音声言語情報処理 20－5
（1998. 2. 5）

統計的言語モデルにおける高頻度形態素連鎖の辞書登録に関する一考察

小林 紀彦　中野 裕一郎　和田 陽介　小林哲則

早稲田大学 理工学部

あらまし:

大語彙連続音声認識において、学習テキスト中に頻出する形態素連鎖語を辞書登録することの効果を検証する。n-gram 言語モデルにおいて認識単位に形態素を利用することが一般的であるが、出現頻度の高い形態素連鎖語を新たな認識単位に加えることにより、言語モデルのテストセットパープレキシティは、形態素単位のみの場合と比較して、最大で約 13%の低下 (bigram) がみられ、連続音声認識実験においても約 4%の認識率の向上が得られた。また、さらにエントロピーを下げると思われる価値の高い形態素連鎖語だけを選出し辞書登録する方法も合わせて検討し、最良の方法ではテストセットパープレキシティで約 17%の低下という結果が得られた。

A Study on Word Unit Selection for LVCSR using Entropy and Frequency of Phrase Observation

Norihiko KOBAYASHI, Yuichiro NAKANO, Yosuke WADA, Tetsunori KOBAYASHI

School of Science and Engineering, Waseda University

Abstract:

We investigated the word unit selection methods for LVCSR (Large Vocabulary Continuous Speech Recognition).

In general, we use morpheme as the recognition unit for n-gram language model. However, some phrases may be more appropriate for the recognition units. We firstly evaluated the effect of selecting frequently appearing phrases as new recognition units. This method reduced 13% of test-set-perplexity and improved 4% of recognition accuracy. Then, we investigated more efficient selection methods in which only valuable phrases affecting the degradation of entropy are selected as new recognition units. The best method reduced 17 % of test-set-perplexity.

連絡先：〒169 東京都新宿区大久保 3-4-1 早稲田大学理工学部 電気電子情報工学科　小林 哲則
TEL: (03)5286-3379 FAX: (03)-9381 E-mail: koba@tk.elec.waseda.ac.jp

1 はじめに

大語彙連続音声認識において、精度の良い言語モデルの利用は不可欠である。そのため近年は n-gram に代表される統計的言語モデルが盛んに利用されている。

英語などの欧米系の言語では、分かち書きされているため n-gram の認識単位として単語を用いることが一般的であり、統計の取りやすさという点からみても合理的である。しかし日本語は単語と単語の間にスペースを入れて書く習慣がないため、単語の定義そのものが一意に決まらない面がある。そこで統計的言語モデルを用いて連続音声認識を行なう場合、認識単位（言語モデルの観点からは、予測単位）をどう設定するかが重要な問題となる。

言語モデルの予測単位に関しては、これまでに、いくつかの研究が行なわれている。政瀧らは、単語クラスを予測単位とする認識システムのために、エントロピーを基準としながら、クラスの分割、単語の連結を行ないながら、予測単位を設定する方法を提案している [5]。伊藤らは、高頻度文字列を予測単位とする方法を提案している [4]。森らは、クロスエントロピーを基準として、予測単位としての文字列、形態素列を選定する方法を、提案している [7]。しかしながら、これらの予測単位の研究では、エントロピーを減少させることに興味の中心がおかれており、大語彙連続音声認識での評価が行なわれていないなどの問題が残されている。

一般的に認識単位が短い場合

- 語彙が少ない（デコーダの負担が少なく、認識は容易）
- 言語モデルの及ぶ範囲が狭い（認識は困難）
- 脱落、挿入が生じやすい（認識は困難）

などの特徴があり、逆に認識単位が長い場合

- 語彙が多い（認識は困難）
- 言語モデルの及ぶ範囲が広い（認識は容易）
- 脱落、挿入が生じにくい（認識は容易）

などの特徴がある。このように、認識（予測）単位の変更は、エントロピーの増減以外にも、認識率と関係する要因に対し影響を持つ。

本研究では、頻度、エントロピーを考慮しながら、形態素の結合によって、認識（予測）単位を決める、幾つかの語彙の決定方式を提案するとともに、これらを、テストセットパープレキシティーと認識率の両面から評価することを試みる。

