# 口語的表現を含む日本語文の形態素解析の実現と評価

1Q－1

竹元義美　福島俊一
(NEC 情報メディア研究所)

## 1 はじめに

形態素解析は、従来、書き言葉中心研究されてきた。しかし、自然言語処理の応用を広げていくためには、口語的表現を含めた広範な文章を扱う必要がある。そのため、話し言葉特有の言い回しを辞書登録して解析する方法がある[1][2]。しかし、口語的表現として話し言葉特有の言い回しだけでなく、2節に示すようなテキスト特有の表記による強調表現も扱う必要がある。

本稿では、口語的表現を含む日本語文の形態素解析について、辞書登録では対応できないテキスト特有の表記による強調表現への対処と評価について述べる。

## 2 口語的表現の形態素解析における問題点

ここでは口語的表現を次の3種類に分けて考える。

- 話し言葉特有の言い回し(例："困っ ちゃう")
- 表記による強調表現
  (例："スミマセン"、"だ ぁー いすき")
- 擬音語・擬態語など(例："ぐげげ～っ")

話し言葉特有の言い回しは、それを整理して辞書登録で対処することができる[1][2][3]。

表記による強調表現は、意図的な片仮名表記によるものと特殊文字を用いるものとがある。これらの強調表現は以下のような問題を引き起こす。

例えば、通常"すみません"と表記するところを著者が意図的に"スミマセン"と片仮名表記したとする。通常は辞書に"スミマセン"は登録されていないから未知語となる。未知語処理では名詞と認定されてしまうことが多い。片仮名列をまとめて名詞とするような未知語処理は、例えば、"コン ピューター"のような誤字を含む単語を判別する能力を持たない。文の理解の観点でも未知語とするよりは辞書中の単語と対応をつけたい。また、通常の表記"だいすき"の中に特殊文字を挿入して"だ ぁー いすき"と表記したとする。このとき特殊文字前後で単語分割に失敗することが多い。

上記のような強調表現のバリエーションは多様であり、話し言葉特有の言い回しのようにすべて辞書登録するのは現実的な策ではない。3節では、表記による強調表現の形態素解析手法を示す。

## 3 表記による強調表現の形態素解析

形態素解析に3.1、3.2節で示す片仮名列置換検索処理と特殊文字置換検索処理とを導入する[3]。形態素解析は、辞書検索→「片仮名列置換検索」→「特殊文字置換検索」→接続検定→候補選択という流れとする。導入する処理は、辞書検索結果としての単語の候補を追加するように働くので、接続検定以降の処理は変更の必要はない。

### 3.1 片仮名列置換検索処理

片仮名列置換検索処理は、辞書から単語が検索されていない片仮名列を平仮名列に置換し、辞書を再検索する処理である。例えば、辞書に"がんば(る)"はあるが"ガンバ(る)"がなかったとしても、"ガンバった"を"がんばった"に直して再検索するので、"ガンバ(る)"という単語の候補が得られる。以下に処理の適用条件・適用範囲などの詳細を述べる。

1. 置換範囲は、片仮名列先頭から最長一致で単語をつないだとき、4文字以上とれない部分とする。
2. 再検索では、置換範囲より前方の字種境界位置(平仮名から漢字や片仮名に変化した位置または句読点の直後など)から置換範囲の末尾までの各文字位置を先頭とする単語を検索する。検索された単語の末尾位置は置換範囲を越えてもよい。
3. 置換範囲の前方より再検索した場合、置換範囲に届かない単語は無効とする(通常検索で既に得られているため)。

### 3.2 特殊文字置換検索処理

特殊文字置換検索処理は、特殊文字("ぁ"、"ぃ"、…、"ぉ"、平仮名列に挿入・追加された"ー"、"～")を削除または置換して辞書を再検索する処理である。

1. 特殊文字を削除して再検索する。
2. 直前の文字との音韻的な規則(お段の直後の長音を"う"に置換するなど)に基づいて特殊文字を置換して再検索する。

---

Implementation and evaluation of a morphological analysis method for colloquial Japanese text

Yoshikazu Takemoto and Toshikazu Fukushima
Information Technology Research Labs, NEC Corp.

