混合主導対話における音声認識誤りに対処するための対話管理

駒谷 和範　　河原 達也
京都大学 情報学研究科 知能情報学専攻
komatani@kuis.kyoto-u.ac.jp

概要

音声認識誤りに対処するための混合主導対話の実現法について述べる。音声認識器の10-best出力を利用して発話内容に関する信頼度(Confidence Measure; CM)を計算し、これに基づいて確認を効果的に行う。また意味カテゴリレベルにおいてもCMを計算し、単語レベルの認識がうまくいかない場合にユーザを誘導する発話を行う。評価実験は、ホテル検索タスクの音声対話システムを初心者24名が使用した際の音声データを用いて行った。意味解釈精度は、第一候補のみを認識結果とする場合と比較して、確認を行わない場合で5%、確認を行う場合では11%向上した。

1 はじめに

音声認識技術の向上を受けて、その応用である音声対話システムの研究が行われている。本研究ではホテル検索をタスクとして、音声対話を通じて情報検索を行うことを目標としている。

音声を通じて計算機と対話を行う際には、音声認識に誤りが生じたり、ユーザがシステムの想定していない発話を行うなどといった問題が頻繁に生じる。これらの問題は、システムの受理できる語彙や文法の範囲を広げたとしても、計算機で人間の音声や言語を扱う場合には本質的に避けられないものであるため、その対処は不可欠である。現在実用的に広く使われている音声対話システムが存在しないのは、この頑健性の欠如が大きな原因の一つであると考えられる。

したがって、実用的な音声対話システムを構築するには、音声認識の性能向上とともに、対話を通じて誤りを解消する枠組みも必要である。システムによって発話が制限されることなく、かつシステム側からも認識誤りに対処するために、混合主導対話(mixed-initiative dialogue)を実現する対話管理方法を考える。混合主導対話とは、ユーザ主導対話とシステム主導対話の長所を取り入れて、基本的にユーザに自由な発話を許しながらも、必要なときにはシステムからユーザに質問したりユーザを誘導したりする対話である。

対話システムの入力である音声認識結果に誤りが含まれる場合に確認を行うことの有効性については、現在までにも研究が行われている。[1][3]では数式を用いて、[2]では計算機同士のシミュレーションを用いて、確認を行う戦略やその有効性に関して議論している。本研究では、実際に対話音声の認識結果を用いて信頼度を計算し、それを用いて効果的な確認や誘導を行うことにより混合主導対話を実現する。

2 音声認識結果の信頼度(CM)の計算

「よく聞き取れなかった言葉について確認を行う」ということは、人間同士の対話でもよく行われることである。したがって、音声認識の結果が「認識誤りである可能性が高い」とわかることは、対話戦略を考える上で非常に有用である[4]。しかし計算機による音声認識は、入力された音声に対して最も尤度の高い単語列を出力するというプロセスであるため、正しい認識結果と認識誤りとを判別するためには何らかの尺度が必要である。そこで、この章では認識結果に対する信頼度(Confidence Measure:CM)を計算する方法を述べる。

2.1 単語に関するCMの計算

音声認識では、入力音声に対して尤度の高い順にn-best解を求めることができる。本研究では、認識エンジンとして本研究室で開発されたJulian[5]を用いており、解ごとにスコアが計算される。そこで、このn-best解のスコアを用いて、単語ごとのCMを求める。本研究では、$n = 10$とした。

1. n-best解の対数スケールの各スコア$score_i (1 \leq i \leq n)$から、最尤解のスコア$score_1$をひいたもの

i	認識結果	p_i
1	あー 施設 に レストラン の 加悦町	.24
2	あー 施設 に レストラン の 桂 の	.24
3	あー 施設 に レストラン の 上賀茂	.20
4	<g> 施設 に レストラン の 加悦町	.08
5	<g> 施設 に レストラン の 桂	.08
6	<g> 施設 に レストラン の 上賀茂	.06
7	あー 施設 に レストラン の カフェ	.05
8	<g> 施設 に レストラン の カフェ	.02
9	<g> 設備 を レストラン の 加悦町	.01
10	<g> 設備 を レストラン の 桂 の	.01

