自然言語処理 128－9
（1998. 11. 5）

構造つきコーパスの共有化に関する一考察

植木正裕, 白井清昭, 徳永健伸, 田中穂積

{ueki,kshirai,take,tanaka}@cs.titech.ac.jp

近年、構文情報つきコーパスが利用可能になり、言語知識や統計情報の獲得に役立っている。しかし、形態素情報つきコーパスに比べて作成が困難なことから、量的には十分とは言えない。また、既存の複数のコーパスの間には、単語境界の認定や品詞体系の粒度などに違いがあることから、獲得した情報を容易には統合できないという問題もある。本論文では、異なる品詞体系でも共有可能な構造情報についての考察を行なう。構造情報を、品詞体系の違いに影響されない品詞体系非依存部分と、品詞体系の違いを吸収するための品詞体系依存部分に分け、品詞体系依存部分の文法のみを入れ換えて形態素・構文解析を行なうことで、コーパスへの形態素情報・構造情報の付与を行なう。

Sharing Syntactic Structures

UEKI Masahiro, SHIRAI Kiyoaki, TOKUNAGA Takenobu, TANAK Hozumi

Department of Computer Science

Tokyo Institute of Technology

{ueki,kshirai,take,tanaka}@cs.titech.ac.jp

Bracketed corpora are a very useful resource for natural language processing, but hard to build efficiently, leading to quantitative insufficiency for practical use. Differences in morphological information, such as word segmentation and part-of-speech tag sets, are also troublesome. An application specific to a particular corpus often cannot be applied to another corpus. In this paper, we sketch out a method to build a corpus that has a common syntactic structure and sets of morphological information based on different tag set schemes. Our system uses a two layered grammar, one layer of which is made up of replaceable tag-set-dependent rules while the other has no such tag set dependency.

1 はじめに

近年、自然言語処理の分野では、大規模なコーパス資源が利用可能になったことから、コーパスを利用したさまざまな研究が行なわれている。しかし、現状では、各コーパスごとに品詞体系が異なり、日本語の場合には単語単位の認定のしかたにも違いがあるという問題がある。このため、複数のコーパスが利用可能であっても、それぞれを別個に利用することしかできない。

これに対して、品詞体系の標準化の試みも行なわれているが、既存の資源を有効に利用するためには、新たに品詞タグをつけ直すか、標準品詞体系との間での変換を行なわなくてはならない。

複数のコーパス資源の共有化の試みとしては、コーパス間の形態素情報のマッピングを行ない、品詞や単語境界を変換する研究が行なわれている [4][5]。これらの研究では、人手で作成した書き換え規則を用いたり、1つの文に対して2種類の品詞体系による形態素情報を付与し、そこから自動的に抽出した書き換え規則を用いたりして、複数のコーパス間での相互変換を行なう。書き換え規則は、単語対単語、あるいは句対句で記述される。異なるコーパスで品詞分類の粒度に差がある場合には、単純な品詞間のマッピングは一意には決まらないため、変換規則は個々の単語ごとに記述される。

しかし、書き換え規則による形態素情報の変換には次のような問題がある。

1. 多品詞語
   通常の形態素解析で生じるような多品詞語だけでなく、品詞分類の粒度の粗いものから細かいものへの変換を行なう際に生じる一対多の変換も多品詞語とみなすことができる。
   これは隣接する単語間での n-gram などを利用することである程度は解消できるが、機能によって細分類された助詞 (格助詞や準体助詞など) などは、一意に決定するのが難しい。
2. 未知語
   書き換え規則中に存在しない語は未知語となる。
   [4] では、品詞が一致する単語での単語対単語の書き換え規則、あるいは、単語長と品詞が一致する語の含まれる句での句対句の書き換え規則を用いることで、これを解決している。句対句の書き換え規則の利用は、隣接する接頭語や接尾語から固有名詞の推定を行なう手法 ([2]) と類似している。しかし、単語対単語の書き換え規則を用いる場合には、前後の情報を用いずに品詞から品詞へのマッピングを行なうため、品詞分類の粒度の粗いものから細かいものへの変換ではノイズを多く含む可能性がある。

[4] では、書き換え規則を適用した結果得られる形態素ラティスは、通常の形態素解析によって得られる形態素ラティスに比べて曖昧性が減ったと報告している。しかし、曖昧性を減らす要因となっているのは、名詞をむやみと短単位に分割しなくなったことであり、多品詞語のところで述べたような機能語の細分類に関してはほとんどの曖昧性が残ってしまっている。

