発話状況に基づく音声認識候補の再順序付け

岩本 秀明　妹尾 正身†　荒川 直哉　横尾 昭男　森元 逞
ATR 音声翻訳通信研究所
†NTT ソフトウェア

1　はじめに

音声認識候補の出現確率を発話状況の情報を用いて再計算し、再順序付けする方法について述べる。発話状況として、前回発話の話者役割、内容語、文末表現および現在の発話の話者役割を用いる。

ここで背景となっている考え方は、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発話状況に依存するということである。音声認識を含めた言語処理では、こうした言語表現の出現確率の発話状況依存性を処理結果候補の尤度計算に用いることができる。

発話の状況依存性については、従来より、質問、要求、およびそれに対する応答といった発話の機能に基づいた、発話タイプ[1]やその表層である文末表現[2]の予測について研究が進められてきた。近年、これを実際に音声認識[3]や音声翻訳システム[4][5]に適用しようとする試みがなされているが、発話タイプや文末表現に加えて、さらに、話題などの補完的な情報を利用する必要性が認識されている[6]。

このような観点も含めて、プラン知識に基づいた強力な手法が従来から提案されている[7]。本研究では、知識の構築と運用に関わる困難さを避けるために、コーパスから比較的容易に獲得できる内容語に着目する。

なお、本研究では音声認識候補の再順序付け(図1)によって発話状況情報の有効性を検証しようとしているが、これは音声認識装置で直接発話状況情報を利用するための予備実験である。

第2節で、発話状況について述べる。3節で、具体的な再順序付け手法について述べる。4節で、その手法を用いた予備実験について報告する。

2　発話状況

本研究では、タスクドメインにホテル予約を取り上げる。これにより、対話の担い手は、ホテル側の担当者と旅行客等の申込者とに限定され、発話状況としては、先行する発話自体が主な情報を持つ。

2.1　発話とその話者役割

以下、現在の発話(の任意の音声認識候補)をU_cとし、前回の発話をU_pとする。また、U_cの話者役割をS_c、U_pの話者役割をS_pとする。今回の実験では、話者役割S_pおよびS_cは、ホテル側の担当者、あるいは旅行客などの申込者のいずれかである。

【例】 S_p = "担当者"
U_p = 「お部屋のタイプはどうなさいますか」
S_c = "申込者"
U_c = 「シングルをお願いします」

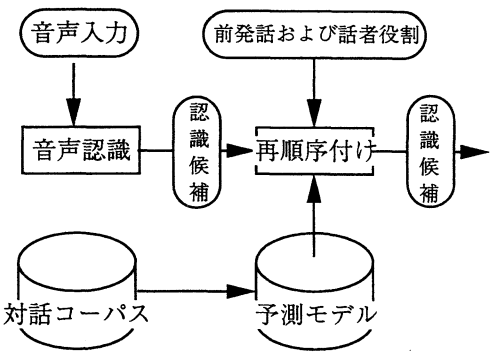

図 1: 構成

2.2 発話の構成要素

前回発話

発話 U_p は、内容語 $\langle V_1, \ldots, V_l \rangle$ と文末表現 E_p を含む。

内容語とは、「部屋の種類」や「料金」などの話題に関わる表層の単語を指す。以下では、自立語および接尾辞を内容語とみなしている。

【例】「部屋」「タイプ」「シングル」（普通名詞）
「キャンセル」（サ変名詞）
「1」「0」（数詞）
「ドル」（接尾辞）
「待」（本動詞）

日本語の文末表現は、質問や要求、それらへの応答などの発話タイプを表現する。発話 U_p の発話タイプ I_p は、文末表現 E_p から推定される。

【例】「～どうなさいますか」$<wh-question>$
「～お願いします」$<action-request><inform>$ etc.

