EDR電子化辞書利用シンポジウム論文集 1995・7・17.

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

植木 正裕　　徳永 健伸　　田中 穂積

東京工業大学情報理工学研究科計算工学専攻

0　はじめに

当研究室では一昨年より、EDR 日本語単語辞書 (以下 EDR 辞書) を用いて形態素解析と統語解析を行なう、一般化 LR 法に基づいた MSLR パーザシステムの研究 [7][2] を行なってきた。しかし、EDR 辞書の情報が文法に十分には反映されてない、後に続く意味解析のことも考慮されていないという問題があった。

前者は、EDR 辞書の連接属性が単に単語間の連接可能性を調べるためだけに用いられていて、各属性そのものの意味が活かされていないという問題である。連接可能性を調べるためだけに用いるのであれば、連接属性は A とか B といったラベルでかまわない。しかし実際には、普通名詞・サ変名詞といったようにそれ自体が意味を持つラベルがついている。そこで、このような情報をもっと有効に活用できないかということが考えられる。

後者に関しては、日本語解析では形態素解析→統語解析→意味解析という流れで解析が進む。形態素解析では動詞の語幹や語尾といったように形態素という単位を扱う。一方係り受け解析[1]では、文節を 1 つの単位として解析を行なう。動詞の連用形であれば連用修飾可能というように必要とする情報も文節単位である。伴光による研究 [7] では形態素解析と統語解析を統合した解析が行なわれたが、統語解析した結果の木は文節がひと塊にはなっていない。

そこで本研究では、これらの問題を解決するために、MSLR パーザで用いる新しい文法の試作を以下の点に留意しつつ行なった。

- 連接属性の情報を文法に反映させる
- 係り受け解析のことを考慮し、形態素列を文節にまとめる文節文法を作成する

EDR 辞書の不備に対する対処もいくつか行なったので、これらについても 2 節で説明する。

[1]係り受け解析は、統語解析と意味解析の中間的なものと我々は考える。

1　文法

1.1　MSLR パーザのオリジナルの文法

MSLR パーザでは、CFG 形式の文法から作った LR テーブルをもとに解析を行なう。まず EDR 辞書を引いて候補となる単語の切り出しを行なうが、その際に品詞と左右の連接属性のセットを考え、細品詞というカテゴリーに分割し、細品詞を先読みとして解析を行なう。

表記	辞書	解析
品詞	名詞	名詞
左連接属性	普通名詞	サ変名詞
右連接属性	普通名詞	サ変名詞

福田は EDR 辞書を対象とした文法 (以下、福田文法) の作成を行なっている [2] が、福田文法には次のような問題点がある。

1. ルールで細品詞をまとめる際に、細品詞の持っている情報が失われる。
   - 普通名詞・サ変名詞などが連続した場合、すべてをただつなげて複合名詞としてしまうので、その複合名詞の中にサ変名詞のような特殊な名詞が含まれているという情報が失われる。

- 接尾語には、名詞接尾語・動詞接尾語などかなりの種類がある (細品詞数は 53) が、それらを細品詞規則によってすべて同じ品詞にまとめている。

2. 統語解析結果を作りはするが、係り受け解析のことを考慮していないために、解析の結果が文節単位になっていない。

1.2 試作した文法

1.1 に述べたように、EDR 辞書を用いる際に、品詞と連接属性を組み合わせた細品詞という細かいカテゴリーを用意した。ここで、連接属性は単なる連接可能性をチェックするためのものではない。例えば以下の例を考えてみる。

1. 工場資材

2. 工場見学

これらはどちらも 2 つの名詞からなる複合名詞である。しかし、1 は「工場」・「資材」ともに普通名詞であるのに対して、2 の「工場」・「見学」の「見学」はサ変名詞である。したがって、「工場見学」の後に「する」という語尾をつけることができる。

普通名詞もサ変名詞も区別せずに扱う場合、複合名詞全体を普通名詞として扱い、

- 普通名詞＋「する (動詞)」と分けて考える。
- 普通名詞にサ変の語尾がつくというルールを追加する。

というの 2 つの方法が考えられる。しかし、サ変名詞＋「する」のサ変動詞の連用形を考えてみると、サ変名詞＋「し」以外にサ変名詞単独でも連用形となる[2]。そのため、「工場見学」＋「、」のようにサ変名詞の直後に読点があった場合、名詞の並列だけでなく動詞の並列となる可能性もある [1]。先ほどの 2 つの解決策ではどちらも複合名詞全体を普通名詞として扱っているので、直後に読点が来ても名詞の並列としか解釈できなくなってしまう。したがって、「工場資材」と「工場見学」はきちんと区別しなくてはならない。そこで文法に次のような規則を考える。

普通名詞 → 名詞列 普通名詞.
サ変名詞 → 名詞列 サ変名詞.

