日本語形態素解析の誤りの回復について

横尾 昭男[*1] 白井 諭[*1] 奥山 信輔[*2] 河村 美砂子[*2] 池原 悟[*3]

[*1]ＮＴＴコミュニケーション科学研究所 [*2]ＮＴＴソフトウエア㈱ [*3]鳥取大学 工学部

1 はじめに

記述された文を対象とした機械翻訳システムなどの自然言語処理システムにおいて，形態素解析は最も入口に位置する処理である。従って，形態素解析において，正しく単語分割を行い，文法的・意味的情報を付与することができなければ，後続の処理は本来の能力を発揮することができない。筆者らは，局所的総当たり法の技術をベースとした形態素解析処理の開発を進め，産業情報に関する新聞記事文に対しては単語レベルでの解析正解率99.5%以上を実現した[1]。

このような精度の達成には，大規模で詳細な情報を持つ単語辞書の開発[2]と精密な解析アルゴリズム[3]を整合させることが不可欠であった。このため，それらをさらに改良して解析精度の向上を図るのは容易なことではない。しかし，新聞記事以外の文を対象にした場合，常にこの精度で解析できるわけではないので，何らかの改良が必要になるというのも事実である。

形態素解析の誤りは，辞書の情報と解析アルゴリズムのバランスに微妙な狂いが生じた場合に起こりやすい。たとえば，接辞処理を強化すると「畜産/物/価格/安定/法」を正しく分割できるようになる半面，「現/代用/語」のような切り間違いが生じる[4]。このような誤りは単語の個別の事情により生じるものであるから，本来なら単語辞書を修正すべきであるが，不用意に修正すると却って精度低下を招きかねないという問題がある。

そこで，本稿では，あらかじめ指定した要注意単語の前後関係に着目して，誤りかどうかを判定し，誤りを回復させる処理を形態素解析の後に付加することにより，形態素解析の精度を向上させる方法を提案する。同様の着眼により，不要な多義を削減する方法も併せて提案する。また，このように，外付けの処理を行うことによって，汎用性には多少欠けるが，一般利用者にも容易に改良が加えられる利点も考えられる。

2 形態素解析の課題

形態素解析の目標として，文字単位99.9%の解析精度と，1000文字/秒の処理速度を，1997年末に達成することが掲げられている[5]。あらゆる対象文に対してこの2つの条件を満たす処理はまだ報告されていないようである。解析精度の点では，どのレベル（たとえば，単語切りの正解，品詞付与の正解，など）まで要求するかによっても評価は異なってくるが，対象文を限定するなどの一定の条件下では目標を達成しているものもある。一方，処理速度の点ではこの目標をクリアしたものが既にいくつか報告されている。

もちろん，この2つの条件を同時に達成するのが最も望ましいが，どのような目的で使用するかにより，条件の重みが違ってくると思われる。たとえば，機械翻訳のようにいくつかの処理が直列に構成されたシステムで使用する場合には，形態素解析が誤れば以降の処理は無力化してしまうため，解析精度の条件の重みが増すと考えられる。これに対して，大規模な言語情報の処理に使用するには，ある程度以上の解析精度は必要であるが，それ以降はむしろ処理速度の比重が増していくと考えられる。

本稿では，日英機械翻訳への適用を念頭に置いて，解析精度の向上策を考える。形態素解析の改良は，辞書の情報と解析アルゴリズムの関係の最適化であると見ることができる。しかし，単語の個別の事情によっては，あらゆるケースを想定した上で相互の最適化が困難な場合があり，また，その詰めを誤ると却って解析精度の低下を招きかねない。そこで，このような個別の事情を形態素解析本体以外の処理で考慮することにより，形態素解析本体の維持の負荷の軽減を図ることを考える。

新たに設ける処理は，解析処理とは独立に単語の個別の事情を考慮するのであるから，解析誤りが生じた特定の単語列に着目して，解析誤りを補

正するためのルールを書いていけばよいことになる。このやり方は，個別にルールを書くため汎用性が乏しい側面は否めないが，逆に副作用の恐れはあまりないと考えられるので，形態素解析本体でかなりの精度が得られる場合においては，形態素解析本体を改良するよりも極めて容易に解析誤りの回復が実現される。

3 誤り回復処理

3.1 基本的な考え方

形態素解析の外付け処理として形態素解析補正と多義絞り込みをルールベースで実現する。ルールは，if-then 形式で記述し，条件部，実行部には，以下の情報を記述する。

条件部
- ・着目単語ごとに着目単語を含む文節，および，前後の数文節の単語の並びを記述
- ・単語情報のすべてが記述できる

実行部

［補正の場合］
- ・正解とする文節，および，単語構成情報

［多義絞り込みの場合］
- ・着目単語の取捨
- ・着目単語の優先度を上げる積極的ルールと優先度を下げる消極的ルール

3.2 ルールの記述法

(1) 単語分割ルール

形態素解析処理の前に入力文字列の字面から単語境界を設定するルールである。条件部には字面を入力し，実行部には適切な位置で区切った表記を入力する。本ルールは，後述する解析補助記号の枠組みを使って実現する。ただし，形態素解析終了時に解析補助記号は出力しない。

