大規模日本語文法構築に関する一考察

野呂 智哉　　岡崎 篤　　徳永 健伸　　田中 穂積

東京工業大学 大学院情報理工学研究科
{noro,okazaki,take,tanaka}@cl.cs.titech.ac.jp

1 はじめに

構文解析において，多様な言語現象を扱うためには大規模な文法が必要となるが，構文木付きの大規模なコーパスがあればそこから大規模な文法を抽出し，ボトムアップに文法を開発することができる。Charniak[2] は，Penn Treebank コーパスから抽出した文法を使用し，人手で作成した文法よりも，特に単語数の多い文において精度のよい結果が得られることを明らかにしている。一方，日本語では完全な構文構造が付与されたコーパスが少なく，括弧付きコーパス等から自動獲得する手法などが提案されている [4]。ところが，このような手法で獲得した文法規則が言語学的に適切であるとは必ずしも言えない。適切な文法を構築するためには人手で構文構造を付与したコーパスを開発する必要がある。しかし，第 2 節で考察するように，人手で構文構造を付与したコーパスから抽出した文法にはコーパス作成者の意図しない解析結果を生む規則が含まれていることが多く，それが曖昧性を無意味に増大させ，構文解析の精度が悪化したり解析時間や使用メモリ量が増大したりする原因となる。そこでこのように曖昧性を増やす原因を分析し，曖昧性を極力抑えるよう文法やコーパスを変更する必要がある。以上のことから，ボトムアップに文法を開発する手順は以下のようになる。

1. 既存の構文構造付きコーパスから文法を抽出
2. 構文解析木の曖昧性を増やす文法規則の分析
3. 分析結果に基づき，構文構造付きコーパスを変更
4. 変更した構文構造付きコーパスから文法を抽出
5. 2〜4 を繰り返す

しかし，こうして得た大規模文法は，実際にはほとんど使われていないと言われている [1]。これまでコーパスからボトムアップに抽出した大規模文法が実用に供せられなかった最大の原因は，解析結果の曖昧性を減少させるために文法やコーパスの変更を行わなかったためであり，2 から 4 の手順を繰り返すことは特に重要であると我々は考える。

本研究では，人手で構文木を付与したコーパスを出発点とし，そこから抽出した文法の問題点を分析し，曖昧性を増やす要因の発見方法を提案し，それに伴う文法の変更基準を与える。また，その基準に基づいて変更したコーパスから抽出した文法を使用して構文解析すると解析結果の曖昧性を大幅に減少させ得ることを実験的に明らかにしている。以上の結果から，曖昧性を極力抑えた実用的な大規模日本語文法をボトムアップに構築することが十分可能であるとの見通しを得ている。

2 大規模なコーパス，文法の問題点

実際に人手で構文構造を付与したコーパスから抽出した文法を使用して構文解析した結果を観察し，曖昧性を増大させる要因を分析すると，大別して以下の 3 種類があると考えられる。

1. コーパス作成者が言語学的に誤った構造を付与
   （例）「開発される」は終止・連体形であるにも関わらず連用句になる（図 1）
2. 同じ構造を付与するべきパターンに対して複数通りの構造を付与（構造に一貫性がない）
   （例）判定詞が直前の名詞句とどの段階で結合するか（図 2）
3. 構文解析時にコーパス作成者の意図しない解析木を生成する規則が存在
   （例）動詞句は連用節にも連体句にもなり得るため，「10年ほど前に体をこわし」が連体句として「長男夫婦」を修飾する構造も可能（図 3）

これら曖昧性を増大させる要因を解決し，コーパスや文法を変更することによって構文解析のための大規模な文法を構築することが可能となるが，曖昧性を増大させる要因をすべて人手で発見することは効率的ではない。次節では曖昧性を増大させる要因を体系的に発見する手法を提案する。

3 曖昧性を増やす要因の発見手法

3.1 構文解析結果から発見する手法

最も一般的な手法は，実際に構文解析を行い，その解析結果を観察することである。しかし，人手で発見することは多大な労力を必要とし，効率的ではない。そこで，曖昧性を増大させる要因となる規則を体系的に発見する手法を考える。

