自然言語処理 28－5
（1981. 11. 20）

動的計画法による漢字仮名混り文の単位切りと仮名ふり

藤崎哲之助（日本アイ・ビー・エム(株) 東京サイエンティフィック・センター）

1. はじめに

日本語の言語処理が大学や会社の研究室の範囲を出て、様々な分野で実用的にその技術が用いられるようになってきたが、そこで依然として重要な基本的問題として日本文の入力の問題と漢字仮名変換の問題がある。

この漢字仮名変換とは通常の日本語の表記法に従いベタ書きされた漢字仮名混り文を入力として、わかち書きされた仮名文を出力するものであり、ベタ書き文を適切な単位でわかち書く単位切りの問題と、それぞれの漢字の読みを文脈より決定する仮名ふりの問題の2つの問題を含んでいる。

この漢字仮名変換の重要性については特に言及するまでもないが、文書よりの索引（KWIC / KWOC）の自動作成、文献検索のための自動キーワード切り出しなどを実現するために不可欠の技術であり、また文書の点字化、将来の文書の自動朗読などにも結びつく技術である。

日本アイ・ビー・エム（株）東京サイエンティフィック・センターでも、従来より将来のオフィス・オートメーション実現のための中核として日本語文書処理システム「ことだま」の研究開発を行ってきているが、[1,2,3,4] そこでも、既存の漢字コード入力された文書に対して「ことだま」の索引自動作成、キーワード文献検索などの文書処理サービスを利用可能とするために漢字仮名変換の研究を行ってきている。

過去においても、この漢字仮名変換には、いくつかの試みが報告されている。[5,6,7] なお、その目的により、特に単位切りにおいては、文節の単位の単位切りを目指すもの、自立語／付属語の単位切りを目指すもの、さらに自立語が複合語である時に語基の単位切りまで目指すものなどの差異があり、それらを一律に比較することはできない。

文節の単位までの単位切りにおいては、字種の変り目（ひらがなからひらがな以外への変り目）を文節の始まりとするという単純な規則が有効であることがよく知られており、この単純な規則だけにより84%の分割精度を得たとの報告が行われている。[8] 単位切りの単位が短くなるとさらに難しくなる。漢語が多く2文字単位で構成されることや、複合語分解の分布を利用し規則と若干の接辞用漢字リストで分割を試みた例もあるが、[9] 最終的には語基の辞書を整備し、分割の際に数万の辞書を参照しなければならない現状である。

一方、仮名ふりにおいても、単位切りと同様に、両隣の文字の字種が漢字以外なら訓読み、その他の場合は音読みとするというような単純な音訓規則を利用すると、ある程度の漢字は読みが一通りしかないことなども手伝って85%程度の精度を得ることはできる。[7]

このように、文節切り、仮名ふり共に簡単な規則で85%程度は容易に達成できるのだが、単位切りの単位をより細くする、精度を高める、などのためにこれらの基本規則の例外を辞書もしくは例外コードとして蓄積することになる。自然言語は奥の深いものであるのでこのようなアプローチは最終的には避けられないと考えられるが、基本的規則の与える精度が低ければ低い程、辞書などの形で蓄積しなければならな

い情報の量は膨大となり、かつ分野への依存度が高まるのは避けられない。

筆者は、従来より文字認識・音声認識などのパターン認識の分野で用いられてきたマルコフ過程の推定アルゴリズムであるViterbiアルゴリズム[10,11]がこの種の日本語の問題に有効であると考え、特に漢字仮名変換の問題への適用を試みた。もちろん、本稿で紹介する方式が漢字仮名変換を完全に解決する訳ではなく、むしろ、この方式が従来までの単純な基本規則に加えて働き、例外辞書や、例外コードの処理を大巾に軽減する可能性があると考える。実験もまだ小規模で、結果もこのアルゴリズムで得られる精度の極値に至っている訳ではないが、この方式の有効性を示す結果が出たこと、また、この方式が漢字仮名変換以外の日本語関連の問題にも有効な手法になり得ると考え報告する。

