構文木コーパスの再構成手法*

7G－1

田代敏久　柏岡秀紀　Ezra W. Brack†
ATR音声翻訳通信研究所‡

1 はじめに

近年、コーパスに基づく自然言語処理が注目を浴びている。頑健で高精度の構文解析の研究には、構文木コーパスを大量に収集し、文法の獲得、既存の文法の再訓練、解析結果の評価等に使用することが望ましい。しかし、構文木コーパスには、どのような文法理論に基づくかにより、様々なバリエーションが存在する。

例えば、日本語の統語論は、主として文節間の係り受けを中心とした統語論(仮に「文節文法」と呼ぶ)と、統語的な主部と補部との関係を中心とした統語論(仮に「一般的な句構造文法」と呼ぶ)とに大別できる。これらの文法による構文木の構造は、図1が示すように、お互いに極めて異なっている。また構文木の構造以前の問題として、単語の分割基準や品詞体系も異なることが多い。

そこで、本報告では、異なる文法理論に基づく構文木コーパスをできるだけ容易に有効利用するために、2つの異なる構文木を調整するための規則（構文木調整規則）を自動的に獲得する手法を提案する。また、獲得した規則によるコーパスの書き換え実験の結果について報告する。

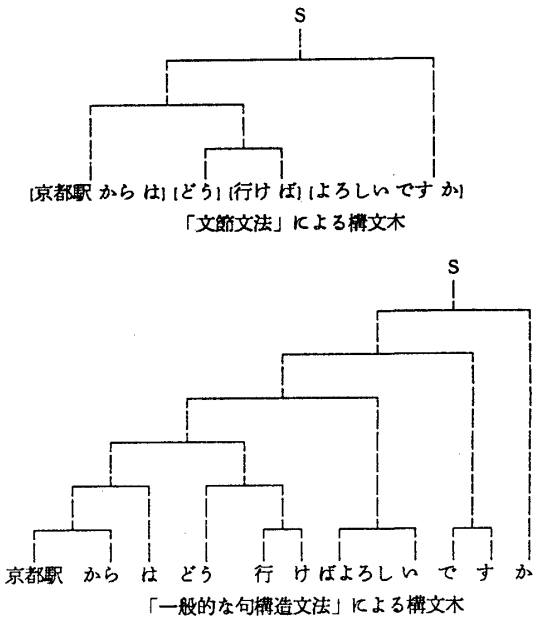

図 1: 構文木の多様性

2 構文木コーパスの再構成

2.1 コーパスの再構成手順

構文木コーパスの再構成は、田代[1]によって提案された形態素情報コーパスの再構成と同様の手法で行なう。この手法は、

1. 同一のテキストに2種類の構文木を付与し、訓練集合を作成する。
2. 訓練集合から構文木調整規則を抽出する。
3. 構文木調整規則を用いてコーパスを書き換える。

という3つのステップから構成される。

2.2 構文木調整規則の獲得

前述の3つのステップの内、もっとも重要かつ困難なのは構文木調整規則の抽出である。我々は構文木調整規則を獲得するために、Eric Brill[2] の Transformation-Based Error-Driven Parsing の手法を応用した。この手法は、コーパスに基づく手法であるが、1) 確率・統計的な知識ではなく、記号的な知識(＝書き換え規則)をコーパスから抽出できる、2) 比較的少数のコーパスで学習できる、という特徴がある。

2.2.1 Brill の手法

先ず、Brill の書き換え規則抽出手法を説明する。この手法は次のような欲張り法(greedy search)を利用している。

1. 「正解ブラケット集合」、「初期ブラケット集合」[1]、「書き換え規則のテンプレート集合」、「正解ブラケットとの比較を行なう評価関数」、を用意する。
2. 「初期ブラケット集合」を現在着目中の(current)ブラケット集合とする。
3. 現在着目中のブラケット集合に対し、すべての書き換え規則を適用してみる。正解ブラケット集合と評価関数を適用しもっともスコアが向上した規則を採用する。どの規則を適用してもスコアが向上しなければ、終了する。
4. もっともスコアが向上した規則を適用した後のブラケット集合を現在着目中の(current)ブラケット集合とする。
5. 3へ戻る。

以上の手続きにおいて、「正解ブラケット集合」をある文法体系で作成された構文木コーパスとし、「初期ブラケット集合」を別の文法体系で作成された構文木コーパスとすれば、2つの異なる文法体系で作成された構文木コーパスの調整規則が抽出できる。

*A Method of Restructuring Bracketed Corpora

†Toshihisa Tashiro, Hideki Kashioka, Ezra W.Black
‡ATR Interpreting Telecommunications Research Laboratories

[1]構文木コーパスに含まれる文を、機械的に、かつ単純(naive)に解析した構文木(ブラケット)。Brill は英語を解析対象としているので、単純な右枝分れ構造を初期ブラケットとしている。

正:(A B C)　誤:(A (B C))

Brill の手法	我々の手法
・書き換え不能 (crossing 不変)	・AB 間の bracket 削除 (precision 増大)

正:(A (B C))　誤:(A B C)

Brill の手法	我々の手法
・書き換え不能 (crossing 不変)	・AB 間に bracket 挿入 (recall 増大)

