自然言語処理　144－14
(2001. 7. 17)

品詞列に基づく構文解析

乾伸雄，小谷善行
東京農工大学工学部情報コミュニケーション工学科
〒184-8588　東京都小金井市中町 2-14-16
{nobu, kotani}@cc.tuat.ac.jp

あらまし

コーパスから得られる品詞のNグラム情報を利用して，品詞列に基づく文脈自由文法規則を用いた日本語構文解析を行う．括弧付きコーパスから規則を収集し，頻度情報から得られる規則の出現確率情報を使い，最適な構文木を生成する．そのため，通常の句構造標識とは異なった句単位を用いる．本稿では，もっとも簡易な句として品詞列自身を用いる．この場合，長い品詞列の出現頻度は低いため，コーパスに出現しなかった句の品詞列の予測が必要となる．そのために，Nグラム情報を用いた手法を提案する．実験結果として，従来提案された手法と同等の精度を得ることをでき，ロバストな解析が実現された．

キーワード　日本語構文解析，Nグラム，確率文脈自由文法

N-gram Based Approach for Syntactic Analysis

Nobuo Inui, Yoshiyuki Kotani
Dept. of Computer Science, Tokyo Univ. of Agri. and Tech.
2-24-16 Nakacho Koganei Tokyo, 184-8588
{nobu, kotani}@cc.tuat.ac.jp

Abstract

In this paper, we propose a PCFG-based Japanese syntactic analysis using part-of-speech N-gram information. This method uses syntactic rules gathered from bracketed corpora and finds an optimal syntactic tree by calculating the occurrence probability obtained from the frequency information. The proposed grammar regards sequences of part-of-speech as clause markers. In this case, it is necessary to forecast occurrence probabilities of long or unseen clauses. N-gram information is used to estimate them, instead of clause-occurrence probabilities, to cope with this issue. Initial experimental results showed that our method achieved the same performance as the previous methods and provided robust processing of syntactic analysis.

key words
Japanese Syntactic Analysis, N-gram Information, Probabilistic Context-Free Grammar

1. はじめに

最近，自然言語処理が対象とする自然言語文の範囲は，新聞のような比較的文法的に整ったものから，口語のようなものまで多岐に渡ってきている．これは，音声認識ソフトが実用的になりつつあることや，これまでコンピュータとは縁遠かった人がインターネットを通して，自分の文を書いたり，他の人の文を読むようになったことが主な理由である．このような様々な表現，今まで非文として対象としていなかった文を処理することが，自動翻訳や検索エンジンなどで要求されてきている．このような対象を扱うために，高精度でロバストな解析手法が必要とされる．

一方，自然言語処理に必要な様々な規則は手作業で作成されてきたが，規則収集，適切な解を得るためのパラメータ調整などに多大な労力が必要となる．そのため，コーパスなどの言語資料を利用する研究がなされている．括弧付きコーパスなどを利用した文脈自由文法における確率を利用した解析や学習[1][7][8][9][10][11]，ニューラルネットワークを利用した非文の判定[2]，確率文脈自由文法から N グラムモデルを推測する研究[3]他，様々な研究において統計的モデルやコーパスを用いた研究が行われている．特に確率モデルは様々な場面で用いられている．例えば，未知語の推定[4]，言語モデルの学習[5]などが行われている．これらの研究より，コーパスから獲得された統計的知識は自然言語処理に有効であることが示されている．

本稿では，このような背景の元で，括弧付きコーパスから構文解析に必要な規則や統計的情報を得る手法を提案する．従来の研究としては，コーパスからの規則の学習[11]の研究がある．この研究では，EDR コーパスを用いて，言語に関するある程度の知識から，文脈自由文法で必要となる句構造標識を定め，文脈自由規則を獲得し解析を行っている．EDR コーパスは細分化された句構造標識を獲得するのに十分なサイズではないため，ある程度の句構造標識の簡略化を行い，解析に用いている．

これに対して，本稿で提案する方法は，句構造標識の代わりに品詞列そのものを文脈自由文法の書き換え規則の左辺に考えることが異なる．これによって，特別な言語知識を導入することなく，解析を行うことができるという利点がある．コーパスさえ入手することができれば，広い範囲の言語表現を解析することが可能となる．

