対話ログによる訓練が可能なインタラクティブ音声理解システムの枠組

中野 幹生 †，東中 竜一郎 †，Matthias Denecke†，須藤 克仁 †，宮崎 昇 ‡，堂坂 浩二 †
† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所
‡ 日本電信電話 (株) NTT サイバースペース研究所

概 要

音声で対話しながらユーザの要求を理解するシステムの構築の枠組について検討する．従来，このようなシステムの構築には，専門家によるドメイン依存の知識源の記述が必要であったが，広く用いられるためには，専門家でなくても高性能なシステムの構築ができる必要がある．本稿では，専門家でなくても初期システム用のドメイン依存知識が記述でき，初期システムとユーザとの対話ログを用いて知識源を訓練することにより，システムの性能を向上できるようなシステム構築の枠組を提案する．

1 はじめに

機械と人間が音声言語でコミュニケーションできるようにする技術は，機械を使いやすくし人々の生活を豊かにすると期待されている．その中で最も重要な技術のひとつは，人間と音声で対話しながら，ユーザの要求を理解する技術である．ユーザの発話を一度で正しく認識・理解できるとは限らないため，問い返しや確認が必要になること，および，ユーザが一度の発話で要求を完全に伝えることができないことから，精度よくユーザの要求を理解するには，単発話の理解だけではなく，ユーザとのやりとりが必要になる．本稿では，そのようなシステムを，**インタラクティブ音声理解システム** (Interactive Speech Understanding system, **ISU**) と呼ぶ. 音声対話システムと異なる呼び方をする理由は，音声対話システムの研究には，ユーザの要求理解だけではなく，ユーザへの情報伝達の方法の研究も含まれるからである．

ISU はどのようなタスクドメインの要求でも理解できることが望ましい．しかしながら，あらゆるドメインの発話が理解できる言語理解モデルの構築法が確立されていないことから，近い将来での実現は難しい．代わりに，特定のタスクドメインの要求を扱うシステム，たとえば，特定のドメインの情報検索・コールルーティングや，テキストコーパスを対象とした質問応答システムなどが研究されてきた．これらの ISU の中には実用レベルに達しているものもあるが，新たなタスクドメイン・タスクタイプの ISU を作るには，専門家の手による，タスクドメイン依存の知識源 (以下**モデル**と呼ぶ) の作成が必要である．ISU が広くインターフェースとして受け入れられるためには，専門家でなくても簡単にモデルが作成できることが必要である．専門知識がなくても簡単に ISU をつくるためのツールキットが開発されている [1, 3, 11]．しかしながら，ツールキットは，プロトタイプを作るのには適しているが，人手をかけてチューニングをしたシステムに匹敵するようなものを作るのは容易ではない．

本稿では，専門家でなくても初期システム用のモデルが作成でき，初期システムとユーザとの対話ログを用いてモデルを**訓練** (training) することにより，システムの性能を向上できるようなシステム構築の枠組を提案する．さらに，従来の ISU に関する研究を整理し，提案する枠組が実現可能であることを示す．

2 インタラクティブ音声理解システム

2.1 インタラクティブ音声理解システムのタスク

ここではまず ISU を定義する．まず，**ユーザ要求**は，ユーザ要求タイプと，スロット (スロット名と値のペア) の集合で表される付属情報からなる．スロット値はアトミックなシンボルもしくはスロットの集合とする．たとえば,「水曜日の 9 時すぎに出るのぞみの指定席が欲しい」というユーザ要求は，次のようにあらわされる．

```
ユーザ要求タイプ: 予約
付属情報:
  [列車種別:    のぞみ
   日にち:      水曜日
   出発時間帯: [時間:          9 時
                時間付加情報: すぎ]]
```

ユーザ要求タイプと各スロット値の間には，共起に制約があるとする．可能なユーザ要求の種類は有限個と

A Framework for Interactive Speech Understanding Systems that are Trainable using Dialogue Corpora. Mikio Nakano, Ryuichiro Higashinaka, Matthias Denecke, Katsuhito Sudoh, Noboru Miyazaki, Kohji Dohsaka (NTT Corporation)

