自然言語処理　132－9
（1999. 7. 23）

全文検索のための字面解析による単語分割

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
〒140-8632 東京都品川区東品川4-5-15
iizuka@trl.mei.co.jp

あらまし

本論文では、辞書を用いない字面解析による単語分割手法を提案する。本手法ではまず、対象文書からルールにより単語抽出を行う。単語抽出は、字種パターンとその前後を抜き出したnグラムを作成して行う。この際、ルールを詳細化して高精度化を図ると共に複数の方式を併用することにより抽出単語数を確保した。このように得られた単語と分割用ルールを相補的に用い、形態素解析に似たアルゴリズムを採用することで、字種変化点にとらわれない精度の高い分割に成功した。本手法はクローズドデータの処理方式での実験の結果として、新聞データに対して適合率90.2%，再現率85.8%を得ることができた。

キーワード　単語抽出，単語分割，コーパス，ルール

Japanese Word Segmentation Using Textual Analysis for Full Text Search

Yasuki IIZUKA

Multimedia Systems Research Laboratory
Matsushita Electric Industrial Co.,Ltd.
4-5-15, Higashi-Shinagawa, Shinagawa-ku, Tokyo 140-8632 Japan
iizuka@trl.mei.co.jp

Abstract

This paper presents a word segmentation method based on a textual analysis. This method does not require any dictionary. The proposed method consists of two steps. The first step is building list of words by filtering string clusters devided by heuristic rules. These heuristic rules mainly utilyzes character types. The second step is segmenting texts based on the extracted word list and the other heuristic rules. The score of evaluation experiment is 90.2% precision and 85.5% recall.

key words　word extraction, word segmentation, corpus, heuristic rule

1. はじめに

膠着語である日本語の処理を考えた場合、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[2][12]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[8]、接続コストの自動学習[11]が提案されている。

特にタグ付コーパスからの文法学習や単語接続（分離）可能性の学習は数多く提案されているが、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処理対象文書と同じ分野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。

一方、コーパスから文字間（文字列間）の分割・接続確率を計算することで、形態素解析によらない単語分割を行う提案もある[7]。

単語抽出の技術は進歩しつつあり、これらの技術を実用システムに適用できる日も近いことをうかがわせる[4][5][6]。

ところでタグ無しコーパスからの単語抽出に統計的手法を用いるものは、適合率を上げようとすると、コーパス量に対して得られる単語数が少なくなってしまうという課題が残る。よって出現頻度の小さな未知語は抽出することが難しい。文書の追加・更新が頻繁にあるテキスト・データベースにおいては、常に新たな未知語が入ってくる可能性がある。しかもそれらの未知語は一度に追加される文書が少ない場合にはサンプル数が充分に取れず、統計的な処理だけで抽出することには困難が伴う。

本研究はこのような背景を踏まえ、全文検索を行うテキスト・データベースの単語分割用辞書整備コスト０を目標とし、ルールベースの抽出単語を用いた単語分割方式を提案するものである。字種に着目したルールベースの単語分割[1]、字種ルールによる抽出単語を使った単語分割[3]、小さな単語リストと大量のテキストからの単語分割[9]の報告が既にあるが、本研究はこれらと比較して

(i) ルールベースでありながら、従来の字種変化点だけに依存した単語分割ではなく、漢字仮名混じり単語の分割も可能とする
(ii) コーパスからの学習が単語知識であるため、学習された内部状態の修正が可能である
(iii) 目的に応じた複数の単語抽出方式と、一般の形態素解析に似たルール順序独立のアルゴリズムを採用し、高精度の単語分割を実現する

という３つの特徴を持ち、複雑なルールを多用することを避けながら、辞書を用いない単語分割の実現を目指した。

本論文ではこの方式について、２章で抽出単語による単語分割の基本的な考え方、３章において単語分割方式、４章で単語抽出方法について述べ、５章において実験と評価、６章で考察を行う。

