音声言語情報処理 34－41
(2000. 12. 22)

質問-応答連鎖からの音声対話システムの誤解の検出

平沢 純一　宮崎 昇　相川 清明
NTT 研究所
〒 243-0198 神奈川県厚木市森の里若宮 3-1
jun@idea.brl.ntt.co.jp
http://www.brl.ntt.co.jp/cs/dug/

あらまし

音声対話システムが何らかの誤解を起こすことは避けられない．ユーザに特別な話し方や操作を強いずにシステムが誤解を克服するには，システムがユーザとのやりとりを利用して自らの誤解を検出し，回復していく技術が必要となる．本稿では，システムの確認（質問）からそれに対するユーザ応答への連鎖に見られる特徴が，システム確認が正しい場合と誤解を含む場合の間で異なっていることを利用して，システムの確認質問に含まれている誤解を検出する精度について報告する．実験の結果，システム-ユーザの発話連鎖（質問-応答連鎖）を用いることで，適合率 88.3%，再現率 87.8%の精度でシステム確認に含まれる誤解を検出することができた．

Detection of Misunderstandings in Spoken Dialogue System using System-User Utterance Sequence

HIRASAWA Jun-ichi,　MIYAZAKI Noboru,　AIKAWA Kiyoaki
NTT Laboratories
3-1 Morinosato-wakamiya, Atsugi, Kanagawa, 243-0198 Japan
jun@idea.brl.ntt.co.jp

Abstract

Conventional spoken dialogue systems cannot avoid misunderstandings. A spoken dialogue system should be able to detect and correct its own misunderstandings in the course of dialogues without asking the user for any special expression or operation. For this purpose, we analysed user-responses following the system confirmation question with/without misunderstandings. Since the user-responses following the system confirmation question with/without misunderstandings differ in some feature parameters, we had good performance to detect whether or not the current system confirmation includes the misunderstandings. In our experimental data, precision 88.3% and recall 87.8% were achieved.

1 はじめに

人間（ユーザ）とコンピュータが音声対話を通じて何らかの仕事（タスク）を遂行できるシステム（音声対話システム）を目指している．ユーザが日常に用いる話しことばで対話できるシステムが実現すれば，ユーザが特別な訓練をせずに使えるインタフェースとなる．しかし，現状の技術では音声の認識・理解の誤りを避けられず，満足にタスクが遂行できないことがある．ここで「特別な訓練を必要としない」という音声対話の利点を保つには，システムに誤解が生じてもあらかじめ決められた発声や特別な操作をユーザにさせるのではなく，あくまで自然な対話のやりとりの中からシステムが自らの誤解を検出し回復させる技術が急務である．

通常の音声対話システムは「確認発話[1]」を行い，それに対するユーザ応答の認識結果によって誤解の検出を試みる．しかしこの方法ではシステムの誤解を検出できるとは限らない．なぜなら，ユーザは yes/no 質問の形式であるシステム確認に対して明示的な「はい/いいえ」で応答するとは限らない [2, 4]．またユーザは訂正のため同じ発話を繰り返すことが多い [10] が，同じ発声であればシステムは同じ誤りを繰り返しやすい．それどころか，誤解を含むシステム確認の後やシステムからの再発声要求に対するユーザ応答は hyper-articulation[8] を起こしやすく認識精度は却って悪くなる [10]．システム確認に対するユーザ応答も他のユーザ発話と同様に誤認識されうるのであって，ユーザ応答の内容（認識結果）のみに依存した誤解の検出は，問題の解決になっていない．

そこで，ユーザ応答の認識結果だけに依存しない誤解の検出に取り組むため，我々はまず wizard of OZ 方式による音声対話実験により，正しいシステム確認と誤解を含むシステム確認に対するユーザ応答のデータを収集・分析した [2, 3]．その結果，システムの正しい確認に対して，ユーザは内容語を用いず「はい」や「そうです」だけで応答するためユーザ応答の継続長が短い傾向にあった．これに対して誤解を含むシステム確認に対するユーザ応答が「いいえ」や「違います」だけなのは 25%弱で，80%近い応答はシステムの誤解を訂正するため同じ内容語の繰り返しを含んでいることがわかった．

