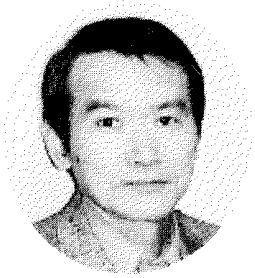

連語を用いた音声認識手法

福岡大学　工学部
教授　工学博士　森　元　　逞

1．日本語における連語

日本語にはかなり固定的な表現が数多く存在する。このような固定的な表現を「連語」と呼ぶが、我々の研究グループでは、多年にわたり新聞や雑誌や教科書など様々なテキストからこのような連語をひろい集め、計算機上に蓄積してきた。収集した連語は付属語的なものと、自立語的なものに分けることが出来る。表１に概要をしめす。

この連語データは種々の用途に利用することが出来る。例えば、これまでワープロのカナ漢字変換にこの連語データを用いる方法を提案し、高い変換性能を達成できることを確認している。

表１　連語データ

分類	例	個数(概数)
付属語性連語	に＋ついて、の＋よう＋に、など＋と＋いっ＋た	2,300
	なけれ＋ば＋なら＋ない、かも＋しれ＋ない、た＋ほう＋が＋よい	
自立語性連語	赤＋の＋他人、目＋の＋毒	70,000
	貰い＋泣き、ラッパ＋飲み	
	相い＋異なる、汗水＋垂らす	
	途方＋も＋ない	
	案＋の＋定、いつ＋に＋なく	
	悪業＋の＋報い、環境＋の＋汚染	
	額＋に＋汗、お手数＋を＋おかけ	
	心＋が＋沈む、気＋を＋吐く	
	態度＋が＋でかい、気＋が＋重い	
	懐＋が＋暖か、愛情＋が＋細やか	
	目＋を＋輝かせ＋て、先＋を＋争っ＋て	
	間尺＋に＋合わ＋ぬ	

2．音声認識に使用される統計言語モデル

ここでは、研究テーマの背景を説明したい。なお、説明の一部で若干の数式を用いることをご容赦いただきたい。図１に一般的な音声認識システムの構成図をしめす。音声は母音や子音の組み合わせでできているが、これらを音素と呼ぶ。音声認識システム内にはこれらの音素のモデルを用意しておく必要がある。このモデルは、実際に人が話した音声を多量に収集し、そこから各母音や子音を切り出して音声分析を行なうことにより作成する。具体的な音素モデルとして、一般的にはHMM(隠れマルコフモデル)と呼ばれる確率的な音素モデルが用いられてい

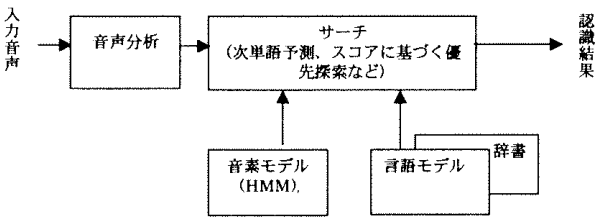

図１　音声認識システムの一般的な構成

る。次に、単語とその読みを記述した辞書を用意する必要がある。「読み」と入力された音声を比較することによって、どの単語が話されたかを認識するわけである。

本研究テーマに直接関連するのは「言語モデル」である。これは、日本語ではどういう単語同士が接続可能かを定義したものである。このような言語モデルを用いずに音声認識を行なうと およそ日本語とは思われない結果しか得られない。

言語モデルとして種々のモデルが提案されているが、現在で最も一般的に用いられているのは「Ｎグラム」と呼ばれている統計的な言語モデルである。Ｎグラムとは、ある単語の生起確率を、その直前に現れたＮ－１個の単語の並び $w_{i-N+1},\cdots,w_{i-1}$ を条件として定義したもの、すなわち条件付き確率 $P(w_i \mid w_{i-N+1},\cdots,w_{i-1})$ を定義したものである。概念的に図示すると、図２のようになる。ある単語列 $w_{i-N+1},\cdots,w_{i-1}$ に対し、その次に出てくる単語 $w_{i1},w_{i2},\cdots$ の確率が定義されている。なお、特にＮ＝１、２、３のものは、おのおのユニグラム(１-gram)、バイグラム(２-gram)、トライグラム(３-gram)と呼ばれている。このような統計言語モデルは、原理的には、多量のテキスト(例えば数年分の新聞記事など)を用意し、その中に個々の単語の並びが何回出現したかをカウントすることによって求めるこ

