スケジューリングタスクにおける 自由発話音声データの収集と音声認識への応用

中筋 知己　樽松 明
電気通信大学 大学院 情報システム学研究科
{nakasuji,kure}@apple.ee.uec.ac.jp

1 はじめに

音声や言語の研究を遂行するための基礎となるデータベース (コーパス) の重要性は広く認識されてきている。第一の目的は、言語モデルを統計的に取得するためである。このためには新聞などの既存の大規模なデータの利用が行なわれているが、これは書き言葉であるため、自然な音声インタフェースのモデルとして用いることはあまり適切とはいえない。話し言葉 (自由発話) のデータベースについては、ATR が旅行会話のタスクで収集を行なう [1] など、規模の大きなものが研究に用いられるようになってきている。

我々は、自由発話音声データベースとして、二人の話者が会話しながら会合などの予定を決定するという、スケジューリングタスクの対話データを約 800 対話の規模で収集した。本稿では、このコーパスを用いて統計的言語モデルを構築し、単語 Trigram model による音声認識実験に応用した結果を示す。

2 単語わかち書きコーパスの構築

対話のタスクは、二人の話者が会合などの任意のスケジュールを決めるものである。**図 1**に例を示す。以前から我々は文部省重点対話領域のスケジューリングタスクのコーパスを収集し用いていたが、同じタスクで、会話数規模のより大きなコーパスを構築し、利用することとした[2]。ドイツの科学技術省のプロジェクトによる Verbmobil データベースの日本語部分である。(ただし、当初のコーパスの書き起こしは、かな漢字表記およびローマ字表記の単語がわかち書きされていないものであった)

```
A10: soodesuka . <h> soredewa <P> yuugata <P> <h>
     /ee/ saNji dewa <P> ikagadeshoo .
B11: /eetto/ jaa saNji karatoyuu kotode kekkoo
     desukeredomo . sorede yoroshiidesuka .
```

図 1: 対話例 (ローマ字書き起こし部)

- 文部省重点対話領域 電通大収録分:
  対話数 13, 発話数 1019
  語彙数 750, 語数 8 千
  品詞情報付き
- Verbmobil データベース日本語部分:
  対話数 805, 発話数 10591
  語彙数 2658, 語数 33 万
  品詞情報なし

この Verbmobil の大規模コーパスを用いて、統計的に信頼できる自由発話の言語モデルの獲得を図った。

2.1 ローマ字テキストの単語セグメンテーション

語の統計量を得るためには、形態素解析処理を行なう必要がある。現在利用できる形態素解析プログラムは JUMAN をはじめとしていくつか存在している。しかし、次のような問題点が存在した。

- 新聞などの書き言葉に対しては有効であるが、自由発話文には不向きである。
- 汎用の巨大な辞書を使用しているため、それだけ誤りが多く発生する。
- 音声認識には読み方が重要なので、解析結果をローマ字に直す必要がある。解析結果につく読みが間違っている場合も多く見られたため、かな漢字をローマ字に直すプログラムを利用するが、これも読み誤りが多い。

そこで、入力をローマ字テキストとし、ローマ字のまま形態素解析し、最後に人手で修正するという方法をとった。その最大の利点は、読み方の間違いが全く発生しないということである。しかもローマ字書き起こしデータが最初からほぼフレーズ (句) ごとに区切られているという情報を有効に利用することができた。

2.2 方法

ローマ字入力形態素解析プログラムの動作を述べる。参照する情報は、次の通りである。

- (品詞 - ローマ字ラベル) 辞書
- (ローマ字ラベル - 出現頻度) 辞書
- 品詞間接続の可否 (可能ならば接続コスト値)

品詞 Bigram は品詞形態素情報をもつ重点対話コーパスから抽出した [3]。品詞 単語 対応表は、ATR Dialogue Database [4] より抽出し、語の活用形それぞれも独立した項目として登録した。品詞の種類は 27 種類とした [5]。語彙は 984 として用いた。単語出現頻度はデータベース構築作業に沿って徐々にフィードバックして用いた。品詞情報なしで 2150 語彙、単語の重みづけに用いた。

