音声言語情報処理 44－28
(2002. 12. 20)

かな・漢字文字列を単位とした言語モデルの検討

金野 弘明† 加藤 正治† 小坂 哲夫† 好田 正紀† 伊藤 彰則††

† 山形大学工学部
〒 992-8510 米沢市城南 4-3-16
TEL: 0238-26-3365
†† 東北大学大学院工学研究科
〒 980-8579 仙台市青葉区荒巻字青葉 05

あらまし 本研究では、形態素解析された単語を単位とせず、文字単位で N-gram 言語モデルを作成した。また、文字単位は言語制約が弱くなるため、評価基準に基づいて文字列を選択し、文字と文字列による N-gram 言語モデルを作成した。文字列の選択基準としては、高頻度の文字列を選択する方法、出現頻度を考慮した相互情報量の値の高いものを文字列と選択する方法、の 2 つを試みた。毎日新聞テキストコーパスと JNAS の音声データベースを用いて、パープレキシティおよび連続音声認識実験における文字誤り率 (CER) を評価した。選択基準としては、相互情報量の方が性能は向上した。単語単位のものと比較してみると性能の改善は見られなかったが、文字単位よりも文字列単位の方が性能が向上した。また、語彙サイズを比較すると、文字、文字列単位は単語単位のものよりも 50%減少している。

キーワード 形態素解析、言語モデル、文字列、出現頻度、相互情報量

A study on language model based on kana and kanji string

Hiroaki KINNO†, Masaharu KATOH†, Tetsuo KOSAKA†, Masaki KOHDA†, and Akinori ITO††

† Faculty of Engineering, Yamagata University
4-3-16 Jonan, Yonezawa 992-8510
TEL:0238-26-336
†† Graduate School of Engineering, Tohoku University
05 Aramaki aza Aoba, Aoba-ku, Sendai 980-8579

Abstract This paper describes a character-based n-gram language model. The proposed model is based on Kanji and Kana character instead of word or morpheme determined by morphemic analysis. To exploit stronger constraint, character strings are used in addition to single characters as basic units of the model. We examined two methods to choose character strings. One method is based on frequency in the training corpus, and the other is based on mutual information as well as the frequency. We carried out experiments to compare perplexities and character error rates (CER) between the proposed model and conventional (word or character based) n-gram model. The results showed that the mutual information based method gave the better performance. Although the proposed model was not superior to the word-based model, it was better than the character-based one. The vocabulary size of the proposed model was about 50% smaller than that of word-based model.

Key words morphemic analysis, language model, character string, frequency, mutual information

1. はじめに

高精度な連続音声認識を行うためには、精度の良い音響モデルに加えて、良い言語モデルの利用が不可欠である。近年、文音声や対話音声認識のための言語モデルとして、N-gramに代表される統計的言語モデルが利用されている。このような統計的言語モデルを作成するためには、大量のコーパスから統計をとる必要がある。英語をはじめとするヨーロッパ系の言語では、単語間にスペースが入るため、統計がとりやすい。一方、日本語のコーパスから統計をとる場合、通常の日本語の文章は英語のように単語ごとに分かち書きされていないため、英語と同じように単語単位のN-gramを構築するためには、事前に形態素解析を行う必要がある。しかし、形態素に区切る際の定義が形態素解析システムによってまちまちであるために、学習・評価などに用いるすべてのコーパスを同じ形態素解析システムで解析しなければならない。また、形態素の定義が違うことにより、N-gramの制約の強さが形態素解析システムに依存してしまう可能性がある。形態素解析を行わずに言語モデルを作成する手法として、音素、音節、文字などをN-gramの単位として用いる手法が提案されている。

本研究では、これまでの形態素解析された単語をN-gramの単位とせず、文字単位でのN-gram言語モデルを作成する[1]。しかし、文字そのものを単位として使うと単語や形態素単位よりもその長さが短い分、言語制約が弱くなる可能性がある。そこで、単語や形態素などの単位に代わり、統計的な選択基準に基づいて文字列を選択し、文字にさらに強い制約を与え、文字と文字列によるN-gramを構成する[2]。この手法を用いれば、形態素解析を行わずに、制約の強い統計的言語モデルを作ることができる。

本稿では、文献[2]の手法に基づいて作成した文字と文字列をそのまま認識の単位とした連続音声認識について検討する。

2. 言語モデル作成

2.1 文字N-gramの作成

形態素解析された単語の語彙(5000語、または、20000語)を文字と読みの組み合わせの列に分けて、それから求めた文字の語彙を用いて、文字単位のN-gram言語モデルを作成する。特殊な文字への分割については以下のような基準に基づく。

