Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

データベース検索音声対話システムにおける対話状態の推定

西村良太[†1]　駒谷和範[†1]

データベース検索を行う音声対話システムにおいて，ユーザの意図を反映し，音声認識誤りに対処して応答生成を行うための対話状態を推定するモデルについて述べる．データベース検索タスクにおいて，対話の状態が「検索条件の指定」「情報の提示要求」の二つを遷移するとモデル化する．この 2 つの状態を対話中から得られる素性に基づき，ロジスティック回帰により予測する．レストランデータベース検索を行う音声対話システムを構築し，7 名の被験者から対話データを収集し，モデルの学習実験を行った．ベースラインシステムでの対話状態の決定精度が 87.1%であるのに対して，学習されたモデルでは，オープンテストで 97.4%であった．また，モデルに用いた素性のうち，どの素性が対話状態の推定に寄与しているかの確認も行った．

Estimation of Dialogue States in Database Search Spoken Dialogue System

Ryota Nishimura[†1] and Kazunori Komatani[†1]

We describe an estimation model of dialogue states in spoken dialogue systems for the database search task. We model dialogues in the database search task as consisting of two states: “specifying retrieval conditions (search)” and “requesting detailed information about specific entries (info.)”. The two states are predicted by a logistic regression classifier based on features obtained from the dialog. We developed a spoken dialogue system for the restaurant database search task and collected dialogue data from seven participants. The experimental result showed that the estimation accuracy was 97.4%. We investigated which features contributed to the estimation of the states.

†1 名古屋大学大学院工学研究科電子情報システム専攻
Department of Electrical Engineering and Computer Science, Graduate School of Engineering, Nagoya University, Japan.

1. はじめに

音声対話システムでは，状況に応じた適切な応答が必要である．これに関しては，ユーザ入力の音声認識結果のみから状況の理解や応答生成を行うもの[1] や，最近では，POMDP により管理される確率的な状態に基づいて応答を生成するものもある[2),3)]．また，現在利用可能な音声認識器では，対話中の文末表現の認識が難しいことから，音声対話システムでは，内容語に重点を置いて入力文を解析し，応答生成を行っている．しかし，文末表現ではない内容語の部分であっても音声認識結果には認識誤りがつきものであり，内容語の脱落や，不必要な内容語の湧き出しによって，ユーザの意図とは異なるシステム挙動や応答を生成してしまう場合がある．これに対して，認識誤りに対応できるように，様々な場合を想定して，人手で細かなルールを作り込むことは非現実的である．

これらの問題に対処するために，本研究では，データベース検索タスク[⋆1] における対話状態の推定に取り組む．対話状態として，データベース検索タスクでの対話が，検索条件の指定と情報の提示要求の 2 状態からなるというモデル[5)] を本研究でも採用する．このモデルは，対話状態を 2 つだけとすることで，対象とするデータベースの内容（ドメイン）に依存しない，対話の概略をモデル化している[5)]．その一方で，これらの状態を推定することで，後述するように，音声認識誤りの誤受理防止やその状態に応じたプロンプト生成が可能となる．

本研究では，この対話状態の推定に，機械学習，具体的にはロジスティック回帰を用いる．この際の素性として，神田ら[5)] が用いていた，直前のシステム応答とユーザ発話から得られる情報に加えて，より長期的な文脈を考慮した情報を用いる．具体的には，対話状態が同じ状態に留まっていた回数を用いる．この情報が必要な理由としては，対話には流れがあり，人間は対話全体での対話状態の変遷を考慮に入れて対話を行っており，これまでにたどってきた対話状態の遷移によって，現在の対話状態が決まると考えられるためである．さらに，この状態推定結果を，言語理解性能の向上，つまり発話の受理棄却性能の向上に用いるだけでなく，その後の対話管理や応答生成に使う方法についても検討する．

⋆1 このデータベース検索タスクは，河原らの分類[4)] における抽象化タスクのひとつであり，システムからユーザ，ユーザからシステムと双方向に情報が流れるタスクとされている．このため，スロットフィリングタスクのように，質問項目や順番を固定して対話を行うことはできない．

Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

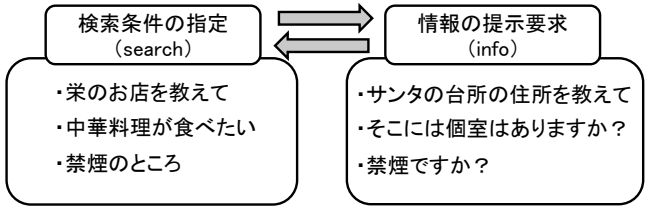

図 1 データベース検索対話における 2 つの対話状態[5)]

