統計的手法による漢字複合語の自動分割†

武田浩一††　藤崎哲之助††

日本語処理における複合語の分割は，機械翻訳，自動インデクシング，文書校正，音声合成等で必要とされる基本的技術であるが，従来より困難な問題であることが指摘されてきた．これは複合語の分割が必ずしも一意でないためであり，最長一致法等の手法による自動分割では十分な分割精度を得ることができなかった．本報告では，漢字複合語をマルコフモデルという確率的情報発生源からの出力であると考え，統計的推定による手法を用いた短単位分割法を提案し，その処理手順と実験結果について述べる．現行の実験システムでは漢字のみからなる一般語しか扱っていないが，本手法の特徴には以下のものがある．1）適用分野で用いられる十分多くの漢字複合語をもとに，正しい短単位が機械的な計算により学習できる．2）複合語の分割に曖昧さがあるときに，最も確からしい分割パターンが求められる．3）基本語の出現頻度順のリストや分布といった計量的データの収集が可能となる．本システムは，JICST より発行されている科学技術論文の抄録データに対して約95%の平均分割精度を達成している．また，あらかじめ用意された辞書の正書項目を利用したり，頻出語の正しい分割パターンを与えるといった各種の改良のもとで約97%の分割精度を得た．今後の課題には，未知語の扱いや，一般的な漢字複合語以外の分割への拡張があげられる．

1. まえがき

日本語文の形態素解析に含まれる複合語の分割は，機械翻訳[4]，自動インデクシング[13]，文書校正[19]，音声合成[11]等で必要とされる基本的技術であるが，困難な問題の一つであった[5]．これは複合語の分割が必ずしも一意でないためであり，従来の最長一致法[7]等の手法による自動分割では十分な分割精度を得ることができなかった．宮崎[12]は係り受け解析法により広汎な複合語を99.8%という高精度で分割しているが，人手により意味情報を含めた辞書項目を整備する作業は一般にコストが高く，適用分野が変化した場合や未知語の扱いが問題となる．

本論文では，漢字複合語をマルコフモデルという確率的情報発生源からの出力であると考え，統計的推定による自動分割法を提案し，その処理手順と実験結果について述べる[1)-3)]．ここで用いた統計的推定法は，英語の連続音声認識[14]や，構文解析木が複数個生じるような曖昧さがある英語の文脈自由文法のもとで，最も確からしい構文解析木を決定する手法[16]において，その有効性が検証されている．日本語の形態素解析に対する統計的手法には既に藤崎[8),9)]の N-gram モデル[18]を用いた方法が知られているが，我々のモデルは漢字複合語の自然な構造を表現しており，造語モデルとして優れているといえる．西野ら[6]はさらに我々の提案したモデル上での複合語の構造解析を試みている．同様の手法は松延ら[10]によっても報告されている．

現行の実験システムでは漢字のみからなる一般語しか扱っていないが，本手法には以下のような大きな特長がある．

1）各適用分野における日本語文書に現れる十分多くの漢字複合語をもとに，正しい単語が機械的な計算により学習できる．

2）複合語の分割に曖昧さがあるときに，その分野における最も確からしい分割結果が得られる[1)]．

3）基本語の出現頻度順のリストや分布といった計量的データの収集が可能となる[2)]．

4）既存の仮名漢字辞書における同音異義語の優先順序の決定や，辞書項目の整備に利用できる．

本手法の実験では，JICST より発行されている科学技術論文の抄録データに対して約95%の平均分割精度を達成している．この精度は処理上の工夫により1%程度改良できた．また，あらかじめ用意された辞書の正書項目を利用した場合についても実験を行い，約97%の精度を得た．辞書項目を利用すると，初期段階から正しい語基や接辞がほとんど得られているため，マルコフモデルに発生する状態数が少なくてすみ，現実の複合語のよりよい近似となっているものと考えられる．したがって，未知語の存在を仮定したり，固有名詞，数詞，かな，英数字等からなる複合語への拡張を考えても，既存の辞書と我々の手法を併用することにより高精度の複合語分割が実現できる．

2. 短単位モデル

日本語文書における一般の漢字複合語のほとんど

† Automatic Decomposition of Kanji Compound Words Using Stochastic Estimation by KOICHI TAKEDA and TETSUNOSUKE FUJISAKI (Tokyo Research Laboratory, IBM Japan Ltd.).

†† 日本アイ・ビー・エム(株)東京基礎研究所

は，2文字の単語（以後，これを語基と呼ぶ）と，接頭辞や接尾辞の働きをする1文字の漢字から構成されていることが知られている[4]．すなわち，このような複合語は

（接頭辞）* 語基（接尾辞）*

という線形表現の語のまとまり（以後これを短単位と呼ぶ）の連接として表される．ここで*は0回以上の繰返しを示している．ただし，簡単のため『人』や『文』のように通常1文字で単独に使われる語も，便宜上接頭辞か接尾辞として扱う．複合語分割の曖昧さは，例えば

信頼性向上

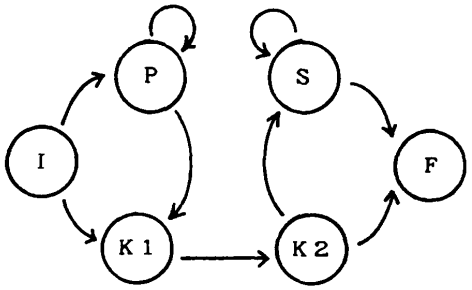
という複合語に対し，

1) 信頼・性 向上
2) 信頼 性向・上

といった異なる短単位を含む複数の連接パターンが存在することにより生じる（語基と接辞の境界を『・』で示した）．日本語の形態素解析では，『信頼性向上』のような長い語の形態素の情報をすべて辞書に登録できないので，『信頼』や『向上』といった短単位の語から正しい解を求めるために，最長一致法[7]や，語基や接辞の使用回数に基づく頻度情報を利用した方法[4]がよく用いられる．ただし，これらの方法による漢字複合語の分割精度は必ずしも十分でなく，実用的な精度を達成するためには，例えば宮崎[12]の人手により入念に整備された情報をもつ辞書が必要であった．

