TR-S-0018

形態素・構文解析用ツールキットMSLRの中国語解析への適用実験

Experiments of Japanese morphological and syntactic analyzer MSLR for Chinese parsing

宮木 衛
Mamoru Miyaki

山本 和英
Kazuhide Yamamoto

坂本 仁
Masashi Sakamoto

2001.2.23

中国語構文木コーパスからMSLRパーザを使用するのに必要な文法、辞書、接続表の各データを抽出するツールを作成した。実際にそれらのデータをMSLRパーザに適用し、中国語文を解析する実験を行なった。さらに、中国語構文木コーパス、MSLRパーザの問題点について考察を行なった。

目次

1 はじめに

本研究では、中国語構文木コーパス (Penn Chinese Treebank, 以下、 PCT) から MSLR 形態素・構文解析器を使用するのに必要な文法規則 (以下、文法)、単語辞書 (以下、辞書)、品詞接続表 (以下、接続表) の各データを抽出し、実際に MSLR の適用を試みて、適用可能であるかを確認し、問題点について検討することを目的とする。

まず、本研究では PCT から MSLR パーザを使用するのに必要な文法、辞書、接続表の各データを抽出するツールを作成した。そして、実際にそれらのデータを MSLR パーザに適用し、中国語文を解析する実験を行った。

最後に、 PCT、 MSLR パーザの問題点について考察し、報告する。

2 用語解説

2.1 中国語構文木コーパス (PCT)

中国語構文木コーパスは University of Pennsylvania(UPENN) の The Chinese Treebank Project が作成したものである[1] 。約 10 万語の中国語 (北京標準語) に対して、単語分割、品詞付与、構造付与を行っている。 PCT は中国国営通信社の新華社通信 (Xinhua News Agency) が発信した 325 記事から作成されている。中国、台湾、香港、シンガポールと米国の研究者から合意が得られるような情報付与を目指している。 2000 年 12 月 4 日に最終版が完成。 LDC(Linguistic Data Consortium) を通じて配布が開始された。現在、さらに新たなデータに対して情報付与を行っている。詳しくは [1][2]を参照していただきたい。なお、 PCT の一部を付録として巻末に添付する。

2.2 MSLR

MSLR の正式名称は「日本語形態素構文解析器 MSLR パーザ・ツールキット」である。これは、東京工業大学大学院情報数理工学研究科計算工学専攻田中・徳永研究室で開発された MSLR パーザ (Morphological and Syntactic LR parser) およびその関連ツールである。このパーザは、主に日本語を対象とし、形態素解析と構文解析を同時に行うパーザであり、形態的な情報と構文的な情報を同時に用いて解析を行うことができるため、形態素解析・構文解析を単独で行う場合に比べて精度が向上すると考えられている。また、構文解析のための知識として文脈自由文法 (CFG) を、形態素解析のための知識として接続表 (文法の前終端記号間の接続可能性を記述した表) を独立に記述することができるため、接続可能性を考慮して文法を記述する必要がないので、必要な文法規則の数を少なく抑えることができることが特徴である。詳しくは文献 [2][3][3] を参照していただきたい。

3 ツール解説

各データは、すべて PCT から抽出したものである。また、記述方法 (書式) は MSLR パーザの説明書に従ったものである。各々データを抽出するプログラミング言語には Perl を使用している。なお、辞書と接続表は 1 プログラムで、文法は 2 プログラムで抽出を行っている。

3.1 文法

文法は品詞を終端記号とする文脈自由文法で記述される。主に構文解析に用いられている。

[1] UPENN は他にも英語、韓国語の Treebank も作成している。
[2] http://www.ldc.upenn.edu/ctb/
[3] http://tanaka-www.cs.titech.ac.jp/pub/mslr

3.1.1 抽出方法

文法の抽出は 2 段階に分けて行っている。まず、文法を抽出する前の段階の処理を以下の方法で行った。

```
( (IP-HLN (NP-SBJ (NP-PN (NR 上海)
 (NR 浦东))
  (NP (NP (NN 开发))
      (CC 与)
      (NN 法制)
      (NN 建设)))
  (VP (VV 同步))) )
```

1. 最初に出現する括弧中の品詞と単語を、 am**(** は数字) に置き換える。(括弧も置き換える)

2. 次に出現する括弧中の文字列が品詞と単語の組み合わせであれば、 1. と同様の処理を、括弧中の文字列が品詞と am**(複数もあり得る) の組み合わせであれば、品詞と am** を am** に置き換える。(括弧も置き換える)

3. 1. 2. の処理を繰り返し、置き換える文字列がなくなった時点で、全ての am** の内容 (置き換えた前の文字列) を出力する。なお、単語に "-NONE-" が存在する行は、この処理を無視している。具体的には、これらの処理を行う前に、 "-NONE-" に関わる全ての単語、品詞を削除している。例を以下に示す。

削除前

...(AS 了)(NP-OBJ (CP (WHNP-1 (-NONE- *OP*))(CP (IP (NP-SBJ (-NONE- *T*-1))(VP (VV 涉及)...

削除後

...(AS 了)(NP-OBJ (CP (CP (IP (VP (VV 涉及)...