2. Viterbi アルゴリズム

Viterbiアルゴリズムとは、離散時間の有限マルコフ過程における推定アルゴリズムであり、図1に示すようなマルコフ過程と、記憶のない雑音を持つ通信路の環境において通信路の出口における観測値$\mathbf{Z}$（以後$\mathbf{Z}$と略す）からマルコフ過程からの出力$\mathbf{X}$（以後$\mathbf{X}$と略す）を推定するものである。問題を単純化するためマルコフ過程の内部状態S_kと出力X_kを同一視すれば$\{S_k \triangleq (X_{k+1}, X_k)\}$通信路に記憶がないので次がいえる。

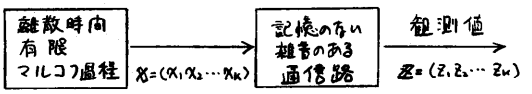
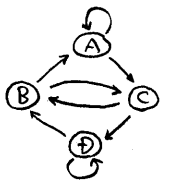


図1　一般モデル

$$P(\mathbf{Z}|\mathbf{X}) = P(\mathbf{Z}|\mathbf{S}) = \prod_k P(Z_k|X_k) \quad (i)$$

さて、$\mathbf{Z}$から$\mathbf{X}$を推定するために、内部状態と時間とをそれぞれ縦軸、横軸とする有向グラフ（trellisと呼ばれる）を作ることが有効である。（図2）



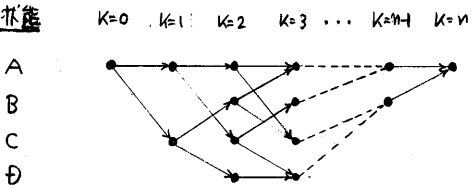


図2　trellisの例

このtrellisは図1のマルコフ過程の内部状態の遷移を時間と共に表記したものであるが、重要なことは観測値$\mathbf{Z}$を発生させる可能性のある$\mathbf{X}$がすべてtrellis上の別ルートとして表わされ、逆に$\mathbf{Z}$を発生させる可能性のない$\mathbf{X}$に対応するルートは含まないように作られていることである。従って、$\mathbf{Z}$から$\mathbf{X}$を推定することは、このtrellis上のルートを1つ選ぶことである。

$\mathbf{Z}$を観測して、それを発生した$\mathbf{X}$の最もらしいものを$\{\mathbf{X}^1, \mathbf{X}^2, \cdots \mathbf{X}^n\}$（それぞれtrellis上のルートに対応）から選ぶのは$P(\mathbf{X}^i|\mathbf{Z})$を最大にするiを求めればよい。ベイズの定理より、

$$P(\mathbf{X}^i|\mathbf{Z}) = \frac{P(\mathbf{X}^i, \mathbf{Z})}{P(\mathbf{Z})} = \frac{P(\mathbf{X}^i)\cdot P(\mathbf{Z}|\mathbf{X}^i)}{P(\mathbf{Z})} \quad (ii)$$

であるので、結局$P(\mathbf{X}^i)\cdot P(\mathbf{Z}|\mathbf{X}^i)$を最大とするiを求めることになる。一方、(i)式と、trellis上の各枝が、X_kに対応しかつ(S_{k+1}, S_k)に対応している

ので、結局

$$P(\mathcal{X}) \cdot P(\mathcal{Z}|\mathcal{X}) = \prod_k P(\xi_{k+1}|\xi_k) \prod_k P(z_k|x_k)$$

を最大にするルートを trellis 上で求めるには、trellis の各枝に

$$-\ln P(\xi_{k+1}|\xi_k) - \ln P(z_k|x_k) \quad \text{(iii)}$$

を距離として与えれば、trellis 上の最短ルートを求めることに帰着できる。そしてこの最短ルートを求めることはよく知られているように動的計画法（ダイナミック・プログラミング）の手法を用いて容易に現実的な計算時間で解くことができる。（図3）