候補発話

現在の発話に対する音声認識候補の形態素列を $U_c = \langle W_1, W_2, \ldots, W_m, \ldots, W_n \rangle$ とする。U_c における文末表現を $E_p = \langle W_{m+1}, \ldots, W_n \rangle$ とする。また、各 W_i のプレターミナルを C_i とする。プレターミナルとは、品詞に活用型や活用形などの情報を加えて細分化した単語のカテゴリである。

3 再順序付け

本手法は、第2節で述べた発話状況から、形態素列として出力される音声認識候補の出現確率を再計算し、その状況に照らして、尤もらしい順序に音声認識候補の並べかえを行なう (図 3)。

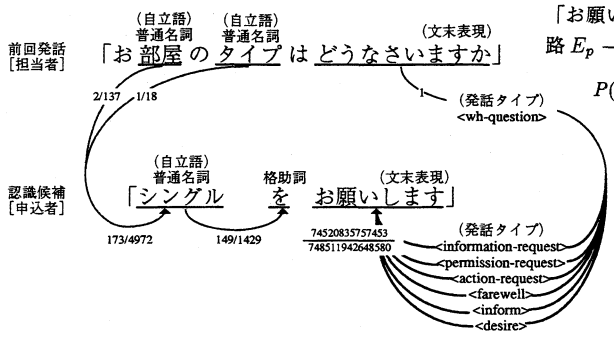

図 2: 確率計算例

3.1 予測モデル

発話タイプおよびプレターミナルが付与された対話コーパスから以下の統計情報をあらかじめ学習しておく。

$P(W_i \mid S_p, V_j, S_c, C_i)$: 前発話の内容語と話者役割を考慮した形態素ユニグラム

$P(C_i \mid C_{i-1})$: プレターミナルバイグラム

$P(I_p \mid S_p, E_p)$: 文末表現に対する発話タイプ

$P(I_c \mid S_p, I_p, S_c)$: 発話タイプのバイグラム

$P(E_c \mid S_c, I_c)$: 発話タイプに対する文末表現

3.2 確率再計算

前発話が与えられたとき、形態素列として出力される音声認識候補の出現確率を以下のように、形態素列と文末表現の出現確率の積として計算する。

$$P(U_c \mid U_p, S_c) = [\prod_{i=1}^{m} P(W_i \mid U_p, S_c, C_{i-1})] P(E_c \mid U_p, S_c, C_m)$$

ここで、W_i が内容語のとき (図 2の「シングル」)、

$$\begin{aligned} P(W_i \mid U_p, S_c, C_{i-1}) &= P(C_i \mid U_p, S_c, C_{i-1}) \\ & \quad P(W_i \mid U_p, S_c, C_i) \\ &\simeq P(C_i \mid C_{i-1}) P(W_i \mid U_p, S_c, C_i) \\ P(W_i \mid U_p, S_c, C_i) &= P(W_i \mid S_p, \langle V_1, \ldots, V_l \rangle, S_c, C_i) \\ &= \frac{1}{k} \sum_{j=1}^{l} P(W_i \mid S_p, V_j, S_c, C_i) \end{aligned}$$

とする。W_i が機能語のとき (図 2の「を」)、

$$P(W_i \mid U_p, S_c, C_{i-1}) \simeq P(W_i \mid C_{i-1}) = P(C_i \mid C_{i-1})$$

となる。前発話の文末表現を条件とした、候補発話の文末表現の出現確率 (図 2の「どうなさいますか」から「お願いします」) を計算するために、可能な全ての経路 $E_p \rightarrow I_p \rightarrow I_c \rightarrow E_c$ を以下のように計算する。

$$\begin{aligned} P(E_c \mid U_p, S_c, C_m) &\simeq P(E_c \mid U_p, S_c) \\ &\simeq P(E_c \mid S_p, E_p, S_c) \\ &= \sum_{I_p} [P(I_p \mid S_p, E_p) \\ & \quad \sum_{I_c} P(I_c \mid S_p, I_p, S_c) \\ & \quad P(E_c \mid S_c, I_c)] \end{aligned}$$

4 評価実験

4.1 評価方法

再順序付け結果に対して、

1. 1位候補の単語正解率
2. 認識候補の各要素の単語正解率を計算し、単語正解率により音声認識候補を並べかえた結果を正解とみなす。その正解の並びとの順序正解率

という2つの基準で評価する。以下に単語正解率と順序正解率とについて述べる。

単語正解率

音声認識結果の評価基準として用いられている単語正解率 [8] について説明する。正解単語列と候補単語列とに対して、正解 (Correct)、置換 (Substitue)、削除 (Delete)、挿入 (Insert) の単語数を DP マッチにより求め、以下の比率を計算する。

$$[\text{単語正解率}] = \frac{[\text{正解}] - [\text{挿入}]}{[\text{正解}] + [\text{置換}] + [\text{削除}]}$$

