音声認識・言語理解システムを用いた音声対話コーパスの収集と分析

Spoken Dialogues via Speech Recognition and Natural Language Understanding Systems: The First Report for the Corpus Collection and an Analysis

小野 正貴[*1] Masaki Ono　本間 健[*2] Takeshi Homma　神田 直之[*2] Naoyuki Kanda

永松 健司[*2] Kenji Nagamatsu　中野 有紀子[*1/*3] Yukiko Nakano

[*1] 東京農工大学 Tokyo University of Agriculture and Technology
[*2] (株) 日立製作所中央研究所 Central Research Laboratory, Hitachi, Ltd.
[*3] 成蹊大学 Seikei University

Spoken dialogue systems are useful and effective in accessing and retrieving information from large text archives like the Web. However, a bottleneck of spoken dialogue systems is speech recognition error. The system needs to select appropriate system responses by considering the speech recognition errors. Focusing on information retrieving conversations in art museums, this study reports a dialogue corpus collection experiment where one of the conversation participants, a guide, receives messages from her/his partner, a visitor, via a speech recognition system and a natural language understanding system. Then, we will analyze the corpus, and predict the types of response behaviors using a machine learning technique.

1. はじめに

膨大な情報がネットワーク上に存在する状況において，誰もが容易に情報アクセスできるユーザインタフェースとして，音声対話システムへの期待は大きい．例えば，インターネット上にあるテキストを検索する情報検索システムは，いくつかの適切なキーワードを入力することにより，ヒットした文書がリストアップされるが，音声を使ってインタラクティブに検索ができると望ましい．また，博物館や展示会場において，展示物についての情報を検索する端末と，音声ガイダンスサービスとを統合することにより，ユーザの情報取得要求に応じてガイダンスを行うことができる，よりインタラクティブなガイダンスシステムが実現可能となるだろう．

音声対話システムの先行研究において，神田ら [1] は，データベース検索時のユーザの行動を検索情報の指定と情報の提供要求に大別し，この情報を音声言語理解に利用する方式を提案している．さらに，翠ら [2] は，この知見を踏まえて，ユーザの検索や質問に応答するモードと，ユーザに有用であると思われる情報をシステム主導で提供するモードとを有する情報案内システムを実装している．

このような情報提供を目的とする音声対話システムでは，音声認識結果が誤っている可能性を考慮しながら，システムの応答を適切に選択することが重要となる [3]．そこで本研究では，音声認識や言語理解の結果が画面に表示される状況において，人間同士の対話データを収録し，確認や会話の修復等のデータを収集するとともに，これらを分析することにより，応答選択方法の知見を得ることを目的とする．音声認識，言語理解は機械が行い，応答選択のみを人間が行う状況を実験的に作り出すことにより，音声対話システムの実装に直接結びつく知見を人間の言語行動から学ぶことができると考えられる．

連絡先：小野正貴（50008646201@st.tuat.ac.jp），中野有紀子（y.nakano@st.seikei.ac.jp）
〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1
成蹊大学理工学部情報科学科

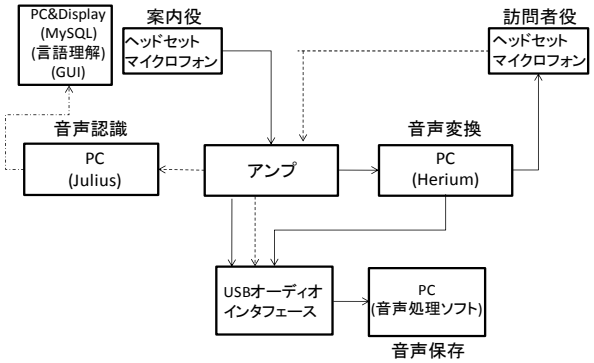

図 1：対話収録環境

本稿では，まず，上記のような特殊な状況での対話データの収集方法について述べ，次にこの対話データを機械学習に適用することにより，大まかではあるが理解の確信度を反映した応答行動が予測できることを示す．