これに対し，我々の手法は，与えられた日本語文書から正しい語基や接辞を学習し，語基や接辞ごとの出現確率の積を用いて最も高い確率で出現する連接パターンを推定しようとするものである．その道具として数学的なマルコフモデルによる短単位の表現を導入した．これを短単位モデル[1]と呼ぶ．日本語文書の各適用分野における正しい語基や接辞，およびその出現確率は未知であるから，モデル上で正しい語基や接辞の学習と，正確な出現確率の計算を行うアルゴリズムを与える必要がある．以下ではその方法について説明する．

図1に短単位モデルを示す．Pは接頭辞，Sは接尾辞，$K1$および$K2$はそれぞれ2文字語基の1文字目と2文字目を表している．IとFは短単位の始まりと終りを示す特別な状態である．実際に現れる漢字複合語（ただ一つの短単位からなるものを含む）を入力として，この短単位モデルは図2のような形に展開される．図2の例では，



I：初期状態　　F：終了状態

P：接頭辞　　S：接尾辞

K1，K2：2文字語基の1文字目と2文字目

図1 短単位モデル
Fig. 1 A primitive word model.

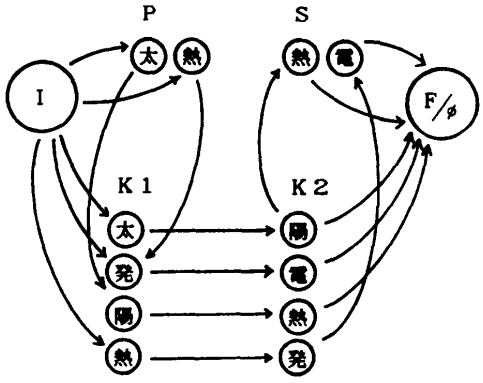


図2 展開された短単位モデル
Fig. 2 An expanded primitive word model.

太陽，太陽熱，発電，熱発電

という語が現れたと仮定したときに得られる展開形を示している．ここでは，Iを除いた各状態はPやSといったラベルLと漢字1文字aの対からなり（これをL/aと書く），状態s_1から状態s_2への遷移により，s_2の漢字が出力として観測されるものとしている．s_2がF/ϕのときには，短単位の区切りを示す特別な記号ϕを出力するが，これは観測されない記号であるとする．（フィルタにより出力系列からϕを取り除いたものが観測系列となるような系を考えている．）太陽熱発電といった複数の短単位からなるものは，短単位モデルにF/ϕからIへの遷移を許すことにより，途中にこの遷移を含むような状態遷移系列に対応する観測系列として扱うことができる．これはF/ϕとIを同一状態とみなすことに等しい．

次章で述べる状態遷移確率推定アルゴリズムを用いて，図2の展開形に状態遷移確率が設定される．これ

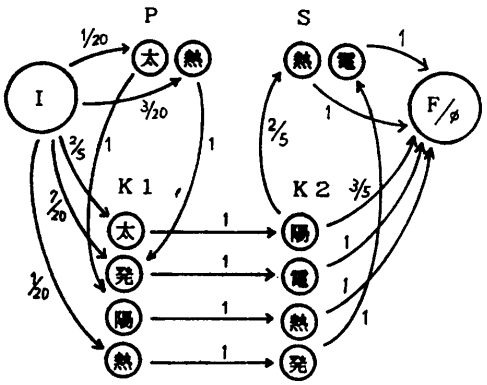


図3 図2に確率を与えた短単位モデル
Fig. 3 The primitive word model of Fig. 2 with associated probability.

を図3に示す．漢字複合語の短単位分割は，この確率付き短単位モデルにおいて最も高い状態遷移確率の積を生じる遷移系列を求めることに帰着される．

例) 図3において『太陽熱発電』という観測系列を発生する状態遷移系列は

(a) $I, P/$太$, K1/$陽$, K2/$熱$, F/\phi$,
$K1/$発$, K2/$電$, F/\phi$

(b) $I, K1/$太$, K2/$陽$, S/$熱$, F/\phi$,
$K1/$発$, K2/$電$, F/\phi$

(c) $I, K1/$太$, K2/$陽$, F/\phi$,
$P/$熱$, K1/$発$, K2/$電$, F/\phi$

(d) $I, K1/$太$, K2/$陽$, F/\phi$,
$K1/$熱$, K2/$発$, S/$電$, F/\phi$

の4種類あり，各遷移系列の生起確率は，(a) 0.0175, (b) 0.056, (c) 0.036, (d) 0.012 となる．したがって，『太陽熱発電』の最も確からしい状態遷移系列は(b)であり，これは『太陽・熱』と『発電』という分割に対応する．

適用分野における漢字複合語は，この短単位モデルから観測される出力であり，個々の漢字複合語の出現頻度はその出力の生起確率によって定まる．したがって，十分多くの漢字複合語とその出現回数をもとに短単位モデルを展開すれば実際に現れる漢字複合語の発生源のよい近似となることが予想される．

図3の確率付き短単位モデルの表記法として，つぎのものを用いる．

(1) $S=\{s_i \mid s_i=L_i/k_i\}$：状態集合
ここで L_i はラベル，k_i は漢字を示す．I は F/ϕ と同一の状態であるとする．

(2) $T=\{t=(s_i, s_j) \mid s_i, s_j \in S\}$：状態遷移集合

(3) 状態遷移 $t=(s_i, s_j)$ に対し，

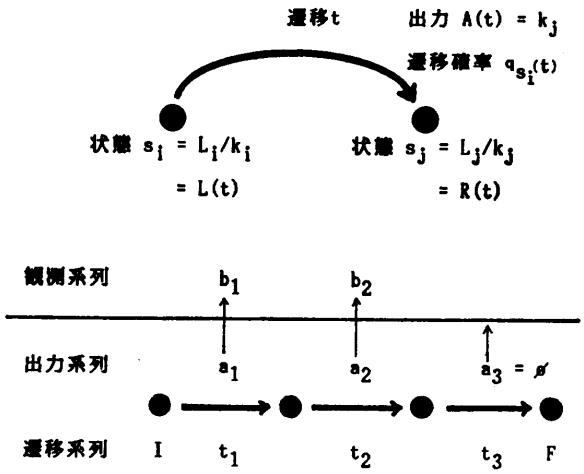


図4 短単位モデルの表記
Fig. 4 Notations for the primitive word model.

