

コーパスからの自動学習と人手での規則作成を 融合させた形の英語品詞タガー

関根 聡

ニューヨーク大学

sekine@cs.nyu.edu

1 序文

ここ数年、コーパスに基いた自然言語処理解析システムの研究が盛んに行なわれている。それらは、教師データが付与されたコーパスから解析のための知識を抽出するというものであり、色々な応用で成功を収めている。しかしながら、著者はコーパスに基づく手法は、コーパスに基いているというその事実から精度の向上には限界があるという考えにいたっている。基本的には、コーパスは無限の言語現象のうちのかなり限られた有限のサンプルでしかないこと、現実世界や人間の知識とのリンクがないことが根本的な原因になっていると考える。このような考え方はこれまでも言及されてはきたが、この論文では代表的なコーパスに基いた手法であるトランスフォーメーション手法において具体的にどのような所に限界が見られ、どのように露見されているか、それらは人手による手法によりどのくらい救えるのかなどを検討した。そして、コーパスに基いた手法と人手によるルール作成を融合させた形の英語品詞タガーによる実験を報告する。このタガーでは、Brillのトランスフォーメーション手法でルールを自動的に抽出し、それを人手によって修正、新規ルールの追加などを行なった。Penn TreeBankによる実験では96.95%の正解率を得、13%のエラー率の向上が見られた。また、システムの出力とPenn TreeBankを比較した所、その食い違いの約1/3がPenn TreeBankのタグ付けの誤りであることが分った。

2 コーパスに基づく方法の問題点

まず、本節ではコーパスに基づく方法の根本的な問題点を列挙し、次節において、Brillの品詞タガー [Brill 1995] を例にどのような誤りが見られるかを報告する。

データ・スパースネス コーパスのサイズというのは有限であり、特に教師付きの学習の場合には、タグ付け作業の負荷から、データのサイズはかなり限られてしま

う。一方、言語現象というのは無限近く存在し、データのスパースネスの問題はどうしても避けられない。この問題を解決するために汎化の手法が試されているが、言語現象には複数の側面があり、単純でシステムティックな方法ではすべての場面に万能な汎化はできないと考えられる。たとえば、1つのシソーラスでは充分ではないのは明らかであるし、学習による汎化という方法も、それ自体データ・スパースネスの問題を内包している。[特異な例] 言語現象には特異な例や例外はいたるところに存在する。一般的な学習手法では、ある低頻度の例が、一般的な例なのか特異な例なのかを区別する方法がなく、そのような特異な例から望ましくないルールを学習してしまうことが考えられる。

[データの誤り] コーパスのタグ付けは人間によって行なわれるため、どうしても誤りが含まれてしまう。

[Ratnaparkhi 1996] は Penn Treebank の品詞の揺れを指摘したが、本論文では、Penn Treebank の品詞付けの正解率は最大 98.7%程度であると推測される実験結果を報告する。自動学習の手法では、これらの誤りを避ける方法はない。

[分野依存性] コーパスはそれぞれの分野があり、分野が違ると言語現象も異なってくる。 [Biber 1993]

[Sekine 1997] ある分野のコーパスから抽出されたルールは、他の分野ではあまり上手く働かないことがあると考えられる。

3 Brill の品詞タガーの問題

Brill の英語品詞タガーを対象に、具体的にどのような問題が露見しているかを観察した。ここでは、Brill のタガーにある WSJ のルールを主な観察対象としている。

[適切な文法的クラスの欠如] 例えば、以下のようなルールがある。「もし、その単語に”1” という文字が含まれていたら、その品詞は NN(名詞) ではなく CD(数字) である。」このようなルールは、0,1,2,3,4,6,7,8 には