2. 確率文脈自由文法

本稿で述べる文法のベースとなる確率文脈自由文法[6]について概説する．次のような文脈自由文法における書き換え規則の集合を考える．

$$\{\alpha \rightarrow \beta \mid \alpha \in V_N, \beta \in (V_N \cup V_T)^+\}$$

V_N :**非終端記号の集合**

V_T :**終端記号の集合**

このとき，ある書き換え規則の発生する確率は，観測データ中で α，$\alpha \rightarrow \beta$の発生する頻度をそれぞれ $f(\alpha)$，$f(\alpha \rightarrow \beta)$と書くと，式(1)によって表される．

$$P(\alpha \rightarrow \beta \mid \alpha) = \frac{f(\alpha \rightarrow \beta)}{f(\alpha)} \quad \cdots (1)$$

例えば，品詞を終端記号，句構造標識を非終端記号と考えると，次のような例が考えられる．

例1）

V_T = { **名詞** , **助詞** , **動詞** , **語尾** }

V_N = { **文** , **連用句** . **用言句** }

文 → 連用句　用言句

連用句 → 連体句　連用句

連用句 → 名詞　助詞

連体句 → 用言句　名詞

連体句 → 名詞　助詞

用言句 → 動詞　語尾

確率文脈自由文法はおのおの規則に割り付けられた式(1)で定義される確率を用い，適切な構文木を式(2)を満たす規則の集まりを選択することで選ぶ方法である．

$$P(r_1 \cdots r_n) = \max_{r_i \in R} \prod_{i=1}^{n} P(r_i) \quad \cdots (2)$$

$P(r_i)$:**ある規則　r_iの発生する確率**

R :**全規則集合**

規則の発生する確率は，括弧付きコーパスを用いることで効率よく得ることができる．また，形態素コーパスを使って，学習によって獲得することも可能である．

確率文脈自由文法の解析では，どのような非終端記号を用いるかによって精度が変わってくる．しかし，一般にどのような非終端記号が最適であるかは知られていない．

3. 品詞列を用いた文法

本稿では，非終端記号に有限な終端記号の列を用い

る. 一般に文脈自由文法では再帰的な構造を許すので, 無限の文を受理することが可能である. これに対して, 本稿で用いる文法は, 有限な終端記号の並びを規則の左辺とし, その分割方法を右辺に書くため, 無限の文を受理することはできない. この問題に対する文法の拡張については 6 節で述べる. 本文法は次のような規則の集合で構成される.

$$\{\alpha \rightarrow \beta \mid \alpha \in V_T^{+}, \beta \in (V_T^{+}\bullet)^{+} V_T^{+}\}$$

V_T : **終端記号の集合**

$\bullet$: **句切り記号**

この文法規則の定義において,「・」は, ある句を部分句に分割するときの部分句の切れ目を表す. 例えば,「名詞助詞動詞語尾」という品詞列(例えば,「今日 は 帰っ た」)では,「名詞助詞」,「動詞語尾」という二つの部分句が存在する. この場合, 次のように規則を記述する.

名詞助詞動詞語尾 → 名詞助詞 ・動詞語尾

長さNの品詞列の場合は, N-1 個の句切れを挿入する場所があるため, $\sum_{i=1}^{N-1} {}_{N-1}C_i$ 通りの句が生成される可能性がある. EDR コーパス[12]のような括弧付きコーパスを使えば, このような規則の頻度を直接得ることができる. EDR コーパスでは句の間の関係を修飾関係などで記述するが, 本稿では句の間の関係は考慮せず, 文の構造だけを抽出する(例1).

