高信頼パーサとプレインテキストコーパスを利用した品詞タグ付け改良規則の自動獲得

平川秀樹[†,☆] 小野顕司[†,☆] 吉村裕美子[†,☆]

大規模コーパスから言語規則や言語知識を獲得するアプローチは，人手による規則開発や知識収集の限界を打ち破るうえでも重要であるが，大規模なタグ付けコーパスを人手を介して準備する手法は，そのコストからいってまだ実際的でない．本論文では，プレインテキストコーパスから，既存の品詞タガーの精度を向上させる品詞判定規則の自動獲得を行う方式を提案する．本方式は，APRAS（Automatic POS Rule Acquisition System）と呼ぶシステムに適用されており，既存の機械翻訳システムの品詞タグ付け規則と構文解析規則という異種の言語規則を組み合わせ利用して，大規模コーパスから品詞判定規則を抽出する．大規模な英文記事コーパスを対象とした実験の結果，獲得された規則は，トレーニングコーパスにない文の 1.7%に対して適用され，そのうちの 78.4%のタグ付け結果に改善が見られた．また，規則対象文のタグ付け処理と構文解析処理にたいして，15.5%の速度向上が見られ，構文解析可能な文の数は，8.0%増加するという結果を得た．

Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpora

Hideki Hirakawa,[†,☆] Kenji Ono[†,☆] and Yumiko Yoshimura[†,☆]

This paper proposes a new unsupervised learning method for obtaining English part-of-speech (POS) disambiguation rules which would improve the accuracy of a POS tagger. This method has been implemented in the experimental system APRAS (Automatic POS Rule Acquisition System), which extracts POS disambiguation rules from plain text corpora by utilizing different types of coded linguistic knowledge, i.e., POS tagging rules and syntactic parsing rules, which are already stored in a fully implemented MT system. In our experiment, the obtained rules were applied to 1.7% of the sentences in a non-training corpus. For this group of sentences, 78.4% of the changes made in tagging results were an improvement. We also saw a 15.5% improvement in tagging and parsing speed and an 8.0% increase of parsable sentences.

1. はじめに

近年，コーパスより様々な言語処理のための知識を自動獲得するための研究がさかんである．特に，英語の品詞タグ付けに関しては，品詞タグの付いたコーパスよりタグ付け規則を統計的手法あるいはそれに文法知識を組み合わせた手法により獲得し，タガー（tagger）を構成する手法が報告されている[1)~6),8),9)]．こうした手法は，十分に大規模なコーパスを用意することで，人手による言語知識（規則）獲得というコストを払わずに，精度の高いシステムを構築できるという利点があるが，その反面，大規模なタグ付き正解コーパスを用意する手間が問題となる．特に，一般的で適用範囲の広い知識の獲得が完了し，より語彙依存的な知識，分野依存的な知識の獲得が問題になるような局面では，データスパースネスの問題が大きくなり，コーパスの準備にかかる手間が大きな問題となる．こうした問題に対する有望な解決策の 1 つは，構築コストが高いタグ付きコーパスの利用に代わって，プレインテキストコーパスを利用する方法を構築することである．

プレインテキストコーパスからの言語知識獲得の研究として，Mikheev は，英語の品詞判定を行う規則の獲得を提案している[6)]．Mikheev の方式は，コーパス中の未登録語に対して，接頭・接尾辞の文字列情報を使って品詞判定を行うというものであり，文脈情報の利用がなくベースとなる情報が限定されている点，また，未登録語という非常に限定された範囲の知識の獲得であるという点が課題である．また，Brill は，教師

† 株式会社東芝研究開発センター
Toshiba R&D Center
☆ 現在，知識メディアラボラトリ
Presently with Knowledge Media Laboratory

付き学習方式（supervised learning）と教師なし学習方式（unsupervised learning），すなわち，品詞タグ付きコーパスならびにプレインテキストコーパスからの品詞判定規則の学習方式を提案し，比較的小規模なタグ付きコーパスから，より精度の良いタガーを構築する手法を示した[8]．この方式では，タグ付きコーパスからの学習（教師付き学習）を行った後，プレインテキストコーパスからの学習（教師なし学習）を適用する．教師なし学習は，可能なタグのセットを絞り込む過程において，より決定されている割合が多い品詞が良い（正解）という仮説を基に行われる．400語の教師付き学習の後では，教師なし学習による精度改善効果は，3.6%（91.8% → 95.4%）あるのに対し，88,220語の教師付き学習では，0.3%（96.5% → 96.8%）と報告されており，効果がかなり減少している．このようにBrillの教師なし学習方式は，学習用タグ付きコーパスの規模が大きくなると，教師なし学習の効果が小さくなり，プレインコーパスの量を増やしても抽出される知識に限界がある点が課題である．教師付き学習を想定せず，プレーンテキストのサイズに応じて知識を獲得できる枠組みが望ましい．