ただし、<g>: filler model

CM_w	(単語)	@	(意味カテゴリ)
1	レストラン	@	施設
0.33	加悦町	@	所在
0.33	桂	@	所在
0.25	上賀茂	@	所在
0.07	カフェ	@	施設

図 1: 単語の信頼度 (CM_w) の計算例

に、定数$\alpha(\alpha < 1)$をかける。

$$scaled_i = \alpha \cdot (score_i - score_1)$$

2. 1. で求めた$scaled_i$を対数スケールから元に戻し、i 番目の解の事後確率 p_iを求める [6]。

$$p_i = \frac{e^{scaled_i}}{\sum_{i=1}^{n} e^{scaled_i}}$$

3. ある単語 w が i 番目の解に含まれるとき$\delta_{w,i} = 1$とすると、w が正しい確率 p_wは、

$$p_w = \sum_{i=1}^{n} p_i \cdot \delta_{w,i}$$

で求める。

この p_wを単語 w の CM(CM_w) とする。

具体例として、「付帯施設にレストランのある宿」という発話の認識結果と、その内容語の CM を計算したものを図 1 に示す。

2.2 意味カテゴリに関する CM の計算

認識された各内容語には意味カテゴリを付与している。これは、有限状態オートマトン (FSA) で記述されている認識文法を意味カテゴリごとにわけておき、どの FSA に受理されたかによって判定したものである。内容語を持つ意味カテゴリは、現在「所在」「付帯施設」など 7 種類である。

この意味カテゴリ (Concept Category) に関しても CM を求める。まず単語の CM の場合と同様に、n-best 解の i 番目の解の事後確率 p_iを求める。ある意味カテゴリ c の内容語が i 番目の解に含まれるとき$\delta_{c,i} = 1$とすると、c が正しい確率 p_iは、

$$p_c = \sum_{i=1}^{n} p_i \cdot \delta_{c,i}$$

で求める。この p_cを意味カテゴリの CM(CM_c) とする。この CM_cは誘導発話の生成の際に用いる (3.2 節)。

3 音声認識誤りに頑健な対話の実現法

3.1 確認発話の生成

3.1.1 内容語の CM を利用した確認発話

2.1 節で述べた内容語に関する信頼度を用いて確認発話を生成する。2 つのしきい値$\theta_1, \theta_2 (\theta_1 > \theta_2)$ を設定すると、確認発話は以下のように行われる。

- $CM_w > \theta_1$
  → そのまま受理する (確認は行わない)
- $\theta_1 \geq CM_w > \theta_2$
  → 直接的に確認を行う
  「○○でよろしいですか？」
- $\theta_2 \geq CM_w$
  → 棄却する

内容語に関する信頼度 CM_wは内容語ごとに計算しているので、一発話内に複数の内容語が含まれている場合でも、その内容語ごとに受理／確認／棄却を決定することができる。全ての内容語が棄却された場合には、再発話をうながすことになる。

3.1.2 対話レベルの知識を利用した確認発話

情報検索というタスクにおいては、検索条件を追加／削除しながら対話が進んでいくため、このような対話の進行から大きく離れた発話は誤りである可能性が高い。その一例として、検索条件がすでに入力されている項目に対して、さらに上書きを行うような発話は、(CM が高いとしても) 認識誤りである可能性が考えられる。実際、短い地名の湧き出し誤りでは、音響的に

CMが高くなってしまうことがある。このようなものに対しても確認を行うことにより、湧き出し誤りによる誤受理を防ぐことが期待できる。

3.2 意味カテゴリのCMを用いた誘導発話

システムを使い慣れないユーザは、システムに自分の発話がなかなか受理されない場合には、どのように自分の要求を発話すればよいかわからなくなることがある。このような場合には、システム側から誘導を行うことが望ましい。

