# Windows-Native な英文品詞タグ付けツールの作成

後藤　一章
大阪大学大学院
k-goto@gs.lang.osaka-u.ac.jp

## 1. はじめに

コーパスに単語の品詞情報を示すタグ(品詞タグ)が付与されていることにより、そのコーパスを利用した言語研究・言語処理研究の可能性は大きく広がるとされている。そのため、現在公開されている多くの英語コーパス[1]には予め品詞タグが付与されている。

一方、各研究者が独自にコーパスを編纂する場合、品詞タグの付与は自らが必要に応じて行うことになる。しかし、コンピュータに比較的馴染みの薄い文科系の研究者にとって、品詞タグの付与は簡易な作業とは言い難い。文科系の研究者においては Windows(または Mac)の利用率が高いと思われるが、品詞付与ツール(品詞タガー)を利用するためには、多くの場合 UNIX やコマンドラインの知識が求められるからである。

そこで、本研究では、文科系研究者による英語コーパスの利用を支援することを目的とし、操作が容易な英文品詞タガーの開発に着手した。本稿では、その進捗状況の報告と評価を行う。

## 2. 背景

英文品詞タガーのタグ付けシステムは、いくつかのタイプに分類される。代表的なものに、隠れマルコフモデルと呼ばれる確率言語モデルを利用した方式がある。この方式では、事前に訓練コーパスから単語の生起確率と品詞の n-gram 生起確率を計測しておき、実際の品詞付与の際には、それらの値を基に、各単語に対する付与確率が最も高い品詞が選択される。この方式の代表的な品詞タガーには Xerox Tagger[2]がある。

これに対し、規則ベースと呼ばれる品詞タガーがある。この方式では、「冠詞の後ろに生起する単語は名詞」といった文法や文脈に基づいた数多くの規則を利用することで、最も妥当な品詞が決定される。こうした規則は従来手作業によって精緻化されてきたが、Brill(1995)で機械学習による自動的な規則の抽出法が提案された。Brill の作成した Rule-Based tagger(a.k.a. 'Brill Tagger')はフリーで公開されており、自然言語処理の分野で広く利用されている。

その他にも、上記の 2 つの方式を併用した CLAWS (Constituent Likelihood Automatic Word-tagging System)[3] などをはじめ、様々な品詞タガーが存在している。しかし、上記に挙げたものも含め、これらのタガーの大半は UNIX や DOS 上でのみ動作可能となっている。そのため、コマンドラインの処理に精通していない文科系の研究者にとっては利用が困難な状況であった。コーパス言語学の進展に伴って、文科系の研究者からも品詞タガーの需要は高まっており、よりユーザフレンドリーな品詞タガーの開発が必要であると考えられる。

## 3. Windows-Native な品詞タガー

### 3.1 開発方針

細谷(1998)の、利用者に「マウスとワープロ操作以外の技能を求めない」といソフト開発方針に従って、次の 3 つの特徴を備えた品詞タガーの作成を試みた。

- Windows-Native
- GUI 形式
- マウス操作のみで利用が可能

利用者人口の高い Windows をプラットフォームとし、インタフェースを直感的に操作が可能な GUI 形式にすることで、操作性の向上を図っている。

[1] 例えば British National Corpus や WordbanksOnline など。
[2] ftp://parcftp.xerox.com/pub/tagger
[3] http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/

また、こうした特徴はWordSmith[4]や茶器[5]などの既存のコーパス検索ツールにも備わっており、ユーザフレンドリーなツールを作成する上で重要な条件になっていると考えられる。

### 3.2 Brill Taggerに基づいた処理プロセス

本ツール(GoTagger)は、タグ付与規則を独自にカスタマイズ可能という利点を有するBrill Taggerを基に作成されている。よって、本ツールの処理プロセスはBrill Taggerとほぼ同一である。つまり、