これによって、普通名詞＋「、」の文節とサ変名詞＋「、」の文節とでは文節の属性が違うことを陽に表すことができる。

同じサ変名詞の例としてさらに次を考えてみたい。

泥沼化する

これを形態素レベルで考えれば、「泥沼 (名詞)」＋「化 (接尾語)」＋「する (語尾)」となる。ここで注目したいのは接尾語「化」である。接尾語の場合、どのような品詞に後続するかで以下のように分類することができる。

名詞接尾語	(日本) 製
動詞接尾語	(走り) 方
形容詞・形容動詞接尾語	(うれし) さ

さらに、接尾語がついたことでその単語全体がどのような新しい品詞に変化したかを考えると以下の表になる。

もとの品詞	新しい品詞	例
名詞	サ変名詞	(泥沼) 化 (する)
名詞	形容動詞	(読書) 好き (だ)
名詞	動詞	(天才) ぶ (る)
動詞	名詞	(走り) 方
動詞	サ変名詞	(書き) 初め (する)
動詞	時相名詞	(帰り) ぎわ
形容詞	名詞	(うれし) さ
形容詞	形容動詞	(厳し) め (だ)
形容動詞	名詞	(静か) さ

この表によれば、「泥沼化する」は「泥沼 (名詞)」＋「化」⇒「泥沼化 (サ変名詞)」となる。「泥沼」と「化」の左右連接属性は以下のようである。

[2]「称する」のように語幹が名詞とならないサ変動詞の連用形は「称し」だけ。

泥沼		化	
普通名詞	普通名詞	名詞接尾語	サ変名詞

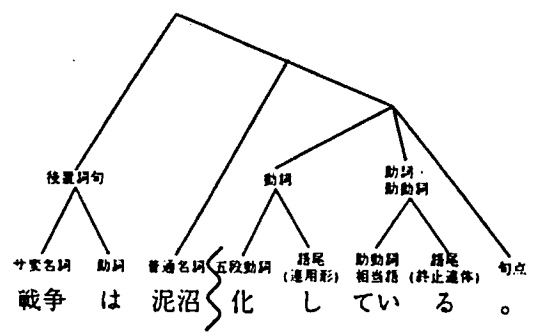
「化」は名詞接尾語であるから名詞「泥沼」と連接可能である。そして、「泥沼化」全体の連接属性は「泥沼」の左連接属性(普通名詞)と「化」の右連接属性(サ変名詞)をそれぞれ左右の連接属性として持つと考えられることから、

泥沼化	
普通名詞	サ変名詞

となり、サ変名詞とほとんど同じように扱ってかまわない。

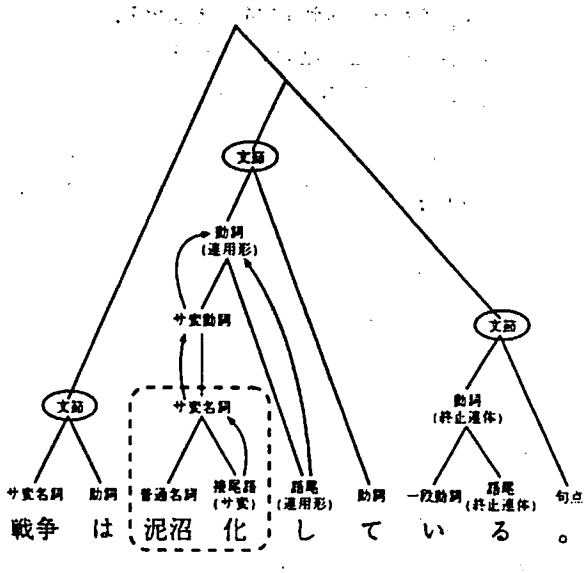
「工場見学」と「泥沼化」の例でわかるように、(接尾語も含む)複合名詞では、サ変名詞のように普通名詞より特徴的なものがある場合、その複合名詞(さらにはその文節)の属性はそのような特徴的な属性を引き継ぐと考えることができる。ただし、これは特徴的なものが複合名詞の最後に来た場合で、「工場見学中」のようにさらに名詞や接尾語が続く場合はあてはまらない。これは先ほどの連接属性を用いた考え方からも明らかである。