テキストコーパスを知識獲得に利用するためのボトルネックは，すでに述べたように正解データを用意するための人的コストが非常に大きいことである．これを避ける1つの方法は，自動的に正解データを構築する手段，あるいは，それに相当するような手法を確立することである．近年の自然言語処理システムは，かなり精度の高い言語的知識を蓄積してきており，こうしたシステムあるいは言語知識自体が，他のシステムあるいは言語知識に対する教師としての役割を果たすことができれば，人手を介することなく，システムや言語知識の改善が期待できる．本論文における基本的なアイデアは，複数の精度の高い言語処理モジュールやシステムを，大規模コーパスを介して付き合わせることにより，人手を介することなく性能の向上を図るという点にある．本手法は，この基本的な性質上，ゼロからの知識獲得を行うのではなく，基本的な知識については，すでに獲得が済んでいる状況を想定しており，開発が進んだ状況，すなわち，一般的知識ではなく，個別知識や分野知識の獲得において特に有効であると考えている．

本論文では，英日機械翻訳システム[10]に実装されている品詞判定モジュールと構文解析モジュールを利用して，既存の品詞判定モジュールの処理結果を修正する規則を自動獲得するシステムAPRAS（Automatic POS Rule Acquisition System）について述べる．2章では，APRASの基本アイデアとその品詞タグ付けへの適用について述べ，3章でAPRASの概要について説明し，4章，5章で，実験と結果について報告する．

2. 既存言語処理モジュールを利用した規則の自動獲得

2.1 基本アイデア

2つの言語処理モジュールA，Bが存在し，各モジュールに対して同じ課題を与え，それぞれの処理の判断結果を比較できるような状況を考える．この処理判断出力の比較結果は，**図1**のバリエーションを持つことになる．

モジュールA，Bの結果が一致するのは，ともに正解のcase1と，ともに不正解のcase4の2つの場合である．モジュールA，Bの精度が高ければ，処理判断結果が一致した場合には，case1である確率が高い．2つのモジュールの処理結果が異なった場合が興味深い場合である．もし，片方のモジュールがつねに正解を出したり，精度が相対的に優れている場合は，そのモジュールを他のモジュールの教師として学習の枠組みを適用できるが，実際には，開発が進んだ状況では2つのモジュールともに精度が高いと想定され，こうした単純な方法を適用することは有効ではない．不一致は，図のcase2，3，5の場合であるが，各モジュールの精度が高いと想定すれば，case5である確率は低い．case2，3の場合は，いずれかのモジュールが正解を出しており，うまく正解を導けば，全体としての精度の向上が期待できる．個々の不一致事例から，その正否を判断することは，人間の教師以外では困難であるが，多数の事例を集積することにより，統計的に判断することを考える．ここでは，与えられた課題に対する解を出す規則の枠組み（適用条件記述と解よりなる）を用意し，その規則を学習するという方法をとる．もし，規則の記述能力が，課題に対して十分であるとすると，不一致事例に対しても適切な規則が存在することになり，適切な規則の獲得ができれば，精度向上

モジュールB ＼ モジュールA	正解	不正解
正解	一致（case1）	不一致（case2）
不正解	不一致（case3）	一致（case4） 不一致（case5）

図1 言語処理モジュールA，Bの処理判断結果の比較結果
Fig. 1 Cases for comparison of output results of linguistic module A and B.

を図ることができる．全体としての戦略は，次のとおりである．

(S1) 不一致事例から，不一致を相殺できる規則の候補を生成する．

(S2) 大規模コーパスから，規則候補の適用条件を満足する事例をすべてリストアップする．

(S3) リストアップした事例から正解を推定し，正解を生成する（と想定される）規則のみを適切な規則として，以後の課題解決に利用する．

以下では，上記のアイデアを，タガーとパーサという2つの言語処理モジュールを対象に，品詞タグ付けという課題に対して適用する☆1．

2.2 品詞タグ付けへの適用

品詞タグ付けという課題は，以下に示すように，タガーとパーサの2つの言語処理モジュールに共通の課題と位置付けることができ，2.1節のアイデアを適用することが可能である．

タガーは，入力文中の単語に対して優先度を付けた品詞を割り当て，それから優先度付けられた系列を生成する機能を有している．ここで，最も優先度の高い品詞列，すなわち，最初に生成される品詞列を第1品詞列と呼ぶこととする．一方，パーサは，1つの品詞列に対して，それが受理可能であるか否かを判定する機能を有している．この2つを組み合わせることにより，入力文に対する品詞列をタガーにより優先度順に生成し，それぞれの品詞列をパーサにより解析するという処理を行うことができる☆2．ここで，パーサが受理可能な文と品詞列を，それぞれ受理可能文（parsable sentence）および受理可能品詞列（parsable POS sequence）と呼び，受理可能な品詞列のうちでタガーの付与する優先度が最大の品詞列を最尤品詞列（most preferable parsable POS sequence）と呼ぶこととする．タガーは，仮説生成器，パーサは，仮説検証器として働いているが，特定の単語に対する品詞タグ付けという課題という観点からみると，第1品詞列中の品詞タグがタガーの処理判断結果，最尤品詞列中の品詞タグがパーサの処理判断結果と考えられる．この観点から，入力文は，次の3種類に分類することができる．

