計算で最尤となった対話行為」「単語信頼度」が現れており，手法 1 とほぼ同じ情報から判別を行っていることが分かる．手法 2 ではこれらに加えて，「類似度が 2 番目のものとの比較」や「各内容語が否定されたことがあるか」といった特徴が導入され，棄却の精度が大幅に向上している．「検索条件の追加」や「検索条件の削除」では誤って棄却と判定されたものが増加したために F 値の低下が見られるが，全体として F 値が 2.5%改善している．

手法 3 では，手法 2 と比べて意味理解誤りがさらに 92 語減少しており，誤り数は 19.8%（= 92/465）

☆☆ F 値 = (2*再現率*適合率) / (再現率+適合率) である．

☆☆☆ 収集した内容語のうち 18 人分を学習用データ，残りの 2 人分を評価用データとして学習・評価を行い，これを 10 通り繰り返した．

削減された．F 値は 3.3 ポイント上昇した．ここでは文脈的特徴を導入することにより棄却に関して最も大きな改善が見られ，手法 2 と比較して F 値で 18.2 ポイントの改善を得た．手法 3 で得られた決定木では，最上位に「類似度計算で最尤となった対話行為」が現れた点では手法 2 の決定木と同様であったが，そのほかにデータベース検索タスクの文脈的特徴が多く現れた．検索条件の値にあたる内容語に対しては，「現在のモード」や履歴の構造モデルに基づく特徴（木構造上での同一属性ノードの深さなど）が上位に現れていた．それ以外の属性名やキー属性の値（店名）に対しては，「現在の検索条件に合致するエントリ数」や「キー属性ならば現在の検索条件に合致するか」「キー属性ならば言及されたことがあるか」などが特徴として有効であった．

ベースラインと比較して提案手法では 163 語の意味理解誤り削減を行えたが，いまだに意味理解誤りを起こすものが 373 語存在している．これらのうち最も多かった誤りは音声認識誤り（「棄却」が正解）を「検索条件の追加」と判定してしまうものであり，133 語存在した．このうち，「四条」が「七条」と認識されるといったような同属性内の誤りが 31 語存在した．これらは提案手法の枠組みではまったく同一の属性を持つため，取り除くことができない．また，対話の初期では対話や文脈の情報が十分得られない．これに起因すると見られる誤りは 18 語見られた．

また，いずれの手法でも「類似度計算で最尤となった対話行為」が決定木の最上位にあり，最も有効な特徴量であった．ここで，本手法が類似度計算の精度から受ける影響を検証するため，類似度計算で用いる想定発話パターンの数と手法 1〜3 の言語理解誤り数の関連を評価した（**図 9**）．想定発話パターンはランダムに選択し，選択された想定発話パターン集合ごとに決定木を再学習した．手法 1 では，想定発話パターン数の減少にともない言語理解誤りが大きく増加しており，想定発話パターンの量が少ないほど類似度計算の精度が悪化することが分かる．手法 2，3 においても想定発話パターン数の減少にともない言語理解誤りが増加するが，手法 1 に比べて影響は小さい．このことから，手法 2，3 では類似度計算の精度が低い場合でも，類似度計算以外の特徴量が有効に情報を補完していることが分かる．

5.4 本手法の他ドメインへの適用

本手法は対象とするデータベースに依存せず適用できる．このことを検証するため，新たにホテルデータベース検索システムを作成した．データベースのエントリ数は 2,004 で，キー属性はホテル名，その他の属性は 7 種類である．このうちホテルタイプ，住所，上限予算，下限予算の 4 属性はレストランデータベースと類似しているが，その他の属性（部屋数，付帯施設，周辺レジャー）は異なる．音声認識部の言語モデルにはレストランシステムと同様に n-gram モデルを用い，語彙サイズは 6,953 語であった．

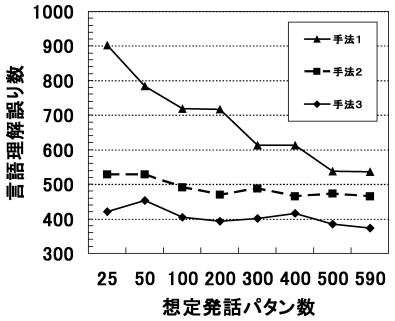

図 9 想定発話パターン数と各手法の精度
Fig. 9 Number of template sentences and classification accuracy.

