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

東中竜一郎　中野 幹生　相川 清明

日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
{rh,nakano}@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

1　はじめに

音声対話システムがユーザと対話しながらユーザの意図を正しく理解しタスクを達成するためには，ユーザ発話を受け付ける度に，その内容を理解して適切に対話状態を更新する必要がある [8]．ここで言う対話状態とは，システムが内部に保持するさまざまな対話に関する情報のことを指し，本稿では文脈と同義とする．例えば，対話状態は対話の各時点でのユーザ意図の理解結果や，ユーザ発話履歴，システム発話履歴などを含む．ユーザ発話のみを用いて発話意図，発話内容を得ることを，音声理解と呼び，現在までの対話状態と，音声理解結果の両方を用いて対話状態を更新することを，談話理解と呼ぶ．一般に，ユーザ発話の音声理解結果は，音声認識結果を精度良く一意に決定する事が困難な事や，構文・意味の曖昧性が存在することから，複数の候補が得られる．しかしながら，音声対話システムが応答を出力するにはユーザ発話の解釈を一意に決定する必要があるため，談話理解の際には現在までの対話状態を参照しつつ，複数の音声理解結果から適切なものを選択する必要がある．

従来の多くのシステムでは，ユーザの発話を受け付けるたびに，談話理解の結果，すなわち対話状態を一意に確定する．しかし，現在までの対話状態とユーザ発話から得られる複数の音声理解結果からは複数の対話状態が導出されるため，曖昧性が存在する．この曖昧性を無視し，ユーザ発話を受け付けるたびに対話状態を一意に確定すると，談話理解の精度が下がる可能性がある．そこで，本稿ではユーザ発話を処理した後にも対話状態の曖昧性を残し，複数の対話状態候補の中から，次の入力との整合性を考慮して対話状態の曖昧性を解消する事により，談話理解の精度向上を試みる．このような複数の対話状態と複数の音声理解結果の組合わせを扱う研究は以前にも報告されているが [9]，対話状態の曖昧性を解消する処理において開発者の直観に基づく規則を用いており，規則の作成コスト，および精度の点で問題がある．そこで本稿では，対話状態の曖昧性の解消のため，音声対話システムとユーザとの対話コーパスから得られた統計情報を用いた談話理解法を提案する．

2　音声対話システムの談話理解

音声対話システムの一般的な構成を図 1 に示す．ユーザ音声の入力に対し，以下のようにシステムは動作する．

1. 音声認識器がユーザの発話音声を入力とし，認識単語列を出力する．
2. 言語理解部は認識単語列を入力とし，構文解析・意味解析等の言語処理を行い，対話行為[1]と呼ばれる表現に変換し，出力する．ユーザ発話と対応する対話行為の例を以下に示す．

ユーザ発話	2 時から 3 時までです
対話行為	[対話行為タイプ: 時間指定 [開始: 2 時][終了: 3 時]]

3. 談話理解部は対話行為と現時点での対話状態を入力とし，対話状態を更新する．
4. 対話管理部は更新された対話状態を入力とし，システムの次発話を決定し，発話文字列を出力する．同時に対話状態を更新する．
5. 音声合成部は対話管理部の出力を受け取り，応答を音声でユーザに伝える．

本稿ではこの中で談話理解部を扱う．本稿では，一つの対話状態と一つの対話行為が与えられたとき，対話状態の更新結果は一意に決まると仮定する．したがって，対話行為は対話状態を更新するコマンドと考えることができる．なお，本稿では，音声理解結果と対話行為を同じ意味で用いる．

3　課題

従来の多くの音声対話システムの談話理解では，対話状態を一意に決定していた．しかしながら，音声認識，音声理解の結果は一般に複数の候補が得られ，したがって対話状態の候補も複数得られる．対話状態を各ユーザ発話の直後に一意に決定するよりも，複

[1] 対話行為は，一般に文に対応するとされることが多いが，自然な発声が多い対話では，完全に文と認めらる発話が少ないため，文節のような小さい単位を，対話行為とみなすこともある．

