研究概要報告書

研究題目	連語データを用いた音声認識手法	報告書作成者	森元　逞
研究従事者	森元　逞、岩瀬　修、首藤　公昭		
研究目的	近年、大語彙（数千～数万語）を対象とした音声認識システムが開発されている。音声認識では音素のモデルとともに、言語モデルが用いられる (図１)。このため、音声認識精度は、音素モデルの良否だけでなく、言語モデルの良否にもに大きく依存する。言語モデルとして最も一般的に用いられているのは、新聞記事などの多量のテキスト（学習データ）から統計的な処理によって抽出された単語バイグラム（確率値つきの単語の２つ組）やトライグラム（確率値つき単語の３つ組）である（図２）。しかしこの統計言語モデルは、３より大きい単語数にわたるような固定的な表現などに対しては、モデル能力が充分とは言い難い。 我々はこれまで長年にわたり、日本語における慣用表現などの比較的固定的な表現(以下、「連語」と呼ぶ)を収集してきた。この連語データは、種々のテキストや各種辞書から研究者の内省をまじえて抽出したものである。これらは、付属語性連語、自立語性連語に大別できるが、合計で７万個にのぼっており、日本語における連語表現のかなりのものをカバーしていると考えている (表１)。 本研究では、この連語データを、従来の統計言語モデルと組み合わせて音声認識に用いる手法について研究する。これにより音声認識精度の向上を狙いとする。 具体的なアプローチとしては、情報処理振興事業協会（IPA）から研究用に配布されている音声認識システム Julius 用の統計言語モデルを用い、これに連語データを組み入れる方法を検討する。この統計言語モデルでは、新聞記事を学習データとして用い、基本的なユニグラムとともに、バイグラム、および逆方向トライグラムが定義されており、語彙数は 5,000 語および 20,000 語の２つのモデルが用意されている（表２）。このように、この言語モデルは規模も大きく、また汎用性も充分高いと判断できる。		

研究内容

1．基本的な考え方

連語と統計言語モデルを組み合わせる方法として種々の方式が考えられる。例えば、ある単語列 W の尤度 St を統計言語モデルから得られる尤度（生起確率）Ss と、連語としての尤度 Sc の線形結合により

$$St = \theta \times Ss + (1-\theta) \times Sc \quad （\theta は 1 以下のある定数）$$

のように求めることが考えられる。しかしこの方法では、連語の尤度をどう設定するかや、θをどう決定するかなどの問題がある。そこで、以下では連語をｎグラムの確率モデルとしてモデル化することを考える。この方法は、全体を統計言語モデルの枠組みで統一的に取り扱うことができるため、見通しが良い。ただし、ｎグラムでは緩連結の連語（例えば、「根が／とても／深い」において「根が／深い」が緩連結連語）のように、必ずしも隣接しない連語を取り扱うには多少問題がある。しかし、このような緩連結連語も隣接して現れることのほうが多いと予想されるから、第１次近似としてはそれほど問題ないものと思われる。以下、連語をｎグラムの確率モデルとして組み込むことについて、その基本的な考え方を述べる。

一般に、ある単語列 $W=w_1, w_2, \ldots w_N$ に対する生起確率は、以下のように求められる。

$$P(W)=P(w_1)\times P(w_2 \mid w_1)\times P(w_3 \mid w_1, w_2)\times \cdots \times P(w_N \mid w_1, w_2, \ldots, w_{N-1}) \quad \cdots (1)$$

ここで、バイグラムの言語モデルを用いるということは、上式の第 3 項以降の項をバイグラムで近似することになり、

$$P_{bi}(W) \quad P(w_1)\times P(w_2 \mid w_1)\times P(w_3 \mid w_2)\times \cdots \times P(w_N \mid w_{N-1}) \quad \cdots (2)$$

として求めることとなる。

しかし、W が連語であるとすれば、この近似は不十分であり、

$$P(w_n \mid w_1, w_2, \ldots, w_{n-1}) >> P(w_n \mid w_{n-1}) \quad (n \geqq 3) \quad \cdots (3)$$

であろうと予想される。ここで（3）式の左辺はｎグラムである。従って、連語 W に関してこのような n グラムを求め、それを言語モデルに組み込むことができれば、性能を改善できることになる。なお明らかなように、トライグラムとの比較においても同様なことが言える。

