ベイズ学習による木接合文法獲得

日本電信電話株式会社
NTT コミュニケーション科学基礎研究所

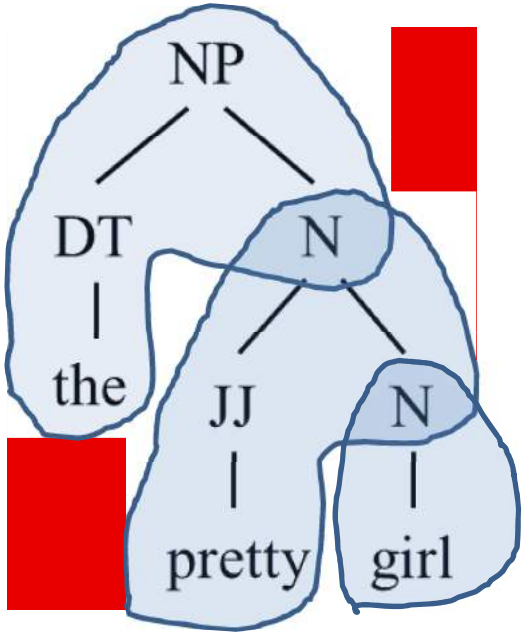
*進藤 裕之, 藤野 昭典, 永田 昌明

文法獲得

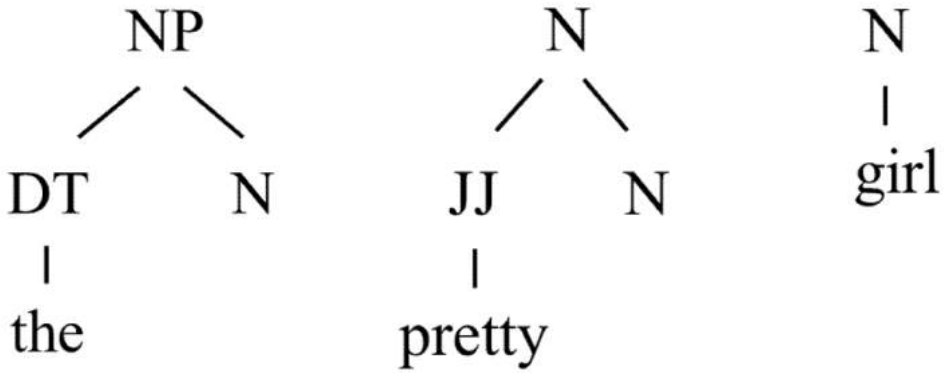
問題設定

構文木から部分木を自動獲得（教師なし学習）する

構文木 (観測データ)



部分木 (未観測データ)



- 構文解析
- 意味解析

研究背景

・木置換文法　　(TSG: Tree Substitution Grammars)

※ TSG は CFG（文脈自由文法）の拡張

木接合文法（TAG）：
木置換文法（TSG） + 部分木の挿入操作

・木接合文法　　(TAG: Tree Adjoining Grammars)

