Vol.2009-SLP-79 No.37
2009/12/22

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その利用

小野正貴[†]　　中野有紀子[††]

音声対話システムは，インタラクティブに情報検索・収集ができる効果的なインタフェースである．しかしながら音声対話システムのボトルネックは音声認識誤りである．そのため，システムは音声認識誤りを考慮しながら，適切なシステム応答を選択することが必須である．そこで本研究では，音声認識誤りを考慮した対話制御方法を検討するための基礎データとして，美術館での情報案内の会話に焦点を当て，情報案内を行う案内者が訪問者発話の音声認識結果の文字列を受け取る状況で，対話コーパス収集実験を行った．また，このコーパスを利用した研究例として，音声認識誤り訂正候補の抽出と，機械学習を用いたシステム応答予測を試みた結果を報告する．

Collecting Spoken Dialogue Data via Speech Recognition and Natural Language Understanding Systems: The First Report of the Corpus Collection and its Applications

Masaki Ono[†]　and Yukiko Nakano[††]

Spoken dialogue systems are useful and effective in accessing and retrieving information interactively.　However, a bottleneck of spoken dialogue systems is speech recognition error. Therefore, the system needs to select appropriate system responses by considering the speech recognition errors. As the basics of studying dialogue management techniques considering speech recognition errors, first, 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. We also report our studies using this corpus; generating candidates for speech recognition error correction, and predicting the types of response behaviors using machine learning.

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した．図 1 で実験のイメージを示す．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．

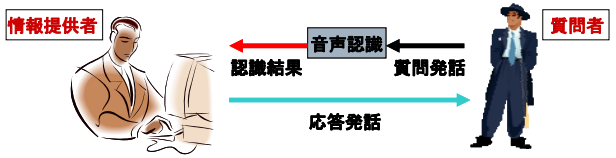

図 1: 対話実験のイメージ図

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，このような誤りを含んだ発話の認識結果を訂正するための候補の選出手法について述べる．最後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．

*[†] 東京農工大学
Tokyo University of Agriculture and Technology
[††] 成蹊大学/東京農工大学
Seikei University/ Tokyo University of Agriculture and Technology

Vol.2009-SLP-79 No.37
2009/12/22

2. 対話コーパス収集

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

収録環境を図 2 に示す．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．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 PC の画面に表示された．

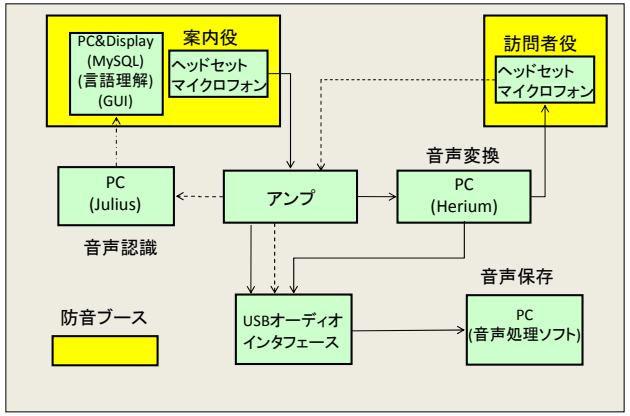

図 2：収録環境

(2) 実験システム

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 3 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，典型的質問文は，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，頻度の高かった文型である．例えば，「＜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．