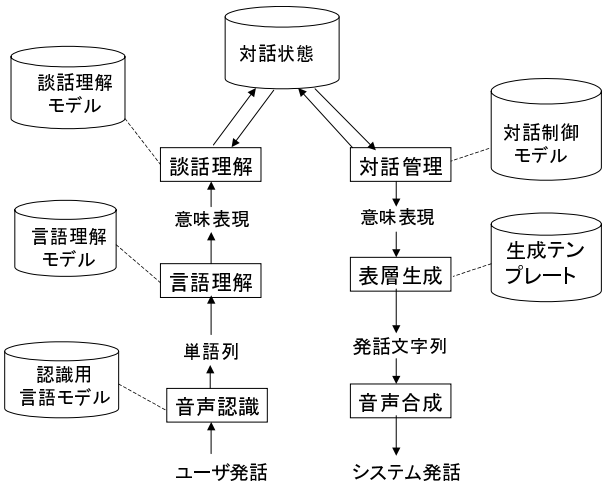

図 1: ISU の一般的な構成

する．

上記の形式であらわすことができないユーザ要求もありえるが，従来の音声対話システムの多くが，上記の形式か，より単純なものを用いており，応用範囲は十分広いと考えられる．

ISU はどのようなアプリケーションの中で用いられるかによって，要求される機能が異なる．時間がかかっても高精度で意図理解を行わなければならない場合もあれば，低精度でもなるべく速く理解結果を得なければならない場合もある．これらの条件を考慮して ISU の評価を行わなければならないが，本稿の主題とずれるのでここでは議論を省略する．

3 対話ログを用いたモデルの訓練

3.1 インタラクティブ音声理解システムの構成

われわれが仮定する ISU の一般的な構成を図 1 に示す (詳細は [5] 参照)．ISU は複数のモジュールからなり，各モジュールがドメイン依存のモデルを参照しながら動作する．

3.2 インタラクティブ音声理解システム構築の枠組

以下にわれわれが考えるインタラクティブ音声理解システム構築の枠組を示す．

(1) ツールキットを用いて**初期システム** S_0 を作成し，$i = 0$ とする．
(2) システム S_i と一般ユーザとの対話を収録する．
(3) 収録されたデータにタグ付けする．
(4) タグ付けされたデータを用いてモデルを訓練する
(5) 訓練されたモデルを用いたシステム S_{i+1} を構築する．
(6) $i = i + 1$ とし，ステップ (2) に戻る．

ステップ (1) の初期システムの作成とステップ (3) のタグ付けは，音声言語処理の専門家でなくても，一般的なプログラミングの知識があればマニュアルを読む程度でできることが必要である．

3.3 訓練に適したモデルの構成

以下では，上記の枠組に適したモデルの構成について議論する．

3.3.1 初期システムのための知識記述

まず，初期システムのためのドメイン依存知識の記述 (以下**初期記述**と呼ぶ) について述べる．初期記述は，以下のものからなる．

キーフレーズ キーフレーズのクラスと各々のクラスに属するフレーズのリストを記述したものである．以下に例を示す．

```
クラス:    駅名
フレーズ:  東京 | 東京 駅 (東京)
           大阪 | 新大阪 | 新大阪 駅 (新大阪)
```

ここで括弧の中は，ユーザ発話の意味表現中で使われるシンボルをあらわす．

発話例 ユーザ発話の例とその意味表現，すなわち言語理解の結果の組で，以下が一例である．

```
発話:        東京駅まで行きたい
意味表現:  [発話タイプ: 目的地指定
             目的地:       東京]
```

談話理解規則 ユーザ発話の意味表現と直前のシステム発話の内容があたえられたとき，理解状態をどう変化させるかを記述した規則の集合．ここで，理解状態とは，その時点までの対話からユーザ要求を理解した結果 (2.1 節で述べた表現であらわされるもの) と，それぞれのスロットが確認済みかどうかの情報を含むデータ構造である．以下は，ユーザ発話が目的地を指定するものと理解したとき，理解状態の目的地スロットを埋めるという規則である．

```
条件:
  意味表現:                [発話タイプ:地名指定
                            地名:        *X]
  直前のシステム発話: 目的地指定要求
アクション:
  対話状態の変化のさせ方: <目的地>=*X
```