正解 正解単語列と候補単語列とで一致する単語数

置換 正解単語列に現れない候補単語の中で、正解単語の代わりに現れる単語数

削除 正解単語列の中で、候補単語列に現れない単語数

挿入 正解単語列に現れない候補単語の中で、正解単語の置換ではない単語数

【例】 c: 正解、s: 置換、i: 挿入、d: 削除

正解「シングル　を　お　願　い　し　ま　す　」
候補「シングル/c で/s お/c 待/s ち/s し/c ま/c す/c」

正解「シングル を　　お　願　い　し　ま　す　」
候補「1/s　0/s ドル/i で/i お/c 願/c い/c し/c ま/c す/c」

正解「シングル　を　お　　願/d い/d し/d ま/d す　」
候補「キャンセル/s が/s ございま/s　　す/c」

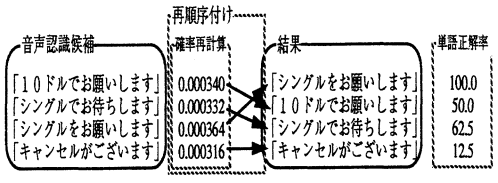

図 3: 再順序付け

順序正解率

本手法の効果を下位の候補間において確認するために、単語正解率からみた正解順序との距離を用いる。以下、そのための尺度である順序正解率について述べる。

音声認識により順位付けされた認識候補あるいは本手法により再順序付けされた候補を S とする。また、S の各要素の単語正解率を計算し、その正解率の高い順に並べかえた候補を T とする。S における順位 i の要素を S_i とする ($1 \leq i \leq n$、n: 要素数)。このとき、S_i の T での順位を $nth(T, S_i)$ として、入れかえ距離 $d(T, i)$ を次のように定義する。

$$d(T, i) = |nth(T, S_i) - i|$$

図 4に、図 3の例における、音声認識候補および再順序付け結果の単語正解率から見た正解順序への入れかえを示す。また、逆順の場合の入れかえも示す。この逆順の入れかえにおいて、12.5% の要素、すなわち、1位から4位への入れかえ距離は 3 である。同時に、100.0% の要素の入れ換え距離も 3 となる。

さらに、S を T に並べかえる、並べかえ距離 $D(T, S)$ を、以下のように定義する。

$$D(T, S) = \frac{1}{2} \sum_{i=1}^{n} d(T, i)$$

ここで、T の順序を逆にしたものを R とすると、R の T への並べかえ距離 $T(S, R)$ は、以下の通りである。

$$D(T, R) = \begin{cases} \sum_{i=1}^{\frac{n}{2}} (2i - 1) & = \frac{n^2}{4} \quad \text{if } n \bmod 2 = 0 \\ \sum_{i=1}^{\frac{n-1}{2}} 2i & = \frac{n^2-1}{4} \quad \text{if } n \bmod 2 = 1 \end{cases}$$