2. 抽出単語を用いた単語分割

本手法の基本的な考え方は次の通りである。

既存の辞書を用いずに、ある文書中に出現した「新しいおもちゃ」という文字列を単語に分解することを考えるとしよう。ルールだけによる解析は難しいが、「新しい」または「おもちゃ」のどちらか、または双方が単語だとわかっていれば分割も可能になる。「新しい」に関しては、同じ文書の別の個所に「新しい素材は、」「新しい話題を」などの文字列が出現していれば、ルールにより抽出可能と考えられる。さらに前者の例では「素材」は別のルールで名詞として抽出可能であるから、「新しい」は名詞に接続するものとして抽出可能であろう。このようにして抽出できた単語を、既存の形態素解析同様に最長一致で当てはめていけば、単語分割点が推定できるはずである。

そこで本手法は、処理対象文書（タグ無しコーパス）から単語抽出を行い、抽出された単語のリストを利用して単語分割を行う（図 2-1）。本方式では、単語抽出と単語分割をクローズド・データで行うことを前提とする。一度何らかのコーパスから学習した結果を汎用的に用いることも検討するが、それでも対象文書ごとに単語抽出を行うことを基本とする。

少ないコーパスから単語抽出を行うには、統計的処理よりもルールによる抽出の方が有効と考えられる。抽出単語の品質を保つため、ルールをある程度詳細化し、その一方で複数の方式を併用することで抽出単語数を確保する。単語分割においては抽出単語を手がかりとして、分割用ルールを併用して精度の向上を図る。単語抽出過程、単語分割過程の双方にルールを用いるが、抽出単語が単語分割のルー

ルにとってトートロジー的にならないよう独立したルールを用い、相補的に働くことを目指す。

単語抽出・分割時に品詞の同定までは行わない。

本手法は、単語抽出ルールに含まれる知識からのブートストラップと捕らえることも可能である。

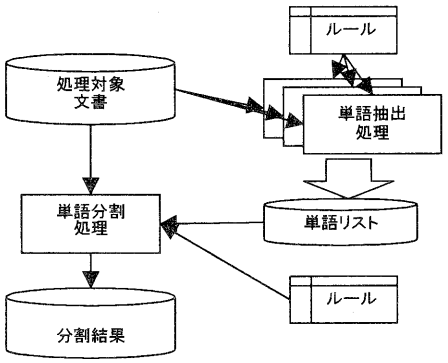

図 2-1 本方式の概念図

2.1. 単語分割の基準

テキスト・データベースにおいて検索対象となる語は名詞が中心と考えられることから、そのインデックス作成における単語分割は、テキスト中に名詞を発見できることが必要である。その一方で、助動詞や接続詞は検索対象にならないと考えても差し支えない。

対象文書分野にもよるが、これらの名詞は漢字や片仮名で記述されていることが多い。逆に検索対象となり得ない助動詞や接続詞は平仮名で記述されている。

漢字で記述される名詞のほとんどは、2〜3 文字の短単位語から形成されていると考えられる。また、検索における再現率を落とさないためには、単語の長さをこの短単位語程度にするべきと思われる。そこで本方式では次のように文章を分割することを目標とする。

(1) 検索対象となる単語の始点終点を発見する。
(2) 漢字単語は 2〜3 文字の短単位語に分割する。
(3) 検索対象とならない接尾辞は分割しない。

品詞別の分割評価基準の概略を表 1に示す。

この基準により、例えば次のように分割する。[1]

[4月]の[都知事][選挙]で[新しい][知事]が[誕生]した。

複合名詞は積極的に 2 文字か 3 文字の語に分割する。これは「二酸化炭素」という単語も、積極的に[二酸化][炭素]のように分割することを意味する。このように定義することで、分割プログラムの実現、および評価が容易になる[2]。

本方式は、以上の基準を元に設計・評価を行った。

品詞	基準
名詞	短単位語に分割／形式名詞は除く
動詞	語幹のみ／平仮名動詞は除く
形容詞	活用語尾を含む
助動詞・指示詞・助詞	除く
副詞	平仮名副詞は除く／名詞を含むものは名詞として
数値	助数詞まで一つとする