2. データベース検索タスクにおける対話状態推定

2.1 データベース検索タスクにおける対話状態

データベース検索対話においては，図 1 に示すように，対話状態は「検索条件の指定 (search)」「情報の提示要求 (info)」の二つから成るとする[5)]．検索条件の指定とは，データベースのエントリを検索するための条件を指定し，候補を絞り込む段階の状態である．ユーザの入力例としては「栄にある中華料理のお店を教えて」などがあり，この場合には，地域「栄」，ジャンル「中華料理」が検索の条件となる．情報の提示要求とは，特定のエントリについての情報を聞き出したい場合の対話である．ユーザの入力例としては「サンタの台所（店名）の住所を教えて」などがあり，この場合には，このお店の情報として，住所が出力される．便宜上，ここではデータベースの内容をレストラン情報に絞っているが，ここでの内容はデータベース検索対話一般に成り立つ．

本研究で用いるデータベースは，表 1 に示すような関係データベースであり，各エントリは，属性とその値の組を持ち，キー属性を 1 つ持っている．この属性名と値の両方を，システムでは内容語とする．

2.2 ロジスティック回帰による対話状態の推定

本研究では，前節で述べた 2 つの対話状態を，機械学習の一手法であるロジスティック回帰に基づき推定する．モデルが無い音声対話システムでは，このような挙動を実現するために，少数のルールを記述して推定する．本研究では，音声認識結果以外に，システムの内部情報も用いて対話状態推定を行う．また，この推定結果から，直接 search と info を決定するのではなく，2 種類の対話状態のうちどちらの状態でありそうかというスコアを出力し，この結果と音声認識結果の信頼度，応答文のスコアなどを統合し，幾つかの出力候補の中から最適なものを選択するなどの利用法もある．他にも，2 種類の対話状態のスコアの差を見て，それに応じて応答文を変更するなども考えられる．

表 1 データベースの項目

属性	値
店名（キー属性）	びすとろさんたのだいどころ（サンタの台所）
住所	愛知県名古屋市西区名駅 3-11-2 サンタウン名駅 2F
地域	名古屋駅前
最寄り駅	近鉄名古屋
ジャンル	居酒屋
食べ物	洋風創作料理
予算	3001 円〜4000 円
座席数	93
説明	一年中クリスマス気分が楽しめるお店
アクセス	名古屋駅より徒歩 7 分 ユニモール 4 番出口〜北へ直進★
飲み放題	あり
食べ放題	なし
個室	あり
クレジットカード	利用可
禁煙	禁煙席無し
貸切	貸切可
ランチ	なし

音声認識結果の認識誤りへの対応としては，次のようなものが考えられる．挿入誤り・置換誤りによって店名が湧き出した場合，モデルが無いシステム，特に対話状態を推定しない場合には，湧き出した店名がそのまま受理され，対話状態が info に遷移してしまう．モデルが有るシステムにて，対話状態の推定結果が正しければ（search であれば），店名が湧き出しであると判断でき，棄却が可能になる．脱落誤りに対しては，logistic 回帰モデルの出力値が確率値であることを利用して，その値によってシステムからのプロンプトを変更できる．例えば，info，search のどちらになるか曖昧な場合には，システムプロンプトを「もう一度言って下さい」とし，曖昧でない場合には，「(info の場合）どのような情報が知りたいのですか？」，「(search の場合）検索条件を指定してください.」として対応することができる．

2.3 素　　性

対話状態推定モデルに用いる素性には，以下に示す対話中に得られる情報を用いる．

(1)　直前の対話状態 [1 素性]

Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

(2) 対話状態が search に留まったターン数，info に留まったターン数 [2 素性]
(3) 直前のシステム応答にて検索結果のリストを提示したかどうかのフラグ [1 素性]
(4) 検索件数フラグ（0 件，1 件，2〜4 件，5 件以上）[4 素性]
(5) スロットが埋まっているかどうかのフラグ [14 素性]
(6) スロットが書き換わったかどうかのフラグ [14 素性]

