自然言語処理 48－2
（1985. 3. 27）

統計的手法を用いた漢字複合語の短単位分割

武田 浩一　　藤崎 哲之助

日本アイ・ビー・エム株式会社　サイエンス・インスティチュート

1．まえがき

日本語処理における複合語の分割は、機械翻訳[1]、自動インデクシング[2]、音声合成[3]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[4]等の手法では十分な分割精度を得ることができなかった。

最近文献[5]において提案されている係り受け解析法は、広汎な複合語を９９．８％という高精度で分割できるが、人手により意味情報を含めた辞書項目を充実させる作業は一般にコストが高く、適用分野が変化した際に未知語の扱いが問題となる。

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。ここで用いた統計的推定法は、英語の連続音声認識[6]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[7]において、その有効性が知られている。

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。[8]
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。[9]

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９７％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場合についても実験を行なった。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、マルコフモデルの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。

従って、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により短期間にかなり高精度の複合語分割システムの構築が期待できる。

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は、２文字の語基と１文字の接頭辞、接尾辞から構成されていることが知られている。[10]即ちこのような複合語は

（接頭辞）＊　語基　（接尾辞）＊

という線形表現の短単位の連接として表わされる。（ここで＊は０回以上の繰返しを示す）。複合語分割の曖昧さは、例えば

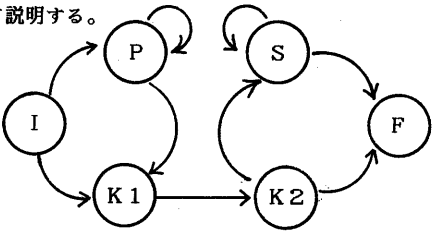
太陽熱発電

のような複合語に対し、

1)　太陽・熱　　発電
2)　太陽　　熱・発電

といった複数の短単位の連接パターンが存在することにより生じる。これらのパターンから正しい解を求めるために、語基や接辞の使用回数に基づく頻度情報がよく用いられる。

我々の手法はこの考えを発展させ、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と呼ぶ。[8] 応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なう必要がある。以下ではその方法について説明する。



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

図１にこの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文字目を表わしている。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（ひと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、

太陽、　太陽熱、　発電、　熱発電

という語が現れたと仮定した時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といったラベルと漢字１文字の対からなり、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／øのときには、短単位の区切りを示す特別な記号øを出力するが、これは観測されない記号であるとする。（フィルターにより出力系列からø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／ø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／øとＩを同一状態とみなすことに等しい。

次節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。これを図３に示す。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。

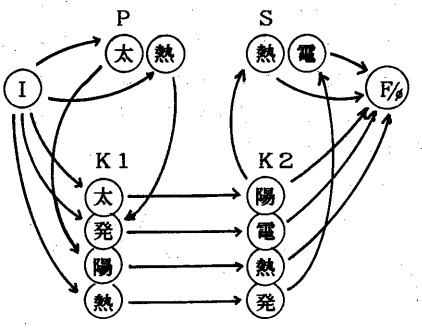


図２　展開された短単位モデル

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／ø，
　　Ｋ１／発，Ｋ２／電，Ｆ／ø

（ｂ）Ｉ，Ｋ１／太，Ｋ２／陽，Ｓ／熱，Ｆ／ø，
　　Ｋ１／発，Ｋ２／電，Ｆ／ø

（ｃ）Ｉ，Ｋ１／太，Ｋ２／陽，Ｆ／ø，
　　Ｐ／熱，Ｋ１／発，Ｋ２／電，Ｆ／ø

（ｄ）Ｉ，Ｋ１／太，Ｋ２／陽，Ｆ／ø，
　　Ｋ１／熱，Ｋ２／発，Ｓ／電，Ｆ／ø

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。　従って、『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。

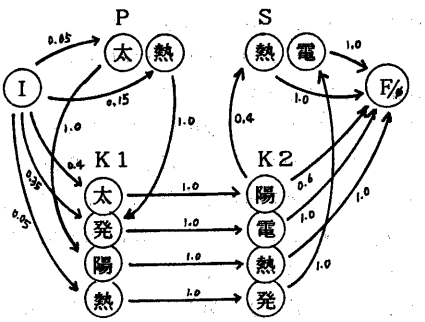


図３　図２に確率を与えた短単位モデル

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。従って、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／ø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、
　$L(t) = s_i$ ：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 $t_{1,n}$
　$t_{1,n} = t_1, t_2, \ldots, t_n$
　$(L(t_1) = I,\ R(t_n) = F/ø)$
　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