存在するが、5と9には存在しない。「数字」という文字のクラスを導入すべきである。同じような例は、Mr.に関するルールは存在するが、Mrs.については存在しないというような肩書だとか、主格代名詞、目的格代名詞、使役動詞、BE動詞、HAVE動詞などのクラスがないということにも見られる。このようなクラスを導入する点については、[Toutanova and Manning 2000]がMEを使うという手法を提案しているが、その導入は直接的ではないため、例えば、ある単語が複数のクラスに属する場合などに混乱が生じる可能性がある。[言語的知識の欠如] Brillのルールでは、“there”に関して2つのルールが存在する。“there”の頻出品詞はEX(~があるというthere are...の用法)であるため、まず“there”という単語にはEXという品詞がふられ、RB(副詞、そこ。例“I went there last Friday”)の用法であろう場合に2つのルールを使って品詞を修正する。2つのルールとは、「もし、次の2つの単語の品詞のいずれかがIN(前置詞など)である場合には、RBにする。」と「もし、次の単語が“.”であれば、RBにする。」というものである。明かにこの2つのルールでは不十分であり、例えば、上記の例文では正しく品詞がタグ付けされない。より良いルールは、「次の単語か前の単語が、BE動詞、HAVE動詞、助動詞のいずれかである場合には、EXとし、それ以外はRBである。」というものであろうと考えられる(もちろん、副詞が間に入る場合なども考慮しなければならないため、実際にはこれより多少複雑である)。同様の現象は、“that”, “all”, “one”などの機能語に見られる。[特異な例とデータの誤り] WSJは実際にはその元の新聞の日付的にも偏っている。WSJの規則には、「“Integrated”という単語があったらNNからNNP(固有名詞)にする」というルールがある。これは、WSJが対象とした頃に“Integrated Resources Inc.”という会社が倒産の危機に陥って、その会社の名前が33回登場していることによる。例えば、もしこのような事件がなかった場合には、次の文は正しく解析される筈である。“Integrated societies were formed in Europe.”このようなルールは、内容語に関して特に多くある。多くの品詞タガーはWSJという閉じた環境で実験されているため、このような問題は露見されないが、コーパスに基づいている方法であるための大きな問題である。(ただし、この論文で報告する実験もWSJで閉じており、このようなルールはそのままにしてある。このようなルールを取り除くことは人間の判断で容易にできるものと考えられる。)

[素性の欠如、複合的な素性] 例えば、大文字が使われ

ているかどうかという点は、そのものが固有名詞かどうかを発見するためだけではなく、他の場面でも有効である。例えば、“America”の次に大文字で始まる単語が来たら、“America”は固有名詞の一部であろうし、そうでなければJJ(形容詞)であると考えられる。有効な素性でBrillで使われていないものが他にも多々存在すると考えられる。

Brillのルールはいくつかのテンプレートの具現化であるが、それは主に1つか2つの条件式を持っているだけである。否定、OR,AND、語彙化されたものなどを含みより複雑な条件式が有効であるような例がある。

4 融合型タガーの作成

Brillの品詞タガーは、品詞タグが付けられたコーパスから自動的に抽出された数百の未知語のためのルールと、各単語の頻出の品詞から、他の品詞への変換を行なうための数百のルールから成る。共に、可読な形式で表現されており、人間が改善するために適している。ルールの変換は、著者1人によって行なわれた。あまりに文法理論に基づいた理想的なルールを作成するよりも、コーパスの例に基き、そこから推測される範囲のルールを作成するにとどめた。

4.1 ルールのフォーマット

ルールのフォーマットは、Brillのものよりも一般化した。ある単語に対して単語の文字列、現在の品詞、辞書に記載されている可能な品詞を組とし、その列を実際の文の一部とマッチングするという方法を取った。また、否定、OR、単語のクラス、などの指定もできるようにした。図1に、thereに関するルールを例として載せる。最初のルールは、「“there”の後に、“be”, “is”,,,などのBE動詞ではないものが来た場合に、1番目の単語(“there”)の品詞をRBにする」という意味である。

4.2 変更点

Brillのタガーに対し、大きく分けて4種類の変更を行なった。

[COMLEX] 未知語を減らす目的で、COMLEX辞書[Grishman et.al 1994]の項目を追加した。すべての項目が頻度1でコーパスに現われたものとして追加した。この追加によって、0.05%の精度向上が見られた。