ここまでを、 1 つのプログラムで処理している。以下に、これらの処理がどのように行われているかについて、上の例を利用して、その一部を以下に示す。

1. ((IP-HLN (NP-SBJ (NP-PN (NR 上海)(NR 浦东))(NP (NP (NN 开发))(CC 与)...
2. ((IP-HLN (NP-SBJ (NP-PN am1 (NR 浦东))(NP (NP (NN 开发))(CC 与)...
3. ((IP-HLN (NP-SBJ (NP-PN am1 am2)(NP (NP (NN 开发))(CC 与)...
4. ((IP-HLN (NP-SBJ am3 (NP (NP (NN 开发))(CC 与)...
5. ((IP-HLN (NP-SBJ am3 (NP (NP am4)(CC 与)...
6. ((IP-HLN (NP-SBJ am3 (NP am5 (CC 与)...

```
am1   (NR 上海)
am2   (NR 浦东)
am3   (NP-PN am1 am2)
am4   (NN 开发)
am5   (NP am4)
        :
        :
```

次に出力されたデータに対して以下の処理を行い、文法の抽出を行う。

1. 括弧の中に非終端記号があるものを探し、あれば、その非終端記号を文法の左辺に出力する。

2. その行において、括弧の中にある am** をそれ以前の行から探し出し、それに対応する終端記号あるいは非終端記号を文法の右辺に出力する。

3. 括弧の中が am** のみである場合は、文の一番始めの文法であるので、左辺には<BUN> を出力する。その他の処理は 2. と同じである。

4. 全ての行が終わるまで、 1. 2. を繰り返す。

以下に、これらの処理がどのように行われているかについて、上記で出力されたデータを用いて示す。

```
am1   (NR 上海)
am2   (NR 浦东)
am3   (NP-PN am1 am2)
am4   (NN 开发)
am5   (NP am4)
am6   (CC 与)
am7   (NN 法制)
am8   (NN 建设)
am9   (NP am5 am6 am7 am8)
am10   (NP am3 am9)
am11   (NP-SBJ am10)
        :
        :

<NP-PN> --> NR NR
<NP> --> NN
<NP> --> <NP> CC NN NN
<NP> --> <NP-PN> <NP>
<NP-SBJ> --> <NP>
        :
        :
```

なお、プログラムの内容、抽出例は付録として巻末に添付する。

3.1.2 抽出後の処理

文法を抽出した後、実際にMSLRパーザに適用する前に以下の処理を人手で行う必要がある。

1. 抽出した文法は重複しているものが多数あるので、これらの重複をを削除。
2. MSLRパーザが示す書式に従って記述した<start> --> <BUN>（この文法は、抽出をする際に、プログラムで自動的に追加される）が文法中のどこかに存在する。この文法は必ずファイルの1行目に記述しなければならないので、その文法を削除し、ファイルの1行目に移動。

3.2 辞書

辞書は、単語とそれに対応した品詞を列挙したデータで、形態素解析の基本単位を集めたものである。辞書の品詞体系は文法の品詞体系と一致していなければならない。

3.2.1 抽出方法

```
( (IP-HLN (NP-SBJ (NP-PN (NR 上海)
 (NR 浦东))
  (NP (NP (NN 开发))
      (CC 与)
      (NN 法制)
      (NN 建设)))
  (VP (VV 同步))) )
```

PCTでは、1行に1つの単語が記述されているため、単語の部分(上の例では上から順番に、上海、浦东、开发、与、法制、建设、同步)と品詞の部分(上から順番にNR、NR、NN、CC、NN、NN、VV)を抽出し、辞書に記述する方法をとっている。また、同じ単語がすでに記述されている場合は、品詞との組み合わせが異なる場合に、その品詞が記述される。また、単語欄"-NONE-"となっている場合は単語ではないので、単語として記述しない。なお、プログラムの内容、抽出例は付録として巻末に添付する。

3.3 接続表

接続表とは、品詞間の接続制約を記述した表である。品詞間の接続制約とは、ある2つの品詞が隣接できるか否かに関する制約である。

3.3.1 抽出方法

```
( (IP-HLN (NP-SBJ (NP-PN (NR 上海)
 (NR 浦东))
  (NP (NP (NN 开发))
      (CC 与)
      (NN 法制)
      (NN 建设)))
  (VP (VV 同步))) )
```

1行に1つの単語(品詞)が記述されているため、1番目の品詞(上の例ではNR)と2番目の品詞(上の例ではNR)に着目し、1番目の品詞列に後続可能な品詞は2番目の品詞列であるので、その情報を抽出し、接続表に記述する方法をとっている。この方法を文の始めから文の終りまで繰り返す処理を行っている。また、同じ品詞がすでに記述されている場合は、後続可能な品詞が記述されている品詞と異なる場合に、後続可能な品詞のみが記述される。また、単語に"-NONE-"が含まれている行はその行を無視し、処理を行っている。なお、プログラムの内容、抽出例は付録として巻末に添付する。

4 MSLRパーザへの適用

4.1 解析前の準備

中国語文の解析を行う前に、LR表、辞書インデックスファイルの作成を行う必要がある。これらについて、簡単に説明する。なお、これらの詳細、作成手順については、文献[2][3]および使用説明書[4]を参照していただきたい。

4.1.1 LR表の作成

MSLRパーザでは文の解析を行う前に、LR表を作成しなければならない。これはMSLRパーザの解析アルゴリズムが一般化LR法に基づいているためであり、LR表作成器を用いて、文法と接続表からLRを作成する。MSLRパーザは、作成されたLR表と辞書を参照しながら入力文の形態素・構文解析を行い、解析結果(構文木)を出力している。LR表作成器の最も大きな特徴は、LR表に品詞間の接続制約を反映させることができる点にある。これは、接続制約に違反する構文木を生成する動作をLR表からあらかじめ除去することに相当する。