(1) バランスした文：第1品詞列が最尤品詞列である文（バランス文）

(2) 対立した文：最尤品詞列が第N品詞列 ($N \geq 2$) である文（対立文）

(3) 受理不可能な文：すべての品詞列が受理不可能である文（受理不可能文）

(1)は，2.1節でいう2つのモジュールの判断が一致した場合，(2)は，2つのモジュールの判断が異なった場合，(3)は，判断の比較ができなかった場合に相当する．

ここで，以降の議論で使用するいくつかの用語について定義を与えておく．

第1品詞（initially tagged POS）：文中の単語が持つ2つ以上の可能な品詞（多品詞）のうち第1品詞列の対応する位置に現れる品詞（タガーにより最も優先度が高いと判定された品詞）

受理品詞（parsable POS）：文中の単語が持つ2つ以上の可能な品詞のうち，最尤品詞列の対応する位置に現れる品詞

焦点単語（focus word）：その単語の第1品詞と受理品詞が異なる単語

品詞コンテキスト：焦点単語の前後の単語の品詞☆3

対立文とその第1品詞列，最尤品詞列，焦点単語およびその品詞コンテキストは，すべて自動的に得ることが可能である．対立文においては，タガーとパーサの解釈の違いが焦点単語の品詞に現れており，その違いを品詞コンテキストの条件で解消する品詞調整規則（POS Adjusting Rule，PA規則と略記する）を次の形で抽出する．

PA規則：W (IPOS) → W (PPOS)：C
W：焦点単語，IPOS：第1品詞，
PPOS：受理品詞，C：品詞コンテキスト

この規則は，「品詞コンテキストCにおいて現れる単語Wにおいて，第1品詞IPOSより受理品詞PPOSを優先する」という意味を持つ．**図2**は，PA規則の生成例を示した図である．焦点単語は，“ranked”で，その第1品詞は動詞（vti），その受理品詞は過去分詞（pp）である☆4．品詞コンテキストは，焦点単語の前後2つの品詞をとっており，この例では，「(name)-(cmm)-$-(cdig)-‘in’」である．「$」は，焦点単語自身の位置を示している．抽出されるPA規則は，抽出に利用されるタガーならびにパーサに依存しないで適用可能な規則ではあるが，規則の抽出が対立文から行われているという点では，タガーの規則などすでにシス

☆1 同じタスクを行う複数のシステム（たとえば，規則ベースのタガーと統計ベースのタガー）を用いるsystem-combinationの方式[11]の出力を対象に学習するなどの方式も考えられる．

☆2 品詞の組合せ爆発の問題を回避するため，実際には，生成する品詞列候補の数は所定の数に限定する．

☆3 ここでは第1品詞列と最尤品詞列における焦点単語のコンテキストが等しい場合のみを対象とする．

☆4 本論文システムでは，動詞の過去形と過去分詞形は，異なった品詞として処理される．

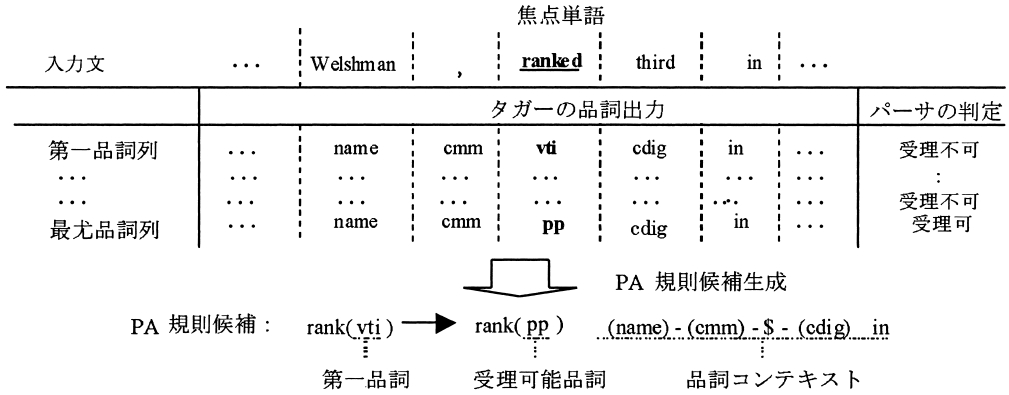


図 2 PA 規則候補生成の方式説明
Fig. 2 PA rule candidate generation.