構文解析の曖昧性とは，同一の単語列に対して付与される解析木が複数存在することである。これには大別して以下の 2 種類のパターンがある。

1. ある単語列 $w_i...w_j$ を同一の非終端記号 X が支配するが，X を展開する文法規則 $X \rightarrow Y_1...Y_n$ と $X \rightarrow Y'_1...Y'_{n'}$ が異なる，または，同一の文法規則 $X \rightarrow Y_1...Y_n$ で展開されるが，各非終端記号 $Y_1...Y_n$ それぞれが支配する単語列の範囲が異なる（図 4(1)）
2. ある単語列 $w_i...w_j$ を支配する非終端記号 X と X' が異なる（図 4(2)）

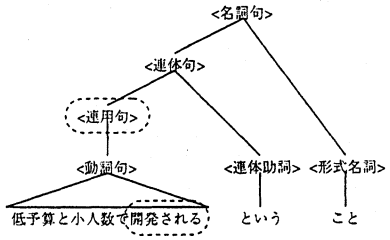
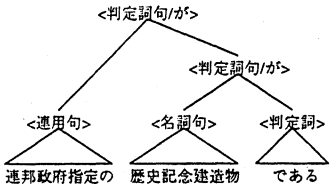


図 1: 誤った構造



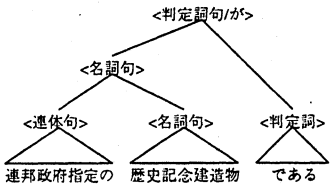


図 2: 一貫性のない構造

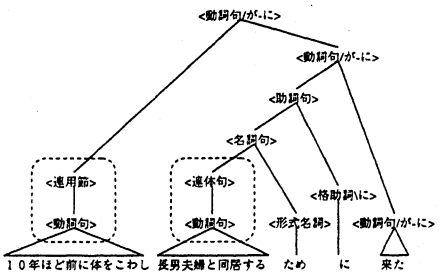


図 3: 意図しない構造

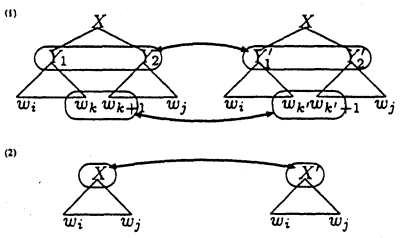


図 4: 曖昧性の出るパターン

我々は構文解析器として一般化 LR モデル（GLR モデル）をベースにした MSLR パーザ [5] を用い，解析結果として出力される圧縮共有統語森 [6] の圧縮もしくは共有されたノードを分析することで，曖昧性を増大させる部分を発見することが可能となる。「5 回分はある」の「5 回分は」の部分の圧縮共有統語森の例を図 5 に示す。ノード番号 146 番の名詞句では数量詞になる規則と複合名詞になる規則のふたつが適用可能であり，1 番目のパターンに相当する。また，ノード番号 182 番の連用句と 183 番の助詞句は同じ「5 回分は」を支配しており，2 番目のパターンに相当する。

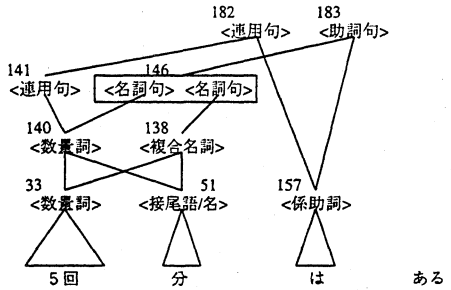


図 5: 圧縮共有統語森の例

3.2 文法規則のみから推測する手法

前節の手法の問題点は，曖昧性を増大させる要因となるパターンを発見できるかどうかは解析する文に左右されることである。そこで，実際に文を解析することなく文法規則のみから推測する手法を考える。