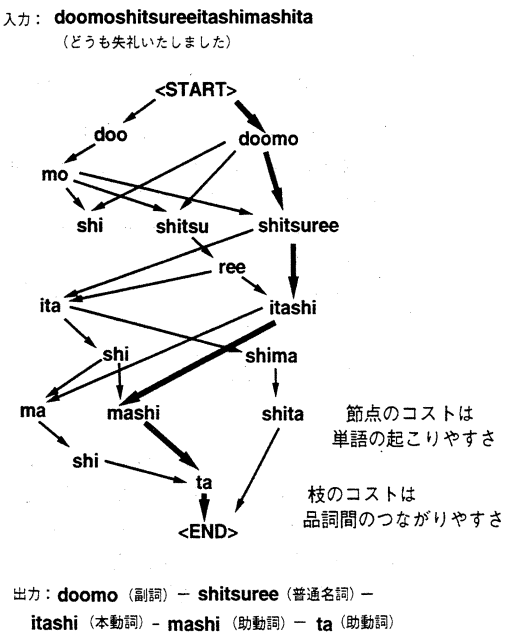

図 2: 解の探索のようす

プログラムの流れは次の通りである。

(1) 辞書を読み込む

(2) 入力ローマ字列の中に含まれうる語を辞書を参照して切り出し、それぞれを 1 つのノード (節点) として用いる (ここで、各ノードは自身の単語コードとフレーズ中での出現位置を情報として保持する)

(3) 探索空間を図 2のような有向グラフとすると、枝は品詞間コスト、ノードのコストを単語の出現頻度として重み付けし、コスト総和による評価に基づいて、最良優先探索 (Best-first search) により解の探索を行なう

(4) 結果を出力する (次の入力があれば (2) に戻る)

2.3 評価

テストセット 10 対話 189 発話に対し、セグメンテーション性能を評価した。文正解率は文全体が一致したものの割合とし 43.4%(82/189) であった。単語正解率は次のように求めた。

$$Correct = \frac{N - S - D}{N} \times 100\ [\%]$$

$$Error = 100 - \frac{N - S - D - I}{N} \times 100\ [\%]$$

N: 正解文の単語数, S: 置換誤り数, D: 脱落誤り数, I: 挿入誤り数

表 1: セグメンテーション単語正解率

参照語数		3111
仮説語数		3076
正解	89.7%	2791
誤り合計	13.0%	404
置換 substitutions	6.5%	201
脱落 deletions	3.8%	119
挿入 insertions	2.7%	84

以上より、単語正解率 89.7% を得たため、人手による修正の負担の軽減を図ることができた。この方法では解を求めるために、最良優先探索を行ない、評価値の最も高い (ここではコストが最小の) 仮説を優先的に展開することによって探索を進めるため、探索効率は良いといえる。また、未知語処理をせず、探索コストの大きすぎる場合は、探索失敗として印をつけて解析を行なわずに出力させた。この未知語検出による棄却率は 5.21% (68/1304) であった。解析の難しい箇所を検出し、積極的に人手作業に解析を委ねることで、結果的にコーパス構築の労力を抑えることができた。

3 統計的言語モデルの音声認識への応用

3.1 言語モデル作成

前節の方法によって、自由発話ローマ字テキストの単語わかち書きテキストデータを作成した。このデータを用いて単語 Bigram/Trigram の統計的言語モデルを作成した。特に、自由発話ということで、ポーズ、雑音、間投詞について、正規化を行ない (内部的には 12 種類のモデルに変換し、各モデルを単語と同様に扱って言語の統計量の取得に用いた [6])、ツールキット [7] を用いて Trigram 言語モデルファイルを作成した。この評価として、次式により、言語の複雑さとしてのテストセット 10 対話におけるパープレキシティを求めた結果を**表 2**に示す。