テムが有している知識を獲得することはなく，抽出に使用したタガーとパーサに対して最適化されているといえる．

以上，2.1 節の (S1) の不一致事例からの規則の生成について述べた．次章では，2 つのモジュールに焦点を当てて，品詞タグ付け規則の自動獲得手法について，(S2)，(S3) を含めてより具体的に説明する．

3. 品詞タグ付け規則自動獲得システム

3.1 APRAS の概要

図 3 は，APRAS を機械翻訳システムに適用した例を示している．APRAS は，規則抽出フェーズと規則適用フェーズの 2 つのフェーズよりなっている．この 2 つのフェーズにおいて同じタガー☆とパーサが使用される．

規則抽出フェーズにおいては，タガーは，既存の品詞決定規則（タガー規則）に基づいて，トレーニングコーパス中の文を解析し，その文に対して可能な品詞系列を生成し，それぞれに対してパーサは，その構文解析可能性をチェックする．対立文が出現すると，PA 規則候補生成モジュールは，PA 規則の候補となる規則を出力する．また，バランス文と対立文に対しては，受理可能品詞列を別途出力しておく．すべてのトレーニングコーパスに対して，この抽出処理を行った後で，規則フィルタリングモジュールが，得られた PA 規則候補の正当性に対する統計的な重みを計算し，信頼度の低い規則を排除して，PA 規則を得る．

規則適用フェーズにおいては，既存の品詞決定規則と，抽出された PA 規則の 2 つを用いてタグ付けが行われる．そして，タガーの出力に対してパーサにより文が解析され，構文解析に成功した場合は，さらに後のモジュールに結果が渡され翻訳される．ここで，PA 規則は，既存の品詞決定規則が適用された後で，その結果に対して優先度を調整するように適用され，基本的には，タガーが構文解析に失敗するであろう無意味な品詞系列を生成することを避けるように働く．このため，タガーにより生成される品詞系列のランキングが改善，すなわち正しい品詞列の候補が優先されて解釈され，それにより，パーサが構文解析に成功するようになったり，より早い段階で正しい品詞系列による解析結果が得られるようになる．

3.2 規則候補生成モジュール

すでに述べたように，PA 規則の候補は，対立文より生成される．バランス文と受理不可能文は，PA 規則の候補を生成しない．ただし，バランス文の単語とその品詞は記録され，後の規則フィルタリングモジュールにより利用される．コーパス中の文を解析して，対立文であった場合，図 2 に示したように，規則生成モジュールは，第 1 品詞列と最尤品詞列を比較して焦点単語を検出し，それぞれの焦点単語に対して PA 規則候補を生成する．品詞コンテキストは，基本的には，焦点単語の前後 2 つずつの品詞より構成されているが，前置詞など特定の単語の場合は，例の「'in'」のように表層単語そのものを使用する．また，文頭，文末など，単語が存在していない部分を含むコンテキストの場合は，空の単語が存在すると見なしている．

この規則生成モジュールにおいては，コンテキストサイズと抽象度レベルという 2 つの主要なファクタを考慮する必要がある．焦点単語に対する品詞コンテキストのサイズを大きくすると，PA 規則の適用条件が

☆ 本システムのタガーは，単語の前後の文脈（単語ラティス）を参照して優先度付けを行う規則ベースのもので，すべての単語に適用される一般規則と，特定の語彙に添付・適用される語彙規則の 2 種類の規則を有している．

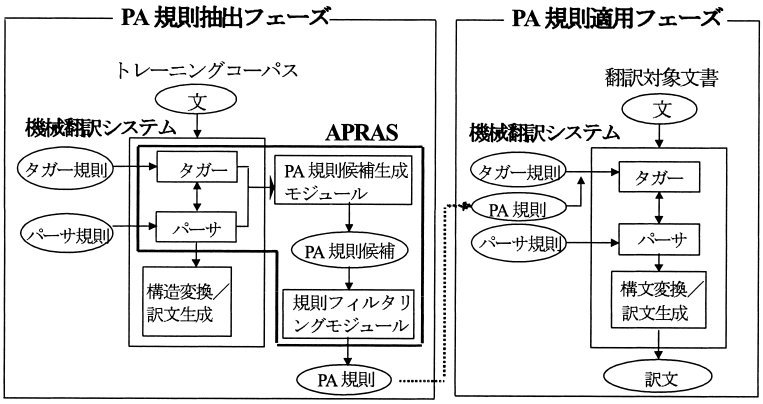


図 3 機械翻訳システムへの APRAS の適用
Fig. 3 Application of APRAS to an MT System.

