自 然 言 語 処 理 139－5
（2000．9．21）

接続確率最小法による教師なし単語分割

飯塚 泰樹

松下電器産業株式会社
マルチメディアシステム研究所
iizuka@trl.mei.co.jp

日本語の処理において、単語への分割は解析処理の最初の課題である。近年はタグ付コーパスの利用により、精度の高い単語分割が実現可能となっている。しかし様々な文書分野へシステムを適用することを考えると、必ずしも常に適したタグ付きコーパスが得られるとは限らない。本論文ではまず、文字間の接続確率を用いることでタグ無しコーパスからの学習により文章を単語へ分割する手法を提案し、その効果を実験により検証する。
さらにこれを応用して、複合語分解のための接続確率最小法を提案し有効であることを示す。

Japanese Word Segmentation by Minimum Connective Probability Method

Yasuki IIZUKA

Multimedia Systems Research Laboratory
Matsushita Electric Industrial Co., Ltd.
iizuka@trl.mei.co.jp

This paper proposes word a segmentation method based on the connection probability between characters. The connection probability are calculated based on plain text data in the application domain. This means that this method does not presume the existence of pre-segmented tagged corpus which most of currently proposed methods need for training.
This paper also proposes a two-step segmentation method applying the above method over the text data which deivided into word candidates by rules.
Effectiveness of these methods are shown by the experimental results of segmenting news paper articles.

1 はじめに

膠着語である日本語の処理において、単語の分割は解析処理の最初の課題となる。例えばテキスト・データベース・システムにおいては、検索インデックス作成のために文章を単語へ分割することが一般的に行われている。

通常、単語分割には辞書を利用した形態素解析処理 [9] が用いられる。形態素解析は、大規模な辞書が利用可能になったこと、未知語の発見／推定技術が発達したことなどにより高い精度の解析が実現されている。しかし形態素解析は辞書の整備や接続コストの整備などが必要であり、多くの場合これらの整備は人手に頼っていた。これに対して近年、頑健な解析を目指して単語や文法の自動獲得 [3]、接続コストの自動学習 [10] が提案されている。

特にタグ付コーパスからの統計的言語モデルの獲得は、高い精度の形態素解析を実現することに成功した [9]。しかしタグ付きコーパスを用いた学習やパラメータ獲得などは、アプリケーション・プログラムの対象文書と同じ分野[1]のタグ付きコーパスが大量に必要であり、コーパスの用意に大きなコストがかかる。システムを様々な文書分野へ適用する (あるいは適応させる) ことを考えると、必ずしも常に適切なタグ付コーパスによる学習が期待できるわけではない。

本研究はこのような背景を踏まえ、辞書整備やタグ付コーパスの用意などのコストを0にすることを目標とするものであり、全文検索システムのための辞書なし教師なし単語分割手法を提案する。

本論文ではこの方式について、2章で課題について、3章において文字接続確率のみによる単語分割方式の実験と評価、4章において文字接続確率を応用した接続確率最小法による複合語分割方式の実験と評価について述べる。

2 課題

我々は辞書整備コスト0を目標として、ルールによる辞書無し単語分割方式 [1](以降ではRule方式と呼ぶ) を提案した。これは、ルールベースで単語抽出を行って自動的に単語リスト (辞書相当のもの) を作成し、作成した単語リストと分割時ルールを相補的に使うことで文章を単語に分割するものであった。しかし図1に示すような複合語の分解では正解を選択する指針が無かった。もちろんルールを拡張して図1が学校の名前を表現していることを規定するようにしてもよいが、際限ないルールの拡張は辞書のメンテナンスよりもコストがかかるだろう。

