# コーパスにおける文脈情報を利用した文法開発支援

川口恭伸, ティラマヌコン・タナラック, 奥村学
北陸先端科学技術大学院大学 情報科学研究科

## 1 はじめに

実用の自然言語処理システムに組み込んで, 実用に耐え得る文法を構築することは困難かつ膨大なマンパワーが必要な作業である. また, あらゆる分野に適用可能で, 過生成でないような文法を作ることは事実上不可能である. そこで, 既存の文法を特定のコーパスに対して修正する作業が必要である.

この作業において人間の作業負荷を軽減するために, 非文解析における誤り訂正のための仮説生成・仮説選択のアプローチと同じように, 解析不能な文を解析可能にするための尤もらしい文法仮説を提示する方法が有効であると考えられる. その手順は, 次のようになる. 1. 解析できない文に対して, 部分解析結果を手がかりにして現在の文法の欠落を仮定し, それを補う文法仮説を生成する.2. コーパスベースの統計的評価値を用いて, 仮説を絞り込む.3. 人間が言語直観に基づいて, 規則として導入する仮説を決定する.

関連する研究として, 清野らによる仮説選択の研究 [1][2] と大谷らによる文法の半自動修正の研究 [3] がある. 前者はより小さく, 他の有力な仮説と競合しないものを優先する評価値によって仮説選択を行ない, 後者は規則の型に関するヒューリスティックスを用いて仮説生成の時点で仮説数を絞り込んだ.

そこで我々は, 局所文脈情報に基づく尺度によって仮説の尤もらしさを決定する手法を提案し, その有効性を調査するために実験を行なった.

## 2 文法開発の手順

我々が採用した文法形式は文脈自由文法である. 文法開発は, まず例文を現在の文法で構文解析し, 解析が失敗した場合にはシステムが解析成功に導くために仮説を生成し, 評価尺度を用いて仮説を絞り込み, 最後に人間が導入すべき仮説を決定する, という手順 (図 1) で行なわれることを想定している.

文法開発支援システムは以下の 2 つのモジュールから構成される (図 2).

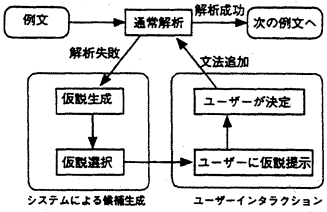


図 1: 文法開発の手順

### 仮説生成モジュール

清野ら [1] に基づくチャート法ベースの仮説生成アルゴリズムを用いた. まず, 現在の文法からボトムアップに生成できる不活性弧はすべて生成しておき, それから, 生成された不活性弧をもとに, 以下のアルゴリズムをトップダウンに適用して仮説を生成する.

**仮説生成アルゴリズム** [1]: 単語位置 $x_0$ から $x_1$ における, ラベル $A$ の不活性弧の仮説 [hypo(A):$x_0$,$x_n$] は, 以下のようなステップによる仮説生成によって導くことができる.

1. 単語位置 $x_0$ から $x_n$ までを覆う $[ie(B_1) : x_0, x_1], ..., [ie(B_n) : x_{n-1}, x_n]$ なる不活性弧列のそれぞれにおいて, $A \rightarrow B_1, ..., B_n$ という新たな規則を仮定し, 不活性弧仮説 $[hypo(A) : x_0, x_n]$ を生成する.
2. 既に存在する $A \rightarrow A_1, ..., A_n$ という形の規則それぞれに対し, $[ie(A_1) : x_0, x_1], ..., [ie(A_{i-1}) : x_{i-2}, x_{i-1}], [ie(A_{i+1}) : x_i, x_{i+1}], ..., [ie(A_n) : x_{n-1}, x_n]$ なる不完全な不活性弧列[1] があるならば, $[ie(A_i) : x_{i-1}, x_i]$ に関してこのアルゴリズムを呼び出す.

このアルゴリズムは, 現在存在しない文法規則を一つだけ仮定することで解析が成功するような不活性弧の仮説を出力する.

### 仮説選択モジュール

仮説生成モジュールでは大量の仮説が生成されるが, その殆んどは正解になりえないものである. 仮説選択モジュールは, 仮説の前後の部分解析結果と, 正

[1] $[ie(A_i) : x_{i-1}, x_i]$ に相当する不活性弧が不足していると考える.