以下では、このような連語の n グラムを求め、バイグラムの言語モデルに組み込む方法について述べる。ここで、求めたｎグラムをそのままのｎグラムとして音声認識システムに組み込むには、システムの探索アルゴリズムに手をいれる必要があり、これは大変である。そこで、以下では連語 W を 1 語と考え、n グラムそのままでなくバイグラムとして組み込む方法を検討する。なお上述のように、同様な方法によってこの n グラムをトライグラム言語モデルに組み込むこともできる。ただし、今回の検討においては、Julius のトライグラムが逆方向のトライグラムであるため順方向に変換する作業が必要になるなどの理由から、まずバイグラムに組み込むこととした。

研究内容

2．連語の妥当性の検証

先に述べたように、連語は日本語の慣用的な表現を研究者の内省を元に抽出したものである。このため、具体的な連語の n グラムを求める前に、まずこれらの連語が統計的にも妥当である（生成確率が充分高い）かどうかを検証することとした。

まず、我々の連語のうち 2 単語連語を取り出し、それに対応する Julius 内のバイグラムの確率を求める。これと、Julius 内に定義されている連語以外の他のバイグラム確率とを比較する。具体的には、2 単語連語に対応するバイグラム確率の対数尤度をとり、それを 1 連語あたりで平均したもの、すなわち

$$-\frac{1}{N_2}\sum_{2\text{単語連語}}\log P(w_2 \mid w_1) \qquad (N_2 \text{は 2 単語連語の数}) \qquad \cdots (4)$$

を求め、その分布を求めた。また、連語以外の 2 単語列についても同様な分布を求め、両者の分布を比較した。

語彙数 5,000 語、カットオフが 0 の場合の結果を図３にしめす。語彙数やカットオフ条件が異なっても傾向はほぼ同じである。横軸は対数尤度にマイナスを付けたものであるから、尤度値が小さい方が確率が高いことになる。縦軸は相対頻度である。これから、2 単語連語を構成する単語相互の連接確率は、それ以外のものよりかなり大きい値を持つものが多いことが分る。ただし、2 単語連語は全部で約 11,000 個あるが、上記で対象とした連語は Julius 内のバイグラムにエントリが存在した約 1,800 個のみである。このようにエントリに存在しなかった連語が多い理由は、恐らく連語データが極めて広範な分野から抽出されたためであろう。しかし、Julius のエントリに存在する連語については上記のような傾向が見られることから、少なくとも Julius が対象としている分野に関しては、連語データは充分有効であると言える。表３に、連接確率値が大きなものの上位 10 個をしめしている。

このように、少なくとも 2 単語連語については、Julius の統計言語モデル（バイグラム）においても充分モデル化されていることが分る。今回はまだ行っていないが、３単語連語とトライグラムを比較した場合も同様な結果が得られるものと思われる。（4 単語以上の連語については、Julius の統計言語モデルに対応するものが定義されていないため、このような検証を行うことは不可能である。）

以上のことから、我々の連語は、統計的にも充分妥当なものが抽出されていると考えられる。

研究内容

3． バイグラム言語モデルへの連語の組み込み

N 単語からなる連語 $W= w_1, w_2, ...,w_N$ においおて W を 1 語と考え、そのバイグラム、すなわち $P(W|a)$および $P(b|W)$を求め（a, b はそれぞれ W に前節および後接する単語)、これを通常のバイグラムと同様に既存のバイグラム言語モデルに組み入れる。

まず前接のバイグラム確率 $P(W|a)$について述べる。W を $\{w_1, w_2, ...,w_n\}$に展開しそれを変形すると、

$$P(W|a) = P(w_1, w_2, ..., w_N|a) = P(w_1|a)P(w_2, ..., w_N| w_1) =\{ P(w_1|a) / P(w_1)\}\times P(w_1, w_2,...,w_N) \quad \cdots (5)$$

となる。この式において、$P(w_1|a)$ および $P(w_1)$は元の言語モデルで与えられているものをそのまま用いる。

ここで $P(w_1, w_2,...,w_N)$をどのようにして求めるかが問題となる。我々が収集・抽出した連語は、先にも述べたように研究者の内省により求めたものであるから、統計的な値は全く得られていない。そこで、ここでは「連語内での単語の n 連接（すなわち n グラム）は、n が大きくなれば 1 に近づく」と仮定することにする。すなわち、

$$P(w_1, w_2,...,w_N) = P(w_1)\times P(w_2 | w_1)\times\cdots\times P(w_N| w_1 , w_2, ..., w_{N-1}) \quad \cdots (6)$$