以下の規則が曖昧性を増大させる可能性があることは容易に分かる。

$$< \text{複合名詞} > \rightarrow < \text{複合名詞} >< \text{複合名詞} > \quad (1)$$

$$< \text{数量詞} > \rightarrow < \text{数量詞} >< \text{読点} >< \text{数量詞} > \quad (2)$$

ところが，以下の規則が曖昧性を増大させる可能性があることは，規則 (1)，(2) に比べると発見が困難である。

$$< \text{名詞句} > \rightarrow < \text{連体句} >< \text{名詞句} > \quad (3)$$

規則 (3) の連体句を以下の規則で展開すると，規則 (1)，(2) に類似した形になる。

$$< \text{連体句} > \rightarrow < \text{名詞句} >< \text{連体助詞} > \quad (4)$$

これを一般化したものを以下に示す。

$$A \rightarrow \alpha \stackrel{*}{\Rightarrow} \beta_1 A \beta_2 A \beta_3 \quad (5)$$

このパターンは曖昧性を増大させる要因となる可能性がある。

構文構造付きコーパスを新しく作る際，それまでのコーパスで使用してきた文法規則では表せない言語現象が現れた場合，それに対処するための新しい文法規則の追加が曖昧性を増大させる可能性があるかどうかを判断する場合に，この手法を利用することができる。しかし，実際に文を解析していないため，曖昧性を増大させるパターンが実在するかどうかは別の手法で確認する必要がある。

4 文法の変更基準

構文解析における曖昧性を増大させる要因を発見したら，その曖昧性を抑えるための文法の変更を行う必要がある。しかし，無計画にすべての曖昧性を抑えるべきではなく，どの曖昧性を抑え，どの曖昧性を残すかと言った文法を変更する際の基準を定めなければならない。そこで，曖昧性を以下のように分類し，それぞれについて考察する。

真の曖昧性：構文構造が異なり，それごとに文が表す意味も異なる。これはさらに以下の3種類に分類する。

- 複合名詞内の構造の曖昧性
- 連用修飾語句の係り先の曖昧性
- 連体修飾語句の係り先の曖昧性

見せかけの曖昧性：構文構造が異なるが，その文が表す意味に違いはない。または，文法が不十分であるために生成される言語学的に誤った構造。

見せかけの曖昧性は極力抑えるべきであるが，真の曖昧性も構文解析で解決できないものについては出力する解析木の構造を制限し，その後の意味解析に先送りするべきであると我々は考えている。

まず，複合名詞内の構造について考える。EDR コーパス 9888 文に含まれる名詞の数と複合名詞を構成する形態素数，文字数を表 1 に示す。複合名詞は 1 文あたり 1〜2 個であり，しかも構成する形態素数は 2〜3 個であることから，それほど曖昧性を増大させるものではないと考えられる。ところが，これは正しく形態素解析した後に構文解析する際の話であり，形態素解析の段階の曖昧性も考慮すると，単漢字が名詞として辞書に登録されている可能性もあるため，本来ならば複合名詞ではない名詞も複合名詞となる可能性がある。また，複合名詞を構成する形態素数も増える可能性がある。複合名詞内の構造の曖昧性は構文解析では解決できない問題であるので，構造を右下がりに制限する文法とする。

表 1: 1 文中に含まれる名詞

1 文あたり		1 複合名詞あたり	
名詞数	複合名詞数	形態素数	文字数
4.52 個	1.14 個	2.34 形態素	9.35 文字

次に連用修飾語句の係り先の曖昧性の解決を考える。ここで述べる「連用修飾語句」とは連用句，連用節，助詞句を指す。このうち，助詞句については動詞の必須格情報を利用することで解決できる。例えば，「子供を連れて買い物に行く」の「子供を」は「連れて」と「行く」のどちらに係るか曖昧であるが，ヲ格を必須格とする動詞は「連れる」だけであるため必須格情報を使えば曖昧性を解消できる。ところが，必須格情報を使うことで逆に曖昧性が増大する可能性もある。例えば，「1 時に太郎が次郎に会う」について，「会う」はニ格を必須格とするが，「1 時に」と「次郎に」のどちらが必須格に相当するかは表層的な情報だけでは判断できない。そこで，表層的な情報だけで曖昧性を抑える効果のあるガ格とヲ格のみを考慮し，その他の必須格情報は用いないことで曖昧性を抑えることとした。