2 高頻度形態素連鎖語の辞書登録による効果

2.1 高頻度形態素連鎖語の抽出

形態素解析済み学習テキスト [9] から頻度の高い形態素連鎖語を抽出し、これを新たな認識単位に加えてみることを考える。形態素の 2 連鎖語の統計をとる場合、例えば、「初めて／の／東京／暮らし／です」という形態素列には、「初めての」、「の東京」、「東京暮らし」、「暮らしです」、と解析される。この手順を全ての学習テキストに行ない、頻度による順位をつける。同様に、3 連鎖語、4 連鎖語、5 連鎖語、6 連鎖語と統計をとり、これら全ての連鎖語から、頻度順に上位から 50、100、250、500、1000 語選択し、新たな認識単位として辞書に登録する。ただし本実験の語彙設定は、毎日新聞 94 年版出現頻度上位 5000 語に限定し、それ以外の語は未知語として扱う。以下に毎日新聞 94 年版 [10]1 年分を解析した結果を示す（表 1、表 2 参照）。

表 1:高頻度形態素連鎖語の内訳

	高頻度付属語連鎖上位				
	50	100	250	500	1000
形態素 2 連鎖	43	82	205	394	781
形態素 3 連鎖	7	17	41	87	174
形態素 4 連鎖	0	1	4	17	39
形態素 5 連鎖	0	0	0	2	5
形態素 6 連鎖	0	0	0	0	1
高頻度付属語組	16	22	44	56	84

表 1 は辞書登録された形態素連鎖語の内訳であるが、およそ 8 割が 2 連鎖語で占められていることが分かる。また、高頻度形態素連鎖語の中に占める付属語の連鎖語の割合は加える形態素連鎖語の数が増えるほど低くなり、付属語のみを考慮した連鎖語を辞書に加えることの効果 [3] の薄さが見てとれる。

表 2:高頻度形態素連鎖語の一例

形態素 2 連鎖	的・な／自民党・の
形態素 3 連鎖	を・求め・た
形態素 4 連鎖	と・発表・し・た
形態素 5 連鎖	の・で・は・ない・か
形態素 6 連鎖	こと・を・明らか・に・し・た

2.2 実験

前述の手順で得られた形態素連鎖語を基本となる 5000 語の辞書に加えて言語モデルを構築する。そしてこれらの言語モデルの能力をテストセットパープレキシティによって評価する。

言語モデルの構築及びテストセットパープレキシティの計算には CMU-Toolkit を利用した。

評価用テキストには、未知語を含まない 5000 語彙からなる文 550 文を使用した。また、認識単位の異なる n-gram の比較を行なうため、得られた一認識単位当たりのエントロピーを平均形態素連鎖数 ($=\sum_i$(認識単位 i の出現確率) * (認識単位 i を構成する形態素数)) で割ったものを補正エントロピーとして求め、これより補正パープレキシティとして算出した (図 1、2 参照)。

高頻度形態素連鎖語を認識単位に加えることにより、パープレキシティは減少することが分かる。高頻度形態素連鎖語 1000 語追加することにより、bigram で約 13 %、trigram で約 31 %の低下となった。

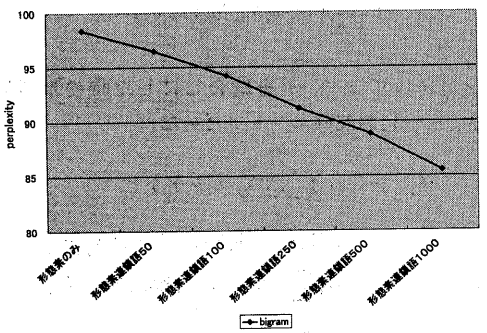

図 1: テストセットパープレキシティ (bigram)