各談話理解規則を適用したときの理解状態の変化の仕方は一意に決まる．すなわち，意味表現は，理解状態を変化させる命令と捉えることができる．

理解状態が満たすべき条件　理解状態が満たすべき条件で，スロット値への制約として記述する．以下は，目的地と到着地は異なるという条件である．

<目的地> != <到着地>

対話制御規則　理解状態に関する条件と，その条件が満たされたときに行う発話の意味表現を記述したものの集合である．合理的であると考えられるすべての規則を記述しておく．

条件:
　理解状態: <目的地>!=nil
　　<目的地確認>=no
システム発話:
　意味表現: [発話タイプ: 目的地確認
　　目的地: <目的地>]

ここで，<目的地>!=nil は，目的地スロットの値が空でないことを示す．

生成テンプレート　システム発話意味表現と文字列の組を記述したものである．以下は目的地を確認する発話を生成するテンプレートである．

意味表現: [発話タイプ: 目的地確認
　　目的地: *X]
文字列: *X までですね？

以上の初期記述を行い，初期システムがデータ収集に耐えられるパフォーマンスを出せるようにするには，ある程度の経験が必要だと思われる．しかし，文法規則を記述する必要や，規則の優先順位を規定する必要がなく，要求される前提知識はあまり多くないため，サンプル記述や記述のためのツールを整備すれば，比較的容易に記述が行えると考えられる．

3.3.2　対話ログへのタグづけ

初期システムとユーザとの対話が収集されたら，以下のタグづけを行う．

- ユーザ発話の書き起こし
- ユーザ発話の意味表現

書き起こしにおける表記・単語切りなどの一貫性を保つため，キーワードおよび発話例およびから作成するツールを用意する．また，意味表現を入力しやすくするためのツールも用意する必要がある．

3.3.3　モデルの構成と訓練

音声認識　音声認識部は，入力中の音声区間を切り出し，その認識結果，すなわち単語列の N-best リストを言語理解部に送る．音声認識部が用いるドメイン依存のモデルとして，単語 n-gram モデル，特に単語クラス n-gram を用いる．初期システムのモデルは，初期記述の発話例から学習する．対話データ収集後は，発話の書き起こしを自動で単語切りしたものを学習データとして用いる．クラス n-gram を用いるのは，初期記述の発話例の量はもちろん，対話データを収集した後でも，データ量がクラスなしの n-gram の訓練には不十分と考えられるからである．

言語理解　言語理解部は，音声認識結果の n-best リストを受け取り，意味表現の m-best リストを談話理解部に送る．n と m が異なるのは，ひとつの単語列から複数の意味表現が得られる場合や，認識誤りなどにより認識結果が理解不能である場合があるからである．言語理解部が用いるモデルには，統計的言語理解モデルを用いる ([10] など)．初期システムのモデルは，発話例を基にした単純なネットワークによる理解モデルと，キーワード抽出を組み合わせたものとする．対話データ収集後は，データ中の発話の書き起こしとその意味表現を用いてモデルを拡張し，理解精度を高める．

談話理解　談話理解部は言語理解結果の m-best リストを受けとり，対話状態の内容をアップデートする．理解結果の曖昧性を表現するため，複数の対話状態の候補が保持されている．各対話状態候補には，談話理解結果とそれまでのユーザ発話の理解の履歴が含まれている．談話理解のモデルは，言語理解規則と対話状態が与えられたとき，一意に対話状態をアップデートするような規則の集合であり，初期記述から直接的に作られる．言語理解結果が m 個あり，対話状態候補が l 個あるならば，ml 個の新しい対話状態候補が得られる．談話理解モデルは，理解状態が満たすべき条件も含む．新しく作られた ml 個の対話状態のうち，この条件を満たさないものは捨てられる．

初期システムにおいては，m, l ともに 1 とすることにより，曖昧性を扱うことなく処理を行う．対話データ収集後，データから得られた談話に関する統計情報を元に，ml 個の対話状態候補にスコアをつけ，最適な対話状態候補を選ぶとともに，スコアの低い候補を捨てる．談話に関する統計情報として，直前までの談話

理解結果とユーザ発話との関係をクラスタリングしたものの生起確率，および，ユーザ発話タイプとシステム発話タイプの連鎖確率 (trigram) を用いる (詳細は文献 [4] を参照されたい)．これらの情報は，ユーザ発話が書き起こされ正しい意味理解結果が与えられた対話データに談話理解規則を適用することによって得ることができるため，談話理解結果のタグ付けを行う必要はない．

