ヒューマンインタフェース 97－22
音声言語情報処理 40－22
（2002. 2. 2）

2つの認識文法を用いた主導権混合型対話制御

安田宜仁 堂坂浩二 相川清明

日本電信電話(株), 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
yasuda@atom.brl.ntt.co.jp

あらまし

本稿では, 認識文法の異なる2つの認識器を同時に使用し, ユーザ発話後にそれらの出力のどちらを利用するかを決定するような主導権混合型対話の制御を提案する. 機械学習を用いて2つの認識器の選択を行った実験の結果を示す. 実験の結果, 適切な文法選択の精度は, ベースラインの 95.5% から, 97.7% に改善することができた.

Mixed-initiative dialogue control using two types of recognition grammer

Norihito YASUDA Kohji DOHSAKA Kiyoaki AIKAWA

NTT Communication Science Laboratories, NTT Corp.
3-1 Morinosato-wakamiya, Atsugi, Kanagawa, 243-0198 Japan
yasuda@atom.brl.ntt.co.jp

Abstract

We propose a new mixed-initiative dialogue control that uses two types of recognition grammer simultaneously and chooses a recognition grammer to use after user-utterance. We also present results of machine learning experiments to choose the appropriate type of recognition grammer. We have improved the accuracy for choosing the appropriate grammer from baseline of 95.5% to 97.7%.

1 はじめに

音声認識技術の発展により，人とコンピュータが音声による会話を通じて特定の仕事を行うような音声対話システムが実用化されてきている．しかし，現在の技術では，大規模な語彙の柔軟な言い回しを含むような話し言葉を認識することは未だ困難であるため，多くの商用の音声対話システムではシステム主導型対話と呼ばれるものが採用されている．システム主導型対話では，対話のやりとりの流れはすべてシステムが決定する．システムはごく限られた範囲内の質問を行い，システムが想定した範囲内での語彙や言い回しのユーザ発話のみを受け付け，それ以外の発話は受け付けない．

こういったシステム主導型対話と対照的なものがユーザ主導型対話である．ユーザ主導型対話では，対話のどの時点でもユーザの発話がある範囲内にあるという仮定は行わない [1]．

システム主導型対話では，事前に決められた対話の流れに従う上に高い認識精度が望めるため確実に対話を進行させることができる反面，システムの想定した範囲外のユーザ発話を無視してしまうという問題がある．さらに，型通りの質問応答を繰りかえすため，慣れたユーザであっても，1度の対話にかかる最低の時間が長いという問題もある．一方でユーザ主導型対話では自由で柔軟はユーザ発話を許容する反面，語彙や言い回しに関する制限が緩いために，認識精度が低くなり，誤認識が原因で対話の進行に支障が出てしまいやすいという問題がある．

このような問題点を解決するために主導権混合型対話と呼ばれるものがある．これはシステム主導型対話とユーザ主導型対話を組合せたものである．対話の局面に応じてシステム主導型対話とユーザ主導型対話を選択し，必要に応じて認識文法を切り替える．主導権混合型対話の音声対話システムでは，たとえば，「フライトは何がありますか?」という質問に対して「行き先を指定してください」といった応答をし，行き先に限定したユーザ応答を求めるといったことを行う [2]．

これまで，主導権の切り替え対話中に検出されるさまざまな cue を用いて主導権を切り替えるような主導権混在型対話制御などが提案されている [3]．

しかし，認識文法の切り替えのタイミングや切り替えた認識文法の内容は，ユーザが対話中に知ることのできる情報ではない．そのため，ユーザは，システムが想定した範囲を越えた語彙や言い回しで応答をしてしまう場合がある．システムが想定した範囲を越えたユーザ発話は認識の対象になっていないので，システムは適切な対処を行うことができず，対話の進行を阻害する要員になっていた．

こういった問題は，システムの発話の時点，つまりユーザの応答の前に認識文法を選択していたために起こっていた．

そこで本稿では，ユーザ主導を前提とした認識文法での認識器とシステム主導を前提とした認識文法での認識器の2つの認識器を同時に使用し，ユーザ発話後にそれらの出力のうちのどちらかを選択するような主導権混合型対話の制御を提案する．ユーザ発話後に文法の選択を行うことで，対話の進行に関する情報や，ユーザ発話の継続長なども含めて判断することが可能となる．さらに，実際の音声対話システムで利用可能な特徴を用い，機械学習を行うことで，認識器の選択が行えるかどうかを検討する．