表 1 分割基準の概略

3. 単語分割方法

3.1. 単語分割手順

まず単語分割方法について考える。

本手法の単語抽出処理では、文書中の全ての単語が抽出されるわけではない。特に助詞や副詞などの単語抽出は行っていない。よって抽出単語リストを用いたとしても、通常の辞書を用いた形態素解析の手法をそのまま適用できるわけではない。そこで本手法では、次の手順で単語分割を行う。

1. 抽出単語の発見
   既に得られた抽出単語リストの単語が、文の中に含まれているかどうかを調べる。形態素解析の辞書引きと同様の処理を行う。
2. 数値表現の発見
   文の中の数値表現を発見する。この処理にはオートマトンを利用し、助数詞まで含めて数値表現とみなす。
3. ルール適用
   字種、格助詞表現、既に発見されている抽出単

[1] 最終結果は単語分割というよりキーワード抽出と呼ぶべきものであるが、不要部分削除の結果このように見えるだけであり、本論文ではこれを単語分割と呼ぶ。
[2] 「四面楚歌」なども分割しようとしてしまうのは問題であるが。

語などを手がかりに、ルールを適用し、単語分割点と単語候補を発見する。

4. 単語へ分割
手順3までで単語の候補が出そろったので、文頭から妥当と思われる候補を順に選んで、文を単語に分割する。

5. 不要部分削除
最終的に不要とみなされる語を削除する。

手順1～3が、辞書を用いた形態素解析の辞書引きに相当する。手順4の単語への分割には、基本的に前方最長一致を選択する A*アルゴリズムを用いる。(図 3-1)

ポインタ p を文の先頭に置く。
if(文の終了){
　終了
}else if(p から始まる単語候補 ω_{pi} がある){
　・ p から始まる全ての単語候補 $\omega_{pi(i=1,2,..)}$ について、その単語候補を選択した場合のスコア $h(\omega_{pi})$ を計算する。
　・ 最も高いスコアを得た単語候補 ω_{ph} を選択し、その単語候補の後へ p を進める。
}else{
　・ p の位置の文字を単語に含まれていないと判断し、次の文字へ p を進める。
}

図 3-1 分割アルゴリズム

ただし図 3-1の中で

$$h(\omega_{pi})=score(\omega_{pi})+max_j\,(score(\omega_{nj(j=1,2,..)}))$$

とし、$score(\omega_{pi})$は単語 ω_{pi} 長さと基本スコアの積とする。基本スコアとは、その単語候補がどのルールによって得られたものか（または抽出単語や数値表現だったか）により与えられる数値である。デフォルトではルールから得られる単語候補よりも抽出単語による単語候補の方が、基本スコアが高くなるように設定されている。 $\omega_{nj(j=1,2,..)}$は ω_{pi} を選んだ時に ω_{pi} の後に接続する単語候補である。

ルールの適用と単語への分割の手順を分離して手順4を独立させたことで、本方式ではルールの記述順序についての厳しい制約は存在せず、自由にルールを追加・削除することが可能となっている。

3.2. ルールの記述

ルールの記述は、単語分割点の発見に関するものと単語候補の発見に関するものとの2種類が可能で、先頭から順番に全てのルールが適用される。

ルールは、字種と文字数とからなる正規表現に似た記述[3]を基本とし、抽出単語やルール適用により既に発見されている単語候補などもパターンとして利用できる。

ルールの記述は次のような仮定に従う。

抽出単語として単語であることがわかっている文字列は、その前後に単語分割点が存在する。字種パターンや助詞と思われる文字列を利用すれば、単語候補が発見できる。ルールによって発見された単語候補の前後にも単語分割点が存在する。複数のルールで同じ個所が単語分割点であると推定された場合、その個所の分割可能性は加算されて高くなる。分割可能性の高い個所間が2～3文字である場合、その2～3文字は単語である。

以上のような仮定をルールとして記述する。記述例を図 3-2に示す。

```
200:(単語開始) → (非単語 漢 3) ← (単語終了) ;
301: 10:(非漢)→ (非単語 漢)← ( “は、” );
```

