音声言語情報処理 15－5
（1997. 2. 7）

単語を認識単位とした日本語ディクテーションシステム

西村雅史　伊東伸泰　山崎一孝　荻野紫穂
日本アイ・ビー・エム（株）東京基礎研究所
e-mail: nisimura@trl.ibm.co.jp

単語を認識の単位とした日本語ディクテーションシステムについて述べる。欧米では既にいくつかのシステムが実用化されているが、日本語においては、単語の概念が明確でないため、N-gram等の言語モデルの導入が容易ではなく、研究が遅れた。最近、形態素を認識単位とした研究が開始されているが、離散単語発声も可能なシステムとなると、ほとんど検討されていない。本論文では、日本人の考える、単語の切り出し方を統計的にモデル化する方法を提案する。この方法によって抽出された単語単位を認識および発声の単位として、不特定話者日本語ディクテーションシステムを構築し、その性能を評価した。約4万語の辞書を用意したところ、新聞記事などのタスクに対しては98%程度のカバレージが得られた。また、男女計20名の各200文ずつの離散単語発声による読み上げ文に対して認識実験を行ったところ、96%以上の認識率が得られた。一方、この際の読み上げ速度は、150～270文字/分に相当し、離散発声にもかかわらず専門オペレータによるキーボード入力よりも1.5倍以上高速であった。

A word-based Japanese dictation system

Masafumi NISHIMURA, Nobuyasu ITOH, Kazutaka YAMASAKI, Shiho OGINO
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.

In this paper, we discuss a word-based Japanese dictation system. In Japanese, word boundaries are ambiguous, because there are no spaces in text. This makes Japanese speech recognition difficult where a large vocabulary is used. Recently such grammatical units as morphemes are used as recognition units, but this doesn't accord with human intuition, and consequently these units cannot be used for isolated utterances. We propose a statistical method for segmenting a text into words on the basis of human utterance units. By using the estimated word units, we developed a speaker-independent Japanese dictation system. Evaluation of the performance of this system showed that the recognition rate was more than 96%, and that the speed of isolated-word speech input was more than 50% faster than that of typing.

はじめに

近年欧米では、統計的言語モデルを用いたディクテーションシステムが、離散単語発声ではあるが、徐々に実用化され始めているが[1]、日本語では、単音節発声の認識とかな漢字変換プログラムを組みあわせたシステムが検討されて以来、音声による日本語文の入力システムは実用化されていない。欧米ではビシネス上、口述筆記への期待が大きかったこともあるが、技術的には、欧米の言語が単語という概念が明確で、単語単位の文章発声が可能であったこと、また、単語のN-gramに代表される言語モデルが経験的に非常に有効に動作したことによる。さらに、音素環境依存型混合連続分布HMMに代表される音響モデルの高精度化が、認識精度を実用可能な領域に押し上げたと言える。

一方、日本語の単語の定義はあいまいで、欧米語のように単語単位の取り扱いには適していない。このため、発声単位と、言語モデルの単位(注:今後、認識単位と呼ぶことにする)を別のものとし、この差異は連続音声認識技術によって解決されるものとして、文字や、解析的最小単位である形態素を認識単位としたシステムが一般的に検討されている[2,3]。しかし、文字や形態素を単位とした場合、1単位当たりの平均モーラ長が短くなる分、たとえ言語モデルの単位当たりパープレキシティが小さくても、音響的には識別が難しいし、長い認識単位に比べると、限定された範囲のN-gram制約では探索範囲の制限効率が悪い。このため、N-gramの拡張方法として、可変長N-gram[4]や、自立語と付属語のN-gram[5]等も検討されている。

このように、認識の単位としては、高頻度でかつ、なるべく長い単位を辞書に持った方が効率がよいことが多い。たとえば、1文字のN-gramよりは、高頻度2文字連鎖のN-gramを用いた方が、一般的には不要な文字連鎖をサーチする回数が減って効率がよい。このような観点から、相互情報量に基づいて単位を合成したり、高頻度文字連鎖を単位とする研究もなされている。しかし、単純に長単位からなる辞書を作ったのでは、種々のタスクに対処しようとすると語彙数が爆発的に増大する恐れがあるし、音声認識の立場からは、発声単位が辞書中の単位より短い場合には単純には対処できないという問題が生じる。

