会話音声認識のための話題別言語モデルの適用

助則 篤志　　樽松 明
電気通信大学情報システム学研究科
{suke,kure}@apple.ee.uec.ac.jp

1 はじめに

音声認識において読み上げ文発話だけでなく自由発話を研究の対象とすることは日常生活において重要である。日常生活で、会議や旅行の場所や時間を相談するような会話には、それぞれ固有の話題が存在する。このような会話をスケジューリングタスクと呼ぶ。

自由会話音声において、ワードスポッティングを用いた話題のラベリングを行なう研究[1]が行なわれ、90%程度の話題正解率が実現されている。本論文は、将来的に、音声の話題抽出と同時に、話題によって使用する言語モデルを振り分ける機能を装備することを念頭に、スケジューリングタスクにおける話題別の言語モデルの適用の有効性を確認するためのものである。

本論文ではスケジューリングタスクの自由発話音声の認識率の向上を行うために、言語モデルを話題別に独立したものを作成し、それぞれの話題に合致したものを使用することの有効性について、実験を通して検討した結果をのべる。言語モデルを話題別に作成することの有効性を示し、音声認識率で評価する。

2 使用コーパス

2.1 コーパス

スケジューリングタスクのコーパスは2人の話者が、海外出張のための打ち合わせということで、そのための日程、交通、宿泊、現地での仕事と観光予定を話し合いによって決定する事を想定して会話したものを録音し、書きおこしたものである。会話の際には、資料として、2人のそれぞれのスケジュール帳、飛行機の時刻表、出張先でのホテルの資料、現地の観光案内資料が手渡され、会話はそれらの下で行われる。テキストコーパスには、かな漢字、ローマ字文節区切り、ローマ字単語区切りの3種類がある。会話の個数は50会話であり、会話の長さは平均81ターンである。使用したコーパスの例を図1に示す。

図 1: テキストコーパスの一部

```
j001ach1_000_BAA_000000: <Smack> ~kasaha
ra~arısa desu .
j001ach2_001_PAA_000000: <B> ~kadowaki ~
masakazu desu .
j001ach1_002_BAA_000000: <B> ~kadowakı <
:<#Rustle> saN:>o <:<#Rustle> mata-se:>
shi-mashi-ta . <P> <e> <P> shucchou no ke
N kime-mashou-ka .
```

2.2 テキストの雑音処理

テキストコーパスには、会話文の他に、多くの種類の雑音が情報として含まれている。雑音の種類は以下の通りである。

表 1: 雑音の一覧

タグ	情報
<#Rustle>	紙をめくる音
<Smack>	舌をならす音
<Throat>	喉をならす音
<Cough>	咳をする音
<B>	息つぎ
<P>	ポーズ(無音)
<#Knock>	ノック、またはマイクへの接触音
<: :>	音声と雑音が重なった部分
<h>	息を吸う、または吐く音
<#>	その他のノイズ。ペンで字を書いている音など

表1に挙げた種類のノイズは、会話中のどこで(どの語の前で、どの語のあとで)挿入されるか

ということが予想できないため、bi-gram 確率は与えないこととした。未知語として処理されるため、back-off 時にごく小さな確率が与えられる。

2.3 言い間違い、言い直し等

表 2: 言い間違い、言い直し等の例

タグ	情報
<!..... >	不正確な発音
+/...../+	話の途中の言い直し
-/...../-	話が完結した後での言い直し

話者が、どこで発音違いをしたり、言い直しをするかということは予想することができないため、表 2の情報は削除し、言い直し、言い間違いが無かったこととしてモデルを作成した。

2.4 話題分けと複合語

図 1のようなテキストコーパスから、話題に分ける作業を人手により行ない、さらに複合語を用いないもの (A) と、用いるもの (B) の 2 種類に分けることとした (表 3)。複合語は、接頭語、接尾語、連語などのほか、一部の助動詞と助詞のつながりを含む。複合語を用いる場合は、図 1の'omata-se-shi-mashi-ta' は 1 語と扱い、複合語を用いない場合は 5 語と扱う。話題分けと、複合語の扱いにより、作成した言語モデルは以下の 12 種類である。