図 3-2 ルール記述例

“単語開始”とは、発見されている単語の終端と次の文字の間に付与されたマークか、またはカッコ等の記号にマッチする。単語終了も同様である。これらの情報は、単語分割点を特定するために使われる。“→”と“←”で囲まれた部分が単語候補となる。個々のルールには、単語分割の手順4で用いるために、ルールから得られた単語候補の基本スコアを指定できる。

4. 単語抽出方法

単語分割を精度よく実現するためには、基本単位となる単語を数多く正確に抽出することで、分割の手がかりを増やす必要がある。このためには、目的に沿った数種類の方法を組み合わせることが有効である。

2.1節の基準で漢字複合名詞を分解するためには、2～3文字からなる漢字名詞を抽出しておく必要がある。(ここではこのような単語を漢字基本名詞と呼ぶ。)

「大変美しい」のような連続した漢字を正確に分割するためには、送り仮名が付く漢字1文字か2文字の用言も抽出する必要がある。

「特産品はおいしいうどんとおもちゃだ。」といった平仮名の連続を切るためには、平仮名のみか漢字

[3] 作成した処理系は、or や not の記述が充分にできるものではない。

平仮名混じりの名詞を可能な範囲で抽出しておく必要がある。

最終的には分割結果に表れない語（2.1章参照）も抽出して分割点発見のために利用する。

表 2に今回採用した単語抽出の方式の一部と、それによって得られる単語の例を示す。２章で述べたように、本方式は分割に対して品詞情報を使わないため、品詞を特定することなく多くの単語を抽出することを目指した。以下では、この幾つかについて抽出方式を説明する。

タイプ	例	パターン
漢字基本	言語 食事	漢字２,３文字
送り仮名付 1-1/1-2/2-1	遠い 右回り 新しい	(漢)+(ひ)+(“い”) など
人名地名	東大阪市	県～市　～さん
名詞一般	おもちゃ	～を
文頭平仮名	しかし	“。”+（ひ 3）+”、”

表 2　単語抽出方式一覧（抜粋）

4.1. 単語抽出基本手順／漢字基本名詞

単語抽出はルール（パターン）を用いて次の手順で実行する。

i.　字種パターン n-gram の作成
ii.　ルールによる取捨選択
iii.　出現回数による選択

以下、漢字２文字名詞の抽出を例にして、上記手順を説明する。

i.　字種パターン n-gram の作成

文書からある字種パターンに合致したものだけを取り出した n-gram を作成する。これをここでは字種パターン n-gram と呼ぶことにする。字種パターン n-gram は、目的とする字種のパターンの前後数文字まで含めて KWIC のように取り出すものである。

漢字２文字名詞の場合は、目的とする字種パターンは図 4-1となる。

前	目的部分	後
漢字以外	漢字２文字	漢字以外

図 4-1　漢字２文字名詞抽出の字種パターン

漢字２文字連続で直前直後が漢字ではないものであり、その前後２文字ずつを含めて取り出す。例えばこの文章を対象として、漢字２文字連続の前後を含めて取り出すと図 4-2のようになる。

前２文字	目的部分	後２文字
この	文章	を対
章を	対象	とし
て、	漢字	２文
続の	前後	を含

図 4-2 字種パターン n-gram の例

ii.　ルールによる取捨選択

字種パターン n-gram の前後の文字に注目し、単語候補をフィルタリングする。漢字２文字単語の場合、目的部分の前については図 4-3の条件を満たすものを排除する。目的部分の後については、図 4-4の条件を満たすもののみを適当とし、これ以外を排除する。