例1:本稿で扱う括弧付き構文

(1) EDR コーパスでの構文

```
(S(t(M(S(t(M(W 1 "´７０S")
                (t(W 2 "ファッション"))))
            (W 3 "が"))
          (t(M(S(t(W 4 "街"))
                (W 5 "に"))
              (t(S(t(W 6 "戻"))(W 7 "り")
                (W 8 "はじめ")(W 9 "て")(W 10 "い")
                (W 11 "る")))))))
      (W 12 "。")))
```

(2) 本稿で扱う構文

```
(((((名詞)(名詞))
    (助詞))
    (((名詞)(助詞))
      ((動詞)(語尾)(動詞)(助詞)(動詞)(語尾))))
  (記号))
```

(3) 抽出される規則(すべて頻度は1)

名詞名詞 → 名詞・名詞
名詞名詞助詞 → 名詞名詞・助詞
名詞助詞 → 名詞・助詞
動詞語尾動詞助詞動詞 語尾
→ 動詞・語尾・動詞・助詞・動詞・語尾
名詞助詞動詞語尾動詞 助詞動詞語尾
→ 名詞助詞・動詞語尾動詞助詞動詞 語尾
名詞名詞助詞名詞助詞 動詞語尾動詞助詞動詞 語尾
→ 名詞名詞助詞・名詞助詞動詞語尾動詞 助詞動詞語尾
名詞名詞助詞名詞助詞 動詞語尾動詞助詞動詞 語尾記号
→ 名詞名詞助詞名詞助詞 動詞語尾動詞助詞動詞 語尾・記号

(4) 句構造文法の例

文 → 用言句 記号
用言句 → 連用句 用言句
用言句 → 動詞 語尾 動詞 助詞 動詞 語尾
連用句 → 名詞 名詞 助詞
連用句 → 名詞 助詞

例1のように, EDR コーパスから(3)の規則を直接獲得することができ, 各規則の適用頻度をカウントすることが可能である. そして, 確率文脈自由文法と同様に, この頻度情報を用いて, 最適な構文木を選択することが可能である. 規則 $\alpha \rightarrow \beta$ において, βは αからしか発生することはないので, 式(3)のように規則の出現確率を括弧付きコーパスからの頻度を利用して算出することが可能である.

$$P(\alpha \rightarrow \beta \mid \alpha) = \frac{f(\alpha \rightarrow \beta)}{f(\alpha)} = \frac{f(\beta)}{f(\alpha)} \quad \cdots (3)$$

一例を次に示す.

P(名詞助詞・動詞語尾 |名詞助詞動詞語尾)

$$= \frac{f(\text{名詞助詞・動詞語尾})}{f(\text{名詞助詞動詞語尾})}$$

f(名詞助詞動詞語尾)
= f(名詞・助詞動詞語尾) + f(名詞助詞・動詞語尾)
+ f(名詞助詞動詞・語尾)
+ f(名詞・助詞・動詞語尾) + f(名詞助詞・動詞・語尾)
+ f(名詞・助詞動詞・語尾)
+ f(名詞・助詞・動詞・語 尾)

定性的に見て, 同じ振る舞いをする規則をまとめ, 異なった振る舞いをする規則を分離するとき精度が向上する. 例1(4)で示した連用句や体言句などはその周囲の環境を意味的に分類したものである. つまり, 連用句の後方には用言句が来ることが期待され, 体言句は節が完結することが期待される. これに対して, 品詞列による表現(3)は, 直接的に品詞列を表現したものであるから, このような外部の環境が直接的に反映されない. そのかわり, 細分化された規則が獲得できることと EDR

表1 EDRコーパスから直接獲得された規則を用いた場合の評価

	種類	文数	解析文数	文受理率	句数	解析句数	正解句数	再現率	適合率
PCFG	学習	18999	18999	1.000	376811	379376	337451	0.896	0.889
	評価	10392	514	0.049	5061	5082	3828	0.756	0.753
品詞列	学習	18999	18999	1.000	376811	381506	338044	0.897	0.886
	評価	10392	514	0.049	5061	5111	3858	0.762	0.755

コーパスのような句構造標識を持たない括弧付きコーパスより規則を獲得できるという利点を持つ. 句構造標識をA, B,・・・とおき, 品詞列をa, b,・・・としたとき, あるaは複数の句構造標識によって受理することができるので, a∈A,B,・・・の関係が成り立つ. このように句となるある品詞列が複数の句構造標識に属する場合は, その句の周囲の環境によって句構造標識が決まり, その区別ができない品詞列による規則で解析を行うと精度が低下する可能性がある. 逆に, ある品詞列が特定の句構造標識にしか属さない場合は, より詳細な特徴を品詞列が提供するので, 精度が向上する可能性がある.