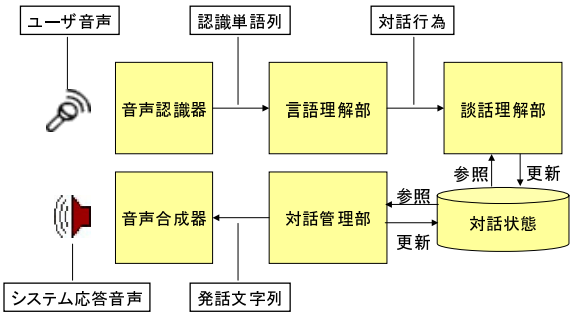

図 1: 音声対話システムの基本構成

数の候補を残しておき，後続する対話から曖昧性を解消する方が良い場合もあると考えられる．

例えば，図 2 に示す対話例は，ユーザの「2 時から」という発話が,「2 時 が」と誤認識されてしまったものである[2]．この段階では，システムは，ユーザ発話が開始時間に関する対話行為か終了時間に関する対話行為かを決定できないため，ユーザ意図の理解結果として両方の場合を保持しながら，相槌をうち，次のユーザ発話を待つ[3]．次発話の「3 時 までです」は一意に終了時間を指定する対話行為であるため，システムが保持するユーザ意図の理解結果それぞれに対して，終了時間の値が更新され，その結果，2 つのユーザ意図の理解結果が作成される．ここで，システムが，先程の対話行為が開始時間に関するものだったと理解できれば，最終的にユーザの意図を正しく理解できる．このように，音声認識の精度が悪い場合，一位の音声理解候補をそのまま信用するより，前後の対話の文脈からもっともらしいものを選ぶ方が良い．本稿では，このように複数の対話状態候補を扱う談話理解法を検討する．ユーザ発話の音声理解結果 (対話行為) も複数の候補が存在するため，対話状態と音声理解結果の組合わせからもっともらしいものを選ぶためのスコアリングが必要となる．

3.1 従来手法

複数の対話状態を扱う従来手法として，ISSS (Incremental Sentence Sequence Search)[5] が提案されている．ISSS では音声対話の現象の一つである，幾つかの音声区間にまたがる発話に対処するため，音声認識・言語理解・談話理解が統合された理解系を持ち，入力として，文だけでなく文の断片（単語，フレーズ）を受け付け，それらの入力の度，逐次的に理解結果を更新する．また，曖昧性を持つ対話行為に対応するには，複数の対話状態をスコア付きで保持することにより，ユーザ発話入力後に，最適な対話状態を決定すればよい．ISSS と n-best 音声認識結果に対する複数の音声理解結果を組み合わせた手法が提案されており [9]，談話理解精度の向上が報告されている．しかし，文献 [9] においては最適な対話状態を決定するために，設計者の直観に基づいて作成された規則を用いて対話状態のスコア付けを行っている．このため，規則の作成およびチューニングに高いコストと専門知識が必要であり，また，誤認識を考慮したルールは手動で作成するのが困難であるといった問題がある．

[2]ここでは説明の簡単のため，n-best 入力は考慮しない．

[3]ユーザ意図の理解結果は，対話状態の一要素である．また，nil というシンボルは値がないことを表す．

4 アプローチ

本稿で提案する談話理解法では，従来法 [9] と同様に，複数の対話状態候補を保持し，それらと n-best 音声認識結果から得られる複数の対話行為[4] との組合せで得られる新たな対話状態候補を，適切に順序付けることにより談話理解を行う．ただし，対話状態の順序付けに，対話コーパスから得られる統計情報を用いる点が従来法と異なる．統計情報には (1) ユーザとシステムの発話の対話行為タイプの連鎖確率，(2) 対話状態とユーザ発話の対話行為の共起関係，の二種類を用いる．それぞれについて，以下に詳しく述べる．

4.1 対話コーパスから統計情報の抽出

対話行為タイプの連鎖確率 永田らは，ユーザ発話の対話行為タイプを直前のいくつかの発話から，対話行為タイプの連鎖確率を用い，統計的に推定することで，音声認識精度を向上させた [4]．永田らの手法は，複数の対話状態候補を扱うものではないが，対話行為タイプの連鎖確率は有用であると考えられるため，連鎖確率として対話行為タイプの N-gram を利用する．具体的には，対話コーパスから得られるシステム発話と，ユーザ発話の書き起こしを，対話行為変換パーサによって対話行為に変換した後，対話行為タイプの N-gram を作成する．