本稿では，システム確認とそれに対するユーザ応答の連鎖に上記で述べたような特徴があることを利用して，システム確認に含まれる誤解を検出する精度についての分析を報告する．つまり，今回のポイントは，(i) 正しいシステム確認の後のユーザ応答は継続長が短く，(ii) 誤解を含むシステム確認の前後では同じユーザ発話が繰り返されやすい，という特徴を用いることで，システム確認で言及されている値の正誤をどの程度推定できるのかを明らかにすることである．

2 関連研究

システムのふるまい[2] の正誤に応じて，その直後のユーザ発話を分析する研究がいくつかある．例えば，システムが誤認識した後のユーザ発声には（通常の発声とは）音響的に異なる特徴がある [8, 12]．Levow[7] は 38 種類の韻律情報を用いて「通常の入力発声」から「システム誤認識の後の訂正発声」を識別する実験を行った結果，誤り率は 25%であった．

音響的な特徴ではなく，言語的な特徴（応答内容）についての分析もある．Boyce[1] はシステム確認が正しい場合/誤っている場合のそれぞれに対するユーザ応答（訂正）の仕方を分析している．Krahmer ら [6] の分析によれば，システムの explicit な確認に対しては，ユーザ応答の形式と内容（単語数，語順，訂正しているかどうか）を用いることで適合率 91%，再現率 95%でシステム確認に誤りがあることを検出できる．但しこれは書き起こしレベルの情報を用いている．

システム確認の直後のユーザ応答だけを分析するのではなく，元のユーザ発話とシステム誤認識の後のユーザ訂正発話との関係も分析されている．Swerts ら [10] は，explicit なシステムの誤り確認に続くユーザ訂正発話のうち 75%は，元のユーザ発話の「繰り返し」であると報告している．田本ら [11] によれば，繰り返しを起こしているユーザ訂正発話は，元のユーザ発話に較べてピッチや継続時間長の増大，及び発話速度の低下が見られる．もしユーザによる訂正や繰り返しを検出できれば，システム確認中の誤解を検出・回復する第一歩となる．

ユーザ音声の繰り返しの検出に関しては，検索シ

[1] confirmation utterance または verification utterance．システムが理解している内容について「月曜日でよろしいですか？」のように確認する．

[2] 狭義のシステム確認発話だけでなく，確認発話以外のシステム音声出力，音声以外のモダリティによるシステム出力も含む．

ステムでのキーワード（単語）入力のインタフェース[3]において，入力発声と訂正発声の「認識結果の尤度の差」と「パワーの時系列ベクトルの距離」を用いて90%弱の適合率と再現率でユーザ音声の繰り返しを検出している[5].

3 実験

システム確認に含まれる誤解を検出する精度を調べるため，wizard of OZ方式により収録した音声対話データを用いた．音声対話データは以前の分析[2, 3]で用いたものと同じである．収録に関する詳細は[2, 3]に述べてあるので，ここでは要点のみを述べるにとどめる．

実験環境. システム質問-ユーザ応答の多様な連鎖を効率的に収集するため，収録実験はwizard of OZ方式を採用したので，システムの誤解の仕方（確認の連鎖パターン，頻度）は実験計画で予め定めてある（表1）．システムの音声には規則合成音声[4]を用い，被験者とシステムのやりとりは音声のみを介して行われた．

対話ドメイン. 対話はform-basedのタスク構造を持つ「会議室の予約」をドメインとした．収録では，システムが常に対話の主導権を持ち，システムが質問を進めることで曜日・予約時間・会議室名の3つのスロットに値を入れて対話が進行する．対話例を図1に示す．

実験方法. 被験者は10〜40代の大学生及び研究者，男女13名を用いた．被験者は「コンピュータと対話することで会議室の予約を取る」よう教示され，被験者ひとりあたり5〜10対話を収録した．

システム発話（質問）. 対話は3つのスロットを埋めるそれぞれの部分から構成され，1つのスロットに関する部分では，以下の4種類のシステム質問が用いられた．

initQ: initial Query. システムがユーザにスロット名を知らせるための質問．システムが新しいスロットに話題を進める時に「何曜日デスカ?」「ドノ会議室ガヨロシイデスカ？」のように質問する．

[3]自由発声の対話インタフェースではなく，システムの認識結果が誤りの場合には同一の単語を繰り返し発声することで誤りを訂正する仕様のインタフェース．

[4]NTTサイバースペース研究所 メディア処理プロジェクトのテキスト合成システム(FLUET)を使用した．

表1: システム質問（確認）の連鎖パターンと頻度.