4. 構文解析

品詞列に基づく規則を用いた構文解析でも, 通常のボトムアップ, トップダウン構文解析を行うことができる. ただし, 後で述べる N グラムモデル導入のため, 式(3)のような確率文脈自由文法(PCFG)タイプの確率式に加えて, 式(4)のタイプの確率式も用いる. Ωは全規則を表す.

$$P(\alpha \rightarrow \beta) = \frac{f(\alpha \rightarrow \beta)}{f(\Omega)} = \frac{f(\beta)}{f(\Omega)} \quad \cdots(4)$$

式(4)の品詞列を用いた解析の場合, 規則が適用されることにより, 品詞列をどのように分割すればよいかが決定される. 句構造標識の場合, ある句構造標識を複数の句構造標識に分解することでは, 一般に品詞列の分割方法は決定されない. 例えば, 用言句を連用句と用言句に分解しても, 実際の品詞の分割方法が決まるわけではない. 品詞の分解が決定的に行われる, つまり品詞列が与えられればその分割方法が決定できるのならば, 式(4)によって最適な構文木が決定できることになる. 式(4)は各品詞列に適用され, おのおのの品詞列が独立に最大の確率を持つ分割方法を決定する(式(5)).

$$P(r_1 \cdots r_n) = \max_{r_i \in R} \prod_{i=0}^{n} P(r_i) \quad \cdots(5)$$

$P(r_i)$: **ある規則　ri の発生する確率**

R : **全規則集合**

5. 品詞列を用いた構文解析実験

実験は EDR コーパスより収集された文を用いて行った. EDR コーパスに収録された 218,100 文の中で, 207,708 文から規則と頻度情報を抽出した. 実際にこの学習用の文から取り出した 18,999 文で精度を評価した. また, 学習に用いていない 10,392 文でも精度を評価した. クロス検定によって評価すべきであるが, 本稿ではこのような実験環境で評価する.

実際に学習用の文から 1,101,051(=A)種類の句を構成する品詞列が獲得されている. 全規則数は 1,166,268(=B)種類となり, 平均的に一つの品詞列から 1.06(=B/A)規則が存在することになる. ただし, ある程度長い品詞列の分割はほとんど一意に決まってしまい, 偏りがある. 表1に実験結果を示す.

当然のことながら, 学習に用いた文は 100%受理することができる. しかしながら, 学習に用いていない文はそうとは言えない. 品詞列の側面から見た場合, 例えば50語からなる語は品詞数が15の場合でも15^{50}の可能性があり得る. 実際にそれほどではないとしても, 品詞長が長くなるにつれて受理できなくなることが予測できる. 表1では, 学習に用いなかった文の受理率は約 5%であり, かなり低いと言える. 再現率, 適合率は式(5)によって計算されているが, 受理された文だけに関して計算している. 学習に用いていない文に関する評価では, 受理率が低いのであくまで参考であるが, 受理されれば, 高精度の解析が可能であることを示している.

$$\begin{cases} \text{再現率} = \dfrac{\text{正解となった句の数}}{\text{コーパス中の句の数}} \\ \text{適合率} = \dfrac{\text{正解となった句の数}}{\text{解析で得られた句の数}} \end{cases} \cdots(5)$$

PCFG 型の解析(式(2))と品詞列の分割型の解析(式(5))に関して, その差はほとんどないと言える. これは, ある品詞列を解析するとき, その部分品詞列の確率を得ることなく分割が可能であることを示している. 式(5)を使った場合, 計算のオーダーは単語数 N, 句が平均的に分割される部分句数 M に対して, $\log_M N$ となり, 高速な構文解析が実現可能である.