と変形できるが、第 3 項以降は

$$P(w_k |w_1, ..., w_{k-1})=P(w_k)+(1- P(w_k))(1-e^{-\lambda(k-1)}) \quad (2\leqq k\leqq N) \quad \cdots (7)$$

が成り立つものと仮定する。なお、この式における λ の求め方については、次節で述べる。

このように（7）を（6）に代入し、さらにそれを（5）に代入して $P(W|a)$を求めるわけであるが、これを既存のバイグラム言語モデルに組み込む際には、この確率値の分を他のバイグラムの確率値から差し引かなければならない。このため、v_i を a に後接する連語以外の単語とすると、元々の言語モデルに定義されている v_i に関するバイグラム確率値 $P(v_i|a)$を

$$\hat{P}(v_i|a) = (1-P(W|a))P(v_i|a) \quad \cdots (8)$$

のように補正する。

最後に、連語 W を 1 語とみなした場合の後接する単語との確率 $P(b|W)$であるが、これは単純に W の最後の単語 w_N のバイグラムを用いて

$$P(b|W)=P(b|w_N) \quad \cdots (9)$$

として求める。

研究内容

４．連語の生起確率の計算

まずλを推定する方法を述べる。（７）式の仮定をより一般的に拡張し、

$$P(w_k \mid w_j, \ldots, w_{k-1}) = P(w_k) + (1 - P(w_k))(1 - e^{-\lambda(k-j)}) \quad \cdots (10)$$

とする。すなわち、連語内のｎグラム（上式では k-j+1 グラム）の確率は、ｎが大きくなれば漸次１に近づくものとする。そうすれば、

$$P(w_k \mid w_{k-1}) = P(w_k) + (1 - P(w_k))(1 - e^{-\lambda})$$

が成り立つから、これから

$$l = -\log \frac{1 - P(w_k \mid w_{k-1})}{1 - P(w_k)} \quad \cdots (11)$$

としてλを求めることができる。なお、これはｋ番目の単語 w_k のユニグラム、バイグラムを使って求めたものであるから、これを λ_k と書くことにすると、連語 W に関するλは各 λ_k (k=2, 3, .., N-1)の平均値として求めることにする。

$$\lambda = (\lambda_2 + \cdots + \lambda_N) / (N-1)$$

$$= -\frac{1}{N-1} \sum_{k=2}^{N} \log \frac{1 - P(w_k \mid w_{k-1})}{1 - P(w_k)} \quad \cdots (12)$$

以上の方法に基づき、各連語データに対するλを求めた。11,131 個の連語のうち、バイグラムのデータが存在したものが 1,370 個であった（存在したデータが少なかった理由は２．で述べた）。このうち、1,103 個の連語についてはλがプラスとなった。なお、残りの 267 個はλがマイナスとなったため、これらは連語として採用するのは不適当であると思われる。

λがプラスとなったものについて、このλを用いてその連語の生起確率を求めた。表４に、この推定した生起確率値が高かった上位 10 個を示している。また、各連語について推定した連接確率値と、バイグラムのみによって計算した生起確率を求めて比較してみた。表５に、前者（推定生起確率）が後者（バイグラムによる生起確率）に比べ大きくなったもののうち上位 10 個をしめしている。これらの結果から、頻度の高い連語には高い生起確率値が付与できているようである。従って、上記のような計算法は充分妥当性があると思われる。