この枠組を用いることにより，文脈に依存した発話理解の曖昧性を対話データの収集とタグ付けのみから行うことができるようになる．

対話制御 対話制御部は最もスコアの高い対話状態候補の内容を基に，システム発話の意味表現を言語生成部に送る．システム発話の内容は，対話状態 (各対話状態候補) に登録される．どのようなシステム発話を行うかは，対話制御モデルによって決められる．

初期システムの対話制御モデルは，初期記述の対話制御規則から，適用可能な規則をランダムに選ぶというモデルである．対話データが収集された後，強化学習を使い, 最適な規則を選ぶための評価関数を学習する [8]．この学習過程では，対話をマルコフ決定過程ととらえ，その遷移確率をコーパスから学習することにより，各々の対話状態で各々の規則に基づく発話を行った場合に，期待されるタスク達成度や対話の短さなどを評価関数として計算する．評価関数の入力値は，規則，理解結果の内容，理解結果の信頼度，対話の履歴などである．一般に強化学習には大量の対話データが必要になる問題があるが，評価関数を，類似の対話状態の評価関数を用いて近似することにより，少量のデータで効率よく学習ができる方法を提案している [2].

言語生成 言語生成部はシステム発話の対話行為表現を受け取り，文字列を音声合成部に送る．ドメイン依存モデルは発話生成テンプレートを用いる．発話生成テンプレートは初期記述からストレートフォワードに作ることができ, 特に訓練する必要はない．

3.4 タグ付け量の削減

以上述べた枠組により，初期システムを使って得られたデータを用いてモデルが訓練できるが，タグ付けの労力がかかるという問題がある．この問題の解決のため，音声認識・理解結果の信頼性尺度を用いる方法がある．信頼性が高い場合には認識・理解結果をそのまま訓練に用いる (教師なし学習)．信頼性が低い場合には優先的にタグ付けを行うことにより，少ないタグ付け量で効率よく訓練を行う (能動学習). これらの手法は認識用言語モデルの訓練に用いられ，有効性が確認されており [6, 7, 9]， 本稿で提案した枠組に導入することが可能である．

4 おわりに

本稿では，専門家でなくても初期システム用のドメイン依存モデルが構築でき，初期システムとユーザとの対話ログを用いてモデルを訓練することにより，システムの性能を向上できるようなシステム構築の枠組を提案し，その実現可能性を示した．今後は，初期記述をさらに容易にする方法，必要なタグ付け量の削減しつつシステムのパフォーマンス向上を達成する方法を検討していく．

参考文献

[1] M. Denecke. Rapid prototyping for spoken dialogue systems. In *COLING-2002*, 2002.

[2] M. Denecke, K. Dohsaka, and M. Nakano. Fast reinforcement learning of dialogue policies using linear function approximation. In *IJCNLP-04*, 2004.

[3] J. Glass, E. Weinstein, S. Cyphers, J. Polifroni, G. Chung, and M. Nakano. A framework for developing conversational user interfaces. In *CADUI-04*, 2004.

[4] R. Higashinaka, M. Nakano, and K. Aikawa. Corpus-based discourse understanding in spoken dialogue systems. In *ACL-2003*, 2003.

[5] 中野, 堂坂. 音声対話システムの言語・対話処理. 人工知能学会誌, 17(3):271–278, 2002.

[6] M. Nakano and T. J. Hazen. Using untranscribed user utterances for improving language models based on confidence scoring. In *Eurospeech-2003*, 2003.

[7] G. Riccardi and D. Hakkani-Tür. Active and unsupervised learning for automatic speech recognition. In *Eurospeech-2003*, 2003.

[8] S. Singh, D. Litman, M. Kearns, and M. Walker. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.

[9] K. Sudoh and M. Nakano. Post-dialogue recognition confidence scoring for improving statistical language models using untranscribed dialogue data. In *ASRU-03*, 2003.

[10] 須藤, 中野. 音声対話システムのための統計的言語理解モデルの構成とその学習. 言語処理学会第 10 回年次大会論文集, 2004.

[11] S. Sutton, R. A. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. W. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. W. Massaro, and M. Cohen. Universal speech tools: The CSLU toolkit. In *ICSLP-98*, pp. 3221–3224, 1998.