

単語連鎖列書き換え規則を用いた形態素解析後処理とその有効性

佐々木 尚人

宮崎 正弘

新潟大学大学院自然科学研究科

1 はじめに

形態素解析システム Maja [1] は、音声処理、機械翻訳など様々な用途を見越して、日本語文を辞書に収録された基本語（短単位）の組み合わせに分割する。音声処理に対しては短単位の分割が望ましいのに対し、機械翻訳に対しては格助詞相当語、複合辞などは、長単位の分割が望ましい。また、他の形態素解析システムでは Maja とは異なる品詞体系、辞書への単語収録単位を持つものもある。

本稿では、Maja で出力される形態素を単語連鎖列書き換え規則を用い、用途に見合った単語の連鎖列を作り出す形態素解析後処理の手法を提案し、その有効性を示す。

2 形態素解析後処理の位置づけ

図 1 に示すように入力された日本語文は Maja により形態素に分割される。次にそれを形態素解析後処理に渡し、単語連鎖列書き換え規則により、用途に見合った単語連鎖列に書き換えたものを構文解析に渡す。

3 単語連鎖列書き換え規則の形式

単語連鎖列書き換え規則の形式を以下に示す。

$$W_{-i}(\text{pos}, \text{sem}, \text{sup}), *, \dots, *, W_{-1}(\text{pos}, \text{sem}, \text{sup}), *, \\ [W_1(\text{pos}, \text{sem}, \text{sup}), \dots, W_j(\text{pos}, \text{sem}, \text{sup})] \\ *, W_{+1}(\text{pos}, \text{sem}, \text{sup}), *, \dots, *, W_{+k}(\text{pos}, \text{sem}, \text{sup}) \\ \rightarrow W_{-i}(\text{pos}, \text{sem}, \text{sup}), *, \dots, *, W_{-1}(\text{pos}, \text{sem}, \text{sup}), *,$$

Postprocessing for Japanese Morphological
Analysis using Word String Rewriting Rules
Naoto Sasaki, Masahiro Miyazaki
Niigata University

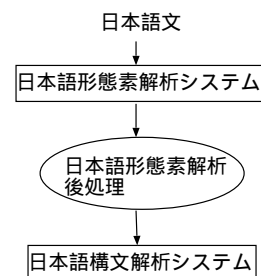


図 1: 形態素解析後処理の位置づけ

$$W_{1'}(\text{pos}, \text{sem}, \text{sup}), \dots, W_{j'}(\text{pos}, \text{sem}, \text{sup}) \\ *, W_{+1}(\text{pos}, \text{sem}, \text{sup}), *, \dots, *, W_{+k}(\text{pos}, \text{sem}, \text{sup}) \\ \{ \}.$$

ここで、 $W_m(\text{pos}, \text{sem}, \text{sup})$ は字面 W_m の単語の品詞が pos、意味属性が sem、補足条件が sup であることを表す。pos は必須の引数、sem、sup は省略可能な引数である。 $W_m(\text{pos}, \text{sem}, \text{sup})$ の代わりに $\langle W_{m1}(\text{pos}, \text{sem}, \text{sup}) / \dots / W_{mn}(\text{pos}, \text{sem}, \text{sup}) \rangle$ と記述することにより、n 個の同形語 $W_{m1}, \dots, W_{mn}$ の OR を表す。[] 内は書き換え対象となる単語連鎖列である。[] の前後にある単語連鎖列は書き換え対象単語連鎖列の前後に共起する単語連鎖列を規定するもので、省略可能である。* は、この部分に長さ 0 以上の任意の単語連鎖列が位置することを示している。{ } は補強項であり、必要に応じて本規則の適用条件を記述する。

本規則により、長さ 1 以上の単語連鎖列 $W_1, \dots, W_j$ の前方の長さ 1 以上の単語連鎖列（構成要素が不連続の離散型でも可）が $W_{-i}, \dots, W_{-1}$ で、後方の長さ 1 以上の単語連鎖列（離散型でも可）が $W_{+1}, \dots, W_{+k}$ で、補強項の条件を満たす場合、単語連鎖列 $W_1, \dots, W_j$ を $W_{1'}, \dots, W_{j'}$ に書き換える。