本論文では、"日本人の考える日本語の単語"を認識の単位とする方法を提案する。この方法の長所としては、ユーザーの考える単位を使用しているため、これを発声単位と一致させることが出来、離散単語発声による文章入力が可能になること、また、一般的には、解析的な最小単位である形態素よりも長い単位を定義できることがあげられる。特に、実用化の観点からは、離散発声による入力を可能にすることには、精度、計算量の両面で大きな意味がある。

一方で、必要語彙数とパープレキシティの増大、ならびに、離散発声時の入力速度の低下が懸念されるが、これについては実験により有効性を示す。

単語の定義方法

計算機による日本語処理を考える場合、処理単位としては形態素を用いるのが一般的である。形態素は、エキスパートが言語現象に関する知識を利用して解析的に決定したものであるから、統計的言語モデルの構築という観点からは、あながち悪い単位とは言えない。ただ、一般のユーザーがその単位を知るはずもなく、音声認識のための発声の単位としては、明らかに適していない。

我々は、欧米語同様に、日本語においても発声および認識の単位として、単語を用いることを検討した。日本語の単語単位はあいまいなものではあるが、単語とよべるような潜在意識的な単位が存在していることも事実である[6,7]。たとえば、「見る」は、解析的には、動詞「見」とその活用「る」とに分けて考えられるが、1つの単語と意識することが自然にできる。

大量のコーパスをこの"日本人の考える単語"単位に分割処理した結果、語彙数、カバレージ、言語モデルのパープレキシティ等の量が、欧米語に対して得られている量と大差なければ、日本語においても、単語単位の認識が可能であると予想される。しかし、数年分の新聞記事といった大量のコーパスを人手で単語単位に分割するというのは、実際には不可能である。

そこで、先ず、限られた量のテキストを用いて、人間の単語単位分割に関する振る舞いを調査し、その結果を既存の形態素解析プログラム[8]の出力と照らしあわせることで、この分割操作を統計的にモデル化した。

単語単位推定用モデルの学習[9]

形態素解析により、テキストは例えば次のように分割される。

研究・所・に・行・っ・た

単語単位の発声を指示した場合、このうち、動詞の活用語尾「っ」や、助動詞「た」はほぼ確実に「行った」と発声されるが、一方、「研究・所」は、結合して発声される場合も、分割して発声される場合もある。そこで、学習データとして、人間の行った単語分割の位置と、形態素解析結果の両方を用い、これらの対応関係をダイナミックプログラミングの手法を用いて推定する。一般的には形態素の方が、人間の行う分割よりも短い単位として定義されているので、基本的には形態素間の分割確率を推定することになる。ただ、例外的に逆の場合も存在するので、高頻度語については形態素中の文字境界についても取り扱う。

先ず、形態素解析により得られる、形態素の品詞（*Pos*, 119種類）、形態素種別（*Kow*, 81種類）および、表記（*Spelling*）の組を $M = (Pos, Kow, Spelling)$ とする。i 番めと*i+1*番めの形態素の間に分割または結合操作 $O = \{Split, Join\}$ が入る確率（分割確率と呼ぶ）は、この2つの形態素のパラメータにのみ依存すると考え、$Pr(O|M_i \rightarrow M_{i+1})$ と定義した。

この確率を、先に求めた対応関係に基づき、すべてのパラメータの組み合わせに対して推定しておく。当然、頻度の低い単語については十分な統計量が得られないので、その場合には、一定の頻度が得られるまで、次のように、順次条件を緩和して推定した確率値で代用することにした。

$$
\begin{aligned}
&Pr(O|M_i \rightarrow M_{i+1}) \\
&= Pr(O|Pos_i, Kow_i \rightarrow Pos_{i+1}, Kow_{i+1}, Spelling_{i+1}) \\
&= Pr(O|Pos_i, Kow_i \rightarrow Pos_{i+1}, Kow_{i+1}) \\
&= Pr(O|Kow_i \rightarrow Kow_{i+1}) \\
&= Pr(O|Kow_{i+1})
\end{aligned}
$$