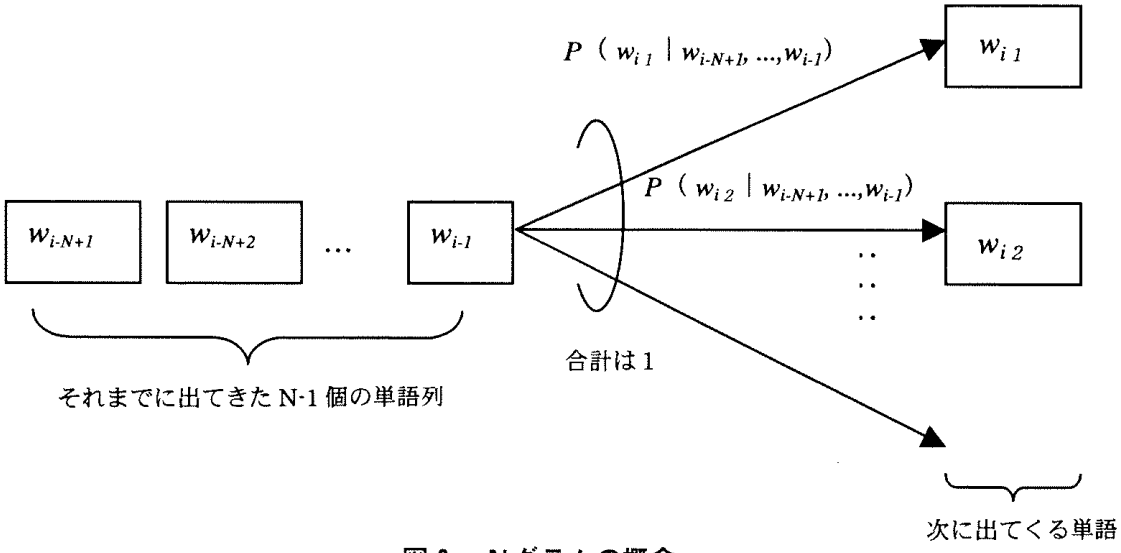

図２　Ｎグラムの概念

とができる。

以上のような単語の生起確率において、Nの長さを長くすればするほど精度の高い確率が得られ、その結果正確な認識ができる。しかし実際は4-gram以上のものを作ることはほとんど不可能である。なぜなら、系列の長さが長いほど、その組み合わせの数が爆発的に増加するからである。例えば、語彙数2千の場合、すべての単語の組み合わせを考えると、バイグラムの数は2千の2乗で4百万、トライグラムであれば80億となり、現在のコンピュータが何とか取り扱える範囲である。しかし、4-gram数になるとその数は16兆にもなり、通常のコンピュータではとても取り扱えなくなる。さらにそのような膨大な数になると、各々の生起確率を学習するには無限大に近い多量のテキストが必要になってしまう。このような理由から、現在の音声認識では、バイグラムないしはトライグラムが用いられている。

さてここまでの話でもうお分かりと思われるが、我々が設定した研究テーマは「音声認識の統計言語モデルに連語データを組み込もう」ということである。上で述べたように、すべての単語組み合わせに関する4-gramや5-gramを組み込むことはその数の多さから不可能に近い。しかし連語についてであれば数はそれほど多くない。そしてもし連語をうまく組み込むことができれば、より精度の高い言語モデルを構築することができ、最終的には音声認識精度を向上させることができることになる。

3．統計言語モデルへの連語の組み込み

(1)連語の生起確率

さて、連語を統計言語モデルに組み込もうとすると、根本的な問題がある。つまり、本文の最初の方にも述べたように、連語は研究者が種々のテキストから1つ1つ拾い出したものであるから、確率値またはそれに類する情報が全く得られていないことである。統計言語モデルに組み込むには、連語に何らかの方法で確率値に近いものを付与しなければならない。それはどうやって求めればよいのだろうか？　1つの方法は統計言語モデルを学習した元のテキストをもってきて、各連語の生起回数を再度カウントし、確率を求める方法である。しかし、統計言語モデルは与えられているが、何らかの理由でその元データが手に入れられない場合、これは実行不可能である。

そこで我々は手に入った統計言語モデルから連語の確率値を近似的に求める方法を採用することにした。連語はかなり固定的な表現であるから、連語を構成する単語が前から順番にでてきたとすると、後ろの単語になるほどその生起確率は1に近くなるであろうと予想される。すなわち、連語 W を構成する単語を $w_1,w_2,w_3,\cdots,w_N$ とした場合、$P(w_1)$、$P(w_2 \mid w_1)$、$P(w_3 \mid w_1, w_2)$、…は図3のようになると思われる。我々は、このような仮定が成り立つとして、連語の確率を求めることとした。