パープレキシティ:

$$F_p(L) = 2^{H(L)}$$

1 単語あたりの言語のエントロピー:

$$H(L) = -\sum_{w_1{}^k} \frac{1}{k} P(w_1{}^k) log_2 P(w_1{}^k)$$

表 2: Bigram/Trigram 言語モデルの統計量

学習	N-gram 登録数			perplexity	
対話数	N=1	N=2	N=3	Bigram	Trigram
190	1557	11503	29799	13.8	6.89
795	2660	24099	73137	17.3	12.5

3.2 言語モデルのカバー範囲

Verbmobil コーパスから得られたデータだけで、テストセット 10 対話についてどれだけの範囲をカバーできているのかを調べた。**図 3, 表 3**に結果を示す。学習セットは 190 および 795 対話とした。

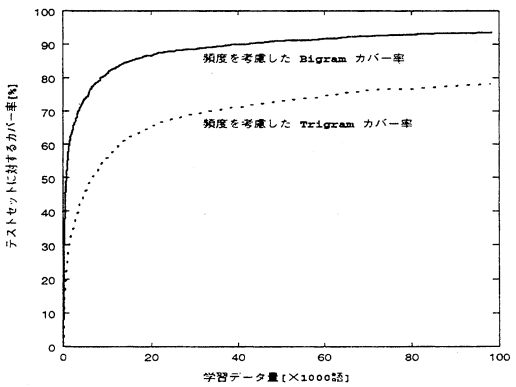

図 3: Bi/Trigram カバー率曲線 (190 対話分)

表 3: テストセットのカバー率

学習対話数	Bigram	Trigram
190	93.7%	78.3%
795	95.5%	83.9%

3.3 音声認識システムへの実装

次に示すような音響分析・音響モデル・言語モデルの条件で認識実験を行なった。認識系は、独 Karlsruhe 大学の Waibel 教授のグループが開発した JANUS システム [8] を用いた。

特徴量 16 メル係数 および 対数パワー
窓幅 5 フレーム, フレーム幅 16[ms], シフト幅 10[ms]
コンテキスト依存モデル, 不特定話者, 男女依存なし
基本モデル数 46: 音素モデル 34 種類,
　　無音および雑音モデル (12 種類)
半連続 HMM 1 モデルあたり 6 状態
スケジューリング対話 190 対話で音響モデルを学習
テストセット 10 対話
話者の男女比= 学習セット (25:53, のべ 120:260)
　　テストセット (2:4, のべ 7:13)
ARPA-NIST 形式言語モデルのサポート
言語モデル中では雑音も単語と同様に扱う
Multi-pass 探索 (1.Bigram 制約により lattice 構築,
　　2.Trigram 制約により再評価)

Verbmobil データのテストセット 10 対話 189 発話に対し、言語モデルを実装して音声認識性能を評価した。単語認識率を**表 4**に示す。ただし、評価の際には雑音モデルをフィルタリングしたものを正解とマッチングさせた。

表 4: 単語認識率 (言語モデル / 学習対話数)

LM	Bigram/190	Trigram/190	Trigram/795
正解	83.8%	84.9%	79.6%

3.4 誤り傾向の分析

認識結果を検討してみると以下のような誤りの傾向が見られた。

- 頻度の少ない人名・固有名詞
  masuda seNsee → mashi ta seNsee
- 間投詞と指示語で発音が似ている場合
  /anoo/(=#HES_ANO#) gogo kara
  → ano #PAUSE# gogo kara
- 間投詞モデルの認識に失敗し、悪影響する場合
  /ma/(=#HES#) saN jikaN → masu naN jikaN

- 一語が二語以上に分割される
  ni ji kara demo → ni ji kara de mo
- 間のび発声 (正解に含めるべき)
  soo desu nee → soo desu ne
  itashi masu → itashi maasu
  itashi masu → itashi masuu