なお、ここで示した条件の適用順序は、あらかじめ木構造で表現し、各節がその条件下での分割確率を持つ構成にした。

言語モデルの学習

先ず、個々のテキストを従来の形態素解析プログラムに通し、タグ付きの形態素解析結果を得る。テキスト中の個々の形態素の2つ組みに対し、先に求めた分割確率木をルートから順次下位の節に向けて探索し、最も多くの条件が一致したところで、その分割確率を得る。なお、すべての分割可能性を考慮すると、長さ*L*の形態素列からなるテキストからは 2^{L-1} 種類の単語分割テキスト*T*がその出現確率 $Pr(T)$ とともに求まることになる。

$$Pr(T) = \prod_{i=1}^{L-1} Pr(O|M_i \rightarrow M_{i+1})$$

全学習テキストに対して、Nつ組の単語出現頻度を、この単語分割テキストの出現確率を重みとしてカウントすれば、単語N-gramモデルの最尤推定値を求めることができる。

なお、音響モデル訓練用あるいは評価用の単語単位読み上げテキストを生成する際には、個々の形態素間でその分割確率に従うサイコロを振り、この分割を行うか否かを決定した。この処理をすべてキストに対して繰り返し行ったコーパスを用いてN-gram言語モデルを推定することもできる。

数字の読み上げ単位

数字については、出現頻度が高く、かつ、単語の種類数が膨大になる。また、上記のような単語切り出しのシミュレーションでは対処できない。常に連続発声を想定するなら、安易に認識単位を定義することもできるが、離散発声を考慮する必要があるので、ここでは、次のような認識単位をあらかじめ定義した。つまり、数字を漢数字で表記したときの、十、千、万、億を位と定義し、数字はそれに先行する数字と、それに続く位で、1つの単語を構成すると考えた。離散発声を行う場合には、位の後には必ずポーズを置く。つまり、

三億・四千・五百・二十・万・三千・五

と分割するようにした。このようにすれば、0～

999,999,999,999までの数値を表現するために必要となる単語数は、57単語に過ぎないし、境界位置に促音が含まれないので、離散発声が可能である。N-gramモデルの学習時には、位別に単語クラスを定義し、クラス内の単語には共通の確率を与えている。

なお、助数詞（～本、～円、等）を含めた場合には促音を境界に含む場合がある。特に出現頻度の高い単位で、促音を含む場合についてはすべての組み合わせを網羅するように語彙を選定した。

認識システムの構成

ディクテーションシステムの構成について述べる。このシステムは、認識精度と、処理速度の観点から、基本的には離散単語発声を認識対象としているが、上記の単語単位の推定方法を用いても、必ずしも、必要とされる単語単位がすべて用意できているとは限らない。たとえば、助詞の前にポーズを置き忘れて文節発声となったり、複合語を分割し忘れるような発声ミスも多々起こりうる。つまり、複数の単語が連続発声されたからと言って、直ちに誤認識となるのではなく、認識率の低下や、応答の遅れといった影響が出たとしても、このうち何割かは正しく認識されることで再入力の手間が省けることが望ましい。そこで、認識のアルゴリズムとしては、離散単語のものではなく、スタックデコーダー[10]と呼ばれる連続音声認識のアルゴリズムを採用した。ただし、各単語の音素表記辞書の末尾には無音を表すモデルを追加しているし、単語間に渡る音素環境も一切考慮していない。システムの構成を図1に示す。

一般的な連続音声認識システムと大きく異なる点は、音響処理部でランクベースラベリング[11]を行っている点と、ファーストマッチ[12]によって、単語の予備選択を行う点である。

音響モデル

音素HMMの各状態遷移に対し、前もって用意した、音素クラスに関する361種類のYes/No Questionを順次適用して、音素環境による判別木を自動生成した。あるノードnを、質問qに基づいて分割するか否かの判断

