自 然 言 語 処 理 149−15
音声言語情報処理 41−6
（２００２．５．２４）

逐次理解を行う音声対話システムの発話理解評価法

東中竜一郎　宮崎昇　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
〒 243-0198 神奈川県厚木市森の里若宮 3-1
{rh,nmiya,nakano}@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp
http://www.brl.ntt.co.jp/cs/dug/

概要

従来、文脈を考慮しないユーザ発話の理解には、CER (Concept Error Rate) または Keyword Error Rate が評価尺度として用いられてきた。しかし、CER では、文脈を考慮してユーザ発話を理解する音声対話システムの発話理解を評価することはできない。本稿では、逐次理解を行う音声対話システムにおいて、対話の各時点での理解状態や理解状態の更新の仕方から得られるさまざまな指標を組み合わせ、システム全体のパフォーマンスとの相関の高い尺度を求めることによって、文脈を考慮した発話理解の評価尺度を作成する手法を提案する。システムのパフォーマンスにはタスク達成時間を用い、理解状態に関して得られたさまざま指標を説明変数、タスク達成時間を被説明変数として、重回帰分析を行った。得られた重回帰式は、比較的よい性能を示し、評価尺度としての有効性を示した。

キーワード: 音声理解, 発話理解, 文脈理解, 音声対話システム

A Method for Evaluating Incremental Utterance Understanding in Spoken Dialogue Systems

Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, Kiyoaki Aikawa

NTT Communication Science Laboratories, NTT Corp.
3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan

Abstract

In single utterance understanding, which does not include discourse understanding, the concept error rate (CER), or the keyword error rate, has been widely used as an evaluation measure for utterance understanding. However, the CER cannot be used for evaluating systems that understand user utterances based on previous user utterances. In this paper, we propose a method for evaluating incremental utterance understanding, which involves speech recognition, language understanding and discourse processing in spoken dialogue systems, by finding a measure that correlates closely with the system's performance based on dialogue states and their way of update. We defined dialogue performance by task completion time, and performed a multiple linear regression analysis using task completion time as the explained variable and various metrics concerning dialogue states as explaining variables. The obtained multiple regression model fits comparatively well and shows validity as an evaluation measure.

keywords: speech understanding, utterance understanding, discourse understanding, spoken dialogue system

1 はじめに

音声認識・合成技術が発達するにつれ、商用サービスが始まるなど、音声対話システムが注目を浴びつつある。音声対話システムには 2 種類ある。一つは、単一のユーザ発話を処理し、文脈を考慮せずに応答を返すもので音声応答システムと呼ばれる。もう一つは複数のユーザ発話、システム応答を文脈の中で処理するものである。本稿の議論の対象である後者のシステムでは、システムはユーザ発話を受け付ける度に、適切に理解状態を更新する必要がある。ここで言う理解状態とは、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す。例えば、各時点までのユーザ発話を処理した結果である理解結果や、各時点での話題、確認済み事項を保持する確認フラグといった、談話に関する情報が含まれる。

文脈を考慮せずにユーザ発話を理解する単発話理解では、CER (Concept Error Rate) または Keyword Error Rate が従来、発話理解の評価尺度として用いられてきた。しかし、CER では、対話履歴を考慮してユーザ発話を理解する音声対話システムの発話理解を評価することはできない。なぜなら、ユーザ発話の理解はユーザ発話前の理解状態に影響を受けるからである。また、発話理解の評価にどのような情報を用いれば良いのかも自明ではない。例えば、理解状態そのものの正しさや理解状態の更新の仕方の正しさといったものが考えられるが、それらをどのように組み合わせて発話理解の評価に用いればよいか明らかではない。

特に我々は発話理解法として ISSS (Incremental Sentence Sequence Search) [1] を提案してきた。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・言語処理・文脈処理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、逐次的に理解状態を更新する。もし、入力が曖昧性を持つ場合は、複数の文脈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は一意に理解状態を決定できる。ISSS を用いた理解を逐次理解と呼ぶ。逐次理解を行う音声対話システムでは、以前の理解状態をもとに次理解状態を決定するため、文脈を考慮に入れた発話理解の重要性が高い。