2. コーパス収集

音声認識や言語理解の誤った情報が伝えられる状況において，それらを修復しながら情報提供を行う会話の収録実験を行った．

2.1 実験手続き

美術館の展示物に関する情報案内システムを想定し，2 名の実験協力者が美術館の案内役となり，訪問者役として参加した実験協力者，それぞれ 10 名ずつと会話を行った．これら 20 ペアにおいて 3 つの条件で，会話を行ってもらい，各ペア約 7 分の会話を 3 会話，全部で 60 会話を収録した．

(1) 収録環境

案内役と訪問者役はそれぞれ別の収録用の防音ブースに入り，ヘッドセットマイクを装着した．案内役の音声は音声変換ツ

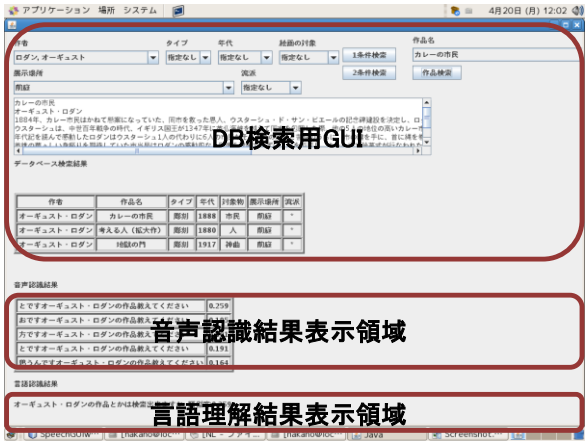

図 2：実験用システム

ール Herium [4] を用いて人工的な印象を与える音声に変換して，訪問者役に伝えられた．各被験者の音声は USB オーディオインタフェース Roland EDIROL UA1000 を経由して，PC のハードディスクに蓄積された．収録環境を図 1 に示す．さらに，訪問者役の音声は，音声認識と簡略的な言語理解にかけられ，その結果が案内役の前に設置された画面に表示された．

(2) 実験システム

さらに，訪問者役の発話情報の提示や，展示物についてのデータベース検索のための GUI を作成し，案内役の画面に統合した．実験システム画面を図 2 に示す．実験システムは以下の 3 つの部分から構成される．

- 音声認識：julius-4.0.2 Windows 版 [5] で処理され，文信頼度の高い順から上位 5 つの候補が画面に表示された．文信頼度は，julius-4.0.2 Windows 版が出力するスコアを事後確率化した値として算出した．音声認識のための言語モデルは予備実験として収集した同じ課題についての人間同士の対話コーパスを用いて作成した．
- 言語理解：あらかじめ登録されている典型的な質問文との類似度を計算し，もっとも類似度の高い典型質問文が言語理解結果として表示された．尚，典型的質問文とは，「＜作品名＞について教えてください」や，「＜年代＞の絵画はありますか」（＜＞内には具体的な作品名や年代が述べられる）のような表現である．同じ実験課題を用いた予備実験において訪問者役の質問を分析し，その中で，頻度の高いものを典型的質問文とした．
- データベース検索用 GUI：案内役が作品の検索を行うための GUI を音声認識，言語理解結果の表示画面と同一画面上に作成した．

(3) 教示

訪問者役には，美術館の入り口にいることを想定してもらい，何らかの視点から作品を見比べることを制約条件とし，案内システムに質問しながら鑑賞したい絵画を 3 つ選ぶよう指示した．また，絵画のリストと地図を紙で提示し，絵画を選ぶための参考にしてもらった．会話終了後，3 つの絵画を選んだ理由を尋ねることにより，訪問者役の課題への取り組み態度の確認と動機づけを行った．一方，案内役には，画面に表示された訪問者役発話の音声認識結果や言語理解結果を見ながら，会話を遂行するよう指示した．つまり，案内役は，誤りを含んだ音声認識や言語理解の結果から適当な応答を決定しなければならない状況であった．

(4) 実験条件

各ペアは，以下の 3 つの実験条件で対話を行った．

音声条件：音声認識結果のみを案内役の画面に表示

音声+言語条件：音声認識結果と言語理解結果の両方を案内役の画面に表示

認識可能リスト提示条件：訪問者役にシステムが認識できる文の一覧を与え，約 2 分間の練習後に音声+言語条件と同じ GUI を用いて会話を遂行

最初の 2 セッションで，音声条件と音声+言語条件を行い（どちらの条件を先に行うかはランダムに割り当てた），第 3 セッションでは，全てのペアが認識可能リスト提示条件で会話を行った．