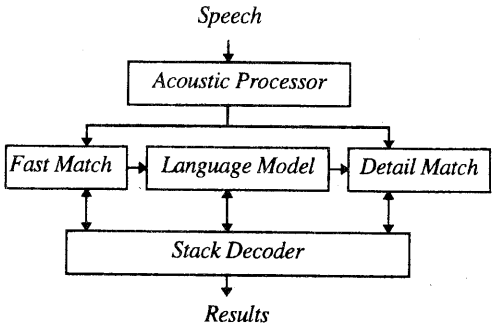

図1 認識システムの構成

には、次の尤度比の尺度を用いた[13]。ここで、yは対象となる特徴量系列を表す。

$$m(n,q) = log\left(\left(Pr_{Ml}(y_l)\,Pr_{Mr}(y_r) / Pr_{Mn}(y_n)\right)\right)$$

なお、本報告では、最大前後5音素までを考慮して推定した、合計1,712個の音素環境依存モデルを使用している。

ランクベースラベリングは、このようにして推定した音素環境依存連続HMMの尤度計算時のダイナミックレンジを押さえるために導入したものである。先ず、入力フレーム毎に、上記音素判別木上のすべてのリーフ上のモデルに対する尤度を推定しておく。これをそのままサーチ時の尤度として使用するのではなく、一旦、尤度順に順位(ランク)を付け、前もって推定してあるランクと尤度の対応表を参照することでデコーディング時に使用する尤度を決定している。

デコーディング

ファーストマッチでは、音素HMM内の状態遷移を無視し、各状態の出力確率の最大値からなる1状態モデルを用いることと、木構造の音素表記辞書を用意しておくことで、高速な単語予備選択処理を実現している。このモデルは、そのスコアFuが、詳細マッチのスコアDuに対し、常に$Fu>Du$という関係を保持している。この条件が満たされる場合、ファーストマッチの1位候補に対する詳細マッチのスコアをファーストマッチの候補リストの打ち切りの閾値として設定すれば、この候補リストには、全単語に対して詳細マッチを行った場合の1位候補が必ず含まれることが保証される。

デコーダーは、スタックアルゴリズムを用いて、入力単語列を推定する。先ず、現時点でのスタックを参照して次に展開すべきパス(単語列)を決め、ファーストマッチを実行する。この候補リストに対して3-gram言語モデルのスコアを併用し、さらに候補を絞る。この結果得られた候補リストに対して詳細マッチングを行う。詳細マッチングの結果に対し再度言語モデルを参照して結果をソートし、順次スタックに積む。この操作を繰り返し、入力単語列を推定する。

データベース

単語単位推定用学習データ

単語単位推定用のモデルを構築するため、新聞記事や、日本語用例集、および電子会議室の発言などから収集した約34,500文について、計17名の被験者が、単語発声単位に分割する作業を行った。

なお、被験者には、

- 不自然にならない限り、できるだけ細かく分割すること。
- 発声できる単位に区切ること。

という指示を与えた。

この結果に基づき、2,829個の節(分割確率)を持つ確率木を作成した。ただし、節として採用するか否かの閾値は当該節の出現回数が50以上のものとした。

言語モデル用学習データ

表1に示す約200万文のテキストデータを3-gram言語モデルの学習用データとした。多様な文体を取り入れるため、産経新聞のデータ量を日経新聞に比べ多くするとともに、口語体の文章として、パソコン通信People上のテキストも利用した。

このうち5%はHeld-out補間法による言語モデルの平滑化に使用している。また、主に3-gramの出現確率の閾値を制御して、サイズの異なる2種類の言語モデル(L1、L2)を作成し、その影響を調べることにした。各モデルの含有セル数を表2に示す。

一方、この学習用データのカバレージが95%以上になるように、単語を選定し、これに数字表現などの用語を適宜加えて39,295語からなる認識対象語彙(40K)を作成した。