4.1.2 辞書インデックスファイルの作成

MSLRパーザで解析を行う際には、高速に辞書引きを行うために、辞書のフォーマットで記述された辞書に対してインデックスを作成し、その辞書インデックスファイルを参照する。MSLRパーザで文の解析を行う場合、必ずこのファイルを作成しなければならない。なお、MSLRパーザの辞書引きモジュールは、奈良先端科学技術大学院大学、松本研究室で開発された高速文字列検索システムSUFARYをベースに作成している。

4.2 中国語文の解析

今回の実験では解析する入力文として、中国語の平文を使用した。

4.2.1 解析結果

解析結果(解析失敗等含む)は7つのパターンに出力された。以下に、その特徴と解析例を述べる。

1. 「accept」(解析成功)が出力され、同時に構文木(解析結果)が出力される。

出力例：
```
### 1 ###
```
上海浦东开发与法制建设同步

```
accept

[<BUN>,[<IP>,[<NP-SBJ>,[<NP-PN>,[NR, 上海],[NR, 浦东]],[<NP>,[<NP>,[NN, 开发]],[CC, 与],
[NN, 法制],[NN, 建设]]],[<VP>,[VV, 同步]]]]

total 92
CPU time 0.2 sec
```

2. 「accept」が出力され、構文木も出力されるが、構文木の数が数えられないというエラーが出力される。

出力例：
```
### 5 ###
```
对此，浦东不是简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，而是借鉴发达国家和深圳等特区的经验教训，聘请国内外有关专家学者，积极、及时地制定和推出法规性文件，使这些经济活动一出现就被纳入法制轨道。

```
accept

[<BUN>,[<IP>,[<PP>,[P, 对],[<NP>,[PN, 此]]],[PU, ， ],[<NP-PN-SBJ>,[NR, 浦东]],[<VP>,[<VP>,[<VP>,
[<VP>,[<VP>,[<VP>,[<DVP>,[<VP>,[<ADVP>,[AD, 不]],[<VP>,[VC, 是],[<VP>,[VA, 简单]]]],[DEV, 的]],
[<VP>,[VV, 采取]]],[PU, “],[<VP>,[<VP>,[VV, 干],[<NP-OBJ>,[<QP>,[CD, 一],[<CLP>,[M, 段]]],[<CP>,
[<IP>,[<NP-SBJ>,[NN, 时间]],[PU, ， ],[<VP>,[<PP>,[P, 等],[<LCP>,[<IP>,[<VP>,[VV, 积累],[AS, 了],
[<NP-OBJ>,[NN, 经验]]]],[LC, 以后]]],[<ADVP>,[AD, 再]],[<VP>,[VV, 制定],[<NP-OBJ>,[NN, 法规],
[NN, 条例]]]],[PU, ” ]],[DEC, 的]],[<NP>,[NN, 做法]]]],[PU, ， ],[CC, 而],[<VP>,[VC, 是],[<NP-PRD>,
[<CP>,[<IP>,[<VP>,[VV, 借鉴],[<NP-OBJ>,[<DNP>,[<NP>,[<NP>,[<ADJP>,[JJ, 发达]],[<NP>,[NN, 国家]]],
[CC, 和],[<NP>,[<NP-PN-APP>,[NR, 深圳],[ETC, 等]],[<NP>,[NN, 特区]]]],[DEG, 的]],[<NP>,
[NN, 经验]]]]],[<NP>,[NN, 教训]]]]],[PU, ， ],[<VP>,[VV, 聘请],[<NP-OBJ>,[<NP>,[NN, 国内外]],
[<NP>,[<ADJP>,[JJ, 有关]],[<NP>,[NN, 专家]]],[<NP>,[NN, 学者]]]],[PU, ， ],[<VP>,[<DVP>,[<ADVP>,
[AD, 积极],[PU, 、 ],[AD, 及时]],[DEV, 地]],[<VP>,[VV, 制定]]]],[CC, 和],[<VP>,[VV, 推出],[<NP-OBJ>,
[NN, 法规性],[NN, 文件]]]],[PU, ， ],[<VP>,[VV, 使],[<NP-OBJ>,[<NP>,[<DP>,[DT, 这些]],[<NP>,
[NN, 经济]]],[<NP>,[NN, 活动]]]]],[<VP>,[<ADVP>,[AD, 一]],[<VP>,[VV, 出现]]]],[<VP>,[<ADVP>,
[AD, 就]],[<VP>,[SB, 被],[<VP>,[VV, 纳入],[<NP-OBJ>,[NN, 法制],[NN, 轨道]]]]]],[PU, 。]]]

can't count the number of trees (more than max value of LONG INT)
CPU time 14.84 sec
```

3. 「accept」は出力されるが、構文木が出力されず、異常終了する。

出力例：
```
### 1 ###
```
上海浦东开发与法制建设同步

```
accept

Segmentation fault (core dumped)
```

4. 何も出力されず、異常終了する。

```
### 1 ###
```
浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，因此大量出现的是以前不曾遇到过的新情况、新问题。

```
Segmentation fault (core dumped)
```

