# 多段解析法による日本語形態素解析の精度

1R－2

白井　諭[*1]　横尾昭男[*1]　池原　悟[*1]　奥山信輔[*2]　宮崎正弘[*3]

[*1]ＮＴＴコミュニケーション科学研究所　[*2]ＮＴＴソフトウェア　[*3]新潟大学工学部

## 1 はじめに

形態素解析は自然言語処理において最も基本となる技術であり，日本語の形態素解析の誤り率は新聞記事などの一般テキストに対して現在は単語単位で１～３％と言われている[長尾94]。実用上は誤り率を１桁程度は減少させたいところだが，精度向上のための具体的方策は必ずしも明らかではない。

筆者らが研究中の日英翻訳システムには，先に提案した多段解析法[宮崎86]による形態素解析処理が組み込まれている。そこで，新聞記事に対する形態素解析処理の走行実験を行ない，解析精度を測定するとともに問題点の分析を行なったので，その結果について報告する。

## 2 多段解析法による形態素解析の概要

多段解析法は，解析精度と処理能力を両立させるため，局所的総当たり法による形態素解析をベースとし，構文や意味の情報が有効となる複合語解析や同形語判別などには，部分的に深く解析することを特徴とする[宮崎86]。処理の概要を以下に述べる。

**①仮文節境界の設定**

字種の変化点（ひらがな→漢字・カタカナ等，句読点→非句読点）に着目して仮文節境界を設定する。仮文節は局所的総当たり法による単語認定を行なうための処理単位である。

**②単語候補の抽出**

単語意味辞書[池原93]を検索し，仮文節内の単語候補を抽出する。仮文節境界を挟んで混ぜ書き語が位置すれば，仮文節境界を補正する。

（例）|売り|出しをけん|制する|　（|は仮文節境界）
　　　　↓　　↓　　↓　（混ぜ書き語抽出）
　　|売り出しを|けん制する|　（仮文節境界補正）

**③ひらがな列分割パターンの抽出**

仮文節内のひらがな列に対する単語候補が連接可能かどうかを文法的に検定し，ひらがな列分割パターンとして可能なものをすべて生成する。

**④ひらがな列分割パターンの絞り込み**

単語意味辞書記載の単語の優先・非優先の指定と仮文節内の自立語数・付属語数を総合的に評価して，ひらがな列分割パターンを一意化する。

**⑤漢字列分割パターンの抽出**

仮文節内の漢字列に対する単語候補が接続可能かどうかを文法的に検定し，漢字列分割パターンとして可能なものをすべて生成する。

**⑥意味的係り受け関係の解析**

漢字列分割パターンのそれぞれに対し，格関係・副詞修飾等の構文的関係や意味属性による単語間の意味的関係を解析する[宮崎84][宮崎93]。

**⑦漢字列分割パターンの絞り込み**

漢字列分割パターンに対して，分割数最小法を基本とし，意味的関係や単語の優先・非優先の指定を加味し評価が高いものを選択する[宮崎84]。

| （例）　分割パターン | 単語数 | 関係数 | 評価値 | 採否 |
|---|---|---|---|---|
| 畜産/物価/格安/定法 | 4 | 1 | 3 | × |
| 畜産/物/価格/安/定法 | 5 | 2 | 3 | × |
| 畜産/物価/格/安/定法 | 5 | 0 | 5 | × |
| 畜産/物/価格/安定/法 | 5 | 3 | 2 | ○ |
| 畜産/物価/格/安定/法 | 5 | 1 | 3 | × |

**⑧単語列分割パターンの抽出**

漢字列分割パターンとひらがな列分割パターンとの文法的な相互制約により，仮文節内の単語分割列パターンを絞り込む。

**⑨単語列分割パターンの決定**

文節境界の設定，同形語の多義の絞り込みなどを行ない，形態素解析の最終結果とする。

本形態素解析で使用する単語意味辞書[池原93]には，一般語12万語のほか，人名・地名・企業名などの固有名詞20万語，専門用語５万語など，合計40万語が収録されている。また，各単語には，3,000項目の意味属性体系に基づく一般名詞意味属性・固有名詞意味属性のほか，約300項目の品詞等の文法情報，上記④⑦で使用する優先・非優先の情報，⑥の構文的情報が付与されている。固有名詞には一般語やその組み合わせたものと同形の語（例えば，平野，八戸）が多いため，上述のような情報が必要となる。

---

**Accuracy of a Japanese morphological analysis based on the multi-level analysis method**

Satoshi SHIRAI[*1], Akio YOKOO[*1], Satoru IKEHARA[*1], Shinsuke OKUYAMA[*2] and Masahiro MIYAZAKI[*3]

[*1]NTT Communication Science Laboratories, [*2]NTT Software Corporation and [*3]Faculty of Engineering, Niigata University

## 3 形態素解析の精度

### 3.1 実験の条件

本稿では，日経産業新聞・リード文965文（情報欄，311記事）を試験文とした。また，試験文中に現れた会社名・人名・商品名のうちの180語に限り利用者辞書に登録した。試験文の概要を以下に示す。

表1 文あたり文字数(平均46.16字/文)

| 字数 | ~10 | ~20 | ~30 | ~40 | ~50 | ~60 | ~70 | ~80 | ~90 | 91~ | 計 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 文数 | 26 | 76 | 147 | 160 | 168 | 153 | 119 | 54 | 32 | 28 | 965 |