このような機能語の細分類は、係り先が決まることで決定できることが多いことから、構文情報を利用することで、品詞体系の変換の精度を上げることも可能である。

本研究では構文情報つきコーパスを利用した品詞タグの変換を行なう手法を提案する。本手法では、構文情報を係り受け括弧つき文に変換し、これを構文解析システムの入力とする。また、解析に使用する辞書や文法は、品詞体系の異なるものに自由に入れ換えることができる。これにより、品詞体系に依存しない共通の構文情報に対して、複数の品詞体系による形態素情報が付与されたコーパスが作成できる。

2 二階層文法

本研究では、品詞体系に依存しない共通の構文構造を記述する文法と、品詞体系の違いを吸収して共通の構文構造とのつながりを記述する

文法とで構成される、二階層文法を使用する。

このような二階層文法を作成するためには、品詞体系への依存／非依存の違いが明確に分離できなくてはならない。ここで図1と2のような2つの構文木を考えてみる。図中の黒丸は、兄弟ノードに単語が含まれず、すべて中間ノードであるようなノードを示している。

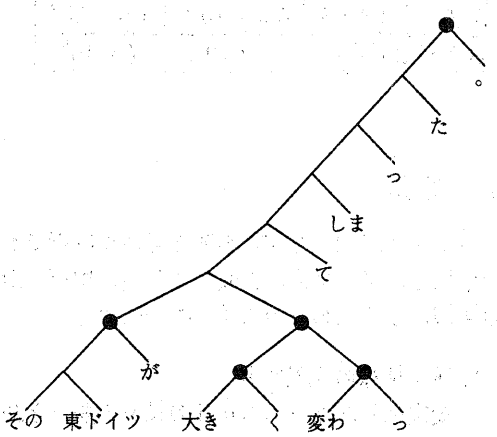


図 1: 構文木の例 (1)

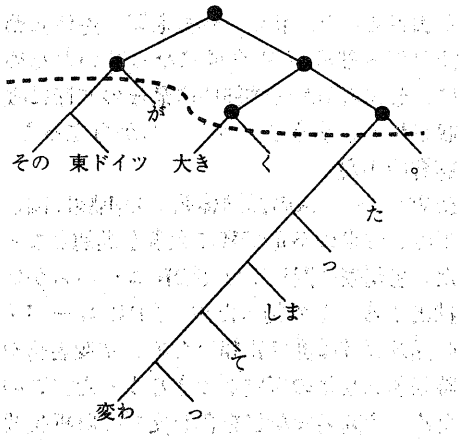


図 2: 構文木の例 (2)

図1ではアスペクトやテンスを表す表現が1つずつ外側に付加された構造になっているため、ほとんどの文法規則の右辺に単語が含まれる。つまり、品詞体系の違いがそのまま文法の違いとなり、品詞体系の交換は容易ではない。

一方、図2では、ルートノード近くのノードにはすべて黒丸がついていて、破線部分の上と下ではっきりと違いがあることがわかる。黒丸のついたノードは文節 (列) を表すノードであり、破線より上の構文木は文節間の係り受け構造を表している。これに対して、破線より下の構造は、それぞれの文節内という狭い範囲での統語構造を表し、品詞体系の違いはこの文節内の統語構造にしか影響しないと考えられる。

このことから、本研究では文節文法に基づいた構文情報をコーパスに付与する。文節文法は、文節間の係り受け構造を表す**文節間文法**と、文節内の統語構造を表す**文節内文法**の2つに大別される。新しい品詞体系でのタグづけは、文節内文法のみを入れ換えることによって行なう。

3 実験

構文情報としては EDR 日本語コーパス [1] を利用した。EDR コーパスに付与されている構文情報は図1に近いため、そのままでは文節文法での括弧づけと交差してしまう。幸い、EDR コーパスの構文情報の各ノードには、図3に示すように、合成関係を表すラベルが付与されている。

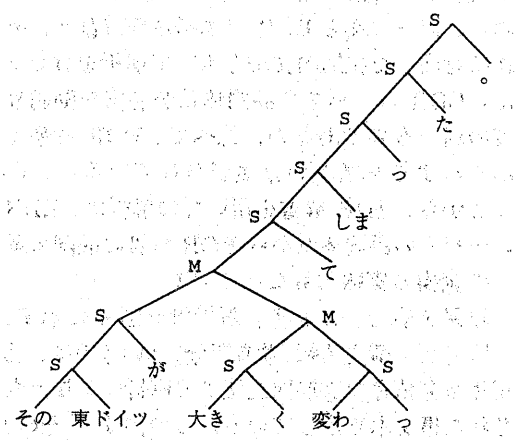


図 3: 合成関係

S 合成は自立語と附属語の合成関係であり、M 合成は係り受けや連体・連用修飾を表す合成関係である。M 合成の外側にある S 合成は、M 合成でカバーされた部分木の最右の S 合成、すなわち、直前にある部分木とマージすること