5. 「overflow」というエラーが出力される。

出力例：
```
### 10 ###
```

尽管浦东新区制定的法规性文件有些比较“粗”，有些还只是暂行规定，有待在实践中逐步完善，但这种法制紧跟经济和社会活动的做法，受到了国内外投资者的好评，他们认为，到浦东新区投资办事有章法，讲规矩，利益能得到保障。

```
too many forests (it must not exceed 100000)
overflow
```
尽管浦东新区制定的法规性文件有些比较“粗”，有些还只是暂行规定，有待在实践中逐步完善，但这种法制紧跟经济和社会活动的做法，受到了国内外投资者的好评，他们认为，到浦东新区投资办事有章法，讲规矩 / ，利益能得到保障。

6. 「failed」(解析失敗) が出力される。

出力例：
```
### 9 ###
```
建筑公司进区，有关部门先送上这些法规性文件，然后有专门队伍进行监督检查。

```
failed
```
建筑公司 / 进区，有关部门先送上这些法规性文件，然后有专门队伍进行监督检查。

7. LR 表ファイルが読み込めない。

出力例：
```
reading the grammar file 'kisoku' Done
reading LR table file 'lr.tab'
too many actions at state 235, lookahead 'P' (it must not exceed 16)
```

なお、各々のエラーの原因等については以降の考察で述べる。なお以降、このような結果が出たことを簡略化するために、各々の結果を番号で表すことにする。(例：「failed」が出力される結果が出た場合、結果 6. と表す)

5 予備実験

本研究では MSLR パーザの動作を確認するため、2 つの予備実験を行った。1 番目は用意された全ての PCT から抽出したデータを基に MSLR パーザに適用した予備実験。2 番目は用意された PCT の内、1 ファイルに限定して抽出したデータを基に MSLR パーザに適用した予備実験である。

5.1 予備実験 1

5.1.1 実験条件

- PCT のファイル名：最終版の PCT ファイル
- PCT のファイル数： 325
- 単語数 (種類)： 10770
- 文法数 (種類)： 3978
- 入力文：ファイル名 chtb_001.fid に含まれる 10 文

5.1.2 実験結果

以上の実験条件から LR 表を作成し、中国語文の解析を行った結果、すべての文に対して結果 7.[4] が出力された。

ただし、“（完）”、“上海”といった短い文あるいは単語単位を入力文とすると結果 1.、結果 7. が出力される。

[4] 番号の説明は第 4.2.1節 (6ページ) を参照

5.2 予備実験 2

5.2.1 実験条件

- PCT のファイル名： chtb_001.fid
- PCT のファイル数： 1
- 単語数： 226
- 文法数： 164
- 入力文：ファイル名 chtb_001.fid に含まれる 10 文

5.2.2 実験結果

解析を行った結果、結果 1. が出力される文が 7 文、結果 2. が出力される文が 1 文、結果 5. が出力される文が 2 文あった。

6 実験

本研究では、文法、辞書、接続表の各々のデータを抽出した。そこで、本節では用意された全ての PCT から抽出したデータを文法の数を頻度を利用することで文法数を削減し、その削減したデータを基に MSLR パーザに適用した実験を行う。

実験では頻度 1 の文法を減らした場合、文法の延べ数を約 5% 減らした場合、文法の延べ数を約 10% 減らした場合について実験を行った。

6.1 実験条件

共通条件

- PCT のファイル名：最終版の PCT ファイル
- PCT のファイル数： 325
- 単語数 (種類)： 10770
- 文法数 (種類)： 3978(参考値)
- 文法数 (延べ数)： 111377 (参考値)

頻度 1 の文法を減らした場合

- 文法数 (種類)： 1765(約 6 割減)
- 文法数 (延べ数)：全体の約 98%
- 入力文：ファイル名 chtb_001.fid に含まれる 10 文

文法の延べ数を約 5% 減らした場合

- 文法数 (種類)： 677
- 文法の頻度数： 7
- 入力文：ファイル名 chtb_001.fid に含まれる 10 文

文法の延べ数を約 10% 減らした場合

- 文法数 (種類)： 303(9 割以上減)
- 文法の頻度数： 26
- 入力文：最終版のファイルに含まれる 3699 の中国語文

6.2 実験結果

頻度 1 の文法を減らした場合

入力した全ての文において、結果 7. の出力結果が得られた。これは予備実験 1 で出た結果とほぼ同じである。

文法の延べ数を約 5% 減らした場合

入力した文の内、結果 1. の出力を得た文が 1 つもなかった。10 文の内、結果 5. の出力が 4 文、結果 6. の出力が 6 文であった。

文法の延べ数を約 10% 減らした場合

結果を表 1にまとめる。なお、結果 1. のうち 244 文が、“（完）”という短い文であった。

表 1: 延べ数を約 10% 減らした実験結果

結果の種類	文数
結果 1.(解析成功 構文木が出力)	525
結果 2.(解析成功 木の数が数えられない)	28
結果 4.(異常終了)	47
結果 5.(「overflow」が出力)	147
結果 6.(解析失敗)	2952

6.3 考察

6.3.1 LR 表のエラーの原因について

LR 表が読み込めないエラーは、文の解析をする以前の問題であり致命的なものである。これは、文法数を削減しない予備実験 1 でも、文法数 (種類) を半分減らしても、変わらなかった。しかし、文法の延べ数を約 5% 減らしたところで、解析成功はしないものの、LR 表のエラーについては解消されたことになる。この結果、LR 表が読み込めないのは文法数（種類）が原因ではないかと考える。