- 発見的手法，最尤推定　[Chang ’03, Chen+ ‘06]
- ベイズ学習による自動獲得

研究概要

木置換文法（TSG）に挿入操作を導入

1. ベイズ理論に基づく部分木の確率モデル

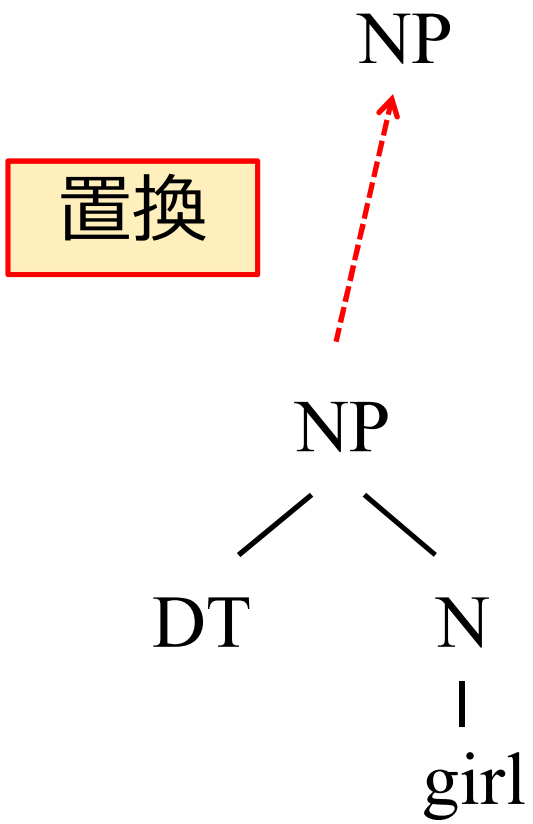
2. 効率的な学習法

※実際には，木接合文法（TAG）のサブセット

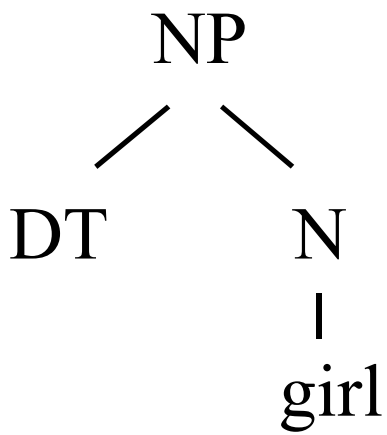
木接合文法（TAG）

NP

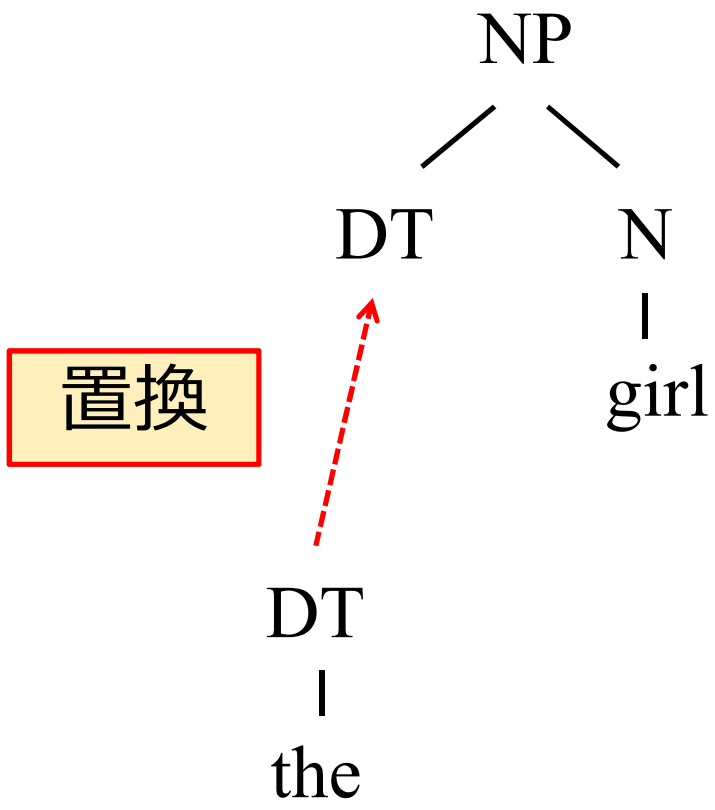
木接合文法（TAG）



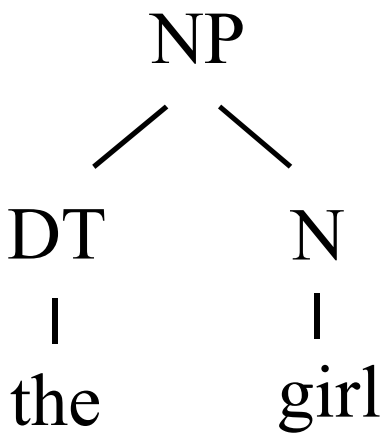
木接合文法（TAG）



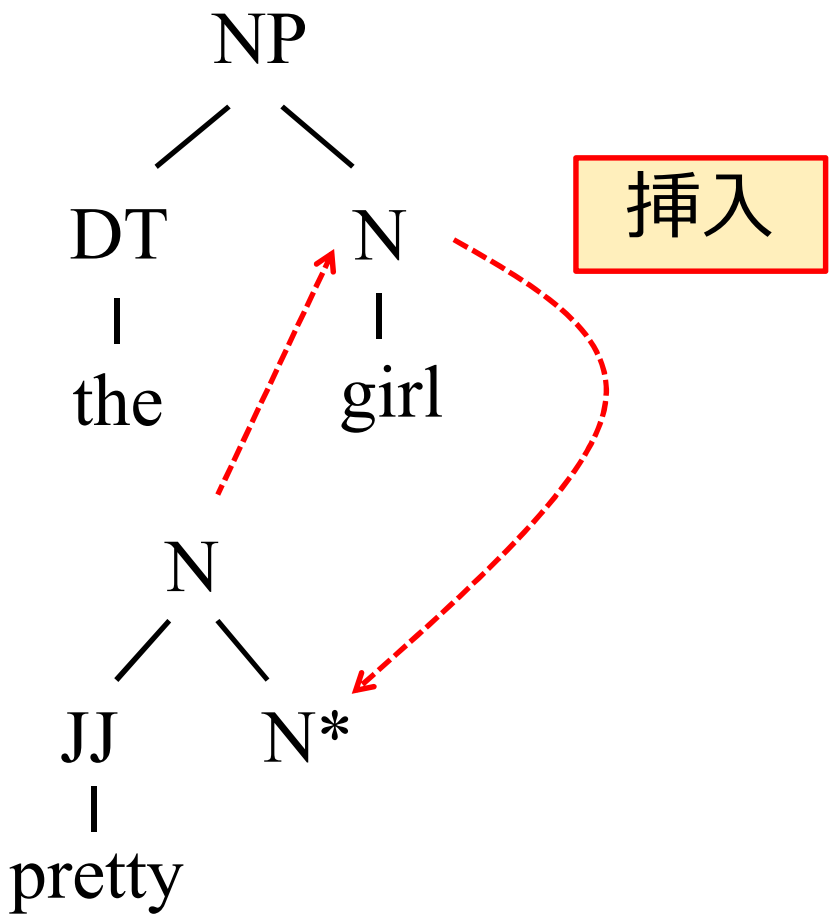