$L(t)=s_i$：t の始状態
$A(t)=k_j$：t の出力
$R(t)=s_j$：t の終状態
$q_s(t)=0$ ($L(t)\neq s$ のとき)
$=t$ の状態遷移確率 ($L(t)=s$ のとき)

(4) 短単位モデルM上の完全パス $t_{1,n}$
$t_{1,n}=t_1, t_2, \cdots, t_n$
$(L(t_1)=I, R(t_n)=F/\phi)$
およびその生起確率

$$p(t_{1,n})=q_I(t_1)\prod_{i=1}^{n-1} q_{R(t_i)}(t_{i+1})$$

ここで完全パスは，一つの漢字複合語に対する状態遷移系列を表している．完全パスが与えられると，出力系列 $a_{1,n}$ は一意に定まり，$A(t_1), \cdots, A(t_n)$ となる．漢字複合語は $a_{1,n}$ より $A(t_i)=\phi$ となる出力を取り除いた観測系列 $b_{1,m}=b_1, \cdots, b_m$ である．図4にこれらの表記を簡単にまとめている．

3. 状態遷移確率推定アルゴリズム

状態遷移確率推定アルゴリズム[2] (以後単に推定アルゴリズムと呼ぶ) は，マルコフモデルにおけるベイズの事後確率推定法を利用したアルゴリズムである．推定アルゴリズムは，

(1) 初期確率の推定

(2) 繰返し計算による確率の更新

の二つのステップに大別される．

前章の図2から図3のような展開を行うために，適用分野に出現する漢字複合語を集めておく．これをト

レーニングデータと呼ぶ．信頼性の高い確率を得るために，トレーニングデータのサイズはできるだけ多いほうがよい．十分多くのトレーニングデータを用いて展開された短単位モデル上で，各状態遷移の発生回数をカウントすれば，（サイコロを何度も振ってそれぞれの目がでる確率を知るように）各状態遷移確率を計算できる．初期確率の計算は以下のように行う．

短単位モデル上では，各トレーニングデータを生じうる完全パスの数は，トレーニングデータの長さにのみ依存する．例えば，長さ2のトレーニングデータ b_1b_2 に対しては，$I, K1/b_1, K2/b_2, F/\phi$ というただ一つの完全パスが定まり，長さ4のトレーニングデータ $z=b_1b_2b_3b_4$ に対しては，

$I, P/b_1, P/b_2, K1/b_3, K2/b_4, F/\phi$

$I, P/b_1, K1/b_2, K2/b_3, S/b_4, F/\phi$

$I, K1/b_1, K2/b_2, F/\phi, K1/b_3, K2/b_4, F/\phi$

$I, K1/b_1, K2/b_2, S/b_3, S/b_4, F/\phi$

の4種類の完全パスがある．

直観的には，上記の z が1回出現したとすれば，これらのパスのうち正しいものを1回通ったものと考えればよいが，正しい短単位に対応したパスが未知であるため，それぞれのパスを1/4回通ったものとする．したがって，各パスに含まれる状態遷移もそれぞれ1/4回通ったものとする．

このような方法で各トレーニングデータを出現させる状態遷移の回数をカウントし，

{状態遷移 (s_i, s_j) のカウントの総和}/{すべての状態 s_{any} についての状態遷移 (s_i, s_{any}) のカウントの総和}

を計算すれば，状態 s_i に達したときに状態 s_j に遷移する確率を求めることができる．5章で述べるように，各パスに重みをつければ，より現実的な初期確率を設定することができる．

初期確率を得ると，繰返し計算により確率値を更新する．この方法により，初期確率として非常に大まかな値を設定しているにもかかわらず，トレーニングデータの生起確率を極大にするような状態遷移確率へ収束することが保証される．

繰返し計算過程は，各完全パスの生起確率と相対出現頻度が計算できるため，各トレーニングデータの出現回数に相対出現頻度を乗じたものをカウントとして同様の計算を行う．トレーニングデータ D は一般に重複を許した複合語 G_i の集合であり，各 G_i はそれを生成する完全パス g_{ij} の集合を用いて $G_i=\{g_{i1}, \cdots, g_{in}\}$ と書ける．

例）トレーニングデータ $D=$ {太陽(100), 発電(50), 太陽熱(20), 熱発電(10), 陽気(10), 熱意(10)} とする．ここで（ ）内は出現回数を示す．このとき $G_1, \cdots, G_6$ は次のようになる．

$G_1=\{g_{11}\}$,

$g_{11}=(I, K1/太), (K1/太, K2/陽), (K2/陽, F/\phi)$

$G_2=\{g_{21}\}$,

$g_{21}=(I, K1/発), (K1/発, K2/電), (K2/電, F/\phi)$

$G_3=\{g_{31}, g_{32}\}$

$g_{31}=(I, P/太), (P/太, K1/陽), (K1/陽, K2/熱), (K2/熱, F/\phi)$

$g_{32}=(I, K1/太), (K1/太, K2/陽), (K2/陽, S/熱), (S/熱, F/\phi)$

$G_4=\{g_{41}, g_{42}\}$

$g_{41}=(I, P/熱), (P/熱, K1/発), (K1/発, K2/電), (K2/電, F/\phi)$

$g_{42}=(I, K1/熱), (K1/熱, K2/発), (K2/発, S/電), (S/電, F/\phi)$

$G_5=\{g_{51}\}$,

$g_{51}=(I, K1/陽), (K1/陽, K2/気), (K2/気, F/\phi)$

$G_6=\{g_{61}\}$,

$g_{61}=(I, K1/熱), (K1/熱, K2/意), (K2/意, F/\phi)$

したがって初期値の設定は例えば $t=(I, K1/太)$ なら

$$q_I(t)=(100/1+20/2)/(100/1+50/1+20/2+20/2+10/2+10/2+10/1+10/1)=11/20$$

となる．この計算結果により**図5**を得る．

繰返し計算の場合は，上記の完全パスに対して次のような相対頻度 $f(g_{ij})$ を得る．ここで $\sum_{j=1}^{n} f(g_{ij})=1$ となる．