6.3.2 「overflow」の原因

ここで、「overflow」の起こる原因について考える。この原因については以下の 3 点が挙げられる。

(a) 抽出されたデータ (辞書、文法、接続表) 形式上の問題がある。

(b) 循環 (tautology) を起こすような文法が存在している。

(c) 文法の組み合わせ爆発が起こりやすくなっている。

まず (a) について考える。これは、実際に抽出しているデータに間違いがないかを調べることで解決できる。筆者は予備実験で使用したデータについて、その調査を行った。その結果、抽出されたデータ形式上の問題点はないと判断した。

次に (b) について考える。これは実際に、文法から循環が起っている文法を探し、そのような文法がなければ解決できる。文脈自由文法において、循環が起こり得る文法の条件は、文法の左辺が 1 品詞かつ右辺が 1 品詞である。すなわち、次のような文法は循環を起こすと考えられる。

```
<A> --> <A>
<A> --> <B>, <B> --> <A>
<A> --> <B>, <B> --> <C>, <C> --> <A>
```

<A> --> <A> のように単独で循環が起こる文法については、予備実験の段階で削除した[5] 。その他の文法の中から循環を起こす文法として次のものがあった。

```
<CP> --> <IP>, <IP> --> <VP>, <VP> --> <NP-PRD>, <NP-PRD> --> <CP>
```

この文法についても予備実験の段階で削除した。この結果、実験では、循環を起こす文法はすべて削除している。逆にそのような文法があると結果 3. が出力されることを確認している。このことから、(b) はないと判断した。

次に (c) について考える。これは、組み合わせ爆発が起こる文法を探せばよいが、本研究ではそれができなかった。しかし、(a) や (b) が原因とは考えられないので (c) が「overflow」になる原因ではないかと判断した。

また、「overflow」と出力される文を分割する実験を行った。これは、もしこれらの原因が組み合わせ爆発によるものであれば、分割して文を解析した時に、「overflow」が出力されないのではないかと考えたからである。分割の方法は “,” や “、” それでも分割できないときは、適当な単語を境に分割することにした。この結果、分割されたすべての文において、「overflow」が出力されなかった。このことから、「overflow」が出力される原因が組み合わせ爆発によるものであることが明確になった。

6.3.3 結果 4. についての考察

この原因は現段階では、はっきりとしたことが言えない。オーバーフロー時のように、組み合わせ爆発が原因とも考えられる。しかし、この 2 つの違いは何かということになると、答えがでない。

そこで、前節と同様に結果 4. となる文を分割する実験を行い、組み合わせ爆発が原因であるという可否について調べることにした。分割の方法は前節で述べたとおりである。この結果、前節と同様、分割されたすべての文において、core ファイルが出力されないことが確認できた。このことから、結果 4. の原因が組み合わせ爆発によるものもあるということがわかった。

しかしながら、2 つのエラーの違いについて、追求することができなかった。

7 まとめ

7.1 Penn TreeBank は MSLR パーザの文法抽出用データとして適切か

本研究では PCT から MSLR パーザに必要な全てのデータを抽出した。これまで、述べたように解析がうまく行われないのは文法に問題があるのではないかと述べてきた。しかし、これは PCT の記述文法が文脈自由文法を取り出すときの相性の悪さにあると考える。MSLR パーザでは文法の記述方法として文脈自由文法を指定している。そのために PCT が、文脈自由文法を抽出することに相応しくない構造をしていれば、MSLR が期待する文法は抽出されないと考えている。

[5] 予備実験の結果は循環を起こす文法をすでに削除している。

7.2 MSLRパーザは中国語文の解析器として適切か

一部の中国語文は解析が成功していることから、MSLRパーザが解析失敗(エラーも含む)のすべての原因があるとは言えない。しかし、問題点はある。それは、エラー処理についてである。実験結果から、解析がうまくいかないのは文法の組み合わせ爆発が起こるということを述べた。しかし、文を分割することにより、解析が成功するまではいかないが、「overflow」が出力されるというエラーは解消された。つまり、解析する文が長くなる(組み合わせ爆発が起こりやすくなる)とな「overflow」どのエラーが起こりやすくなっている。このことから、MSLRパーザは文法の組み合わせ爆発に弱く、さらにエラー処理にも問題があるのではないかと考えた。

8 残された課題

残された課題について以下にまとめる。

- 組み合わせ爆発が起こる文法を探し、データ抽出の段階でそれらを修正するツールを作成する必要がある。
- 各々のデータを抽出したツールの見直し。
- 生成確率の高い順に構文木を出力する(PGLRモデルを用いた)実験を行う。

参考文献

[1] Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Shizhe Huang, Tony Kroch, and Mitch Marcus. "Developing Guidelines and Ensuring Consistency for Chinese Text Annotation" Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-2000), 2000.

[2] 田中穂積, 竹澤寿幸, 衛藤純司. MSLR法を考慮した音声認識用日本語文法 –LR表工学(3)–. 情報処理学会研究会報告, SLP15-25, pp.145-150, 1997.

[3] 白井清昭, 植木正裕, 橋本泰一, 徳永健伸, 田中穂積. 自然言語解析のためのMSLRパーザ・ツールキット. 自然言語処理, Vol.7, No.5, pp.93-112, 2000.