以上のシステムを用いて，10 人の被験者からデータの収集を行った．データの収集手順は 5.1 節と同様である．この際には，レストランデータベースのデータから学習した決定木を利用した．この結果，1,426 の内容語を含む 1,271 発話が得られた．単語正解精度は 83.2%であった．

これらのデータを用いて，言語理解の精度を検証した．比較手法として，前節の手法 1，2，3 に加え，新たに以下の手法 4 を導入した．

- **手法 4：**学習データにはレストラン検索システムで収集した 2,803 内容語を用い，評価データとしてホテル検索システムで収集した 1,426 内容語を用いる．提案手法に従い，すべての特徴量を用いて対話行為推定と内容語の受理/棄却を行う．

手法 2，3 では，10-fold クロスバリデーション☆により評価結果を得た．手法 2，3 では決定木の枝刈りのパラメータを 20 通り試行し，最適なものを評価結果としている．また手法 4 では，パラメータ調整を行わずデフォルトの値を用いた．以上の結果を**表 5** に示す．

まず，手法 2 と手法 3 を比較すると，ホテルドメインでも言語理解精度が向上していることが分かる．これは，対象とするデータベースによらず本手法によって言語理解精度の向上が得られることを示している．

☆ 得られたデータのうち 9 人分を学習データ，1 人分を評価データとして評価を行い，これを 10 回試行した．

表 5 ホテル検索システムにおける内容語ごとの意味理解精度（F 値）
Table 5 Classification accuracy of language understanding in hotel system.

	手法 1	手法 2	手法 3	手法 4
検索条件の追加	0.917	0.940	0.964	0.930
情報の提示要求	0.978	0.967	0.983	0.990
検索条件の削除	0.822	0.711	0.756	0.933
棄却	0.287	0.318	0.504	0.527
計	0.888	0.890	0.926	0.924

さらに手法 3 と手法 4 では言語理解精度がほぼ等しい．このことは本実験で行ったように，事前データとして他のドメインで収集したものが利用できることを示している．すなわち本研究で提案した対話文脈的な制約は，新たなデータ収集なしに他ドメインにも適用可能である．音声対話システムにおいては個々のドメインで対話データを十分に収集するための労力が大きいため，この点は非常に重要である．

6. ま と め

本稿では，関係データベースを検索するタスクにおいて，音声認識誤りの棄却や解釈の曖昧性の解消のために，対話文脈的な制約を取り入れる手法について述べた．まず，データベース検索タスクにおける対話のモデル化を行った．さらに，これらのモデルから得られる特徴量を用いて決定木を構成し，言語理解部に組み込むことで，対話文脈的な制約を言語理解結果に反映させた．

被験者による評価実験の結果，提案するモデルを利用することで言語理解精度が向上することを確認した．さらに，得られた決定木が，学習したドメインとは異なるドメインのデータベースにおいても同様の効果を発揮することを確認した．以上の結果から，本手法は関係データベースを検索するタスクではドメインによらず有効であることが示された．

謝辞 本研究の一部は，科学研究費補助金（基盤研究（A），特定領域「情報学」，若手研究（B）），21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」の支援を受けた．最後に，日頃より議論をしていただく奥乃研究室の皆様に感謝いたします．