木接合文法（TAG）



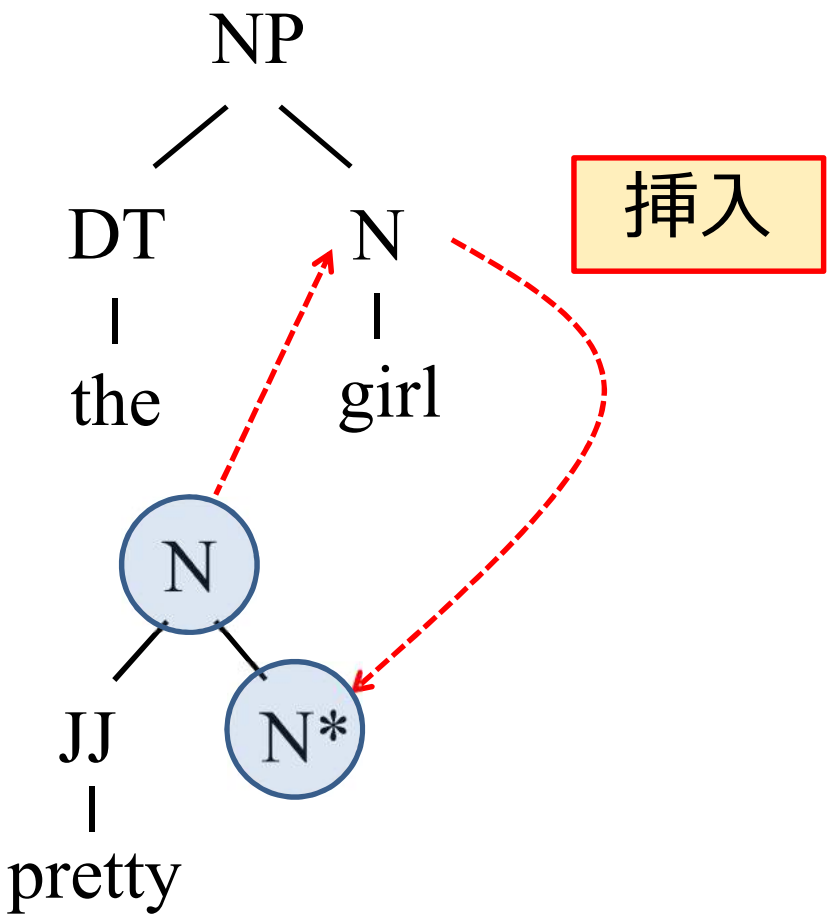
木接合文法（TAG）



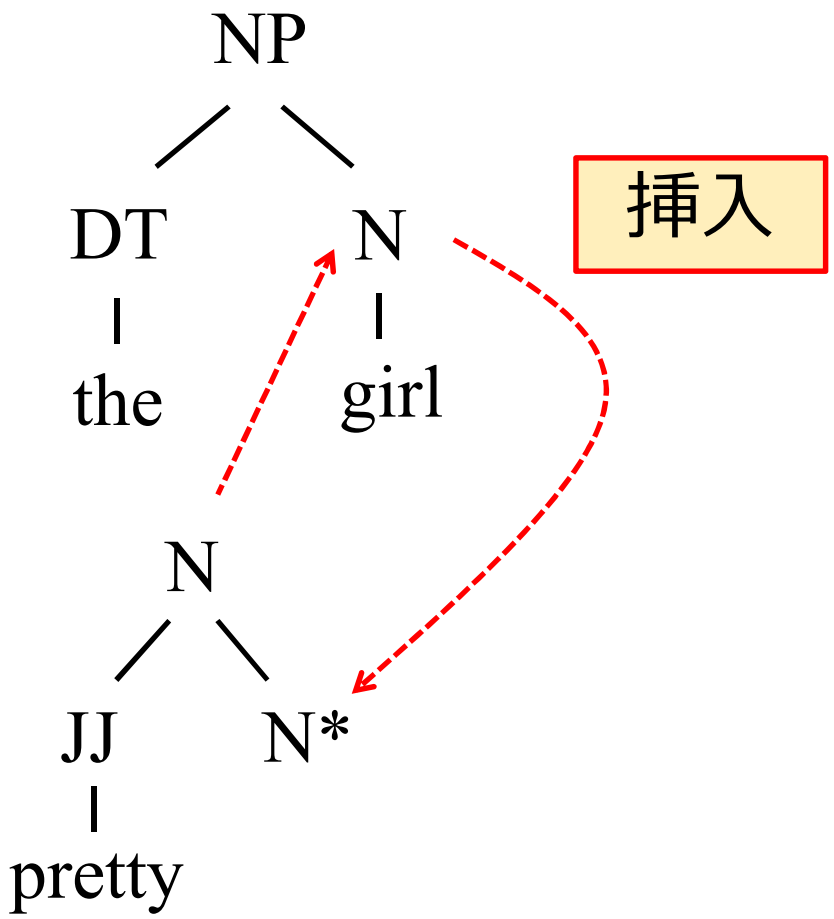
木接合文法（TAG）



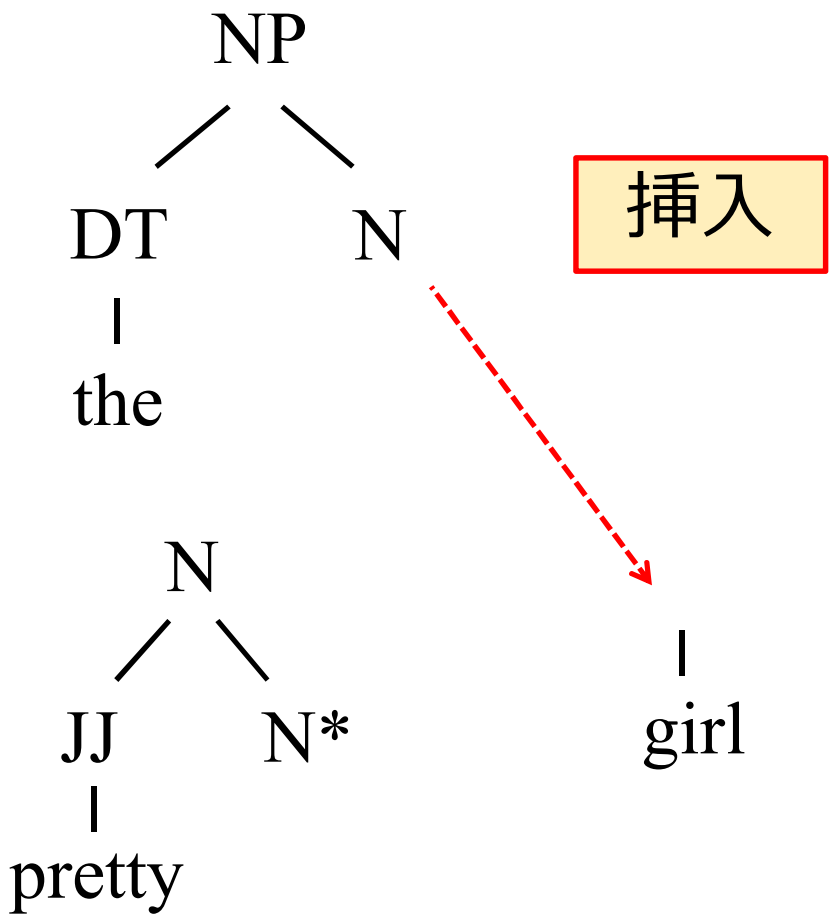
木接合文法（TAG）



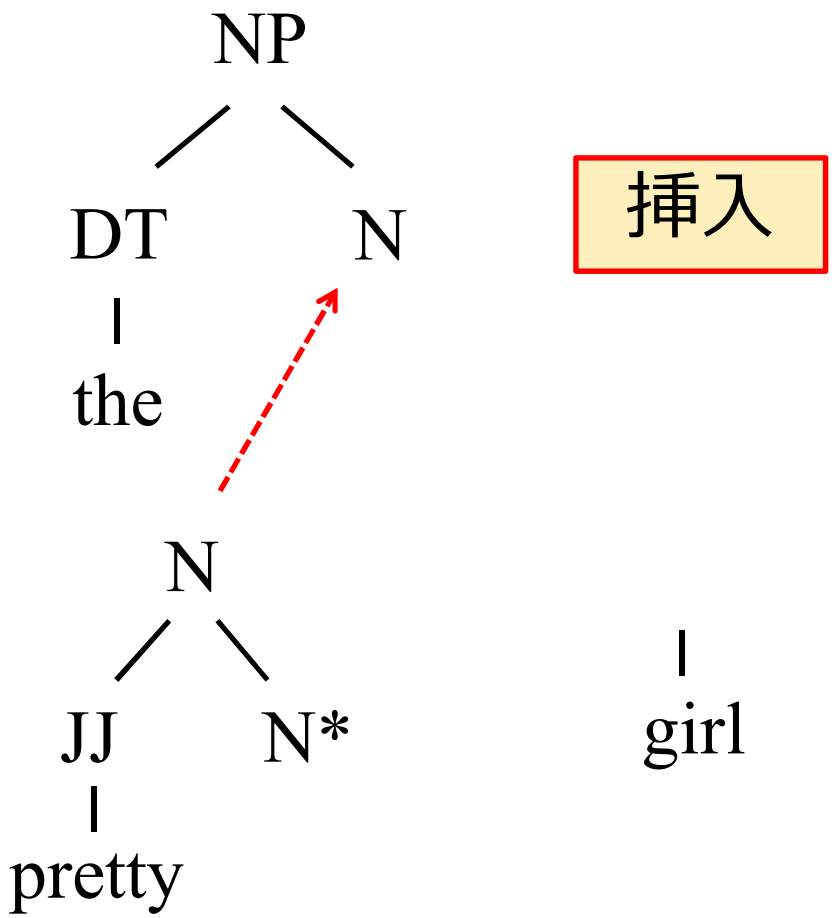
木接合文法（TAG）



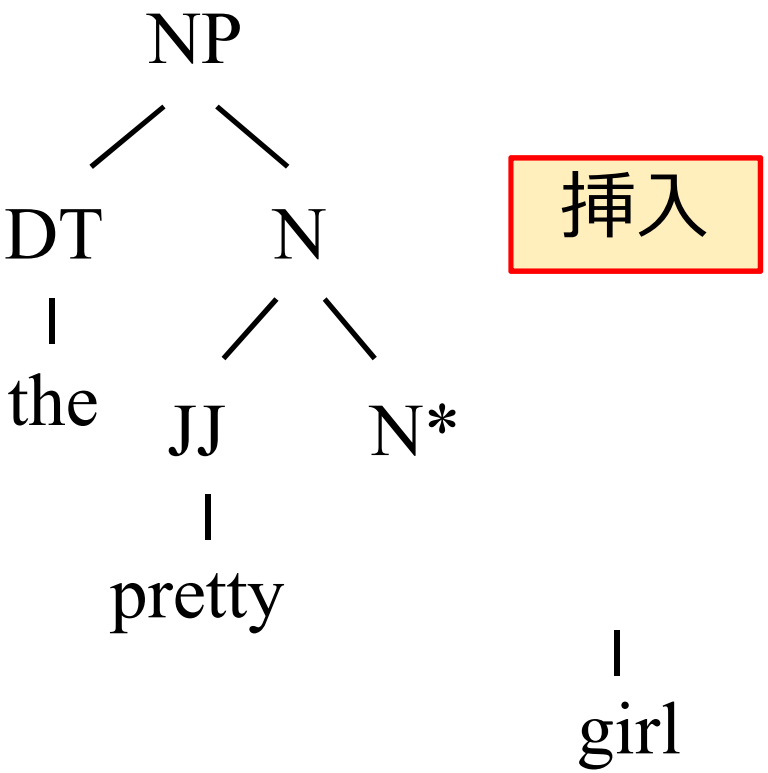
木接合文法（TAG）



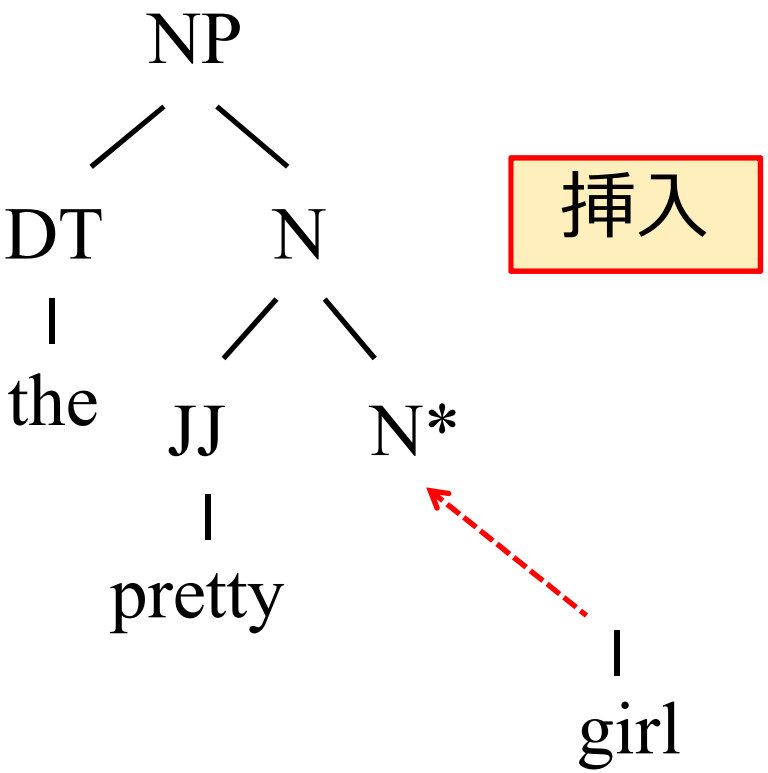
木接合文法（TAG）



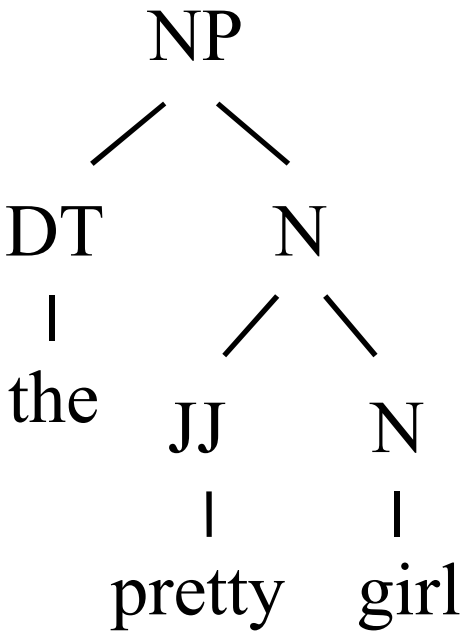