表1 言語モデル用学習データの内訳

データ名	総文数
日経新聞	415K
産経新聞	1,291K
EDR[14]	169K
電子会議室の発言	173K

表2 各言語モデルのセル数

モデル名	2-gram	3-gram	総セル数
L1	2,772K	2,196K	5,008K
L2	2,784K	12,443K	15,267K

音響モデル用学習データ

20代～60代の男性110名、女性105名が、それぞれ内容の異なる120文（計25,800文、355,454単語）を離散単語発声した音声データを用いて不特定話者用の音響モデルを作成した。各読み上げテキストは、新聞記事、手紙、小説などから音素バランスを考慮しつつ多様な文体を選択した。なお、これらの読み上げテキストは前もって、単語単位に分割する処理を行った後、人手で修正作業を行い、ポーズ位置をすべて指定してある。また、全体の約10%の文章については、句読点や、記号も読み上げている。音声採取は比較的静かな室内で、接話型のマイク（Shure SM-10A）を用いて行った。

実験結果と考察

カバレージとパープレキシティ

言語モデルの性能を、人間が実際に単語単位に分割したテキスト計約4,000文を用いて調査した。内容は、言語モデルの学習にも用いた、新聞２種類、電子会議室の発言に加え、学習領域外の、ビジネストークの書き起こし文と、小説である。テスト文はすべてオープンデータである。また、単語分割を行った被験者も学

習時とは異なる。

パープレキシティ推定時の3-gram言語モデルはL1およびL2を使用し、1-gramあるいは2-gram言語モデルを使用した場合についても同様に調査した。結果を表3に示す。なお、クローズの場合よりもカバレージが向上しているのは、クローズデータには、形態素解析誤りや、単語の分割誤りにより生じた未知語が相当数含まれていたためである。

表3 40K語彙のカバレージとパープレキシティ
(3-gram言語モデル：上段はL1、下段はL2を使用)

データ名	Coverage	Perplexity		
		1-gram	2-gram	3-gram
日経新聞	98.0	2737.2	159.6	99.2
				91.0
産経新聞	95.5	2074.6	227.4	180.8
				170.8
電子会議室	95.2	2545.6	321.6	272.4
				258.6
ビジネス	97.5	1762.8	158.6	127.9
トーク				125.7
小説	94.6	1902.4	250.8	210.4
				202.6

日経新聞に対しては、98%のカバレージが得られているが、姓名を対象から除くと98.8%、さらに、被験者の数字単位の切り間違い、記号も除くと、カバレージは99.1%に達する。ビジネストークに対しては学習領域外のタスクであるにも関わらず97.5%のカバレージが得られたが、その他の３つのタスクに対しては95%程度に過ぎなかった。日経新聞以外では、姓名以外の未知語として、擬態語、擬音語などのカタカナ語、あるいは英語のカタカナ表記の揺れが目につく。また、丁寧語、「～れる」「～られる」の部分の揺らぎ、「お」「御」などの接頭語の接続、「～って」等の言い回し、略名などが数多く未知語となっている。その他にも、複合動詞、複合名詞、３文字語などが、分割されていないために未知語となっていた。

一方、パープレキシティの観点から見ても、日経新聞がこの５つのタスクの中で最も小さい値であったのに対し、同じ新聞でも、産経新聞は２倍近い値となっている。訓練用のデータ量は、むしろ圧倒的に産経新聞が多く、N-gramのデータ不足とは考えにくい。これは、それぞれの新聞がカバーしている領域の広さや、文体の多様性の影響が出ているのであろう。特に、日経新聞ではかなり長い単位の複合語が多用されていることが、N-gram言語モデルにとっては有利に働いていると思われる。その傾向は、2-gramと比較した場合の3-gramのパープレキシティの減少率が特に大きい点にも見て取れる。

一方、電子会議室の発言は学習領域内のタスクにも関わらず、最も難しいタスクとなっている。これは、単語のカバレージの低さにも現れているが、これらが、口語体の文章であり、学習時に十分な精度の形態素解析結果が得られず、未知語の出現頻度が高かったこと、また、根本的に、学習データの絶対量が不足していたことにも起因している。さらに、文中に記号や、表記上の誤りなども多く、それらの前処理にも課題が残る。

逆に、ビジネストークのパープレキシティが小さいのは、括弧などの記号がなく、また、文体が比較的平易であった上に、内容が新聞に多く見られるような経済上の用語を多く含んでいたためと思われる。一方、小説は括弧などの記号が多く、内容的にも、学習データにはほとんど含まれていない会話文が主体であった。