4 単語書き換え規則の適用

4.1 連語の抽出

3で述べた規則において、長さ2以上の単語連鎖列 $W_1, \dots, W_j$ を一つの単語 $W_{1'}$ に書き換えることによって、以下の例に示すように、短単位語に分割された単語連鎖列を一つの単語にまとめることができる。

- 格助詞相当語：
[に (格助詞), よつ (五段ラ行動詞), て (既定の助動詞)]
→ によって (格助詞).
- 複合辞：
[こと (形式名詞), が (格助詞), でき * (一般動詞)]
→ ことができ * (助動詞).
(* は活用語尾を表すワイルドカード文字)
- それ以外の連語：
[こと (普通名詞), に (格助詞), よる (五段ラ行動詞), と (接続助詞)]
→ ことによると (副詞).

4.2 同形語の判別

3で述べた規則を用いて、長さ1以上の単語連鎖列 $W_1, \dots, W_j$ 中の単語の同形語セット $\langle W_{m1} / \dots / W_{mn} \rangle$ を一つの単語 $W_{m'}$ に絞り込むように書き換えることによって、以下の例に示すように、同形語の前後に共起する単語連鎖列 (連続型 / 離散型) によって同形語を判別することができる。

「鳩は動物で、鳥である。」のように、「で」の直後が読点で、後方に名詞 + 断定の助動詞がある場合、断定の助動詞と判別。

W_{-1} (名詞), [$\langle$ で (格助詞)/で (断定の助動詞)],
 W_{+1} (読点), *, W_{+2} (名詞), W_{+3} (断定の助動詞)
→ W_{-1} (名詞), で (断定の助動詞), W_{+1} (読点), *, W_{+2} (名詞), W_{+3} (断定の助動詞)

「その事故は偶然によるもので故意によるものではない」のように「で」の直前の名詞と同じ名詞 + 断定の助動詞が後方にある場合、断定の助動詞と判別。

W_{-1} (名詞), [$\langle$ で (格助詞)/で (断定の助動詞)], *,
 W_{+1} (名詞), W_{+2} (断定の助動詞)
→ W_{-1} (名詞), で (断定の助動詞), *, W_{+1} (名詞), W_{+2} (断定の助動詞) { $W_{-1} = W_{+1}$ }.

4.3 形態素誤りの修正

3で述べた規則を用いて、以下の例のような特定の単語連鎖列において生じる単語認定や単語区切りの誤りを修正できる。

- 品詞認定の誤り修正
[で (格助詞), も (係助詞), あ * (五段ラ行動詞)]
→ [で (断定の助動詞), も (係助詞), あ * (断定の助動詞)].
- 単語区切りの誤り修正
[残念 (形動名詞), な (助動詞), がら (普通名詞)]
→ [残念 (形動名詞), ながら (接続助詞)]

4.4 他の形態素解析システムの出力への変換

3で述べた規則を用いて、Maja の出力結果を、品詞体系や辞書への単語収録単位の異なる他の形態素解析システムの出力に変換できる。

- Maja 出力の ALTJAWS [2] 出力への変換例
その/本/は/読む/こと/は/でき/ない
→ その/本/は/読む/ことはでき/ない

5 おわりに

本稿では短単位語の組み合わせに分割された形態素解析結果に単語列書き換え規則を適用し、その一部を異なった単語連鎖列に書き換える枠組みを提案した。長単位語である連語としてまとめるだけでなく、同形語判別や形態素解析誤りの修正も行うことができる。またこの手法により Maja の出力を他の形態素解析システムへの出力へ変換する機能を持たせることも期待できる。

参考文献

- [1] 尾嶋、宮崎：日本語形態素解析システムにおける部分的再試行機構の導入とその効果、情報処理学会第58回全国大会 1E-4(1999)
- [2] 白井、横尾、松尾、大山：日英翻訳のための日本語解析技術、NTT R&D、Vol.48、No.12、pp.1399-1404(1997)