$f(g_{11})=f(g_{21})=f(g_{51})=f(g_{61})=1$

$f(g_{31})=1/3$

$f(g_{32})=2/3$

$f(g_{41})=1/2$

$f(g_{42})=1/2$

よって同様に $t=(I, K1/太)$ を求めると

$$q_I(t)=\{100\times f(g_{11})+20\times f(g_{32})\}/\{100\times f(g_{11})+50\times f(g_{21})+20\times f(g_{31})+20\times f(g_{32})$$

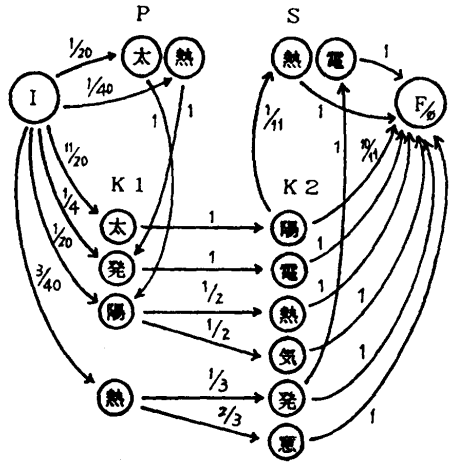


図 5 初期確率の設定
Fig. 5 Calculation of initial probability.

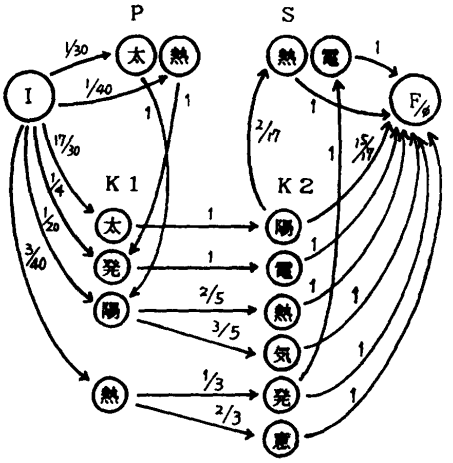


図 6 繰返し計算による確率の更新
Fig. 6 Update of probability by repetitive calculation.

$$+10\times f(g_{41})+10\times f(g_{42})+10\times f(g_{51})+10\times f(g_{61})\}=\frac{17}{30}$$

となる．この結果を図 6 に示す．

繰返し計算を何度も行うことにより，現実の漢字複合語をうまく近似できると考えられる．この計算法は，英語の連続音声認識の分野で使われその有効性が知られている Forward-Backward アルゴリズム[14]と等価であり，同アルゴリズムのもつ次の数学的性質[15]を満足する．

［確率値の単調性と収束性］

与えられた短単位モデルMにおける各トレーニングデータ G_i の生起確率を $p(G_i, M)=p(g_{i1}, M)+\cdots+p(g_{in}, M)$ とし，トレーニングデータ集合 D の生起確率を $p(D, M)=p(G_1, M)+\cdots+p(G_m, M)$ とする．（ここで $p(g_{ij}, M)$ は上例の $p(g_{ij})$ をM上で計算したものとする．）また 1 回の繰返し計算 U により確率値を更新された短単位モデルを $M'=U(M)$ とする．このとき，

$$p(D, M)\leq p(D, M')$$

が成立する．さらに，ここで等号が成立するのは，

$$M=M'$$

のとき，かつそのときのみである．

したがって繰返し計算の順次適用により，トレーニングデータの生起確率を極大にする確率値をもつ短単位モデルへ近づけることができる．計算機上で有限の数値を扱う場合は，Mのエントロピー等の計算により，Mと M' の近似度を求め推定アルゴリズムの停止性を保証するとよい．

繰返し計算過程は初期確率から機械的に更新確率を生成するので，現実に現れる漢字複合語をできるだけうまく近似するためには，適切な初期値を決定することが重要である．これについては 5 章で述べる．

本報告で示した推定アルゴリズムの計算は，簡単のため完全リストを個々に求めるように書いているが，文献 15) の議論とトレリス[8]というデータ構造を用いることで，トレーニングデータの長さに比例した時間で行える．

4. 漢字複合語アルゴリズム

漢字複合語分割アルゴリズム（以後，分割アルゴリズムと呼ぶ）は，状態遷移確率の積を求める際にその対数をとり，対数の加法性を利用して効率よく生起確率が最大の完全パスを求めるものである．これは動的計画法[8]あるいはマルコフモデル上の Viterbi アルゴリズム[17]として知られている手法を利用している．

例）図 5 の短単位モデルを用いて

『太陽熱発電』$=b_1b_2b_3b_4b_5$

を分割する．

STATE^i で i 番目の出力 b_i までを観測したときに到達しうる状態のリストを示し，PRED$^i(s)$ で STATE^i 中の状態 s に到達しうる状態遷移系列のうち，その生起確率が最大のものを示す．

STATE^1 および PRED^1 は次のようになる．

STATE$^1=\langle P/太, K\,1/太\rangle$

PRED$^1(P/太)=\langle(I, P/太)\rangle$,

$p(\mathrm{PRED}^1(P/太))=1/20$

$\mathrm{PRED}^1(K\,1/太)=\langle(I, K\,1/太)\rangle$,

$p(\mathrm{PRED}^1(K\,1/太))=11/20$

同様に STATE^2 および PRED^2 は次のようになる.