これらの素性を採用した理由は，素性（1）は，同じ対話状態が連続して起こりやすいと考えられるためであり，素性（2）は，素性（1）と同じ理由である．素性（3）は，システムから検索結果（店名リスト）を提示した直後には，ユーザは店名を発話し，info に移行することが多いためである．素性（4）は，検索件数が絞り込まれれば，info に移行するという傾向を表す．検索件数によって分けているのは，それぞれでシステムの挙動が異なるためである．システムの挙動については，3.4 節にて後述する．素性（5），(6）でのスロットとは，対話中に得られた内容語を保持するために，各属性に対して 1 つずつ用意されている．スロットが用意されている内容語は，表 1 に示した 17 項目のうち，住所，説明，アクセス以外の 14 項目である．3 つの項目を含めないのは，これらが，info だけに対する内容語であり，また店名のように，対話状態の遷移に重要な役割をはたしてはおらず，有効な素性ではないと考えられるためである．スロットは，同じ属性の内容語が入力された場合には上書きされ，対話中に埋まる，書き換わるということが起こる．このスロットにて，ユーザから得られている情報が確認できる．特定の対話状態によく現れる内容語（属性）がある可能性があるため，素性に採用した．

これらの素性がとる値は，素性（1）は，1 を search，2 を info とする．素性（2）は，例えば search に 3 ターン留まっていれば，search に留まったターン数は 3 とする．この時，info に留まったターン数は 0 とする．その他の素性は，条件に合う場合には 1，合わない場合には 0 とする．

3. データ収集に用いた音声対話システム

音声対話システムの概略図を，図 2 に示す．このシステムを用いて，対話ログを収集し，モデルの構築を行う．

3.1 音声認識部

音声認識エンジンには，Julius[⋆1]を用いた．言語モデルには，クラス n-gram モデルを使用した．言語モデルの学習には，Yahoo!知恵袋の料理・グルメ・レシピカテゴリから得られたテキスト（441,872 文）と，想定発話言語モデル（CFG 文法）から生成したテキスト（8,477 文）を用いた．検索用データベースに含まれる店名 (<shop>)，地域 (<location>)，最寄り駅 (<station>)，ジャンル (<genre>)，食べ物 (<food>) をクラス単語とし，Yahoo!知恵袋テキスト中に含まれるクラス単語を，以下のようにクラスラベルに置き換えて，言語モデルの学習を行った．

例）吉野家に行こう → <shop> に行こう

そして，これらのクラスに対応するデータベース内の全単語（例での「吉野家」）を，音声認識用の単語辞書に追加した．最終的な語彙サイズは，39928 単語であり，各クラス内の語彙サイズは表 2 に示されている．