(2)生起確率の推定

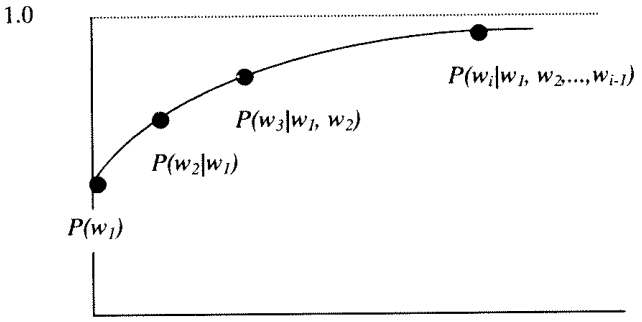

図3　連語の確率に対する仮定

ベースとなる統計言語モデルとして、IPA(情報処理振興事業協会）からフリーソフトとして配布されている汎用日本語音声認識システムJuliusの言語モデルを用いた。この言語モデルは数年分の新聞記事を元に作成されており、ユニグラム、バイグラムおよび逆向きのトライグラムが定義されている。

連語の確率が図3のようであるとすれば、最低ユニグラム、バイグラムの2つの確率値があればそれから図3の曲線が推定でき、その曲線を外挿することにより連語の確率を推定することができる。もちろん、トライグラムの確率値を使ったほうがより正確な確率値を求めることができるが、まず第1ステップとして、ユニグラム、バイグラムのデータのみを用いることとした。

このようにして求めた連語の確率のうち、確率値が大きい連語のいくつかを表2に示す。直感的に、出現頻度の高い連語に高い確率が付与されており、上記の推定方法はうまく働いているようである。

表2　推定された連語の確率（上位10個）

確立(対数)	連語
−3.262206	何＋と＋なれ＋ば
−3.498861	それ＋に＋も＋かわら＋ず
−3.599385	か＋と＋いって
−3.603696	一＋呼吸＋置い＋て
−3.644741	日＋を＋追っ＋て
−3.791955	そうでなければ
−4.030524	これ＋と＋言って
−4.103192	こう＋で＋なけれ＋ば
−4.192297	持っ＋て＋生まれ＋た
−4.287625	可能＋性＋が＋高い

(3)元の統計言語モデルへの組み込みと性能評価

さて、連語の生起確率が推定できたわけであるが、次はこれを元の言語モデルへどう組み込むかが問題である。まずは、N単語からなる連語 $W=w_1,w_2,\cdots,w_N$の各部分系列を$P(w_i \mid w_1, w_2,,\cdots, w_{i-1})$のようにNグラムとして登録する方法が考えられる。しかしそうすると、音声認識エンジン自体の認識アルゴリズムを大幅に変更しなければならない。そこで、連語全体を1つのかたまりとし、$P(W \mid w_{prev})$、$P(w_{after} \mid W)$のようなバイグラムとして登録することとした。（ただし、w_{prev}およびw_{after}は、おのおのWに前接および後接する単語とする。） なお、この連語に対するバイグラムをどう計算するかについては、話が細かくなるので省略したい。

さて、そのようにして連語を組み込んだ言語モデルを作成し、その性能評価を行なった。言語モデルの評価方法として広く用いられているのは評価文集合（テストセット）に対する複雑

度（パープレキシティ）を用いる方法である。複雑度というのは、大まかに言えば、ある単語の次に生起する平均単語数を、単語の生起確率を考慮して求めたものであり、複雑度の値が小さいほど言語モデルとしては性能が良いことになる。新聞記事から抽出したいくつかの評価文集合に対して複雑度を求めてみた。その結果、ある評価文集合では複雑度が７％程度改善されたが、別の評価文集合では、若干（１％程度）ではあるが悪化してしまった。この悪化の原因としては色々考えられるが、多分以下のような事項が理由であろうと考えられる。

(1)今回は連語の生起確率をユニグラム、バイグラムのみから推定したが、その結果、推定した確率にかなりの誤差が混入した。

(2)連語を１語として取り扱っている。このため評価文中に、連語とほとんど同じ意味であるが、一部だけ異っているものが現われても連語とみなされない(例えば、「影＋も＋形＋も＋見え＋ない」という連語に対し、「影＋も＋形＋も＋見当たら＋ない」という表現が現われた）場合があった。

(1)に関しては、今後はトライグラムなどを使用することにより、さらに推定精度を改善して行くことが必要になる。また(2)の問題に関しては、連語に表現の揺れなどを取り組む方法について検討する必要がある。

4．おわりに

連語を音声認識システムの言語モデルを組み込む方法について研究を進めてきた。この結果、一部の評価文に対してはある程度の改善効果が確認できた。しかし、全体としてはまだ充分な性能が得られていない。今後ともさらに研究を進めていく予定である。

最後になったが、本研究をご支援いただいた㈶サウンド技術振興財団に心から感謝の意を表したい。