(1) 大阪-市立-山田-中学校
(2) 大阪市-立山-田中-学校

図 1: 複合語の分解の複数の候補

一般にこのような問題は統計的言語モデルを採用することで解決可能である [6]。統計的言語モデルにより単語分割問題は (1) で表現できる。$P(W)$ は通常、単語 N 個組である N-gram モデルで近似する。N-gram モデルの利点は、タグ付コーパスから自動学習出来ることであり、一般的には 2-gram と 3-gram が採用されることが多い。3-gram の場合は (2) 式のようになる。

$$\hat{W} = \arg\max_{W} P(W) \quad (1)$$

$$P(W) = \prod_{i=1}^{m} P(w_i|w_{i-1}w_{i-2}) \quad (2)$$

しかし実用に際して、システムを様々な文書分野へ適用させることを考えると、必ずしも常にタグ付コーパスによる学習が期待できるわけではない。例えば、新聞記事で学習した結果を医療文書に適用した場合の信頼度は低いものとなるであろう。

この問題は、精度の高い単語分割プログラムを作成するためには単語分割された大量のコーパスが必要であり、単語分割された大量のコーパスを作成するためには精度の高い単語分割プログラムが必要という鶏と卵の関係に帰着される。

[1] 新聞一般、工学、医学、経済などのこと

3 文字間接続確率による単語分割

3.1 原理

前章の課題は、言語モデルとして単語列のモデルを採用したことに起因する。そこで単語列ではなく、文字列について検討してみる。

文字に注目してみると、一般に単語を構成する文字列は全ての文字の組み合わせの単語が存在するわけではないので、文字の組合せの出現は等確率ではない。

「ある文字列がユニットを構成している場合、その前後には様々な文字が出現する」という仮定は多くの単語抽出の研究で用いられている [2][4][5]。本研究でもこの仮定を採用し、これを文字間の接続確率で考える。すなわち、ある文字列がユニットを構成している場合、その前後には様々な文字が出現するわけであるから、ユニットの文字とその前後の特定の文字との接続確率は低くなるはずである。と同時に、そのユニットの構成文字間の局所的な接続確率は相対的に高くなると考えられる。

形式的に表現するなら、文字列 $C_1...C_n$(以降では C_1^n と表記する) が単語 W_xを構成する時、この単語 W_xの後に任意の文字 C_{n+1}が出現する確率 $P(C_{n+1}|C_1^n)$ は、単語を構成する文字間の接続確率 $P(C_i|C_1^{i-1})$(ただし $1 < i \leq n$) より相対的に低くなると考えられる。

$$Pc(C_i|C_1^{i-1}) > Pc(C_{n+1}|C_1^n)$$

上記確率は、本来なら文字列の長さでの正規化が必要であり [7]、一様な比較は難しい。しかし日本語の単語の分割点を発見するためだけなら、n が一定の文字 N-gram による近似が可能と考えられる。すなわち、着目している点の前後の固定長文字列だけによる比較を行う。

文字間の接続確率は、ある程度の量の文書を集めることができれば、そこから統計的に調べて計算することができる。文書データベースを構築するような状況ならば、単語分割対象であるデータベースに登録する文書自身から、文字間接続確率を統計的に調べて計算することができる。この計算値は日本語全体について調べた場合の確率値とは違うものであろうが、確率値を調べた文書、あるいは類似の文書の分割に適用するのに適した性質を持つため望ましいと考えられる。

文字 N-gram の計算に、計算資源の問題から 2〜4-gram を使うことは、日本語の短単位語の平均文字列長の観点からも望ましいだろう。

よって以上の原理を用いることで、単語分割されていない文書から文字間接続確率を調べ、その文字間接続確率を使うことで、文書を辞書を用いることなく単語単位へと分割することが可能と考えられる。

3.2 計算方法詳細

上記の原理の計算方式を次のように定式化する。

方式 1 (文字間接続確率による単語分割法) 文字間接続確率に *N-gram* 近似を用い、文字 C_{i-1}と文字 C_iの間の確率 $Pc(C_{i-1} \sim C_i)$ として、長さ n の文字列 $C_{i-n}...C_{i-1}$が出現したという条件のもとで長さ m の文字列 $C_i...C_{i+m-1}$が出現する確率として定義する。*(n,m は固定)* この確率が閾値 T以下の部分を分割点とする。

$$Pc(C_{i-1} \sim C_i) = Pc(\underbrace{C_i...C_{i+m-1}}_{m\text{ 個}} | \underbrace{C_{i-n}...C_{i-1}}_{n\text{ 個}}) \tag{3}$$

