音声対話システムにおける談話理解部のオフライン評価法

東中竜一郎　中野 幹生

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

1　はじめに

音声対話システムがタスクを達成するためには，ユーザの意図を正しく理解する必要がある．ユーザの意図は一度の発話で理解できるとは限らない．そのため，文脈を考慮した発話の理解，すなわち談話理解が必要となる．音声対話システムの談話理解部を評価するためには，ユーザとのやり取りを扱うことから，実ユーザによる対話実験を行う必要があり，コストが高いという点で問題があった．

本稿では，音声対話システムにおける談話理解部を，実ユーザによる対話実験を用いずに，対話コーパスを用いて評価する手法 (オフライン評価法) を提案する．本手法により，談話理解部の評価コストを削減することが可能になれば，今後の談話理解部の改善に有益であると考えられる．

本稿では，まず音声対話システムにおける談話理解について簡単に触れた後，扱うべき問題点を説明する．その後，我々のアプローチを示し，対話コーパスによる談話理解部の評価実験とその結果について述べる．最後にまとめと今後の課題を述べる．

2　音声対話システムにおける談話理解

図 1 に音声対話システムの基本的な構成を示す．音声対話システムは，主に音声認識部・言語理解部・談話理解部からなる発話理解部と，内容生成部・表層生成部・音声生成部からなる発話生成部から構成される [10].

音声認識部は，ユーザ発話音声を受け取り，認識単語列を出力する．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，対話行為と呼ばれる発話意図，発話内容の理解結果を出力する．単独のユーザ発話から発話意図，発話内容を得ることを音声理解と呼ぶ．

談話理解部は，音声理解の結果である対話行為と，現時点での対話状態を入力とし，対話状態を更新する．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，ユーザ発話履歴，システム発話履歴などを含む．

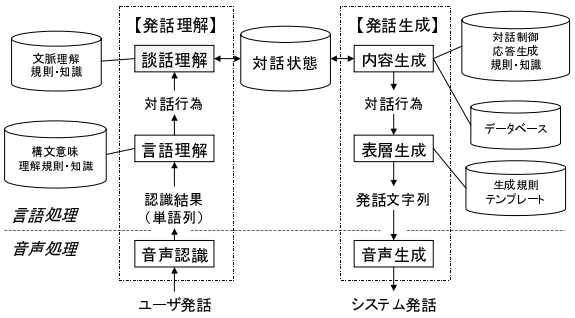

図 1: 音声対話システムの基本構成

発話生成部は，更新された対話状態を元に次発話を生成する．適切に応答するには，システムは出来るだけ適切に対話状態を更新する必要がある．

3　問題

音声対話システムの各構成部は，特定のドメインにおいてタスクを達成するためにドメイン依存知識を保持している．音声認識部であれば，認識語彙や言語モデルがドメイン依存知識である．ドメイン依存知識を個別に改善すると，他の構成部に悪影響を及ぼすことも考えられる．そのため，改善にあたってはシステム全体のパフォーマンスを考慮する必要がある．全体のパフォーマンスとは，例えば，タスク達成率やタスク達成時間などで計測される，システム全体の良さである [9, 1, 4].

システム全体のパフォーマンスを考慮して評価するためには，実際に，評価したい構成部を組み込んだシステムを用いて対話実験を行う必要がある．しかし，被験者を雇わなくてはいけないため，コストが高いという問題がある．また，パフォーマンスは被験者の個人差によるところが大きく，少ない被験者数では正当に評価することが困難である．

そこで，音声認識部や言語理解部は一般的に，対話コーパスを用いて評価を行う．以下に手続きを示す．

1. ある時点でのシステムを用いて，対話音声を収録する．
2. ユーザの発話音声に対し，書き起こしや対話行為のタグ付けを行う．

3. 収録された音声に対する Word Error Rate(WER) や Concept Error Rate(CER) を用いて評価する [3].