木接合文法（TAG）



木接合文法（TAG）

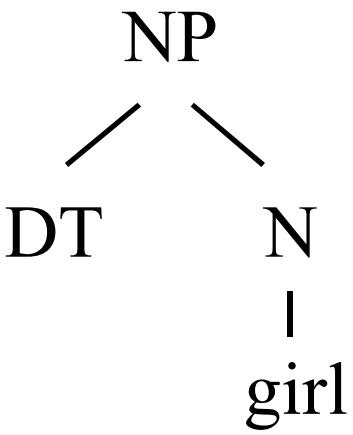


木接合文法（TAG）

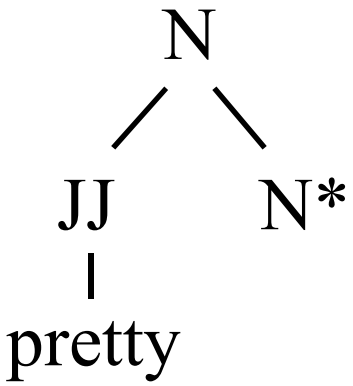


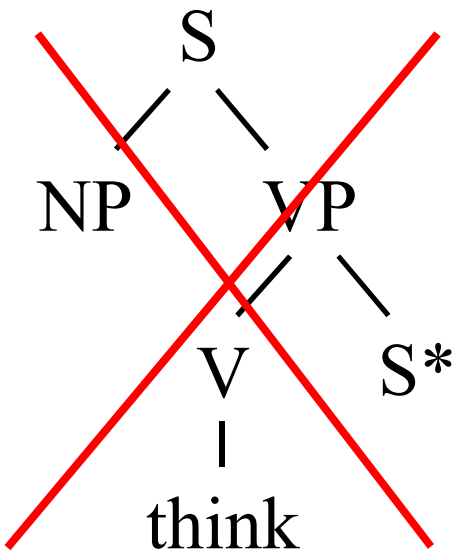
部分木の仮定

置換用の部分木

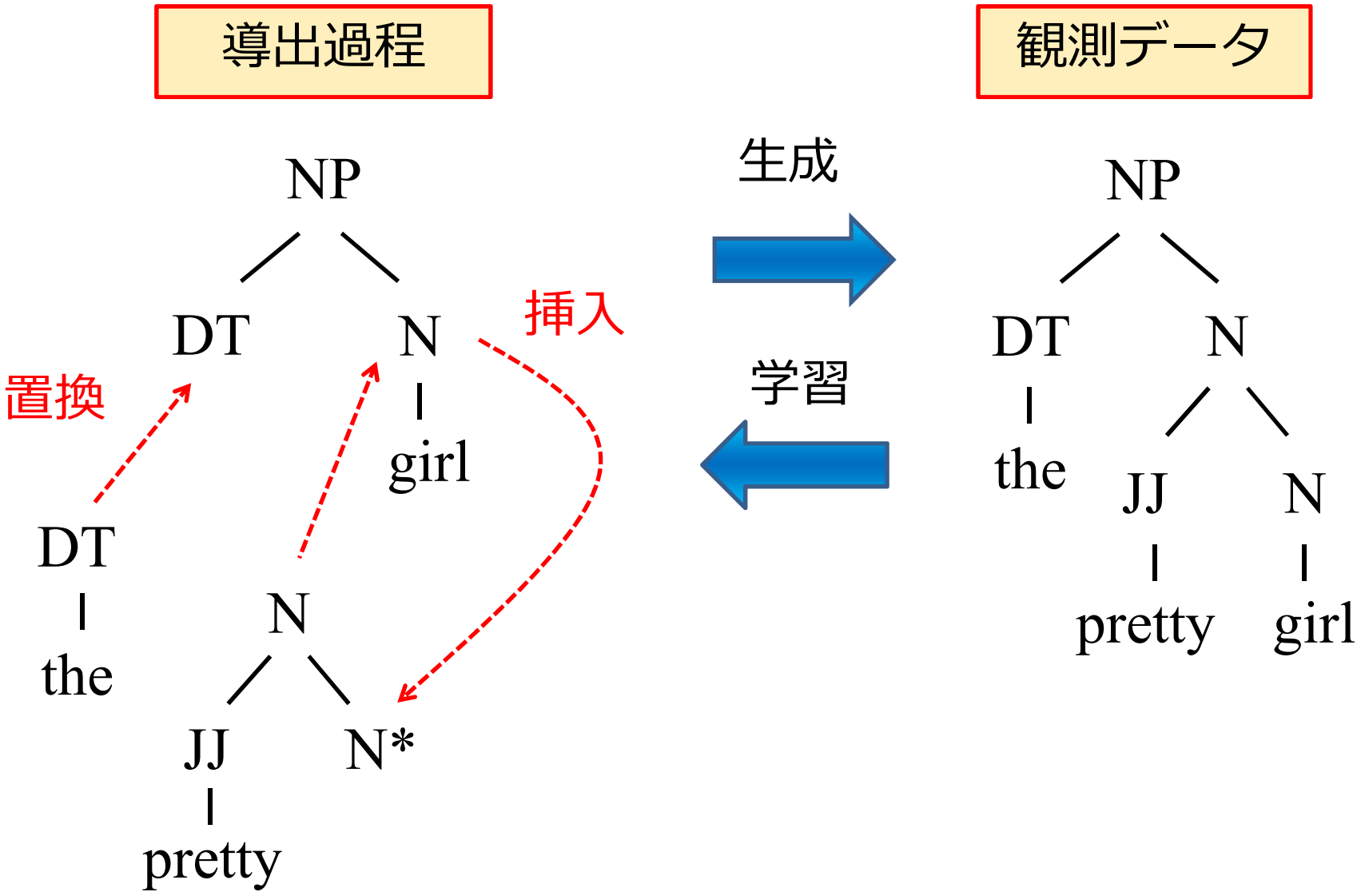


挿入用の部分木





生成と学習

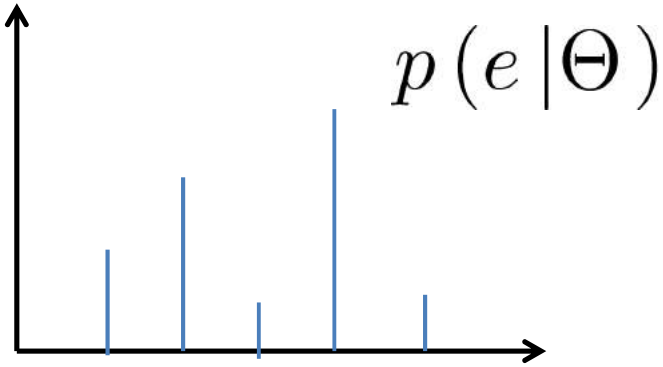

導出過程
観測データ
生成
学習
挿入
置換
NP
DT
N
girl
DT
the
N
JJ
N*
pretty
NP
DT
N
the
JJ
N
pretty
girl