対話状態と対話行為の共起確率 対話コーパスから，対話の各時点での対話状態と，その直後のユーザ発話を抽出する．ユーザ発話は対話行為に変換され，対話状態と直後の対話行為のペアを作成する．対話行為タイプの連鎖確率が対話における大まかな対話の流れを表すのに対し，本共起情報は，対話行為による対話状態の変化といった，局所的で詳細な対話の流れを表すことができる．単純な対話状態と対話行為の bigram では，対話状態に含まれる内容の複雑さのため，データがスパースになるので，対話状態が保持するユーザ意図の理解結果が対話行為によってどのように更新されるか，更新される項目はすでにシステムによって確認された項目であるかなどの共起の仕方を，8 つの 2 値 (表 1) からなる 256 のクラスで表し，それぞれのクラスの生起確率を求める[5]．

[4]一般に一つの音声認識結果からは対話行為の列が得られるため，より正確には対話行為列と呼ぶべきである．また，対話行為列中に，音響尤度比が極端に低い対話行為が含まれる場合，その対話行為を理解しない方がよい可能性があるため，その対話行為を含まない場合の対話行為列も候補に含める．

[5]一発話目はユーザが比較的自由に話す（自由発話である）ため，最初の対話状態を更新する対話行為は特別なものとして扱う必要がある．

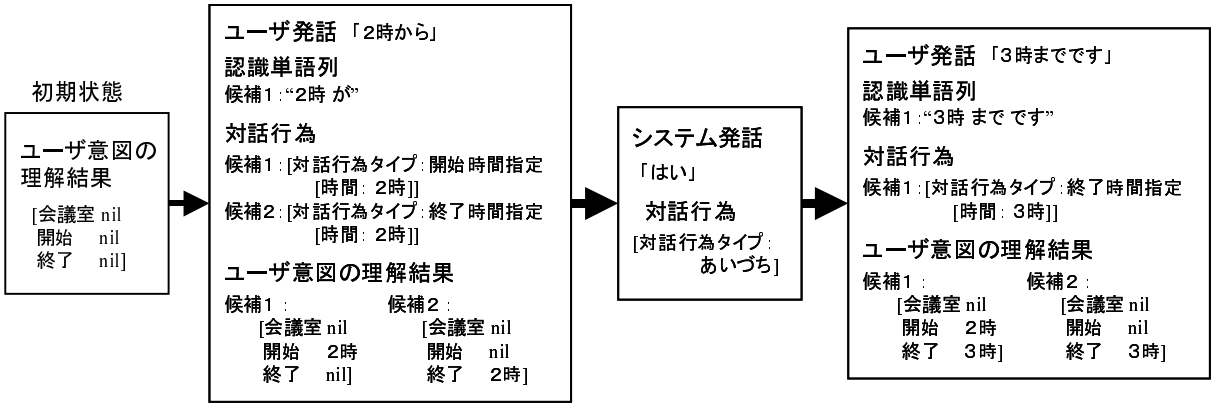

図 2: 複数の対話状態を用いたシステムにおける対話例

ユーザ意図の理解結果はスロットと呼ばれる属性値対から構成されるフレーム表現 [1] とする．

番号	説明
1.	直前に質問したスロットを更新するか
2.	確認中のスロットを更新するか
3.	これまでに確認したスロットに関する対話行為か
4.	値を持たないスロットに関する対話行為か
5.	値を持つスロットに関する対話行為か
6.	値を持つスロットに関する対話行為で，更新後も値が同じか
7.	値を持つスロットに関する対話行為で，値を変更するか
8.	最初の対話状態についてか