6. Nグラム情報を利用した確率の計算

表1の結果が示すように，品詞列に基づいた構文解析は，形態素情報を用いなくてもある程度の精度の解析が期待できる．しかしながら，受理率が低いことへの対処が必要となる．様々な方法が考えられるが，本稿ではNグラム情報を利用する．

句の生成される確率に関して，品詞 k グラムを用いて，式(4)を式(6)のように近似する．

$$P(\alpha \rightarrow \beta) = \frac{f(\beta)}{f(\Omega)}$$
$$\approx P(w_1)P(w_2 \mid w_1)\cdots P(w_n \mid w_{n-k+1}\cdots w_{n-1}) \quad \cdots(6)$$
$$\beta = w_1\cdots w_n \text{：品詞列}$$
$$P(w_1) = \frac{f(\beta)g(w_1)}{f(\Omega)}$$
$$P(w_i \mid w_{i-k+1}\cdots w_{i-1}) = \frac{g(w_{i-k+1}\cdots w_i)}{g(w_{i-k+1}\cdots w_{i-1})}$$
$$g : \text{品詞列} w_1\cdots w_n \text{における部分品詞列の頻度}$$

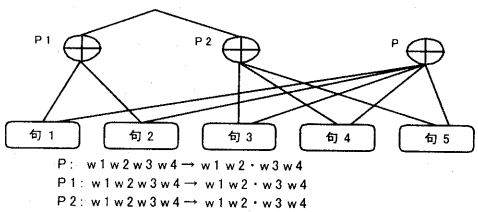

図1 句による統計量の分類

式(6)は各規則で計算されるため，解析を行う品詞列に複数の規則が適用可能となる．この場合，実際にどの規則を適用するかの判断が必要となるとともに，規則を選ぶための計算時間が問題となる．そのため，規則を構成する品詞列の類似性によってグループ化することが考えられる．

図1は規則のグループ化の概念を示す．ただし，本稿では，最も簡単な形，すなわち，すべての句を一つにNグラムの集まりにまとめた形で扱う．これは，最適なモデルではないが，品詞列に基づく解析が最低限どの程度の精度を示すかの目安になる．最終的に，N グラムによる構文解析には式(7)を用いる．

$$P(\alpha \rightarrow \beta) \approx P(w_1)P(w_2 \mid w_1)\cdots P(w_n \mid w_{n-k+1}\cdots w_{n-1}) \quad \cdots(7)$$
$$\beta = w_1\cdots w_n \text{：品詞列}$$
$$P(w_1) = \frac{h(w_1)}{f(\Omega)}$$
$$P(w_i \mid w_{i-k+1}\cdots w_{i-1}) = \frac{h(w_{i-k+1}\cdots w_i)}{h(w_{i-k+1}\cdots w_{i-1})}$$
$$h : \text{規則における品詞列の出現回数}$$

7. Nグラムを用いた構文解析実験

5 節と同様に，式(7)を用いた構文解析実験を行った．結果を表2に示す．EDR コーパスで用いられている品詞分類は荒いものであるので，そのままNグラム情報にすると，例えば，開き括弧と閉じ括弧の対応や助詞の機能的な違いなどを表すのに適切ではない．そのため，次の5種類のNグラム情報を比較することにする．

a. 品詞分類(15種類)だけ扱う
b. 助詞だけ形態素レベルで扱う
c. 記号だけ形態素レベルで扱う
d. 語尾だけ形態素レベルで扱う
e. 助詞と記号を形態素レベルで扱う

また，N グラムで推定する品詞列の候補は，その品詞列が EDR コーパスに含まれる文で句として現れなかったものに限る．つまり，部分的には N グラムで品詞列を分解し，部分的には品詞列の規則で分解するということを行った．これによって，比較的短い品詞列の分解を高精度に行うことが期待できる．

実験では品詞5グラムを用いている．ただし，受理しない文があらわれるのを防ぐため，5グラムまでの線形補間で式(6)の個々の連接確率を表現する．

表1では評価用の学習に使用していない文が約 76%程度の再現率で予測可能であることを示していた．しかしながら，表2では，最大でも約 68%程度の再現率であった．この原因としては，Nグラム情報を句ごとではなく，

表2 Nグラムを用いた解析結果

	種類	文数	解析文数	受理率	句数	解析句数	正解句数	再現率	適合率
品詞情報だけ	評価	10392	10392	1.000	206989	212480	140625	0.679	0.662
助詞は形態素	評価	10392	10392	1.000	206989	209747	141762	0.685	0.676
記号は形態素	評価	10392	10392	1.000	206989	212523	140589	0.679	0.662
語尾は形態素	評価	10392	10392	1.000	206989	212320	139295	0.673	0.656
助詞，記号は形態素	評価	10392	10392	1.000	206989	210044	142084	0.686	0.676