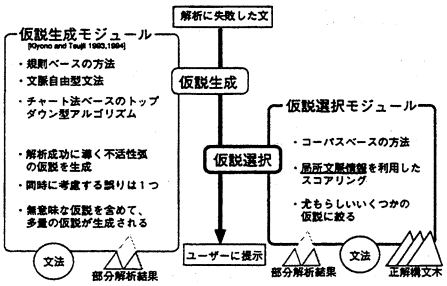


図 2: システムの構成

解構文木集合における同カテゴリの局所文脈情報を手がかりに仮説の尤もらしさをスコアリングし, 仮説の絞り込みを行ない, ユーザーに提示すべき仮説を選択する. 正解構文木集合とは, 現在の文法で正しい解析が行なわれた文の構文木の集合である.

## 3 局所文脈情報を用いた仮説スコアリング

局所文脈情報とは, あるカテゴリの左右に出現するカテゴリのことである. 同じラベルを持つカテゴリは局所文脈情報が類似しているという仮定から, コーパスからの自動文法獲得の研究 [4][5][6] において品詞列からカテゴリを推定するための統計データとして用いられている.

これらの研究では局所文脈情報として語彙カテゴリ (品詞) を用いているが, 我々の仮説選択のタスクにおいては, 仮説の左右には部分解析結果である不活性弧が存在する (図 3) ので, この情報も利用することができる. 我々のタスクで局所文脈情報として語彙カテゴリのみを考慮する場合, 一つの仮説に対してラベルと局所文脈情報のセットは一つしか得られず, データスパースネスの問題が出やすいと考えられる. そこで左右に連接する語彙カテゴリおよび非終端記号 (部分解析結果の不活性弧) の両方を考慮する尺度を考案した.

我々が提案する評価値は, 部分解析木における仮説と文脈とのセットが正解構文木集合の中で出現する可能性が高いほど, その仮説は尤もらしいと判断されるべきだというヒューリスティックに基づいている.

まず, 局所文脈情報として語彙カテゴリのみを考慮する場合の評価値 $Score_{lc}$[2] を定義する. 次に, 局所文脈情報として語彙カテゴリと非終端記号を共に考慮する場合の評価値 $Score_{lc+nt}$[3] を定義する.

[2] $lc$ : lexical category 語彙カテゴリ
[3] $nt$ : nonterminal symbols 非終端記号

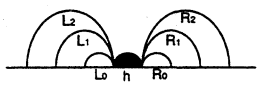


図 3: 左右連接カテゴリ

**評価値 $Score_{lc}$ : 語彙カテゴリのみを考慮する場合**

ラベル $Cat_h$ を持つ 不活性弧仮説 $h$ の評価値は, 左右に接する語彙カテゴリを $L_0(h), R_0(h)$ とすると, 正解構文木集合における, $L_0(h), Cat_h, R_0(h)$ の出現頻度 $N(L_0(h), Cat_h, R_0(h))$ , 文脈 $L_0(h), R_0(h)$ の出現頻度 $N(L_0(h), R_0(h))$ に関して以下のように定義する.

$$Score_{lc}(h) = \frac{N(L_0(h), Cat_h, R_0(h))}{N(L_0(h), R_0(h))}$$

**評価値 $Score_{lc+nt}$ : 語彙カテゴリと非終端記号を考慮する場合**

不活性弧仮説 $h$ の左右に接する語彙カテゴリおよび非終端記号 (図 3) をそれぞれ $L_0(h), ..., L_m(h)$, $R_0(h), ..., R_n(h)$ が存在するとき, $h$ の評価値 $Score_{lc+nt}$ を以下のように定義する.

$$Score_{lc+nt}(h) = \sum_{j=0}^{m}\sum_{k=0}^{n} \frac{N(L_j(h), Cat_h, R_k(h))}{N(L_j(h), R_k(h))}$$

## 4 評価実験

局所文脈情報を用いたスコアリングの有効性を評価するために, 不完全な文法と大規模コーパスを用いて実験を行なった.

実験で初期文法として用いた文法は, タナラックらの研究 [4] で獲得された文法である. これは,EDR 英語コーパス [7] の品詞 (語彙カテゴリ) および括弧情報をもとに自動獲得された文法であり, 非終端記号にはラベルとして機械的に割り振られた記号がつけられている (表 1参照). 文法形式は文脈自由文法であり, 規則数は 272, 各規則の右辺には最大 4 つの語彙カテゴリまたは非終端記号が存在する (表 2参照).

コーパスは,EDR 英語コーパス [7] 中の 48100 文を用いた (表 3). 例文それぞれには, 文法と一致する品詞 (語彙カテゴリ) と, 括弧情報が付加されている. 実験では, 語彙カテゴリのみを用いて構文解析と仮説生成, 仮説選択を行ない, 括弧情報は選択された仮説の正しさの評価に用いた.