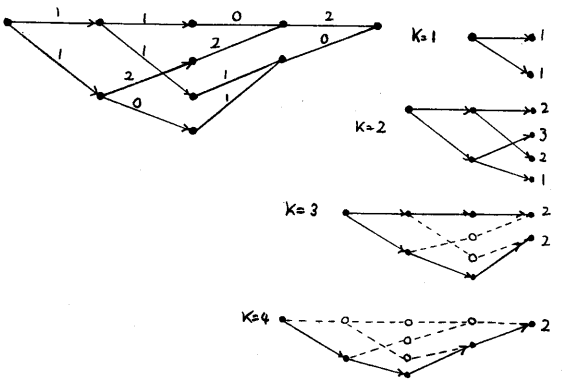


図3 動的計画法による最短パスの探索

3. 漢字仮名変換の形式化

漢字仮名変換の問題をこの Viterbi アルゴリズムに形式化するには、それがマルコフ過程よりの出力文字列 $\mathcal{X}$ が雑音のある通信路を経て文字列 $\mathcal{Z}$ に変形された図1の枠組みで考える必要がある。従って、勿論奇異ではあるが、そもそも単位切りされた仮名書き文 $\mathcal{X}$ が存在しており、それが雑音のある通信路を経て漢字仮名混りのベタ書き文として観測されると考える。当然、ここでの雑音とは、$\mathcal{X}$ 内の仮名の部分文字列が漢字文字列に確率的にお変換えられること、また、$\mathcal{X}$ 中の単位切りの空白が欠落することである。特に単位切りに関しては、応用の目的から、文節単位切り、語基レベルまでの単位切りなど様々なものが考えられる訳だが、それらのいずれもは、通信路上の雑音の設定により形式化できる。

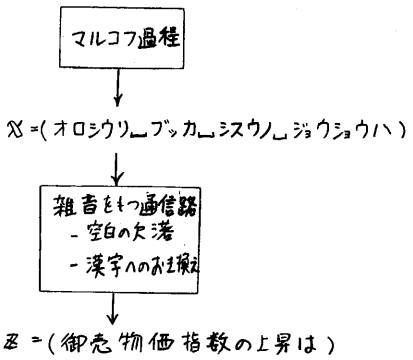


図4 漢字仮名変換のモデル

従って、この枠組での漢字仮名変換とは、観測された $\mathcal{Z}$ を発生し得る全ての単位切りされた仮名書き文を全ルートとして含む前述の trellis を作り、各 trellis 上の枝の距離を (iii) 式により与え、その上での最短ルートを探索することになる。最短ルートが得られれば、そのルートに対応する $\mathcal{X}$ が、漢字仮名混りベタ書き文 $\mathcal{Z}$ の最も確からしい単位切りと仮名訓りである。

3.1 雑音

図4の枠組における雑音は $\mathcal{X}$ 上の部分仮名列が漢字、ひらがな等にお変換えられるもの、また、$\mathcal{X}$ 上の空白が欠落するものの2種を考えねばならない。前者においては、より一般的には次のような確率的雑音 E を考えればよい。

$$E = \{x^* \to z^*,\ p(z^*|x^*)\}$$

このEは任意長の仮名列が、任意長の漢字列におき換えられるので、確率付きの国語辞典を用意することとなる。但しここでの確率とは P(漢字正書|読み)である。　このようなEが与えられれば観測文字列 $\mathcal{Z}$ より、可能な $\mathcal{X}$ をすべてのルートとして持つ trellis を作ることは容易に行える。　すなわちEの逆変換 E' を作れば、$\mathcal{Z}$ 上の部分文字列に E' に含まれる変換を順次適用することですべての $\mathcal{X}$ を含む有向グラフを作れる。(図5)

$$E = \begin{pmatrix} x_1\,x_2 \to z_1\,z_2, & 0.3 \\ x_1\,x_2 \to z_2, & 0.7 \\ x_1 \to z_1, & 1.0 \\ x_2 \to z_3\,z_4, & 1.0 \\ x_3 \to z_4, & 0.9 \\ x_3 \to z_2\,z_3, & 0.1 \end{pmatrix} \quad \text{など}$$

$\mathcal{Z} = (z_1\,z_2\,z_3\,z_4)$

⇩

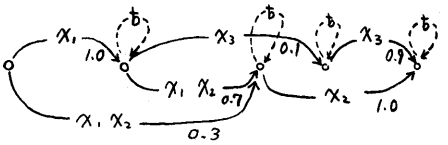


図5　すべての $\mathcal{X}$ を含む有向グラフ

図5において各枝の下に付した数字は雑音Eの変換の起る確率で(iii)式の第2項に対応する。

空白が欠落する雑音は、図5に点線で示した位置に空白を発生するループを加えればよい。　但しこのループは2度以上連続して回わらないので、展開して、ループを含まない有向グラフに展開できる。(図6)

