

決定木学習による音声対話システムのための話者交代判定法

佐藤玲 東中竜一郎 田本真詞 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

{sato,rh,nakano}@atom.brl.ntt.co.jp, {tamoto,aik}@idea.brl.ntt.co.jp

概要

本稿では、ターン取得のタイミングを音声対話システムが決定するためのアルゴリズムを機械学習により作成する手法を提案する。本手法は、人手によりターン取得タイミングの正解をラベルづけされた、実ユーザとシステムとの対話コーパスを学習に用い、ターン取得判定の決定木を作成する。現状の多くの音声対話システムとは異なり、ポーズだけではなく、認識結果や理解結果、韻律情報などのさまざまな特徴量もターン取得判定に利用する。その結果、ユーザが発話の途中でポーズを置いても、途中で不要な割り込みをされずに発話を続けられる。予備実験の結果、ユーザのポーズで必ずターンを取るとしたベースラインアルゴリズムと比較して、決定木学習法を用いた本手法が優れていることが示された。

キーワード: ターン取得, 決定木学習, 音声対話システム

Learning Decision Trees to Determine Turn-Taking by Spoken Dialogue Systems

Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, Kiyoaki Aikawa

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

Abstract

This paper presents a method for learning algorithms to decide the timing of turn-taking in spoken dialogue systems. This method uses a decision tree learned from the corpus of dialogues between human users and systems in which desirable turn-taking behaviors are annotated by hand. It utilizes a variety of attributes, such as recognition and understanding results and prosodic information. Unlike most of the existing systems, it enables spoken dialogue systems to decide the timing of turn-taking based on not only pauses but also other features, so that users can speak to the system even if they put pauses in the middle of their utterances. The result of a preliminary experiment shows that the learned decision tree outperforms the baseline strategy, which takes turn at every user pause.

keywords: turn-taking, decision tree learning, spoken dialogue system

1 はじめに

近年の音声認識や言語処理技術の発展によって、人とコンピュータが音声対話を通してやり取りを行えるシステムの構築が可能になってきている。

これらのシステムはほとんどの場合、一定長のポーズを検出した時にターン（発話の番）を取り、システムの発話後、ターンをユーザに譲渡する。このようなシステムとの対話では、ユーザは自然に話すことができない。なぜなら、ユーザはシステムに伝えたいことをすべて話し終えるまで、ポーズを置かないように注意しなければならないからである。

本研究の目的は、音声対話システムのターン取得のタイミングを決定することである。人間同士の対話コーパスを分析することにより、ターン取得のタイミングと相槌のタイミングを調査する研究は、従来よりなされている。しかし、音声対話システムのターン取得に関する研究は少ない¹。我々はこれまでに、システムのターン取得に関する問題に取り組み、ポーズに加えシステムの理解結果を用いたターン取得法を開発した [1]。しかし、同手法によるターン取得の仕方は開発者の直感に基づいて考え出されたものであり、コーパス等のデータに基づいたものではない。

本稿では、人間と音声対話システムとの間の対話のコーパスを用い、システムがターンを取るかどうかを判定するアルゴリズムを、機械学習により求める方法を提案する。ユーザのポーズをシステムがターン取得すべき箇所の候補とし、ターンを取るポーズと、そうでないポーズを区別するアルゴリズムを学習する。ターン取得を決める要因として、ポーズ直前のユーザ発話の認識結果やそれまでの理解結果、ポーズ直前のユーザ発話の韻律など、多くの特徴量が関わると考えられるため、決定木学習を用いた。決定木学習はさまざまな要因を合わせてクラス分けするのに適した手法である。

以下では、関連研究を述べたあと、提案手法を説明し、さらに、会議室予約システムによる対話コーパスを用いた予備実験の結果を報告する。

2 関連研究

人間同士の対話を分析し、話者交代、すなわちターン取得のタイミングについて考察されている。たとえば小磯ら [6] は人間同士の対話を分析し、人間は統語情報と韻律情報が手がかりとして、ターンを取るかどうかを決めていると報告している。田本と川端 [2]