本稿では、逐次理解を行う音声対話システムにおける発話理解の評価尺度を、対話の各時点での理解状態や理解状態の更新の仕方から得られるさまざまな指標を組み合わせ、システム全体のパフォーマンスとの相関の高い尺度を求めることによって作成する手法を提案する。具体的には、音声対話システムのパフォーマンスをタスク達成時間とし、理解状態に関して得られる指標を説明変数、タスク達成時間を被説明変数として、重回帰分析を行った。得られた重回帰式は、比較的良い性能を示し、評価尺度としての有効性を示した。

2 章で、本研究の課題について詳細に述べ、3 章で、本研究のアプローチと理解状態に関するさまざまな指標について説明する。4 章で、実システムを用いた対話実験と、対話データの分析結果について述べ、最後に、結論と今後の課題を述べる。

2 課題

複数のユーザ発話を処理し、逐次的に理解状態を更新するようなシステムでは、初期理解状態（通常空）はユーザ発話ごとに次状態に更新されていく。例えば、初期理解状態はユーザ発話 1 の処理後、理解状態 A になる。理解状態 A はユーザ発話 2 の処理後、理解状態 B に更新される (図 1)。

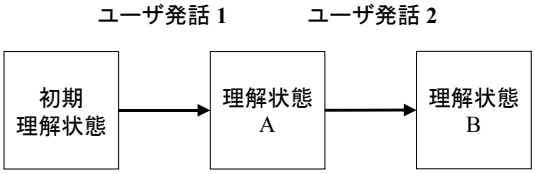

図 1: 理解状態の更新

理解状態を考慮せずに発話理解を行う音声応答システムでは、CER が発話理解の評価尺度として使用できるが、逐次的に理解状態を更新するようなシステムでは、以前の理解状態が更新の仕方に影響を与えるため、CER は使用できない。具体的に言うと、理解状態 B は初期理解状態の時点でユーザ発話 2 を処理しても得られない。理解状態 A という状態において、ユーザ発話 2 が処理されてのみ、理解状態 B に更新される。また、CER で発話理解の評価が可能なシステムでは、ユーザ発話を処理した後の正解の理解状態は明らかであるが、過去の理解状態を考慮すると、正解の理解状態がどういうものであるかが自明ではなくなる。例えば、理解状態 B はその時点での理解状態としては間違っているかもしれないが、部分的に正しく更新されたかもしれない。

ユーザが「3 時から」（ポーズ）「4 時まで」と言った場合を考えよう。「3 時から」が「2 時から」に誤認識されたとすると、次の「4 時まで」の処理後も誤

認識の結果が残ってしまう (図 2)。こういった場合、時点時点での理解状態の正しさか、理解状態の更新の正しさか、どちらをどのように発話理解の評価に用いればよいのか自明でない。また、このようなシステムにおける発話理解を評価する手法は現在のところ提案されていない。

今後、より正確に理解状態を更新するシステムの構築や、談話処理ルールの自動獲得などに発展させるためにも、文脈を考慮した音声対話システムの発話理解の評価尺度が必要である。

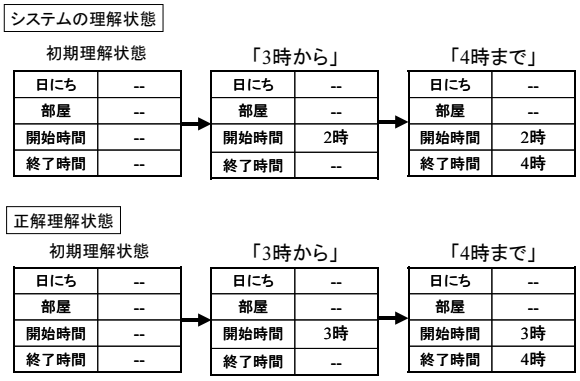

図 2: 理解状態更新の例：システムの理解状態と対応する正解理解状態

3 アプローチ

課題に対処するため、音声対話システムの各時点での理解状態と理解状態の更新について考えられる指標を多数列挙し、それら指標の値と音声対話システム全体のパフォーマンスとの相関が高くなるような尺度を、重回帰分析を用いて求める。