表 1: 初期文法の例

```
l1n1 -> adv,noun    l1n2 -> adj,noun
l1n1 -> adv,l1n1    l1n2 -> adj,l1n1
l1n1 -> adv,l1n2    l1n2 -> adj,l1n2
l1n1 -> art,noun    l1n2 -> adj,l1n8
 ...        ...      ...        ...
```

表 2: 初期文法のデータ

| | |
|---|---|
| 規則数 | 272 |
| 非終端記号数 | 55 |
| 終端記号 (語彙カテゴリ) 数 | 18 |
| 右辺の項数の最大 | 4 |
| 右辺の項数の平均 | 2.04 |

## 4.1 準備

初期文法を用いてコーパス全文を構文解析した結果, 29195 文 (コーパス全体の 60.7%) においてコーパス元々つけられている括弧と交差しない構文木を生成した. この括弧付け非交差の構文木から局所文脈情報を抽出した.

また, 解析が失敗した文が 5082 文あり, これに対して仮説生成を行なった結果, コーパスに付けられた括弧付けと交差しない構文木を構成する仮説を生成したのは 3212 文あった. そこでこの 3212 文を実験コーパスとして, 仮説選択の実験を行なった.

## 4.2 正解の基準

仮説の正解を判定するために,"コーパスに付けられた括弧付けと交差しない構文木を生成できる" という条件を用いた. ただし, 現在の評価値はいずれも不活性弧仮説の子カテゴリに関する評価値を導入していないため, 本実験では, 同じ位置で同じラベルを持つ仮説をまとめて 1 つの仮説とみなし, 仮説内部に非交差構文木がひとつでもあれば正解とした.

## 4.3 結果と考察

表 4は, 評価値 $Score_{lc+nt}$を用いた場合に, スコアの高いものから上位いくつかの仮説をユーザーに提示したとき, そこに正解が含まれるかどうかを文単位で数えたものである. 提示する仮説数を上位 1 つだけにした場合, 提示した仮説の中に正解を含んでいた文が 1362 あり, 提示する仮説数を上位 5 つに拡大するとさらに 875 文で正解が含まれるようになり, 上位 10 個に拡大するとさらに 397 文で正解が含まれるようになる, ということをこの表は示している.

評価値 $Score_{lc+nt}$を使用することによって, 1 位の仮説のみを提示した場合,3217 文中で 1362 文において正解を提示することができた. 上位 10 個の仮説を提示すると $1362 + 875 + 397 = 2634$ 文に対して正解を提示することができた. これは正解仮説が生成できた文のうちの 81.9%をカバーする.

表5は, 一文に対して提示する最大の仮説数を決めたときに, 提示された仮説の中にどの程度正解が含まれているかを表している. 提示する仮説数を 1 に絞った場合は, $\frac{1362}{3217} = 42.3\%$の確率で正解が提示され, 提示する仮説数を 10 にした場合提示された仮説のうち $\frac{1362+3349+3409}{3217+11288+12846} \times 100 = \frac{8120}{27351} = 29.7\%$ が正解であった.

表 6と表 4, および表 5と表 7 をそれぞれ比較すると,$Score_{lc}$より $Score_{lc}$の方がやや良い結果になっている. しかし, その差はそれほど大きくない.

$Score_{lc}$と $Score_{lc+nt}$で, それほど性能に差がつかなかったのは, $Score_{lc+nt}$のスコアにノイズが入っているためではないかと考えられる. すなわち,1. 仮説の左右に接するカテゴリは, 全てがその仮説と共存するとは限らない.2. 複数存在する文脈はそれぞれ独立ではない, という問題があり, これがスコアに関してノイズとなっている可能性がある.

表 3: コーパスのデータ

| | |
|---|---|
| 文数 | 48100 |
| 平均単語数 | 10.76 |
| 最大単語数 | 30 |

表 4: 正解を提示できた文の数 ($Score_{lc+nt}$)

| 提示仮説数 | 正解を提示できた文の数 (A) | $\frac{(A)}{全文}$ |
|---|---|---|
| 1 | 1362 | 42.3% |
| 5 | 875 | 27.1% |
| 10 | 397 | 11.7% |
| 20 | 279 | 8.6% |
| 30 | 98 | 3.0% |
| 50 | 91 | 2.8% |
| 51- | 114 | 3.5% |
| 全体 | 3217 | (100%) |

