発話状況の情報を音声認識に用いた音声対話システム

6Ｈ－7

荒川直哉・加藤直人・竹澤寿幸・森元逞

㈱ＡＴＲ音声翻訳通信研究所

1　はじめに

発話状況あるいは文脈の情報は、音声言語処理において意味論または語用論的な解析のために用いることができる一方、音声認識や構文解析においても制約あるいは統計的選好として利用することができる。ここで発話状況の情報としては、話者の役割（サービス提供者か享受者かなど）、先行する発話列の発話タイプ（<question> <inform> など）、対話構造（答えられていない質問の有無など）、話題、などが考えられる。我々は発話状況の情報を音声言語システムに統合的に利用していく研究を進めているが、今回は音声対話システムへの、文脈情報を用いて音声認識候補の再順序づけを行う機構に関する報告を行う。

2　音声対話システム

現在、ユーザからの音声入力に対してシステムが音声で応答する音声対話システムを試作している[1]。このシステムのタスクは「奈良近辺のホテルの予約業務」である。システム構成の概観を図１に示す。今回の報告では音声言語統合処理部での発話状況管理部からの情報の利用について述べる。

音声認識＋構文解析部（図１）はHMM法による統計的な音声認識とLR表による構文解析と同時進行的に行うことにより、文法的に不適切な認識結果を排除する（HMM-LR法）。

今回、我々は文脈情報を用いて、音声認識＋構文解析部の出力結果（通常１つの発話に対して複数の認識候補を含む）のスコアリングを行い、再順序づけを行うような仕組みを導入した。文脈情報は発話状況管理部から提供される。

発話状況管理部は、話者情報（誰が話しているのか）の管理、発話タイプの推定、話題の抽出、対話構造の認識、照応・省略解析を行う。今回の報告ではこれらのうち、前者４項目の情報の利用について述べる。

3　文脈情報の種類とその音声認識への利用

特定の言語表現があらわれる確率はそれらが用いられる文脈に依存すると考えられる。例えば、金額が話題になっている場合には「ＸＸ円」などの金額の値に関する表現があらわれやすいであろう。音声認識では認識結果候補である言語表現の出現確率を候補のスコア付けに用いる（後述）ので、こうした出現確率の文脈依存性の利用が可能である。

3.1 話者の役割情報の利用

話者の役割とは例えばサービス提供者か享受者かなどの区別のことである。サービス提供者と享受者は特定の言語表現の使用頻度が異なる（例えばサービス提供者の方が丁寧な表現を用いる）と考えられる。我々の対話システムにおいては音声認識の対象となる話し手は常に対話システムに話しかける人間のユーザであるから、音声認識用のデータを求める際に、サービス享受者側の発話のみから統計を取ることにより、音声認識の性能向上を図っている。

3.2 発話タイプ推定の利用

ここで発話タイプとは、発話の機能あるいは発話によって話者が行おうとする行為の分類を指す。我々は発話タイプとして <question> <inform> など３１種類を用いている。日本語の場合、発話タイプは文末表現と関連がある（例えば <question> に分類される発話は「か」で終わることが多い）ため、音声認識の対象となる発話がどのような発話タイプを持つかの確率を計算することができれば、その発話の文末表現が現れる確率を計算することが可能になる [2]。発話タイプの出現確率は先行する発話の

Spoken Dialog System with Context-Sensitive Speech Recognition
Naoya ARAKAWA, Naoto KATOH, Toshiyuki TAKEZAWA, Tsuyoshi MORIMOTO
{arakawa, katonao, takezawa, morimoto}@itl.atr.co.jp
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, 619-02 JAPAN