[4] 東京工業大学大学院情報理工学研究科計算工学専攻田中・徳永研究室. 日本語形態素構文解析器MSLRパーザ・ツールキット使用説明書 Ver.1.0. 1999.
(http://tanaka-www.cs.titech.ac.jp/pub/mslr)

付録 A　データ

付録 A.1　プログラムファイル

- 文法 1

```
#! /usr/local/bin/jperl

## start

$dir = "/DB/LTDB6/010111/LDC2000T48/data";

@files = sort (`cd $dir/; find . -name "*.fid" -type f -print`);

foreach $file (@files){

  open (IN, "$dir/$file");

  while(<IN>){

      chomp;
      s/<.*>//g;
      s/\t//g;
      s/^\s+//g;
      $space = $_;

      if($space =~ /^\(/){

  $moji .= $space;
      }
  }
}

close IN;

if($moji =~ /\(-NONE-\s[^\)]+\)/){

    $moji =~ s/\(-NONE-\s[^\)]+\)//g;

    while($moji =~ /\(\S+\s+\)/){

$moji =~ s/\(\S+\s+\)//g;

    }
}

$moji -~ s/\(WHNP-\S+\s\S+\)//g;

$bangou = 1;

while($moji =~ /\([^\(^\)]*\)/){

    $moji =~ s/(\([^\(^\)]*\))/ am$bangou /;
    my($a) = ($1);
    $kisoku[$bangou] = $a;
    print "am$bangou   $kisoku[$bangou]\n";

    $bangou = $bangou + 1;

}

#for($s = 1; $s < $bangou; $s = $s + 1){

#    print "am$s   $kisoku[$s]\n";

#}
```

- 文法 2

```
#! /usr/local/bin/jperl

## start

$i = 0;

($file_name) = @ARGV;

open (IN, "$file_name");

while(<IN>){

    chomp;

    s/\(\s+/\(/g;
    s/\s+\)/\)/g;
    s/\s\s/ /g;
    s/\(FRAG\s\s/\(FRAG /;

### ゴミ取り ### #ゴミがある場合はここで修正

    s/<\s//g;

#################

    $jisyo_yobi[$i] = $_;
    $jisyo = $_;

    /^(am\d+)\s/;
    my($bangou) = ($1);

    $bangou_yobi[$i] = $bangou;
    $rensou{$bangou} = $jisyo;

    $i = $i + 1;

}

close IN;

open (IN, "$file_name");

print "<start> --> <BUN>\n";

while(<IN>){

    $m = 0;
    $b = 0;

    #print $_;

    chomp;

    s/\(\s+/\(/g;
    s/\s+\)/\)/g;
    s/\s\s/ /g;
    s/\(FRAG\s\s/\(FRAG /;

### ゴミ取り ### #ゴミがある場合はここで修正

    s/<\s//g;

#################

    $tai = $_;

    #print "$tai\n";

    if($tai =~ /\(am\d+/){
```

```
$left = "< >";
$tai =~ s/^am\d+\s\s\(//;
$tai =~ s/\)/ \)/;

while($b == 0){

    unless($tai =~ /^\)/){

$tai =~ s/^(am\d+)\s+//;
my($right_bangou) =($1);

$rensou{$right_bangou} =~ /(\(\S+)\s+/;
my($right_goku) = ($1);

if($rensou{$right_bangou} =~ /\(\S+\s+am\d+/){

    $right .= " "."<".$right_goku.">";

}
else{

    $right .= " ".$right_goku;
}

    }

    if($tai =~ /^\)/){

$right =~ s/\(//g;

print "<BUN> -->$right\n";
$b = 1;
$right = "";
    }

}
    }

    if($tai =~ /\(.+\s+am\d+/){

$tai =~ s/^am\d+\s\s(\(\S+)\s+//;
my($left) = ($1);
#print "$left\n";
        $tai =~ s/\)/ \)/;

while($m == 0){

    unless($tai =~ /^\)/){

$tai =~ s/^(am\d+)\s+//;
my($right_bangou) = ($1);

$rensou{$right_bangou} =~ /(\(\S+)\s+/;
my($right_goku) = ($1);

if($rensou{$right_bangou} =~ /\(\S+\s+am\d+/){

    $right .= " "."<".$right_goku.">";

}
else{

    $right .= " ".$right_goku;
}


    }

    if($tai =~ /^\)/){
```

```
$left =~ s/\(//g;
$right =~ s/\(//g;

print "<$left> -->$right\n";
$m = 1;
$right = "";
    }

}
    }
}

close IN;
```

- 辞書

```
#! /usr/local/bin/jperl

## start

$none = "-NONE-";

$dir = "/home/xmmiyaki/TreeBank/test";

@files = sort (`cd $dir/; find . -name "*.fid" -type f -print`);

foreach $file (@files){

  open (IN, "$dir/$file");

  while (<IN>){

chomp;
s/<.*>//g;
s/\t//g;
s/^\s+//g;

#print "$_\n";

$yobi = $_;

#if($yobi =~ /\(PP\s[^(^)]+\)/){

    #print "$yobi\n";
    #print "file_name $file\n";

#}

if(/\(([^(^)]+) ([^(^)]+)\)/){
  my($pos, $word) = ($1, $2);

  if($pos ne $none){

      unless($jisyo{$word} =~ /^$pos;|;$pos;/){

  $jisyo{$word} .= "$pos;";

      }
  }
      }
    }
}

close IN;

while (($key, $value) = each(%jisyo)){

    $value =~ s/;$//g;

    print "$key\t$value\n";

}
```