$\mathrm{STATE}^2=\langle K\,1/陽, K\,2/陽\rangle$

$\mathrm{PRED}^2(K\,1/陽)=\langle(I, P/太), (P/太, K\,1/陽)\rangle$,

$p(\mathrm{PRED}^2(K\,1/陽))=1/20$

$\mathrm{PRED}^2(K\,2/陽)=\langle(I, K\,1/太), (K\,1/太, K\,2/陽)\rangle$,

$p(\mathrm{PRED}^1(K\,1/陽))=11/20$

STATE^3 では F/ϕ を経由する遷移

$K\,2/陽 \rightarrow F/\phi(=I) \rightarrow P/熱$

$K\,2/陽 \rightarrow F/\phi(=I) \rightarrow K\,1/熱$

があり,

$\mathrm{STATE}^3=\langle S/熱, K\,2/熱, P/熱, K\,1/熱\rangle$

$\mathrm{PRED}^3(S/熱)=\langle(I, K\,1/太), (K\,1/太, K\,2/陽), (K\,2/陽, S/熱)\rangle$,

$p(\mathrm{PRED}^3(S/熱))=1/20$

$\mathrm{PRED}^3(K\,2/熱)=\langle(I, P/太), (P/太, K\,1/陽), (K\,1/陽, K\,2/熱)\rangle$,

$p(\mathrm{PRED}^3(K\,2/熱))=1/40$

$\mathrm{PRED}^3(P/熱)=\langle(I, K\,1/太), (K\,1/太, K\,2/陽), (K\,2/陽, F/\phi), (F/\phi, P/熱)\rangle$,

$p(\mathrm{PRED}^3(P/熱))=1/80$

$\mathrm{PRED}^3(K\,1/熱)=\langle(I, K\,1/太), (K\,1/太, K\,2/陽), (K\,2/陽, F/\phi), (F/\phi, K\,1/熱)\rangle$,

$p(\mathrm{PRED}^3(K\,1/熱))=3/80$

これを $\mathrm{STATE}^6=\langle F/\phi\rangle$ まで行うと,

$\mathrm{PRED}^6(F/\phi)=\langle(I, K\,1/太), (K\,1/太, K\,2/陽), (K\,2/陽, S/熱), (S/熱, F/\phi), (F/\phi, K\,1/発), (K\,1/発, K\,2/電), (K\,2/電, F/\phi)\rangle$

なる解をえる.

ここで厳密に同じ生起確率をとる複数の PRED が存在するときには,解の曖昧さは解決できない.このような曖昧さは語基単位の N-gram[18] や意味情報によらないと解消できないと考えられる.ただし,5章の実験ではこのような場合は起こらなかった.

分割アルゴリズムは文献 15) の議論により,与えられた複合語の長さと短単位モデルのラベルの数の多項式に比例した時間で計算を終えることが示せる.ここである一つの漢字 a を生成できる状態の数は高々四つ ($P/a, K_1/a, K_2/a, S/a$ の四つ) であるから,結果的にこの計算は複合語の長さに比例した時間で終えられることが示せる.

5. 実験システム

前章までの手法を計算機上に実現したので,その実験結果について述べる.

システムの構成は図7に示すようになっている.適用分野は JICST の電気工学編 (Vol. 26) のテープ22巻よりなる技術論文の抄録である.ここから確率計算を行うための漢字連続 (これをトレーニングデータと呼ぶ) を機械的に抽出した.これらの漢字連続には漢字複合語としては正しくないもの (『実験及(び)』や『以上特(に)』) も含むが,以後これらの漢字連続を漢字複合語と呼ぶ.抽出された漢字複合語の延べ出現回数と異なり語数を表1に示す.トレーニングデータとして使用したのは,延べ2回以上現れた,長さが2,3,4の漢字複合語である.このような部分集合を用いた理由は

1) 漢字複合語の特徴は長さ2,3,4のものでだいたい近似できる.これ以上の長さのものはほとんど1回のみ出現しただけである.

2) 1回のみ出現したものを除外することで,正し

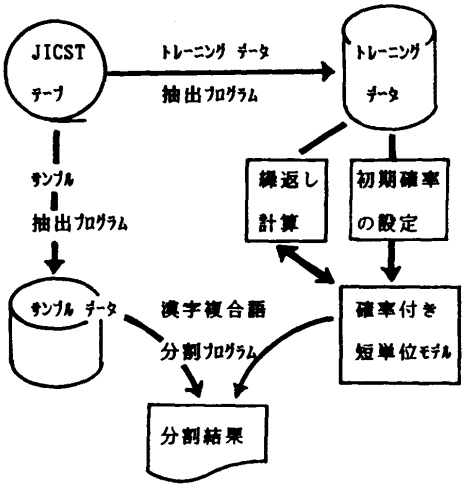


図7 実験システムの構成
Fig. 7 Overview of an experimental system.

表1 漢字複合語の延べ出現回数と異なり語数
Table 1 Total number of occurrences and number of distinct words of Kanji compound words.

文字数	1	2	3	4	5以上
延べ出現回数	351,879	659,985	177,918	163,518	126,506
異なり語数	1,157	11,724	30,584	57,136	88,020
1回のみ出現	134	4,288	17,692	38,369	75,330

漢字列の長さ	パターンと出現比率	
2	IK_1K_2F	1.0
3	IPK_1K_2F	0.5
3	IK_1K_2SF	0.5
4	$IPPK_1K_2F$	0.1
4	$IK_1K_2IFK_1K_2F$	0.1
4	$IPK\ K_2SF$	0.7
4	IK_1K_2SF	0.1

図 8 完全パスのパターンと出現比率
Fig. 8 Patterns and probability of complete paths.