は、人間同士の対話を分析することによって、ターン取得判定に必要な情報を調べている。しかし、人間とコンピュータとの対話は、多くの点で人間同士の対話とは異なる [3]。その上、音声対話システムの場合、音声認識誤りや言語理解誤りは避けられない。よって、人間同士の対話の分析結果は、参考にすることはできても、人間とコンピュータとの対話にそのまま適用できるとは限らない。音声対話システムによるターン取得の研究例として、Bell ら [4] は、統語情報に基づいて、ターン取得のタイミングを決定する音声対話システムを構築し、文末表現の利用可能性を示唆している。しかし、音声認識誤りは考慮されておらず、システムの評価も行われていない。

3 アプローチ

我々はシステムがターンを取得する箇所の候補を、ユーザのポーズ、すなわち、ユーザの発話区間の終了時とした。これは、システムがユーザの発話中に割り込む場合、不快に感じるユーザがいるという実験結果があるからである [5]。これは、システムの割り込み発話が常に不適切であるという意味ではないが、現状では割り込み発話の適切なタイミングを決定することは難しいと考えられる。本稿では、ポーズ位置でのみシステムがターンを取るべきかを決定するアルゴリズムの作成を目的とする。

人間同士の対話に関しては、小磯らによって、韻律や文脈、統語情報や対話履歴がターン取得タイミングの決定に関与していると報告されている [6]。これらすべての特徴量が音声対話システムのターン取得に利用できるとは限らない。また、他の特徴量が効果的である可能性もある。しかし、音声対話システムがターンを取るタイミングの決定には、多くの特徴量が関わっていると考えること自体は妥当な仮定であると考えられる。上記の例のように、ターン取得のタイミングに関係する特徴量は、記号で表されるものもあれば、数値で表されるものもある。そこで、我々は記号も数値も合わせて扱うことが出来る点を考慮して、決定木学習法 ID3 [7] を用いた。決定木は人間と音声対話システムによる対話コーパスから学習により作成される。

3.1 クラス

学習により作成される決定木によって出力されるクラスは、「ターン取得する」か、または「ターン取得しない」のどちらかであるとする。決定木を学習

¹本研究では、相槌はターンを取得する発話とみなさない。

ラベル	発話数
両名ともターン取得で一致	2,503
両名の応答が不一致	307
両名とも相槌で一致	1,435

表 1: 2 名のラベラーによる正解ラベルの一致率

収録システムとラベラー間の一致	発話数
一致	3,001
不一致	937

表 2: 収録システムとラベラー間の一致

により作成するためにはクラス分けのための正解が必要であり、人間と音声対話システムによる対話コーパスに正解ラベルを付与しなければならない。ラベル付けは、システムとの対話を行っていない人間の被験者が行う。「ターン取得する」か、または「ターン取得しない」のどちらかのラベルがポーズごとに付与される。

3.2 特徴量

ここでは、決定木学習に利用すべき幾つの特徴量を述べる。

まず、ポーズの直前のユーザ発話長は、システムの適切なターン取得に影響を与えると考えられるため、ポーズ直前のユーザ発話長を特徴量として採用した。次に、人間同士の対話の分析により、ポーズの有無に関わらず、言語情報、特に文末表現がターン取得に重要な役割を果たしていることが分かっていることから [6]、ユーザ発話の認識結果も特徴量として用いる。また、ユーザ発話の韻律情報もターン取得のタイミングと関係があることが示されていることから、韻律情報も用いる。最後に、ターン取得のタイミングはシステムに伝えられた情報量にも依存すると考えられるので、ユーザ発話長、認識単語数、およびポーズ時点までの部分的な理解結果 [1] もターン取得の決定に効果的であると考え、これらの特徴量を学習に用いる。

4 実験

本章では、我々のアプローチの有効性を示すために行った実験の詳細とその結果について述べる。

4.1 対話データ収集

学習に用いるトレーニングデータと評価用のテストデータ収集のため、人間と音声対話システム間の対話を収録した。収録システムは、対話管理モジュールとして、音声対話ツールキット WIT [8] を用いて構築した。音声認識器には Julius [9]、音声合成器は NTT サイバースペース研究所の Final Fluets [10] を使用した。認識語彙数は 161 であり、また認識器は 0.75 秒より長い無音区間をユーザのポーズとみなす。対話ドメインは会議室予約である。

被験者はインストラクションに基づき、1 つまたは 2 つの日付において、1 つまたは 2 つの会議室を、ある時間からある時間まで予約する。対話ドメインとしては非常に小さいものであるが、もし語彙数等を増やし、タスクが複雑なものになれば、音声認識率は低下し、被験者は自然に話すことを諦めて、短い発話に終始する可能性がある。その点を考慮し、予備実験として比較的簡単なタスクを選んだ。