次の条件を排除	排除されるパターン例		
前が数字	第一	回目	は、
前が漢字平仮名	聴き	手側	は、
前が特定の平仮名	せい	剤約	７グ

図 4-3 フィルタリング条件１

次の条件のみ選択	パターン例		
直後が句読点	ない	場合	、こ
直後が格助詞相当	た。	東京	は、
後２文字が「から」等	た。	東京	から

図 4-4 フィルタリング条件２

iii.　出現回数による選択

以上のようにフィルタリングした結果について、目的部分だけを取り出し、ソートして、出現回数を調べる。その結果、ある文字列がＮ回以上出現していた時、この文字列を単語と認めて単語リストに加える。現在、Ｎを２としているが、Ｎを１より大きく設定しているのは、非常に珍しい組み合わせや、コーパスとなる文書原稿の校正ミスと思われる間違った語の抽出を避けるためである。

4.2. 送り仮名付 1-2 型語の抽出

漢字１文字に送り仮名２文字が続く語として、形容詞の基本型（終止形）などがある。（ここではこれを送り仮名付 1-2 型語と呼ぶ。）このような単語の抽出には、漢字基本名詞の場合と同様の手順を踏み、まず図 4-5以下のように字種パターン n-gram を作成する。

前	目的部分			後
漢字以外	漢字1文字	平仮名1文字	“い”	平仮名以外

図 4-5 送り仮名付 1-2 型語抽出の字種パターン

フィルタリング条件は、前後の条件はこのままで、目的部分の中心の平仮名一文字を経験から得られた条件で取捨選択する。これにより、「頭がい骨」「腹ばい状態」などを削除する。新聞の場合、「妻あい子さん」のような記述があるため、これらもフィルタリングするが、「下さい」などはそのまま含めるものとする。

目的部分の最後が「く」のパターン（「美しく」など）についても、形容詞連用形などを含む語の集合として同様に抽出する。

4.3. 地名・人名の抽出

地名・人名は漢字基本単語抽出では抽出できない場合があるため、別の方法を用意する。

都市名抽出には、基本的には図 4-6のパターンを用いる。

前	目的部分	後
漢字2字+県	漢字m文字	市・町など

図 4-6 都市名抽出の字種パターン

パラメータmを1～3の間で変化させることで、市の名前や町の名前を得ることができる。得られた市の名前を元に、ブートストラップの手法でさらに町の名前を得ることができる。

人名については、漢字 N 文字に続く“さん”などを手がかりに人名全体の候補を抽出し、前方の M (M=2,3) 文字に頻出する文字列がある場合、それらを苗字の候補とする。後方についても同様に頻出文字列を調べ、名前部分の候補とする。

4.4. その他の単語の抽出

このほかに、文頭（句点直後）から漢字か読点までの数文字の平仮名を接続詞などとして抽出する[4]。「～を」「～は、」などのパターンから取り出した3～4文字の文字列からは、文字列前方の分散の度合によりフィルタリングをすることで、漢字仮名混じり名詞を抽出する。

一度抽出した3文字の漢字基本名詞から前方2文字、後方2文字に頻出する文字列を調べることで、接辞や語幹の推定が行える。

さらに、既に抽出した漢字基本名詞を使い、その前後の字種パターンにより、名詞に接続する単語の抽出が可能である。

5. 実験と評価

前章までに述べてきた方式をもとに、単語の抽出・分割実験を行った。単語の抽出については、C のプログラムにより字種パターン n-gram をディスク上に作成し、これを on-disk のまま awk,sort,uniq などを用いて処理した。単語の分割については、3章で述べた手順を実行する C のプログラムを作成した。

実験の処理対象文書には朝日新聞の 94 年データ(本文 約 223Mbyte 約1億1千万文字)を用いた。分割結果例を図 5-1に示す。[5]