表 5: 提示された仮説の正解率 ($Score_{lc+nt}$)

| 提示仮説数 | 正解仮説数 | 仮説数 | $\frac{正解仮説数}{仮説数}$ |
|---|---|---|---|
| 1 | 1362 | 3217 | 42.3% |
| 5 | 3349 | 11288 | 29.8% |
| 10 | 3409 | 12846 | 24.7% |
| 20 | 4469 | 22105 | 19.5% |
| 30 | 2606 | 17743 | 15.0% |
| 50 | 2832 | 27001 | 11.2% |
| 51– | 6176 | 102015 | 6.2% |
| 全体 | 24203 | 196214 | 13.3% |

表 6: 正解を提示できた文の数 ($Score_{lc}$)

| 提示仮説数 | 正解を提示できた文の数 (A) | $\frac{(A)}{全文}$ |
|---|---|---|
| 1 | 1340 | 41.6% |
| 5 | 1006 | 31.2% |
| 10 | 277 | 8.6% |
| 20 | 225 | 7.0% |
| 30 | 111 | 3.5% |
| 50 | 121 | 3.8% |
| 51- | 136 | 4.2% |
| 全体 | 3217 | (100%) |

表 7: 提示された仮説の正解率 ($Score_{lc}$)

| 提示仮説数 | 正解仮説数 | 仮説数 | $\frac{正解仮説数}{仮説数}$ |
|---|---|---|---|
| 1 | 1340 | 3217 | 41.6% |
| 5 | 3368 | 11288 | 29.8% |
| 10 | 3134 | 12846 | 24.3% |
| 20 | 4300 | 22105 | 19.4% |
| 30 | 2673 | 17743 | 15.0% |
| 50 | 3033 | 27001 | 11.2% |
| 51– | 6315 | 102015 | 6.2% |
| 全体 | 24203 | 196214 | 12.3% |

## 5 おわりに

本稿では, 既存の文法で解析可能な文の構文木から抽出した局所文脈情報を用いて仮説選択のためのスコアリングを行なう方法を提案した. 実験では, 初期文法で解析できなかった文に対して, 上位 10 仮説に絞った場合に 81.4%の文に対して正解を含む仮説を提示できた.

今後の研究課題として, 以下のような事項が挙げられる.

- 子カテゴリに関するスコアの考慮
  現在のスコアリング手法では, 同じ位置にある同じラベルを持つ仮説は, すべて同じスコアになってしまうという問題がある. 例えば, 文全体を覆う不活性弧仮説[4]に関して, すべて同じスコアを割り当ててしまう. そこで, 仮説の持つ子カテゴリについて何らかのスコアを与えることでこの問題を解決できると考えられる.

- 非終端記号にラベルの付いた構文付きコーパスにおける実験
  仮説の正解の基準として括弧付けの非交差を用いたが, これだと適切な部分にに適切なカテゴリラベルを割り当てることができたかどうかの判定ができない. より厳密な判定をするために, 非終端記号にラベルの付いた構文付きコーパスを使用して実験を行なう必要があると考えている.

- 規則仮説の仮説選択, 提示
  現在の実装では, 一文に関する不活性弧仮説をユーザーに提示するモデルになっている. しかし, 各文に対して仮説生成モジュールが必ず正解を含む仮説を生成できるわけではない. そこで, 複数の文, あるいはコーパス全体に対して尤もらしい規則仮説を提示し, ユーザーに提示する方法を検討する必要がある.

[4] トップダウン型の仮説生成アルゴリズムを採用しているので, このような仮説は多数存在する.

## 参考文献


[1] Kiyono, M.,Tsujii, J., Combination of Symbolic and Statistical Approaches for Grammatical Knowledge Acquisition, Proc. of 4th Conference on Applied Natural Language Processing, ACL, pp72–77, 1994.

[2] Kiyono, M.,Tsujii, J., Hypothesis Selection in Grammar Acquisition, Proc. of 15th International Conference on Computational Linguistics (COLING '94), pp837–841, 1994.

[3] Ootani, K., Nakagawa, S., A Semi-Automatic Learning Method of Grammar Rules for Spontaneous Speech Proc. of Natural Language Processing Pacific Rim Symposium '95 (NLPRS '95) pp514–519, 1995.

[4] ティラマヌコン・タナラック, 奥村学, 括弧付きコーパスを利用した文法獲得手法, 情報処理学会研究報告, 96–NL–116, pp85–92, 1996.

[5] Brill, E., Marcus, M., Automatically acquiring phrase structure using distributional analysis, Proc. of Speech and Natural Language Workshop, pp155–159, 1992.

[6] 森信介, 長尾真, 統計によるタグ付きコーパスからの統語規則の獲得, 情報処理学会研究報告, 95–NL–110, pp79–86, 1995.

[7] (株) 日本電子化辞書研究所, EDR 電子化辞書仕様説明書 (第 2 版), 1995.