2 従来の問題点

従来の主導権混合型対話制御では，システムがユーザに対して項目を限定した質問や確認を行えば，ユーザはその範囲での応答を行うものだという前提のもとに，項目を限定した質問や確認を行っていた．しかし，限定的な質問や確認を行った場合でも，質問や確認の対象がユーザの意図しているものと外れているものだったり，ユーザが質問や確認より前の対話のやりとりでまだ終了していないものがあると思った場合には，システムからの情報要求や確認の範囲外のことを伝達しようと考えることが起こり得る．システム発話の度にシステムの認識語彙が切り替えられているということや，さらにはその認識語彙が行われたことを知る手段を持たないユーザがシステムが行なった限定的な質問や確認とは異なる内容の発話を行うことが考えられる．

このようなシステムが想定した範囲を越えたユーザ発話は認識対象になっていないので，システムは適切な対処を行うことができず，対話の進行を阻害する要因になっていた．

図2の対話例では，システムは「時間」あるいは「日付」を求めているため，ここで想定してないもの，たとえば会議室名などは認識対象にはなっていない．しか

U1	会議室を予約したいんですけれども、出来ますでしょうか
S1	はい
S2	いつ使用されますか？
U1	2月6日です
S3	いつのご使用ですか？
U2	2月6日の、午前9時30分から午後15時までです
S4	いつ使用されますか？
U3	2月6日の、9時30分から15時30分まで、第10会議室を使用したいんですけれども
S5	いつのご使用ですか？
U4	2月6日になります 時間は午前9時30分から、15時までの使用となります
S6	いつのご使用ですか？
U5	2月6日になります、時間は9時30分から15時までです
U6	と、第10会議室を借りたいんですけれども

図 1: 対話例

し実際には発話 U3 や 発話 U6 のようにユーザがこれらの情報も含めて応答することがある.

一方で, 大語彙で柔軟なユーザ主導的文法を常に用いることも考えられるが, システム主導型の文法を用いた場合に比べて, 認識精度が低くなってしまう. 仮に全ユーザ発話に占めるシステム想定外のユーザ発話の割合が, 低い認識精度という欠点を補う程度に高いのであれば, 常にユーザ主導的文法を用いることは有用であるだろう. しかし, 現在の技術では, ユーザの発話の自由度を増やした場合の認識精度は, ユーザが想定外の質問する割合の低さに見合うほどで高くない. さらにタスクの規模が大きくなればなるほど, 大規模な語彙と多様な言い回しの可能性があり, ユーザ主導型文法での認識精度は下がっていってしまう. したがって常にユーザ主導的文法を用いるという戦略は現状では適切ではない.

3 提案法

もしユーザがシステムの想定内の応答をした場合にはシステム主導的文法を, ユーザがシステム想定外の応答をした場合にはユーザ主導的文法を選択することができれば, 想定内の発話の場合での高い認識精度を犠牲にすることなく, 想定外の自由な発話を許容することができると考える.

本稿で提案する方法は, システム発話直後ユーザ発話以前に認識文法を選択するのではなく, 常にユーザ主導を前提とした認識文法での認識器とシステム主導を前提とした認識文法での認識器の 2 つの認識器を同時に使用し, ユーザ発話後にそれらの出力のうちのどちらかを選択する方法である.

認識文法の選択を, ユーザ発話以前に使える情報だけでなく, たとえば, 発話の継続長や, システム発話と認識結果の対応といったユーザ発話後に使える情報を含めて判断することで, 適切な文法の選択が可能にあると考えられる.

適切な認識器の選択を行うために, 機械学習の手法を用いてどちらの認識器の結果を採用するかどうかを決定する. このために使う特徴としては, 5 節に示すように, 現在の音声対話システムが実際に動作中に取り扱うことができる情報を用いる.

4 コーパス

会議室の予約を行うことができる音声対話システムを用いてコーパスを収集した. この音声対話システムは, form-based なタスク構造をもち, 予約内容は 場所, 開始時間, 終了時間, 日付けの 4 つの項目から構成され, 場所は 2 箇所まで同時に指定することができる.

ユーザから伝達された予約内容に対しては, 少なくとも一度はシステムからの確認が行われる. また, 予約内容を構成するには不足している情報があれば, ユーザに対して「何時からですか?」のような情報要求のための発話を行う.

確認の戦略は 2 種類ある. 一方は新しく伝えられた項目がある度にその項目を確認する方法で, もう一方はすべての情報が得られるまではユーザへの情報要求を繰り返し, 必要な情報がすべて得られた時点でまとめて確認を行う方法である.