- 接続表

```
#! /usr/local/bin/jperl

# setuzoku.pl -- #########

## start

$mae = 0;

$dir = "/home/xmmiyaki/TreeBank/test";

@files = sort (`cd $dir/; find . -name "*.fid" -type f -print`);

foreach $file (@files){

    open (IN, "$dir/$file");

    while (<IN>){

chomp;

s/<.*>//g;
s/\t//g;
s/^\s+//g;

s/^\(\(/\( \(/;
$goku = $_;

if($goku =~ /^\(\s\(/){

    unless($setuzoku{$mae} =~ /\$\s/){

unless($mae =~ /0/){

    $setuzoku{$mae} .= "\$ ";
}

$mae_flag = 0;
$mae = "";
$ato = "";

    }
}

if($goku =~ /\(([^\(^\)]+) [^\(^\)]+\)/){
    my($pos) = ($1);

    $pos =~ s/\s//g;

    unless($pos =~ /-NONE-/){

if($mae_flag == 1){

    $ato = $pos;

    unless($setuzoku{$mae} =~ /\b$ato\b/){

$setuzoku{$mae} .= $ato." ";

    }

$mae = $ato;

}

else{

    $mae = $pos;
    $mae_flag = 1;
}
```

```
    }
}
    }
}

close IN;

while(($key, $value) = each(%setuzoku)){

    $value =~ s/\s$//g;

    if($value =~ /\$\s/){

$value =~ s/\$\s//;
$value .= " "."\$";
    }

    print "$key : $value\n";
}
```

付録 A.2 抽出したデータの一部

- 文法

```
<start> --> <BUN>
<ADJP> --> <ADJP> <ADJP>
<ADJP> --> <ADJP> PU <ADJP>
<ADJP> --> <ADJP> PU <ADJP> PU <ADJP>
<ADJP> --> <ADVP> <ADJP>
<ADJP> --> <NP-PN> <ADJP>
<ADJP> --> <NP-PN> <QP> <ADJP>
<ADJP> --> <NP> <ADJP>
<ADJP> --> <NP> <QP> <ADJP>
<ADJP> --> <QP> <ADJP>
<ADJP> --> AD
<ADJP> --> AD <ADJP>
<ADJP> --> JJ
<ADJP> --> JJ CC JJ
<ADJP> --> JJ JJ
<ADJP> --> JJ JJ JJ
<ADJP> --> JJ PU JJ
<ADJP> --> JJ PU JJ PU JJ PU JJ
<ADJP> --> PU <ADJP> PU <ADJP> PU <ADJP> PU
<ADJP> --> PU JJ PU
<ADVP-WH> --> AD
<ADVP> --> <ADVP> PU <ADVP>
<ADVP> --> <ADVP> PU <ADVP> CC <ADVP>
<ADVP> --> <ADVP> PU <ADVP> PU <ADVP>
<ADVP> --> <PP> <ADVP>
<ADVP> --> <PP> <PP-TMP> <ADVP>
<ADVP> --> AD
<ADVP> --> AD AD
<ADVP> --> AD CC AD
<ADVP> --> AD PU AD
<ADVP> --> AD PU AD PU AD CC AD
<ADVP> --> CS
<ADVP> --> JJ
<BUN> --> <CP-Q>
<BUN> --> <CP>
<BUN> --> <FRAG-OBJ>
<BUN> --> <FRAG>
<BUN> --> <FRAG> <IP-HLN>
<BUN> --> <FRAG> PU
<BUN> --> <IP-HLN>
<BUN> --> <IP-Q>
<BUN> --> <IP-SBJ-HLN>
<BUN> --> <IP>
<BUN> --> <IP> <IP>
<BUN> --> <IP> PU
<BUN> --> <NP-HLN>
<BUN> --> <NP-HLN> PU <IP-HLN>
<BUN> --> <NP-PN-HLN>
<BUN> --> <NP-PN-SBJ> <VP> PU
<BUN> --> <NP>
<BUN> --> <PP>
<BUN> --> <PP> PU
<BUN> --> <PRN-HLN>
<BUN> --> <QP-HLN>
<BUN> --> <UCP-HLN>
<BUN> --> PU
```

- 辞書

郡 NR
哈克 NR
六十三亿 CD
俞宜国 NR
９·７１２ CD
张雅心 NR
国产化 NN
共和军 NR;NN
轻工业 NN
外汇 NN
赵海娟 NR
６月 NT
刘水玉 NR
分钟 M
谈到 VV
筹资者 NN
十万 CD
上万 CD
所得税 NN
搞好 VV
阳台 NN
偏离 VV
造成 VV
改组 VV;NN
吉林 NR
增值率 NN
傅全有 NR
素质 NN
二百九十一点八亿 CD
外交 NN
联营 JJ
落户 VV
巴拉 NR
十五 CD
,PU
海基会 NR
上午 NT
一百二十亿 CD
关系稿 NN
台帐 NN
减少 VV
代收费 NN
覃立模 NR
分子 NN
兼任 VV
南昆 NR
意味 VV
外经 NN
有效 VA;JJ;AD
板门店 NR
有些 PN;DT;AD
自己 PN
速度 NN
意外 NN
怀成波 NR
宽敞 VA
弹跳 NN
冬训 NN
求同存异 VV
上学 VV
七百八十四 CD
有限 JJ;VA
有线 JJ
非洲 NR
需求 NN
外界 NN
联姻 VV
落后 VA;JJ
深表 VV