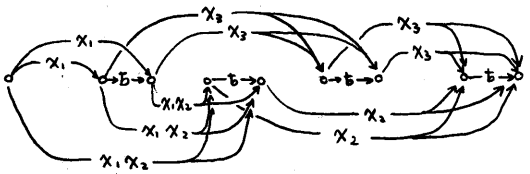


図6　空白欠落を考慮した有向グラフ

従って、図6のグラフ上の全ルートにより可能な $\{\mathcal{X}^i\}$ 及びそれぞれに対する $P(\mathcal{Z}|\mathcal{X}^i)$ を得ることができる。

3.2 n-gram

雑音Eより各 $\mathcal{X}^i$ に対する $P(\mathcal{Z}|\mathcal{X}^i)$ を得ることはできたが、最適な $\mathcal{X}^i$ を選ぶためには、更に各単位切り仮名書き文の生ずる確率 $P(\mathcal{X}^i)$ を知らねばならない。[(ii)式]　しかし、これを自然言語で分野を限らず直接的に得ることはほぼ不可能に近い。

n-gramモデルとは、文字列上の文字の発生確率がそれ以前の $n-1$ 文字に依存するとするモデルで、自然言語のふるまいを近似するのに有効であることが知られている。[12]　特に、このモデルを用いるなら、内部状態を過去 $n-1$ 文字の出力に対応させたマルコフ過程として、問題をViterbiアルゴリズムに帰着することができる。　従って、動的計画法で現実的に解くことができるようになる。

もちろん、自然言語をマルコフ過程で近似することに抵抗はあるが、文献1の外にも文字認識[13]・暗号解読[14]・音声認識[15]などでその有効性が実証されている。　もちろん、nを大きくとれば、それと共に近似はよくなり、遠距離の文脈依存性を持つことになるが、逆に内部状態の数が m^{n-1} (m はアルファベットの数)であるので計算は困難となる。また、$P(x_n|x_1x_2\cdots x_{n-1})$ を計算するための n 文字組の数も m^n となるので、確率の信頼性を保つのが困難となる。

特に、漢字仮名変換などにおいては、漢字とそれの読みとの対応が遠く隔れた文脈に依存することは少ないと考えられるので小さいnで十分近似が可能と思える。

n-gramで近似を行うことにより、(ii)式における第1項は $\mathcal{X}^i = (x_0^i x_1^i x_2^i \cdots x_n^i)$ に対して次のようになる。

但し x_0^i は文頭記号とする。

$$P(x^i) = P(x_1^i | x_0^i) \times \prod_{j=0}^{n-2} P(x_{j+2}^i | x_j^i\ x_{j+1}^i) \quad \text{(iv)}$$

[n-gram の n=3 として]

これは図6の有向グラフ上の1つのルート x^i に対応するので、図6の有効グラフより trellis を作成すれば各頂点が trellis 上の枝の距離 (iii)式の第1項に対応する。 従って3.1の雑音で述べた(iii)式の第2項相当の $P(z^*|x^*)$ と合わせて、trellis 上の各枝の距離が与えられることになり、Viterbi アルゴリズムに帰着できた。

3.3 文脈に依存する雑音への配慮

3.1で述べた雑音としては、仮名文字列の漢字列へのおき換え、空白の欠落の2種があったが、それらの発生確率は文脈に依存しないとしてあった。しかし、それらの発生確率が文脈に依存しないというのは、日本語の現実には合致しない。 特に、文献8に見られるように単位切りのための空白と両隣の文字種の間には強い関係があるし、また、仮名列から漢字列のおき換えは特に漢字列の長さが1であるときしおき換えが漢知辞典型のとき) 両隣の文字種と強い関係がある。(例えば、訓型のおき換えを行う雑音は両隣がひらがなにおき換えられる場合には起り易い、等)

このように雑音のおき換えの発生確率がごく近い文脈に依存して変わり得ることを無視するのは適切でない。 そこで、この雑音の依存性を表現するのに、前述の E を文脈依存型のおき換え形式、例えば $\alpha x^* \beta \to \alpha z^* \beta$, $P(z^*|x^*)$、にすることはできない。 従って、何らかの別の手段を行う必要があるが、次のように解決することができる。