1. LEXICONによるタグ付け
2. LEXICAL RULEによるタグ付け
3. CONTEXTUAL RULEによるタグ付け

の順で処理が実行される。

LEXICONによるタグ付けでは、各単語に対する生起確率が最も高い品詞が付与される。例えば、moveという単語の品詞候補には名詞と動詞の可能性があるが、訓練コーパス内で動詞としての頻度が高ければ、動詞タグが優先的に付与されることになる。続く、LEXICAL RULEによるタグ付けでは、訓練コーパスに生起しなかった未知語に対しての品詞付与が行われる。LEXICAL RULEの多くは、単語の形態素情報を手がかりとしている。最後に、前後の単語や品詞情報を手がかりとしたCONTEXTUAL RULEによって品詞が決定される。例えば先ほどのmoveの場合、通常は動詞タグが付与されるが、次のようなCONTEXTUAL RULEが適用される場合、名詞タグに変更される[6]。

VB NN PREV1OR2TAG DT

### 3.3 Brill Taggerとの相異

次の2つの特徴は、本ツールとBrill Taggerの重要な差異である。

<u>Brill Taggerのみに搭載されている機能</u>

タグ付け規則の自動学習

<u>本ツールのみに搭載されている機能</u>

Multiword単位によるタグ付け

Brill Taggerには、Brownコーパスと WSJコーパスから獲得された規則ファイルが予め添付されているが、これに加えて、各自の用意したタグ付きコーパスを利用して、新しい規則の自動獲得が可能である。しかし、これには、正確に品詞タグが付与された大規模なコーパスが必要となり、さらに、Brill Taggerのマニュアルによれば、24 - 72時間もの処理時間を要するとされている。こうしたコストの高さを考慮し、現時点では本ツールに規則獲得の機能は搭載していない。

一方、本ツールには、multiwordの概念が導入されている。Multiwordとは、“according to”のように、「副詞＋前置詞」と捉えるより、2語で1つの「前置詞」と捉えた方が妥当な単語の組み合わせのことである。Multiword単位のタグ付けはCLAWSなどでも採用されており、タグ付けの精度を向上させることが報告されている(Garside and Smith, 1997)。しかし、例えば“a little”が「副詞」であるか「冠詞＋形容詞」であるかは文脈次第であり、multiwordの導入には課題も残されている。そこで、現時点においては、多義性の生じにくいmultiwordのみを適用の対象としている。Multiwordを利用した場合のタグ付与結果は次のようになる[7]。

***Because/IN_M<IN> of/IN_M<IN>*** the/DT reason/NN ,_, there/EX are/VBP ***a/DT_M<JJ> lot/NN_M<JJ> of/IN_M<JJ>*** birds/NNS .

### 3.4 ツールの現状

図1・図2に示した本ツールのスクリーンショットを参照しながら、機能の概要について説明する[8]。

1. ユーザはまず、タグ付けを行うファイルを選択する。複数のファイルを同時に選択することが可能であり、異なったフォルダ内に存在するファイルも一括で処理が可能である。
2. 次に、「Settings」内で、使用する規則ファイルの選択を行う。現時点では、規則ファイルのうち、LEXICONとCONTEXTUAL RULEをカスタマイズして使用することが可能である。