表 1: 対話状態と対話行為の共起の仕方

4.2 対話状態のスコアリング

対話行為の内容に基づいて対話状態を更新した場合，対話状態は一意に更新されるため，音声理解の結果である対話行為が l 個存在し，対話状態 m 個が存在した場合，$m \times l$ 個の新たな対話状態が作成される．このとき，新しく作成される対話状態 S_{t+1} のスコアを以下のように定義する．

$$S_{t+1} = S_t + \alpha \cdot s_{nbest} + \beta \cdot s_{ngram} + \gamma \cdot s_{col}$$

ここで，S_t は更新前の対話状態のスコアであり，s_{nbest} は対話行為の n-best 順序に関するスコア，s_{ngram} は対話行為タイプの連鎖確率に関するスコア，s_{col} は対話状態と対話行為の共起確率に関するスコア，α，β，γ はそれぞれ重み係数である．

4.3 対話状態列の順序付け

作成された対話状態は前項で算出される優先度によって順序づけられ，最も優先度の高い対話状態をその時点での最尤な対話状態とし，それに基づいてシステム応答が行われる．対話状態数の最大値 (対話状態ビーム幅と呼ぶ) を，スコアの低い対話状態を捨てることにより一定に保ち，リアルタイムで理解を行う．

5 実験

5.1 対話コーパスの作成

音声対話システムと人との対話コーパスを作成した．対話データ収集は音声対話システムを過去に使ったことのないユーザを対象に，簡易防音を施した部屋で行われた．システムは音声対話システム作成ツールキット WIT[6] を用いて作成した会議室予約システムである．被験者は実験者の指示に基づき，日付，開始時間，終了時間，会議室名をシステムに伝え，会議室を予約した．音声認識エンジンとして Julius3.1p[3] を付属の音響モデルと共に用いた．言語モデルは，受付可能なフレーズから作成した単語 trigram である．システム応答の音声合成には NTT サイバースペース研究所の FinalFluet[7] を用いた．本システムは，システム設計者が作成した談話規則に基づき理解をおこなう．一被験者につき 16 対話収録し，その結果 15 名（男性 10 名，女性 5 名）の被験者から 240 の対話データを収集した．タスク達成に 3 分以上かかった対話は失敗とし，その場で対話を打ち切った[6]．

対話コーパスからの統計情報の抽出 それぞれの対話セッションに関して，ユーザ発話，システム発話の開始時間と終了時間，およびユーザ発話前後のシステムの対話状態を対話記録（ログ）に保存した．ユーザ音声とシステム音声は録音され，すべてのユーザ音声は書き起こされた．書き起こしとシステム発話記録から，CMU-Cambridge Toolkit[2] を用い，対話行為タイプの trigram を作成し，対話状態と，直後のユーザ音声の書き起こしから，対話状態と直後の対話行為の共起確率を求めた．共起の仕方は全部で 17 パタンあった．

[6] タスク達成率は 78.3%(188/240) であった．

5.2 提案法を用いたシステムによる対話実験

提案手法の有効性を検証するため，本アプローチを談話理解部に組み込んだシステムを作成し，対話実験をおこなった．音声認識エンジンは Julius3.3p1 であり，5-best の認識単語列を出力する．言語モデルは，対話コーパスの書き起こしより作成した単語 trigram を用いた．今回は，s_{nbest} は n-best 順序の逆数の対数，s_{ngram} は対話行為 trigram の対数尤度，s_{col} は共起確率の対数とし，実験的に $\alpha = \beta = \gamma = 1$ とした[7]．対話状態のビーム幅は 15 であった．一被験者につき 16 対話収録し，その結果 16 名（男性 7 名，女性 9 名）の被験者から 256 の対話データを収集した．タスク達成に 3 分以上かかった対話を失敗としたとき，タスク達成率は 77.3%(198/256) であった[8]．また，収集した対話データ (220 対話，1541 発話) の中から，スコアリングにおいて 2 位以下であった対話状態が，次発話により 1 位になった回数を数え上げたところ全部で 120 回 (7.79%) あった．

