3R-01 PGLR法を用いた構文木付きコーパスの自動構築

白井清昭　　今井宏樹　　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

1 はじめに

統計的自然言語処理においてしばしば問題となるのが，学習に用いる言語資源が十分にないというデータスパースネス問題である．特に，様々な言語資源の中でも，自然言語文に構文木が付与された構文木付きコーパスは，確率一般化LR法 [1] (Probabilistic Generalized LR Method，以下PGLR) のような構文構造に関する統計情報を始め，様々な種類の統計情報の学習に用いることができる．実際，英語においてはPenn Tree Bank [4]，日本語においてはEDRコーパス [5] や京大コーパス [2] など，構文的な情報が付加されたコーパスの公開が進んでいる．ところが，単語の形態素区切り及び各形態素の品詞を付与した品詞付きコーパスなどに比べて，構文木付きコーパスは例文に構文木までも付与しなければならないために作成コストが高く，大規模なコーパスを全て人手によって作成することは難しい．

構文木付きコーパスの作成コストを削減する一つの方法として，平文に対して構文木を付与する代わりに，例文に対して自動的に構文木を付加してからその構文木を修正することが挙げられる．人手修正前に付与された構文木が正しければ正しいほど，コーパス全体の構文木の修正に要する作業量は減少する．本研究では，人手による後修正を行うことを前提に，その人的負担を軽減するために，高い精度で自動的に構文木を付与することを目的とする．

2 テキストへの構文木の自動付与

本研究では，例文に構文木を自動的に付与するために，パーザとPGLRモデルの2つを用いる．前者は例文に付与する構文木の候補を生成するために，後者は生成された各構文木の候補の確率を計算し，例文に付与するべき最も尤もらしい構文木を選択するために用いられる．

ここで問題となるのは，PGLRモデルをどのように学習すればよいのかということである．学習用言語資源として構文木付きコーパスを用いるようないわゆる教師ありの学習を行うことはできない．なぜなら，ここではPGLRモデルを学習することのできる構文木付きコーパスは存在しないことを前提とし，そのような構文木付きコーパスを構築することを目的としているからである．

Automatic Construction of a Structural Annotated Corpus using the PGLR model
SHIRAI Kiyoaki, IMAI Hiroki, TOKUNAGA Takenobu, TANAKA Hozumi
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama Meguro-ku, Tokyo, 152 JAPAN

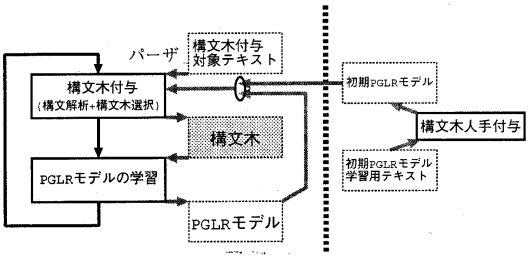

図 1: 構文木の自動付与

そこで本研究では，(1) 人手によって付与された構文木から初期PGLRモデルを学習し，また (2)PGLRモデルの学習と構文木付与を交互に繰り返すことによってこの問題を解決する．本手法における構文木自動付与の流れを図1に示す．

初期PGLRモデルの学習

少量の文 (以下，これを初期PGLRモデル学習用テキストと呼ぶ) に対して人手によって構文木を付与し，それらから初期PGLRモデルを学習する．PGLRモデルと同じく構文的な統計情報の学習を目的としたInside-Outsideアルゴリズムによる確率文脈自由文法のパラメタ推定の場合，教師なし学習 [3] よりも教師あり学習 [6] の方が優れた結果を残していることなどから，本研究ではたとえ少量でも教師データを人手で作成した方がよいと考える．

PGLRモデルの学習と構文木付与の反復

パーザを用いてコーパスの各例文 (以下，構文木付与対象テキストと呼ぶ) を構文解析し，1.の初期PGLRモデルによって1位の確率が与えられる構文木を付与する．さらに，構文木付与対象テキストに付与された構文木をもとにPGLRモデルを再学習し，再学習されたモデルをもとに再び構文木の付与を行う．また，この作業を繰り返し行う．

初期PGLRモデルの学習データ量は十分ではないので，大規模な構文木付与対象テキストを利用し，それらに付与された構文木をPGLRモデルの学習に用いる．また，PGLRモデルの学習とテキストへの構文木の付与を交互に繰り返すのは，初期PGLRモデルでは学習量が少ないために見落されていた構文的な統計情報が，上記のような反復を繰り返すうちにPGLRモデルに反映されることを期待している．

3 評価実験

2節で提案した手法を評価する予備実験を行った．

構文木付与対象テキストとしてEDRコーパス [5] を使用した．EDRコーパスには既に構文木が付与されているが [1]，我々はこのコーパスの例文に新たに構文木を付与することを想定して実験を行った．実験の手順を以下に示す．