- 長音は1音素なので区切らない
- 促音は1音素で区切る
- 拗音は1音節なのでそれ以上区切らない
- 英語の頭文字を用いた略称は区切らない
- 英語のアルファベットで読むものは区切る
- 読みが区切れないものは1文字として扱う

表1 選択される文字列の例

長さ	文字列の例
2	てい した して いる する った から ない …
3	ている してい として ていた ること って い …
4	している について っている れている していた …
5	されている については なっている 明らかにし …

2.2 文字列候補の選択

文字の語彙から選択基準を用いて得られる文字列を用いて、文字列単位のN-gram言語モデルを作成する。文字列の選択方法としては2つの手法を試みた。一つは長さ2〜5の文字列候補を出現頻度の大きい順に選択する。もう一つは相互情報量に基づいて、文字列の長さ2の文字列候補を選択する。相互情報量$I(a_i, a_{i+1})$は、以下の式で与えられる。

$$I(a_i, a_{i+1}) = \{P(a_i, a_{i+1})\}^{\frac{1}{n}} \log \frac{P(a_i, a_{i+1})}{P(a_i)P(a_{i+1})} \quad (1)$$

ここで、a_iは文字、$P(a_i)$、$P(a_i, a_{i+1})$はそれぞれ、文字、文字列の出現確率、nは文字列の出現確率による重みの大きさを制御するパラメータである。nの値を変えて、それぞれについて、相互情報量の大きい順に上位500/1000/2000/.../8000個の文字列を選択する。記号(、,。,・…)は、評価の際には考慮しないので、文字列には含めない。出現頻度に基づいて選択する場合も同様である。

出現頻度で選択される文字列の例を表1に示す。また、出現頻度と相互情報量の選択基準による文字列候補の比較を表2に示す。2つの選択基準は、語彙サイズ5000語、文字列の長さは2とし、相互情報量のパラメータnは1とする。

表2より、全体をみると、出現頻度による選択基準では機能語の文字列が多く選択され、相互情報量による基準では機能語と内容語の両方の文字列が選択されている。

2.3 文字列N-gramの作成

学習テキストには毎日新聞'91〜96(ただし、94年は1〜9月分)を用いた。この中で語彙に含まれる単語のみを文字に分割し、それ以外の単語は未知語とするコーパスをまず作成した。次に、文字単位のコーパスを2.2節で選択した文字列候補を用いて左最長一致法で文字列にまとめたコーパスを作成した。作

表 2　選択基準による文字列候補の比較

順位	出現頻度	相互情報量
1	てい	てい
2	した	する
3	して	した
4	いる	して
5	する	から
6	った	った
7	から	いる
8	ない	など
9	こと	日本
10	って	こと
11	など	ない
12	され	され
13	ある	って
14	では	ある
15	れた	二十
16	二十	問題
17	いた	さん
18	とし	統領
19	ると	東京
20	れる	大統
21	にな	れた
22	日本	首相
23	いて	委員
24	には	関係
25	ため	経済
26	この	ソ連
27	さん	政府
28	なっ	によ
29	れて	ため
30	るこ	れる
31	られ	つい
32	によ	られ
33	まで	受け

成したコーパスと辞書を用いて文字列 N-gram 言語モデルを構築する。カットオフは 0、back-off には Witten-Bell discounting を用いる。

3. 文字列 N-gram のパープレキシティ

文字列単位の言語モデル、及び、それと比較のために文字単位、単語単位の作成した言語モデルのパープレキシティを文字単位に換算した。評価テキストには 5000 語、20000 語で閉じた文を用いた。文数はそれぞれ 100 文であり、毎日新聞 94 年 10 月～12 月から抽出したものである。評価データの詳細を表 3 に示す。出現頻度、相互情報量に基づく選択基準による文字列 trigram のパープレキシティをそれぞれ表 4、表 5 に示す。文字列単位のパープレキシティは、文字列候補数のうち最もパープレキシティの低いものを示した。

表 4 より、5k、20k ともに単語単位の方が文字単位よりも性能が良く、文字列単位の性能は単語単位

表 3　評価データ

評価データ	日本音響学会連続音声データベース
5k 閉じ	男性 10 名 100 文
	1799 文字 1215 単語
	1 単語=1.48 文字、1 文=約 18 文字
20k 閉じ	男性 23 名 100 文
	2875 文字 1872 単語
	1 単語=1.54 文字、1 文=約 29 文字