[単語][分割][の][新しい][技術][を][調査][する]。

図 5-1 単語分割結果の例

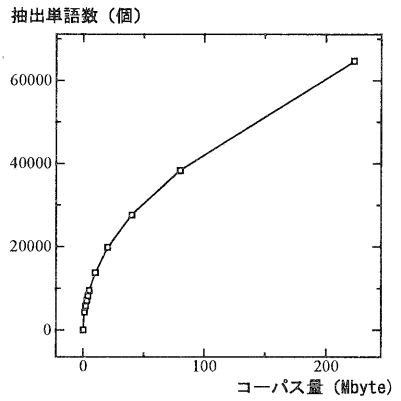

図 5-2 単語抽出実験結果

コーパス量を変化させた時の抽出単語数の変化を、図 5-2に示す。グラフ上のプロットは、コーパス量が 1,2,3,4,5,10,20,40,80,223Mbyte の時に相当する。抽出単語総数は、全データ(223MByte)の時に 64,661 個だった。このうち 4 章で説明した漢字 2 字基本名詞は 18,182 個で、JUMAN[12]の辞書を正解とした場合の適合率は 87.3%だった。送り仮名付 1-2 型語では、字種パターン n-gram が 442 個、フィルタリングによって最終的に得られた語は 175 個で、目視の結果、適合率 84.6%であった。間違った

[4] これは分割結果には表れない。

[5] この例は不要語削除をしていない状態。

ものは、「目ない」(切れ目ない)「応なく」(いや応なく)など27個があった。

このように抽出した単語を用いて、単語分割を行った。単語抽出に用いるコーパス量を変化させた時の単語分割精度測定結果を図 5-3に示す。

単語分割精度の測定は、単語分割正解との比較により再現率と適合率を求めることで行った。単語分割正解は対象文書をJUMANで解析した上で、2章の分割基準に沿って、プログラムと手で修正を加えて作成した。

正解はコーパスの先頭部分1Mbyte以内から任意に抽出した38記事319文で、いずれもクローズド・データである。

分割時ルールを148個[6]使った実験で、単語抽出コーパスが223MByteの時に、分割精度として適合率90.2%、再現率85.8%という結果を得た。

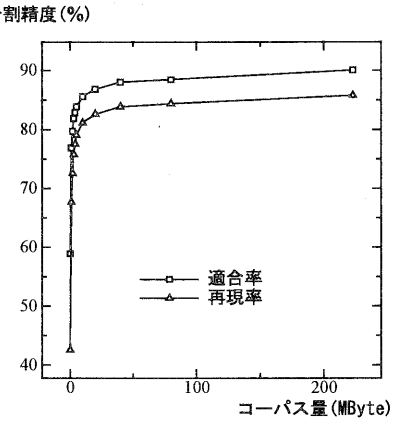

図 5-3 単語分割実験結果

6. 考察

まず単語抽出について検討する。漢字基本名詞の適合率が悪いのは、「亜大」などの省略語、「易水」などの珍しい語、地名の「原尻」などがJUMANの辞書に載っていないことから、単純な比較には無理があったためである。しかし「子氏」(「佐藤ひろ子氏」の部分から抽出)「陰様」(「お蔭様で」の部分から抽出)などの誤りもあり、200語のサンプリング調査では適合率98%程度と推定される。これは今回採用した複数の単語抽出方式の中で、最も効率の良いものであった。

一方、送り仮名付1-2型語では、前章で説明した通りここから得られた単語数は175個と少ないが、ノイズもかなり混じっていた。抽出数が少なかったのは、そもそも日本語の中で、このような条件に合致する形容詞などの語彙が（名詞に比べて）少なかったからと推測される。

単語分割は、単語抽出の元となるコーパス量が0の場合でも適合率60%程度となっているが、これはルールだけによる分割の精度である。

単語分割の精度を落としている大きな原因は2つあり、その一つは平仮名を含む名詞である。「子ども」「初もうで」などが相当し、仮名混じり名詞については単語抽出が充分ではないことを示している。もう一つの原因は用言の活用語尾である。今回の基準策定の際、動詞と形容詞では基準を異なるものにしてしまったことから、処理もやや乱雑になっていることは否めない。

単語抽出はルールを詳細化すれば、抽出単語の適合率向上が見込まれるが、その一方で抽出単語数は減少する。では抽出精度と抽出単語数のバランスを変化させると、単語分割にどのように影響するのだろうか。