都市	+5.145E-01	本宮	+2.611E-04
都下	+5.164E-03	本人	+1.818E-03
都電	+5.164E-03	本部	+2.769E-03
都内	+1.549E-02	本義	+2.611E-04
都合	+3.460E-01	本生	+2.611E-04
都会	+2.582E-02	本分	+2.611E-04
都政	+5.164E-03	本造	+2.611E-04
都立	+5.164E-03	本気	+2.611E-04
都度	+5.164E-03	本会	+2.611E-04
都入	+5.164E-03	本文	+7.035E-01
都制	+5.164E-03	本場	+2.611E-04
都心	+2.065E-02	本立	+2.611E-04
都民	+5.164E-03	本屋	+2.611E-04
都島	+5.164E-03	本名	+2.611E-04
都議	+5.164E-03	本法	+7.860E-02
都税	+5.164E-03	本物	+1.305E-03
都落	+5.164E-03	本当	+4.700E-03
都営	+5.164E-03	本社	+2.089E-03
都側	+5.164E-03	本性	+2.611E-04
都庁	+5.164E-03	本来	+1.279E-02

図 9 語基 $K1\ K2$ の状態遷移確率表
Fig. 9 State transition probability for primitive words $K1\ K2$.

い漢字複合語と認められないものの多くを取り除くことができる.

3) 実験のために要する計算量をなるべく効果的に減少させる.

といったことがあげられる.

さらに**図 8** に示すように長さ 2, 3, 4 の漢字複合語は, 一般的な分割パターンの比率が知られており[4], これを参考にして図 8 のような完全パスとその出現比率を用いた.

繰返し計算を終えたときの状態遷移確率表の一部を**図 9** に示す. 繰返し計算を 4, 5 回行うと, 確率値はほぼ収束した.

[漢字複合語の分割結果と検討]

JICST の文献抄録から無作為に抽出した長さ 3 以上の延べ約 2,500 語の漢字複合語のサンプル数種に対し分割アルゴリズムを適用した. 分割例は**図 10** のようになった. また, 文字長ごとの平均分割精度は**表 2** のようになった. 全体の平均分割精度は約 95% である. これは次章で述べる改良の結果, 約 97% にまで向上した.

一様振幅周波数特性	121212S12	-1.27093E+01
一端子周波数依存性	12S12S12S	-1.43907E+01
一般化方程式誤差法	12S12S12S	-1.24734E+01
一般化相互相関器法	12S1212SS	-1.22991E+01
一般化可到連性空間	12SP12S12	-1.45771E+01
一般化優度比統計量	12S12S12S	-1.64295E+01
一般化反作用原理及	12S12S12S	-1.53643E+01
一般化逆固有値問題	12SS12S12	-1.43205E+01
一般産業用電気機器	1212S12SS	-1.31461E+01
一般産業用電気設備	1212S1212	-1.38083E+01
一般的多領域信頼性	12SP1212S	-1.28993E+01
一般的感度解析問題	12S121212	-1.09410E+01
一般料金改正用料金	121212S12	-1.70158E+01
一般形伝達関数実現	12S121212	-1.27132E+01
一階連立微分方程式	121212125	-1.83198E+01

1 は K_1, 2 は K_2 を示す. I と F は省略している. 生起確率の値は 10 を底とする対数をとっている.

図 10 漢字複合語の分割例
Fig. 10 Decomposition of Kanji compound words.

表 2 文字長と平均分割精度
Table 2 Average accuracy of Kanji compound word decomposition according to the length of words.

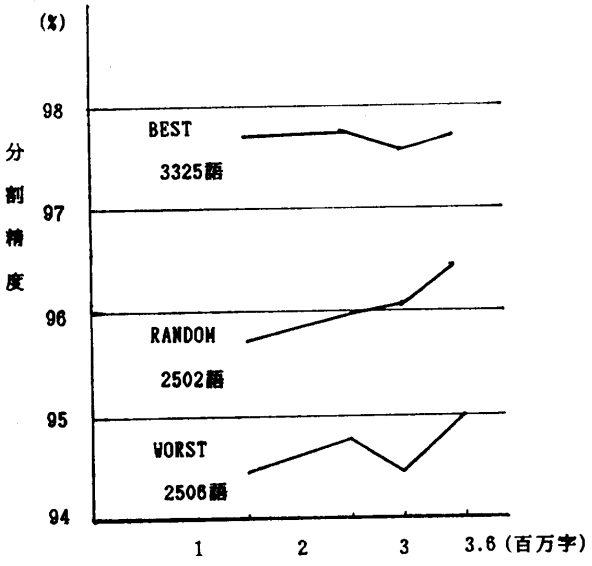
文字数	3	4	5	6
平均分割精度	99.04	95.38	95.02	89.65

文字数	7	8	9	10
平均分割精度	91.17	88.00	86.09	83.19

ここで正解としたものは, 2 文字の語基と接辞が正しく認識されているものである. 1 文字の語基が 2 文字の語基と結合しているものは, 我々のモデルではP あるいはSのいずれかとして認識され, この係りかたが正しいものも正解とした. 数詞はPとして認識される.

表 3 が示すように, 分割精度はトレーニングデータの量が増加するにつれほぼ向上している. トレーニングデータから抽出された BEST というサンプルは, それらのデータが短単位モデルの確率計算にちょうど反映されていることより, 本手法で達成しうる最良の分割精度を近似していると考えられる. RANDOM というサンプルは無作為抽出により平均分割精度を調べるのに用いたものの一つである. WORST というサンプルは, トレーニングデータを取り出すために使用した JICST のテープ 22 巻以外のテープから抽出したもので, 本方法の分割精度の下限を近似するもの

表 3 トレーニングデータ量と分割精度
Table 3 Amount of training data and accuracy of decomposition.