システム質問の回数（タイミング）ごと initQ	sys-confirm1	sys-confirm2	質問タイプごと
initQ(360)	→ φ(67)	→ φ(67)	initQ(360) CR (131) ME (131) ME2 (18) RE (131)
	→ CR (84)	→ φ (64)	
		→ CR(20)	
	→ ME (98)	→ φ (61)	
		→ ME (19)	
		→ ME2(18)	
	→ RE(111)	→ φ (50)	
		→ CR(27)	
		→ ME(14)	
		→ RE (20)	
initQ(360)	cnfm1(293)	cnfm2(118)	(771)

CR: Correct Recognition. ユーザが伝えたのと同じ，正しい値による確認質問．ユーザ発話が正しく認識されている場面が想定されている．

ME: Misrecognition Error. ユーザが伝えたのと異なる，誤った値による確認質問．ユーザ発話が誤って認識された場面が想定されている．

RE: Rejection Error. システムによる再質問．システムが認識結果を得られなかった場面を想定している．「モウ一度オ願イシマス」「何曜日デスカ？(initQと同じ質問)」のようにユーザに再発声を求める．

1つのスロットに関する部分の対話ではこの4種類のシステム質問を組み合わせて，10パターンの連鎖が用意された．各スロットでは最初に必ずinitQが行われ，その後に0〜2回のシステム質問（CR, ME, RE）が行われる．表1に連鎖パターンとそれぞれの頻度を示す．

4 結果と分析

利用した対話データは13名の被験者による120対話(360スロット分)，総発話数は2128(ユーザ発話823，システム発話1305)．823のユーザ発話のうちシステムの「最初の質問（initQ)」「再質問（RE)」「挨拶」に対するユーザ応答を分析対象から除き，値に関するyes/no質問である「正確認（CR)」「誤確認（ME)」に対するユーザ応答289（CR確認の後134，ME確認の後155）を分析の対象とした[5].

分析では，着目するユーザ応答だけでなく，それに先行する発話の情報も用いた（図2）．以後，分

[5]システム確認数（CR確認131，ME確認149，表1の最右列）とユーザ応答数（CR確認の後134，ME確認の後155）が一致しないのは，1つのシステム確認に対して複数のユーザ応答が存在する場合があったためである．

発話番号	話し手	回数 (confirm#)	確認タイプ	スロット (slot)	発話
(1)	SYS			挨拶	会議室予約システムデス
(2)	SYS	initial-Query	initQ	曜日	何曜日デスカ?
(3)	USR	response0			月曜日です
(4)	SYS	confirm1	ME	曜日	火曜日デスカ？
(5)	USR	response1			違います．月曜日です
(6)	SYS	confirm2	ME2	曜日	木曜日デスカ？
(7)	USR	response2			いえ，月曜日です
(8)	SYS	initial-Query	initQ	時間	何時カラデスカ？
(9)	USR	response0			午前 9 時からです
(10)	SYS	confirm1	CR	時間	午前 9 時デスカ？
(11)	USR	response1			はい

図 1: **実験で収集された対話例.** 各発話ごとに，確認回数（confirm#），確認タイプ，言及しているスロット（slot）などの情報が付与される．

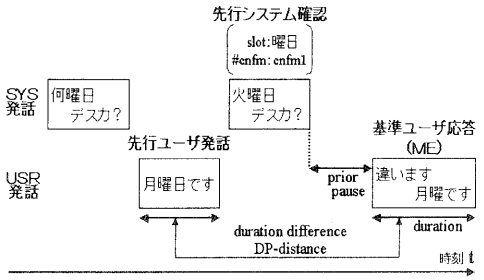

図 2: 分析対象とする発話連鎖の関係と使用する特徴パラメータ.