厳しくなり，規則の精度向上が期待できる．しかし，トレーニングコーパスから得られる PA 規則候補の頻度が減少するため，統計的な規則フィルタリングが難しくなり，統計的な信頼性を確保するためには，より大規模なコーパスを必要とするようになる．また，この種の単語の窓（word window）のサイズは，必ずしも大きい方がよいということもないため，ここでは，前後 2 つの単語を見ることとした．本論文では，PA 規則は，この汎用の 1 パタンのみに統一しているが，Brill は，4〜6 パタンの変換規則（transformation）テンプレートを人手でアプリオリに与えている．変換規則の起動環境（triggering environment：本論文での文脈条件に相当）には，非常に粗いものも含まれている．このため，獲得される規則は，その規則単体での正しさという観点では必ずしも保障できるものではないと思われる．Brill の考え方では，一連の変換規則を順次かけてゆくことで，ある変換規則でいったん誤った品詞がアサインされても別の変換規則で，再度別の品詞がアサインされるというプロセスが発生する．このため，規則適用の順番が非常に重要であり，それを Transformation-Based Error-Driven Learning という正解コーパスに対する誤り最少化というクライテリアで決定する手法をとっている．本論文では，実際には基本的な規則はすでに獲得されているという状況を想定しており，Brill のように複数の規則が適用されるということがほとんどなく，規則の適用順序の制御を行う必要性はない．

2 つめの抽象度レベルとは，品詞コンテキストにおける品詞の抽象度をどの程度にするかという課題であり，抽象度の高い（粗い）タグを用いると規則の適用条件が緩み適用範囲が広がるが，正確性が低下する恐れがある．品詞のクラスを階層化するなどして，種々のレベルでの抽象化を行うなどの方法も考えられるが，ここでは，文献 7) と同様に，タグセットとしては，いわゆる品詞レベルのタグ，細分化された品詞タグ（subdivision level），および一部の語については，表層レベルのものを使用している．特に，機能語である前置詞，be 動詞，have などは，単語そのものが焦点単語の品詞選択に重要な影響を与えると考えられるため，これらは単語表層語のレベルで品詞コンテキスト条件に記述した．結果として，名詞 16 種類，動詞 17 種類，前置詞ならびに群前置詞 410 種類，形容詞・副詞 70 種類など，全体で 513 種類の品詞タグを利用している．また，PA 規則自体は，上記の例のように特定の単語（rank）に対する語彙規則としている．単語の表層でなく品詞を採用することにより，より一般的な規則とすることも可能であるが，汎用規則の抽出が目的ではないため，PA 規則は，語彙規則としている．

3.3 規則フィルタリングモジュール

本節では，PA 規則候補生成モジュールにより生成された規則候補のうち，不適格な規則を取り除く規則フィルタリングモジュールについて述べる．

1 つの対立文から PA 規則を抽出することは，1 つの事例に合致する規則を抽出することであり，この個別の事例のみからでは，その規則（知識）の正否の判定はできない．そこで，2.1 節の (S2)，(S3) で述べたように，コーパスから得られる多くの事例を観測することにより，正しい規則の抽出を考える．PA 規則の文脈条件が焦点単語の正解の品詞を判定するに十分な弁別力を有するとすれば，文脈条件と焦点単語を同じくする事例を集めれば，焦点単語の品詞は 1 つになると考えられる．タガーの規則に不備があり，この品詞の判定に誤りがあると，パースに失敗し，対立文や

$POS_{W,C}X$	A
$POS_{W,C}X \rightarrow POS_{W,C}Y$	B

図 4 構文解析過程におけるコンテキスト C 中の単語 W の品詞推移
Fig. 4 Transition of POS of W in parsing process for context C.

受理不可能文になる可能性が高い．パーサは，文全体の品詞列を見て判断を行うため，特定の文脈条件に対する種々多数の事例を集めれば，焦点単語の正しい品詞に対して受理可能とし，そうでない品詞に対して受理不可能とする可能性が高くなる．この仮説を利用することにより，規則の良し悪しの判定が可能となる．いくつかの基準を考えることができるが，APRAS では，「変更率（adjustment ratio）」と呼ぶ指標を用いている．

図 4 は，品詞コンテキスト C: P1-P2-$-P3-P4 中の単語 W に対する構文解析の様子を示したものである．図の A の場合は，タガーが最初に W に対して品詞 X を割り当て，その構文解析に成功したというバランス文のケースである．B の場合は，品詞 X では構文解析に失敗し，その後品詞 Y を割り当てたときに構文解析に成功したという対立文の場合である☆．N_a と N_b をそれぞれ，A の場合と B の場合の文の数とすると，品詞コンテキスト C における単語 W の品詞 X から品詞 Y に対する変更率は，次の式で表される．

$$\text{変更率}_{W,C}(X \rightarrow Y) = \frac{N_b}{(N_a + N_b)}$$

パーサが，特定の品詞コンテキスト中の焦点単語に対する正しい品詞を持つ品詞列に対してはその大半を受理し，誤った品詞を持つ品詞列に対しては大半を受理不可能と判断するだけの精度を有していると仮定すれば☆☆，次のような対応関係が成立すると想定される．