[文の最初の単語] 文の最初の単語に、小文字化された単語の各品詞の頻度に1/10の重みを掛けあわせたもの

(there EX RB) (!be|is|'s|are|'re|was|were|have|has|had * *) > 1=RB
 (there EX RB) (* RB *) (be|is|'s|are|'re|was|were|have|has|had * *) > 1=EX
 (There|there RB EX) (* MD *) > 1=EX
 (There|there RB EX) (* RB *) (* MD *) > 1=EX
 (Is|Are|Was|Were|Have|Has|Had * *) (there RB EX) > 2=EX

図 1: there に関するルール

を加えた。この変更によって 0.02%の精度向上が見られた。

[相対的頻度の閾値] 品詞の変更の際に新しい品詞の頻度が 20 以下である場合に、それが前の頻度の 1/100 以下であったら、品詞の変更は行なわないようにした。ただし、ルール毎に明示的にこの規則を無効にすることができる。例えば、"'s"の頻出の品詞は POS であり、PRP の頻度の 900 倍あるが、let の後に表われた場合にはほぼ確実に PRP であるというような時には明示的に PRP にできるようにしてある。この変更により、0.09%の精度向上が見られた。

[ルールの追加変更] この変更が、精度向上の大部分を占めている。すべてを説明することは不可能なため、大雑把にトピックを紹介する。

Brill のタガーでは、ルールを適用する順番は重要である。ファイルの上にあるルールから順に適用してゆくため、一度変更されたルールでも再度変更される可能性が存在する。しかしながら、そのままの状態では、同じようなルールがファイルのいたる所に散在し、可読性が良くない。したがって、同じような種類のルール (例えば、VBN と VBD の間の曖昧性) は、その内部のルールの順序は保持したまま、ひとつの場所に集めた。それぞれの種類のルールは、初期の段階で 10 個から 25 個、追加変更をした後では 10 個から 50 個のルールになった。また、このための精度の劣化はほとんど見られなかった。

図 1 に “there” に関するルールを載せた。前述した通り、Brill のルールでは “there” に関するルールは 2 個であったのを図にあるように 5 つ (or を展開すると 29 個) のルールに変更した。この変更では、コーパスの KWIC や、Brill のタガーでの誤りや変更過程での誤りの分析などを参考にした。

このような変更は、“there” だけではなく、Brill の誤りなどから推測できるすべてを対象とした。ルールの数については、表 1 に載せる。最終的なルール数は 562 であるが、その中には OR が 2292 回、not が 161 回使用されているため、Brill のルールフォーマットで数えると 3000 個以上になると推測される。

システム	ルールの数
最終的なシステム	562 (>3000)
Brill (TH=2)	510

表 1: Context Rule の数

4.3 実験結果

実験では、Penn Treebank の構文木もついている 25 のディレクトリにあるデータを以下の 4 つに分けて行なった。4 つというのは、Brill のタガーの未知語学習のための TRAIN(75%) と、Brill のタガーの変換ルールの学習および、ルールの手動での追加変更のための参考データの TUNE (15%) と、定期的に性能のチェックを行なうための CHECK(5%) と、テストのための TEST(5%) である。TEST のデータは、Brill のタガーの性能評価と、その後 2 回の合計 3 回しか使用していない。

表 2 に実験結果を載せる。我々のシステムは Brill の

system	accuracy
システム	96.95%
Brill のタガー	96.51%

表 2: 実験結果

タガーに比較して 0.44%の性能向上、約 13%の間違え率の向上が見られた。我々の結果はこれまで報告されている単一の英語品詞タガーの性能 (96.6% : [Brill 1995] と [Ratnaparkhi 1996]) よりも優れている。

5 食い違いの分析

システム開発中に、著者は Penn Treebank のデータに多くの誤りがあることに気がついた。この誤りが我々のシステムにどのように影響を与えているかを見るためにシステムの結果と Penn Treebank のデータの食い違いのデータを 100 個無作為に取り出し、どちらが正しいかを調査した。この判定は、英語を母国語とし計