対話コーパスによる評価は，実ユーザによる対話実験の評価と一致するとは限らないが，音声認識部や言語理解部に関しては，対話コーパスに対する個別のパフォーマンスが，システム全体のパフォーマンスと直感的に相関があるとして，広く受け入れられている．

談話理解部の評価に関しては，ユーザとのやり取りを扱うことから，実ユーザによる対話実験が必要とされており，コストが高い [2]．対話コーパスを用いて，談話理解部を評価することが可能であれば，今後の改善に有用であると考えられる．なお，文献 [11] は対話コーパスを用いて談話理解部を評価しているが，システム全体のパフォーマンスとの相関は考慮していない．

4 アプローチ

談話理解部は，ユーザとシステムのやり取りから，迅速かつ正確にユーザの意図を理解することが望ましい．対話コーパスを用いた評価を考える場合，コーパスにおける発話列を順に理解し，なるべく短いユーザとシステムのやり取りから正解のユーザ意図が理解できる談話理解部が，性能が高いと考えることができる．図 2 に，天気情報案内システムにおける対話を例に用いたオフライン評価の概要を示す．評価の流れは次のようになる．

1. ある時点でのシステムを用い，対話コーパスを収録する．コーパスには，対話収録時のシステム発話とユーザ発話 (認識結果を含む)，対話収録時のユーザ意図の理解結果が含まれ，ユーザ発話毎にユーザ意図の正解がタグ付けされているとする．

2. $i = 0$ とする．

3. 評価対象の談話理解部を用いて，対話収録時の i 番目のユーザ発話を理解する．

4. 談話理解部が出力するユーザ意図と，正解のユーザ意図が一致するかをチェックし，一致すればステップ 5 へ進む．一致しなければ i 番目のシステム発話を理解し，$i = i + 1$ とし，ステップ 3 に戻る．

5. i は推定ターン数であり，この値を比較することによって，談話理解部の評価を行う．

図 2 における，談話理解部 A と談話理解部 B は，それぞれ $i = 2$，$i = 3$ において正解のユーザ意図を理解した．そのため，どちらの談話理解部も収録時のものより優れており，A の方が B よりも性能が高いといえる．

もちろん，談話理解部が異なればシステムの確認発話内容も変化すると考えられる．例えば，談話理解部 B におけるユーザ発話 U1 のユーザ意図の理解結果は < 今日, $null$，天気 > である．そのため，実際にはシステムの次発話は「明日の京都の天気についてですね？」(S2) とはならないと考えられる．

しかし，システムがユーザ意図の理解結果の可能性をすべて保持しながらユーザ発話を理解したとすると，選択可能な複数のユーザ意図の理解結果の中から，システムは < 今日, $null$，天気 > を選択したと考えることができる．その場合，システムは理解結果候補として < 明日，京都，天気 > も保持しているため，談話理解部の候補の選び方によっては S2 を発話することもあり得る．結局，システムはユーザ意図の正解にたどり着くまで，誤った確認等を繰り返すのであるから，談話理解部はそれらの情報もユーザ意図の確定に利用する必要がある．従って，誤った確認をもうまく扱える談話理解部はよい談話理解部といえる．

次節では，実際に対話コーパスを収録し，複数の異なる談話理解部のオフライン評価を行った実験について述べる．

5 実験

5.1 対話コーパス

対話コーパスとして，天気情報案内システムにおける 120 対話を収録した．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，被験者は実験者の指示に基づき，都道府県名か市名，日付，情報種別（天気・気温・降水確率のいずれか）をシステムに伝え，天気情報の提供を受けた．

音声認識エンジンとして Julius3.3[6] 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[8] を用いた．ユーザ発話はすべて書き起こされ，対話行為および，発話後のユーザ意図の正解がタグ付けされた．

5.2 談話理解部の作成

コーパスベース談話理解法 [5] を用いて，複数の談話理解部を用意した．コーパスベース談話理解法とは，コーパスから得られた統計情報を理解に用いる手法であり，統計情報をどの程度の重みで適用するかによって，理解の仕方を変動させることが可能である．重みを大きくすることで，よりコーパスに沿った理解を行うようになる．