次に、言語モデルのサイズの影響についてであるが、総セル数がL1に比して3倍以上あるにも関わらず、L2の効果はいずれのタスクに対しても小さい。記憶量や計算速度に制限がある場合、3つ組みより2つ組みのデータを充実させた方が、パープレキシティの観点からは効率がよいと言える。

不特定話者単語認識率

学習時とは異なる20代～60代の男女各10名が、単語単位に発声した、各200文（入力単語総数=52,612）を用いて認識実験を行った。なお、ポーズ位置は学習データと同様に読み上げテキスト上に指定してあり、記号や句読点もすべて読み上げている。

内容は、新聞記事150文(日経、産経、EDRから各50文)と、電子会議室の発言50文であり、すべてオープンデータである。なお、40K語彙によるカバレージが100%になるように文章を選んでいる。タスク別の誤認識率を表4に示す。

表4　認識実験結果 (言語モデルはL1を使用)

タスク	入力単語総数	Perplexity	置換 (%)	挿入 (%)	脱落 (%)	誤認識率 (%)
新聞記事	41,952	98.4	2.6	0.2	0.4	3.2
電子会議	10,660	318.1	5.3	0.2	0.4	5.9
全体	52,612	124.8	3.1	0.2	0.4	3.7

発声速度=0.9～1.6単語/秒 (平均1.2)

このテスト文に対し、言語モデルL1を使用した場合のパープレキシティは表3で示した結果に近いものである。なお、この実験では同音同義語は、すべて正解として数えている。

また、言語モデル未使用時の誤認識率は、61.7% (ただし、同音異義語を除く音響上の誤りは33.9%) である。さらに1-gramだけを適用すれば11.6%、さらに2-gramも考慮すれば、4.6%まで誤認識率は低下しており、L1では3-gramまでを考慮したことによる改善は小さい。なお、L2を使用すれば誤認識率は2.8%まで改善されたが、この場合のパープレキシティは58.6であった。

一方、このデータを使い、文毎のパープレキシティと、3-gram言語モデルを用いたことによる誤認識の削減量(音響モデルのみによる誤認識率との差)の関係を調べた。その結果が図2である。このように音響モデルだけによる認識結果との差分を用いることで、エントロピー ($\log_2 Perplexity$) との相関が強くなったが、それでも相関係数 r=0.6に過ぎない。なお、このような正規化を施す前の誤認識率とエントロピーとの相関は r=0.45であった。

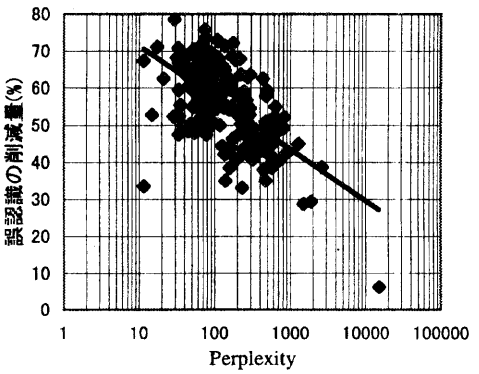

図2　言語モデルに起因した誤認識の削減量とPerplexityの関係

入力速度の比較

離散単語発声による入力速度を、単音節発声、連続発声、ならびに、かな漢字変換を使ったキーボード入力と比較した。結果を表5に示す。なお、単音節入力に関しては、かな漢字変換後の文字数に換算してある。

表5　入力速度の比較

入力方法	入力速度(文字/分)
キーボード入力 (初心者) [15]	40
キーボード入力 (専門オペレータ) [15]	100～150
単音節発声	40～75
離散単語発声	150～270
連続発声	280～510

このように、読み上げ文ではあるが、離散単語発声でも、専門オペレータのキーボード入力に比べ、おおむね1.5倍以上、初心者と比べれば4～7倍程度高速に入力できており、キーボードとかな漢字変換プログラムに代わる、有効な日本語入力手段となりうることがわかる。