3.1 理解状態に関する指標

ある時点における理解状態は、ユーザ発話処理後、次の理解状態に更新されるが、この一連の理解状態更新の流れを理解単位と定義し、ユーザ発話前の理解状態を初期理解状態、ユーザ発話後の理解状態を最終理解状態と呼ぶ。システムの理解状態はフレーム表現で表される。同表現は従来の音声対話システムなどでよく用いられる表現であり [2]、フレームはスロットと呼ばれる属性-値対により構成される。初期理解状態、最終理解状態はそれぞれ初期フレーム、最終フレームとも呼ばれる。ユーザ発話によって理解状態は適宜更新され、更新された理解状態をもとに、システムは適宜応答を生成し、次のユーザ発話を待つ。

対話の各時点での理解状態と理解状態の更新についての指標に関してであるが、まず理解単位の理解状態に関する指標を考案し、複数の理解単位で構成される対話に関しても、同指標を用いることにする。値としては、理解単位から算出された指標の値の相加平均を用いる。

理解単位における理解状態に関する指標は、システムがユーザ発話を処理した後の理解状態を仮説フレームとし、その時点における正解（人手で作成）を正解フレームとする時、正解フレームと仮説フレームを比較することによって得る。

比較には 2 種類ある。一つは、時点時点での理解状態の値の直接比較である。フレーム内のそれぞれのスロットが値を持つか、値が異なるか、同じであるかを比較する。この比較により、仮説フレームのスロットそれぞれについて、4 種類のラベルを付与することができる (表 1)。

ラベル	名前	説明
C	正解	仮説フレームのスロットと正解フレームのスロットの値が同じ。
I	挿入誤り	仮説フレームのスロットが値を持っており、正解フレームのスロットが値を持っていない。
D	削除誤り	仮説フレームのスロットが値を持っておらず、正解フレームのスロットが値を持っている。
S	置換誤り	仮説フレームのスロットと正解フレームのスロットの値が異なる。

表 1: 仮説フレームのスロットそれぞれに対するラベリング

もう一つの比較は、仮説フレーム、正解フレームの初期フレームからの変化同士を比較するものである。この比較により、仮説フレームのそれぞれのスロットについて、5 種類のラベルを付与できる (表 2)。

上記 9 種の仮説フレームに対するラベルを用い、理解単位に対し 9 つの指標を得る。前述のように、対話全体を表す指標に関しても同指標を用い、対話に関するそれぞれの指標の値は、理解単位におけるそれぞれの指標の相加平均を用いる。また、10 番目の指標として、スロット正解率が 100% であった理解

ラベル	名前	説明
CU	更新正解	仮説フレームにおけるスロットが正解フレームと同様に正しく更新された。
CL	非更新正解	仮説フレームにおけるスロットが正解フレームと同様に正しく更新をされなかった。(更新しないのが正解)
UD	更新削除	正解フレームにおけるスロットは変化したが、仮説フレームのスロットは変化しなかった。
UI	更新挿入	仮説フレームにおけるスロットは変化したが、正解フレームのスロットは変化しなかった。
US	更新置換	仮説フレームにおけるスロットがそれぞれ異なる値に更新された。

表 2: 仮説フレームのスロットそれぞれの変化に対するラベリング

単位の全理解単位数における割合を用いる。以下に指標を列挙する[1]。

1. **スロット正解率**

$$\frac{C}{\text{総スロット数}}$$

2. **挿入誤り率**

$$\frac{I}{\text{総スロット数}}$$

3. **削除誤り率**

$$\frac{D}{\text{総スロット数}}$$

4. **置換誤り率**

$$\frac{S}{\text{総スロット数}}$$

5. **スロット誤り率**

$$\frac{\text{誤りスロットの合計数}}{\text{総スロット数}} = \frac{I+D+S}{\text{総スロット数}}$$

6. **スロット更新精度**

$$\frac{\text{正しく更新されたスロット数}}{\text{仮説フレームにおいて更新されたスロット数}} = \frac{CU}{CU+US+UI}$$