2.2 収録データ

上記のような手続きにより遂行された各対話に関して，実験システムのログ，会話音声，およびその書き起こしを収集した．システムのログには，以下の情報がタイムスタンプとともにファイルに書き込まれ，ログファイルとして出力された．

- 音声認識候補の形態素リスト
- 音声認識候補の信頼度
- 最も類似度の高かった典型質問文とその類似度の値

さらに，ログファイルと書き起こしファイルを統合し，実際に起こった対話とシステムが認識した結果とを対応付けて見ることができる統合ファイルを作成した．統合ファイルの一例を図 3 に示す．ここで，点線で囲まれた部分はログファイルの内容であり，下線が引かれた「訪問者:」，「案内役:」の部分は，音声を書き起こした結果である．音声認識結果には部分的に誤りが含まれているものの，ある程度発話内容が推測できることが見て取れる．

3. 応答タイプの予測

前節で述べた対話収集実験では，認識誤りや理解誤りが数多く発生しているにも関わらず，すべてのペアで課題が完了しており，対話の失敗を防ぎながら訪問者役の質問が明確化されてゆく過程が観察された．例えば，作者や作品名などキーワードとなる言葉が正しく認識されているようであれば，「ルノワールの絵画ですか?」のように，認識されたキーワードの確認を行う，あるいは彫刻というキーワードが認識された場合，それに付随していたと思われる情報を得るために，「どの彫刻でしょうか?」のように，確認の範囲を限定するような質問を行うといったことが数多く観察された．

そこで，ログと書き起こしを統合したコーパスデータから，案内役の応答行動の予測を行い，本予測結果が対話システムにおける応答選択に有用な情報となりうるかを検討する．

3.1 応答ペアの抽出と応答の分類

(1) 応答ペアの抽出

訪問者役の発話に対する案内役の応答を予測するのが目的であるため，訪問者役のターンとそれ続く案内役のターンのペアを応答ペアとし，分析の単位とした．図 3 に示す例が 1 つの応答ペアである．ただし，訪問者のターンが「はい」や「わかりました」のような同意や相槌のみである場合は，これに対して案内

121.9425 125.0175 訪問者: ほかに[題材]人物の[分野]彫刻はありますか

音声認識結果 128.10547（認識時刻）:ASR:
第一位候補: 館+カン+名詞 に+ニ+助詞 [題材]人物+ジンブツ+名詞 の+ノ+助詞 [分野]彫刻+チョーコク+名詞 が+ガ+助詞 あり+アリ+動詞 ます+マス+接尾辞　文信頼度:0.238
第二位候補: 館+カン+名詞 に+ニ+助詞 [題材]人物+ジンブツ+名詞 の+ノ+助詞 [分野]彫刻+チョーコク+名詞 は+ワ+助詞 で+デ+動詞 が+ガ+助詞　文信頼度0.229
第三位候補: 館+カン+名詞 に+ニ+助詞 [題材]人物+ジンブツ+名詞 の+ノ+助詞 [分野]彫刻+チョーコク+名詞 が+ガ+助詞 あり+アリ+動詞 ま+マ+名詞　文信頼度:0.205
第四位候補: と+ト+助詞 に+ニ+助詞 [題材]人物+ジンブツ+名詞 の+ノ+助詞 [分野]彫刻+チョーコク+名詞 が +ガ+助詞 あり+アリ+動詞 ます+マス+接尾辞　文信頼度:0.167
第五位候補: と+ト+助詞 に+ニ+助詞 [題材]人物+ジンブツ+名詞 の+ノ+助詞 [分野]彫刻+チョーコク+名詞 は +ワ+助詞 で+デ+動詞 が+ガ+助詞　文信頼度:0.161

128.1211（認識時刻）:
第一位候補: 館に[題材]人物の[分野]彫刻があります　文信頼度:0.238
第二位候補: 館に[題材]人物の[分野]彫刻はでが　文信頼度:0.229
第三位候補: 館に[題材]人物の[分野]彫刻がありま　文信頼度:0.205
第四位候補: とに[題材]人物の[分野]彫刻があります　文信頼度:0.167
第五位候補: とに[題材]人物の[分野]彫刻はでが　文信頼度:0.161