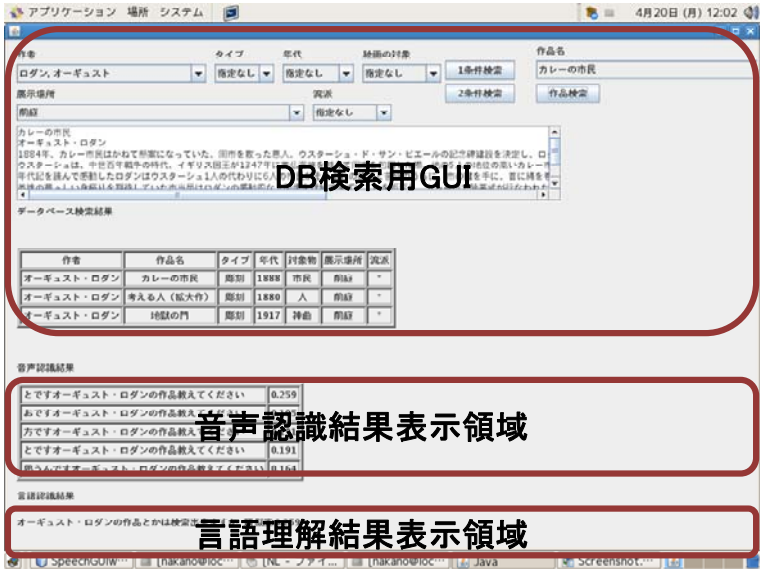

図 3：実験用システム

(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

システムログと書き起こしを統合したものを図 4 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．ここでは，「ほかに」が「館に」や「どこに」に誤認識されるなど，文末の表現に乱れがあるものの，「人物の彫刻」について知りたいという部分は理解できるため，ある程度発話内容が推測できることが見て取れる．

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: 形態素解析結果
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）： 認識文

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 4：システムログと書き起こしの統合

3. コーパス利用例 1：音声認識誤りの誤り訂正候補提案

図 4 に示したように，本実験において収集したコーパスには，認識誤りがあるにもかかわらず案内者が正しい発話を予測したことが若干だが見られた．これは人間の判断で文脈情報，音響情報を用いて正解を推定していたと考えられる．このことからこのコーパスの文脈情報と音響情報を用いて音声認識誤りの訂正候補の抽出を試みる．

3.1 キーワード

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(DB) 登録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．表 1 に DB の登録例を示す．つまり，これらの単語は全てキーワードと認定される．また，図 4 で[]で示されている単語がキーワードである．

3.2 DB からの関連語の抽出

キーワードとキーワードに関連する単語をまとめて関連語群と呼ぶ．関連語は以下の方法で決定する．

(1) 現在音声認識の対象となっている訪問者の発話の直前にある案内役ターン中の発話リストを取得．

(2) この発話リスト中に含まれるキーワードのリストを取得.

(3) このキーワードリスト中の各キーワードについて以下の方法で, DB から関連語群を決定.

(3-1) ターゲットとなるキーワードをキーとして, DB を検索し関連語候補を得る. 例えば, ターゲットとなるキーワードが「オーギュスト・ロダン」である場合, DB 登録 1 と 2 をあわせたもの(ただし重複は除く)が関連語候補となる. つまり, 関連語候補は, オーギュスト・ロダン, カレーの市民, 彫刻, 十九世紀, 市民, 前庭, 考える人, 人, の 8 単語である.

(3-2) (3)で求めたキーワードリスト中の全ての単語について(3-1)の処理を行う.

(3-3) 上記の処理で求めた単語リストから重複を取り除いたものを関連語群とする.

	作品	作者	分類	時代	対象物	場所
1	カレーの市民	オーギュスト・ロダン	彫刻	十九世紀	市民	前庭
2	考える人	オーギュスト・ロダン	彫刻	十九世紀	人	前庭
3	ゲッセマネの祈り	ルカス・クラーナハ	宗教画	十六世紀	橄欖山	本館 2 階

表 1:DB 登録例

3.3 対話履歴を用いた関連語群の拡張

前節では, 直前の案内役のターンからのみ関連語を抽出していたが, 対話の文脈を考慮するには, より以前の発話中に含まれるキーワードからも関連語を抽出する必要があると考え, 3 つ前のターンまでを文脈として取り入れ, これらの中に含まれるキーワードから得られる関連語を関連語群に加えた.

3.4 音声認識誤りの誤り訂正候補の抽出

誤認識された結果と前節で述べた各関連語との類似度を以下の方法で算出し, 誤認識の訂正候補抽出を試みた. 図 5 において訂正候補抽出の手法を図で示す.