5.3 複数対話状態の有効性の検証実験

文脈を複数保持することの有効性を検証するため，音声認識候補の入力を 1-best に固定し，対話状態のビーム幅を 1 にしたものと 30 にした場合について，対話実験をおこなった[9]．一被験者につき 16 対話収録し，その結果 28 名（男性 4 名，女性 24 名）の被験者から 448 の対話データを収集した．対話状態のビーム幅が 1 のシステムと 30 のシステムのタスク達成率はそれぞれ 88.3%，91.0%であった．収録されたデータにおけるタスク達成時間を，同一被験者の，同一タスク[10]での対話のペアにおいて，ビーム幅 1 のシステムとビーム幅 30 のシステムの平均タスク達成時間は，それぞれ 107.66 秒，95.86 秒であり，pairwise の t 検定により統計的に有意な差があった (p=0.0043)．

6 考察

談話理解部作成のコスト 本アプローチを用いた対話システムにおけるタスク達成率が最高で 91.0%(1-best 入力，対話状態ビーム幅 30 の場合) と高いことから，本アプローチは，人手による談話理解規則の作成を十分削減する効果があるといえる．統計ベースの談話理解部の作成には，初期システムを構築し，対話コーパスを作成する必要があるが，一度作成すれば，専門知識を必要せずに，談話理解部の構築が可能である．

複数対話状態の有効性 複数対話状態の有効性の検証実験の結果から，対話状態を複数保持することが，より高精度な談話理解に寄与することが示された．

複数音声認識候補による弊害 1-best のみを用いる場合の方が，5-best を用いる場合よりも，タスク達成率においてよい値を示した．このことは，n-best 次数を増やすことによって，対話状態に共起しやすい間違った認識結果が n-best の下位に出現することが原因と考えられる．

7 結論と今後の課題

本稿では，音声対話システムの談話理解において、複数の文脈と複数の音声認識候補をから得られる複数の理解候補を統計モデルを用い順序付ける手法を提案した．対話実験の結果，談話理解部作成のコスト削減が実現され，複数の対話状態を保持することの優位性も示された．

今後の課題としては，対話行為タイプの連鎖確率と対話状態と対話行為の共起確率以外の統計情報の利用や，s_{nbest}, s_{ngram}, s_{col} を足し合わせる際の重みの最適化が挙げられる．

謝辞

本研究において，有益なアドバイスを頂いた村瀬洋メディア情報研究部長ならびにマルチモーダル対話研究グループの諸氏に感謝します．

参考文献

[1] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd. GUS, a frame driven dialog system. *Artif. Intel.*, 8:155–173, 1977.

[2] P. Clarkson and R. Rosenfeld. Statistical language modeling using the CMU-Cambridge Toolkit. In *Proc. Eurospeech*, 1997.

[3] A. Lee, T. Kawahara, and K. Shikano. Julius – an open source real-time large vocabulary recognition engine. In *Proc. Eurospeech*, pp. 1691–1694, 2001.

[4] M. Nagata and T. Morimoto. First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203, 1994.

[5] M. Nakano, N. Miyazaki, J. Hirasawa, K. Dohsaka, and T. Kawabata. Understanding unsegmented user utterances in real-time spoken dialogue systems. In *Proc. 37th ACL*, pp. 200–207, 1999.

[6] M. Nakano, N. Miyazaki, N. Yasuda, A. Sugiyama, J. Hirasawa, K. Dohsaka, , and K. Aikawa. WIT: A toolkit for building robust and real-time spoken dialogue systems. In *Proc. SIGDIAL*, pp. 150–159, 2000.

[7] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima. A Japanese TTS Syntem Based on Multi-form Units and a Speech Modification Algorithm with Harmonics Reconstruction. *IEEE Transactions on Speech and Processing*, 9(1):3–10, 2001.

[8] 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, vol.17 No.3:200–207, 2002.

[9] 宮崎, 中野, 相川. n-best 音声認識と逐次理解法によるロバストな音声理解. 情処研報 SIG-SLP-40, pp. 121–126, 2002.

[7] 対数の底は 10 とした．

[8] 本システムは，膨大な対話状態数を考慮するため，理解に比較的時間がかかった．

[9] ビーム幅を大きく設定しすぎると，計算量が増えるとともに，スコアの低い理解結果が悪影響を及ぼす可能性があるため，経験的に 30 とした．

[10] 会議室の予約内容のパタンを指す．