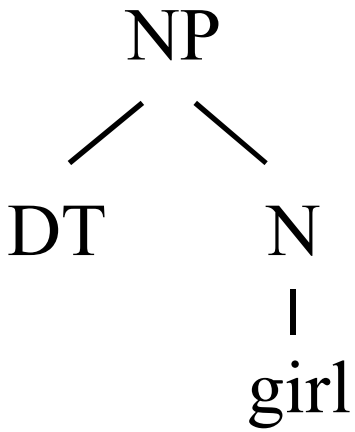
確率モデル

確率モデル（部分木の生成モデル）

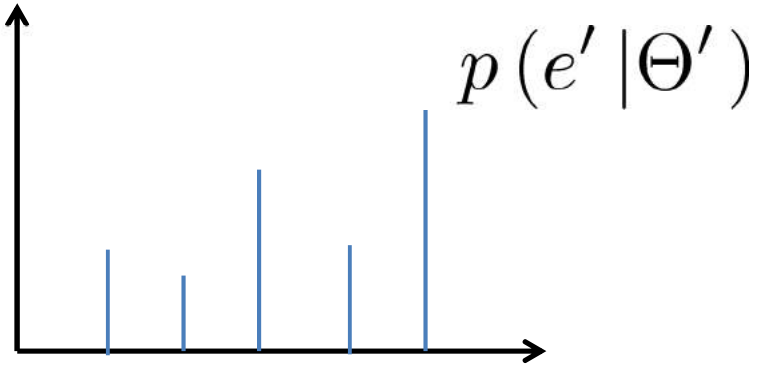
事前確率：Pitman-Yor Process [Pitman and Yor '97]