変更率が高い ⇔ 単語 W の品詞は Y が正しい
変更率が低い ⇔ 単語 W の品詞は X が正しい

このように，高精度パーサの判定結果を統計的に反映した変更率は，PA 規則候補の良し悪しを判定するための指標として利用することができると考えられる．この仮説の正否判定や適切な変更率のスレショルド値は，PA 規則候補を実際に実験・調査することにより実証的に検証する．変更率の利用において重要な点は，変更率は，使用するパーサに対して，ある意味で最適化されているということである．この意味は，変更率により選ばれた PA 規則が品詞の判定規則として誤っていた場合でも，その PA 規則の適用がパーサの出力に対して直接悪影響を与えることは非常に少なく，処理時間を短縮できると予想できることである．これは，その PA 規則の適用により構文解析にトライされなくなる品詞系列は，変更率の定義から，もともと構文解析に成功する可能性が非常に低く，また，その場合，バイパスされた構文解析時間が節約されるという現象がおこるためである．誤った PA 規則の抽出は，通常，もともとのパーサ規則，すなわち，構文解析知識に欠落や欠陥があった場合に起こり，その PA 規則の対象となる文に対しては，もともとのパーサ規則で構文解析しても誤った結果となることが多い．

また，生成される PA 規則は，タガー規則やパーサ規則の変更により変化するため，原理的には，それらが変更された場合に再度行うのが適切であるが，PA 規則の生成は，すべて自動に行えること，ならびに，誤った規則の適用も悪影響を与えないことから，タガーやパーサ規則の開発時にそのつど行うのではなく，新バージョンの固定などのタイミングで行えばよい．また，PA 規則はタガー規則の語彙規則として取り込むことが可能であり，人手により適切な規則のみを取捨選択し，着実な改善を行うことができる．実際，PA 規則の抽出は，タガー規則やパーサ規則の問題抽出にも役立ち，規則のチェック・改良などの開発作業にも有効である．

4. 規則獲得実験ならびに評価

4.1 規則抽出実験

3 章で述べた手法を評価実験するため，トレーニングコーパスとして英語ニュース記事（6,684,848 文，530 MB）☆☆☆に適用したところ，300,438 個の異なった PA 規則候補が得られた．頻度の低い規則は，信頼度の高い変更率を得ることはできないと考え，6 以下の頻度を持つ規則候補は対象外としたところ，17,731 個の PA 規則候補が得られた．

次に，3.3 節で述べた，変更率による規則抽出法の正当性を確認するために，抽出された PA 規則候補のうち，頻度 10，20，30 の規則候補を，それらが得ら

☆ ここでは，品詞 X と品詞 Y の 2 つの可能性を持つ場合についてのみ言及するが，単語 W が 3 つ以上の品詞を持つ場合に対しても適用可能である．

☆☆ 使用したパーサにより受理される品詞系列の品詞精度は，実験対象より抽出した単語約 6,000 語に対して 99%以上であった[2]．

☆☆☆ Financial Times (1992–1994, approx. 210,000 documents, 530 MB) in NIST Standard Reference Data TREC Document Database: Disk4 (Special No.22), National Institute of Standards and Technology, U.S. Department of Commerce (http://www.nist.gov/srd).

変更率(%)	総数	正	(%)	誤	(%)	不定
0-9	20	0	(0)	19	(95)	1
10-19	25	3	(12)	17	(68)	5
20-29	24	4	(17)	10	(42)	10
30-39	16	8	(50)	2	(13)	6
40-49	15	10	(67)	1	(7)	4
50-59	15	7	(47)	4	(27)	4
60-69	15	15	(100)	0	(0)	0
70-79	17	15	(88)	1	(6)	1
80-89	16	15	(94)	1	(6)	0
90-99	18	14	(78)	3	(17)	1
100	29	23	(79)	2	(7)	4
総数	210	114		60		36

図 5 変更率と抽出された PA 規則候補の正当性
Fig. 5 Adjustment ratios and the validity of extracted rules.

れた元の文を参照しながら，人手により次の 3 種類に分類した．

正：どの文に適用しても正しい．抽出した規則そのものが正しい規則である．

誤：どの文に適用しても誤り．パーサ規則の不備などに起因して抽出された誤った規則．

不定：正でも誤でもなく，与えられた品詞コンテキストでは，焦点単語の品詞が一意に正誤が決定できないもの，あるいは，どちらの品詞も誤りであるもの．

次は,「不定」の例である．

trading (PS) → trading (N):
(N)-'of'-$-(DET)-(N)
(PS：現在分詞，N：名詞，DET：冠詞)

単語 "trading" は，".. index features represent a more convenient and liquid way of trading an index basket than ...," のような文においては，現在分詞であるが，"By the close of trading the deal was quoted at 99.82 bid." のような文においては名詞である．

図 5 は，分類の結果を示している．図から明白なように，30%以下の変更率を持つ規則候補については，不正解の規則が正解の規則を上回っており，30%より上の規則では，その逆となっている．特に変更率が 60%以上の規則においては，大半の規則が正解規則となっている．この結果から，変更率を基にした PA 規則選択のフレームワークは正当性があるといえ，60%をスレショルドとして以降の実験を行うこととした．上記の実験をベースにすると，変更率 60%をスレショルドとして PA 規則候補のフィルタリングを行うと，PA 規則の 86%が正しい規則，7%が誤った規則，7%がどちらともいえない規則になると予想される．また，変更率が 30%以下の規則については，タガーにより初期にアサインされる品詞を変更しないように制御する変更抑制型の PA 規則として適用することが考えられるが，これに関しては，次の課題とし，今回の実験では省略した．

PA 規則例

group(V) → group(N) : (DET)-(N)-$-','-(VP)

DSM, the chemicals **group**, declined Fl 2.50 or 2.3 per cent to Fl 103.60.