系列を表わしている。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、$A(t_1),\ldots,A(t_n)$ となる。漢字複合語は $a_{1,n}$ より $A(t_j)=o$ となる出力を取り除いた観測系列 $b_{1,m}=b_1,\ \ldots,\ b_m$ である。図4にこれらの表記を簡単にまとめている。

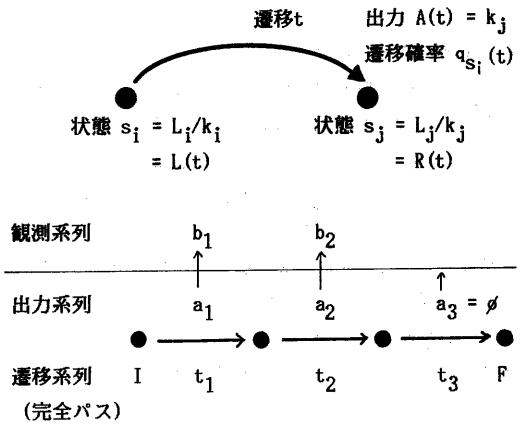


図4　短単位モデルの表記

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[9]（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2つのステップに大別される。

前節の図2から図3のような展開を行なうために、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なう。

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1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が一回出現したとすれば、これらのパスのうち正しいものを一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。従って、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節で述べるように、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なう。

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。（やや人為的なトレーニングデータであるが、マクロなふるまいを少数のデータで理解できるようにしている。）ここで（ ）内は出現回数を示す。このとき $G_1,\ \ldots,\ G_6$ は次のようになる。

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

g_{11} = (I,K1/太),(K1/太,K2/陽),(K2/陽,F/ø)

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

g_{21} = (I,K1/発),(K1/発,K2/電),(K2/電,F/ø)

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

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

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

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

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

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

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

g_{51} = (I,K1/陽),(K1/陽,K2/気),(K2/気,F/ø)

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

g_{61} = (I,K1/熱),(K1/熱,K2/意),(K2/意,F/ø)

従って初期値の設定は例えば t = (I,K1/太) なら

$$q_I(t) = \frac{1*100/1 + 1*20/2}{\begin{array}{l}1*100/1 + 1*50/1 + 1*20/2 \\ + 1*20/2 + 1*10/2 + 1*10/2 \\ + 1*10/1 + 1*10/1\end{array}} = 11/20$$

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を得る。

$f(g_{11}) = f(g_{21}) = f(g_{51}) = f(g_{61}) = 1.0$

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

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

（以下省略）

よって同様に t = (I,K1/太) を求めると

$q_I(t) = 17/30$

となる。この結果を図６に示す。

繰返し計算を何度も行なうことにより、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[6]と等価であり、同アルゴリズムのもつ次の数学的性質[11]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_xの生起確率を$p(G_x,M) = p(g_{x1},M)+...+p(g_{xn},M)$とし、トレーニングデータ集合Tの生起確率を $p(T,M) = p(G_1,M) + ... + p(G_x,M)$とする。（ここで$p(g_{xy},M)$は上例の$p(g_{xy})$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、

$$p(T,M) \leq p(T,M')$$

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。

従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

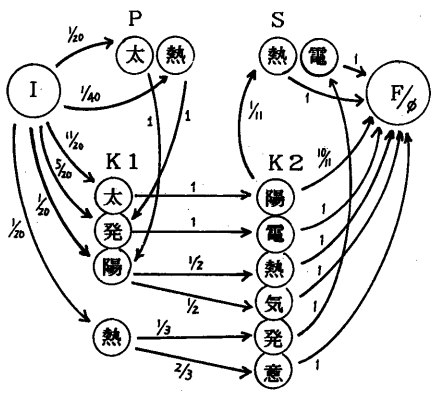


図５　初期確率の設定

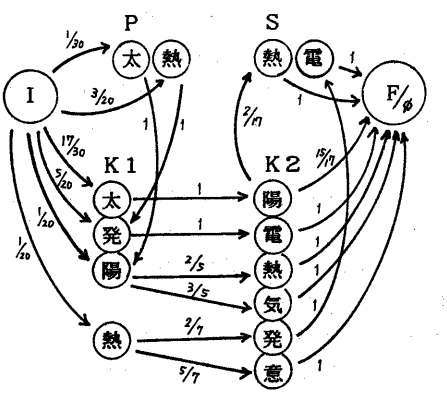


図６　繰返し計算による確率の更新

要である。これについては５節で述べる。

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[11]の議論とトレリス[12]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行なえる。