(1) 類似度計算手法

レーベンシュタイン距離（編集距離）と N-gram を用いて音声認識結果の音素と関連語の音素の類似度を求め, その類似度によって誤認識の訂正を試みる.

以下に, 書き起こしと音声認識結果, および音声認識結果の音素列を示す.

書き起こし：十九世紀の作品はどこにありますか
音声認識結果：途中九世紀の作品はどこにありますか
音声認識結果[音素]：tochuukyuuseikinosakuhiNhadokoniarimasuka

まず, 音声認識結果[音素]を N-gram（N は比較対象の関連語の音素の数）によって分割する. 例えば,「十九世紀」の音素数"jyuukyuuseiki"は 13 なので, n=13 で, 認識結果を分割すると以下のようになる.

N-gram を適用後：tochuukyuusei, ohuukyuuseiki, huukyuuseikin, ・・・

これらの要素すべてと比較対象の関連語のレーベンシュタイン距離を求める.
これによって認識結果のすべての音素パターンについて関連語とのレーベンシュタイン距離を求めることができる.

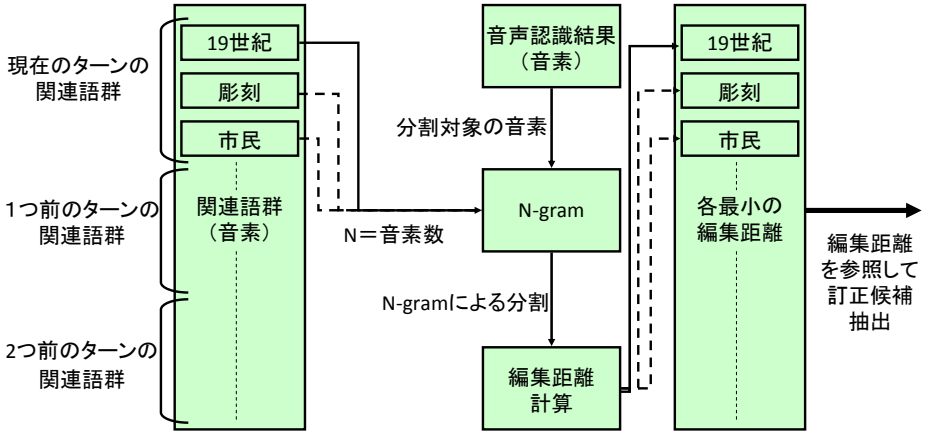

図 5：訂正候補抽出の手法

3.5 音声認識誤りの誤り訂正の精度評価

音声認識誤りがある状況において, 提案方式による誤り訂正の精度評価を行った. まず, 誤りを含む訪問者発話の認識結果を抽出し, それに対する音声からの書き起こしを正解データとする. 当該発話の音声認識結果 5-best において正解のキーワードが含まれない場合, 5-best それぞれに対して, 訂正候補昇順 5 個を抽出する. そして, これらの候補の中に正解のキーワードが含まれていた場合, 訂正候補が正解であるとみなす. ここでは収集したコーパスの 59 対話を用いた. 結果を表 2 に示す.

音声認識誤り数	誤り訂正の正解数	%
95	61	64.2

表 2:関連語による誤り訂正の正解率

表 2 のように精度は 65%弱と低い水準にとどまっている. しかしコーパスを調べた結果, 誤りを含んだ認識結果に対して案内役が実際に誤り訂正を行った割合は 30%未満であるが, 本手法はすべての認識誤りについて訂正を試みたために結果的に高い訂正率が得られなかったと考えられる. 人間の訂正行動との詳細な比較を行い, 本手法の評価を再検討すべきだと考える.

4. コーパス利用例 2：応答タイプの予測

第 2 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 (図 4) 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.

4.1 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 4 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, 案内役の応答では, 訪問者の発話内容を確認することが多く, その際, 会話中のキーワードが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, 案内役による確認の表現を以下の 3 種類に分け, これを機械学習の対象とした.