⋆1 http://julius.sourceforge.jp/

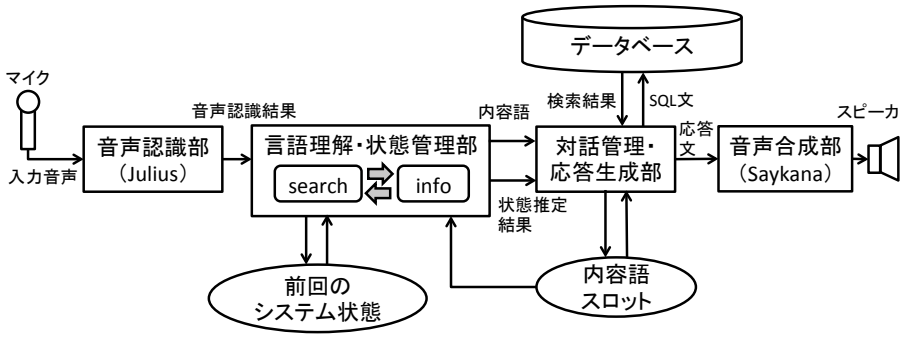

図 2 音声対話システムの概略図

表 2 言語モデルの語彙サイズ

種類	語彙サイズ
店名	2399
地域	19
最寄り駅	229
ジャンル	22
食べ物	82
その他	37177
合計	39928

3.2 言語理解・状態管理部

このシステムは，対話データ収集のためのベースラインシステムであるため，対話状態は

Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

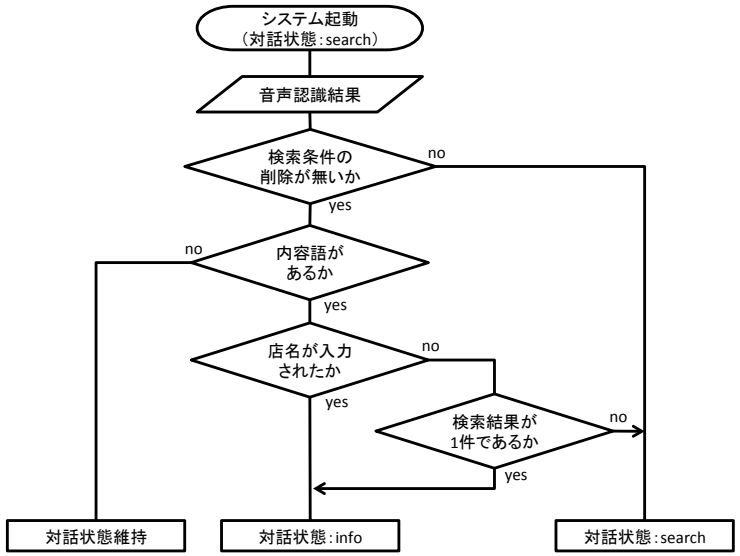

図 3 ベースラインシステムでの状態決定のフローチャート

音声認識結果のみを用いて決定する．将来的にはこの部分で，2 章で述べた，ロジスティック回帰を用いた対話状態の推定を行う．

内部状態は 2.1 節で述べたのと同様,「search（検索条件の指定）」「info（情報の提示要求）」とし，ベースラインシステムでは，図 3 のフローチャートに従って決定される．対話状態の初期値は search とする．ユーザによって店名が入力されるか，もしくは検索候補が 1 件に絞りこまれると，店名を決定し，info に遷移する．検索条件の削除が行われた場合は，search に遷移する．info であっても，検索条件の削除を行うことができ，これはユーザが残った検索条件によってお店を検索することを目的としていることから，対話状態は search である．ユーザ入力に内容語が含まれていない場合には．前回の対話状態を維持する．

3.3 データベース

レストラン検索データベースには，愛知県にある 2,398 件分のデータが含まれている．店名は，かな表記を用いている．データベースソフトウェアには，H2 Database[⋆1] を用い，データベースの検索には，PostgreSQL プロトコルを用いており，SQL 文にて検索を行う．システムで用いる属性の数は，表 1 に示した 17 種類である．

3.4 対話管理・応答生成部

対話管理・応答生成部では，内容語と対話状態推定結果を受け，必要であればデータベースの検索を行い応答を生成する．

対話状態が search の場合には，ユーザから入力された検索条件をスロットに保存する．すでに埋まっているスロットに関する内容語が入力された場合には，新しい値にてスロットを上書きする．これらのスロットの値を用いてデータベースの検索を行う．例として，地域スロットに「大須」，ジャンルスロットに「和食」が入っている場合に生成される SQL 文を以下に示す．

SELECT * FROM DATABASE WHERE ”small_area:name” LIKE ’%大須%’ and ”genre:name” LIKE ’%和食%’

（データベース内では，属性名は英語で表記されており，地域は small_area:name，ジャンルは genre:name となっている．）

この検索結果の件数に応じて，以下のような応答を返す．

- 5 件以上：[例] 5 件のお店が見つかりました．他の検索条件を指定してください．
- 1 件〜4 件：[例] 2 件のお店が見つかりました．びすとろさんたのだいどころと，しゅんはなびです．
- 0 件：お店は見つかりませんでした．検索条件を変更してください．

検索結果が 1 件であった場合には，その店名をスロットに保存する．

対話状態が info の場合には，特定のお店について，ユーザから入力された内容語に対する情報を出力する．例えば「住所を教えて」という入力に対しては,「住所は，愛知県名古屋市 … です.」と出力する．

ユーザ入力から内容語が得られなかった場合には，対話状態に対応したプロンプトを出力する．対話状態が「search」であれば,「どんな条件のお店を探しますか？」と出力する．対話状態が「info」であれば，スロットに店名がない場合には「どのお店について知りたいのですか？」と出力し，店名がある場合には「サンタの台所 (店名) について，何が知りたいですか？」と出力する．