表 3: 言語モデルの種類

モデル名	話題	複合語
全体モデル A	全て	×
全体モデル B	全て	○
挨拶モデル A	挨拶	×
日程モデル A	日程	×
飛行機モデル A	飛行機	×
ホテルモデル A	ホテル	×
観光モデル A	観光	×
挨拶モデル B	挨拶	○
日程モデル B	日程	○
飛行機モデル B	飛行機	○
ホテルモデル B	ホテル	○
観光モデル B	観光	○

3 言語モデルの作成

言語モデルは、49 会話 (No1 ～ 49) のテキストコーパスを用い bi-gram モデルを作成した。bi-gram 確率は、

$$P(w_i|w_{i-1}) = \frac{P(w_{i-1}w_i)}{P(w_{i-1})} \approx \frac{N(w_{i-1}w_i)}{N(w_{i-1})} \quad (1)$$

として求めたものである。[2][3]

また、back-off 平滑化のディスカウンティングには Good-turn-discounting[4] を用い以下の様に計算した。

長さ 2 の単語列のうち、n 回出現するものの種類を R_n とする。

$N(w_{i-1}) = n$ とするとき、N を単語 2 つ組の総出現回数として

$$P(w_i|w_{i-1}) = \begin{cases} d_n f(w_i|w_{i-1}) & \text{if } n > 0 \\ \beta(w_{i-1})\alpha P(w_i) & \text{else if } N(w_{i-1}) > 0 \\ P(w_i) & \text{otherwise} \end{cases} \quad (2)$$

ただし、

$$\alpha = (1 - \sum_{N(w_{i-1}^{i})>0} P(w_i))^{-1} \quad (3)$$

$$d_n = \frac{(n+1)R_{n+1}}{nR_n} \quad (4)$$

$$\beta(w_{i-1}) = 1 - \sum_{N(w_{i-1}^{i})>0} P(w_i|w_{i-1}) \quad (5)$$

4 言語モデルの評価

作成した言語モデルを、perplexity. バイグラムのヒット率. 未知語率のそれぞれの値、音声認識結果によって評価する。評価用に用いたコーパスは 1 会話 (No.50) である。単語パープレキシティは、ある単語 1 個が出現する平均的な確率の逆数で定義され、

$$PP = (P(w_1 w_2 ... w_n))^{-\frac{1}{n}} \quad (6)$$

となる。バイグラムのヒット率、未知語率はそれぞれ、テストセット中の bi-gram の中で言語モデ

ル中に存在するものの確率、テストセット中の語彙の中で、言語モデル中に存在しないものの確率である。

4.1 perplexity 等による評価

12 個の言語モデルのうち、日程モデル A を例にとって、話題の合致した日程文をテストセット文としたときの perplexity. バイグラムのヒット率. 未知語率を示したものを図 2.3.4に示す。

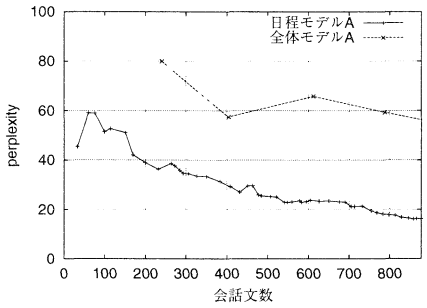

図 2: 日程文に対する perplexity

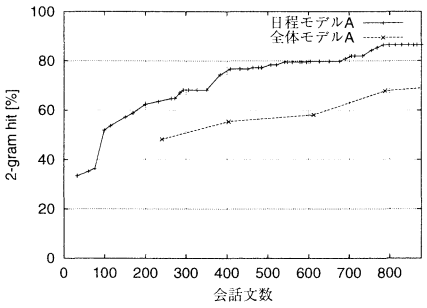

図 3: 日程文に対する bi-gram hit

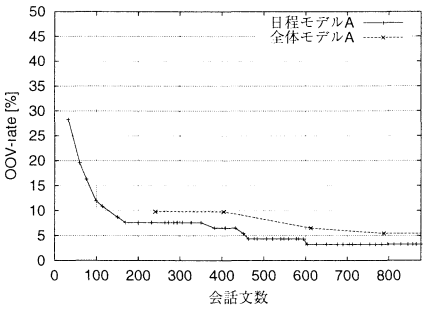

図 4: 日程文に対する OOV-rate

図 2の通り、同サイズのモデルとして比較した場合、話題が一致したモデルを用いた場合に perplexity が低下していることを確認できた。また、ここで示していない他の話題についても、話題が一致したモデルで、平均 30 程度 perplexity が低下している。