ところが式 (3) の計算のためには、n+m gram の統計を取る必要がある。「京都」と「東京都」の区別を考えるなら、$n \geq 2$ または $m \geq 2$ が必須である。$n \geq 2$ かつ $m \geq 2$ とすると 4-gram (またはそれ以上) の文字組の統計計算が必要になり、非常に大きな記憶空間を必要とするため実用的ではない。そこで (3) 式の計算を次式で近似することにする。

$$Pcf(C_i|C_{i-n}^{i-1}) \otimes Pcb(C_{i-1}|C_i^{i+m-1}) \tag{4}$$

(4) 式は、n 個の文字列が出現した後に特定の文字が出現する順方向の確率である Pcfと、m 個の文字列が出現する前に特定の文字が出現する逆方向の確率である Pcb の演算結果である。演算⊗は何らかの演算であり、積、相加平均、加重平均などが考えられる。(4) 式の確率値は、Pcfは n+1 グラム、Pcbは m+1 グラムの統計を取れば計算できる。

方式名	計算式 (4) について	結果
前向き単純 3-gram	n=2, m=1	78.3%
3-gram 相加平均	n=2, m=2, $(P(C_i\|C_{i-2}^{i-1}) + P(C_{i-1}\|C_i^{i+1}))/2$	82.9%
3-gram の積	n=2, m=2, $P(C_i\|C_{i-2}^{i-1}) \times P(C_{i-1}\|C_i^{i+1})$	83.5%

表 1: 実験した計算方式

3.3 実験と評価 (1)

前節の原理を検証するため、実験と評価を行なった。計算機の資源の関係から、式 (4) について n=m=2 として 3-gram を計算することが実際的と考えられる。また演算⊗については相加平均と積を計算した。

全体の処理手順を図 2に、今回実験した確率の計算方式を表 1に示す。3-gram 確率の計算は最尤推定を使い、クローズドデータのみを用いて評価する。確率テーブルはスパースであるため、圧縮して記憶する。実験の対象データとして CD-毎日新聞 95 年版データを用いた。分割例を図 3に示す。

1. 分割対象文書から文字間接続確率を最尤推定して確率テーブルに記憶。
2. 分割したい文の各文字位置について、文字間接続確率をテーブルより調べ、
3. その値が設定した閾値 T以下の場合に分割する。(ただし句読点の前後は常に分割するものとする。)

図 2: 処理手順

分割正解として京大コーパス [11] を参照し、これと実験結果を比較する。分割点が正解と同じ場所に出現したかどうかで、再現率と適合率を計算する。本方式では、閾値 Tを変化させることで再現率と適合率は連続的に変化する (図 4)。閾値 Tを高くすれば分割が細かくなり、再現率は 1 に達っするが適合率が悪くなる[2]。そこで閾値 Tを変化させた時の F 値[3]の最大値をとって精度とした。再現率と適合率が交差する付近で F 値も最大に達っする。

表 1の右端に確率計算式による分割精度 (80MByte で学習) の測定結果を示す。表 2に学習データ量よる分割精度 (3-gram の積で計算) の変化を示す。

衆院/の/選挙/制度/も/変わり/、/政治/は/大きな/節目/を迎え/ている/。

図 3: 分割例

学習データ量	確率テーブルの大きさ	分割精度
2MByte	9MByte	81.7 %
20MByte	31MByte	83.0%
80MByte	65MByte	83.5%