音声対話システムの理解状態は、フレーム表現によって表わされ、スロットと呼ばれる属性と値の対の集合より構成される [11]。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、対話に関する情報を含む。また、システムは逐次的にユーザ発話を理解する。その結果、システムの理解状態はユーザ発話を処理する度に一意に決定され、システムはターンを取るかどうかの判定に、ポーズ直後の理解状態を用いることができる。収録に用いたシステムがターンを取るかどうかは、以下のヒューリスティクスを用いて決定している。

- † ユーザが特定の発話によって、明示的にシステム応答を要求するか、システムが予約に必要な情報をすべて得た場合、システムは質問や確認のため、ターンを取得する。その他の場合は、システムは相槌のみを行う。

以降、このシステムを収録システムと呼ぶ。今回収録システムを用い、19 歳から 35 歳までの 24 名の被験者を対象に、210 対話収録した。総発話区間数は 4,768 であった。男女比は 1 対 1 で、一被験者につき約 10 対話を収録した。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 69% であった。

	正解精度 (%)
提案手法 (テストセットによる評価)	83.9
提案手法 (予測値)	86.6
収録システム	76.2
ベースライン	63.7

表 3: クロスバリデーションによる決定木の正解精度

4.2 人手による正解のラベル付け

次に、正解クラスのラベル付けを行った。対話研究者ではない被験者（以下、ラベラー）は収録対話を聞き、システムがターンを取るべきだと考えられるポーズに、ターン取得のラベルを付与した。システムが相槌を打つか、何もしなくて良いと考えられるポーズには相槌ラベルが付与された。このラベル付けの結果は正解クラスとして、機械学習に利用する。1 対話につき 2 名のラベラーがラベル付けを行った。2 名のラベラーのラベルの分布を表 1 に示す。両者の一致率は 92.8% であった。このことは、ほとんどの場合、人間のラベラー同士のターン取得の判定は一致し、ラベラーの付与したラベルをターン取得の正解として使うことが可能であることを示している。

本研究では、2 名のラベラーのラベルが一致したときのみ、正解クラスとして利用した。その他の場合は、トレーニングデータにもテストデータにも用いなかった。ラベルの一致の数は表 2 に示す。収録システムのターン取得正解率は 76.2% であった。

4.3 特徴量

学習には、幾つかの種類に分類される総数 114 の特徴量を利用した。それらを以下に列挙する。

カテゴリ: 統語・意味カテゴリ。認識結果の中に、あるカテゴリの単語が含まれている場合、そのカテゴリ特徴量の値は yes とし、そうでなければ no とする。機能語は細かいカテゴリに分類した。全部で 30 のカテゴリを用いた。

最終語: 認識結果の最終語の統語・意味カテゴリ。上述の統語・意味カテゴリのものと同一、30 の特徴量で表現した。

単語数: ユーザ発話の認識結果における単語数。

理解状態: ユーザ発話理解後のシステムの理解状態。特徴量として、理解状態を表現するフレームのスロット値を用いる。収録システムでは、理解状態に 8 つのスロットを用いるため、特徴量数は

8 で、特徴量それぞれの値は、スロットの値が埋まっているかどうかにより empty か non-empty かのどちらかとする。

理解状態変化: ユーザ発話前とユーザ発話後の理解状態を比較することによる、各スロット値の変化。

発話長: ユーザ発話の長さ。

韻律的特徴量: 野口と伝 [12] の相槌個所の発見に関する研究で用いられたピッチとパワーのパラメータ。彼らは 17 のピッチパラメータを利用しているが、我々は同様の 17 のパラメータに加え、1 つの信頼度を用いた。短い発話では、ピッチを抽出することが難しいため、信頼度はユーザ発話長に依存した値とし、長い発話では信頼度が高くなるように設定した。全く同じパラメータがパワーに関しても使用された。すなわち、パワーパラメータ数は 18 である。

4.4 評価

我々は前述の特徴量に基づいて決定木を作成し、他のヒューリスティクスと本方式を比較して、本手法の評価を行った。決定木の作成には、C4.5 アルゴリズムを利用した [7]。C4.5 のオプションはすべてデフォルト値を用いた。学習された決定木の性能を評価するために 10-fold cross validation を実行した。その結果を表 3 に示す。正解は枝刈りをした木を用い計算した。