で、図1の構文木を図2の構文木に変換できる。実験ではこのようにして変換したものを入力として用いた。

また、新しく付与する品詞体系としてはRWCコーパス[3]で用いられている品詞体系を用いた。RWCコーパスは、EDRコーパスと違い、それに対応した辞書がないため、コーパスに出現した単語をそのまま辞書として用いた。ただし、固有名詞をはじめとする名詞類は、十分な語数がカバーできないと考えられるため、EDRコーパスから抽出した名詞類を辞書に追加した。

文法に関しては、文節間文法は各文節に係り属性と受け属性(連用修飾{する, される}など)を付与し、係り属性と受け属性の組合せが適切で、係り受けの交差がない、すべての構文木を生成する文法を使用した。文節内文法は、RWCコーパスの品詞体系に対して、形態素列を文節にまとめあげる規則だけを記述した。この際、文節間文法での解析に必要となる係り受け属性が文節に付与されるようにした。

また、今回の実験ではEDRコーパスの構文情報を文節文法に即した形式に変換したため、構文情報の変換の正しさなどの検証をするために、同一の文をEDR日本語単語辞書と、それに対応した文節内文法を用いて解析を行なった。EDRコーパスの品詞体系が名詞や動詞程度の粗いものであるのに比べて、EDR辞書ではそれよりも細かい分類がされている。このことから、EDR辞書を用いての解析は、EDRコーパスの品詞体系からEDR辞書の品詞体系への変換の実験にもなっている。

結果を表1、および、解析例を図4に示す。

図4で、構文情報(共有部分)というのは、品詞体系非依存の文節間文法での解析により生成された構文木を表している。この例は、「その」「東ドイツが」「大きく」「変わって」「しまった。」の5つの文節によって文が構成されている。また、構文情報(EDR辞書品詞体系)および構文情報(RWC品詞体系)は、それぞれの品詞体系に依存した文節内文法での解析により生成された文節内統語構造を表している。なお、1形態素で構成される文節については省略してある。

表1: 解析結果

品詞体系	EDR	RWC
解析終了	69,504	51,810
解析失敗	80,338	98,032
構文木数(最少)	1	1
構文木数(最多)	5.9×10^{10}	1.6×10^{9}
構文木数(平均)	1×10^{6}	7×10^{5}

4 考察

表1を見ると、EDR辞書を用いた解析でも半分以上の文で解析に失敗している。解析失敗の原因は主に次のようなものであった。

1. 構文情報の変換誤り
   変換誤りはコーパス自体の情報の誤りと思われるものが多い。M合成の外側にあってS合成されていた単語を、直前の部分木に組み込んだ際に、括弧づけに交差が起こった例があった。特に、形式名詞に連体修飾節が係る部分がS合成になっていたために、形式名詞が修飾節中の最後の文節に吸収されてしまうというケースが目立った。
2. 辞書の不足
   RWCコーパスの品詞体系での実験に関しては、辞書の不足が特に大きな問題となった。名詞類に関してはEDRコーパスから抽出することで補ったが、EDRコーパスの品詞づけは非常に粗いため、サ変名詞や時相名詞などの区別ができなかった。このため、サ変名詞などを含む文での解析失敗が目立った。また、RWCコーパスの品詞体系では、動詞を語尾つきで1形態素としているため、コーパス中に出てこなかった活用形のエントリーが不足するという問題もあった。
3. 文法のカバレージの不足
   表1の結果を比べると、EDR辞書の品詞体系を用いた解析の方が、RWCコーパスの品詞体系を用いた解析に比べて、解

析が終了した文の数が2万近く多い。しかし、解析できた文を比較すると、EDR 辞書の品詞体系では解析できなかったのに、RWC コーパスの品詞体系では解析できたものが12,926 もあった。EDR の品詞体系でこの約1万文が解析できなかった理由の1つとして文法のカバレージの不足が挙げられる。今回の実験では、文節内文法の記述の際には、特に相互の参照をせず、別個に作成を行なった。そのため、一方の文法では受理されるが、他方では受理されないというケースがみられた。

このうち 3 などについては、文法記述の精度を上げることで、今後改良できると考えられる。

2 に関しては、名詞類に関して特に不足が目立ったことから、内容語は変換規則で単純に置き換え、機能語は構文情報から特定するという方法も考えられる。

また、括弧つき入力文の解析に関してもいくつかの問題があった。

図 4 の解析例を見ると、「変わってしまった。」の部分が 2 つの文節に分解されていることがわかる。入力文の括弧つきでは、図 2 のように「変わってしまった。」全体で 1 つの文節を構成することを期待している。