表 2: 学習データ量と分割精度の関係

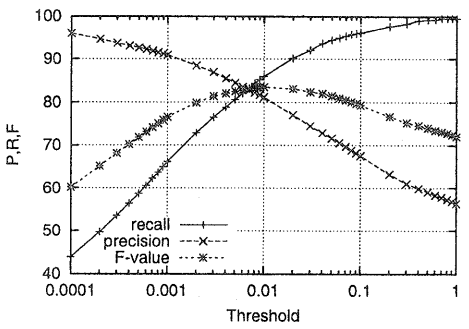

図 4: 閾値と再現率・適合率・F 値の関係 (1)

[2] 句読点で強制的に分割しているため、0 にはならない。
[3] β=1 で固定とする

3.4 考察 (1)

各計算方法による精度は表 1の右端に示した通りだが、前向き単純3-gramによる計算よりも、3-gramによる4-gram近似の方が高い精度を実現できた。

実験結果では、片仮名単語が切れすぎる傾向にあることが判明した。これは片仮名が外来語(あるいは音)を表記した長い単語であるのに、接続確率だけで分割すると漢字単語と同じ平均単語長で切られてしまうからである。本実験では文章全体において閾値Tを固定とし、局所的にTに差を設けることは行なわなかったが、平仮名部分や片仮名部分といった文字種により、閾値Tをヒューリスティックスで補正することは可能かもしれない。しかしどの文字種が何文字続いた時に閾値をどれだけに設定すればよいかといった操作は、人手で行おうとすると非常に難しいだろう。

また実験結果では「名詞+助詞」や「助詞+接頭辞」が一つの単語と認識されることが多かった。これは学習コーパス中にどれだけそのような表現が出現したかを示しているが、これが必ずしも正解とは一致しないことが問題である。

ところで学習データ量を変化させた時の様子は表2にあるが、学習データ量が大きいほど最適な閾値Tは小さくなる傾向にある。これは学習データ量の増加に従いN-gramの組合せが増えるためと考えられる。本実験では閾値Tをパラメータとして変化させているが、実システムを作成するためには最適な閾値を決定する手段が必要であろう。閾値を決定には、閾値を変化させた時の平均単語長に着目することが有効な手段になると予想される。

以上のように文字接続確率だけによる単語分割は、およそ単語らしい部分での分割が可能であるが、精度にあと一歩の課題が残る。

4 接続確率最小法による単語分割

4.1 原理

前章ではモデルとして文字モデルを採用し、局所的な文字間接続確率だけで単語分割が可能かどうかを検証した。これはある文字列の「単語らしさ」については何ら情報を持たないものであった。

一方、Rule方式[1]は単語抽出から出発することで、2章で示したように単語間接続確率の欠如といった課題を持っていた。

そこでこれら2つの方式の融合し、高い精度の単語分割を実現することを検討した。基本方針として、単語分割の枠組にはRule方式を用いることにし、複合語分解や未知語相当の語句周辺の解析には前章の方式を取り入れる。

すなわちRule方式で単語分割解の候補を作成し、その解の候補の選択を文字間接続確率を用いて行なう。原理として次の前提を設ける。

「複数の単語分割候補のうち、文字間接続確率が小さい所で分割されているものが解として尤もらしい。」

これを接続確率最小法を呼ぶことにする。

4.2 計算方法詳細

接続確率最小法の基本形は次の通りである。

方式 2 (接続確率最小法:基本形) 既存の辞書によりある文の単語分割解の候補が複数得られたとする。この候補の集合を$X = \{x_1, x_2, ,\}$とする。文の文字間位置を$i : i = 1, 2, 3, ,$とし、文字間位置iにおける文字間接続確率をP_iとする。単語分割される文字間位置の集合$S_x, x = x_1, x_2, ,$は分割解の候補により違うものとする時、

$$\hat{x} = \arg\min_{S_x} \sum_{i \in S_x} Pi \qquad (5)$$

となる候補を選択する。[4]