今期の研究は以上のところまでである。今後さらに以下の研究を継続して実施する。

（１）求めた連語の確率モデルを実際のバイグラム言語モデルに組み込む。

（２）テストセット・パープレキシティを求めたり、具体的な音声認識実験などを行って、作成した言語モデルの性能を評価する。

（３）同様な手法により、トライグラム言語モデルへの組み込みを行う。

研究概要報告書

研究のポイント	（1）情報処理振興事業協会（ＩＰＡ）から研究用にリリースされてる音声認識用統計言語モデルを用いて、収集した連語の確率的な妥当性を検証する。 （2）連語を１つの単語とみなし、これを既存の統計言語モデルに組み込む方法を明らかにする。 （3）実際に上記ＩＰＡの統計言語モデルに組み込み、その性能評価を行う。
研究結果	（1）収集した連語のうち、2 単語連語について、その確率的な妥当性を検証した。 （2）ある連語Wを１単語とみなし、これをバイグラム言語モデルないしトライグラム言語モデルに組み込む手法を明らかにした。 （3）バイグラム言語モデルに組み込むにあたり、連語Wの生起確率を求める必要がある。この生起確率を推定する方法を考案した。 (4) 上記（3）の手法により、実際に連語の生起確率を計算した。その結果を見てみると、一般的に使用頻度が高いと思われる連語は確率が高くなっており、上記手法は妥当であると思われる。
今後の課題	現段階では、研究のポイントで挙げた項目のうち、（2）までが完了している。今後は（3）を行うことになる。具体的には、以下のような項目を進める。 （1）連語を実際のバイグラム言語モデルに組み込む。 （2）テストセット・パープレキシティを求めたり、実際の音声認識実験などを行って、性能評価を行う。 （3）同様な手法により、トライグラム言語モデルへの組み込みを行う。 さらに、今回の研究では連語を構成する単語はお互いに隣接していることが条件となっているが、この条件を緩和する方法についても今後の課題として検討したい。

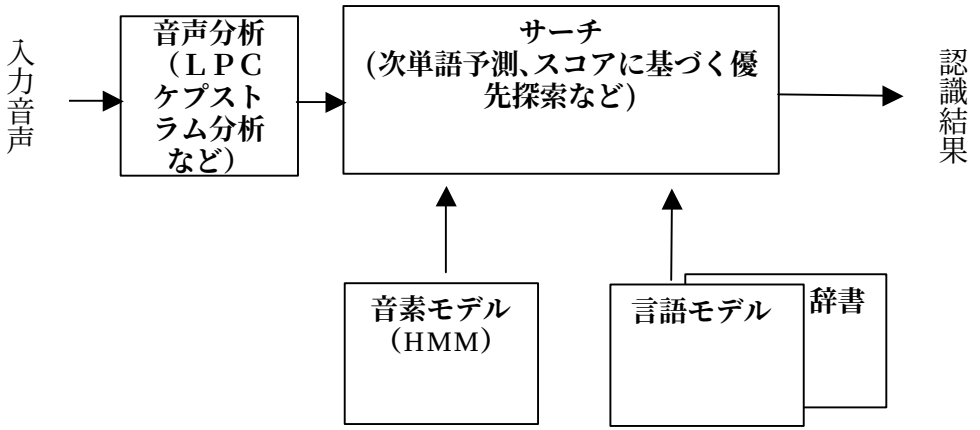

図１　音声認識システムの一般的な構成

先行単語	後続単語	確率
あいまい	さ	-0.9477
あいまい	すぎる	-2.6198
あいまい	だ	-1.3045
あいまい	だっ	-1.4535
あいまい	で	-1.0214
あいまい	です	-2.6198
あいまい	と	-1.8665
：	：	：

図２　統計言語モデルの例（バイグラム）

確率の欄は、実際の確率値の対数

表１　連語データ

分類	小分類		例	個数
付属語性連語	関係表現		に／ついて、の／ように、などと／いった	960
	助述表現		なければ／ならない、かも／しれない、たほうが／よい	1,400
自立語性連語	強連結	名詞性	赤の／他人、目の／毒	55,000
		サ変名詞性	貰い／泣き、ラッパ／飲み	
		動詞性	相い／異なる、汗水／垂らす	
		形容詞性	途方も／ない	
		副詞性	案の／定、いつに／なく	
		その他	四文字熟語、～する、など	
	緩連結	名詞性	悪業の／報い、環境の／汚染	15,000
		サ変名詞性	額に／汗、お手数を／おかけ	
		動詞性	心が／沈む、気を／吐く	
		形容詞性	態度が／でかい、気が／重い	
		形容動詞性	懐が／暖か、愛情が／細やか	
		副詞性	目を／輝かせて、先を／争って	
		その他	間尺に／合わぬ、心胆／寒からしむ	

表２　Ｊｕｌｉｕｓの言語モデル

語彙数	カットオフ	バイグラム数	トライグラム数
５，０００	０	１，２９７，０００	１１,７６６,０００
	１	７８７，０００	４，８５１，０００
２０，０００	１	１，２３８，０００	４，７３４，０００
	４	６５８，０００	１，５９３，０００