入力形式は以下のとおりである。

```
if（単語 表記）then（単語 表記 表記 ...）
```

例１：

```
if（単語 "現代用語"）
then（単語 "現代" "用語"）
```

(2) 誤り回復ルール

形態素解析処理の終了後に，ルールに従って形態素を補正したり，多義を絞り込んだりする処理である。

入力形式は以下のとおりである。

```
if   ((文節 数字（単語 数字 条件式)...）
      (...)...）
then ((文節 数字
        （単語 数字 動作 付加情報...)...）
      (...)...）
```

各要素の意味は以下のとおりである。

文節	文節であることを表す識別子
単語	単語であることを表す識別子
数字	条件部と実行部で文節や単語を対応させるための識別子
条件式	パラメータキーワードによる指定
動作	辞書検索，新規作成，多義優先／非優先，削除を指定。省略可
付加情報	パラメータキーワードによる指定

条件部，実行部には内部に必ず１つ以上の文節を記述し，文節の内部には必ず１つ以上の単語を記述する。文節と単語には，条件部と実行部での対応をつけるために番号を付与する。実行部において動作が記述された場合は，その動作を行う。付加情報が記述された場合は，指定された情報を設定する。

文節や単語の位置を記述するために，以下のキーワードを設定した。

*	０個以上の任意の単語または文節 条件部においてのみ使用可能である
先頭	文節または単語の先頭
最後	文節または単語の最後

誤り回復ルールのルール記述の例を以下に示す。なお，例においては，品詞コードの代わりに対応する品詞名で説明する。

例２：

```
if   ((文節 1（単語 1 表記 = "当初"）
             （単語 最後））
      （文節 2（単語 先頭）
             （単語 2 表記 = "予算"）））
then ((文節 2（単語 先頭）
             （単語 1 品詞 = "連体詞型名詞"）))
```

この例では，文節の最後の単語の表記が「当初」で，次の文節の先頭の単語の表記が「予算」であった場合に，「当初」を２番目の文節の先頭に移動し，その品詞を連体詞型名詞にしている。

3.3 ルール作成支援

3.2節で述べたように，本形態素解析処理の出力結果の詳細な情報を使えば，かなり細かいところまで誤りの回復制御を行うことができる。しかしながら，形態素解析処理を単なるパッケージとして使うユーザにとっては，上述のようなルールを作成するのは，面倒なことである。

そこで，形態素解析処理の一般的な出力である，単語の字面と品詞だけをルールの条件部と実行部に記述するだけで，3.2節で述べたルール形式に変換するルール作成支援の枠組みも作成した。この支援ツールには，ルール文法チェッカも組み込まれており，ルールの誤り箇所を指摘してくれる。なお，もちろん，この作成支援ツールを使ってルールを作成した後で，詳細な制御情報を付け加えることも可能である。

(1) 単語分割ルール

ルールの基本形は，

条件部 = 実行部

である。条件部，実行部とも，単語の表記のみを記述し，実行部の単語表記を '/' で区切る。

表記 = 表記/表記

例３：

入力

```
現代用語 = 現代/用語
```

ルール出力

```
if   (単語 "現代用語")
then (単語 "現代" "用語")
```

(2) 誤り回復ルール

ルールの基本形は，(1)と同様，

条件部 = 実行部

である。条件部，実行部とも，表記，品詞コード，標準表記を記述する。標準表記は省略可能である。また，'|' で文節境界を，'/' で単語境界を記述する。１つの形態素（単語）の記述形式は，

表記（品詞コード [,標準表記]）

である。

例４：

入力

```
洗い(動詞転生名詞)|直(副詞)|す(五段動詞終止形)
=
|洗い(五段動詞連用形)/直す(五段動詞終止形)
```

ルール出力

```
if   ((文節 1 (単語 1 表記="洗い"
                 品詞 = "動詞転生名詞")
            (単語 最後))
      ((文節 2 (単語 先頭)
            (単語 2 表記="直"
                 品詞 = "副詞")
            (単語 最後))
      ((文節 3 (単語 先頭)
            (単語 3 表記="す"
                 品詞 = "五段動詞終止形")
            (単語 最後)))
then ((文節 1 (単語 先頭)
            (単語 1
                 品詞 = "五段動詞連用形")
            (単語 2 (検索 表記="直す"
                 品詞 = "五段動詞終止形"))
            (単語 3 削除)))
```

4 形態素解析の多義

4.1 多義の種類

形態素解析で誤りを引き起こすのは，入力された日本文に多義があり，形態素解析処理においてその多義の選択を誤るのが１つの原因である。まず，その多義の種類について，概観する。

