2H--7

EDR 日本語単語辞書を用いて 形態素解析と統語解析を行なうシステム[1]

植木 正裕　徳永 健伸　田中 穂積[2]
東京工業大学情報理工学研究科計算工学専攻[3]

1 はじめに

従来の自然言語解析システムではそれぞれ独自の辞書や文法を用いていたが、これらの資源の開発には多大なコストがかかるため、研究者間の資源の共有が求められている。辞書については、EDR 日本語単語辞書 [3] と呼ばれる大規模な電子化辞書がすでに開発されている。ところが、この辞書を対象とした文法はまだ十分に整備されていない。本研究では EDR 辞書を対象とした日本語文法の作成し、これを用いた形態素解析と統語解析を行なうシステムの開発を目的とし、研究を行なった。本稿では、その概要を述べるとともに、作成した文法を用いての解析実験の結果を報告する。

また、実験過程で、EDR 辞書にもまだ不備と思われる点があることがわかったので、それらについての対処方法を述べる。

2 EDR 辞書について

EDR 辞書は、基本単語 20 万語・専門単語 10 万語を有する大規模辞書である。各単語のエントリーには、品詞・読みや連接属性などの情報が与えられている。

連接属性とは、各単語がどのような単語と隣合わせになり得るかということを表すもので、連接規則によって接続可能性が定義されている。さらに、例えば接尾語の場合、連接属性を見ることで、名詞接尾語・動詞接尾語などの区別ができる。

3 文法

2 節でも述べたように、連接属性を見ることで品詞をより細かく区別することができる。そこで、それを文法に反映させることにした。

好き	
左連接属性	右連接属性
名詞接尾語	形容動詞ダ活用 3

これは接尾語「好き」の例である。「好き」は名詞接尾語なので、「読書」(左右連接属性 = サ変名詞) との連接が可能である。「読書」と「好き」が接続してできる「読書好き」の連接属性は次のようになる。

読書 好き	
左連接属性	右連接属性
サ変名詞	形容動詞ダ活用 3

また、同じ名詞接尾語でも次のような例もある。

泥沼 化	
左連接属性	右連接属性
普通名詞	サ変名詞

このように、同じ品詞でも働きの異なるものがあるので、それを文法にきちんと記述をした。

4 辞書の不備への対処

作成した文法で実際に解析実験を行なったところ、EDR 辞書の不備と思われる点がいくつか見つかった。

1. 連接規則に誤りがある
2. ひらがなと漢字のような表記の違いに対応しづらい
3. 登録されている固有名詞の数が少ない

本稿では 1 についてのみ説明する。詳細は [2] を参照してほしい。

4.1 連接規則の誤り

東欧		訪問	
左連接属性	右連接属性	左連接属性	右連接属性
固有名詞	固有名詞	サ変名詞	サ変名詞

これは複合名詞の例であるが、「東欧訪問」は実際に可能であるにもかかわらず、連接規則では連接が不可能となっている。そこで、このようなものをコーパスから自動的に抽出して、それをもとに連接規則を修正することを考えた。

使用したコーパスは EDR コーパスである。EDR コーパスは次のような特徴がある。

[1] a system to integrate morphological and syntactic analysis using EDR japanese basic word dictionary
[2] Masahiro UEKI, Takenobu TOKUNAGA, Hozumi TANAKA
[3] Tokyo Institute of Technology, Graduate School of Information Science & Engineering, Department of Computer Science

- 文が形態素に区切られている
- 形態素情報から連接属性が特定できる

そこで、EDR コーパスから連続する 2 単語の連接属性の組を取りだし、それが一意に決まったものについて連接規則と照らし合わせてみた。その結果、連接規則では連接が不可能な連接属性の組が 455 組発見された。しかし、コーパス作成は人手で行なわれているため入力ミスなどもある。そこで、得られた結果のうち、特に頻度の高かった 50 組について、入力ミスなどによるものを人手で除去し、残った 13 組について連接規則を修正した。

5 実験と評価

5.1 評価方法

システムの評価に用いた日本語文は EDR コーパスよりランダムに選んだ 100 文である。文の長さは 18 文字から 71 文字、平均 36.4 文字であった。パーザには、GLR 法を用い、形態素解析と統語解析を統合した MSLR パーザ [1] を用いた。解析精度を評価するにあたっては、スコアづけを行ない、スコアが 1 位の木のみに関して次のような基準で評価した。

- 形態素の区切り・品詞がすべて正しいか
- 文節の区切りがすべて正しいか

5.2 実験結果と考察

連接規則の修正と辞書引き部での処理によって、解析可能な文の数は 50 から 79 に、そして形態素区切り・品詞の正解数も 36 から 55 に増加している。特に、固有名詞の処理が効いていると思われる。しかし、それでも全体の 2 割について解析結果が得られていない。これは次のような理由によると考えられる。

- 文法の不足
- 連接規則の間違い
- 辞書引きの失敗 (ひらがなと漢字の混ぜ書きなど)

ちなみに、EDR 辞書を juman 用の辞書に変換することで juman でも EDR 辞書の情報が利用できる。EDR 版 juman では 100 文すべての解析が可能だが、形態素区切り・品詞共すべて正解のものは 30 文であった。

6 まとめ と 今後の課題

本研究では、連接情報を有効に利用した文法を作成し、辞書の不備に対して対処を行ない、EDR 辞書を用いて形態素・統語解析を行なうシステムを構築した。解析結果の得られないものが約 2 割あるが、正解率は入力の 5 割、解析結果が得られたものの 7 割であった。

今後の課題としては、

1. 辞書の整備
2. 統語解析失敗からの回復方法の検討
3. 実験の大規模化による、本システムの有効性の検証

などが挙げられる。

参考文献

[1] 伴光昇, 福田譲, 白井清昭, 田中穂積. 圧縮統語森上での形態素解析候補の絞り込み -品詞列統計情報の利用-. 1994 年度 人工知能学会全国大会 (第 8 回) 論文集, pp. 527–530, 6 1994.

[2] 植木正裕. EDR 辞書を用いて日本語文の形態素解析と統語解析を行なうシステム. EDR 電子化辞書利用シンポジウム論文集, pp. 33–39, 7 1995.

[3] 日本電子化辞書研究所. EDR 電子化辞書利用マニュアル, 第 2.1 版, 1994.

連接規則の修正	×	×	×	×	○	○	○	○
ひらがなのヒューリスティクス	×	×	○	○	×	×	○	○
未登録語の処理	×	○	×	○	×	○	×	○
解析可能文数	50	66	61	75	52	69	64	79
形態素区切り・品詞共全正解文数	36	50	41	53	38	52	43	55
文節区切り全正解文数	36	49	42	54	40	52	45	56

表 1: 実験結果