（注）トライグラムは逆方向トライグラムが定義されている。

（注）表１、表２のいずれの表においても個数は概数である。

（構成図、ブロック図、フローチャート、データ表、グラフ、写真など研究内容の補足説明にご使用下さい）

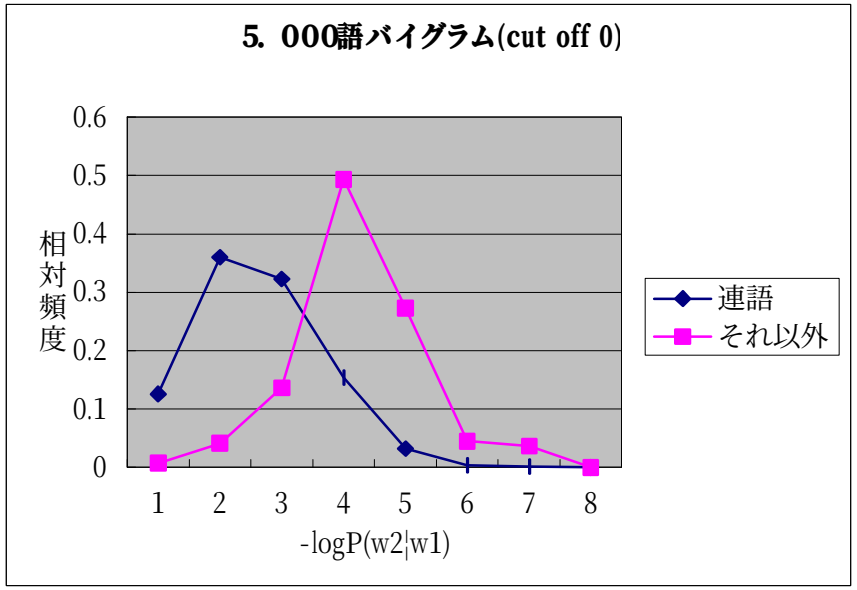

図３　2 単語連語の対数尤度と、連語以外のバイグラムの対数尤度の比較（ただし、対数尤度のマイナスをとっている）

表３　2 単語連語で確率（対数尤度）の大きなもの（上位 10 例）
（語彙数 5,000 語）

順位	対数尤度	2 単語連語
1	-0.0111	いけ-ない
2	-0.0191	従業-員
3	-0.0476	明らか-に
4	-0.0549	一緒-に
5	-0.5720	多角-的
6	-0.0588	特捜-部
7	-0.0654	評-議会
8	-0.0700	消極-的
9	-0.0957	対照-的
10	-0.1070	同時-に

（注：フローチャート図，ブロック図，構成図，写真，データ表，グラフ等 研究内容の補足説明にご使用下さい。）

表４　生起確率が大きな連語（上位 10 個）

順位	生起確率（対数尤度）	連語
1	-3.262206	何+と+なれ+ば
2	-3.498861	それ+に+も+かかわら+ず
3	-3.599385	か+と+いって
4	-3.603696	一+呼吸+置い+て
5	-3.644741	日+を+追っ+て
6	-3.791955	そう+で+なけれ+ば
7	-4.030524	これ+と+言って
8	-4.103192	こう+で+なけれ+ば
9	-4.192297	持っ+て+生まれ+た
10	-4.287625	可能+性+が+高い

表５　バイグラムによる生起確率に比べ、推定による生起確率が大きな連語（上位 10 個）

順位	①推定した生起確率（対数尤度）	②バイグラムによる生起確率（対数尤度）	差（①－②）	連語
1	-10.5326	-20.3417	9.8090	仏-の-光-より-金-の-光
2	-9.5874	-19.3516	9.7642	苦労-を-苦労-とも-思わ-ない
3	-12.0355	-21.6704	9.6348	目-で-見-て-口-で-言え
4	-10.2019	-19.7705	9.5685	海-とも-山-とも-つか-ず
5	-5.6375	-15.1977	9.5602	面-と-向かっ-て-の
6	-17.1392	-7.6028	9.5363	後-の-世-まで-伝える
7	-5.7031	-15.0940	9.3908	犠牲-者-が-出る
8	-7.9759	-17.2969	9.3209	一-人-相撲-を-取る
9	-4.6985	-13.8546	9.1560	万-に-一つ-の-可能-性
１０	-6.2041	-15.3498	9.1456	首-を-長く-し-て-待つ