音声合成には，Saykana (AquesTalk)[⋆2] を用いた．

⋆1 http://www.h2database.com/html/main.html

⋆2 http://www.a-quest.com/quickware/saykana/

Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

表 3 音声認識率（%）

単語正解率	単語正解精度	置換	削除	挿入	単語誤り率	文誤り率
78.3	71.5	5.5	6.2	6.8	28.5	55.0

表 4 ベースラインシステムによる分類結果の混同行列

		分類結果	
		search	info
実際のデータ数	search	389	72
	info	8	152

4. 実　　験

4.1 学習用データの収集

今回構築した音声対話システムを用いて，7 名の被験者（本研究室所属の学部生と院生）から対話データを収集した．対話を行う前に，まず被験者に，以下の項目について書かれた実験の説明書を読んでもらった．

- 対話ログ収集実験についての説明
- 音声対話システムで受理可能な入力例
- お店の絞り込みに使える条件

検索候補が 4 件以下になると，システムが店名リストを提示することも，ここで教示している．

実験中のシステムの出力は，音声出力の他に，画面に「応答文」,「現在の絞り込み条件（スロット値）」を表示させた．

被験者は 3 つの課題に沿って対話を行った．課題は「今日は一人で栄に来ています．ガッツリしたものが食べたい気分です．食べたいものを 2 つ考えて，それぞれについて，1 件お店を選んで，住所を確認してください.」のようなものである．提示した課題が完了できた時点で対話を終了した．

収集された対話は，全部で 21 対話（被験者 7 名，各 3 対話）あり，被験者による発話数の合計は 638 発話であった．

4.2 学習用データの音声認識率

収集した対話データに対して，書き起こしを行い，音声認識率を算出した．表 3 に結果を示す．これは，対話実験にて得られた，全 638 文の結果である．被験者 7 名の間の音声認識率は，単語正解率で 70.1%〜86.3%，単語正解精度で 60.3%〜80.3%であった．

4.3 学習用データへの正解タグ付け

収集した対話データに対して，想定される対話状態の正解タグを人手で付与した．付与した正解は，search，info の 2 種類である．ここで，ユーザの入力の中には，あいさつ（ありがとう．ありがとうございました），了解（わかりました），ひとりごとのような，search，info のどちらにも分類されない発話（4 発話）と，笑い，咳のみの発話（13 発話）が含まれていたため，これらは分析対象外として，学習用データから削除した．残った 621 発話分のデータにてロジスティック回帰関数の学習を行う．このデータの正解タグの数の分布は，search が 461，info が 160 であった．

ベースラインシステムにより分類を誤っていた箇所は，80 箇所であり，その内訳を，表 4 に示す.「(正) search → (誤) info」では，店名の湧き出しが起こり，誤って info に遷移してしまうものが多かった．その他の誤りとしては,「他のお店を教えて」や「○○を削除して」と入力され，条件を削除して search に遷移するべきところで，認識誤りをして，search に遷移できないものがあった．

「(正) info → (誤) search」では，店名が認識されず，対話状態が info に遷移できなかったものがほとんどであった．1 箇所，リストから店名を指定する時に，ユーザが「二つ目の店」と発話し，システムが対応していなかったため，店を指定できなかったというケースがあった．

4.4 ロジスティック回帰関数の学習

ロジスティック回帰関数の学習には，線形分類器学習ソフトウェアである LIBLINEAR[6] を用いた．素性は，2.3 節に挙げたものを用いている．LIBLINEAR で学習する際には，L1-regularized logistic regression を用い，収束基準の epsilon には，0.5 を用いた (デフォルト値は，0.01)．収束基準については，デフォルト値のまま学習を行うと，少数の素性にて過学習が起こるため，経験的に値を決定した．

4.5 実 験 結 果

収集した対話データを用いて学習した対話状態推定モデルの分類精度を表 5 に示す．クローズドは，全 621 発話分のデータにて学習し，同データにてテストを行った結果である．10-fold は，10 分割交差検定を行った結果である．ベースラインシステムは，対話収集に用いた対話システム，つまり図 3 の決定規則により出力された結果である．

ロジスティック回帰モデルの係数を，表 6 に示す．学習の結果，係数が 0 にならなかった（素性として効果のある）素性は，36 素性中 22 素性であった．係数が正の値の素性は

Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

表 5 モデルの学習結果

種類	分類精度 (%)
クローズド	97.6
10-fold	97.4
(ベースラインシステム)	87.1

表 6 ロジスティック回帰モデルの係数

素性	係数
直前の対話状態	0.074
search に留まったターン数	0.19
info に留まったターン数	-0.63
リストを提示したかどうかのフラグ	-1.51
検索件数フラグ（0 件）	0.37
検索件数フラグ（1 件）	-0.37
検索件数フラグ（2〜4 件）	-0.40
検索件数フラグ（5〜件）	1.48

search の決定に関するもの，係数が負の値の素性は info の決定に関するものである．検索件数フラグは，5 件以上の場合には，強く search になることが示されている．これは検索結果が 5 件以上の場合には検索結果のリストが提示されないためである．2〜4 件の場合には，検索結果のリストが提示されるため，info の傾向が強くなっている．