おわりに

日本人が単語と考える単位の切り出し方を統計的にモデル化することで、単語を認識単位とするディクテーションシステムを構築した。この単位はそのまま発声単位として利用できるため、日本語においても離散単語発声によるディクテーションが実現できる。

このように、単語の切り出し方の揺らぎを反映した単位を使用したにも関わらず、4万語程度の語彙を用意すれば、幅広いタスクに対し、95%以上の単語カバレージが得られること、また、3-gram言語モデルのパープレキシティは、小さいものでは100程度、大きなものでも、300以下に押さえられていることがわかった。

また、20名の話者の離散単語発声を用いて、この不特定話者用のシステムを評価したところ、読み上げ文に対してではあるが、パープレキシティ98.4の新聞タスクに対しては96.8%、パープレキシティ318.1の口語

体のタスクに対しても94.1%の精度が得られた。

さらに、入力速度を比較したところ、離散発声にも関わらず、オペレータのキーボード入力よりもおおむね1.5倍以上速い入力が可能であることがわかった。

今後はこの認識単位を連続発声に対しても適用し、その認識性能を調べるとともに、他の、形態素などの認識単位と比較してその優位性を検証したい。また、言語モデルの観点からは、主に口語体の文章に関する学習データの充実を図っていきたい。

謝辞

データ使用を許可してくださった、産経新聞社、日本経済新聞社、ならびに(株)ピープルワールドカンパニーに感謝します。

参考文献

[1]加藤雅浩, "音声がパソコンの入力手段に、キーボード、マウスを補完", 日経エレクトロニクス, no.627, pp.153-162, (1995-01)

[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・漢字文字連鎖確率に基づく統計的言語モデルの利用", 信学論, J77-A, 2, pp.198-205, (1994-02)

[3]松岡達雄, 大附克年, 森岳至, 古井貞熙, 白井克彦, "新聞記事データベースを用いた大語い連続音声認識", 信学論, J79-D-Ⅱ, 12, pp.2125-2131, (1996-12)

[4]政瀧浩和, 松永昭一, 匂坂芳典, "連続音声認識のための可変長連鎖統計言語モデル", 信学技報, SP95-73, pp.1-6, (1995-11)

[5]磯谷亮輔, 嵯峨山茂樹, "付属語のN-gram、自立語のN-gramを用いた音声認識", 音響講演論文集, 2-Q-6, pp.95-96, (1993-03)

[6]時枝誠記, "日本語文法口語編", 岩波全書, (1950)

[7]西村雅史, 大嶋良明, 野崎広志, "日本語Dictation Systemのための統計的言語モデルに関する一考察", 情処講演論文集, 3R-7, pp.2-117-118, (1995-09)

[8]丸山宏, 荻野紫穂, "正規文法に基づく日本語形態素解析", 情処学論, 35, 7, pp.1293-1299, (1994)

[9]伊東伸泰, 西村雅史, 荻野紫穂, 山崎一孝, "人の発声単位を考慮した日本語言語モデルの検討-日本語における単語とは", 自然言語処理, 116-9, pp.57-64 (1996.11)

[10]L.R.Bahl, F.Jelinek, R.L Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE. Trans. Pattern Analysis and Machine Intelligence, PAMI-5, 2, pp.179-190 (1983)

[11]L.R.Bahl, P.V. de Souza, P.S.Gopalakrishnan, D.Nahamoo, M.A.Picheny, "Robust methods for using context-dependent features and models in a continuous speech recognizer", Proc. ICASSP'94, pp.533-536 (1994)

[12]L.R. Bahl, S.V.De Gennaro, P.S.Gopalakrishnan, R.L.Mercer, "A fast approximate acoustic match for large vocabulary speech recognition", IEEE Trans. Speech and Audio Processing, SAP-1, 1, pp.59-67 (1993)

[13]L.R.Bahl, P.V.de Souza, P.S. Goplalkrishnan, M.A. Picheny, "Context dependent vector quantization for continuous speech recognition", Proc. ICASSP'93, pp.632-635(1993)

[14]EDR電子化辞書仕様説明書, (株)日本電子化辞書研究所

[15]長尾真, "日本語情報処理", 電子通信学会