---

[4] http://www.lexically.net/wordsmith/

[5] http://chasen.naist.jp/hiki/ChaKi/

[6] 各記号はVB(動詞原型)、NN(名詞)、PREV1OR2TAG(1つか2つ前の品詞)、DT(冠詞)を表す。

[7] Mは Multiwordであることを示し、Multiwordとしての品詞は<>で囲まれている。

[8] 現時点(2005/02/07)におけるヴァージョンは0.3

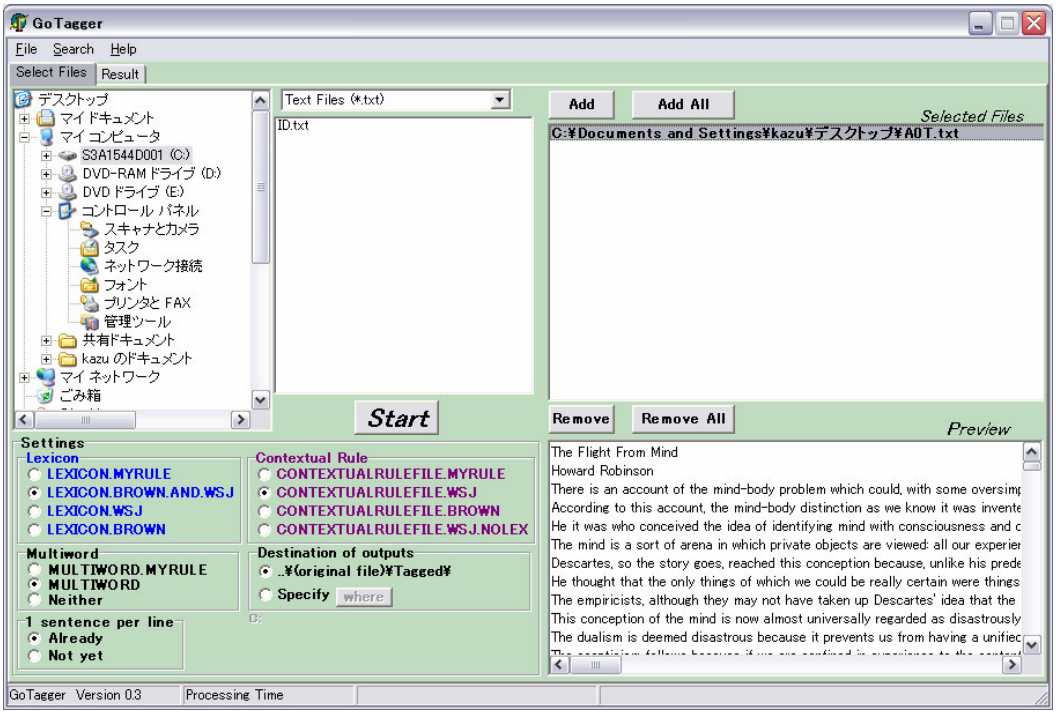


図 1　タグ付与を行うファイルとパラメータの選択画面

3. 「Multiword」の設定部分では、Multiword 単位のタグ付けを行うか否か選択することが可能である。「Destination of outputs」においては、タグ付与されたファイルの出力先を指定する。特に指定しなければ、オリジナルのファイルがある場所に tagged という下位ディレクトリが作成され、その中に出力されることになる。また、Brill Tagger と同様、本ツールでもタグ付与を行うデータは 1 行 1 文になっている必要があり、「1 sentence per line」の設定を Not yet にしておくことで、内部で 1 行 1 文に自動的に変換される。
4. START を押すとタグ付けが開始される。

選択したすべてのファイルのタグ付与が終了すると、出力ファイルを表示する図 2 の画面に切り替わる。画面右側に表示されているリストは、本ツールで使用されている Penn Treebank のタグセットである。

現時点における課題は、Multiword の種類が少数なことである。今後は辞書やコーパスを分析し、多義性の少ない multiword データの拡充を行う予定である。

なお、ツールの作成は Delphi6 Personal で行った。

### 3.5　評価

本ツールのタグ付与規則は Brill Tagger のファイルを利用しているが、実際にタグ付与を行うアルゴリズムは独自に作成されている。また、Brill Tagger に搭載されていない本ツール独自の機能として、タグ付与の事前に行う必要がある tokenization[9]や 1 行 1 文への自動変換がある。これらの理由によって、両ツールの結果にはわずかな相異が見られる。こうした差異は、例えば次のような文において現れている(B は Brill Tagger、G は本ツールを表す)。

[B] It/PRP is/VBZ different/JJ from/IN the/DT smile/NN spoken/VBN of/IN above/IN
[G] It/PRP is/VBZ different/JJ from/IN the/DT smile/NN spoken/VBN of/IN above/RB

この例においては、Brill Tagger では above に前置詞(IN)のタグが誤って付与されているが、本ツールでは適切な副詞(RB)タグが付与されている。一方、次の文では、Brill Tagger によって付与された形容詞(JJ)のタグが適切である(NN は名詞)。