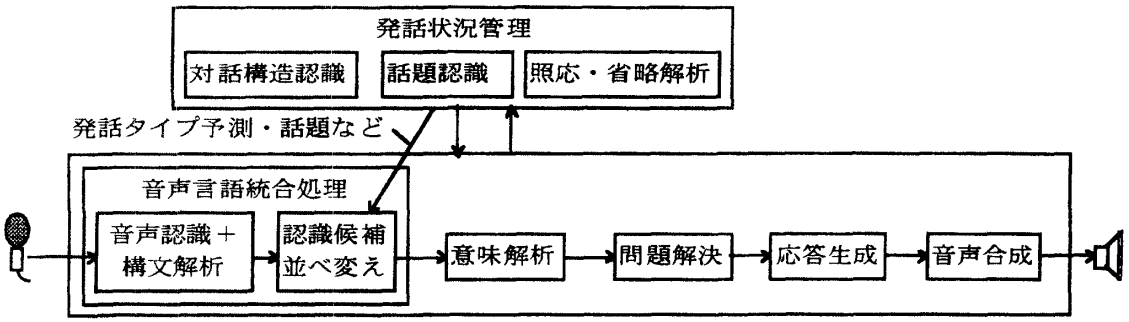

図1

文末表現から統計的に求められる。推定された発話タイプに基づいて計算される文末表現の出現確率は音声認識候補の出現確率を計算するために用いられ、結果的にその再順序づけに利用される。

3.3 話題情報の利用

特定の話題について話している場合、特定の語彙の出現頻度が高くなると考えられる。例えば病気が話題になっている場合には「体温計」といった表現が現われやすいであろう。

話題としては例えば「部屋の種類」、「ルームタイプ」という表現に対して *RoomType* という話題を設定したり、あるいは対話中で係助詞（「は」など）に先行する名詞句などをそのまま用いたりする。音声認識に対して有用な話題のリストはタスクドメインを考慮して人為的に選んだり、後続の発話中の特定の表現と生起相関の高い表現を用いたりすることを試みている [3]。

3.4 対話構造の利用

ここで対話構造とは「質問－応答－補足」といった“インタラクション構造”を指す [4]。音声認識において対話構造を考慮する理由は、ある質問が答えられていない場合には、その質問項目（＝話題）に関連するような言語表現が現れる可能性が高いのではないかと推測できるからである（図2）。

3.5 音声認識の「言語モデル」

統計的な音声認識では、言語表現の認識を行う際に、その言語表現が生起する確率を計算することが一般的に行われる。よく用いられる「言語モデル」はNグラムであり、w_i をi番目の形態素とするとその生起確率は $P(w_i|w_{i-N+1},\ldots,w_{i-1})$ で表される。ここで w_i の生起確率を計算する際に今まで述べてきたような文脈をCと表わせば、$P(w_i|w_{i-N+1},\ldots,w_{i-1},C)$ となる。言語表現全体の生起確率は $\prod_i P(w_i|w_{i-N+1},\ldots,w_{i-1},C)$ で表わされる。

文脈情報による音声認識候補の並び変えにおいては、各候補について言語表現の生起確率を文脈付きNグラム（通常 N=2）により再計算し、音響スコアと統合した結果を用いてソーティングを行うことを試みている。将来的には文脈付きNグラムを音声認識本体に組み込むことも考えられる。

参考文献

[1] 巖寺俊哲, 竹澤寿幸, 田代敏久, 加藤直人, 石崎雅人, 森元逞: “ポーズ単位に基づく音声言語統合処理と発話状況管理の統合 －音声対話システムの試作－,” 1996年度電子情報通信学会総合大会論文集 情報・システム1, pp.331-332 (1996).

[2] M. Nagata, T. Morimoto: “An Information-Theoretic Model of Discourse for Next Utterance Type Prediction,” *Transaction of Information Processing Society of Japan*, 35 no.6, pp.1050-1061 (1994).

[3] N. Katoh, T. Morimoto: “Statistical Method of Recognizing Local Cohesion in Spoken Dialogues,” COLING96, 2, pp.634-639 (1996).

[4] 巖寺俊哲, 石崎雅人, 森元逞: “対話のインタラクション構造と話題の認識,” 情報処理学会自然言語研究会報告 104-16, pp.119-126 (1994).

客　：「いくらですか。」
ホテル：「シングルですか、ツインですか。」
客　：「シングルです。」
ホテル：「八千円です。」

対話構造

客が「シングルです。」と発話した時点ではまだ最初の金額に関する質問が答えられていないので、継続して金額についての表現が出現する確率が高い。

図2