音声認識には連続音声認識コンソーシアムの 2000 年度版 Julius[4] を用いた. テキスト音声合成には, NTT サイバースペース研究所の FLUET[5] を用いた. 対話数は 216 で, ラベル付け対象のシステムターン数は 3649 である. 被験者数は 18 才から 65 才までの大学生を中心とする 36 人で, 男女比は 1:1 である. 披験者

	開始	終了	書き起し文	応答種類ラベル
U1	6.61	8.83	会議室の予約をお願いしたいんですが	
S1	10.21	11.91	いつ使用されますか?	:date
U2	12.45	15.08	ええ、11月28日です	
S2	16.35	18.48	8時からデスネ?	:no :date
U3	19.17	21.93	いえ、11月28日です	
S4	23.37	27.04	11月28日デスネ?	:yes
U4	27.18	27.53	はい	

図 2: ラベル付けされた対話の一部

には予約するべき内容を事前に指示してあるため, システムが確認を行なったり情報を要求する項目に対して, 被験者が答をもたないという状況は発生しない.

システムが行なう情報要求発話と理解内容の確認発話を対象に, 対応する応答がシステムが想定した範囲内だったのか, それともシステムが想定した範囲外だったのかを人手で付与した. 図4にラベル付けをした対話断片を示す.

たとえば, システム発話 S1 では, ユーザは日付を答えているので, 応答種類ラベルには日付を表わす :date が付与される. 相づちはラベル付けの対象にしていない.

データ収集時のシステムは, 対話中での認識文法の切り替えは行わず, 対話を通じて同一のユーザ主導的認識文法を用いた.

収集したすべてのユーザ発話を再度システム主導的文法で認識を行うことにより, 各ユーザ発話に対して2つの認識結果を用意した. このシステム主導的文法は, 収録時に用いた文法からシステムが言及した項目以外を除いて作成した. たとえば, システム発話 S1 のようにシステムが日付けまたは時刻を要求している場合には, システム主導的文法には会議室名は含まれない. 同様に, S2 では時刻を確認しているので, 日付や会議室名は含まれない. 「はい」「いいえ」といった対話の進行に必要な一般的な語彙はどちらの文法にも含まれている.

対象としたシステム発話のうち, 応答が想定外であったものは, 4.5% であった. そのうち, システムの想定内のことに加えて, それ以上の情報をユーザが伝えた例は 42.4%, システムの想定内のこととは無関係のことをユーザが伝えた例は 57.6% であった.

5 判定に利用する特徴

5.1 対話の進行に関する特徴

対話の進行に関する特徴として, システムが言及した属性, システム発話の種類, 同一の属性に対して連続何度目の言及かを用いる.

システム発話の種類 システム発話は確認なのかそれとも情報要求なのか. 今回対象としたシステム発話は, 確認と情報要求の2種類に分けられる. この特徴単体では, ユーザがシステムの想定外の発話した例は, 確認の場合で 72.7%, 質問の場合で, 27.2% であった.

システムが言及した属性名 直前のシステム発話において, 確認あるいは情報要求の対象となった属性名. 場所と時刻, 時刻と日付, 場所, 時刻, 日付 の5種類. 開始時刻と終了時刻はまとめて時刻として扱った.

同一の属性に対して連続何度目の言及か 同一の属性に対して, システムが確認や質問を連続何度目の言及かどうか. たとえば, 図2の対話例での S2 の発話であれば, 時刻を含む言及は連続2度目なので 2 となる.

5.2 音声認識・言語理解に関する特徴

音声認識・言語理解に関する特徴として音声認識スコアと, 理解結果を用いる.

音声認識スコア 2つの認識器の音声認識スコアのうち, どちらのスコアの方が高いのか, あるいは同じなのか. 音声認識スコアは, 認識候補上位 10 候補を求め,

		想定内		想定外		
	使った特徴	正解率	再現率	適合率	再現率	適合率
ベースライン	システム主導型のみ (従来の主導権混合型対話)	95.5%	100.0%	95.5%	0.0%	-
	認識結果の一致のみ	95.3%	96.4%	98.6%	53.9%	31.8%
	音声認識スコアのみ	91.9%	94.8%	96.6%	36.6%	26.6%
決定木作成	音声	90.5%	97.2%	92.8%	18.3%	37.9%
	認識・理解	95.3%	100.0%	95.3%	0.0%	-
	対話	95.3%	100.0%	95.3%	0.0%	-
	音声 + 対話	96.7%	99.8%	96.9%	35.0%	87.5%
	音声 + 認識・理解	96.3%	99.7%	99.6%	24.1%	77.8%
	認識・理解 + 対話	96.7%	99.0%	97.6%	44.4%	66.7%
	音声 + 認識・理解 + 対話 (すべて利用)	97.7%	99.8%	97.8%	52.6%	90.9%

表 1: 想定外発話と想定内発話の分類の正解率, 再現率, 適合率

各候補について音声フレーム毎に認識結果の viterbi path 上における状態と最も高いスコアを与える状態との対数出力尤度比を求め, フレーム長で正規化したものを用いた.