次に連続音声認識実験を行なった。

音響モデルとして、表 3 に示す音素ごとのモデルを採用した。

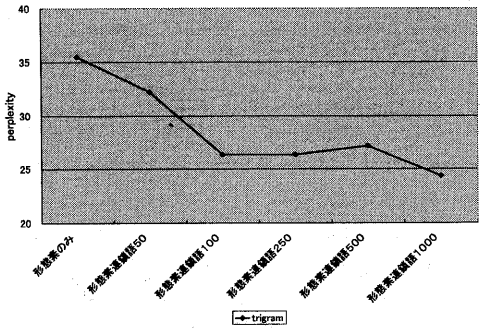

図 2: テストセットパープレキシティ (trigram)

表 3:使用した音素モデル

分類	ラベル
母音	a i u e o
長母音	aa ii uu ee oo
促音	q
撥音	N
子音	b ch d f g h j k m n p r s sh t ts w z
や行	ya yu yo yaa yuu yoo
無音	silB silE

各音素モデルは、状態数 5 (実質 3)、混合数 4、パラメータは mfcc12 次元と log パワー、およびその差分の計 26 次元とした。

初期モデル学習データとして ATR 音声データベースの単語データ (5240 単語×男性 4 話者、時間情報を含んだ音素遷移ラベルあり、16MHz にダウンサンプリングしたもの) を使用し、連結学習用データとして、音響学会連続音声データベースの ATR 音素バランス文の A セット (50 文×男性 8 話者、音素区間なしのラベルあり、16MHz サンプリング) を使用した。

探索アルゴリズムとして、onepass viterbi アルゴリズム [6] を使用した。

認識対象文は、日本音響学会新聞記事読み上げ音声コーパスより、5000 語だけで構成された 94 文 (男性話者 10 人) とした。

また認識率、認識精度を比較するため、認識結果は全て形態素単位で統一することにした。

以上の条件で行なった連続音声認識実験 (bigram) の結果が図 3 である。

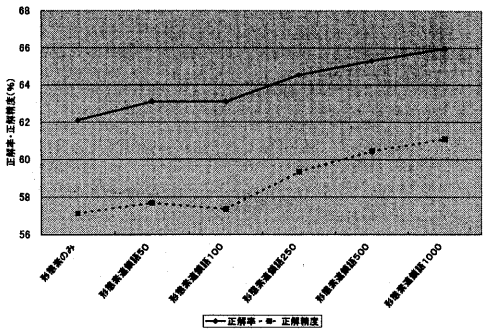

図 3: 認識結果

高頻度形態素連鎖語を辞書登録することで、辞書登録する語数に関わらず認識率の向上が得られた。特に連鎖語の頻度上位1000語を辞書登録することにより、認識率、認識精度ともに約4%上昇した。これらより、出現頻度の高い連鎖語を辞書登録することの効果が確認できた。

3 高価値形態素連鎖語の選出

3.1 価値のある形態素連鎖組の選出

高頻度形態素連鎖語の辞書登録により認識率を上げることができたが、すべての高頻度形態素連鎖語がパープレキシティを下げるため、あるいは認識率を上げるためにに働いているかは分からない。そこで、高頻度形態素連鎖語の中からさらにエントロピーを下げると思われる価値の高い形態素連鎖語だけを選びだし辞書登録することを検討する。今回は価値の高い形態素連鎖語の選出方法として次の2つの方法について試行した。

3.2 高価値形態素連鎖組選出方法 1

高価値形態素連鎖組選出方法1は、組にした方が、次に続く語のエントロピー（平均情報量）が下がる連鎖語を選び出す方法である。

直前の単語が x の時、次単語 Y_i がもつエントロピー H は、

$$H(Y_i|x) = -\sum_i P(Y_i|x)\log P(Y_i|x)$$

である。高頻度形態素連鎖語を ab として、その ab の価値（$Value$）をテキスト中の単語列 a, b, Y_i より、

$$Value = P(a,b)\{H(Y_i|b) - H(Y_i|ab)\}$$

として求め、この $Value$ を高頻度形態素連鎖語 ab の価値とする。高頻度形態素連鎖語上位50000についてこの $Value$ を求め、価値の高い順に上位50、100、250、500、1000個を選び出す。