[1]指標の式における C,I,D,S,CU,CL,UI,UD,US はそれぞれ同ラベルを付与されたスロット数を指す。

7. **更新挿入誤り率**

$$\frac{\text{仮説フレームにおいて更新されたスロット数}}{\text{正解フレームにおいて更新されなかったスロット数}} = \frac{UI}{CL+UI}$$

8. **更新削除誤り率**

$$\frac{\text{仮説フレームにおいて更新されなかったスロット数}}{\text{正解フレームにおいて更新されたスロット数}} = \frac{UD}{CU+US+UD}$$

9. **更新置換誤り率**

$$\frac{\text{仮説フレームにおいて正しく更新されなかったスロット数}}{\text{正解フレームにおいて更新されたスロット数}} = \frac{US}{CU+US+UD}$$

10. **音声理解率**

$$\frac{\text{スロット正解率が 100\%である理解単位数}}{\text{全理解単位数}}$$

3.2 音声対話システムのパフォーマンス

本研究の対象である音声対話システムはタスク達成型のものであるため、タスク達成に要した時間をパフォーマンスの尺度として用いる。アンケート調査などで得られるユーザ満足度も尺度として考えられるが、タスク達成時間とユーザ満足度の間に関連が高いことも指摘されており [3]、タスク達成時間をパフォーマンスの尺度として用いることは妥当であると考えられる。

なお、タスク達成時間はシステム応答の仕方（対話戦略）に影響される。発話理解とタスク達成時間の関係のみに着目する場合、複数の対話戦略を用い、その影響を排除する必要がある。また、タスク達成時間はタスクの内容にも依存する。認識しにくい単語をタスクに含むものであると、より時間がかかる。よって、タスク達成時間はタスクと対話戦略によって正規化する必要がある。

4 実験

4.1 データ収集

これまでに述べた対話に関する 10 指標の値とタスク達成時間の関連を調べるため、対話実験を行い対話データを収集し、分析した。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。データ

収集に用いた音声対話システムは音声対話システム作成ツールキット WIT[4] を用いて作成した。タスクドメインは会議室予約で、被験者はインストラクションに基づき、1 つか 2 つの日付において、1 つか 2 つの会議室を、ある時間からある時間まで予約する。タスク（予約内容）は 5 パターン用意した。音声認識エンジンとして Julius3.1[5] 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 N-gram である。システム応答に用いた音声合成エンジンには Final Fluet[2][6] を用いた。システムの認識語彙数は 160 で、それぞれカテゴリと意味素性を与えられ、システムの辞書に登録された。構文解析規則数は 18 で、言語・文脈処理ルール数は 38 であった。

システムの理解状態にはフレーム表現を用い、ドメイン依存の 6 つのスロット（ドメインスロット）と、談話の進行に関係する情報（話題、確認フラグ、ユーザの直前の動作など）を保持する 3 つのスロット（談話スロット）から構成される。談話スロットのラベリングが困難なことから、分析はドメインスロットのみを対象にした。また、複数の対話戦略を用いる必要があるため、今回 2 つの対話戦略を用意した。一つはシステムが予約に必要な情報をすべて得るかユーザが明示的にシステム応答を要求するまでユーザ発話を受け付けるもので、もう一つはユーザ発話を受け付ける度にその内容に関して確認を行うものである。

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理解状態が対話記録（ログ）に保存された。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。一被験者につき 10 対話収録し (5 つのタスクパターン、2 つの対話戦略)、その結果 18 名（男性 9 名、女性 9 名）の被験者から 180 の対話データを収集した。バージインを除く総発話区間数は 3595 であった。タスク達成に 5 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。その結果タスク達成率は 63.6%(112/176[3]) であった。なお、タスク達成時間を用いることのできないタスク達成失敗の対話は分析対象外とした。

4.2 正解フレーム

正解フレームを求めるためには人手によるラベリングの多大な労力が必要である。そのため、我々はまず初期フレームと書き起こしを入力として、最終フレームを出力するようなシミュレーションシステムを構築し、その最終フレームを人手で修正し正解フレームとすることで、ラベリングの手間を大幅に軽減した。