1. EDRコーパスからランダムに N_{init} 個の文を選択して初期PGLRモデル学習用テキストとした．また，これらの例文に対して人手によって構文木を付与したとして，EDRコーパスに付与された構文木から初期PGLRモデルを学習した．
2. EDRコーパスから1.とは異なる10,000文をランダムに選択して構文木付与対象テキストとした．これらを構文解析し，1.で学習した初期PGLRモデルによって1位の確率が与えられる構文木を例文に付与した．また，各例文に対して，PGLRモデルによる確率が上位 T_{top} 位の構文木をもとにPGLRモデルを再学習し，再学習されたPGLRモデルによって1位の確率が与えられた構文木を付与することを繰り返した．

2.における反復過程において，EDRコーパスにおける構文木を正しい構文木とみなし，構文木付与対象テキストに付与された構文木がどれだけ正しいかを評価した．評価尺度として括弧付けの適合率 [6, 7] (正しい構文木中の全ての括弧付けと矛盾しない括弧付けの割合)を用いた．結果を図2,3に示す．

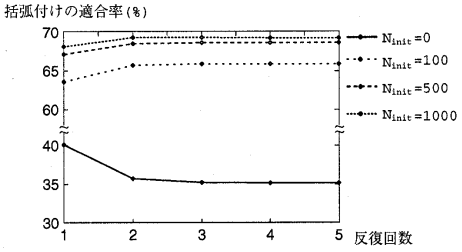

図 2: N_{init} と括弧付けの適合率の相関 ($T_{top} = 1$)

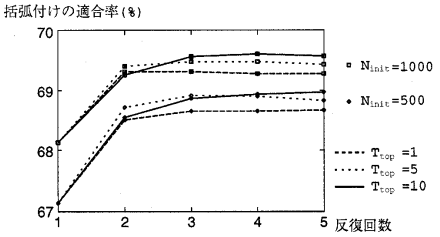

図 3: T_{top} と括弧付けの適合率の相関

図2は $N_{init} = 0, 100, 500, 1000$ のときの括弧付けの適合率の変動を示している．N_{init} の値を大きくするにつれて，括弧付けの適合率も向上することがわかる．特に $N_{init} = 0$ とそれ以外の場合とで括弧付けの正解率が大きく異なる．このことから，高い精度でコーパスに構文木を付与するためには，たとえわずかな数の例文でも初期PGLRモデルを学習するための構文木を人手によって付与した方がよいといえる．一方，PGLRモデルの学習と構文木の付与を反復した結果，$N_{init} = 0$ の場合を除き，括弧付けの適合率が上昇することがわかる．しかしながら，2回程度の反復で適合率の上昇は収束し，適合率の著しい向上は見られなかった．

図3は T_{top} の値を変化させたときの括弧付けの適合率の変動を示している．この結果から，T_{top} の値を大きくすればするほど括弧付けの適合率は向上し，またその上昇もゆるやかであることがわかる．これは T_{top} の値を大きく設定することにより，初期PGLRモデルを学習する際にあまり出現しなかった事象の確率を高く推定するように学習が行われたためと推察できる．

4 おわりに

構文木付きコーパス作成時の人的負担を軽減することを目的に，テキストに対して高い精度で構文木を付与するための手法として，初期PGLRモデルを学習するための少量の構文木を人手で作成すること，およびPGLRモデルの学習と構文木付与を交互に繰り返すことがある程度有効であることを予備実験により確認した．

本手法により自動付与した構文木は人手によって修正されるが，人手修正された構文木が増えるにつれて，これらを用いてPGLRモデルを再学習することにより，人手修正された構文木が付与されていない例文に対してさらに高い精度で構文木を自動付与できると考えられる．今後は，このような本手法を改良する手法をいくつか考案し，構文付きコーパスの構築に応用していきたい．

参考文献

[1] K. Inui, V. Sornlertlamvanich, H. Tanaka, and T. Tokunaga. Probabilistic glr parsing: A new formalization and its application to structural disambiguation. 自然言語処理, Vol. 5, No. 3, pp. 33–52, 1998.

[2] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会論文集, pp. 58–61, 1997.

[3] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer speech and languages*, Vol. 4, pp. 35–56, 1990.

[4] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.

[5] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第2版. Technical Report TR–045, 1995.

[6] F. Pereira and Y. Schabes. Inside-Outside reestimation from partially bracketed corpora. In *Proceedings of the ACL'92*, pp. 128–135, 1992.

[7] 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.

[1] 但し，構文木の内部ノードには文脈自由文法の非終端記号に相当するような構成素ラベルは付与されていない．