現在、この点について予備的に調べているが、チューニングされた現時点を基準に、抽出する単語数を増やして品質が低下した場合も、抽出する単語の品質を上げて抽出単語数が減った場合も、分割精度が低くなる傾向が観測されている。最適ポイントがどのような状態かは経験からしか得られていない。

しかし図 5-3及び図 6-1から読めるように、抽出単語数がある数に達っすると安定した分割精度が得られることが判明している。本実験の新聞データの場合、このポイントはコーパス量が20Mbyteの付近であった。

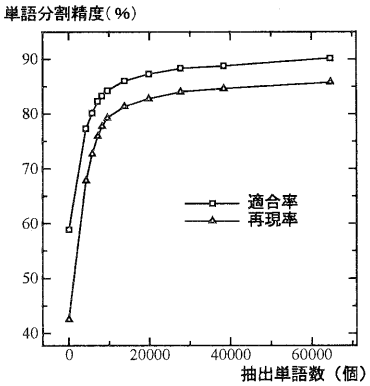

図 6-1 抽出単語数と分割精度

[6] ルール記述にはor表現が使えないため、orの部分を複数のルールに分けて記述している。

7. おわりに

本論文では、単語抽出の結果を用いた単語分割について述べた。実験の結果、まとまった文書があれば充分な単語を抽出することが可能であり、この単語を用いることで、辞書を用意しなくとも文を単語に分割できることが確かめられた。

統計処理によるパラメータ学習／新語抽出を用いた形態素解析が統計的と呼べるならば、本方式の単語分割の精度は確率的と呼べるものであろう。単語はコーパス中に複数回出現するが、多数回出現することで単語抽出されるのではなく、ただ一回(実際には N(固定)回)条件にあてはまる形で出現することで抽出される。そして抽出されればそれがすぐに分割に利用されるからである。

その意味では本方式では、最初に大量の単語を含んだ良質のコーパスが存在するか、または初期状態として基本的な単語のリストをあらかじめ持っていることで、精度の高い単語分割が実現できると思われる。

初期の目標である少ないコーパスからの単語抽出は、まだ課題として残っている。データベースの追加・更新をモデルとし、初期状態で与えられるコーパスからの単語抽出結果とその後に与えられる追加コーパスからの単語抽出結果について、単語分割精度に与える影響を解析する必要があるだろう。(これは図 6-1の右半分の状態の解析に相当する。)

また、単語分割結果から抽出単語へフィードバックする機構[9]についても検討する必要がある。

単語分割の精度は、およそ 80%程度にまで落としても検索の精度に大きな影響はないとする報告がある[10]。本手法は、適合率で 90.2%、再現率でも 85.8%を達成している。基準が甘く副詞などが充分に取れていないといった課題が残るが、検索システムのための単語分割方式として本手法は有効であると考えられる。

今後、先に述べたように、単語抽出の精度と抽出単語数のバランスが単語分割精度に与える影響の調査などを通し、ヒューリスティックスの改良を行いながら単語分割精度の向上を目指すつもりである。

謝辞

本研究で利用したコーパスは、朝日新聞９４年データを用いています。利用を許可していただいた朝日新聞社に深く感謝いたします。

参考文献

[1] 鈴木惠美子 ： 統計調査に基づく文字列パターンを用いた日本語文自動分割,
信学会論文誌 D-II Vol.J79 No.7 (1996)

[2] 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・自然言語処理研究会, 94-NL-101-10(1994)

[3] 下畑光夫 杉尾俊之 ： 文字種切り出しと複合語分解によるキーワード抽出,
情処・自然言語処理研究会, 97-NL-120-13 (1997)

[4] 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)

[5] 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)

[6] 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)

[7] 中渡瀬秀一 ： 正規化頻度による形態素境界の推定,
情処・自然言語処理研究会, 96-NL-113-3(1996)

[8] 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)

[9] 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・自然言語処理研究会 97-NL-121-2, (1997)

[10]多田智之 金岡秀信 ： 「形態素解析の検索システムに与える影響」
言語処理学会 第４回年次会 発表論文集,(1998)

[11]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)

[12]日本語形態素解析システム JUMAN
http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html