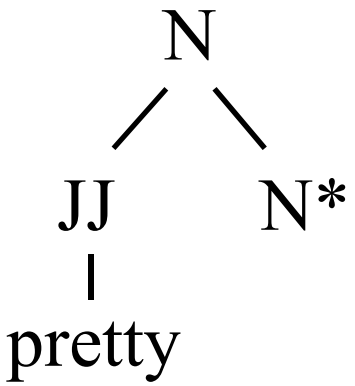
置換用の部分木





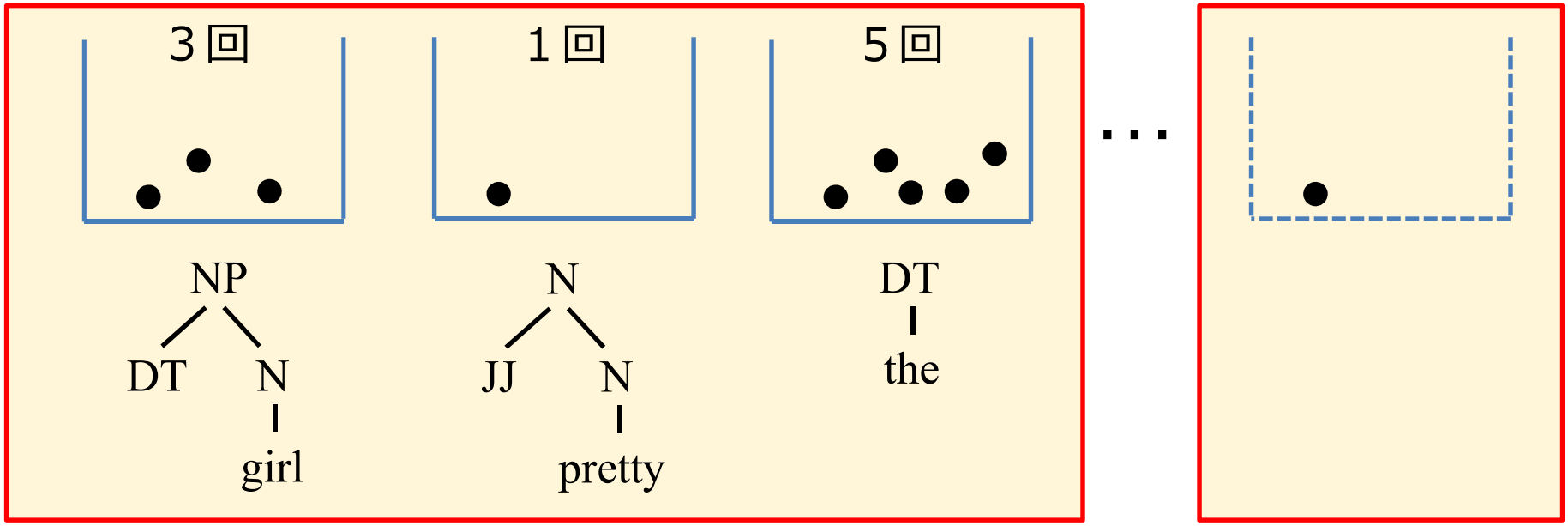
挿入用の部分木





確率モデル（部分木の生成モデル）

Pitman-Yor Process　に基づく部分木の生成モデル



・球数に比例した確率で箱を選択　あるいは

・ある確率で新たな箱（＝部分木）を作成

・今までに生成された少数の部分木を繰り返し利用

・「コンパクト」かつ「データに適応的」な部分木の確率モデル

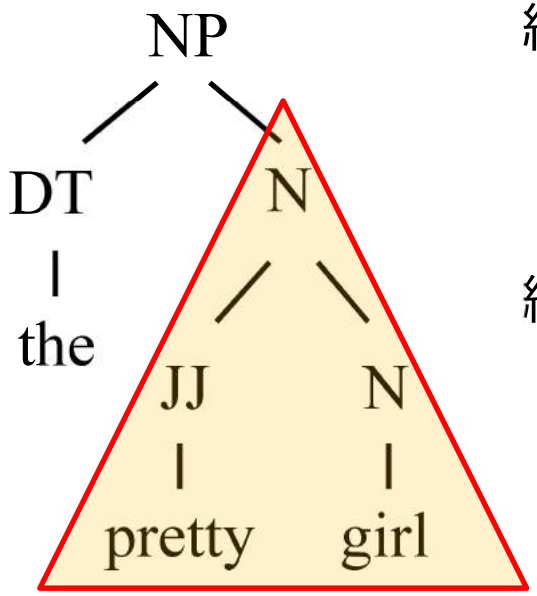
学習

学習

ブロック化 Metroporis-Hastings法（MCMCの一種）

１．構文木の内側確率の計算（≒HMMの前向き確率）　**動的計画法**

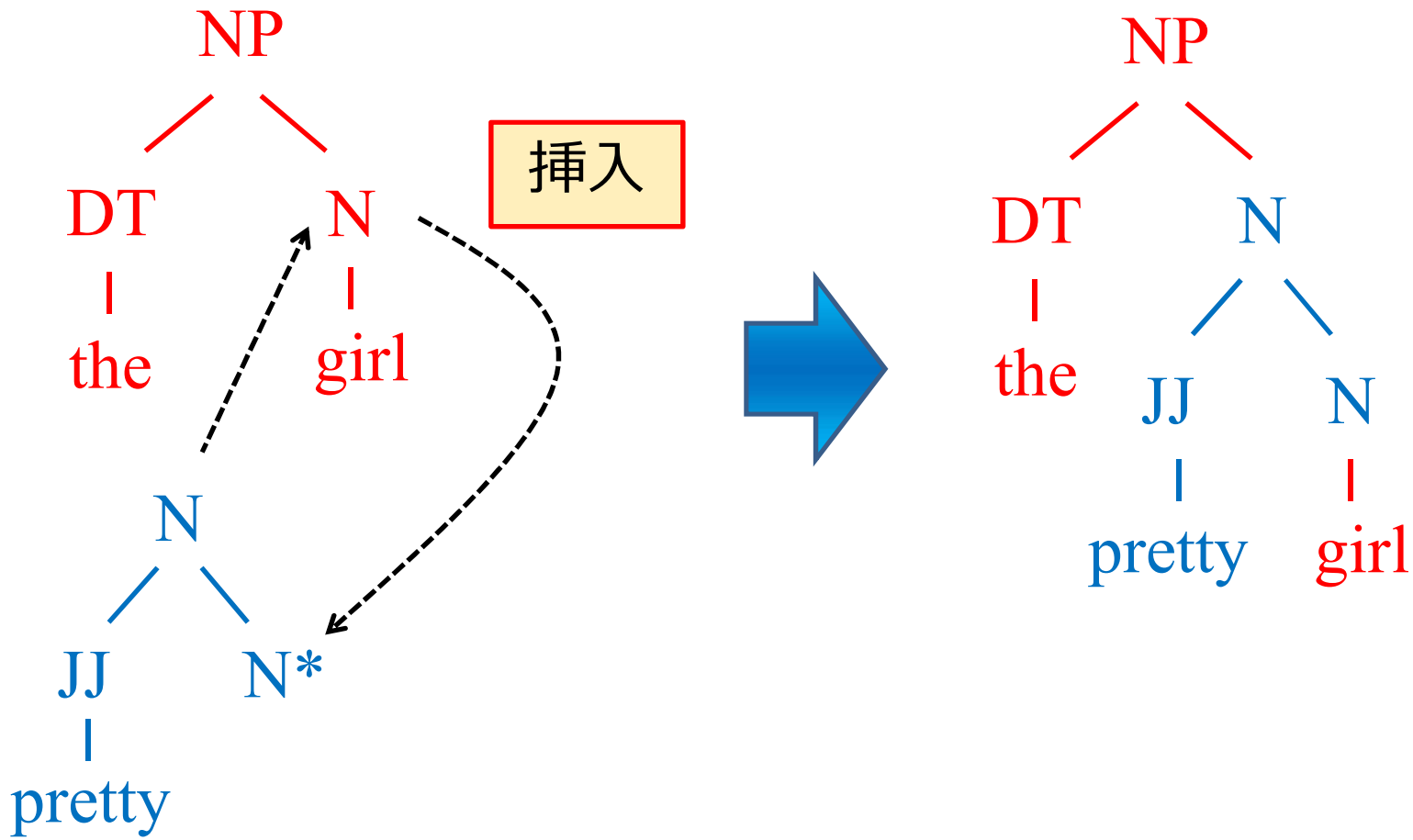
ex. “N”から始まり“pretty girl”を出力する確率：



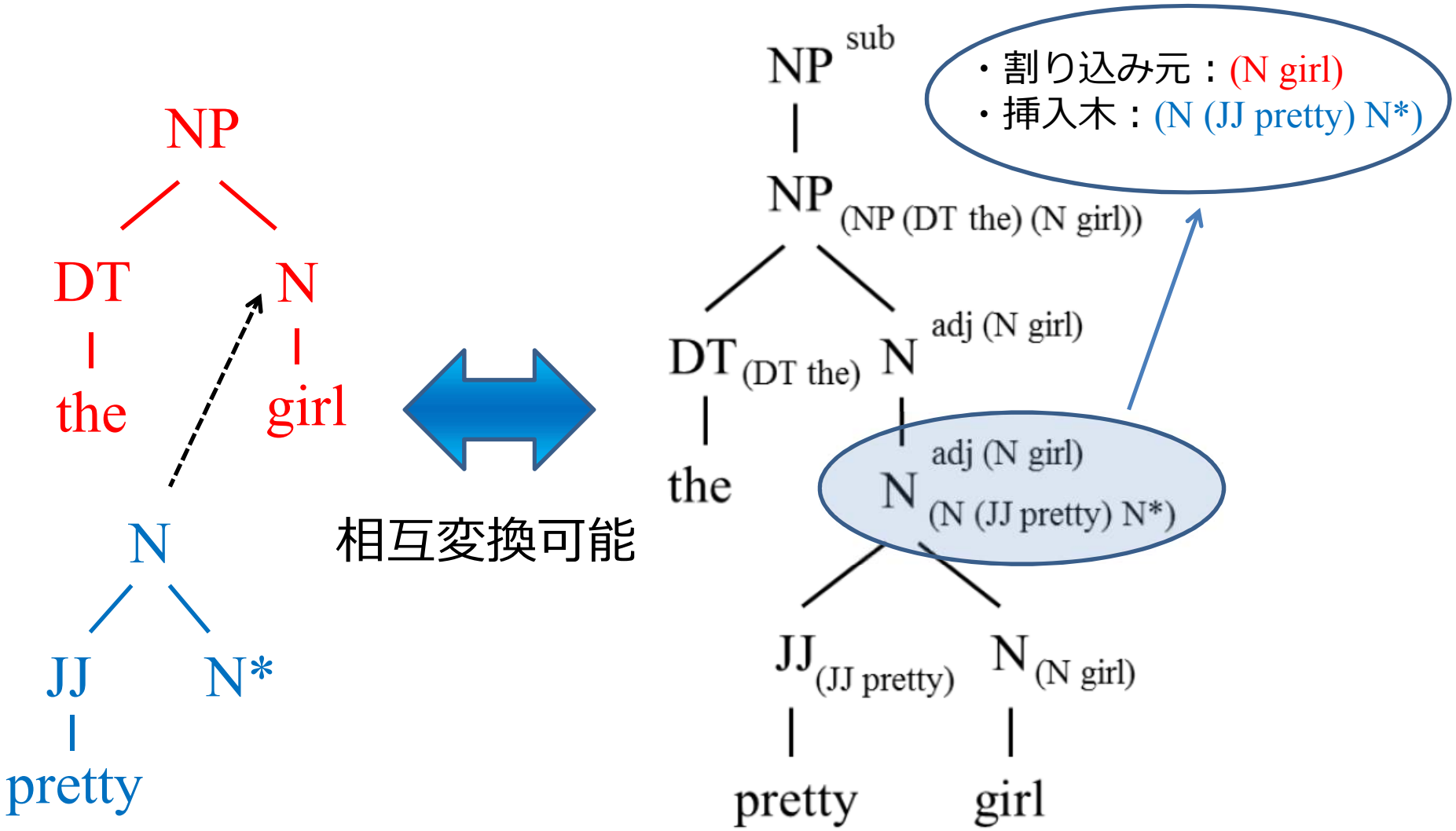
経路１：P (N → JJ N) × P (JJ → pretty) × P (N → girl)

経路２：P (N → JJ N, JJ → pretty) × P (N → girl)