＜応答タイプ＞

- 完全繰り返し表現：訪問者の発話をほぼ完全に繰り返す場合
  例：訪問者「ロダンの考える人の展示場所を知りたいのですが」
  案内役「ロダンの考える人の展示場所をお知りになりたいのですか?」
- キーワード付表現：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
  例：「ロダンの彫刻ですか」
- キーワード無し表現：応答発話中にキーワードが含まれない場合. キーワード以外の名詞を用いた確認の発話等
  例：「どのような絵をお探しですか」

4.2 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, 5-best, 3-best 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[6]を設定した.

- キーワードに関連する特徴値
  例：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
- キーワード以外の名詞, 指示詞に関連する特徴値
  例：音声認識結果の 第 3 位候補までに出現する名詞の数
- 典型的質問に関連する特徴値
  例：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

4.3 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 [7] による実装を用いた. 予測精度は 5 回の交差検定の結果を採用している. まず, 3 種類の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ず, 特に完全繰り返し表現の予測ができていなかった. これは, 対話の失敗を避けるために, より安全な対話方略であるキーワードを用いた確認の発話を行うことが多かったためと考えられる.

そこで, 完全繰り返し表現とキーワード付表現をまとめ, 情報が少ない状況で行われるキーワード無し表現との区別が可能であるか否かを評価した. その結果を表 3 に示す. この場合は, 全体の予測精度も 75%まで向上している.

カテゴリ	Precision	Recall	F-Measure
完全繰り返し表現 or キーワード付表現	0.794	0.839	0.816
キーワード無し表現	0.667	0.597	0.63

表 3：2 カテゴリでの予測結果

5. 全体の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，音声認識誤りの訂正候補を抽出することを試み，その結果，65%ほどの誤りを訂正できることがわかった．今回の手法では文脈情報を関連語群を追加するという形で簡易的に与えていたが，今後は，文脈情報の付与方法を工夫するなどして精度向上を目指す．2 つ目のコーパス利用例として，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，認識できたキーワードを積極的に使用した確認か，相手へ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かどちらが次に出現するのかを 75%の精度で予測できることがわかった．しかし，キーワード無し表現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．

最後に，本研究では応答行動の予測と音声認識誤りの訂正を独立の研究として行ったが，今後は音声認識誤りの誤り訂正を応答行動の予測に役立てていきたい．具体的に一例を挙げると，発話の認識結果が誤認識により理解できない内容になっているときは，応答方法として，「もう一度言ってください」といった反復依頼をする等，会話の内容が発展しない行為が繰り返されることが今回の実験コーパスでは複数みられた．提案手法を用いると，このような場合において誤り訂正候補の結果により何度も同じ関連語が抽出されると想定され，これらの語を使った応答を行うことにより発話者の意図に沿った方向で，会話を発展させることができると考えられる．

6. 謝辞

本研究は（株）日立製作所中央研究所との共同研究の一環として実施されたものである．本研究を遂行するにあたり有益なコメントをくださった，日立中央研究所本間健氏，神田直之氏，永松健司氏，さらには本研究を御支援くださった大淵康成氏に対し深甚の謝意を表する．

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 *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) 小野正貴, et al., 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, 第 23 回人工知能学会全国大会,1I1-1, 2009.
7) Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2nd ed. 2005, San Francisco: Morgan Kaufmann.

著者紹介

小野正貴
2008 年金沢大学工学部情報システム工学科卒業，現在東京農工大学大学院工学府情報工学専攻ユビキタス&ユニバーサル情報環境専修修士課程に在学中，優れたユーザインタフェースであるとして音声対話システムに取り組んでいる．

中野有紀子（正会員）
1990 年東京大学大学院教育学研究科修士課程修了．同年，日本電信電話株式会社入社．2002 年 MIT Media Arts & Sciences 修士課程修了．(独) 科学技術振興機構社会技術研究開発センター専門研究員，東京農工大学大学院工学府特任准教授を経て，2008 年 4 月より成蹊大学理工学部情報科学科准教授．博士（情報理工学）．