Ciga, the hotels **group**, fell L47 to L1,902 after new s that it remained in the red in 1991 while turnover was little changed at L49 0bn.

reports(N) → reports(V) : ','-(NAME)-$-'from'-(NAME)

A SURVEY released yesterday shows that unemployment in the former Soviet Union this year may reach the level of the 1930s US depression, AP **reports** from Washington.

A Japanese opposition MP resigned yesterday to take responsibility for links to a scandal, Reuter **reports** from Tokyo.

related(VP) → related(PP) : (N)-(CC)-$-(N)-(PNC)

The sanctions, imposed after years of US complaints, are directed towards Indian exports of pharmaceuticals, chemicals and **related** products.

Job Training and **Related** Services.

opened(PP) → opened(VP) : (N)-(N)-$-'at'-(DIG)

The Liffe bund futures contract **opened** at 88.55 and reached a high of 88.64 before closing at 88.49.

The Liffe bund futures contract **opened** at 88.46 and closed at 88.52 on a volume of about 20,000 contracts.

further(V) → further(ADV) : (N)-(AUX)-$-(V)-(DET)

The US administration's recent actions against Canadian exports could **further** damage the popularity of the Mulroney government.

A provision at the year end for middle management redundancies could **further** spoil the picture.

long(ADJ) → long(ADV) : 'not'-'have'-$-'to'-(V)

Park may not have **long** to prove himself.

They did not have **long** to wait, however.

品詞記号の説明

ADJ: 形容詞, ADV: 副詞, CC: 等位接続詞,
DET: 冠詞, DIG: 数詞, N: 名詞, NAME: 固有名詞,
PNC: 句点記号（除：コンマ）PP: 過去分詞,
V: 動詞（過去形以外）, VP: 動詞過去形.

図 6 抽出された PA 規則の例
Fig. 6 Examples of extracted PA rules.

以上の条件により，17,731 個の PA 規則候補から 4,496 個の PA 規則を得た．**図 6** に抽出された PA 規則の例を，その抽出元となった例文とともに示す．

4.2 規則適用実験

PA 規則を適用することにより，次を期待することができる．

(1) 構文解析初期の段階で受理可能な品詞系列をとらえることによる処理時間の短縮

(2) タガーの精度ならびにパーサの精度の向上

PA 規則適用の効果を測定するため，学習コーパスとして未使用の英文記事コーパス（146,229 文 22.6 万語）に対して，前章の実験により抽出された 3,921

	PA 規則非適用	PA 規則適用
タグ付け時間(sec.)	79.40	88.49 (+9.09, +11.5%)
構文解析時間(sec.)	252.17	191.73 (−60.44, −24.0%)
総合処理時間(sec.)	331.57	280.22 (−51.35, −15.5%)
構文解析可能な文	64.0%	72.0%

図 7 処理時間ならびに構文解析可能な文の割合
Fig. 7 Processing time and parsable sentence ratio.

個の PA 規則☆を既存のタガーに組み入れて品詞タグ付け処理を行った結果，対象文全体の 1.7%にあたる 2,421 文，全単語の 0.11%にあたる 2,476 単語において PA 規則の条件が満足され，規則の適用が行われた．これらの文に対して，PA 規則を適用する場合と適用しない場合のそれぞれに対してタガーで品詞付け実験を行った．この実験では，WorkStation SUN Ultra U1E/200 を使用し，処理時間の差の測定，構文解析に成功した文の数の測定を行った．この結果は，**図 7** に示されている．

タグ付けに必要な時間は，11.5%増加したが，構文解析時間は，24%，解析全体でも時間は，15.5%それぞれ減少し，構文解析に成功した文の数は 8.0%増加した．別途，テスト用コーパスから PA 規則の適用の有無とは無関係に選んだ 400 文に対し，PA 規則を利用する場合と利用しない場合の処理速度の差を測定したところ，全体処理時間 0.36%増との数値を得た．規則適用時間の増加分と構文解析時間の低減分が相殺されて，全体処理時間の増加はほとんどなかった．また，PA 規則を語彙タガー規則に組み込んでしまうことで，規則適用のオーバヘッドはほとんどなくなると考えられる．

PA 規則の適用によりタグ付けの結果が異なる単語の事例のうち，524 個について規則の効果をサンプリング調査した結果，次のようなデータを得た．

- 改善：411 (78.4%)
- 悪化：84 (16.0%)
- 改善でも悪化でもなし：29 (5.5%)

(A) 改善例

語彙レベルの知識獲得の有効性を示す改善事例を以下にあげる．

(A-1) Following German unification, the Soviet Union has agreed to withdraw 380,000 soldiers and their 220,000 family members from eastern Germany by the end of 1994.