4.2 音声認識結果

作成したそれぞれの言語モデルを用いた音声認識結果を示す。男性話者の音響モデルと雑音モデル [5] を用いた。正解文 nijuu hachi nichi no suiyoubi made に対する認識結果は以下の通りとなった。

表 4: モデル別の認識結果

言語モデル	認識結果
全体モデル A	juu hachi ee chuu su juu bi made
日程モデル A	juu hachi ee chuu suiyoubi made
全体モデル B	juuhachi ee chuu su juu bi made
日程モデル B	juuhachiji ee juu suiyoubimade

4.3 単語認識率

それぞれの話題の文を認識対象の文とした時に、各言語モデル別の単語認識率の変化を、言語モデルを用いない際の単語認識率を基準として表 5に示す。

5 考察

- 言語モデルとして複合語を扱うかどうかは、一概にはどちらが良いとはいえない結果がでているが、複合語を用いることで、認識率の低下はきたさないことがわかった。
- 言語モデルを話題別に作成することによる認識率の向上が見られた。ただし、話題によってはそれほど結果が良くならなかったものがある。これらの点は以下のような点によって解消できると思われる。
  - 単語クラスをまとめた品詞 N-gram を採用する事。(とくに固有名詞に対して)
  - bi-gram から tri-gram へ移行すること。
  - 学習データの増加

表 5: 単語認識率の変化 (%)

モデル	全ての文	挨拶文	日程文	飛行機文	ホテル文	観光文
全体モデル A	+4.31	+1.30	+18.48	+1.30	± 0.00	+2.30
話題合致モデル A		+14.32	+20.58	+1.25	+8.19	+2.54
全体モデル B	+1.51	+4.80	+19.91	+5.65	+1.63	−4.48
話題合致モデル B		+9.56	+28.15	+6.04	+11.47	+4.95

– 発音間違いや外来語に対する柔軟性のある辞書を作成すること。

6 おわりに

連続音声認識の性能を向上するために、スケジューリングタスクの自由発話データを用いて、言語モデルを話題別に作成し、これを適用する効果を調べた。話題は、「挨拶」、「日程」、「飛行機」、「ホテル」、「観光」の 5 個に分けた。話題別に作成した言語モデルを適用することは、perplexity 値を用いても、実際に音声認識に用いても効果をはっきりと認めることができた。話題別言語モデルは、話題に合致したものからのみ学習しているため、容量も小さく済み、利点は大きい。ただし、話題によって効果の高いものとそうでないものの差があるため、さらに改良の余地があるといえる。

言語モデルによる効果は見られたものの、今回使用したスケジューリングタスクに対する音声認識システム全体をみた場合、まだ認識率の低さが目立つため、今後は、多方面での改良が望まれる。言語モデルの面からみた場合、固有名詞の扱い、ノイズの問題、言い間違いや言い直しの問題等を改善することにより、システム全体の能力の向上が期待できる。

今後、ワードスポッティングによる話題抽出 [1] の技術を利用し、人手によらずに話題分けを行い、自動的に言語モデルを切り替える事を課題と考える。

参考文献

[1] 五十嵐基仁:「自由会話音声におけるラベリング手法」. 日本音響学会平成 9 年度秋期研究発表会講演論文集,pp.137-138(1997.9)

[2] 中川聖一:「確率モデルによる音声認識」, コロナ社

[3] F.Jelinek:"Self-organized language modeling for speech recognition",Language Processing for Speech Recognition,pp.450-506.Merc- el,Dekker,Inc.(1990)

[4] S.M.Katz: "Estimation of probabilities from sparse data for language model component of a speech recognizer",IEEE Trans.ASSP.
vol.35,pp.400-401(1987)

[5] 北村道大、樽松明:「自由発話音声における HMM を用いた音声と雑音の識別」, 日本音響学会講演論文集,2-Q-14,(1996.9)