最後に連体修飾語句の係り先の曖昧性を考える。連体修飾語句の係り先の曖昧性は，連用修飾語句の場合のように必須格情報を使用する効果は期待できない。そこで，複合名詞の場合のように右下がりに制限すべきである。ところが，連体修飾語句の係り先の曖昧性は，大別して以下の 2 種類がある。

- 連用修飾語句の範囲を変えないもの
  - [[新しい 環境] への適応能力]
  - [新しい[環境への適応能力]]
- 連用修飾語句の範囲を変えるもの
  - [[１００年の 歴史] を持つ祭り]
  - [１００年の[歴史を持つ祭り]]

連用修飾語句の係り先については構造の制限を行わないので，連体修飾語句の係り先の制限は連用修飾語句の範囲を変えない場合に限る。例えば，「１０年近くの歴史を持つ東郷神社の市」の場合，「１０年近くの」の係り先が「歴史」になる場合と「東郷神社」，「市」になる場合で，「持つ」を係り先とする助詞句が，前者は「１０年近くの歴史を」，後者は「歴史を」となり，「持つ」の連用修飾語句の範囲が変わるので，「歴史」を係り先とする構造（図 6(a)）と「市」を係り先とする構造（図 6(b)）の 2 通りが可能となるが，構文解析の段階では 1 通りに絞ることはしない。

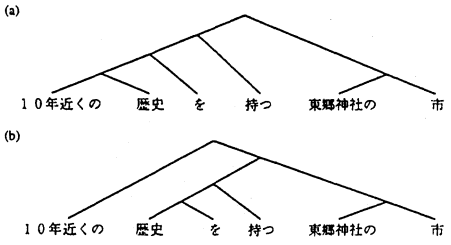


図 6: 連体修飾語句の係り先の制限

5 評価実験

以上の変更基準に従って，EDR コーパス 486 文に人手で構文構造を付けたコーパスに対して変更を施し，変更前と変更後のコーパスからそれぞれ文法を抽出した。486 文の平均形態素数は 20.66 形態素，文法規則数は変更前が 865 規則，変更後が 968 規則である。この文法を使用し，MSLR パーザ [5] で同じ 486 文を解析し，出力された解析木数を 1 文あたりの形態素数別にまとめたグラフを図 7 に示す（縦軸は対数尺度である）。ただし，入力文は品詞列とし，構文解析を行う。これより，特に形態素数の多い文について，曖昧性が大幅に抑えられていることが分かる。

次に，確率一般化 LR モデル（PGLR モデル）[3] を使用し，486 文を訓練データ，同じ 486 文を評価データとして実験を行った。出力された解析木がどれだけ正しいかを評価する尺度を以下のように定義する [4]。

$$\text{括弧付けの再現率} = \frac{\text{正しい括弧付けの数}}{\text{コーパスの構文構造に含まれる括弧付けの数}} \quad (6)$$

$$\text{括弧付けの適合率} = \frac{\text{矛盾しない括弧付けの数}}{\text{解析木に含まれる全ての括弧付けの数}} \quad (7)$$

$$\text{文の正解率} = \frac{\text{正しい解析木が含まれる文の数}}{\text{受理した文の数}} \quad (8)$$

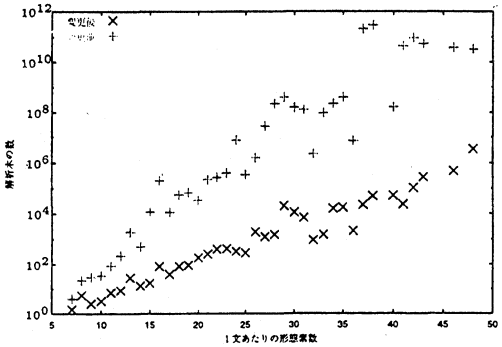


図 7: 1 文中に含まれる形態素数別に見た解析木数

生成確率 1 位の解析木のみを比較した結果を表 2 に，10 位以内から 100 位以内まで 10 位刻みの順位以内に正しい解析木が含まれる文の数の割合をグラフにしたものを図 8 に示す。これより，変更後のコーパスから抽出した文法の方が出力する解析木の数が大幅に抑えられるだけでなく，正しい解析木の生成確率も上位に入るようになることが分かる。