統計情報の重みは二つのパラメタにより表現される．一つは対話行為の連鎖確率に関する β であり，も

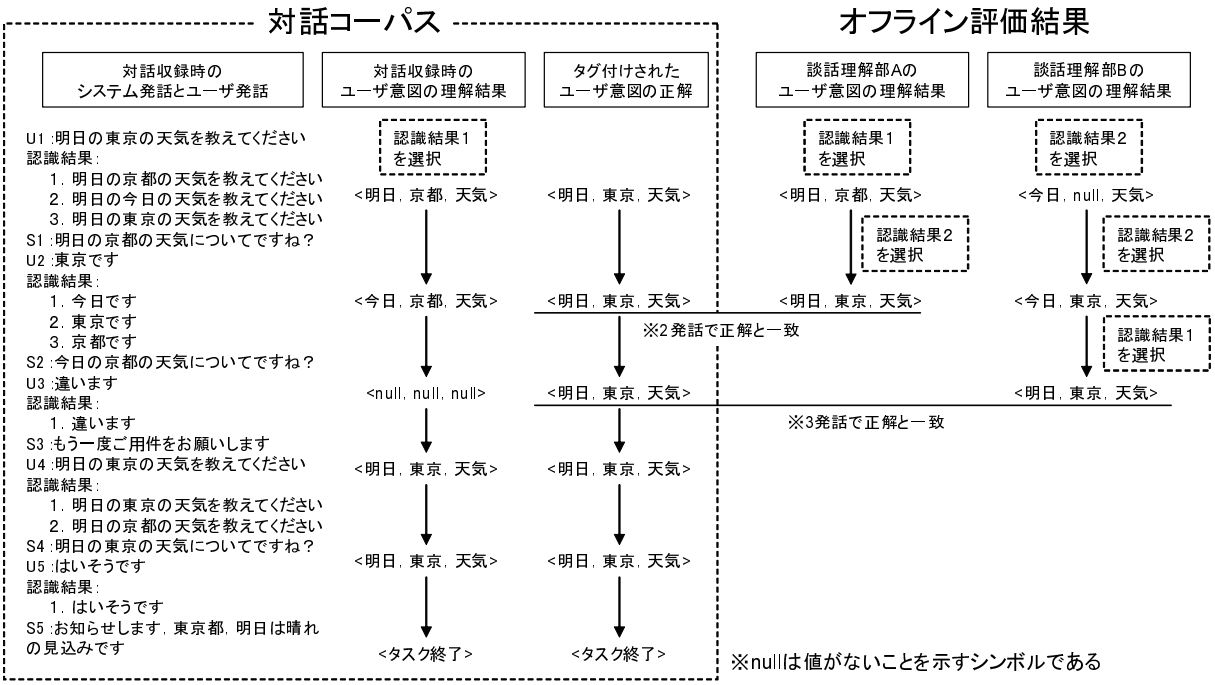

図 2: オフライン評価の概要

う一つはユーザ意図の変化の仕方に関する γ である．今回，β と γ をそれぞれ 0.0 から 2.0 まで 0.1 ずつ変動させ，組み合わせにより，全部で 441(21×21) の談話理解部を作成した．$\beta = 0, \gamma = 0$ の場合は，統計情報を全く用いない談話理解部であり，常に音声認識結果が一位のものを優先して理解に用いる．

なお，統計情報は対話コーパスのうち，ランダムに選択した 30 対話 (約 500 発話) から抽出した．

5.3 オフライン評価実験

統計情報の抽出に用いた 30 対話を除いた，90 対話を対象としてオフライン評価実験を行った．441 の談話理解部を用いて 90 対話を評価したため，全部で 39690 対話をオフラインで行ったことになる．それぞれの対話について，正解のユーザ意図を保持するまでのユーザ発話数を求めた．また，最後までユーザ意図の正解に達することができなかった対話は，タスク失敗としてカウントした．

5.4 結果

図 3 は 441 の談話理解部それぞれのタスク達成率を元に作成した等高線図である．ここでいうタスク達成率とは，オフライン評価において，収録時の談話理解部より早く（または同時に）ユーザの意図に到達できた率である．色が濃いほどタスク達成率が高いことを示している．最もタスク達成率が高い談話理解部は，$\beta = 0.4, \gamma = 0.1$ のものであり 95.1%であった．また，$\beta = 0.0, \gamma = 0.0$ の場合は，93.2%であった．音声認識率が高い場合にはこの 2 つのパラメタセットで全く差がない．しかし，音声認識率が低い場合には，3 対話分の差がみられた．全体の傾向としては β と γ が 0 周辺の時にタスク達成率が高く，γ を大きくするに従って，タスク達成率が低下する．