意味カテゴリのCMを利用すると、単語に関する認識がうまく行かない場合に有用な誘導を行うことができる。その例を図2に示す。

発話：「所在が大阪府の宿」
正解：大阪府@所在

認識結果：
1: 所在 が ポートアイランド の <g>
2: 所在 が ポートアイランド の <g>
3: 所在 が 大阪府 の <g>
4: 所在 が 大阪府 の <g>
5: 所在 が 大阪市 の <g>
6: 所在 が 大阪市 の <g>
7: 所在 が 岡崎 の <g>
8: 所在 が 岡崎 の <g>
9: 所在 が 大原 の <g>
10: 所在 が 大原 の <g>

CM_w	(単語)@(意味カテゴリ)
0.38	ポートアイランド@所在
0.30	大阪府@所在
0.13	大阪市@所在
0.11	岡崎@所在
0.08	大原@所在

CM_c	意味カテゴリ
1	所在

図2: CMを用いた誘導発話が有効な例

図2の例では、内容語に関するCMがどれもしきい値（θ_2）よりも低いので受理や確認は行われないが、意味カテゴリ[所在]のCMは高い。このような場合には、単純に入力を全て棄却して「もう一度言ってください。」と言うよりも、「所在がどこですか？」とユーザを誘導する方が、次発話の認識の際の語彙を絞り込むことにより認識誤りを少なくすることができる。

また、意味カテゴリのCMが高いのに内容語の認識が出来ない状況が続く場合には、内容語の部分が未知語である可能性が高いと推定できる。未知の地名が発声されたと推定できる場合には、「都道府県名（あるいは市町村名）から指定してください。」とユーザ発話を認識可能な語彙の範囲内に誘導することも考えられる。

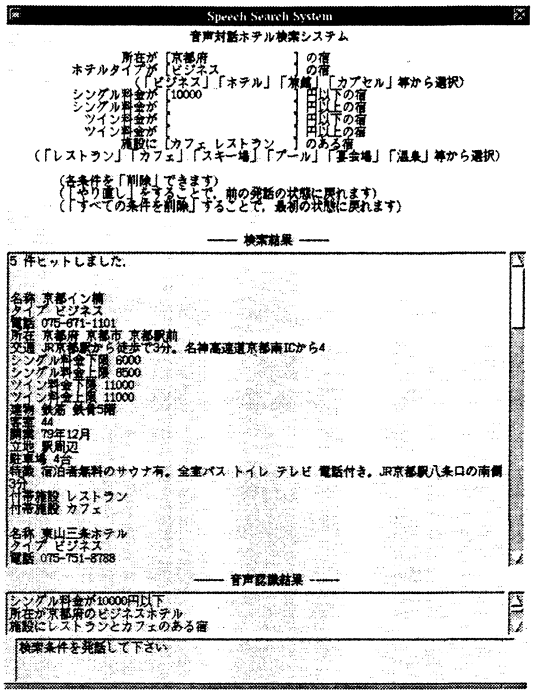

図3: ホテル検索システムのGUI[7]

4 評価実験

4.1 実験データ

音声対話システムを使ったことのない24名の話者に対して、関西地区のホテル検索システムであることや検索可能な項目、項目の削除の仕方などを教示し、全体で約120分間、GUI[7]（図3）の付いたシステムを使用してもらった音声を収録した。その音声データを1.25秒毎に区切り、雑音や息の音だけの部分を取り除いた結果、全部で705発話（約29発話/人, 最大64, 最小11）となった。

少しの雑音や助詞の省略などがあるものも含めて、システムが受理可能であると考えられる発話は581発話で、全体の82.4%であった。残りの124発話は、語彙外・文法外・タスク外・発話の断片などであるが、以下ではこれらも含めて実験・評価を行っている。

4.2 評価基準

認識結果の意味解釈精度の評価基準として、誤受理率(False Acceptance; FA)とスロットエラー(Slot Error;