４．漢字複合語分割アルゴリズム

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。これは動的計画法あるいはマルコフモデル上のViterbiアルゴリズム[13]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

$$『太陽熱発電』 = b_1 b_2 b_3 b_4 b_5$$

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。

$STATE^1$ = < P/太, K1/太 >

$PRED^1$(P/太)=<(I,P/太)>, p($PRED^1$(P/太))= 1/20

$PRED^1$(K1/太)=<(I,K1/太)>, p($PRED^1$(K1/太))= 11/20

同様に$STATE^2$および$PRED^2$は次のようになる。

$STATE^2$ = < K1/陽, K2/陽 >

$PRED^2$(K1/陽) = <(I,P/太),(P/太,K1/陽)>,
p($PRED^2$(K1/陽))= 1/20

$PRED^2$(K2/陽) = <(I,K1/太),(K1/太,K2/陽)>,
p($PRED^1$(K1/陽))= 11/20

$STATE^3$ではF/øを経由する遷移

K2/陽 - F/ø(=I) - P/熱

K2/陽 - F/ø(=I) - K1/熱

があり、

$STATE^3$ = <S/熱, K2/熱, P/熱, K1/熱>

$PRED^3$(S/熱) = <(I,K1/太),(K1/太,K2/陽),
(K2/陽,S/熱)>,
p($PRED^3$(S/熱)) = 1/20

$PRED^3$(K2/熱) = <(I,P/太),(P/太,K1/陽),
(K1/陽,K2/熱)>,
p($PRED^3$(K2/熱)) = 1/40

$PRED^3$(P/熱) = <(I,K1/太),(K1/太,K2/陽),
(K2/陽,F/ø),(F/ø,P/熱)>,
p($PRED^3$(P/熱)) = 1/80

$PRED^3$(K1/熱) = <(I,K1/太),(K1/太,K2/陽),
(K2/陽,F/ø),(F/ø,K1/熱)>,
p($PRED^3$(K1/熱)) = 1/80

これを$STATE^6$ = <F/ø>まで行なうと、

$PRED^6$(F/ø) = <(I,K1/太),(K1/太,K2/陽),
(K2/陽,S/熱),(S/熱,F/ø),
(F/ø,K1/発),(K1/発,K2/電)>

および t = (K2/電,F/ø) なる解をえる。

ここで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[14]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節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［13］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。

５．実験システム

前節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

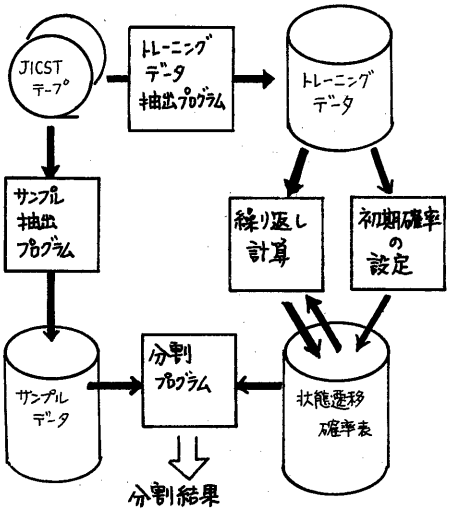


図７　実験システムの構成

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、漢字複合語切り出しプログラムにより、トレーニングデータを機械的に抽出した。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。このようなサブセットを用いた理由は

１）漢字複合語の特徴は長さ２，３，４のもので大体近似できる。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。

文字数	1	2	3	4	5以上
延べ出現回数	351879	659985	177918	163518	126506
異なり語数	1157	11724	30584	57136	88020
１回のみ出現	134	4288	17692	38369	75330

表１　漢字複合語の延べ出現回数と異なり語数

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[1]、これを初期確率の設定にした。

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回行なうと、確率値はほぼ収束した。

長さ	パターンと出現比率	
2	IK_1K_2F	1.0
3	IPK_1K_2F	0.5
	IK_1K_2SF	0.5
4	$IPPK_1K_2F$	0.1
	IPK_1K_2SF	0.1
	$IK_1K_2FIK_1K_2F$	0.7
	IK_1K_2SSF	0.1

図８　完全パスのパターンと出現比率

[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２５００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。これは次節で延べる改良の結果、約９７%にまで向上した。

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。

太平	+4.743E-03	分周	+3.057E-03	気候	+3.102E-02
太陽	+9.952E-01	分科	+7.411E-04	成長	+2.991E-01
富士	+1.000E+00	分母	+3.149E-03	成金	+4.641E-07
前年	+1.136E-02	分離	+7.196E-02	成人	+7.197E-03
前部	+1.747E-03	分担	+3.376E-03	成行	+6.991E-09
前方	+6.021E-02	分散	+1.324E-01	成信	+7.157E-04
前後	+9.557E-02	分析	+1.460E-01	成分	+2.597E-01
前進	+2.010E-02	分裂	+3.243E-03	成文	+1.439E-07