例えば図5では、解の候補(1)の分割点の集合は$S_{(1)} = \{2, 4, 6\}$であり、解の候補(2)の分割点の集合は$S_{(2)} = \{3, 5, 7\}$である。よって$P_2 + P_4 + P_6$と$P_3 + P_5 + P_7$を比較し、その小さい方を解として選択する。

$Pi > 0$であることから、(5)式は分割数最小法と同様な効果も得られる。しかし必ずしも分割数最小の候補が正解とは限らない。そこで定数Thを導入することで、(5)式は容易に次のようのに拡張できる。

方式 3 (接続確率最小法:応用形 1) 全ての文字位置の集合を$N = \{i : i = 1, 2, ,\}$とし、単語に分割さ

[4]和ではなくて積として計算することも可能である。ただし積にすると、基本式のままでは分割数最小法と同じ振舞いにはならない。

(1) 大阪-(P_2)-市立-(P_4)-山田-(P_6)-中学校
(2) 大阪市-(P_3)-立山-(P_5)-田中-(P_7)-学校

図 5: 複合語の分解の複数の候補

れる文字位置の集合をS_xとする時、単語分割される文字位置の場合は文字間接続確率P_iを、そうでなければ定数Thを加算し、その和をもって解の候補のスコアとする。このスコアが小さい方の候補を選択する。

$$\hat{x} = \arg\min_{S_x} \sum_{i \in N} Qi \tag{6}$$

$$Q_i = \begin{cases} P_i, & i \in S_x \\ Th, & i \notin S_x \end{cases} \tag{7}$$

応用形 1 により、分割数最小法にとらわれず自由な分割数で分割できる。

ところで前章の文字間接続確率による単語分割法は、辞書を全く使わない単語分割を可能とした。辞書を使わないということは、未知語が存在しないことに等しい。そこでこの特徴も融合の対象として検討した。

式 (6) は文字間だけではなく単語そのものにもスコアを与えることで、次のように拡張することができる。

方式 4 (接続確率最小法:応用形 2) 単語分割候補 x に出現する単語列を $W_x = w_1, w_2, ,$ とする。辞書中の単語の集合をDとする時、単語w_jに対してそれが辞書に載っている場合はスコアUを、そうでなければスコア$V(V > U)$を与えることとして、式 *(6)* を拡張した式 *(8)* によって計算されるスコアが小さい候補を解として選択する。

$$\hat{x} = \arg\min_{S_x, W_x} \left(\sum_{i \in N} Qi + \sum_{w_j \in W_x} R_j \right) \tag{8}$$

$$Q_i = \begin{cases} P_i, & i \in S \\ Th, & i \notin S \end{cases} \tag{9}$$

$$R_j = \begin{cases} U & w_j \in D \\ V & w_j \notin D \end{cases} \tag{10}$$

$$U < V \tag{11}$$

字種等にとらわれず自由な未知語推定を行い[5]、これを解の候補の作成に利用して応用形 2 で計算することで、未知語への対応が可能となる。この場合、未知語部分は 3章の方式で分割し、その他の部分は (6) 式で分解することと同等になる。この時、定数Thが3章の閾値Tに相当する。

4.3 実験と評価 (2)

以上の原理を検証するため、接続確率最小法の式 (8) と Rule 方式 [1] とを組み合わせたプログラムを作成し、実験を行なった。

プログラムの処理の概要を図 6に示す。

(1) 単語抽出を行ない単語リストを作成。
(2) 文字間接続確率を計算。
(3) 以下を文に関して繰り返し実行。
  (3.1) 単語リストと未知語推定を基に
    分割解候補を作成。
  (3.2) 文字間接続確率を用いて候補を選択。

図 6: 処理手順

手順 (1) と手順 (3.1) は Rule 方式のプログラムをそのまま用いた。手順 (1) の単語リストの作成は CD-毎日新聞 95 年版データの本文 92MByte から、手順 (2) の確率テーブルの作成は同データの本文 80MByte から行なった。手順 (3.1) の未知語推定は、任意の文字列を未知語とみなして、解に加える方式とした。手順 (3.2) の文字間接続確率については、3-gram の積を用いた。