実際の解析例を見てみたい。



これは福田文法による解析例である。この例では「泥沼化」を1つの名詞とした場合は解析が失敗するらしく、「化する」を動詞として扱ってしまっている。そのため「泥沼」と「化」の間で木が大きく分かれたような構造になっている。

それに対し、今回試作した文法では、



という解析木が最良のスコアを持ち、「泥沼」と「化」では「化」のサ変名詞としての属性が残り、さらに「し」と一緒になって動詞の連用形の属性が加わる。最終的に、「戦争は」「泥沼化して」「いる」という文節に分かれ、「泥沼化して」の文節は動詞の連用形の属性を持つことが分かる。(この例では「泥沼化」にさらに語尾がついているため、サ変名詞であることは最終的には意味を持たない。)

2 辞書の不備への対処

1節で述べたようにして試作した新しい文法を用いて解析を行なったところ、EDR 辞書の不備と思われる点がいくつか発見された。

- 品詞間の連接可能性を表す連接規則に誤りがある
- 登録されている固有名詞の数が少ない
- ひらがなと漢字のような表記の違いに対応しづらい

この節ではこれらに対する対処について述べる。

2.1 連接規則の誤り

新しい文法ではＡＢという単語の並びが可能でも、Ａの右連接属性とＢの左連接属性を調べると連接不可能なため解析に失敗する例が見つかった。

中曽根首相の東欧訪問、倉成外相の
南太平洋諸国歴訪は、ソ連をにらん
だ日本の外交攻勢として伝えられた。

これは EDR コーパスからの例であるが、「東欧」「訪問」という複合名詞の部分について連接属性のつながりを見てみると、

東欧	訪問
固有名詞	サ変名詞

となっている。ところが、連接規則では固有名詞とサ変名詞は連接が不可能となっている。このように実際は連接できるにも関わらず、連接規則では連接が不可能となっているものがあった。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。

EDR コーパスから、連続する単語の組をすべてとりだし、連接規則と照らし合わせてみた。その結果、連接規則では接続が不可能な連接属性の組が 455 組発見された。しかし、コーパス作成の際の入力ミスなどもあるので、それらを人手で取り除き、残った 13 組について連接規則を修正した。修正例を以下に挙げる。

固有名詞	サ変名詞	東欧 訪問
普通名詞	形容詞	信心 深(い)
接尾語(単位)	数字	(1)基1 4 (億円)

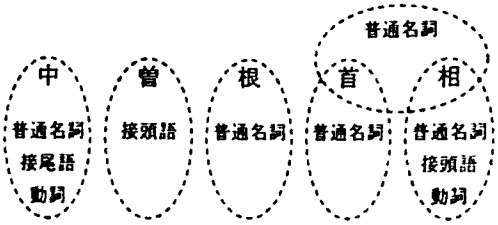
2.2 固有名詞の扱い

EDR 辞書は基本単語辞書ということで、国名などは入っているが人名などの固有名詞がほとんど含まれていない。固有名詞の表記は、漢字・カタカナ・アルファベットのものが大部分である。本システムでは漢字表記の固有名詞の扱いを考慮した。

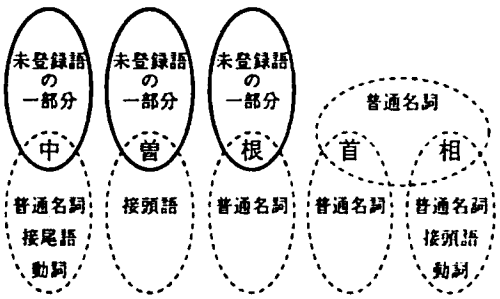
例えば、次の例を考える。

中曽根首相は概して財界では人気が
ある。

この例では、「中曽根」が辞書に登録されていない。「中曽根首相」の部分の辞書引き結果を見ると次のようになっている。



「首相」の部分は 1 文字ずつのエントリーの他に「首相」という 2 文字のエントリーも見つかっている。しかし、「中曽根」の部分は 1 文字ずつのエントリーしか見つかっていない。そこで、漢字列中で漢字 1 文字のエントリーしか見つからなかった部分について、次のように未登録語の一部の可能性があるというラベルをつけることにした。