と考えている．

次に分割誤りのパターンについて検討する．誤りの分類には長さ2から8のデータをそれぞれ無作為に500個抽出して用いた．

誤りは大別して次のものがある．

1) $K1\ K2$ とすべきところを PP あるいは SS としたものおよびその逆
濃厚溶液 ($P\ P\ K1\ K2$)
均一線路 ($K1\ K2\ S\ S$)
相対論的電子 ($K1\ K2\ K1\ K2\ K1\ K2$)

2) $P\ K1\ K2$ とすべきところを $K1\ K2\ S$ としたものおよびその逆
標本共分散行列($K1\ K2\ S\ K1\ K2\ K1\ K2$)
速算法 ($P\ K1\ K2$)

3) …SP… と …SS… との誤り
給水用深井戸 ($K1\ K2\ S\ S\ K1\ K2$)

4) $K1\ K2\ P\ K1\ K2$ と $K1\ K2\ S\ K1\ K2$ との誤り
前頭葉下部 ($K1\ K2\ P\ K1\ K2$)

5) 固有名詞，3文字以上の語基
大阪市地下鉄 ($P\ K1\ K2\ K1\ K2\ S$)
国際度量衡検定所 ($K1\ K2\ S\ K1\ K2\ K1\ K2\ S$)

6) 分割できなかったもの，漢字以外の文字と結合するもの
前照灯(＊＊＊)
用材料開発上考慮 ($P\ K1\ K2\ K1\ K2\ S\ K1\ K2$)

7) 短単位モデルでは扱えないもの
米国対日本 ($K1\ K2\ S\ K1\ K2$)
三台 ($K1\ K2$)

このうち特に目立ったのが $K1\ K2$ と SS との誤りだった．分割できなかったもののほとんどはトレーニングデータに現れなかった語基を含んでいた．PP や SS と判断して誤ったものの多くも同様であった．これらはトレーニングデータの量をもっと多くすることで解決できる．またカナ，英字，アラビア数字等を語基や数詞として扱うことでおかしな漢字連続を複合語として処理する可能性を大きく減少させられる．4) の失敗例は12件で，曖昧な分割に対して誤ることはそう多くなかった．3文字以上の固有名詞や語基，『対』，および1文字の語基が接辞とともに現れるものは現在の短単位モデルの限界と考えられ，今後解決すべき課題である．また，固有名詞を一般名詞と区別できることが望ましいが，これを現在の手法で扱うのは困難であると考えられる．

[基本語辞書の作成]

分割アルゴリズムは短単位モデルを複合語の認識系として用いるが，短単位モデルより順次

結果 +8.858E-03	利用 +4.593E-03	適用 +3.097E-03
方法 +8.821E-03	構造 +4.510E-03	動作 +3.092E-03
計算 +8.028E-03	記述 +4.330E-03	評価 +3.006E-03
測定 +7.833E-03	時間 +4.264E-03	関数 +3.002E-03
特性 +7.289E-03	比較 +4.176E-03	考察 +2.941E-03
制御 +7.107E-03	研究 +4.131E-03	効果 +2.935E-03
場合 +7.031E-03	信号 +4.044E-03	電圧 +2.928E-03
開発 +6.049E-03	方式 +3.885E-03	理論 +2.917E-03
設計 +6.006E-03	紹介 +3.836E-03	試験 +2.862E-03
問題 +5.917E-03	技術 +3.789E-03	機能 +2.818E-03
検討 +5.898E-03	影響 +3.773E-03	温度 +2.682E-03
解析 +5.828E-03	処理 +3.648E-03	実現 +2.676E-03
回路 +5.558E-03	条件 +3.493E-03	効率 +2.617E-03
可能 +5.415E-03	提案 +3.408E-03	電力 +2.553E-03
実験 +5.319E-03	変換 +3.352E-03	最適 +2.518E-03
構成 +5.249E-03	出力 +3.338E-03	決定 +2.517E-03
使用 +5.153E-03	応用 +3.336E-03	性能 +2.460E-03
必要 +5.004E-03	関係 +3.219E-03	報告 +2.435E-03
装置 +4.883E-03	変化 +3.130E-03	発生 +2.434E-03
説明 +4.793E-03	電流 +3.129E-03	従来 +2.409E-03

図 11 基本語辞書
Fig. 11 2-Character primitive word dictionary with probability.

$I \rightarrow K\,1/b\,1 \rightarrow K\,2/b\,2 \rightarrow F/\phi$

の状態遷移系列を取り出し，その生起確率の高い順に並べれば，頻度順の基本語辞書を得る．これを**図11**に示す．

6. 短単位分割法の改良および辞書の利用

分割アルゴリズムの失敗例をもとに，推定アルゴリズムの初期確率の設定法を再検討した．初期確率設定以降のプロセスは，与えられた初期確率にのみ依存するから，初期確率の設定時に次のような工夫[3]を行った．

1）いくつかの頻出語に対する正解を与える．

2）明らかに現れえないと思われる完全パスの生成を禁止する．

3）接辞として用いられる漢字を指定する．

1)の正解の付与は，長さ3，4のトレーニングデータに対して試みた．この結果前章のサンプル群に対して最高 1.4%，平均 0.6% の分割精度の改良をみた．正解を付与することの効果は，前章のパターンの重みをかえることよりも，誤った語基の生成を防ぐという点で重要であった．

本手法では各トレーニングデータに対してすべての可能な完全パスのパターンを生成してしまうため，冗長な状態を生じやすい．したがって，2)については接頭辞や接尾辞が連続して3個以上現れないという制限を設けた．これによりトレーニングデータの長さを4以上にして，できるだけ多くのデータを処理するようにしても完全パスの数が急激に増加することがなくなる．

3)については，あらかじめ接頭辞や接尾辞となる漢字のリストを用意した．(ただし1文字で語基となる可能性のある一般語もリストに加えている.) このリスト中にない漢字 k が P/k や S/k となる状態を含む完全パスの生成を禁止した．

これらの情報は極めて容易に付加できるものである．さらに同様に考えて次のような方法を試みた．

4）トレーニングデータから2文字のデータをすべて取り出してテーブルを作成しておく．3文字以上のトレーニングデータから完全パスを生成するときには，その完全パス中の $K\,1/k$-$K\,2/k'$ に対応する語基 kk' がすべてテーブルに含まれるもののみ，その生成を許す．こうして生成できるパスの数が0個になったものは，すべての完全パスを生成させる．

これは文献4)で3文字の漢字複合語を分割するときに使われた方法を完全パスに適用したものである．実際に用いたトレーニングデータ（長さが3，4のもの，異なり語で約3万語）のなかで許される完全パスの数が0になったものは約540語に過ぎなかった．**表4**に上記の方法の適用による平均分割精度の変化を示す．

表4 各種の方法による平均分割精度の変化
Table 4 Improvements of decomposition.