析で焦点をあてるユーザ応答を「基準ユーザ応答」と呼び，基準ユーザ応答を誘引したシステム確認を「先行システム確認」，基準ユーザ応答の 1 つ前のユーザ発話を「先行ユーザ発話」と呼ぶ（図 2）．

基準ユーザ応答にはそれぞれ先行システム確認の確認タイプに応じて CR，ME いずれかのラベルが付与される（図 2の例では ME）．基準ユーザ応答と先行システム確認の関係，基準ユーザ応答と先行ユーザ発話の関係を分析するため，それぞれの基準ユーザ応答について以下の特徴パラメータを計算した．

4.1 分析に利用した特徴パラメータ

言及スロット（slot）：先行システム確認が言及しているスロット名を表す．“曜日”，“時間”，“部屋名”，“挨拶[6]” のいずれかのカテゴリ値を取る（図 1の第 5 コラム）．

確認回数（confirm#）：先行システム確認が同一スロット内での何回目のシステム確認であるかを表す．1 回目の確認（confirm1），2 回目の確認（confirm2）などの値を取る（図 1の第 3 コラム）．

継続長（duration）：基準ユーザ応答の発話継続長（単位 sec）．

先行話し手（prior speaker）：基準ユーザ応答のひとつ前の発話（直前発話）の話し手を表す．通常のユーザ応答はシステム質問に対する応答なので prior speaker=SYS となるが，複数のユーザ発話が連続する場合には prior speaker=USR となる．

先行ポーズ（prior pause）：直前発話の終点から基準ユーザ応答の始点までの時間（単位 sec）．prior pause<0 の場合は，基準ユーザ応答が直前発話に重なっている（ラッチング）．

継続長の差（durational difference）：基準ユーザ応答と先行ユーザ発話の発話継続長の差（単位 sec）．dur.diff.<0 なら先行発話が短く，dur.diff.>0 なら先行発話が長いことを示す．

2 ユーザ発話の音響的距離（DP distance）：基準ユーザ応答と先行ユーザ発話との音響的な距離．動的計画法（DP マッチング）により発話継続長を正規化した 2 つの発話のケプストラム系列の間のユークリッド距離を求めた．値（距離）が 0 に近い[7] ほど 2 つの音声は類似している，すなわち，繰り返しが起こっていると考えることができる．

[6]但し「slot=挨拶」は分析対象には含まれない．

[7]2 つの音声の継続長の差が大きく，DP マッチングにおける整合窓から外れると非常に大きな値を出力するようになっているので，分析には 0.0≤DP dist.<10.0 のものだけを用いた．

表 2: CR/ME 確認に対する特徴量の分布（平均と標準偏差）と有意差検定

特徴量 (単位)	平均 (S.D.) CR に対して	ME に対して	平均間の差	t 値	p
duration(sec)	0.34 (0.19)	1.03 (0.46)	−0.69	−16.29	< 0.0001
prior pause(sec)	0.53 (0.63)	0.61 (0.51)	−0.08	−1.24	0.22
dur. diff.(sec)	0.69 (0.47)	0.13 (0.56)	0.56	9.11	< 0.0001
DP dist.	0.52 (0.33)	0.56 (0.43)	−0.04	−0.34	0.74
特徴量 (値)				χ^2 値	p
slot(曜日, 時間, 部屋名)	カテゴリー変量			0.84	0.66
confirm#(confirm1,confirm2)				1.17	0.56
prior speaker(SYS,USR)				3.80	0.05

4.2 分析結果（特徴パラメータの分布）

289 の基準ユーザ応答（CR 確認の後 134，ME 確認の後 155）に関して，上記のそれぞれの特徴パラメータを計算し，CR の場合と ME の場合での特徴の違いを分析した（表 2）．分析のポイントは (i) 正しいシステム確認の後のユーザ応答（CR）は継続長が短く，(ii) 誤解を含むシステム確認（ME）の前後では同じユーザ発話が繰り返されやすい，かどうかである．

表 2には，基準ユーザ応答が CR に対する場合と ME に対する場合に関して，連続変量（時間や距離）ではそれぞれの平均，及び平均の間に有意な差があるかどうかの t 検定の結果を示した．カテゴリ変量（スロットやタイミング）ではそれぞれの頻度の分布に有意な差があるかどうかの χ^2 検定の結果を示した．