前述の E はおきかえ $(x^* \to z^*; P(z^*|x^*))$ の集合として定義されていたが、それを次のように新たに定義する。

$$E = \{ (x^*, t \to z^* \ ; \ P(z^*|x^*) \} \quad \text{(v)}$$

新たな E における t は、おき換えの種類を表わす記号で、国語辞典型おき書え (z^* の長さ2以上)、漢知辞典音型 (z^* の長さ1で x^* が z^* の音読みである時)、漢知辞典訓型 (z^* の長さ1で、x^* が z^* の訓読みの時)、ひらがな型 (x^*='ヒ', z^*='ひ' 等の時)、カタカナ型、特殊記号型の区別を行う。(それぞれを以後, W, O, K, H, N, S と区別する。)

雑音 E を拡張すると同時に、n-gram における アルファベットも次のように拡張する。 すなわち、これまでの議論では n-gram は x_i すなわち仮名文字の出現確率であったのだが、それを拡張して雑音 E のおき換え規則の左辺の和集合をアルファベットとして考える。 従って従来 $x_i\ x_{i+1}, \cdots, x_{i+n-2}$ から x_{i+n-1} の出現確率を予想するものであったが、今度は n-gram は

$$P((x_5 x_6, t_3) | (x_1, t_1), (x_2 x_3 x_4, t_2)) \quad \text{(vi)}$$

[n-gram の n=3 のとき, $t \in \{W, O, K, H, N, S\}$]

のように仮名の並びの推定 (上の例では $x_1, x_2 x_3 x_4 \to x_5 x_6$) を行うと同時に字種の並びの推定も行う能力をも持つ。($t_1\ t_2 \to t_3$) 当然空白もアルファベットの1つとして含まれるので、文字種の文脈による空白の欠落も n-gram に配慮されることになる。

この新たな雑音 E と n-gram における形式化では、雑音は文脈に依存せずに発生するが、それに対する補償が n-gram により完全に行われることとなる。 これは図1の構造で

通信路における文脈依存性をマルコフ過程に押しやったものであり、Viterbiアルゴリズムの枠組を用いる大きな利点である。

4. 実験

本稿で紹介した漢字仮名変換の新しい方式の有効性を確めるために小規模の実験を行った。

4.1 確率付き漢和辞典

3.1 及び 3.2 に述べたようにこの方式では一般に雑音として国語辞典型の変換を許す訳だが、本実験では、(V)式の z^* の長さを 1 に制限し、確率付きの漢和辞典を用意した。(図7)

E = ハツ,O → 発, p_1
ハッ,O → 発, p_2
パツ,O → 発, p_3
ハツ,K → 初, p_4
イン,O → 引, p_5
イン,O → 印, p_6
ヒ,K → 引, p_7
ヒキ,K → 引, p_8
。
。
。

図7 確率付き漢和辞典

これは国立国語研究所の調査[16]を土台として若干の漢字を加え、現在約2000漢字を含んでいる。なお、音便形など(発→パツ,ハッなど)に関する補正を行うと同時に、用言に用いられる漢字の送りがな(引→ひ,ひく,取→と,とりなど特に連用形)に関する補正を行っているので、通常の漢和辞典より読みの種類が多くなっている。

4.2 n-gram

一方、(iii)式の P_1 項を得るためのn-gramとしては、理想的にはnの大きい方がよいのだが、本実験では $n=3$ と設定した。また実際に(vi)式の形の3-gram(tri-gram)を得るデータ準備の手間を軽減するため、(vi)式のtri-gramを次式のように近似した。

$$P((x_5 x_6, t_3) \mid (x_1, t_1), (x_2 x_3 x_4, t_2))$$

$$\Downarrow$$

$$P(x_5 x_6 \mid x_1, x_2 x_3 x_4) * P(t_3 \mid t_1 t_2)$$

$$\Downarrow$$

$$P(x_6 \mid x_4 x_5) * P(x_5 \mid x_3 x_4) * P(x_4 \mid x_2 x_3) * P(x_3 \mid x_1 x_2) * P(t_3 \mid t_1 t_2) \qquad \text{(vii)}$$