この方法は、形態素 a と b を組にすることで、次の語が予測しやすくなるものを選択することと等価であると考えられる。

3.3 高価値形態素連鎖語選出方法 2

高価値形態素連鎖語選出方法2は、形態素 a と b を組にすることで、次の単語が予測しやすくなるものを選ぶ方法と、組にすることで ab が予測されやすくなるものを選ぶ方法の2つをあわせたものである。つまり、選出方法1の予測しやすさに、高頻度形態素連鎖語 ab の予測されやすさを加味して価値を計算したものである。

テキスト中の単語列 X_i, a, b, Y_j において、この部分のエントロピーは、

$$\begin{aligned} H_1 &= -\sum_{i,j} P(X_i,a,b,Y_j) \\ &\quad \times\{\log P(a|X_i) + \log P(b|a) + \log P(Y_j|b)\} \\ &= -\{\sum_i P(X_i,a,b)\log P(a|X_i) \\ &\quad + P(a,b)\log P(b|a) \\ &\quad + \sum_j P(a,b,Y_j)log(Y_j|b)\} \end{aligned}$$

となり、語 a と b を結合しにした時の単語列 X_i, a, b, Y_j の部分エントロピーは、

$$\begin{aligned} H_2 &= -\sum_{i,j} P(X_i,a,b,Y_j) \\ &\quad \times\{\log P(ab|X_i) + \log P(Y_j|ab)\} \\ &= -\{\sum_i P(X_i,a,b)\log P(ab|X_i) \\ &\quad + \sum_j P(a,b,Y_j)\log P(Y_j|ab)\} \end{aligned}$$

となる。

ここで、連鎖語 ab を採用した時のエントロピー変化の期待値として $H_1 - H_2$ を求めることを考える。

$$
\begin{aligned}
H_1 - H_2 \;=\; & -\{\sum_i P(X_i, a, b) \\
& \times \{\log P(X_i|a) - \log P(X_i|ab)\} \\
& + \sum_j P(a, b, Y_j) \\
& \times \{\log P(Y_i|b) - \log P(Y_j|ab)\}\}
\end{aligned}
$$

この $H_1 - H_2$ を高頻度形態素連鎖語 ab の価値とする。高頻度形態素連鎖語上位 50000 についてこの値を求め、価値の高い順に上位 50、100、250、500、1000 個を選び出す。

3.4 比較実験

それぞれの選出方法で得られた形態素連鎖語を基本となる 5000 語彙に加えて言語モデルを構築し、それぞれの能力をテストセットパープレキシティにより評価した。ここでも認識単位の異なる n-gram の比較のため、前述の補正を施した (図 4)。

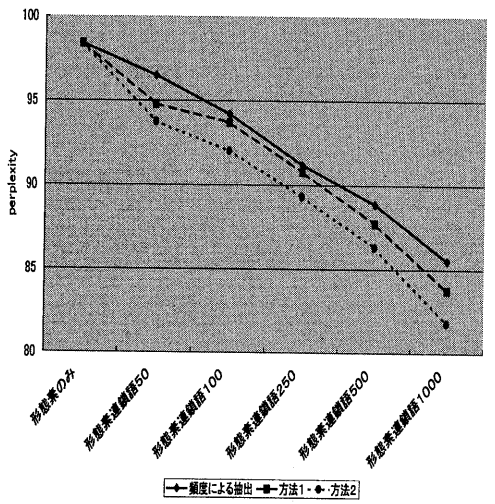

図 4: テストセットパープレキシティ (bigram)

高頻度形態素連鎖語選出の場合と比較して、パープレキシティは各方法ともに大幅な低下はみられなかったが、最も良い値が得られた方法 2 では最大約 17%のパープレキシティの減少が確認された。