表 2を見ると，基準ユーザ応答の発話継続長（duration）の平均は正しいシステム確認（CR）に対する時の方が有意に短い（0.34sec）ことがわかる．また，2 つのユーザ発話の継続長の差（dur.diff.）の平均では ME 確認をはさんだ場合の方が有意に短く（0.13sec），似た長さの発話が繰り返されていることがわかる．しかし 2 ユーザ発話間の音響的距離（DP dist.）は CR/ME どちらの確認に対する場合も差があるとは言えなかった．

4.3 システム確認中の誤解の検出

では，システム確認が正しい場合と誤解を含んでいる場合の間で，システム確認とその前後のユーザ発話が異なる特徴を示すということを利用して，システム確認で言及されている値が正しいか誤っているかをどの程度推定できるのだろうか？そこで決定木学習の手法（C4.5[9]）を用いて，分析に用いた各特徴パラメータ単独で，システム確認中の誤解を検出する精度を調べた．精度は検出の推定誤り率，及び適合率，再現率[8] を計算した．精度の計算には交差検定を行った（表 3）．

表 3: 各特徴量ごとの誤解検出精度

特徴量	推定誤り率	適合率	再現率
slot	48.4%	52.4%	95.6%
confirm#	46.4%	53.7%	100.0%
duration	12.4%	93.4%	83.2%
prior pause	36.0%	64.2%	74.5%
prior speaker	46.4%	53.7%	100.0%
dur. diff.	25.9%	78.2%	72.9%
DP dist.	46.4%	53.7%	100.0%
ALL	13.5%	88.3%	87.8%

表 3の推定誤り率を見ると，基準ユーザ応答の継続長（duration）で 12.4%，継続長の差（dur.diff.）で 25.9%の誤り率を達成している．これは，(i) 正しいシステム確認の後のユーザ応答（CR）は継続長が短く，(ii) 誤解を含むシステム確認（ME）の前後の 2 ユーザ発話は長さが似ている，という特徴がシステム確認中の誤解の検出に役立ったことを示している．さらに，今回分析に使用したすべての特徴パラメータを併用すると，全体で適合率 88.3%，再現率 87.8% の精度でシステム確認中の誤解を検出できた（表 3の ALL 行）．この時，事例から学習された決定木をプロダクションルールの形式で表現したものを図 3に示す．

[8]適合率（precision）とは，検出器が「誤解を含んでいる」と検出した事例のうち，実際にシステム確認が誤解を含んでいた事例の割合．再現率（recall）とは，システム確認が誤解を含んでいる事例のうち，検出器が「誤解を含んでいる」と正しく検出できた事例の割合．

```
Rule 1      if     ( 継続長 ≤0.36 )                    then CR
Rule 18     else if ( スロット=部屋名 )
                   ∧ ( 継続長 ≤0.81 )
                   ∧ ( 先行ポーズ ≤0.23 )              then CR
Rule 8      else if ( スロット=曜日 )
                   ∧ ( 回数 = cnfm1 )
                   ∧ ( 継続長 ≤0.65 )
                   ∧ ( 継続長差 >0.21 )                then CR
Rule 10     else if ( スロット=時間 )
                   ∧ ( 継続長 ≤0.65 )                  then CR
Rule 11     else if ( スロット=部屋名 )
                   ∧ ( 継続長 ≤0.65 )
                   ∧ ( 継続長差 ≤ −0.13 )              then CR
Rule 20     else if ( 継続長 >0.81 )
                   ∧ ( 回数 = cnfm2 )                  then ME
Rule 12     else if ( 回数 = cnfm1 )
                   ∧ ( 継続長 >0.65 )                  then ME
Rule 9      else if ( 継続長 >0.36 )
                   ∧ ( 先行ポーズ >0.28 )              then ME
                   Defalut                             ME
```

図 3: システム確認中の誤解を検出する決定木（全特徴量を利用．プロダクション規則の形式で表現．規則は上から順に適用する．）

5 考察

対話はインタラクティブな過程なので，システム発話の内容に応じてユーザの応答にも変化が生じる．今回は，システム確認とその前後のユーザ発話の関係の特徴を調べることで，システム発話の内容（誤解を含むかどうか）を推定できることを示した．