評価は 3章と同様に京大コーパスのを正解として参照し、F 値の最大値を精度として用いた。閾値による F 値の変化を図 7に示す。実験の結果、精度として F 値 91.4%を得た。この時、接続確率最小法を用いない Rule 方式単独では 90.8%だった。

しかし Rule 方式は独自の基準により単語分解を行なっているため、この基準に沿って評価する必要がある。京大コーパスを修正して作成した正解に基き、

[5]例えば辞書にない 2~4 文字の文字列を全て未知語とみなすなど

再評価を行なった結果、Rule 方式単独で精度 92.4%、接続確率最小法の併用で精度 93.1%を得た。

また、特許公開広報 CD-ROM から IPC 分類 C セクションのみを取り出したデータを用いて、60MByte のデータからの学習によるクローズドデータの評価では、独自に作った正解との比較で精度 95.6%を達成している。

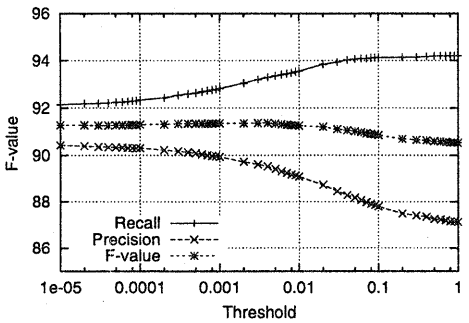

図 7: 閾値と再現率・適合率・F 値の関係 (2)

計算方式＼評価	京大コーパス	独自基準
Rule 方式単独	90.8%	92.4%
+接続確率最小法	91.4%	93.1%

表 3: 接続確率最小法の実験結果

4.4 考察 (2)

実験より、Rule 方式 [1] は単独の場合に比べて、接続確率最小法によりわずかに改善されていることがわかった。改善の効果が低いのは、接続確率最小法の効果が表れるのが主に複合語分解の部分であり、文章全体から見て Rule 方式だけで複合語分解に失敗していた部分が小さいことに起因していると考えられる。

また Rule 方式は平仮名語の抽出は充分には出来無いため、平仮名部分の単語分割は文字間接続確率に頼ることが多かった。しかし、平仮名部分の分割は文字間接続確率を使っても間違うことが多いようである。

例えば「～記録であり、」という文について正解では「であり」を一つの単語 (判定詞) としているが、接続確率による分割プログラムは「～記録/で/あり/、」のように分割する。逆に正解では「～/する/だけ/で/、」と分割する所を、プログラムは「～/する/だけで/、」と分割する。このように分割数が多かったり少なかったりする部分があるため、分割閾値だけではプログラムの動作をコントールすることができず、全体の精度の改善に寄与していないと考えられる。

Rule 方式に比べて接続確率最小法の効果があった部分の例を図 8 に示す。

兵庫/県北部→兵庫県/北部
神戸市/立/本山/第一小→神戸/市立/本山/第一小

図 8: 分割結果の改善箇所の例

このように漢字の複合語の分解には効果があった。本研究の目的は、全文検索のインデックス作成用の単語分割の実現である。筆者らのグループがある日本語全文検索システムの利用実体を調査した結果、検索システムの Query に占める名詞の割合は 9 割以上であり、しかも全体の 8 割以上が漢字名詞か片仮名名詞であった。検索システムとして再現率を重視するなら、複合名詞の正確な分解が必要であり、本方式は意味を持つと考えている。

ところで式 (8) は未知語が含まれていても分割可能であるという特徴を持つ。これは文書の処理の過程で未知語の抽出も可能ということを意味する。(本実験の場合、未知語とは図 6の手順 (1) のルールによる単語抽出で抽出できなかった単語のことを差す。) 新聞データの処理を行いながら抽出した単語の例を図 (9) に示す。これは抽出された単語から漢字で始まるものだけを抜き取って例示したもので、誤りも含まれている。未知語抽出として用いることが出来るかどうかの評価には、分割に用いる辞書の整備をした上で実験する必要があると考えられる。