正:(A (B C))　誤:((A B) C)

Brill の手法	我々の手法
・AB 間のブラケット移動 (crossing が減少)	・AB 間のブラケット削除 (precision 増大) ・AB 間にブラケット挿入 (recall 増大)

表 1: Brill の手法との相違点

2.2.2 任意の木構造の書き換え

Brill[2] により提案された書き換え規則は、二分木の構造を変化させるだけの機能しか持っていない (書き換え規則を何回適用しても、ブラケットの総数は変化しない)。しかし、構文木コーパスは、必ずしも二分木で表現されているとは限らない。そこで、我々は任意の木構造 (ブラケット) を調整できるように、次のような 12 種類の書き換え規則のテンプレートを用いることにした。

{ADD | DELETE} {LEFT | RIGHT} BRACKET
{BEFORE | BETWEEN | AFTER} TAG-X,TAG-Y

書き換え規則の変更に伴い、評価関数も変更した。Brill は Black[3] により定義された構文木の評価方法の内、crossing error rate のみを用いている。我々はこれを改め、より適切な評価基準である再現率 (recall) と適合率 (precision) を用いることにした。表 1に、Brill の手法との相違点を示す。

2.2.3 単語分割・品詞体系の差異への対応

Brill の手法においては、品詞情報を利用するのは、書き換え規則の適用時のみなので、2 つのコーパスの品詞体系が異なっていても問題ない。一方、Black の構文木の評価方法は、英語を想定して作成されたため、2 つの構文木の終端記号の数が同じことを期待している。そこで、我々の実験では、単語を文字のリストに展開して対応した。

3 実験

実験は、ATR の対話データベース [4] に付与されている係り受け情報を基に作成した文節文法による構文木と、ATR の音声翻訳システムの音声言語パーザ [5] で用いられている単一化に基づく日本語句構造文法による構文木を対象とした。

まず、ＡＴＲの対話データベースベースから 300 文を選び、2 種類の構文木情報を付与し、内 100 文を訓練集合、200 文を試験集合とした。200 文の平均文字長は、17.1 文字であり、内 70 文は 20 文字以上の文である。

次に訓練集合に付与された 2 つの構文木情報から構文木調整規則を抽出した。なお、書き換え実験は 2 つの構文木間双方向で行なった。

また、元の構文木の情報を使わずに構文解析を行なった場合と比較するために、機械的に作成した左枝分れ構造からも調整規則を抽出した。

表 2は、句構造文法の構文木から文節文法への構文木への書き換え実験の結果である。調整規則の適用により、元の構文木間の差異がかなり減少したことが分かる。また、左枝分れ構造からの学習の結果と比較すると、文字長の影響が少なくなっている。これは、我々の手法が、元のコーパスの情報を有効利用していることを示していると思われる。

	実験対象	再現率	適合率
学習なし	オープン (全体)	51.8%	31.2%
コーパスから学習 (241 ルール)	オープン (全体)	82.5%	86.4%
	(２０文字未満)	87.7%	92.1%
	(２０文字以上)	79.6%	83.2%
	クローズ	93.8%	96.5%
左枝分れから学習 (343 ルール)	オープン (全体)	62.0%	69.3%
	(２０文字未満)	74.2%	82.4%
	(２０文字以上)	55.1%	61.7%
	クローズ	84.1%	91.5%

表 2: 句構造文法から文節文法

	実験対象	再現率	適合率
学習なし	オープン (全体)	31.2%	51.8%
コーパスから学習 (289 ルール)	オープン (全体)	68.8%	72.9%
	(２０文字未満)	72.9%	72.8%
	(２０文字以上)	66.5%	72.9%
	クローズ	86.7%	87.8%
左枝分れから学習 (335 ルール)	オープン (全体)	61.6%	63.9%
	(２０文字未満)	74.8%	71.3%
	(２０文字以上)	54.1%	59.0%
	クローズ	78.7%	79.9%

表 3: 文節文法から句構造文法

表 3は、逆方向への書き換え実験の結果である。この実験においても、前述の実験と同様な結果が得られた。しかし、句構造文法の構文木は、文節文法の木とくらべ複雑なため、全体の数値は低くなっている。

4 おわりに

今後はさらに大規模な実験を行なうと共に、英語のコーパスの書き換えも行なう予定である。

参考文献

[1] Tashiro, T., Uratani, N., Morimoto, T, "Restructuring Tagged Corpora with Morpheme Adjustment Rules", COLING94, 1994.

[2] Brill, E., "Automatic Grammar Induction and Parsing Free Text: Transformation-Based Error-Driven Parsing," ACL93, 1993.

[3] Black, E., et al, "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars", DARPA Speech and Natural Language Workshop, 1991.

[4] Sagisaka, Y., Uratani, N., "ATR Spoken Language Database," The Journal of the Acoustical Society of Japan, Vol. 48, 12, pp. 878-882, 1992

[5] Nagata, M. and Morimoto, T.: "A Unification-Based Japanese Parser for Speech-to-Speech Translation", IEICE Trans. Inf. & Syst., Vol.E76-D, No.1, pp.51-61 ,1993.