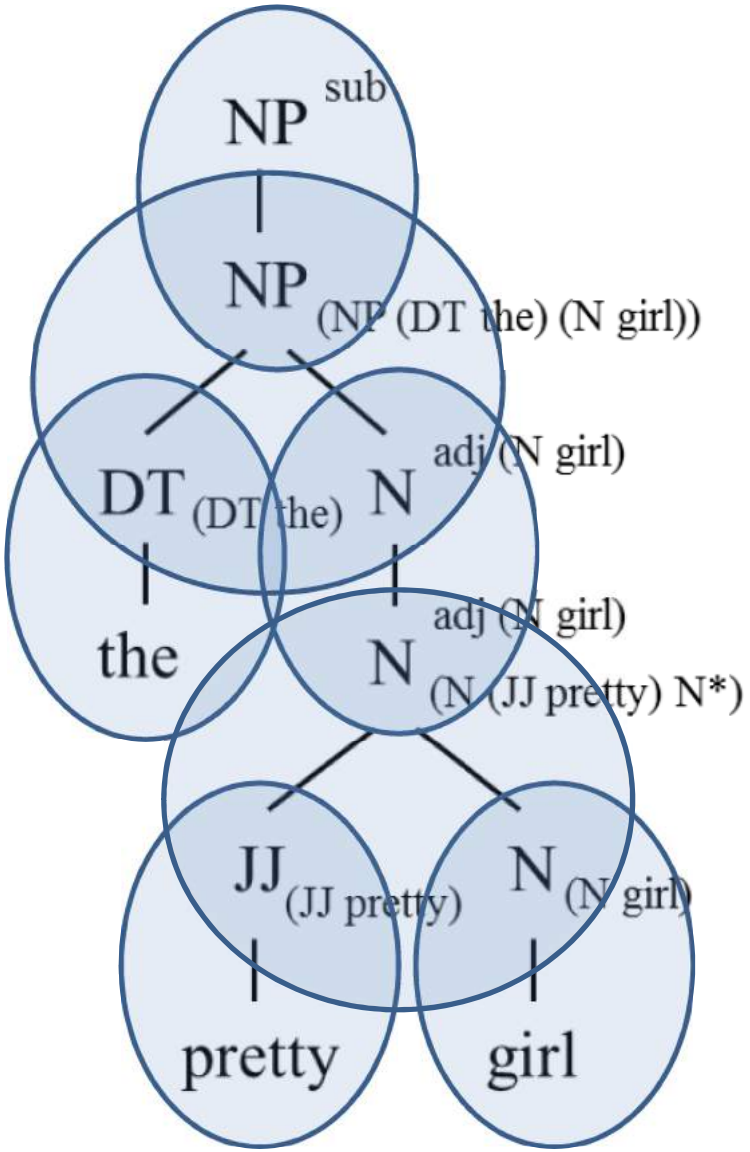
挿入操作の問題点



解決策：ノードのラベル付け



解決策：CFG に分解

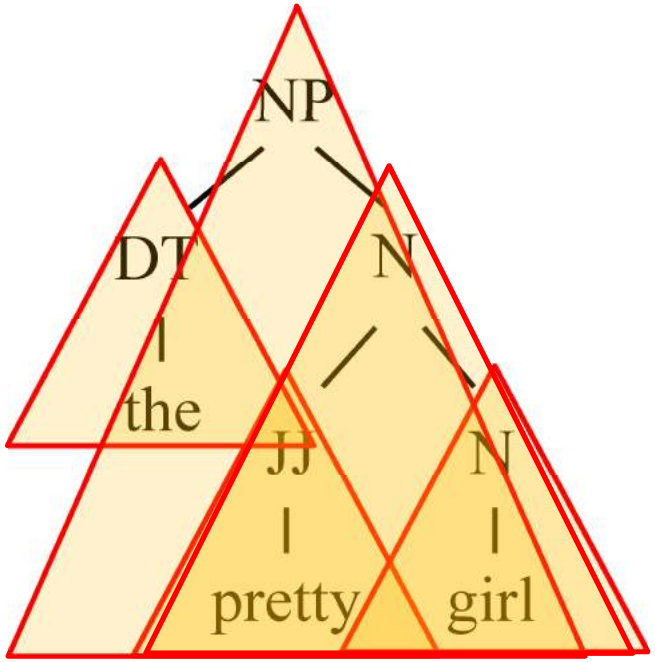

NP
sub
NP
(NP (DT the) (N girl))
DT
(DT the)
N
adj (N girl)
the
N
adj (N girl)
(N (JJ pretty) N*)
JJ
(JJ pretty)
N
(N girl)
pretty
girl