括弧つきの入力文の解析では、括弧の付与された範囲をちょうどカバーする部分構文木ができる解析だけを受理し、括弧と交差が生じるような解析は受理しない。しかし、今回の実験では、この部分構文木がちょうど 1 つの文節をカバーするという制限を設けなかったため、この例のように 2 つ以上の文節に分解されたものがあった。また、これとは逆に、1 つの文節が 2 つ以上の括弧を含んでいたものもあった。

5 結論

構文解析に用いる文法を、品詞体系非依存の文節間文法と品詞体系依存の文節内文法とに分けることで、文節間の係り受け関係を表す共有の構造情報と、品詞体系ごとに異なる文節内の統語構造情報および形態素情報を付与したコーパスを作成することができた。

助詞などの機能による細分類は、係り先の情報が利用できることでうまく選択することができたが、複合名詞などは通常の形態素解析と同じ曖昧性が生じてしまった。今後は、規則による書き換えと、構文情報を利用した解析の両者の利点を組み合わせた手法を考察したい。

また、構文解析システムへの入力文に、構文情報を表す括弧を付与することで、構文情報を制約として用いたが、期待していたほどの制約がかからなかった。この点についても、構文情報の与え方と利用の仕方の両面から検討を行ないたい。

参考文献

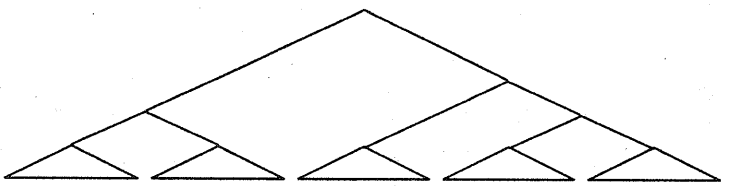
[1] EDR. 電子化辞書仕様説明書 第2版. Technical report, 日本電子化辞書研究所, 3 1995.

[2] 木谷強. 固有名詞の特定機能を有する形態素解析処理. 情報処理学会 自然言語処理研究会, Vol. 90, No. 10, pp. 73–80, 7 1992.

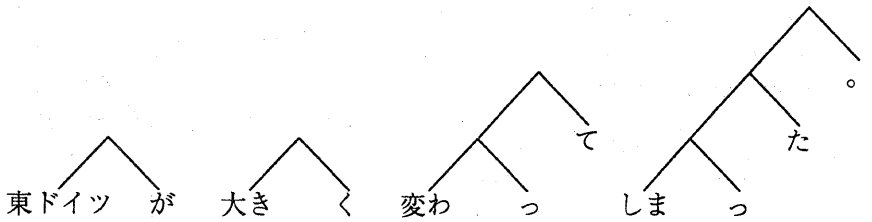
[3] テキストグループデータベースワークショップ. RWC テキストデータベース報告書. Technical report, 技術研究組合 新情報処理開発機構, 3 1997.

[4] 田代敏久, 森元逞. 形態素情報付きコーパスの再構成手法. 情報処理学会論文誌, Vol. 37, No. 1, pp. 13–22, 1 1996.

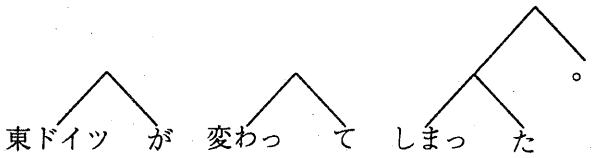
[5] Simone Teufel. A support tool for tagset mapping. In *Workshop SIGDAT (EACL 95)*, 1995.



構文情報 (共有部分)



構文情報 (EDR 辞書品詞体系)



構文情報 (RWC 品詞体系)

	EDR コーパス	EDR 辞書品詞体系	RWC 品詞体系
その	連体詞	連体詞	連体詞
東ドイツ	名詞	固有名詞	名詞 固有名詞
が	助詞	格助詞	助詞 格助詞
大き	形容詞	形容詞語幹	形容詞 連用テ接続
く	語尾	形容詞語尾 連用形く	
変わ	動詞	ラ行五段動詞語幹	動詞 五段・ラ行
っ	語尾	ラ行五段動詞語尾 連用音便形	連用タ接続
て	助詞	接続助詞	助詞 接続助詞
しま	動詞	ワ行五段動詞語幹	動詞 五段・ワ行促音便
っ	語尾	ワ行五段動詞語尾 連用音便形	連用タ接続
た	助動詞	助動詞た 終止・連体形	助動詞 特殊型 見出し形
。	記号	句点	記号

形態素情報

図 4: 解析例