一つにまとめた(図1)ことが原因として考えられる. 本稿では扱っていないが, N グラム情報を連用句や体言句といった句構造標識の単位でまとめることができれば, 更に性能が向上することと考えられる. また, [14]ではRWCコーパス[13]に含まれる形態素の連接情報を利用した構文解析を提案した. このコーパス自体には構文情報が含まれていないので, 小規模な括弧付きコーパスを利用して確率を補正した構文解析に適する確率情報を用い, 再現率約 72%, 適合率 61%の精度を出している. このように, 何らかの確率の補正を行うことでも精度を向上することができると考えられる.

白井ら[11]では, 本稿と同じく EDR コーパスから抽出される品詞列を日本語のヒューリスティックな知識から句構造標識に変換して, 文脈自由文法を得ている. これを使って, 約 62%の再現率を得ている. 本稿で述べた方法は, 日本語に関する情報を使わず, 単にコーパス中に記述された情報を使っているという特徴がある. また, 白井らの方法に比べて, 約 5%程度の精度の向上が得られている.

8. おわりに

本稿では, 括弧付きコーパスから得られる句を構成する品詞列を用いた構文解析の方法を提案した. また, ロバストな解析を実現するための N グラムモデルの導入方法を提案した. 結果として, 従来提案された規則および確率を学習したシステムと同程度以上の精度を出すことができた.

今後の課題として, N グラムモデルの細分化による精度の向上を考えていきたい.

謝辞

本研究は日本学術振興会平成13年度科学研究費補助金(課題番号 12780266)の支援を受けて行われた.

参考文献

[1] Pereira F., Schabes Y.: Inside-Outside Reestimation from Partially Bracketed Corpora, *30th ACL*, pp.128-135, 1992

[2] Lawrence S., Files C.L., Fong S.: Natural Language Grammatical Inference with Recurrent Networks, *IEE Trans. on Knowledge and Data Engineering*, Vol.12, No.1, pp.126-140, 2000

[3] Stolcke A.: Bayesian Learning of Probabilistic Language Models, *Ph. D thesis, University of California at Berkeley*, 1994

[4] Weischedel R., Meteer M., Schwartz R., Ramshaw L., Palmucci J.: Coping with Ambiguity and Unknown Words through Probabilistic Models, *Computational Linguistics*, Vol.129, No.2, pp.359-382, 1993

[5] Suppes P., Botter M., Liang L.: Machine Learning Comprehension Grammars for Ten Languages, *Computational Linguistics*, Vol.22, No.3, pp.329-350, 1996

[6] Manning C.D., Shutze H.: Fondations of Statistical Natural Language Processing, *MIT Press*, 1999

[7] Beil, F., Carrol, G., Prescher, D., Riezler, S,. Rooth, M.: Inside-Outside Estimation of Lexicalized PCFG for German. *37th ACL*, pp.269-266, 1999

[8] Bod, R., Kaplan, R.: A Probabilistic Corpus-Driven Model for the Lexical-Functional Analysis. *COLING-ACL'98*, pp.145-151, 1998

[9] Charniak, E.: Statistical Parsing with a Context-free Grammar and Word Statistics, *AAAI'97*, pp.598-603, 1997

[10] Collins, M.: A New Statistical Parser Based on Bigram Lexical Dependency, *34th ACL*, pp.184-191, 1996

[11] Shirai, K., Tokunaga, T. and Tanaka H.: Automatic Extraction of Japanese Probabilistic Context Free Grammar From a Bracketed Corpus, *Journal of Natural Language Processing*, 4(1): pp.125-146 1997 (In Japanese)

[12] EDR: EDR Electric Dictionary Manual Ver. 1.5, 1996

[13] Toyoura, J., Tokunaga, T. and Isahara, H.: Development of RWC Text Database Tagged with Classification Code, *IPSJ Technical Report*, NL-114-5, 1996 (In Japanese)

[14] Inui N,, Kotani Y.: Robust N-gram Based Syntactic Analysis Using Segmentation Words, *15th PACLIC*, pp.333-343, 2001