言語理解結果 128.3086: NLU:[分野]彫刻に興味があるんですが文信頼度:0.238

134.4925 135.8750 案内役: [題材]人物の[分野]彫刻を
136.1875 137.0050 案内役: ご案内しますか

図 3：システムログと書き起こしの統合

役が具体的な内容を持つ確認行為を行うことはないので，応答ペアとはみなさない．

(2) 案内役の応答の分類

次に，抽出された応答ペアにおける，案内役の応答の分類を行った．予備的な検討において，施設の案内を目的とした情報提供会話において，データベースの項目となっている単語が，会話中のキーワードとして重視されていることが明らかになったので，我々はこの点に着目し，以下の(a)〜(j)の応答分類カテゴリを作成した．これらのカテゴリは分析者の判断により各応答に付与された．尚，本研究では，キーワード選定の恣意性を排除し，選定基準を明確化するために，キーワードはデータベース項目となっている単語（展示場所，作者名，題財，年代，流派，作品名）のみに限定した．図 3 で[]で表している単語がキーワードである．また，各形態素の認識結果は，「表記+読み+品詞」の形式で出力されている．

さらに，応答カテゴリを案内役による訪問役の発話理解の確信度という観点から 3 群に分け，これを機械学習の対象とした．

- 発話理解高確信度群：発話全体の理解に確信があると思われる応答
  (a) 終助詞「ね」を伴った案内役による訪問者発話の復唱
  (b) 訪問者役に確認をすることなく情報提供を行う
- キーワード認識高確信度群：1 つ，あるいは複数のキーワードを用いて，確認の応答を行う場合
  (c)キーワードのみの確認の発話
  (d)キーワードを含む確認の発話
  (e) キーワードを含む WH 疑問文
  (f) 典型的質問にキーワードを埋め込んだ確認の発話
- 低確信度群：応答発話中にキーワードが含まれない場合
  (g) キーワード以外の名詞を用いた確認の発話
  (h) キーワード以外の名詞を用いた WH 疑問文
  (i) 同一発話の反復要求
  (j) 当該応答ペアに直接関係しない発話

例えば，図 3 の例における，案内役の「人物の彫刻をご案内しますか」という発話は，典型的な質問文の型には一致していないが，キーワードを用いた確認となっているので，カテゴリ(d)に分類される．

3.2 特徴量の設定

訪問者役の発話に対して，案内役が上記のどのグループの応答行動を選択するのかを予測することを目的とし，認識候補間のキーワード，名詞等の一致の度合い，典型的な質問（2.1 節(2)を参照）との一致の有無と，一致有りの場合，その種類，前応答ペアからの履歴情報，直前のやり取りの種類等に関する計 20 種類の特徴量を設定した．

- キーワードに関連する特徴値
  (1) 音声認識結果の第 2〜第 5 位候補と第 1 位候補とのキーワード一致率の平均値
  (2) 音声認識結果の全ての候補で一致したキーワードの数
  (3) 全ての候補で一致したキーワード数/見つかったキーワード数
  (4) 見つかったキーワード数－全ての候補で一致したキーワード数
  (5) 音声認識結果の第 3 位候補までで一致したキーワードの数
  (6) 音声認識結果の第 3 位候補までのでのキーワードの一致率

(7) (5)と(2)の差
(8) 現応答ペアの(2)の中で，前応答ペアの(2)と一致するキーワードの割合
(9) 現応答ペアの(5)の中で，前応答ペアの(5)と一致するキーワードの割合

- キーワード以外の名詞，指示詞に関連する特徴値
(10) 音声認識結果の 第 3 位候補までに出現する指示詞の数
(11) 音声認識結果の 第 3 位候補までに出現する指示詞の種類
(12) すべての候補で一致した名詞の数
(13) 音声認識結果の第 1 位候補に出現する名詞の中で第 2～第 5 位候補中にも現れるものの割合
(14) 音声認識結果の第 3 位候補までで一致した名詞の数
(15) 音声認識結果の第 1 位候補に出現する名詞の中で第 2，第 3 位候補中にも現れるものの割合
- 典型的質問に関連する特徴値
(16) 音声認識結果の第 1 位候補は典型的質問であると認定できるか
(17) (16) が真である場合，その典型的質問のカテゴリ名
(18) (16)で認定された典型的質問と音声認識結果の第 2～5 位候補において認定された典型的質問との一致率
(19) 音声認識結果の第 3 位候補までての典型的質問の一致率
(20) (19)が 100%であった場合，そこで認定された典型的質問のカテゴリ名