表2 文あたり形態素数(平均22.07個/文)

| 個数 | ~ 5 | ~10 | ~15 | ~20 | ~25 | ~30 | ~35 | ~40 | ~45 | ~50 | 51~ | 計 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 文数 | 22 | 81 | 157 | 204 | 172 | 141 | 101 | 57 | 17 | 4 | 9 | 965 |

表3 文あたり文節数(平均8.59文節/文) （補足）

| 個数 | ~ 5 | ~10 | ~15 | ~20 | 21~ | 計 |
|---|---|---|---|---|---|---|
| 文数 | 217 | 476 | 222 | 40 | 10 | 965 |

試験文全体の集計
・文字数=44,544
・形態素数=21,303
・文節数= 8,282

### 3.2 解析精度

本稿では，試験文における本来の形態素境界と解析処理により設定された形態素境界とに着目して解析精度を評価する。前節で述べた条件下で形態素解析を走行させた結果に含まれる誤りを種類別に集計すると次のようになる。

表4 解析誤りの種別とその度数

| 誤りのタイプ | 形態素境界 | | 文節境界 | | 品詞 |
|---|---|---|---|---|---|
| | 切断 | 不切 | 切断 | 不切 | |
| 件数 | 28 | 41 | 31 | 21 | 40 |

この結果，以下に示す評価値が得られる。

**Ⓐ形態素境界の設定**
（形態素単位）
・境界設定再現率 = 20,297／20,338 = 99.80%
・境界設定適合率 = 20,297／20,325 = 99.86%
（文字単位）
・境界設定正解率 = 43,519／43,579 = 99.86%

**Ⓑ文節境界の設定**
・境界設定再現率 = 7,296／7,317 = 99.71%
・境界設定適合率 = 7,296／7,327 = 99.58%

**Ⓒ品詞の設定**（正しく解析された形態素に対して）
・品詞正解率 = 21,194／21,234 = 99.81%

以上から，本形態素解析処理の精度は，品詞の設定を含めると，形態素単位で99.5%程度とみられる。

## 4 今後の課題

### 4.1 解析誤りの分析

例えば「現代/用語」が「現/代用/語」と分割されると形態素切断誤り2件と形態素不切誤り1件が含まれるが，関連があるため，1件の誤り要因に集約できる。こうして集約した要因101件を2節の処理ステップ別に集計すると次のようになる。

表5 解析誤りの処理ステップ別集計

| 処理ステップ | 件数 | 代表例（3.2節の誤り再現とその件数）× 発生件数 |
|---|---|---|
| ①仮文節境界の設定 | 0 | —— |
| ②単語候補の抽出 | 20 | 中心に(形不切1)×11，十分/の|一(形切断2, 文切断1)×4 |
| ③ひらがな列分割Pの抽出 | 12 | のに(形不切1)×8 |
| ④ひらがな列分割Pの絞り込み | 2 | ても(形不切1)×2 |
| ⑤漢字列分割Pの抽出 | 1 | 内/・/外線(「内」の品詞1)×1 |
| ⑥意味的係り受け関係の解析 | 24 | 同|～(文切断1)×5，米(品詞1)×4 |
| ⑦漢字列分割Pの絞り込み | 3 | 中でも(形不切1)×1 |
| ⑧単語列分割Pの抽出 | 14 | 現/代用/語(形切断2, 形不切1)×1 |
| ⑨単語列分割Pの決定 | 25 | で(品詞1)×8，と(品詞1)×7 |
| 合計 | 101 | （注）/ は形態素境界，| は形態素境界かつ文節境界 |

### 4.2 改良の可能性

表5の問題点は，辞書との連携（②），分割パターン抽出・絞り込みルール（③④⑤⑦⑧⑨），意味的関係解析ルール（⑥）の3種類に大別される。現在，改良作業を実施中であるが，表5の代表例など60件程度はルールの改良により解決される見込みであり，最終的な解析精度は形態素単位で99.8%をクリアできると考えている。また，残る誤りに対しても誤り回復のメカニズムも構築中であり[白井94]，実用上十分な精度が得られる見通しである。

## 5 おわりに

多段解析法による形態素解析の精度は，形態素単位で現状99.5%であり，改良により99.8%が達成できる見込みであることを報告した。現在進めている改良の結果については別の機会に報告する。

**＜謝辞＞**
本検討にご協力くださった梅田美砂子氏を始めとするＮＴＴソフトウェアの各位，並びに，日本語単語意味辞書の構築・改良にご協力くださった阿部さつき氏，矢部孝幸氏を始めとするＮＴＴアドバンステクノロジの各位に感謝する。

**＜参考文献＞**
[池原93] 池原，宮崎，横尾：日英機械翻訳のための意味解析用の知識とその分解能，情処論 Vol.34 No.8 (1993.8)
[宮崎84] 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 (1984.11)
[宮崎86] 宮崎，大山：日本文音声出力のための言語処理方式，情処論 Vol.27 No.11 (1986.11)
[宮崎93] 宮崎，池原，横尾：複合語の構造化に基づく対訳辞書の単語結合型辞書引き，情処論 Vol.34 No.4 (1993.4)
[長尾94] 長尾 ほか：自然言語処理技術のこれからの課題，「自然言語処理の技術動向」調査報告会 (1994.3.30)
[白井94] 白井，池原，松尾，兵藤：日本文書き替え処理における制御機能の構成について，第49回情処全大 4K-11 (1994.9)