これによって通常の辞書引きだけではうまくいかなかった固有名詞についてもある程度解析ができるようになった。

2.3 ひらがな表記の扱い

固有名詞と同じく辞書のエントリーの問題として、ひらがな表記の単語の問題がある。EDR 辞書には漢字表記のエントリーは多いが、ひらがな表記のエントリーは少ない。しかし、普段文章を書く場合に、漢字表記のあるものをすべて漢字で書くとは限らない。

ついでに言えば、全く同じ記事でも、
見出しやレイアウトを変えてある。

この例は文法的にも正しく、どこにも不自然なところはない。しかし、この文は解析に失敗する。実は「ついでに」というエントリーがないのである。EDR 辞書を引くと「序(で)に」というエントリーしかない。

EDR 辞書には読みに関する情報もある。「序(で)に」の読みは「ついでに」である。これはまさにひらがな表記と等しい。そこで、MSLR パーザでは

ひらがな部分に関しては、表記だけでなく読みでも辞書を引くようにしている。しかし、ひらがな部分をすべて読みとして引いた場合、辞書引き結果が多くなる。漢字表記があるものについて、ひらがな表記のされやすさを考えてみると次のようになると思われる。

普通名詞 < 動詞・副詞・形式名詞 etc

そこで、ひらがな表記されやすいものを候補として優先するようにした。これにより、ひらがな表記された場合でも解析できるようになったが、「せん光[3]」のような混ぜ書きがうまくいかない。

3 実験と評価

本研究では、係り受け解析を考慮して、文節とその属性に注目した文節文法の試作を行なった。この節では解析実験の結果とその評価を述べたい。

3.1 評価方法

今回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。

解析精度を評価するにあたっては、

1. 文節の区切りが正しいか
2. 形態素の区切りが正しいか
3. 形態素に正しい品詞がふられているか

を基準とした。ただし、複合名詞をどこでいくつの名詞として区切るかは、意味解析なしには決められないので、複合名詞内では 2 の基準は適用しない。また、「買い入れ」のように動詞の連用形が名詞となるもの、「かゆみ」のように形容詞語幹に語尾がついて名詞となっているものなどについては、名詞となっていても、動詞あるいは形容詞＋語尾となっていても正解とし、2 や 3 は厳密には適用しなかった。

さらに、本システムとの比較のために、

1. 福田文法を用いた MSLR パーザ
2. EDR 版 juman

の 2 種類についても実験を行なった。

1 は福田による MSLR パーザ用の文法で、EDR 辞書を対象として作られている。本研究では、この文法の問題点を改善することが目的の 1 つでもあったので、文法の能力の評価のためにこれと比較した。

また、juman は、通常は juman 専用の辞書を用いて解析を行なうが、EDR 辞書をあらかじめ juman 用の辞書と同じ形式に変換することで、辞書引き部分に EDR 辞書を用いることができる。EDR 辞書を用いたシステムとしての評価のためにこれを用いた。

3.2 実験結果と考察

juman は、解析結果として候補を 1 つしか出力しないため、本システムも複数のスコアづけを行ない、もっともスコアの高いもの 1 つを選んで、それを評価に用いた。

表 1 が実験結果である。形態素区切り・品詞の正解数は、1 文中のすべての形態素について区切りと品詞が正解だった文の数を表している。文節区切りの正解数も同様である。

同じ EDR 辞書を用いていても juman ではすべての文に対して解析結果が得られているのに対して、本システムでは約 3 割が失敗している。これは juman では形態素解析のみを行ない統語解析は行なっていないので、正しい形態素の並びでなくとも連接可能であれば解析に成功するためである。また、形態素解析に失敗した部分を未定義語として処理もする。本システムでは、文節間の係り受けなどについては決めていないものの、文節をひとまとめにするように解析木を作っている。その過程で解析に失敗することがある。解析に失敗する理由としては、

- 品詞の連接情報が間違っている部分があると、reduce に失敗する。
- 漢字とひらがなというように、辞書と入力文で表記が異なるために、辞書引きですでに失敗している