計算法	平均分割精度(%)
正解の付与	96.3
正解の付与＋フィルタ	96.8
テーブル検索	96.6
テーブル検索＋正解の付与	97.3

二重窓	+3.00456E-11	一次巻線	+2.99495E-11
1 2 K		N K 1 2	
三相用	+1.63854E-10	一致精度	+3.90180E-15
N P K		1 2 1 2	
十分早	+2.35362E-14	二重拡散	+8.26694E-11
1 2 K		N K 1 2	
東京湾	+1.35656E-11	三相系統	+1.76373E-10
1 2 K		N P 1 2	
大企業	+2.09067E-07	十分配慮	+1.21587E-13
P 1 2		N K 1 2	
中空系	+9.17583E-11	市外区間	+2.15373E-12
1 2 K		1 2 1 2	
小粒子	+3.02443E-07	大地表面	+7.59520E-11
P 1 2		P 1 2 K	

N は数詞を示す，K は1文字の語基，1と2は2文字語基の1文字目と2文字目を示す．I と F は表示していない．

図12 辞書を用いた完全パスの生成
Fig. 12 Generation of complete paths using a dictionary.

既存の辞書がある場合は，上のテーブルをそのまま辞書と置き換えればよい．語基の学習効果は失われるが，これにより正しい語基，接辞のみからなる短単位モデルを得る．**図12**にこの方法で生成する完全パスの例を示す．

7. まとめと今後の課題

短単位モデルと統計的推定法による漢字複合語の分割手法について述べた．現在は一般漢字複合語のみしか扱っていないが，本手法は今後の改良や辞書の利用によって，広汎な複合語を高精度で短単位に分割できると予想している．本手法は正しい短単位を自動的に学習できるという大きな特長があるが，短単位モデルに含まれるものは漢字1文字の接辞と漢字2文字の語基に限られている．したがって，前章の最後に示したようにこの学習機能を犠牲にして，既存の辞書から接

頭辞P，接尾辞S，語基Wに相当する語を既知とすることで，漢字2文字以外の語基，カタカナ語や固有名詞を含めたすべての短単位に我々のモデルを拡張できる．

謝辞 短単位モデルの検討や実験システムの実現，分割結果の検討，分割失敗例の分類等に御協力いただいた鈴木恵美子さん，沼尾雅之氏，西野哲朗氏（現在東京電機大学）に深謝いたします．

参考文献

1) 鈴木，武田，沼尾，藤崎：統計的手法による漢字列の短単位分割，第29回情報処理学会全国大会論文集，4J-1 (1984).
2) 鈴木，武田，沼尾，藤崎：統計的手法による基本漢字辞書作成法，第29回情報処理学会全国大会論文集，4J-2 (1984).
3) 武田，鈴木，藤崎：漢字複合語自動分割の一手法，第30回情報処理学会全国大会論文集，5G-6 (1985).
4) 長尾，辻井，山上，建部：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, pp. 514-521 (1978).
5) 中野，野村：日本語の形態素分析，情報処理，Vol. 20, No. 10, pp. 857-872 (1979).
6) 西野，藤崎：漢字複合語の確率的構造解析の試み，第34回情報処理学会全国大会論文集，1W-8 (1987).
7) 野村，森：漢字かな変換システムの試作，信学論，Vol. J66-D, No. 7, pp. 789-795 (1983).
8) 藤崎：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，情報処理学会自然言語処理研究会資料，28-5 (1981).
9) 藤崎：自然言語の曖昧さの取扱いの研究，東京大学大学院情報工学科博士論文 (1985).
10) 松延，日高，吉田：日本語確率文法における書き換え規則の確率の推定について，情報処理学会自然言語処理研究会資料，55-4 (1986).
11) 宮崎，大山：日本文音声出力のための言語処理方式，情報処理学会論文誌，Vol. 27, No. 11, pp. 1053-1061 (1986).
12) 宮崎：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol. 25, No. 6, pp. 970-979 (1984).
13) 諸橋：自動索引付け研究の動向，情報処理，Vol. 25, No. 9, pp. 918-925 (1984).
14) Bahl, L. R., Jelinek, F. and Mercer, R. L.: A Maximum Likelihood Approach to Continuous Speech Recognition, *IEEE Trans. PAMI*, Vol. PAMI-5, No. 2, pp. 179-190 (1983).
15) Baum, L. E.: An Inequality and Associated Maximumization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes, *Inequalities*, Vol. III, pp. 1-8 (1972).
16) Fujisaki, T.: A Stochastic Approach to Sentence Parsing, *Proc. of Computational Linguistics*, pp. 16-19 (1984).
17) Forney, G. D., Jr.: The Viterbi Algorithm, *Proc. of IEEE*, Vol. 61, pp. 268-278 (1973).
18) Shanon, C. E.: Prediction and Entropy of Printed English, *Bell Syst. Tech. J.*, Vol. 30, pp. 50-64 (1951).
19) Takeda, K., Fujisaki, T. and Suzuki, E.: CRITAC—A Japanese Text Proofreading System, *Proc. of Computational Linguistics*, pp. 412-417 (1986).

(昭和62年3月17日受付)
(昭和62年6月11日採録)

武田 浩一（正会員）

昭和33年生．昭和56年京都大学工学部情報工学科卒業．昭和58年同大学院修士課程修了．同年日本アイ・ビー・エム(株)に入社．同社東京基礎研究所にて自然言語処理，文書データベースの研究に従事．現在カーネギー・メロン大学機械翻訳センター客員研究員．電子情報通信学会，日本ソフトウェア科学会，ACM 各会員．

藤崎哲之助（正会員）

昭和22年生．昭和44年東京大学工学部計数工学科卒業．昭和46年同大学院修士課程修了．同年日本アイ・ビー・エム(株)に入社．同社東京基礎研究所にて自然言語処理，人工知能の研究に従事．現在 IBM トーマス・J・ワトソン研究所にてパターン認識の研究に従事．工学博士．電子情報通信学会，日本ソフトウェア科学会，計量国語学会，ACM 各会員．