理解結果 2つの認識器の出力を構文解析した結果を項目単位 (場所, 時刻, 日付) で表わしたものを対象に, システム主導的文法での結果, ユーザ主導的文法での結果, そして2つの出力が項目単位で一致しているかどうか (0, 1). もしもユーザがシステムの想定外の応答をしていて, しかも誤認識が起こっていないのであれば, 2つの出力は異なる結果を出すはずである.

5.3 音声に関する特徴

音声の特徴としてユーザ発話の継続時間を用いる.

発話継続時間 直前のユーザ発話の継続時間. システムが求めた範囲内でユーザが応答するようなシステム主導型の対話と違い, ユーザが自由に複数の項目について言及するような場面では, ユーザがより多くの情報を伝えようとして, 発話長が長くなることが考えられる.

6 評価

6.1 ベースライン

機械学習による方法との比較のため, 従来の主導権混合型対話に相当する方法, 音声認識スコアのみを使う方法, 認識結果の一致を見る方法の3つのベースラインを用意した.

従来の主導権混合型対話に相当する方法では, 確認あるいは情報要求を行なった範囲内でのみユーザ発話があることを前提とする.

音声認識スコアのみを使う方法では, システム主導的文法とユーザ主導的文法の2つの文法の選択はユーザ発話後に行うが, 5.2 節で述べた音声認識スコアのみを利用し, ユーザ主導的文法をもつ認識器の音響スコアが高い場合のみユーザ主導的文法の認識器の出力を採用し, それ以外の場合はシステム主導的文法の認識器の出力を採用する.

認識結果の一致を見る方法も, 音声認識スコアのみを使う方法同様, 認識文法の選択はユーザ発話後に行う. 5.2 節同様に 2 つの認識器の理解結果の項目単位での一致だけを見て, 2つの出力が項目単位で一致していれば, システム主導的文法を採用し, そうでなければユーザ主導的文法を採用する.

6.2 決定木の作成

機械学習の手段として, 決定木を用いる. 5節で挙げた特徴を使い, 決定木を作成した. 決定木作成のアルゴリズムには C4.5[6] を使用した.

5節で挙げたすべての特徴を使ったものに加えて, 比較のため対話の進行に関する特徴, 認識・理解に関する特徴, 音声に関する特徴を単体で, あるいはこれらの組合せを使って作成した決定木も作成した.

全データのうち 3/4 を訓練データとし, 1/4 を評価データとした.

評価の結果を表1に示す. 表より, すべての特徴を利用した決定木の結果は最も高い精度であり, 従来の主導権混合型対話に比べ, 2.2% の精度向上が見られる.

今回の実験では想定外の発話の割合が 4.5% と少なく, ベースラインが 95.5% と高かった. これは, 被験者は事前に予約内容を把握していて, その予約内容はシステムの知識構造と一致していた. このことはシステムの想定外の発話の割合を減らしていたと思われる. 実際に対話システムが利用される場面では, ユーザの要求内容と, システムの知識構造は必ずしも一致するとは限らず, そのような場合, システムの想定外の発話はもっと増えることが予想される.

7 おわりに

本稿では, 混合主導型の対話システムにおいて, ユーザ発話を高い精度で理解することを目的として, システム主導的文法とユーザ主導的文法の選択をユーザ発話後に行う方法を提案した. 対話の進行に関する特徴, 音声認識・言語理解に関する特徴, 音声に関する特徴を利用して決定木を作成することで, 97.7% の精度でユーザ発話後に適切な文法を選択することができた.

今後は, システムが言及した内容に対して必ずしもユーザが答をもってないようなことも考えられような複雑なタスクを用いて有用性を検討する必要がある.

謝辞 日頃よりご指導いただく, 当研究所メディア情報研究部 村瀬洋部長, 音声認識スコアの利用についてご助言いただいたマルチモーダル対話研究グループ宮崎昇氏に感謝します.

参考文献

[1] Gorin, A. L., Riccardi, G. and Wright, J. H.: How may I help you?, *Speech Communication*, Vol. 23, No. 1/2, pp. 113–127 (1997).

[2] Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., DiFabbrizio, G., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M., Ruscitti, P. and Walker, M.: The AT&T-DARPA COMMUNICATOR Mixed-Initiative Spoken Dialog System, in *ICSLP* (2000).

[3] Chu-Carroll, J.: MIMIC: An adaptive mixed initiative spoken dialogue system for information queries (2000).

[4] http://www.lang.astem.or.jp/CSRC/.

[5] http://www.ntt.co.jp/news/news98/9802/980216b.html.

[6] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1992).