S の T に対する順序正解率は、上記の逆順並べかえ距離を基準にして、以下のように定義する。

$$[\text{順序正解率}] = \frac{D(T, R) - D(T, S)}{D(T, R)}$$

図 3、4における音声認識候補および再順序付け結果の順序正解率は、それぞれ以下の通りとなる。

$$(4 - 3)/4 = 0.25, \quad (4 - 1)/4 = 0.75$$

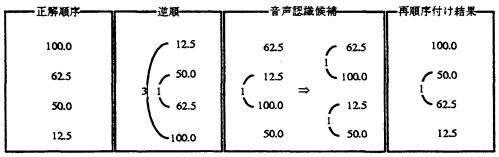

図 4: 並べかえ距離

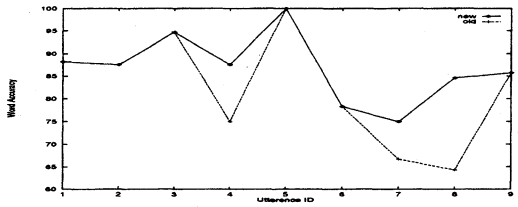

図 5: 単語正解率

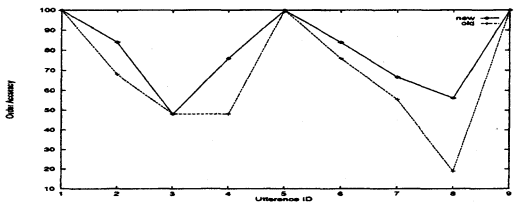

図 6: 順序正解率

4.2 実験結果

実験結果を図5、6および表1に示す。

図5、6は、3節で述べた手法の評価実験を行なったとき、最も精度が高かった対話データである。横軸は、発話番号であり、図5、6の縦軸はそれぞれ、単語正解率と順序正解率である。これらの図からも明らかなように、本手法では、1位候補の単語正解率が同じ場合でも、2位以下の候補の並べかえがありうる。順序正解率では、その部分の効果も含まれている。

表1に1位候補の単語正解率および全ての候補の順序正解率を示す。3節の手法から、内容語あるいは発話タイプに関連する計算を省いて、実験条件を異ならせた。

実験には、ホテル予約会話の7データを用いた。これらは、発話数90、単語数1324である。学習データは、旅行会話の199データを用いた。ホテル予約の7データは、この199データに含まれる。順序正解率では、音声認識候補が1つしかない場合や音声認識候補の各正解率が全て等しい場合を除外し、50発話で評価した。

表 1: 実験結果

実験条件	単語正解率 (%)	順序正解率 (%)
音声認識出力	80.09	66.59
内容語あり発話タイプあり	79.50	70.39
内容語あり発話タイプなし	80.54	69.90
内容語なし発話タイプあり	78.85	62.86

5 考察

本手法と音声認識結果とでは1位候補の単語正解率に有意な差は見られないが、順序正解率は3.8% 向上した。

単語正解率および順序正解率どちらかみても、内容語を使用しない場合が、精度が低い。これは、内容語を用いることの有効性を示している。また、これは同時に発話タイプを使用する場合の精度が低いことを示している。しかし、両者を組み合わせた場合の順序正解率が最も高い。内容語と発話タイプとは、次発話予測方法として、相互に補う仕組みで統合できる可能性がある。

6 おわりに

従来、内容語の次発話予測への効果は明らかでなかったが、発話状況に基づいた音声認識候補の再順序付け実験を予備的に行ない、発話間の内容語が音声認識の尤度計算に効果の高いことを示した。

今後は、発話間の内容語と発話タイプとを統合的に考慮した、音声認識あるいは言語解析の尤度計算方法を検討していく。

参考文献

[1] 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

[2] N. Katoh, T. Morimoto, "Statistical Method of Recognizing Local Choesion in SpokenDialogues", Proc. of COLING-96, Vol.2, pp.634-639, 1996

[3] 巌寺 俊哲, 竹澤 寿幸, 石崎 雅人, 森元 逞, "次発話予測による音声認識結果の再順序付け", 情報処理学会全国大会, 1996-09

[4] Reithinger, N. Engel, E., Kipp, M and Klesen, M., "Predicting Dialogue Acts for a Speech-to-Speech Translation System", Proc of ICSLP-96, 1996

[5] Lavie, A., Levin, L., Qu, Y., Waibel A., Gates, D., "Dialogue Processing in a Conversational Speech Translation System", Proc of ICSLP-96, 1996

[6] 荒川 直哉, 竹澤 寿幸, 加藤 直人, 森元 逞, "発話状況の情報を音声認識に用いた音声対話システム", 情報処理学会全国大会, 1997-03

[7] 山岡 孝行, 飯田 仁, "階層型プラン認識モデルを利用した次発話予測手法 – 話し手の意図を表す表現についての音声認識結果曖昧性の解消", 電子情報通信信学会論文誌, vol. J76-DII, No.6, 1993

[8] Cambridege University Engineering Department Speech Group and Entopic Research Laboratories Inc., "HTK : Hidden Markov Model Toolkit V1.5", 1993