表 4　出現頻度に基づく文字列 trigram の PP

語彙		5k	20k
文字列単位	最大長:2	**16.51**(3k)	**17.73**(1k)
	3	16.69 (1k)	17.78 (2k)
	4	16.72 (1k)	17.87 (2k)
	5	16.71 (1k)	17.82 (2k)
文字単位		17.84	19.30
単語単位		12.34	16.04

():文字列の選択個数

表 5　相互情報量に基づく文字列 trigram の PP

語彙		5k	20k
文字列単位	n 0.5	16.42 (2k)	17.39 (2k)
	1	16.21 (3k)	17.26 (8k)
	1.5	16.25 (2k)	17.28 (7k)
	2	16.18 (2k)	**16.93**(7k)
	2.5	**16.07**(3k)	16.96 (7.5k)
	3	16.21 (3k)	17.48 (3k)
	3.5	16.22 (6k)	17.57 (5k)
文字単位		17.84	19.30
単語単位		12.34	16.04

():文字列の選択個数

と文字単位の中間であることがわかる。選択基準で比較してみると、出現頻度では文字列の最大長 2 がもっとも良い性能を示したが、表 5 より、文字列の長さが同じ条件である相互情報量の結果のほうが方がより良い性能となった。また相互情報量のパラメータ n の値を変えることで改善が見られた。

4. 文字列 N-gram による認識実験

4.1 trigram の性能評価

(1) 評価データに未知語を含まない場合

作成した文字列 trigram による音声認識実験を行った。デコーダは 2 パス方式で、第 1 パスに単語間 triphone を利用する。実験条件を表 6 に示す。評価テキストにはパープレキシティの算出と同様に語彙サイズ 5k 語、20k 語で閉じた文を用いた。評価は文字単位で行うので、単語、文字列の認識結果は文字に直してから評価する。出現頻度、相互情報量に基づく選択基準のの文字列 trigram による文字誤り率 (CER) の結果をそれぞれ表 7、表 8 に示す。表 7、8 では、文字列の最大長、相互情報量の式 (1) のパラメータ

表 6　認識実験条件

音響モデル	音素カテゴリ　35
	HMnet2000 状態+無音 3 状態
	16 混合分布
言語モデル	語彙　5k: 5267 語、2340 文字
	20k:20777 語、4879 文字
	第 1 パス:bigram
	第 2 パス:trigram
デコーダ	
ビーム幅	単語内ビーム幅:300
	単語間ビーム幅:200
	仮説数制限:5000
第 1 パス	単語　:言語重み 20
	:挿入ペナルティ -5
	文字、文字列　:言語重み 25
	:挿入ペナルティ -3
第 2 パス	言語重み　:-5〜50(5 刻み)
	挿入ペナルティ:-50〜10(5 刻み)

表 7　出現頻度に基づく文字列 trigram の文字誤り率 (%)

語彙		5k	20k
文字列単位	最大長:2	**7.21**(3k)	**8.39**(1k)
	3	8.10 (1k)	8.78 (2k)
	4	7.79 (1k)	8.97 (2k)
	5	7.85 (1k)	8.66 (2k)
文字単位		7.91	9.91
単語単位		3.89	7.61

():文字列の選択個数

表 8　相互情報量に基づく文字列 trigram の文字誤り率 (%)

語彙		5k	20k
文字列単位	n 0.5	7.08 (2k)	8.12 (2k)
	1	7.21 (3k)	**7.84**(8k)
	1.5	6.76 (2k)	8.24 (7k)
	2	6.64 (2k)	8.00 (7k)
	2.5	6.57 (3k)	7.96 (7.5k)
	3	**6.51**(3k)	8.96 (3k)
	3.5	7.02 (6k)	8.74 (5k)
文字単位		7.91	9.91
単語単位		3.89	7.61

():文字列の選択個数

n の各値に対して、文字列の選択個数に、文字の語彙を加えたものが、文字列単位の言語モデルの語彙となる。

文字列 trigram は、文字 trigram よりも高い認識性能が得られ、その効果は、語彙サイズ 20k の方が、5k の場合より大きい。選択基準の比較では相互情報量の方が文字列 trigram の認識性能が高い。出現頻度による文字列最大長の最適値は、5k、20k ともに 2 である。また、文字列の選択個数は語彙サイズ 5k では 3000 個、20k では 1000 個がよい。相互情報量の