参 考 文 献

1) Araki, M., Komatani, K., Hirata, T. and Doshita, S.: A Dialogue Library for Task-oriented Spoken Dialogue Systems, *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp.1–7 (1999).
2) *C5.0.* http://rulequest.com/index.html
3) Chu-Carroll, J. and Carpenter, B.: Vector-based Natural Language Call Routing, *Computational Linguistics*, Vol.25, No.3, pp.361–388 (1999).
4) Gabsdil, M. and Lemon, O.: Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems, *Proc. 42nd Meeting of the Association for Computational Linguistics* (*ACL'04*), pp.343–350 (2004).
5) Pradhan, S.S. and Ward, W.H.: Estimating Semantic Confidence for Spoken Dialogue Systems, *Proc. ICASSP*, Vol.1, pp.233–236 (2002).
6) Walker, M., Langkilde, I., Wright, J., Gorin, A. and Litman, D.: Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You?, *Proc. 1st Conf. on the North American Chapter of ACL* (*NAACL00*), pp.210–217 (2000).
7) Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T.J. and Hetherington, L.: JUPITER: A Telephone-Based Conversational Interface for Weather Information, *IEEE Trans. Speech and Audio Processing*, Vol.8, No.1, pp.85–96 (2000).
8) 音声ポケロケ．http://www.lang.astem.or.jp/bus/
9) 李　晃伸，山田真士，西村竜一，鹿野清宏：公共音声情報案内システム「たけまるくん」の運用および収集発話の分析，情報処理学会研究報告，2004-SLP53-9 (2004).
10) 李　晃伸，河原達也，鹿野清宏：2 パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法，情報処理学会研究報告，2003-SLP-49-48 (2003).
11) 中川聖一，堀部千寿：音響尤度と言語尤度を用いた音声認識結果の信頼度の算出，情報処理学会研究報告，2001-SLP-36-13 (2001).
12) 河原達也，住吉貴志，李　晃伸，坂野秀樹，武田一哉，三村正人，山田武志，西浦敬信，伊藤克亘，伊藤彰則，鹿野清宏：連続音声認識コンソーシアム 2001 年度版ソフトウェアの概要，情報処理学会研究報告，2002-SLP-43-3 (2002).
13) 入江友紀，松原茂樹，河口信夫，山口由紀子，稲垣康善：意図タグつきコーパスを用いた発話意図推定手法，人工知能学会研究会資料，SIG-SLUD-A301-03 (2003).
14) 由浅裕規，水野智士，伊藤敏彦，甲斐充彦，小西達裕，伊東幸宏：状況と文脈を利用した音声対話型車載インタフェースの構築と評価，情報処理学会研究報告，2003-SLP-49-34 (2003).
15) 東中竜一郎，中野幹生，相川清明：複数文脈を用いる音声対話システムにおける統計モデルに基

づく談話理解法，情報処理学会研究報告，2003-SLP-45-17 (2003).
16) 駒谷和範，河原達也：音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理，情報処理学会論文誌，Vol.43, No.10, pp.3078–3086 (2002).
17) 駒谷和範，河原達也，清田陽司，黒橋禎夫，Fung, P.：柔軟な言語モデルとマッチングを用いた音声によるレストラン検索システム，電子情報通信学会技術研究報告，SP2001-113 (2001).

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)

神田 直之（正会員）

2004 年京都大学工学部情報学科卒業．2006 年同大学院情報学研究科知能情報学専攻修士課程修了．在学中は音声対話システムの研究に従事．現在，株式会社日立製作所勤務．

駒谷 和範（正会員）

1998 年京都大学工学部情報工学科卒業．2000 年同大学院情報学研究科知能情報学専攻修士課程修了．2002 年同大学院博士後期課程修了．京都大学博士（情報学）．同年より京都大学情報学研究科助手．情報処理学会平成 16 年度山下記念研究賞，FIT2002 ヤングリサーチャー賞受賞．電子情報通信学会，言語処理学会，人工知能学会，ACL 各会員．

尾形 哲也（正会員）

1993 年早稲田大学理工学部機械工学科卒業．日本学術振興会特別研究員，早稲田大学理工学部助手，理化学研究所脳科学総合研究センター研究員，京都大学大学院情報学研究科講師を経て，2005 年より同助教授．博士（工学）．この間，早稲田大学ヒューマノイド研究所客員助教授．人間とロボットのインタラクションと協調，神経回路モデル等の研究に従事．2000 年度日本機械学会論文賞，IEA/AIE-2005 最優秀論文賞等を受賞．RSJ，JSME，JSAI，IEEE 等会員．

奥乃 博（正会員）

1972 年東京大学教養学部基礎科学科卒業．日本電信電話公社，NTT，JST 北野共生システムプロジェクト，東京理科大学理工学部情報科学科を経て，2001 年 4 月より京都大学大学院情報学研究科知能情報学専攻教授．博士（工学）．この間，スタンフォード大学客員研究員，東京大学工学部客員助教授．人工知能，音環境理解，ロボット聴覚の研究に従事．1990 年度人工知能学会論文賞，IEA/AIE-2001，2005 最優秀論文賞，IEEE/RSJ IROS-2001 Best Paper Nomination Finalist，第 2 回船井情報科学振興賞等受賞．人工知能学会，日本ソフトウェア科学会，日本認知科学会，日本ロボット学会，ACM，AAAI，IEEE，ASA，ISCA 各会員．本学会英文図書出版委員．人工知能学会，日本ロボット学会評議員．『インターネット活用術』（岩波書店），『*Computational Auditory Scene Analysis*』（共編，LEA），『*Advanced Lisp Technology*』（共編，Taylor & Francis）ほか．