4.3 結果

最終的に、ログ、書き起こしに不良があった対話を除き、残った 108 対話のログを分析に用いた。理解状態に関して得られた 10 指標の値を説明変数、タスクパターンと対話戦略によって正規化されたタスク達成時間を被説明変数 (Y) として、変数増加法によるステップワイズ回帰分析を行った。最終的に変数として 10 指標のうち 7 指標が用いられた。

結果、次の重回帰式が得られた。

$$\begin{aligned} Y \quad = \quad & -4.19 - 12.49x_1 + 12.77x_2 - 0.03x_3 \\ & -17.74x_4 + 4.54x_5 + 2.11x_6 + 2.98x_7 \quad (1) \end{aligned}$$

ここで Y は タスク達成時間の予測値，x_1 は 挿入誤り率，x_2 は 置換誤り率，x_3 は スロット更新精度，x_4 は 更新挿入誤り率，x_5 は 更新削除誤り率，x_6 は 更新置換誤り率，x_7 は 音声理解率である。

決定係数は 0.57 で自由度調整済みの決定係数は 0.54 であった。RMSE (Root Mean Square Error) は 0.63 であった。今回得られたモデルは比較的よい性能を示すと言え、評価尺度としての有効性を示していると考えられる。また重回帰式によるタスク達成時間の予測値とタスク達成時間の実測値の分布を示す図を図 3 に示す。

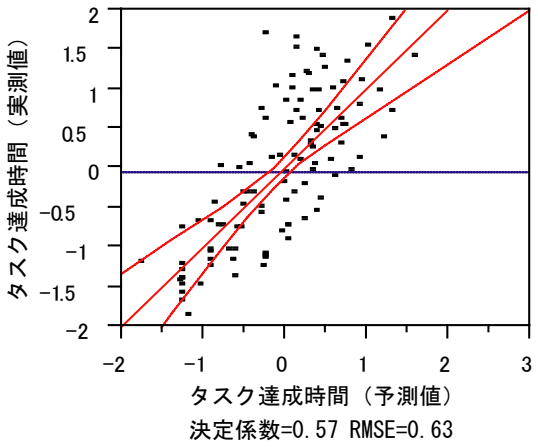

図 3: タスク達成時間の予測値とタスク達成時間の実測値の分布

[2]NTT サイバースペース研究所提供
[3]システムに誤動作が生じた 4 対話はデータから除いた。

また、タスク達成時間に対する 10 指標の相関係数を表 3 に示す。更新削除誤り率 が 0.62 と比較的高い相関があり、続いてスロット更新精度の-0.45 であった。

	タスク達成時間
スロット正解率	−0.40
挿入誤り率	−0.07
削除誤り率	0.29
置換誤り率	0.40
スロット誤り率	0.40
スロット更新精度	−0.45
更新挿入誤り率	0.15
更新削除誤り率	0.62
更新置換誤り率	0.24
音声理解率	−0.42

表 3: タスク達成時間に対する 10 指標の相関係数

5 結論と今後の課題

本稿では、逐次理解を行う音声対話システムの発話理解評価法を提案した。具体的には理解状態に関して得られるさまざま指標を説明変数、タスク達成時間を被説明変数として重回帰分析を行い、その結果得られる重回帰式を評価尺度として用いる。得られた重回帰式は比較的よい予測値を与え、評価尺度としての有効性を示した。また、対話の理解状態と理解状態の更新の両方を評価に用いる方が良いことが分かった。

今後の課題であるが、今回実験に使用したタスクドメインが比較的小さいことが挙げられる。ドメインを大きなもの、例えばフライト予約などに変更しても同様の結果が得られるかは不明である。またドメインが架空の会議室予約であることから被験者の対話に対するモチベーションの低さも問題である。今回、対話のパフォーマンスとしてタスク達成時間を用いたが、ユーザ満足度などの指標も今後用いる必要があるだろう。

以上に述べたようにさまざまな問題点は残るが、対話実験による結果、我々のアプローチは逐次発話理解の評価法として有望であると考えられる。

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。また重回帰分析に関してご助言を頂いた、環境理解研究グループの深山篤氏に感謝いたします。

参考文献

[1] 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*, 1999, pp. 200–207.

[2] 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.

[3] 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.

[4] 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*, 2000, pp. 150–159.

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

[6] 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.