従って、(vi)式を得るために、字種のtri-gramと仮名文字のtri-gramが必要であった。

字種のtri-gramは信頼のおけるデータが手に入らなかったので、文献8の坂本氏の調査を基にして簡単なモデルによる推定値を加え代用した。
一方、仮名文字のtri-gramは、仮名書きして1200～1500字程度の文章を36ケ集め、単位切り仮名書きし、仮名文字の3つ組を集計して作成した。文章の分野としては、新聞の社説約半分、コンピュータ関連のマニュアル、雑誌記事が約1/4、残りが、社内規則である。
ここでの単位切りは文節切りに加えて、4文字以上の複合語内を2文字ずつまたは3文字の語基に単位切りすることを行った。(図8) 従って後に示す結果の単位切りもその方式が反映している。

チホウ ブンケンノ リネンヲ イカニ
ダイイチジ チホウ セイドデ チョウサカイガ、アタラシ
アリカタニツイテ トウシンアンヲ マトメタ。ヨカッチュウニ
トウシンアンノ ナイヨウハ、クニ、チホウヲ ツウジタ
オオキナ ハンラトナッテイル ソノタメ トシ ケイカク ノウチ
キョウセイ ケンケンヲ チホウニ イショウ、コッコ 而
チホウ デサキ キカンヲ ハイシ トウゴウスル サラニ チホ
キョウセイ ケイタイ キョウセイ イインカイ セイドキ
コンポンテキ モンダイモ テイキサレテイル

図8 訓練用データ

特に単位切り出れた訓練用データから仮名の3つ組を集計する際に、幼音、長音、濁音、半濁音などは先行する文字が限定されるので、独立した仮名文字とは扱わなかった。 従って、むしろ、モーラ単位のtri-gramであるといえる。

4.3 解析

解析プログラムはPL/1で約3000ステップで図9の構造である。 入力文に対し、まず図7の確率付き漢和辞典を引き、図6の形の{文}を含む有向グラフを作成する。 次に有向グラフの図9のウインドウ部分をtrellisに部分的に展開する。(trellis全体を展開するのはメモリーを食うため) このウインドウを有向グラフの前から後ろへ移動すると同時に動的計画法により最短ルートが得られる。
図10に、入力列「再上昇に転じた米金利」に対して得られる図6形式の有向グラフを例として示す。

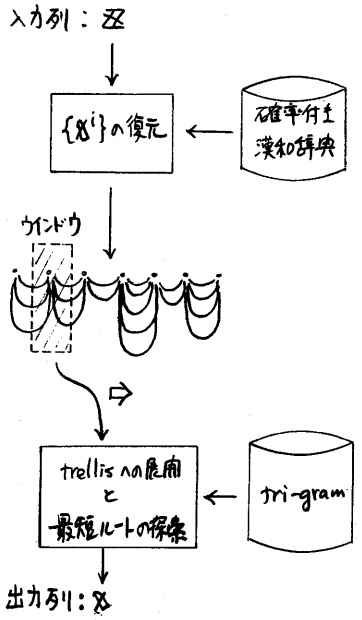


図9 解析の流れ

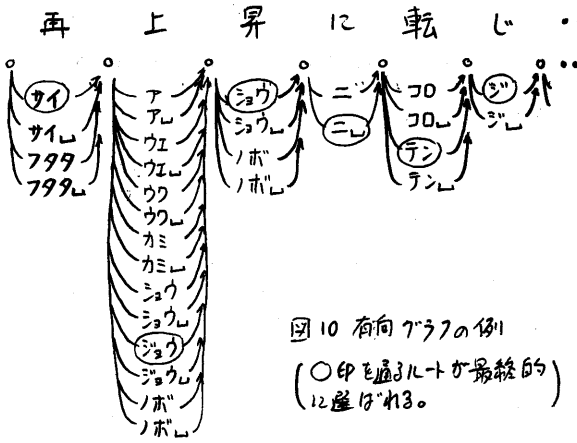


図10 有向グラフの例
(○印を通るルートが最終的に選ばれる。)

5 結果と考察

新聞の社説を適当に選んで本方式での漢字仮名変換を試みた例の一部を図11に示す。

(図11)

引き続き高水準を続けている。一月の卸売物価指数は完成品・準耐久消
() () 引(ヒ)き(キ)続(ツヅ)き(キ) 高(コウ)水(スイ)準(ジュン)を(ヲ) 続(ツヅ)け(ケ)て(テ)い(イ)る(ル)。(.) 一(イチ)月(ガツ)の(ノ) 卸(オロシ)売(ウリ) 物(ブッ)価(カ) 指(シ)数(スウ)は(ハ) 完(カン)成(セイ)品(ヒン)。(.) 準(キ)耐(セツ) 消(チョウ)整(セイ)済(ザイ)。(.)