表 9　異なる語彙サイズによる 20k 閉じ評価データの文字誤り率 (%)

評価データ		20k 閉じ	
語彙		5k	20k
OOV		単語 6% 文字 1%	- -
文字列単位	n 1	**10.18**(8k)	**7.84**(8k)
	1.5	14.84 (7k)	8.24 (7k)
	2	15.11 (8k)	8.00 (7k)
	2.5	15.15 (3k)	7.96 (7k)
	3	15.15 (3k)	8.96 (3k)
文字単位		15.93	9.91
単語単位		13.09	7.61

():文字列の選択個数

式 (1) のパラメータ n の最適値は、語彙サイズによって異なり、5k では 3、20k では 1 がよい。また、文字列の選択個数は、語彙サイズ 5k では 2000〜3000 個、20k では 7000〜8000 個がよい。5k の文字列の語彙サイズは、単語の語彙サイズとほぼ同じである。20k では、文字列の語彙サイズが単語の語彙サイズよりも少ないにもかかわらず、単語単位に近い認識精度が得られたと言える。

表 2 より、選択基準による文字列候補の結果を考慮すると、内容語の文字列が多く選択されると認識性能が向上すると考えられる。

(2) 評価データに未知語を含む場合

単語より小さい単位を用いた言語モデルは語彙外単語 (OOV) の認識に用いられることがある [6]。ここでは、語彙サイズ 5000 語の trigram 言語モデルを用いて語彙サイズ 20000 語用の評価データの認識実験を行う。実験方法は 4.1 節 (1) と同様である。表 9 に各単位の文字誤り率を示す。文字列単位の選択基準は 4.1 節 (1) で用いた相互情報量を用いて、式 (1) のパラメータは n=1/1.5/…/3 までとする。

表 9 より、単語単位、文字単位よりも文字列単位が最も良い性能が得られた。

4.2　4-gram 言語モデルによる性能評価

もう 1 つの実験として、文字、文字列 4-gram による認識実験を行った。実験方法は、4. 節の実験で作成された単語グラフを trigram でリスコアして得られる N-best 候補文 (100/200/.../1000 個) に対し、4-gram で再度リスコアし、最もスコアが高い文を認識結果とする。言語モデルの作成方法、実験条件、評価データは 4 節と同様である。4-gram ではカットオフを行う。カットオフの値は 0〜3 とする。文字 4-gram、文字列 4-gram の結果を表 10 に示す。表 10 では、カットオフと n の各組み合わせにおいて、N-

表 10　文字、文字列 4-gram による文字誤り率 (%)

語彙	カットオフ	文字単位	文字列単位				
			1	1.5	2	2.5	3
5k	0	7.01	7.15	7.02	**6.76**	7.29	7.28
	1	**6.64**	6.83	7.53	**6.76**	7.53	7.53
	2	7.15	7.34	7.72	**7.15**	8.04	7.66
	3	7.21	**7.21**	7.59	7.47	8.10	8.17
20k	0	**9.32**	8.24	8.78	8.39	**8.16**	8.78
	1	9.60	8.16	8.31	**8.00**	8.35	9.32
	2	9.71	8.86	8.94	**8.70**	**8.70**	9.40
	3	10.02	8.90	8.70	**8.66**	8.90	9.25

best 候補文の N を変えて、誤り率の最も小さいものを示している。

表 10 と 4.1 節の表 8 を比較すると、文字単位の 4-gram は trigram よりも性能が向上した。また、文字列単位の比較では、4-gram によるリスコアは表 10 と比較して性能が悪化した。文字列単位では trigram の評価範囲で充分と考えられる。カットオフは 5k で 0、20k で 1 のときに最も効果が現れた。なお 5-gram 言語モデルについて同様の実験を行ったが、文字 4-gram の性能を改善することができなかった。文字単位では 4-gram までで充分と考えられる。

5. 考　　察

N-gram 言語モデルのパープレキシティによる評価は認識実験の文字誤り率と同じような傾向が見られた。文字列 N-gram 言語モデルの性能は、単語単位と文字単位の N-gram 言語モデルの中間であったが、語彙サイズが大きい方が、単語単位の性能との差が小さくなった。選択基準では、相互情報量に基づく選択基準による文字列 N-gram 言語モデルの方が性能が向上した。また、各パラメータ、選択個数などの最適値は語彙サイズが 5000 語と 20000 語で異なった。各単位での比較、考察を以下に示す。