4.6 対話状態の推定により改善される例

今回学習したモデルにて，対話状態の推定がベースラインシステムから改善された例は，74 例であった．逆に，ベースラインシステムの結果が改悪された例は，7 例であった．改悪された例としては，ユーザが「二千円ぐらいで食べられるお店を教えて」と search の発話をしたが，info と推定した場合などである．このことから，対話状態推定モデルを用いることで対話状態の推定がより良く行えることが示されている．

今回用いたデータにて，対話状態の推定により有用となる例を以下に示す．

(1) 店名が湧き出して info になっていた場合

対話例

- ユーザ入力：「ランチがあるお店を教えて」
  （認識誤り：ばんちゃがる（店名）の店教えて）

(a) ベースラインシステム：「ばんちゃがるについて，何が知りたいんですか？」

(b) モデル有りシステム：「検索条件を入力してください.」

正しく search と推定できれば，店名が湧き出しであると判断でき，結果を棄却できる．

(2) 店名が脱落・置換し認識されずに info にならなかった場合

対話例

- ユーザ　：やきとりちゅうべぇ（店名）
  （認識誤り：しっとり中です）

(a) システム：検索条件を入力してください．

(b) システム：店名を教えてください．

正しく info であると推定できれば,「どのお店について聞きたいんですか？」など，対話状態に適したプロンプトを出力できる．

(3) 「全部削除」「○○を削除」「他のお店を教えて」などが認識されず，info に留まった場合

対話例

- ユーザ　：全部削除
  （認識誤り：全部卓上）

(a) システム：お店について，何が知りたいんですか？

(b) システム：検索条件を入力してください．

正しく search であると推定できれば，削除と言っていた事実は回復できないものの「検索条件を入力してください.」など，対話状態に適したプロンプトを出力できる．

5. ま と め

本稿では，データベース検索タスクの音声対話システムにおいて，音声認識誤りに対処して応答生成を行うための対話状態を推定するモデルについて述べ，対話状態の推定実験を行った．データベース検索音声対話システムを構築し，被験者から収集した対話データを用いて実験を行った結果，対話状態の推定精度は，ベースラインシステムによる決定精度を上回ることが確認された．また，この対話状態推定部をオンラインで動作するシステムに組み込むことで，店名の湧き出しの棄却や，適切なシステムプロンプトが出力できる見込みがあることが確認された．

今後は，音声対話システムを用いたことがない被験者から，より多くの対話データを収集し，そのデータを用いてモデルの学習を行う．また，今回構築したモデルを組み込んだ状態の対話システムを用いてデータを収集し，分析を行う．

Vol.2012-MUS-94 No.20
Vol.2012-SLP-90 No.20
2012/2/4

謝　　辞

本研究は JST 戦略的創造研究推進事業さきがけの支援を受けた．言語モデルの学習には，ヤフー株式会社が国立情報学研究所に提供した『Yahoo! 知恵袋データ』を利用した．

参 考 文 献

1) 西村竜一，西原洋平，鶴身玲典，李　晃伸，猿渡　洋，鹿野清宏：実環境研究プラットホームとしての音声情報案内システムの運用，電子情報通信学会論文誌，Vol.J87-D-II, No.3, pp.789–798 (2004).
2) Williams, J.D. and Young, S.: Partially observable markov decision processes for spoken dialog systems, *Computer Speech and Language*, Vol.21, No.2, pp.393–422 (2007).
3) 南　泰浩，森　啓，目黒豊美，東中竜一郎，堂坂浩二，前田英作：対話データの統計量を用いた POMDP による対話制御，情報処理学会研究報告. SLP， Vol.2009-SLP-79, No.15, pp.1–6 (2009).
4) 河原達也，荒木雅弘：知の科学　音声対話システム，オーム社 (2006).
5) 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対話文脈を利用した音声言語理解，情報処理学会論文誌， Vol.47, No.6, pp.1802–1811 (2006).
6) Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research*, Vol.9, pp.1871–1874 (2008).