ベースで年率約二〇%も上昇した。新方式の採用後、しばらく
() () ベー(ヘー)ス(ス)で(デ) 年(ネン)率(リツ)約(ヤク) D(D)。(.) も(モ) 上(ジョウ)昇(ショウ)し(シ)た(タ)。(.) 新(シン) 方(ホウ)式(シキ)の(ノ) 採(サイ)用(ヨウ)後(ゴ)。(.) し(シ)ば(バ)ら(ラ)く(ク)。(.)

落ち着いていた通貨供給量や銀行貸し出しの増勢にも、昨年末から
() () 落(オ)ち(チ)着(ツ)い(イ)て(テ)い(イ)た(タ) 通(ツウ)貨(カ) 供(キョウ)給(キュウ)量(リョウ)や(ヤ) 銀(ギン)行(コウ)貸(カ)し(シ)出(ダ)し(シ)の(ノ) 増(ゾウ)勢(セイ)に(ニ)も(モ)。(.) 昨(サク)年(ネン)末(マツ)か(カ)ら(ラ)。(.)

再び強まる傾向がみられる。FRBとしては一月の卸売物価の動きが
() () 再(フタタ)び(ヒ) 強(ツヨ)ま(マ)る(ル) 傾(ケイ)向(コウ)が(ガ) み(ミ)ら(ラ)れ(レ)る(ル)。(.) A(A)と(ト)し(シ)て(テ)は(ハ) 一(イチ)月(ガツ)の(ノ) 卸(オロシ)売(ウリ) 物(ブッ)価(カ)の(ノ) 動(ウゴ)き(キ)が(ガ)。(.)

確認した物価の騰勢持続、原油の値上がりがそれをさらに強めそうな
() () 確(カク)認(ニン)し(シ)た(タ) 物(ブッ)価(カ)の(ノ) 騰(トウ)勢(セイ) 持(ジ)続(ゾク)。(.) 原(ゲン)油(ユ)の(ノ) 値(ネ)上(ア)が(ガ)り(リ)が(ガ) そ(ソ)れ(レ)を(ヲ)さ(サ)ら(ラ)に(ニ) 強(ツヨ)め(メ)そ(ソ)う(ウ)な(ナ)。(.)

情勢を前に、改めてここで政策方針を明示する必要を感じたのであろう。
() () 情(ジョウ)勢(セイ)を(ヲ) 前(マエ)に(ニ)。(.) 改(アラタ)め(メ)て(テ) こ(コ)こ(コ)で(デ) 政(セイ)策(サク) 方(ホウ)針(シン)を(ヲ) 明(メイ)示(ジ)す(ス)る(ル) 必(ヒツ)要(ヨウ)を(ヲ) 感(カン)じ(ジ)た(タ)の(ノ)で(デ)あ(ア)ろ(ロ)う(ウ)。(.)

アフガニスタン問題の発生に伴う米国の軍需生産の拡大、情報関連の
() () ア(ア)フ(フ)ガ(ガ)ニ(ニ)ス(ス)タ(タ)ン(ン)問(モン)題(ダイ)の(ノ) 発(ハッ)生(セイ)に(ニ) 伴(トモナ)う(ウ)米(ベイ)国(コク)の(ノ) 軍(グン)需(ジュ) 生(セイ)産(サン)の(ノ) 拡(カク)大(ダイ)。(.) 報(セン)関(リャク) 連(フッ)情(ジ)の(ノ)。(.)