(1) 単語分割の多義

日本文中に漢字文字列，ひらがな文字列などがある場合，どこで単語を分割するかの多義がある。たとえば，「畜産物価格安定法」は，

畜産 産物 物価 価格 格安 安定 定法

のように，連続するどの２文字で分割しても，それらが単独で意味を持つ単語である。さらに，

産 物 価 格 安 定 法

の各１文字は，接頭語または接尾語としても使われるので，この複合語を意味的に正しく分割するのは難しい。

(2) 品詞，標準表記の認定の多義

このタイプの多義については，日経産業新聞965文（情報欄リード文）に対する形態素解析結果に含まれる多義520例を対象として分類し，その中から出現件数の多い10事例を抽出した[6]。これを基に，誤り回復の枠組みの検討を進めた。以下にその抽出事例を示す。

(1) 自動詞，他動詞の区別（例：開く）
(2) 自動詞の連用形，自動詞転生名詞
　　（例：受け）

(3) 他動詞の連用形，他動詞転生名詞
（例：取り付け）
(4) 自動詞転生名詞，他動詞転生名詞の区別
(5) ひらがなに対して，複数の漢字表記があるもの（例：かける → 掛ける,賭ける,駆ける）
(6) 漢字に対して，複数の読みが当てはまるもの（例：開く → ひらく，あく）
(7) 形容詞に対し，述語として使われているか，副詞的に使われているかの区別（例：近く）
(8) 品詞は同じで，意味属性のマッピングが違うもの（例：日（＝ 暦日 / 非暦日））
(9) 一般名詞，形式名詞の区別（例：もの）
(10) 助動詞，格助詞の区別（例：で（だ））

4.2 多義絞り込みルール

多義絞り込みは，実行部の動作で指定する。例を以下に示す。

```
例5：
  if   ((文節 1
          (単語 1
             一般名詞意味属性 ∈ (財産 生命))
          (単語 2 表記 = "を"))
        (文節 *)
        (文節 2 (単語 3 表記 = "かける")))
  then ((文節 2
          (単語 3
             (優先 標準表記 = "賭ける")))

例6：
  if   ((文節 1
          (単語 1
             一般名詞意味属性 ∉ (財産 生命))
          (単語 2 表記 = "を"))
        (文節 *)
        (文節 2 (単語 3 表記 = "かける")))
  then ((文節 2
          (単語 3
             (非優先 標準表記 = "賭ける")))
```

例5では，前方の文節に一般名詞意味属性が『財産』または『生命』に含まれる単語があり，その直後に「を」がある場合に，標準表記として「賭ける」を指定している形態素を優先する。

例6では，逆に，「を」の直前の単語の一般名詞意味属性が『財産』にも『生命』にも含まれない場合に，標準表記として「賭ける」を指定している形態素の優先度を下げる。

5 単語分割多義の解消

単位文や句の開始位置と終了位置を入力文中に明示することにより，構文解析や意味解析における依存関係の解析の手助けとなる手法として，解析補助記号が知られている。ところが，この記号を入れることによって，副作用として形態素の境界を示す効果が生じている。そこで，この副作用を積極的に活用することを考え，独立記号を導入することを考える。

本処理では，形態素解析が認定する文節境界に解析補助記号が現れた場合の扱いとして3種類を想定した。解析補助記号は「透明形態素」であり，これにより単語の分割が変化することはあるが，これを入れたことによりその前後の単語間の連鎖の検定には影響を与えないようにする。

(1) 右側（後方）文節の先頭に置く
（依存関係の開始を示す）
(2) 左側（前方）文節の末尾に置く
（依存関係の終了を示す）
(3) 前後どちらの文節にも含まれずに，独立した文節となる
（単語の境界を示す効果がある）

上記(1),(2),(3)の解析補助記号として，どの文字（記号）を使うかは，形態素解析に先立って指定する。

解析補助記号は，日本語の解析処理が完了した段階で文節構造から削除される。

6 おわりに

単語レベルでの解析正解率99.5%以上の形態素解析処理を対象に，単語の前後関係に着目した誤り回復の外付け処理を付加することにより，形態素解析の精度向上を図る手法を提案し，その処理系を実現した。

今後は，ルールの数を増やし，大量の日本文に対し試験を行い，その効果を確認する予定である。

参考文献

[1] 白井,横尾,池原,奥山,宮崎:多段解析法による日本語形態素解析の精度，第50回情処全大 1R-2 (1995)
[2] 横尾,宮崎,阿部,池原,白井,細井:日英機械翻訳における意味解析のための単語辞書，言語処理学会第3回年次大会 A2-2 (1997)
[3] 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, (1986)
[4] 白井,池原,河岡,上田:日本文書き替え処理における制御ルールの機能別構成，第47回情処全大，6P-4, (1993)
[5] 長尾編: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)
[6] 白井,池原,井上:近接単語の並びに着目した形態素解析多義の絞り込み，第52回情処全大 5B-5 (1996)