次に連続音声認識実験を行なった。実験条件は前章と同じである。認識結果を図 5 に示す。

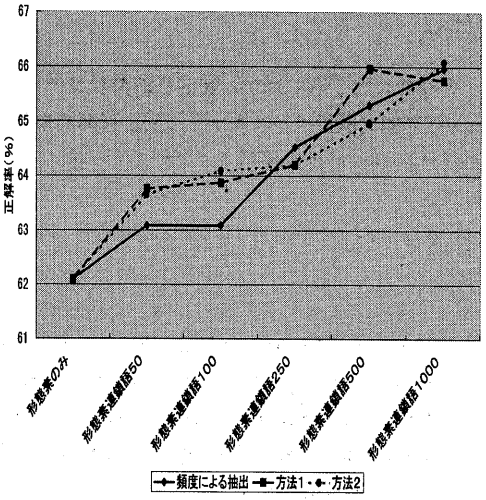

図 5: 認識結果

提案された各選出方法において、連鎖語を辞書登録することにより、認識率の向上はみられたが、高頻度の連鎖語を登録した場合と比較して、選出方法によって多少の差はあるもののほとんど変わらない結果となった。ただ、登録する連鎖語の数が比較的少ない (100 語程度) 場合には、認識率の向上がうかがえる。

4 まとめ

形態素の連鎖語を認識単位に加えることによる効果を検討した。学習テキスト中に出現する高頻度形態素連鎖語を抽出し、辞書登録することにより、形態素のみの場合と比較して、認識率の向上が確認された。

また、頻度順に連鎖語を追加するのではなく、連鎖語として採用する価値の高そうな語の選出法についても検討した。その結果、テストセットパープレキシティにおいては低下がみられたが、認識率ではいずれの方法も頻度による選出方法と大きな差は見られなかった。

今後は、音響モデルやサーチアルゴリズム等を改善しながら引続き検討を続けたい。

謝辞

本研究には、CD-ROM 版毎日新聞記事データ、RWC 形態素解析ツールキット、ASJ 新聞記事読み上げ音声コーパスを使用させていただいた。

参考文献

[1] T.Kawabata, T.Hanazawa, K.Itoh, and K.Shikano "Japanese Phoneic typewriter using HMM phone recognition and stochastic phone-sequence modeling" IEICE Trans. Information and Systems, vol.E74, no.7, pp.1783-1787, 1991.

[2] 山田智一, 松永昭一, 川端豪, 鹿野清宏 "音声認識における仮名・漢字文字連鎖に基づく統計的言語モデルの利用" 信学論（A）, vol.J77-A, no.2, pp.198-205, 1994.

[3] 小林紀彦, 中野裕一郎, 肥田木康明, 小林哲則 "統計的言語モデルにおける付属語の扱いに関する一考察" 音講論, 2-1-6, pp.59-60, Sep.1997.

[4] 伊藤彰則, 好田正紀 "かな・漢字文字列の連鎖統計による言語モデル" 信学論, D-*II*, Vol.J79-D-*II*, No.12, pp.2062-2069, Dec.1996.

[5] 政瀧浩和, 松永昭一, 匂坂芳典 "連続音声認識のための品詞・単語可変長 N-gram" 音講論, 1-P-17, pp.195-196, Mar.1996.

[6] C.H.Lee,L.R.Rabiner "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," IEEE Trans. Acoust., Speech, and Signal Proc., Vol.37,No.11,pp.1649-1658, Nov. 1989.

[7] 森信介, 山地治, 長尾眞 "予測単位の変更による n-gram モデルの改善" 信学技報, NLC97-48, SP97-81, pp.35-42, Dec.1997.

[8] 森大毅, 阿曽弘具, 牧野正三 "再現性を考慮した文字列に基づく統計的言語モデル" 信学技報, NLC97-47, SP97-80, pp.29-34, Dec.1997.

[9] "RWC テキストデータベース第 1 版 RWC テキストデータベース報告書" 技術研究組合 新情報処理開発機構, 1996.

[10] "毎日新聞 CD-ROM 1994 年版" 毎日新聞社, 1995.