図９　語基K1K2の状態遷移確率表

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	12121212S	-1.83198E+01
一週間以上連続動作	P12121212	-1.74522E+01
一週間連続運転可能	P12121212	-1.72723E+01
一軸性誘電体導波路	12S12S12S	-1.64958E+01
一軸異方性媒質表面	1212S1212	-1.58473E+01

1は K1，2は K2 を示す
Iと下は省略している

（正起確率の10の対数をとったもの）

図10　漢字複合語の分割例

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

表２　文字長と平均分割精度

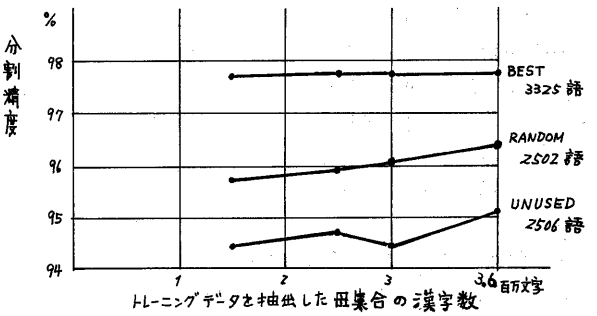


表３　トレーニングデータ量と分割精度

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。UNUSED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）K1K2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）PK1K2 とすべきところを K1K2S 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）K1K2PK1K2 と K1K2SK1K2 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K1K2 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あらわ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

I - K1/b1 - K2/b2 - F/ø

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。

結果	1.0521E-02	利用	5.3991E-03
方法	1.0433E-02	構造	5.2915E-03
計算	9.3953E-03	記述	5.0812E-03
測定	9.1419E-03	比較	4.9569E-03
特性	8.3925E-03	時間	4.9319E-03
制御	8.3182E-03	研究	4.8452E-03
場合	8.2705E-03	信号	4.7382E-03
開発	7.0966E-03	方式	4.5625E-03

図11　基本語辞書

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[15]を行なった。

1）いくつかの頻出語に対する正解を与える

2）明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。この結果前節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、前節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。従って、(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K1/k-K2/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［１］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。

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

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、広汎な漢字複合語を高精度で短単位に分割できると予想している。

謝辞

短単位モデルの検討や実験システムの実現、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 哲朗氏に深謝いたします。

計算法	平均分割精度（%）
正解の付与	９６．３
正解の付与＋フィルター	９６．８
テーブル検索	９６．６
テーブル検索＋正解の付与	９７．３

表４　各種の方法による平均分割精度の変化

超小形化	+5.15229E-16	静的解析	+5.16855E-09
ＰＰＫＳ		ＰＫ１２	
超小形化	+4.76898E-14	静的解析	+8.99210E-13
Ｐ１２Ｋ		ＫＰ１２	
超小形化	+4.01644E-14	静的解析	+1.01313E-10
Ｐ１２Ｓ		ＫＳ１２	
既設三相	+2.28259E-13	静的解析	+9.23266E-14
１２ＮＫ		１２１２	
劣化原因	+1.23365E-12	直列制御	+4.82769E-18
ＫＰ１２		ＫＰ１２	
劣化原因	+5.06863E-17	直列制御	+2.40215E-16
ＫＳ１２		ＫＳ１２	
劣化原因	+1.09179E-11	直列制御	+1.28232E-10
１２１２		１２１２	

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

参考文献

[1] 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

[2] 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

[3] Miyazaki,Goto,Ooyama and Shirai:Linguistic Processing in a Japanese-Text-to-Speech System, Proc. of ICTP,pp.315-320,1983

[4] 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

[5] 宮崎：係り受け解析を用いた複合語の自動分割法、情処論、Vol.25,No.6, pp.970-979,1984

[6] Bahl and Jelinek:A Maximum Likelihood Approach to Continuous Speech Recognition,IEEE Trans. on PAMI,Vol.PAMI-5,No.2,pp.179-190,1983

[7] Fujisaki:A Stochastic Approach to Sentence Parsing,Proc. of COLING, pp.16-19, 1984

[8] 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、59年後期情処全大、4J-1,1984

[9] 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、59年後期情処全大、4J-2,1984

[10] 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

[11] Baum:An Inequality and Associated Maximumization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes, Inequalities,Vol.III,pp.1-8,1972

[12] 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、自然言語処理研究会、28-5,1981

[13] Forney,Jr.:The Viterbi Algorithm,Proc. of IEEE,Vol.61,pp.268-278,1973

[14] Shanon:Prediction and Entropy of Printed English,Bell Syst. Tech. J., Vol.30,pp.50-64,1951

[15] 武田、鈴木、藤崎：漢字複合語自動分割の一手法、60年前期情処全大、5G-6,1985