[9] Penn Treebank の tokenization の規則に基づいて行われている。

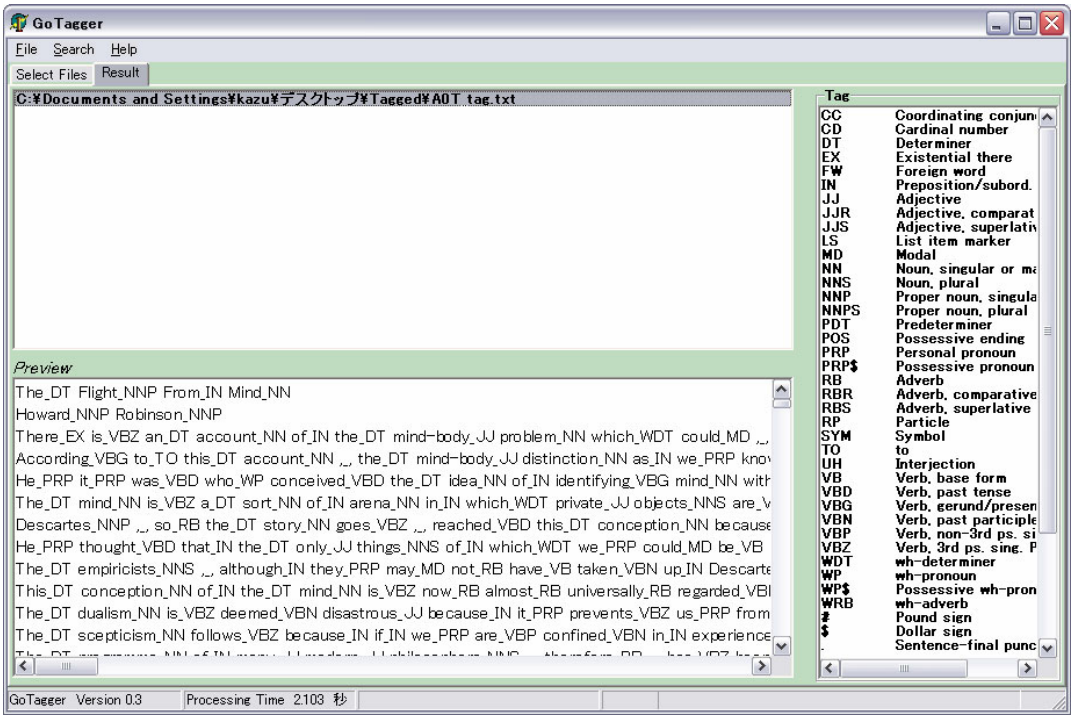


図 2　タグ付与が行われたファイルの一覧

[B] He/PRP confronts/VBZ the/DT outside/JJ world/NN
[G] He/PRP confronts/VBZ the/DT outside/NN world/NN

ただし、両ツールに付与された約 4 万語の品詞タグの比較調査を行ったところ、両タガーの一致率は 99.8%となり(表 1)、その誤差は極めて少ないと言える。

表 1　Brill Tagger と本ツールのタグの比較調査

| | 度数 | 割合 |
|---|---|---|
| 品詞タグが一致 | 42,242 | 99.8% |
| 品詞タグが不一致 | 83 | 0.2% |
| 計 | 42,325 | 100.0% |

また、本ツールがタグ付与を行う処理速度は、表 2 が示すように、約 13,000 語/秒であった[10]。

表 2　本ツールが品詞付与に要する時間

| 総語数(語) | 処理時間(秒) |
|---|---|
| 39,600 | 3.61 |
| 75,468 | 5.34 |
| 136,230 | 9.85 |

[10] WindowsXP Home Edition、PentiumM1.4GHz、512MB RAM の環境における結果。

## 4.　まとめ

本稿では、Windows-Native な英文品詞タガーの作成状況についての報告を行った。本ツールは英文品詞タガーとして新たな技術を提案するものではないが、コーパスの利用に取り組む文科系研究者の一助になるものと考える。本ツールは筆者の Web ページ[11]の「公開中のツール」においてフリーで公開されており、今後は利用者から広くフィードバックを得る予定である。

## 参考文献

細谷行輝(1998)「21 世紀の外国語教育-マルチメディア授業支援システム『新世界』の活用」、『言文だより』、大阪大学言語文化部、Vol.15.

Brill, E. (1995) "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", *Computational Linguistics*, Vol. 21(4), 543-565.

Garside R., and Smith N. (1997) "A hybrid grammatical tagger: CLAWS4", *Corpus Annotation*, In Garside et al., Longman, 102-121.

[11] http://uluru.lang.osaka-u.ac.jp/~k-goto/index.html