駒ケ根市, 過多, 長さは, 郵船, 郎氏, 独創的な, 政調, 管区, 俊史, 文化賞 , 三浦, 県初, 豊朗, 米新, 直ちに, 見せた,

図 9: 抽出された未知語の例

5 終わりに

文字接続確率を利用した、教師無し単語分割について 2 つの方法を述べた。

4章で述べた接続確率最小法については Rule 方式による単語分割の拡張として実験と評価を行なったが、単語分割解の候補を作成する方法はこれに限定されるものではないことに注意したい。よく整備された辞書を用いる一般的な形態素解析でも、接続確率最小法で補強することが可能なのである。

その意味では、接続確率最小法は 4章のような実験評価ではなく、辞書を用いた形態素解析で評価するべきだろう。あるいは複合語部分のみで実験評価するべきかもしれない。

タグ無しコーパスからの学習 (いわゆる教師無し学習) の研究は多くない。[8] では再推定可能な統計的単語モデルを構築し、単語リストを再推定しながら不適当な単語を排除して教師無しの単語分割を実現している。[7] では文字列に対する正規化頻度を計算し、これを文書の最初から適用していく限界頻度法を用いることで単語分割を実現している。

これらに対して本方式は、計算量の少なさが特徴の一つと考えている。[7] の方式では最も長い単語の長さまでの正規化頻度をあらかじめ計算しておく必要があるが、本方式では 3-gram の確率値の計算だけで済むため、そのコストは比較的小さいのではないだろうか。

本論文ではクローズドデータの実験のみを行い、オープンデータでは試していない。学習データに無い文字組みの問題を避けるためであり、これは N-gram 言語モデルを用いた場合の共通の課題である

一般に N-gram 言語モデルは、文章全体について N-gram 確率の積を取るため、加算法、線型補間法、バックオフ法などの補正で良好な結果を得ている。しかし本方式では確率は局所的に評価されるため、加算法が意味を無さないことは自明である。同様に線型補間やバックオフについてもその効果は未知数であり、これらは今後の確認課題としたい。

参考文献

[1] 飯塚泰樹 : 全文検索のための字面解析による単語分割, 情処 NL 研究会 132-9, 1999

[2] 森信介 長尾眞 : nグラム統計によるコーパスからの未知語抽出, 情処・論文誌, Vol.39-7, 1998

[3] 森信介 長尾眞 : タグ付きコーパスからの統語規則の獲得, 情処・論文誌, Vol.37-9, 1996

[4] 新納浩幸 井佐原均 : 疑似Nグラムを用いた助詞的定型表現の自動抽出, 情処・論文誌, Vol.36-1, 1995

[5] 下畑さより 杉尾俊之 永田淳次 : 隣接文字の分散値を用いた定型表現の自動抽出, 情処・自然言語処理研究会, 95-NL-110-11, 1995

[6] 山本幹雄 : 統計的言語モデル - 理論と実験 - , 言語処理学会 第 5 回年次大会 チュートリアル資料, 1999

[7] 中渡瀬秀一 : 統計的手法による単語の切出しについて, 信学技報 NLC95-68, 1995

[8] 永田昌明 : 単語頻度の再推定による自己組織化単語分割, 情処 NL 研究科 121-2, 1997

[9] 永田昌明 : 前向きＤＰ後向きA* アルゴリズムを用いた確率的日本語形態素解析, 情処 NL 研究科 101-10, 1994

[10] 竹内孔一 松本祐治 : 隠れマルコフモデルによる日本語形態素解析のパラメータ推定, 情処・論文誌, Vol.38-3 1997

[11] 京都大学テキストコーパス, http://pine.kuee.kyoto-u.ac.jp/nl-resource/corpus.html