が挙げられる。これについては 2.1 や 2.3 で対処法を述べたが、まだうまくいかないものも多い。

[3]漢字表記は「閃光」

	本システム 実験 1[a]	本システム 実験 2[b]	本システム 実験 3[c]	福田 文法	EDR 版 juman
解析可能	74	76	87	68	100
形態素区切り・品詞正解	49(66.2%)[d]	51(67.1%)	73(83.9%)	10(14.7%)	30(30.0%)
形態素区切り 1 つ誤り	12	12	7	10	14
品詞 1 つ誤り	6	6	1	6	19
文節区切り正解	46(62.2%)[e]	49(64.5%)	79(90.8%)	—	—
1 ヶ所切れていない	9	9	3	—	—
1 ヶ所切れすぎている	9	8	2	—	—

[a]連接規則修正前

[b]連接規則修正後

[c]固有名詞なし

[d]形態素区切り・品詞正解 / 解析可能

[e]文節区切り正解 / 解析可能

表 1: 実験結果

しかし、解析の正解率としては EDR 版 juman を凌いでおり、解析が終了した文のうち約 7 割で形態素の区切り・品詞が完全に正しい結果が得られ、EDR 辞書を用いるシステムとしては優れた精度を示した。

文節区切りに関しては、juman ではそのような出力ができないので、本システムのみでの評価となるが、形態素区切り・品詞の正しかったものについては、ほとんどのものが文節区切りも正しい結果が得られた。

実験 2 は、実験 1 と同じ文を 2.1 で修正した連接規則を用いて解析したものである。

また、EDR 辞書は人名などの固有名詞を含まないので、参考として固有名詞を含まない文での実験も行なってみた。(実験 3) 固有名詞を含まない文ではかなり精度が上がっている。失敗したものにはひらがな表記の単語を含むものが多かった。

4 まとめ

本研究では、EDR 辞書を用いて、

- 連接情報を有効に利用した文法を用い、
- 係り受け解析への利用しやすさを考慮した

形態素・統語解析システムを構築した。

解析の終了しないものが 2 割強あるが、正解率は入力の 5 割、解析が終了したものの約 7 割で、EDR 辞書を用いるシステムとしてはよい結果を示した。

しかし、まだ以下の問題がある。

1. 途中で解析木がなくなると解析全体が失敗する
2. ヒューリスティックによって正解が落されてしまう
3. 連接情報だけでは扱えない例外がある

2 に関しては、名詞と副詞のように品詞の曖昧性がある場合には文節の区切れも曖昧になって、正解が 2 つ以上得られるはずであるが、ヒューリスティックにより枝刈りを行なうと正しいものまで枝刈りされてしまうことがある。

3 の例としては「お勧めする」を挙げておく。「勧めする」は文法的に正しくないが、「お勧めする」なら正しい。しかし、「お」と「勧め」あるい

は「勧め」と「する」のように隣合う形態素の間の関係だけでは説明できない。あくまでも「お～する」と呼応する形になってはじめて文法的に正しくなるのである。

これらの解決は今後の課題である。

参考文献

[1] 秋山典丈. 形態素解析で残る曖昧性を考慮した日本語文の係り受け解析. 修士論文, 東京工業大学, 1995.

[2] 福田譲. 日本語文法開発に関する研究. 卒業論文, 東京工業大学, 1994.

[3] 亀田雅之. 簡易日本語解析系 Q_J P. 情報処理学会 自然言語処理研究会, Vol. 94, No. 4, pp. 25–32, 3 1993.

[4] 木谷強. 固有名詞の特定機能を有する形態素解析処理. 情報処理学会 自然言語処理研究会, Vol. 90, No. 10, pp. 73–80, 7 1992.

[5] 益岡隆志, 田窪行則. 基礎日本語文法 –改定版–. くろしお出版, 1992.

[6] Tanaka Hozumi, Tokunaga Takenobu, and Aizawa Michio. Integration of morphological and syntactic analysis based on lr parsing algorithm. 自然言語処理, Vol. 2, No. 2, pp. 59–74, 4 1995.

[7] 伴光昇. 形態素解析と統語解析の統合処理システムに関する研究. 修士論文, 東京工業大学, 1994.

[8] 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.