- 接続表

```
NR : NR NN NT PU AD ETC OD DEC DEG CD VV DT JJ LC VC P CC VA PN VE CS M MSP NT-SHORT BA NN-SHORT SB $
NT : NT NN LC AD PU VV VC P CC DT NR CD JJ NT-SHORT DEG VE BA DEC VA PN CS NR-SHORT OD LB $
SP : PU
BA : NR JJ AD NN PN PU DT CD VV M NT P
AD : AD VV VC PU DEV SB P VA VE DT NN LC CD NR NT JJ BA DEG LB PN CC CS OD M MSP
PU : NN NR VV AD JJ P DEC CC NT DT CD CS VA PU PN VE MSP LC VC BA SB NT-SHORT DEG LB OD FW M ETC
NR-SHORT SP $
CC : NN VC NR VV AD P NT VA DT CD JJ LC VE PU PN BA OD SB NR-SHORT
CD : M NN PU VV LC VC JJ CC AD P DEG NR NN-SHORT DEC VA CD PN $
VV : VV AS NN NR DEC PU CD JJ CC DT AD P NT PN VA LC VC DEG DEV OD BA VE LB MSP ETC FW SB DER M
NR-SHORT SP $
DEC : CD JJ VC NN NR PU OD VE DT P AD NT VV M VA PN CC
M : NN VV PU JJ VE DEG AD LC DT CC CD DEC SB VA P NR VC M LB NT AS PN DEV ETC OD $
LB : NN CD NR PU WHNP-3
LC : VV AD NR PU DEG P NN JJ DEC OD SB DT CD MSP VC VE PN VA CC SP NT LC $
MSP : VV AD VA P PU
DEG : JJ NN CD NT PU P DT NR VA OD PN NN-SHORT VV AD CC M LC
P : PN VV AD DT NN NT JJ NR PU CD VA P BA OD VE NR-SHORT SB
OD : M JJ PU NN DEC NR
AS : VV NR DEC NN CD DT JJ AD P NT VA PU PN VE M OD SP
JJ : NN PU M CD JJ DEG VV P NR CC DT AD PN VA
SB : VV AD
CS : NR VE VV NT NN AD BA VA CD P DT OD PN
VA : DEV PU CC DEC VA NN CD ETC VV AS MSP LC DER SP DEG P AD NR DT $
DT : NN M CD AD JJ P VE PU PN VA NR DEG VV LB DT
NN-SHORT : NN CC CD PU NN-SHORT
VC : CD NT VA VV JJ NN P NR DT FW AD PU PN VE M OD
DER : AD
VE : CD VA JJ NN NR AD AS VV DT DEC P MSP PN PU M
ETC : NN PU CD JJ LC VV NR AD DT DEG P
NN : CC NN VV NR PU ETC DEC DEG VC LC JJ AD CD P PN VE NT VA DT OD FW MSP SB BA NT-SHORT LB CS DEV
 SP NN-SHORT M $
WHNP-3 : NN
NR-SHORT : NR P NN NT
FW : PU CD $
PN : PU AD VV P VC NN DEG NT CC CD PN DEC DT VA VE NR MSP JJ LC LB OD SB
NT-SHORT : CC
DEV : VV P AD
```

付録 A.3 PCT の一部

```
<DOC>
<DOCID>DNPIN.19971211.0067</DOCID>
<HEADER>
<DATE>1997-12-11</DATE>
</HEADER>
<BODY>
<HEADLINE>
( (NP-HLN (NN 改稿)) )
</HEADLINE>
<TEDNPT>
<P>
( (IP (NP-TPC (CP (WHNP-1 (-NONE- *OP*))
  (IP (NP-SBJ (-NONE- *pro*))
      (VP (NP-TMP (NT 今))
  (VP (VV 播)
      (NP-OBJ (-NONE- *T*-1))))))
      (QP (PU （)
  (FW ＣＢ)
  (CD ００３０１１)
  (PU ）)
  (CLP (M 号)))
      (IP-TTL-APP (PU “)
  (NP-SBJ (NP-APP (NP-PN (NR 中国)
 (NN 人民)
 (NN 银行))
  (NP (NN 副行长)))
  (NP-PN (NR 高德柱)))
  (VP (VV 谈)
      (NP-OBJ (DNP (NP (NN 金融)
       (NN 改革))
   (DEG 的))
      (NP (NN 成就)
  (CC 与)
  (NN 前景))))
  (PU ”))
      (NP (NN 稿)))
      (PU , )
      (NP-SBJ (UCP (NP (NN 标题))
   (CC 及)
   (LCP (QP (CD 一)
    (CLP (M 段)))
(LC 中)))
      (NP (PU “)
  (NP-PN (NR 中国)
 (NN 人民)
 (NN 银行))
  (NP (NN 副行长))
  (PU ”)))
      (VP (VV 改为)
  (NP-OBJ (PU “)
  (NP-PN (NR 中国)
 (NN 银行))
  (NP (NN 副行长))
  (PU ”))
  (PRN (PU （)
       (NP (ADVP (AD 共))
   (NP (QP (CD 两))
       (NP (NN 处))))
       (PU ）)))
      (PU 。)))
( (FRAG (NP-PN (NN 新华社)
       (NN 对外部))
(NP (NT 十二月)
    (NT 十一日)) ) )
</P>
</TEDNPT>
</BODY>

<!-- ************************************************************ -->
```

```
<!-- * Created from chtb_045.sgm on Wed Jun  7 14:13:23 EDT 2000 * -->
<!-- * Log in: root starts tagging at Sat Jun 10 15:24:29 2000. * -->
<!-- * Logout: root stops  tagging at Sat Jun 10 15:26:33 2000. (1424) * -->
<!-- ************************************************************** -->
```