表1にその社説全体における結果を示す。

a	総文字数	1072
b	総漢字数 (漢数字は一以外は含まない)	461
c	誤って読まれた漢字数	23
d	読みふり正答率 (c/b)	95.01%
e	単位切り位置総数	273
f	余分に挿入された空白数	7
g	単位切り忘れた個数	16
h	単位切り正答率 $\left(\frac{e-f-g}{e}\right)$	91.58%

(注) ここでの正しい単位切り位置とは文節内区切りと、複合語内語基内区切りである。

表1 結果

表1の結果は、訓練データの少なさ、(約45K仮名文字)に対して十二分に満足できるものと考えられる。訓練データを蓄積することにより、さらに精度は向上すると思われる。

並に、この実験もしくは形式化の現界を示す誤りも若干見られる。例えば「国家的計画、」が「コッカテキ」ケイガ、」のように変換された。これは(vi)のtri-gramを(vii)のように近似したことに起因していると考えられる。すなわち(vii)式の近似により tri-gramで評価される仮名列がすでにそれらの変換されるべき字種を忘れてしまっている。従って、この例でも「計画」より得られる仮名列「ケイカク」と「ケイガ」の比較で、後者が「系が」、「計が」などと圧倒的に頻度の高い別の同音表現のために優先されてしまっている。

別な例で興味深いのは「歩合」がいつでも「フゴウ」と呼まれてしまうことである。これも「歩合」から得られる「ブアイ」、「フゴウ」の比較において、「符号」、「符合」、「互号」、「富豪」などのより頻度の高い同音語にひきずられていると思われる。

単位切りが91.58%とあまりよくないのは主として長い複合語分解の誤りによっている。例えば、「連邦準備理事会」は「連」邦」準備理」事会」のように分割された。これは、現在のモデルの複合語分解が、漢語は2文字単位が多いなどの複合語に関するいくつかの常識を用いず、単に空白を仮名列内、字種内に含んだ場合のtri-gramによる評価だけで行っているからと考えられる。これは近い将来の課題である。

おわりに

本稿で紹介したViterbiアルゴリズムによる漢字仮名変換は実用的な観点からは精度が十分でないとの見方もあろう。しかしまえ書きにも述べたように、この上に例外辞書、例外コードなどを積み上げるための土台として十分有効と考える。また、モデル自体にもまだ幾多の本質的改良の可能性がある。実験も小規模でこのモデルの性能を十二分に引き出していない点もある。しかし、本稿で紹介したViterbiアルゴリズム及びそれでの漢字仮名変換の形式化は全く新しく、他の関連問題でも十分に有効な手段であると考えたので報告を行った。

なお、最短ルートの探索パスのプログラム作成で、当センターの客員研究員瀬川清氏(早稲田大学)に協力いただいた事を感謝する。

参考文献

[1] 藤崎他、日本語文書処理システム「ことだま」-概念と設計思想、東京サイエンティフィック・センター・レポート、N:G318-1512, 1980

[2] 藤崎他、日本語文書処理システム「ことだま」の仮名漢字変換、情報処理論文誌, vol 23, No.1, 1982 (予定)

[3] 籐崎、諸橋、ことだま文書処理システムの文書検索機能，IBMレビュー 85,1981 (予定)
[4] 大河内他、仮名漢字変換のための文法解析、情報処理計算言語学研究会資料 25-4,1981
[5] 田中、漢字かなまじり文を全文カナ書き、ローマ字書きに変換するシステムについて、国立国語研究所報告 (34), 1969
[6] 中野、言語研究のための日本語データ入力システム、日本語情報処理シンポジウム報告集、1978
[7] 荒木、他、JICSTの実用的全自動漢字ーカナ変換システム、K-KACSについて、情報処理、Vol 20. No.10, 1979
[8] 坂本、文節の認定、日本語情報処理シンポジウム報告集、1978
[9] 長尾、他、国語辞書の記憶と日本語文の自動分割、情報処理、Vol 19, No.6, 1978
[10] G.D.FORNEY,JR., The Viterbi Algorism, Proceedings of IEEE, Vol.61, No.3, 1973
[11] A.J.Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorism, IEEE Trans. Information Theory, vol.IT-13, 1967
[12] C.E.Shannon, Prediction and Entropy of Printed English, Bell Sys. Tech. J., Vol.30, 1951
[13] J.Raviv, Decision Making in Markov Chains Applied to the Problem of Pattern Recognition, IEEE Trans. Information Theory, Vol IT-3, No.4, 1967
[14] L.R.Bahl, et.al., Disambiguation of Stenotype and Other forms of Ambiguous Text by Use of Contextual Statictics, personal memo, IBM T.J.Watson Res. Ctr, 1980
[15] J.Jelinek, Continuous Speech Recognition by Statistical Methods, Proc. of the IEEE, Vol.64, No.4, 1976
[16] 現代新聞の漢字、国立国語研究所、秀英出版，1976