単語と文字

単語単位の方が性能が良かった。語彙は文字単位の方が少ないが、やはり文字の欠点である制約の弱さによって誤り率に差がでた。未知語がある場合の実験では、単語単位の方が性能がよかった。これは、未知語を含まない場合と同じ傾向である。

単語と文字列

単語単位の方が性能が良かった。しかし、語彙サイズが 5000 語より 20000 語では文字誤り率の差が少なくなった。文字列単位の語彙は単語単位より少ない語彙で単語であることを考えると、文字列単位の方が性能が良くなる可能性がある。また、文字列の選択基準として、相互情報量は出現頻度より性能が向上した。相互情報量で長さ 2 以上の文字列を選択することによって性能がさらに改善される可能性がある。

文字と文字列

5000 語、20000 語ともに文字列単位の方が性能が良かった。文字に対する制約の弱さを文字列によって改善できた。また、「てい」と「ている」などの文字列による重複、文字列にしか出現しなかった文字などを検討することで性能がさらに改善される可能性がある。

選択基準の比較は選択基準において、相互情報量の方が良い。4-gram の実験では、文字単位で若干の改善が見られた。

6. ま と め

かな・漢字文字列の N-gram を検討した。本手法を用いれば、形態素解析を行わずに、文字を単位とした言語モデルよりも強い制約をもつ言語モデルを生成できる。文字列候補の選択基準として、出現頻度に基づく方法と相互情報量に基づく方法を用いて、文字列の選択個数を変えて認識実験を行った。その結果、文字単位よりも高い精度の trigram 言語モデルを得ることができた。また、4-gram による認識実験の結果、文字単位では性能が向上したが、文字列単位では悪化した。未知語がある場合の実験では、評価セットの OOV 率は少なかったこともあるが、文字列単位の最も性能が良かった。選択された文字列の上位には「した」「てい」などの機能語の文字列が多く現れた。選択基準として相互情報量を用いた場合には「日本」「問題」などの内容語が上位に現れた。このことが、選択基準による性能の差であると考えられる。

今後は、相互情報量による選択基準において、長さ 2 以上の文字列候補を検討し、精度の向上を目指す。また、今回は単に選択基準の値の順に文字列を選択したが、[4] のような文字列選定アルゴリズムなどを用いて文字列候補の最適化を行う。

文　　献

[1] 山田, 松永, 川端, 鹿野: 「音声認識における仮名・漢字文字連鎖確率に基づく統計的言語モデルの利用」, 信学論,vol.J77-A,No.2,pp198-205,1994

[2] 伊藤, 好田: 「仮名・漢字文字列の連鎖確率による言語モデル」, 信学論,vol.J79-D-II,No.12,pp2062-2069,1996

[3] 金野, 加藤, 伊藤, 好田: 「仮名・漢字文字列を単位とした音声認識の検討」, 春音講論, Vol. Ⅰ,pp155-156,2002
[4] 和田 , 小林, 中野, 小林: 「大語彙連続音声認識における連鎖語の追加による語彙拡大の効果」, 情報処理学会誌, Vol.40 No.04, pp.1413-1220 (1999-04)
[5] 堀, 加藤, 伊藤, 好田: "状態クラスタリングによる HM-Net の構造決定法の検討", 信学論,vol.J81-D-Ⅱ,No.10,pp.2239-2248(1998).
[6] L. Galescu, "Sub-lexical language models for unlimited vocabulary speech recognition", Tech. Report of IEICE, SP2002-30, pp. 37-42 (2002-05)
[7] 森, 粕谷: 「superword モデルに基づく話者交替関連表現の抽出と予測力評価」 情報処理学会研究報告, SLP-36-9 (2001-09)
[8] 金野, 加藤, 伊藤, 好田: 「仮名・漢字文字列を単位とした言語モデルの検討」, 秋音講論, Vol. Ⅰ,pp139-140,2002
[9] S. Matsunaga and S. Sagayama, "Variable-Length Language Modeling Integrating Global Constraints", Proc. Eurospeech 97,pp. 2719–2722, Sep. 1997.
[10] S. Deligne and F. Binbot, "Language modeling by variable length sequence: Theoretical formulation and evaluation of multigram", Proc. ICASSP95, pp.169-172 (1995)
[11] 伊藤、牧野: 「音声認識のための文節構造モデルとその制約について」, 情報処理学会,vol95-SLP-6-7,pp.43-50,May 1995
[12] 伊藤、好田: 「文字列パターンの N-gram による文節モデルの検討」, 信学技報,Vol.95,No.431,pp19-24,December 1995