3.3 応答行動の予測とその精度評価

以上の特徴量から応答行動のタイプを予測するモデルを SVM を用いて生成し，その精度評価を行った．ここでは，収集した対話のうち，音声+言語条件の 20 対話分のデータのみを使用した．SVM は weka [6] による実装を用いた．予測の評価結果を表 1，表 2 に示す．予測精度は 5 回の交差検定の結果を採用している．まず，各応答タイプを予測対象とした場合では，65%と低い予測精度しか得られなかった．カテゴリごとの予測精度を見てみると，特に発話理解高確信度群の予測ができていなかった．これは，確信度が十分高い場合でも，対話の失敗を避けるために，より安全な対話方略をとり，キーワードを用いた確認の発話を行うことが多く，そのため，発話理解高確信度群とキーワード認識高確信度群とを区別することが難しいからだと考えられる．

そこで，発話理解高確信度群とキーワード認識高確信度群とを区別せず，これらをまとめて中高確信度群とし，低確信度群との区別が可能であるか否かを評価した．その結果を表 2 に示す．この場合は，全体の予測精度も 75%まで向上し，両カテゴリの F-measure も向上している．

表 1:3 カテゴリでの予測結果

カテゴリ	Precision	Recall	F-Measure
発話理解高確信度群	0	0	0
キーワード認識高確信度群	0.669	0.874	0.758
低確信度群	0.627	0.627	0.627

表 2:2 カテゴリでの予測結果

カテゴリ	Precision	Recall	F-Measure
中高確信度群	0.794	0.839	0.816
低確信度群	0.667	0.597	0.63

4. まとめと今後の課題

本研究では，音声認識や言語理解の結果が画面に表示される状況において対話データを収録し，確認や会話の修復等のデータを収集するとともに，これらを用いた分析の一例を報告した．まず，データ収集では，音声認識と言語理解の有無と認識語彙の提示条件を変えることにより，3 つの条件で対話収録を行った．収録された音声を書き起こし，音声認識や言語理解結果と統合することによりコーパスを作成した．次に，このデータを用いて，案内役の応答行動の予測を試みた．その結果，比較的理解の確信度が高いと思われる応答，すなわち，確認なしか認識できたキーワードを積極的に使用した確認を行う場合と，同一発話の反復依頼やキーワード以外の名詞を用いた確認発話等，理解の確信度が低い状態で行われる応答行動とを区別することがある程度可能であることがわかった．

しかし，これは十分な精度とはいえない．特に，高確信度群のみを予測することができないこと，低確信度群の予測精度に問題があることなどが課題である．今後は，応答行動の予測モデルの精度向上が必要である．また，さらに詳細な質的分析を行うことによって，認識された語のうち，どのような語を重視するのか，あるいは誤認識があっても，文脈上大きな影響がなく，特に修正する必要のない言葉とはどのような性質をもつのか等を明らかにすることにより，情報検索・提供の音声対話における，適切な応答決定方式の確立を目指す．

参考文献

1. 神田直之, et al., *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, 2006. **47** (6): p. 1802-1811.
2. 翠輝久, et al., *質問応答・情報推薦機能を備えた音声による情報案内システム*. 情報処理学会論文誌, 2007. **48**(12): p. 3602-3611.
3. Paek, T. and E. Horvitz, *Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems*, in *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, S.E. Brennan, A. Giboin, and D. Traum, Editors. 1999, American Association for Artificial Intelligence: Menlo Park, California. p. 85-92.
4. *Herium 音声変換ツール*. [cited; Available from: http://www.sp.m.is.nagoya-u.ac.jp/people/banno/spLibs/herium/index-j.html.
5. *julius-4.0.2*. [cited; Available from: http://julius.sourceforge.jp/forum/viewtopic.php?f=13&t=53.
6. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2nd ed. 2005, San Francisco: Morgan Kaufmann.