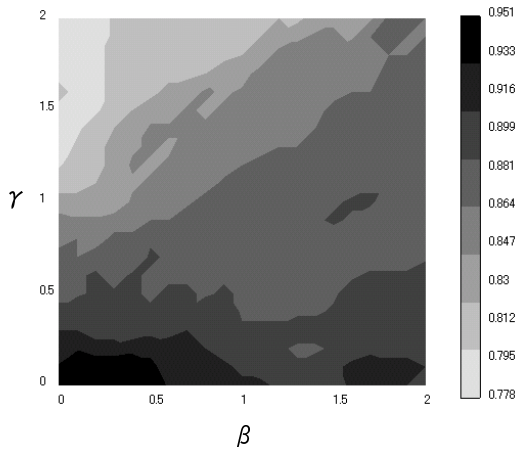

図 3: タスク達成率の等高線図

図 4 はタスク達成した対話のみを対象とした時の，ユーザ意図の正解を得るまでの平均発話数の等高線図である．色が濃いほど平均発話数が多いことを示している．最も平均発話数が少ない談話理解部は，$\beta = 0.1, \gamma = 1.0$ のものであり 3.53 であった．また，$\beta = 0.0, \gamma = 0.0$ の場合は，3.81 であった．全体的に，β が小さく γ が大きい時に平均発話数が少ない．

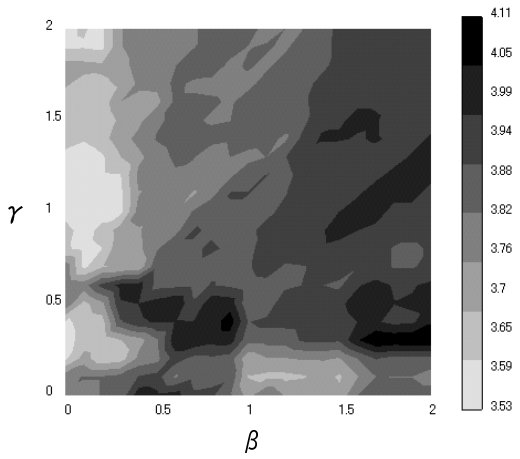

図 4: ユーザ意図の正解を得るまでの平均発話数の等高線図 (タスク達成した対話のみが対象)

図 3 と図 4 の比較から，統計情報の重みが小さい方が，タスク達成率は高く，タスクが達成が可能な対話に関しては，統計情報の重みが有効に機能していると考えられる．

6 まとめと今後の課題

本稿では，音声対話システムにおける談話理解部を，実ユーザによる対話実験を用いずに，対話コーパスを用いて評価する手法 (オフライン評価法) を提案し，複数の談話理解部をオフライン評価により比較した．オフライン評価における評価が実ユーザを用いた対話実験による評価と一致するかどうかは，実ユーザを用いた対話実験により検証される必要がある．今後検証を行う予定である．

また，オフライン評価のほかにシミュレーション対話による評価も考えられる．シミュレーション対話は，ユーザの振る舞いを真似る模擬ユーザと，評価したい理解部を持つシステムとを，コンピュータ上で擬似的に対話させる手法である [7]．シミュレーション対話の場合，模擬ユーザの振る舞いの妥当性が問題となるが，今後検討していく．

謝辞

本研究において，有益なアドバイスを頂いた牧野昭二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] G. Damnati. Evaluating speech recognition in the context of a spoken dialogue system : critical error rate. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 280–283, 2002.

[2] E. A. Filisko. A context resolution server for the galaxy conversational systems. Master's thesis, Massachusetts Institute of Technology, 2002.

[3] J. Glass, J. Polifroni, S. Seneff, and V. Zue. Data collection and performance evaluation of spoken dialogue systems: The MIT experience. In *Proc. ICSLP*, pp. IV:1–4, 2000.

[4] R. Higashinaka, N. Miyazaki, M. Nakano, and K. Aikawa. Evaluating discourse understanding in spoken dialogue systems. In *Proc. Eurospeech*, pp. 1941–1944, 2003.

[5] R. Higashinaka, M. Nakano, and K. Aikawa. Corpus-based discourse understanding in spoken dialogue systems. In *Proc. 41st ACL*, pp. 240–247, 2003.

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

[7] R. Lopez-Cozar, A. D. la Torre, J. C. Segura, and A. J. Rubio. Assessment of dialogue systems by means of a new simulation technique. *Speech Communication*, 40(3):387–407, 2003.

[8] 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.

[9] M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems.*, 2000.

[10] 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 No.3:200–207, 2002.

[11] 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. 音声言語情報処理 (SLP-40), pp. 121–126, 2002.