学習

ブロック化 Metroporis-Hastings法（MCMCの一種）

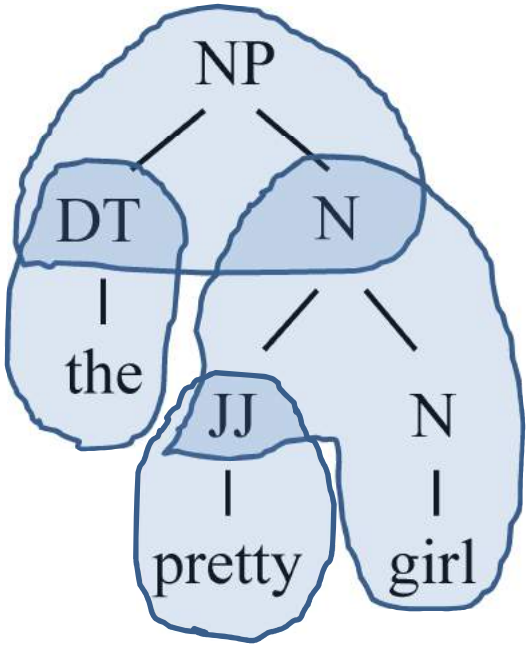
１．構文木の内側確率の計算（≒HMMの前向き確率）　動的計画法

ex. “N”から始まり“pretty girl”を出力する確率：



学習

２．内側確率に基づいて部分木を生成（サンプリング）



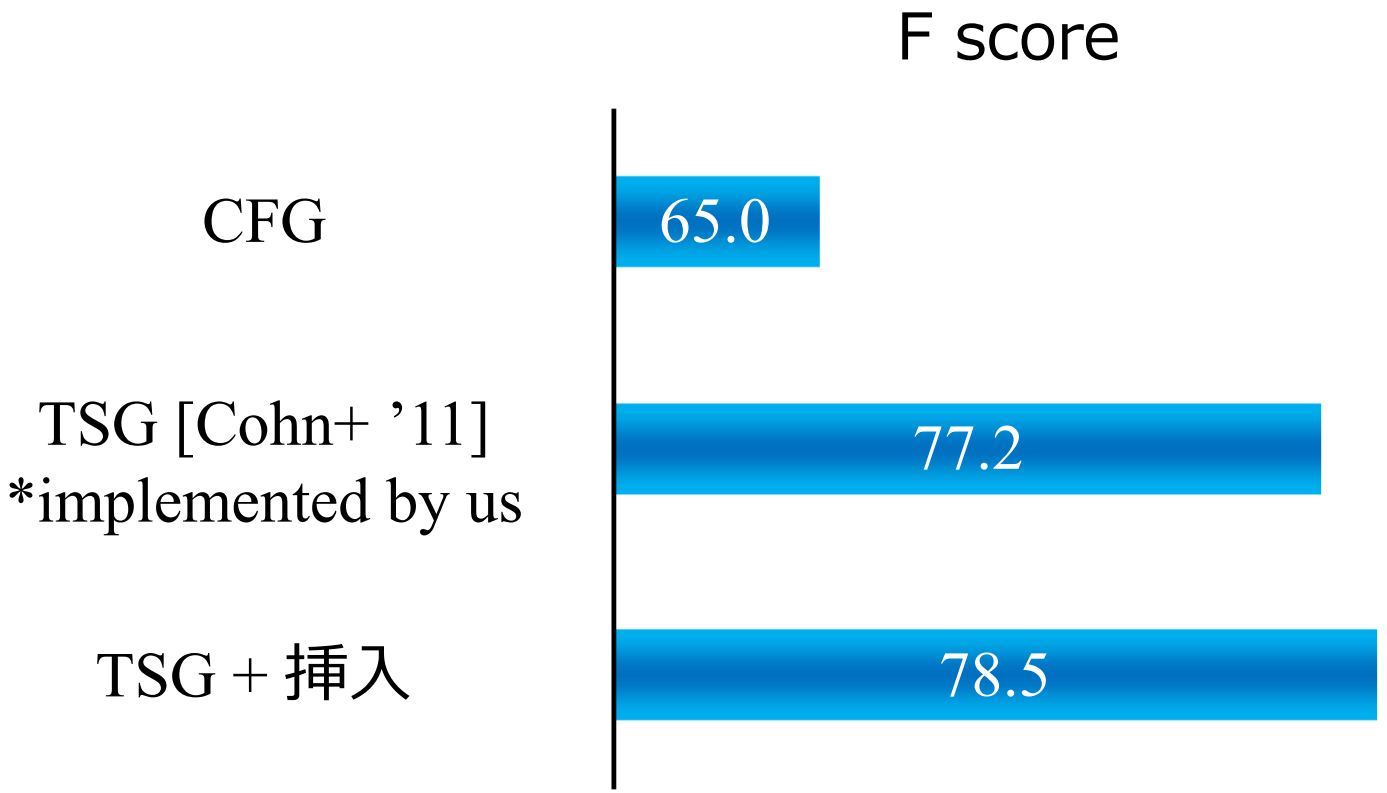
３．Metropolis-Hastings法で部分木集合を受理または棄却

受理：ステップ２の部分木集合をモデルに加える
棄却：ステップ２のサンプルは捨てて以前のサンプルに戻す

実験結果

実験結果（小規模データ）

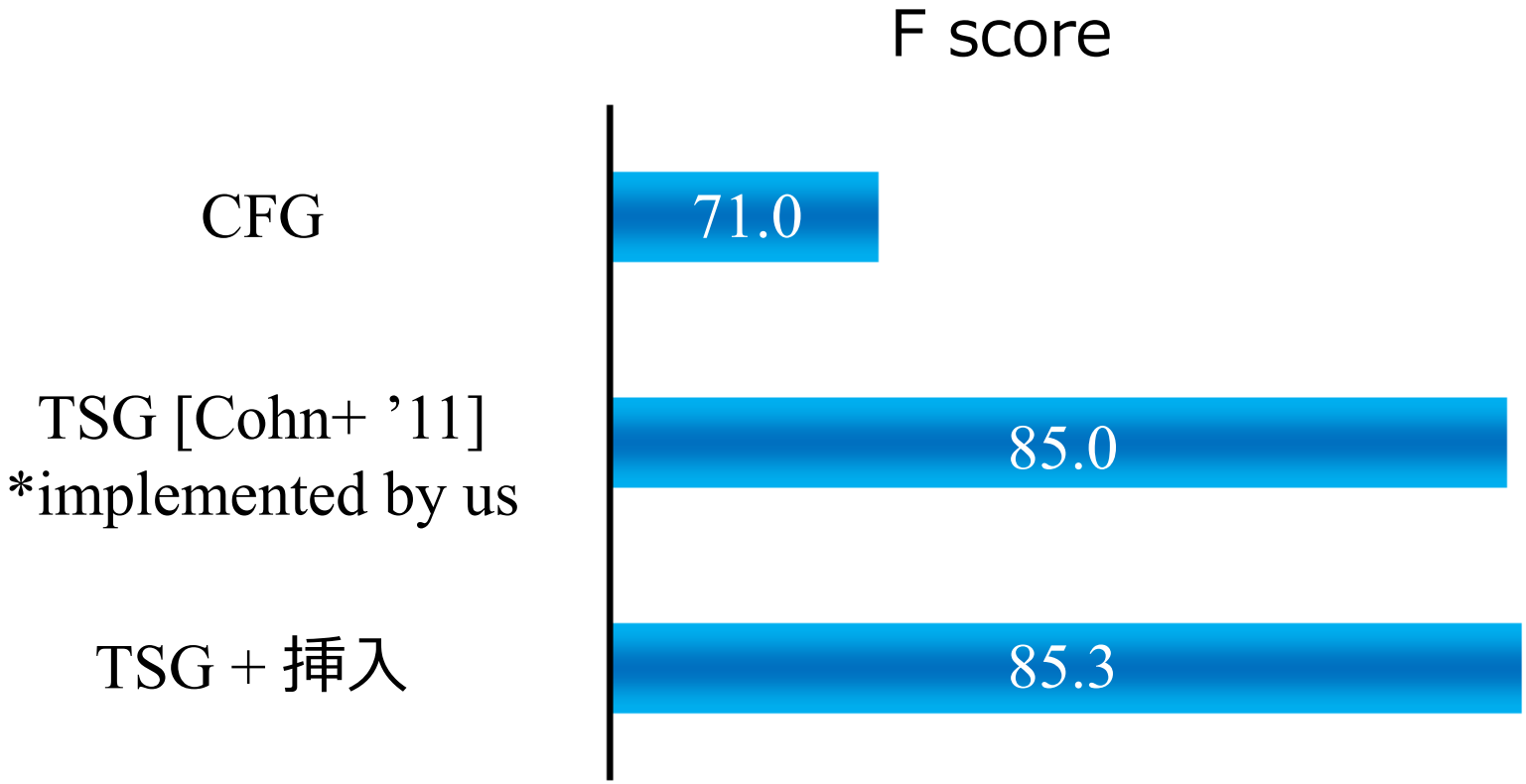
獲得された部分木を用いて構文解析実験



WSJ Penn Treebank (training: sec. 2, test: sec. 22)

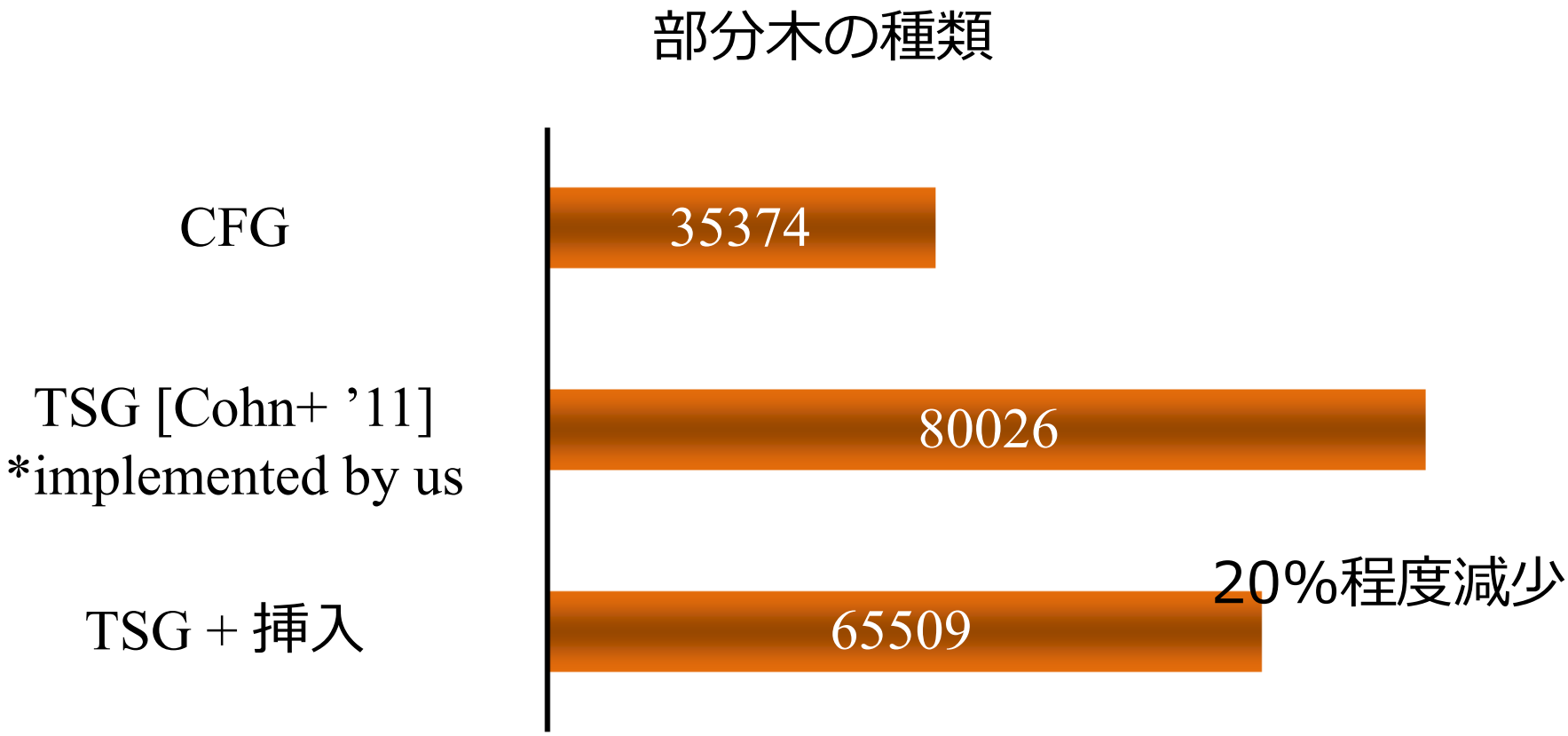
実験結果（標準データ）

獲得された部分木を用いて構文解析実験



WSJ Penn Treebank (training: sec. 2-21, test: sec. 23)

実験結果（標準データ）



WSJ Penn Treebank (training: sec. 2-21, test: sec. 23)

実験結果

挿入操作によって獲得された部分木

(NP (NP) (: -)

(NP (NP) (ADVP (RB respectively)))

(PP (PP) (, ,))

(VP (VP) (RB then))

(VP (VP) (RB not))

(QP (QP) (IN of))

(S (S) (: ;))

まとめ

木置換文法に挿入操作を導入

- Pitman-Yor Process に基づく部分木の確率モデル
- 動的計画法を利用した効率的な学習法

実験結果

- 少量の学習データでは，挿入操作により構文精度向上
- 学習データが増加すると，TSG とほぼ同等の構文解析精度
- 部分木の種類：20%程度減少