3.5 考察

一般に語彙 V のテキストデータについて十分な N-gram を学習するには V^N の学習テキストが必要であると言われている。**3.2**で述べた言語モデルのテストセットでは、$V = 310$ であったので、Bigram モデルでは $V^2 = 9.6$ 万、Trigram モデルでは $V^3 = 3$ 百万程度の語数の学習テキストが必要であるといえる。今回用いたデータ規模では、最大でも 33 万語程度であったので、Trigram を十分に学習したとはいえない。だが、Bigram に限ってみれば、200 対話分で必要基準を満たすようである。このときカバー率は 93.7% であった。また、この 4 倍の量を学習した場合 1.8% 増加して 95.5% となった。いずれにせよ、学習テキストは多いに越したことはないが、なかなか十分量を得ることは難しい。そこで平滑化手法による Trigram の推定が必要であり、今回のモデルにも用いられている。

音声認識の結果によると、予想通りパープレキシティの小さい値をとる Trigram モデルが Bigram モデルを上回る認識性能を示した。一方、学習データを 4 倍程度増やしたモデルでは逆に認識性能は低下した。この原因は、まず、言語モデルの規模が大きくなり、パープレキシティの増加に見られるように、予測すべき語の数が多くなり、誤りの数が増大したことがあげられる。次に、出現頻度の高い語と低い語との出現確率の格差が増大し、相対的に出現頻度の低い語の認識が不利になることが考えられる。言語モデルの最適化の方法についてさらに検討を進めていく必要がある。

次に、誤認識例についてみると、品詞情報を考慮していないことが誤りの原因になっていることがいえる。クラスに基づいたモデルを設計し、部分的に品詞と同様な機能の導入を検討していく必要がある。

また、ポーズや間投詞モデル、ガーベージモデルについてみると、これらは言語モデルの統計量を取得するうえで、無視できないものとなっている。ガーベージモデルの学習にはさまざまな発音の入力を与えるため、あまり信頼性を得られないが、ポーズに関してはほぼ定常的な音源であるため、認識結果には信頼性があると考えられる。しかし正解と必ずしもマッチしていない。これは、データベースのポーズ位置が必ずしも正確ではないことが原因である。書き起こしデータベース作成時に、ポーズのラベルを人手で付与しており、基準が主観的になることに起因するものもある。

4 まとめ

本稿では、自由発話音声対話データベースとしてスケジューリング対話を用い、統計的言語モデルを作成して音声認識に適用した。この認識実験において単語正解率 80% 前後という高い認識性能を得ることができ、統計的言語モデルは自由発話の認識においても非常に有効な言語モデルであることが確認された。

今後の課題としては、自由発話対話システム実現のために有効な他の情報を活用して、言語モデルを改善していく必要がある。

謝辞

本研究においてご支援頂いた Alex Waibel 教授をはじめ、Interactive System Laboratory (独 Uni-KA/ 米 CMU) の皆様に感謝致します。

参考文献

[1] Nakamura et. al. "Japanese Speech Databases for Robust Speech Recognition", ICSLP 1996.

[2] 樽松 明, 中筋 知己. "スケジューリングタスクの自由発話音声の特徴", 電子情報通信学会全国大会, 1996.3

[3] 三木 清一, 田代 敏久, 竹沢 寿幸. "バイグラムを用いた日本語形態素解析における各種探索方法の比較検討", ATR Technical Report, 1993.8

[4] 江原, 井ノ上, ほか. "ATR 対話データベースの内容", ATR Technical Report, 1990

[5] 浦谷ほか. "音声データベースにおける日本語形態素解析マニュアル", ATR Technical Report, 1993.9

[6] Schultz,Rogina: "*Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition*", ICASSP 1995.

[7] Ries, Suhm, Rosenfeld: "*The CMU Statistical Language Modeling Toolkit*", 1995

[8] Waibel et.al., *JANUS-II Translation of Spontaneous Conversational Speech*, ICASSP 1996.