システムがユーザ発話の終わりで常にターンを取る手法をベースラインアルゴリズムと呼ぶことにする。ベースラインアルゴリズムのターン取得正解率は、63.7% であった。収録システムのターン取得正解率は 76.2% であった。一方、我々の手法の正解率は 83.9% であった。各特徴量の効果を評価するため、学習に利用する特徴量を変化させてターン取得正解率の変化を調査した。その結果を表 4 に示す。理解状態の特徴量とカテゴリの特徴量がシステムのターン取得には有効であることが示されたため、これらの特徴量だけを使つての学習も行つた。その結果、理解結果とカテゴリの特徴量の両方を利用すれば、正解率が下がることはないが、理解結果の特徴量だけでは、不十分であることが分かった。

表 5 に、統語・意味カテゴリと理解結果を特徴量として決定木を作成した場合の根から葉までのパスのうち、カバー率の高いものをいくつか示す。表中の slot で終わっているものは、理解状態特徴量である。また、categories で始まっているものは、カテゴリ特

用いた特徴量	テスト セットに おける正 解精度 (%)	予測され る正解精 度 (%)
フルセット	83.9	86.6
カテゴリ以外	81.8	85.2
最終語とカテゴリ以外	84.2	86.8
単語数以外	84.9	86.8
理解状態以外	78.7	82.2
理解状態変化以外	84.6	86.8
発話長以外	84.4	86.9
韻律情報以外	85.5	86.7
カテゴリと理解状態のみ	85.2	85.6
理解状態のみ	76.4	75.1

表 4: 学習に利用する特徴量の変化によるターン取得正解率の変化

微量である。例えば, *start_time_slot* 特徴量は理解状態の開始時刻のスロット値が埋まっていることを意味している。また, *category_request_aux* 特徴量はユーザ発話の認識結果が要求を意味する *shitainodesuga* のような補語を含んでいることを意味している。これらの特徴量は対話ドメイン依存性が強いので, ここでは詳細に説明しない。

5 議論

本実験で用いた対話のドメインは比較的小さく, またタスクが現実の状況に基づくものではないという制限はあるが, ターン取得正解率がベースラインアルゴリズムと比較して改善されたため, 本実験結果は本手法の有用性を示唆している。また, これらの結果は, ユーザのターンの終点を決定しなくても, 途中の理解結果を出力する逐次理解が必須なものである可能性を示している。

表 5 が示すように, 今回作成した決定木はドメインに強く依存し, その意味を解釈することは難しい。そこでより簡潔で一般的なターン取得のルールを獲得するためには, 我々が考える以外の特徴量を模索する必要があると思われる。同様な実験を他のドメインで行うことにより, 簡潔で一般的な特徴量を見つけ出せる可能性があると考えられる。

今回の実験に使用したシステムでは, 韻律的な特徴量が決定的ではないことを示す結果となったが, この予備実験の結果だけからどんなシステムでも韻律が有用ではないと結論付けることはできない。韻律特徴の抽出精度の向上に加えて, 他のドメインでも実験を行う必要があるだろう。しかしながら, 小磯らが行った人間同士の対話研究の結果を考えると, ターン取得のタイミングを決定するには, 韻律の特徴だけではやはり不十分であると考えられる。

6 おわりに

本稿では, ユーザがポーズを置いたとき, 音声対話システムがターンを取るべきかどうかを決定するためのアルゴリズムを学習によって求める方法を提案した。その結果, 予備的な実験ではあるものの, 本方式は, 我々が用いた会議室予約のタスクでは効果的であることが示された。

今後は他の対話ドメインで実験を行うことに加え, 学習された決定木をシステムに組み込み, ユーザ満足度とともにタスク達成率や達成時間を調べることで, 本方式の効果を評価する計画である。

謝辞

日頃よりご指導頂くメディア情報研究部の村瀬洋部長, 有益な示唆を頂くマルチモーダル対話研究グループの諸氏に感謝致します。

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] M. Tamoto and T. Kawabata, "Analysis of Speaker Transition in Cooperative Task Dialogues," in *IPSJ-SLP, 96-SIG-SLP-12-3, Information Processing Society of Japan*, 1996, pp. 13–18, (in Japanese).
- [3] C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman, "Comparing several aspects of human-computer and human-human dialogues," in *Proc. SIGDIAL*, 2001, pp. 48–57.