表 2: 解析結果の評価（生成確率 1 位の解析木のみ）

	括弧付けの		文の
	再現率	適合率	正解率
変更前	96.95%	96.92%	61.93%
変更後	97.18%	97.20%	66.26%

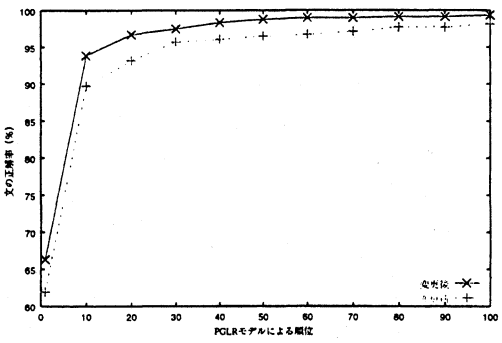


図 8: 各順位以内に正しい解析木が含まれる文の数の割合の変化

6 おわりに

多様な言語現象を扱える大規模な文法を開発するためには構文構造付きコーパスからボトムアップに抽出すべきであるが，構文解析の曖昧性を無意味に増大させるなど問題が多く，実用に供されていないのが現状である。しかし，曖昧性を増大させる要因を分析し，文法やコーパスを変更することを繰り返すことによって，構文解析のための実用的な大規模文法を構築できると我々は考えている。

本研究では，構文解析のための大規模日本語文法をボトムアップに構築するために以下の 3 点を考察した。

- 人手で作成した構文構造付きコーパスから抽出した文法の問題点
- 曖昧性を増大させる要因となる規則の発見手法
- コーパス，文法を変更する際の基準

以上の考察結果に基づいてコーパスを変更し，そこから抽出した文法で構文解析を行った結果，曖昧性が大幅に減るだけでなく，正しい解析木が生成確率順位の上位に入ることを明らかにしている。以上の結果から，大規模なコーパスから抽出した文法規則を，本論文で提案した指針に基づき変更することにより，大規模であるにも関わらず，構文解析結果の曖昧性を極力抑え，実用になる日本語文法を構築することが十分可能であるとの見通しを得た。

今後の課題を以下に示す。

- 今回は使用できるデータが 486 文と非常に少なかったために PGLR モデルによる実験において訓練データと評価データを分けることができなかったが，データを増やしてより厳密な実験を行うべきである。
- 複合名詞内の構造の曖昧性や連体修飾語句の係り先の曖昧性については構造を制限したが，これを再解析する手法を考える。
- 今回は構文解析における曖昧性のみを考慮したが，形態素解析の段階の曖昧性を抑えることも考慮に入れる必要がある。
- さらに曖昧性を抑えるために，素性構造を利用することを考える。

参考文献

[1] James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. Toward conversational human-computer interaction. *AI Magazine*, Vol. 22, No. 4, pp. 27–37, 2001.

[2] Eugene Charniak. Tree-bank grammars. In *The 13th National Conference on Artificial Intelligence*, pp. 1031–1036, 1996.

[3] Kentaro Inui, Virach Sornlertlamvanich, Hozumi Tanaka, and Takenobu Tokunaga. Probabilistic GLR parsing: A new formalization and its impact on parsing performance. 自然言語処理, Vol. 5, No. 3, pp. 33–52, 1998.

[4] 白井清昭, 徳永健伸, 田中穂積. 括弧付きコーパスからの日本語確率文脈自由文法の自動抽出. 自然言語処理, Vol. 4, No. 1, pp. 125–146, 1997.

[5] 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.

[6] Masaru Tomita. *Efficient Parsing for Natural Language*. Kluwer Academic Publishers, 1986.