システム確認が正しい場合，ユーザはシステムの確認を肯定するだけでよく，とりたてて余計なアクションを追加する必要がない．それに対してシステム確認が誤っている場合，ユーザは何とかしてシステムが誤解していることをシステムに伝えなければならないので，もう一度同じことを繰り返したりする．システムは自分自身の確認内容が正しいのか誤解を含んでいるのかを予め知ることはできない．しかしそれぞれの場合にシステム確認の前後でどんなことが起こるのか，その違いを知っていれば，観察可能なユーザの様子から自分の誤解に気づくことができる．

今回の分析結果では主に継続長に関わる特徴量（duration や dur.diff.）がシステム確認中の誤解の検出に貢献した．ユーザ発話の繰り返しを検出する目的で DP マッチングによる 2 つの音声パターンの間の距離を用いたが，必ずしもシステムの誤解の検出には役立たなかった．これは，この特徴量が同じ発話内容だけれども発話に別の表現が付加されてしまう例（「違います月曜日です」など）に対して必ずしも有効でなかったことや，正しいシステム確認の前後でのユーザの「はい」の繰り返しを検出してしまうことなどが原因と考えられる．今後，多様な対話データ中からユーザによる訂正の繰り返しを検出するには，より頑健かつ高精度な特徴量へと改善を加える必要がある．

実用的な音声対話システムでは，システム確認に含まれる誤解を検出しただけではタスクは遂行されない．ユーザの伝えたい内容（正しい値）を確定させるため，誤解の回復も考慮に入れたシステムでなければならない．

謝辞 日頃よりご指導いただくメディア情報研究部萩田紀博部長，有益な示唆をいただく対話研究グループの諸氏，実験と分析にご協力いただいた，木間良子さん，酒巻隆治さんに感謝致します．

参考文献

[1] S. J. Boyce. Designing the user interface for a natural spoken dialogue system. In *the 7th International Conference on Human-Computer Interaction, (HCI International '97)*, pp. 367–370, 1997.

[2] 平沢, 宮崎, 中野, 相川. 音声対話システムの誤解に対するユーザ応答の分析. 平成 12 年春季 音響学会講演論文集 3-8-10, pp. 85–86, 2000.

[3] J. Hirasawa, N. Miyazaki, M. Nakano, and K. Aikawa. New feature parameters for detecting misunderstandings in a spoken dialogue system. In *ICSLP-00*, Vol. 2, pp. 154–157, 2000.

[4] B. A. Hockey, D. Rossen-Knill, B. Spejewski, M. Stone, and S. Isard. Can you predict responses to yes/no questions? yes, no, and stuff. In *Eurospeech97*, 1997.

[5] 今井, 井ノ上, 橋本, 米山. 未知語処理のための繰り返し音声検出手法. 電子情報通信学会技術研究報告 SP99-26, pp. 1–6, 1999.

[6] E. Krahmer, M. Swerts, M. Theune, and M. Weegels. Problem spotting in human-machine interaction. In *Eurospeech99*, Vol. 3, pp. 1423–1426, 1999.

[7] G.-A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *COLING-ACL98*, 1998.

[8] S. Oviatt, G.-A. Levow, M. MacEachern, and K. Kuhn. Modeling hyperarticulate speech during human-computer error resolution. In *ICSLP96*, 1996.

[9] J. R. Quinlan. *C4.5: Programs for Machine Learning*. The Morgan Kaufmann, 1992.

[10] M. Swerts, D. Litman, and J. Hirschberg. Corrections in spoken dialogue systems. In *ICSLP-00*, Vol. 2, pp. 615–618, 2000.

[11] 田本, 川森, 相川. 自律型音声対話システムによる対話データの自動収集. 人工知能学会研究会資料 SIG-SLUD-A001-6, pp. 19–24, 2000.

[12] 宇津木, 竹内. 音声認識装置が音声を正しく認識しない事態における 人間の発声の変化. 人間工学, 31(4):287–293, 1995.