表 1: 1-best 解のみを用いた手法との比較

	1best 解のみ	確認無	確認有
FA+SErr(%)	51.5	46.2	40.0

SErr) の和を用いた。誤受理率は、認識誤りを誤って受理してしまう率で、受理するしきい値を下げると増加する。スロットエラーは、正解が受理する部分に含まれていない率で、1 から適合率 (precision) を引いた値である。したがって、受理するしきい値を上げていくとスロットエラーは増加する。

$$FA = \frac{\text{受理した中で誤っていたスロット数}}{\text{受理したスロット数}}$$

$$SErr = 1 - \frac{\text{受理した正解スロット数}}{\text{実際の正解スロット数}}$$

4.1 節のデータを全て用いて、これらの和が最小となるように単語に関する信頼度 (CM_w) に関するしきい値を定めた結果、$\theta_1 = 0.9$ となった [8]。同様に、誤棄却率 (False Rejection; FR) と、確認を行う範囲での誤受理率の和が最小となる点として、$\theta_2 = 0.6$ とした。

$$FR = \frac{\text{誤って棄却したスロット数}}{\text{棄却したスロット数}}$$

正解数は発話単位ではなく、内容語 (=スロット) を単位とした。全正解数は 804 であった。

4.3 音声認識結果の第一候補のみを用いる場合との比較

認識結果の第一候補を用いる方法と本手法の意味解釈精度を比較した結果を表 1 に示す。「確認無」は θ_1 以上は受理、これ以下は棄却したもので、$\theta_1 = 0.9$ のときの FA(θ_1)+SErr(θ_1) の値である。「確認有」は $\theta_1 \geq CM_w > \theta_2$で確認を行った場合で、$\theta_1 = 0.9$, $\theta_2 = 0.6$ のときの FA(θ_1)+SErr(θ_2) の値である。

表 1 のように、10-best 解から CM_wを計算したことで精度が 5.3%向上し、また CM_wを利用して確認発話を行うことにより精度をさらに 6.2%向上させることができた。

4.4 意味カテゴリ推定の有用性

$\theta_2 = 0.6$ の場合、単語の CM からでは認識結果の候補が得られなかったスロットは 148 個あった。このうち、意味カテゴリの CM_cが 0.9 以上でかつ正解であるものは 34 個であるため、単語 CM だけでは棄却されていたスロットのうち 23%に対して有効な誘導発話を行うことができた (FA は 49%)。

5 まとめ

音声認識誤りに対してより頑健に対話を進めるために、内容語と意味カテゴリの 2 レベルで信頼度 (CM) を計算し、それを用いた対話戦略について述べた。

意味理解率は n-best 解を用いて CM を計算することにより、確認を行わない場合でも 5.3%、確認を行う場合では 11.0%の向上が見られた。また、意味カテゴリに関する CM を用いることにより、単語レベルの CM では棄却されるスロットのうちの 27%に対して、有用な誘導発話を生成する方法を示した。今後この意味カテゴリの CM の計算法を改良し、評価する予定である。

謝辞 本研究に対して、小笠原科学技術振興財団の支援を受けた。

参考文献

[1] 新美康永, 小林豊. 音声認識の誤りを考慮した対話制御方式のモデル化. 情報処理学会研究報告, 95-SLP-5-7, 1995.

[2] Watanabe T., Araki M., Doshita S.. Evaluating Dialogue Strategies under Communication Errors using Computer-to-Computer Simulation. Trans. of IEICE, Info & Syst., Vol.E81-D. No.9, pp.1025-1033, 1998.

[3] 新美康永, 西本卓也, 荒木雅弘. 確認対話の制御方式の効率と音声認識システムの性能との関係. 情報処理学会研究報告, 99-SLP-27-17, 1999.

[4] T.Kawahara, C.-H.Lee, B.-H.Juang. Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification. IEEE Trans. on Speech and Audio Processing, Vol.6, No.6, pp.558-568, 1998.

[5] 李晃伸, 河原達也, 堂下修司. 文法カテゴリ対制約を用いた A*探索に基づく大語彙連続音声認識パーザ. 情報処理学会論文誌, Vol. 40, 4, pp.1374-1382, 1999.

[6] G.Bouwman, J.Sturm, L.Boves. Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARISE Project. Proc. of ICASSP99, 1999.

[7] 田中克明, 河原達也, 堂下修司. 汎用的な情報検索音声対話プラットフォーム. 電子情報通信学会技術研究報告, SP98-109, NLC98-45 (98-SLP-24-14), 1998.

[8] 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた頑健な混合主導対話の実現法. 情報処理学会研究報告, 00-SLP-30-9, 2000.