3. 再検索は、削除・置換位置より前方の字種境界位置から削除・置換位置までの各文字位置を先頭とする単語を検索する。ただし、削除・置換位置をまたがる単語でない場合は無効とする。

## 4 評価

### 4.1 評価の手法と結果

週刊誌テキストを対象に、片仮名あるいは特殊文字を含む箇所(句読点で区切られた範囲)427件の形態素解析結果を、3節の処理の導入前後について比較・分析した。その結果を表1、2に示す。また、各表では、A.検索された単語またはその組合せでカバーされる箇所と、B.未知語が含まれる箇所とを分けた。2節で述べた考えから、Bは誤りとみなしている。()内の数字は、片仮名や特殊文字を含む箇所のうち、表記による強調表現の件数を示す。なお、強調表現でない特殊文字は、擬音語・擬態語に含まれるものである。

片仮名列置換検索の導入によって、以前は意図的な片仮名表記のために解析に失敗していた箇所84件のうちの63件(75%)が正しく解析できた。特殊文字置換検索の導入では、以前は特殊文字による強調表記のために解析に失敗していた箇所11件のうちの4件(36%)が正しく解析できた。それらの改善例は以下の通りである。

- ゲン(固有名詞)コツ(名詞)→ゲンコツ(名詞)
- バツグン(未知語→形容動詞)
- アブナイ(未知語→形容詞)
- どー(未知語)する→どうする(サ変動詞)

しかし、表記による強調表現にもかかわらず改善されなかった箇所(片仮名: 21件、特殊文字:7件)が残った。さらに、以前は正しく解析されていた片仮名列が片仮名列置換検索の導入によって解析に失敗する副作用も発生した(6件)。

### 4.2 考察

上述の副作用6件は、次の例のように、置換検索処理結果の単語を含む候補が解析結果となってしまったものである。

- ホンダ(固有名詞)の(格助詞)→ホン(名詞)だの(並立助詞)

このような副作用は、候補選択の段階で置換検索で得られた単語の評価値を悪くするような処理によって除去できる。

表記による強調表現にもかかわらず改善されなかった28件の原因分類を示す。

a. 平仮名表記が辞書にないため(17件)
b. 未知語の作成条件の見直しまたは評価値の調整が必要なもの(5件)
c. 片仮名列中の長音を特殊文字とみなしてないため(1件) (例)“キツーイ”
d. 特殊文字が単語の末尾にあるもの(5件)
(例)“いいですよねー”

dは、3.2節の仕様で特殊文字削除位置をまたがる単語のみを有効としたので処理できなかった。単語の末尾についた特殊文字は、直前で検索された単語に含めてしまう処理により改善できる。

以上から、片仮名列置換検索では副作用6件すべてが除去できる見込みである。また、特殊文字置換検索では解析に失敗していた箇所11件のうち、未改善dの5件は容易に除去でき、9件(82%)が対処できるようになる見込みである。

表 1: 評価結果(片仮名)

| | A.検索語 | | | | B.未知語 | |
|---|---|---|---|---|---|---|
| 処理 | 正 | | 誤 | | 誤 | |
| 導入前 | 287(3) | | 24(18) | | 96(66) | |
| 処理 | 正 | 誤 | 正 | 誤 | 正 | 誤 |
| 導入後 | 281(3) | 6(0) | 15(15) | 9(3) | 48(48) | 48(18) |

表 2: 評価結果(特殊文字)

| | A.検索語 | | | | B.未知語 | |
|---|---|---|---|---|---|---|
| 処理 | 正 | | 誤 | | 誤 | |
| 導入前 | 3 | | 0 | | 17(11) | |
| 処理 | 正 | 誤 | 正 | 誤 | 正 | 誤 |
| 導入後 | 3 | 0 | 0 | 0 | 4(4) | 13(7) |

## 5 おわりに

口語的表現を含む日本語文の形態素解析について、片仮名列置換検索処理と特殊文字置換検索処理を導入し、解析精度が向上することを示した。

今後は、より大きなテキストで評価し、文節候補選択時の評価値調整など、考察で述べたような処理を実現してゆく。

### 参考文献

[1] 竹下他,情処42全大1C-3,1991
[2] 荻野,計量国語学第19巻第1号,1993
[3] 竹元他,情処46全大1B-2,1993