算言語学で長く研究を行なっている Grishman 教授が、Penn Treebank の品詞の定義 (“Part-of-Speech Tagging Guidelines for the Penn Treebank Project”) に慎重に従って行なった。表 3 にその結果を示す。「その他」は定義に従っても曖昧である物や、両方が不正解のものを含む。食い違いの約 1/3 が Penn Treebank の間違えで

Penn Treebank が正しい	57
システムが正しい	39
その他	4

表 3: Analysis of Mismatches

あることが分り、システムが得られる最高性能の限界がかなり近いということが分る。この結果に基づく、Penn Treebank のアノテーションの精度は 98.7% でシステムの本来の精度は 98.1% であることが推測される。

ただし、この結果を分析すると、システムが定常的に間違っているパターンが発見される。そのひとつは、VBD と VBN の曖昧性である。たとえば、the authority reported last week の report は VBD であるが、the figure reported last week では VBN であろう。これらは、動詞の格の意味的な情報が文全体の構文情報を利用しないと解決できない。

6 議論

コーパスに基づく方法について分析し、人間の知識の重要性を強調してきた。この 2 つの手法はお互いに補完的であると考えられる。学習手法では、人間が簡単に扱えないくらいの数多くの素性を統一的に扱うことができ、また系統的なスコアや統計的なデータを提供することができる。一方、人間はその場その場の文脈や状況に即した一般化の規則を見付けるのに優れていると考えられる。また、特異な例やデータの誤りも人間でないと見付けにくいであろう。

今回のアプローチが成功した理由のひとつに、ルールの可読性がある。これまで、トランスフォーメーション方式のルールの特徴のひとつは可読性にあるという報告がされていたが、実際に詳しく読んだという報告は著者の知る限りない。この可読性は、SVM や ME などでは得られないので、それらの学習手法では直接的に融合型のシステムを作成するのは不可能であろう。

学習手法と人間のルール作成に関して興味深い報告がある [Ngai and Yarowsky 2000]。人間がコーパスにタグ付けし、それから学習を行なうのと、人間が最初からルールを作成するのではどちらが効率的かという課題に

ついての報告である。それによると、ある程度の精度の NP チャンカーを作成するためには、学習手法を利用した方が有利であったと報告されている。それに対し、本研究の結果に基づく、一般的な学習手法のみでは精度向上に限界があり、それを越えるためには人間の知識が必要になるであろうということである。つまり、最後の数パーセントの精度は、人間の知識を利用した方が学習手法を追及するよりも効率的であろうという結論である。

我々は同様の考えに基づき、コーパスに基いた方法と人間の知識を利用した方法の 2 つを融合させた形で、チャンカー、固有表現タガー、係り受け解析などを作成してゆく予定である。

参考文献

- [Biber 1993] Douglas Biber “Using Register-Diversified Corpora for General Language Study” *In the Journal of Computational Linguistics, Vol.19 No.2* 1993
- [Brill 1995] Eric Brill “Transformation-based Error-driven Learning and Natural Language: A case study in part of speech tagging” *In the Journal of Computational Linguistics, Vol.21 No.4* 1995
- [Grishman et.al 1994] Ralph Grishman, Catherine Macleod and Adam Meyers “COMPLEX Syntax: building a Computational Lexicon” *In the proceedings of COLING* 1994
- [Ngai and Yarowsky 2000] Grace Ngai, David Yarowsky “Rule Writing or Annotation: Cost-efficient Resource Usage for Base noun Phrase Chunking” *In the proceedings of ACL* 2000
- [Ratnaparkhi 1996] Adwait Ratnaparkhi “A Maximum Entropy Model for Part-of-Speech Tagging” *In proceedings of EMNLP* 1996
- [Sekine 1997] Satoshi Sekine “The Domain Dependence of Parsing” *In the proceedings of ANLP* 1997
- [Toutanova and Manning 2000] Kristina Toutanova, Christopher D. Manning “Enriching the Knowledge Source Used in a Maximum Entropy Part-of-Speech Tagger” *In the proceedings of WVLC* 2000