適用規則：following (ADJ) → following (PS):
--$-(ADJ)-(N)
(「**」は空単語を示す.「following」は文頭位置にある．PS：現在分詞)

(A-2) The deficit stemmed from a trading loss of Pounds 1.88m at four businesses which the group has discontinued.

適用規則：stemmed (PP) → stemmed (VP):
(DET)-(N)-$-'from'-(DET)

(A-1) の学習は，文頭位置で分詞構文をとりやすい「follow」の特徴が統計的に抽出された結果ととらえることができる．(A-2) は動詞の過去形と過去分詞形が同形の動詞が名詞と前置詞の間に挟まれているパターンで，動詞に他動詞用法がある場合はどちらが正しいかを一般的に定義するのが難しい代表的なケースの 1 つである．しかし「stem」という語に着目することで，「stem(自動詞)+from …」という用法を統計的に浮かび上がらせることができ，上記規則を獲得している．

(B) 悪化例

悪化した 84 個の場合において，43 個は，適切なパーサ規則（構文解析知識）が存在しないために誤った解析結果から抽出された PA 規則により導かれていた．(B-1) にその一例をあげる．

(B-1) The NUT is the only teachers' union opposed to the setting up of a pay review body for teachers.

適用規則：setting (PS) → setting (N):
'to'-(DET)-$-'up'-'of'

「the 現在分詞 up of ...」が正しいが,「the 現在分詞 up」の部分を名詞句として解析する規則が不足していたため,「setting」は名詞として解析を通過し，上記規則が獲得される結果となった．この際,「up」は「setting」にかかるものとして解析されていない．繰返しになるが，この種のタグ付けの悪化は，パーサの解析まで考慮すると，最終的には同じ結果となるため，たとえば翻訳プロセス全体の結果としては，差とならないことが多い．

悪化の残り 32 個は，特有のヘッダ部分の構造の認識の不備や文のセグメンテーション誤りが起因して発生していた．(B-2)，(B-3) にそれぞれの例をあげる．

(B-2) London Page 8 Photograph German finance minister Theo Waigel (left) talks with Bank of France governor Jacques de Larosiere at the Washington meeting (Omitted).

適用規則：photograph (V) → photograph (N):

☆ 本実験では，4,494 個の PA 規則のうち，573 個が削除されている．これらは，複合語（例. 'that is' (adverb)）と非複合語の系列（例. 'that (pronoun)+is (verb)'）のように異なったコンテキストを有するものである．

(NU)-(DIG)-$-(N)-(N)

（NU：単位名詞．副次的影響：Germanが形容詞から名詞へ変化）

(B-2) は本来「Photograph」の後で大きく切れるべきもので,「German」以下から写真の内容説明の文が始まっている．規則抽出実験で用いたコーパスにはこのような写真説明の文が多く含まれている．このような文がつながって 1 文として処理され，その結果,「Photograph」に対してタガー規則は，動詞を第 1 品詞としており，上記の規則を獲得した．これ自体は正しい規則獲得であるが，本来 2 文として解析すべきものを 1 文としているため,「Photograph」に名詞が優先されたことの副次的影響として,「German」に対して名詞が優先されるという副作用が出た．

(B-3) FT 01 MAY 91 / Scots population rise

適用規則：Scots (ADJ) → Scots (N):

(DIG)-'/'-$-(N)-(V)

(B-3) は「/」で文の構成が大きく区切れるが，区切らずに 1 文として処理された．ここで使用したパーザ規則では「/」の処理が限定されており，数字の後の「/」の直後の語を形容詞として解析する規則が用意されていなかったため，名詞としての解析結果を導き，これが蓄積されて上記規則の獲得に至った．(B-2), (B-3) にあげたような誤りの大半は，コーパス処理プロセス（文抽出）の改良や解析規則の改良により回避できると考えている．

上記の実験の規則の精度を基に計算すると，PA 規則の適用におけるタグ付けに対する改善率としては，62.4%（78.4% − 16.0%）を得ることができる．PA 規則は，コーパス中の単語の 0.11%に対して適用されることから，全体では，0.07%の品詞付け精度の改善を期待することができる．この値の妥当性を見るため，テストコーパス中の 5,630 語に対して，PA 規則を適用した場合としない場合の結果を比較することによりタガーの精度を比較した．この実験では，98.60%のタグ付け精度が，98.65%に向上し，0.05%の改善が確認でき，ほぼ上記の結果と一致した．PA 規則は，語彙依存の規則であるため，規則適用の条件を満たす文の率はかなり低いといえるが，PA 規則の数が増加するにつれて，PA 規則が適用される文の数も増えてくると予想される．トレーニングコーパスのサイズを増加させることにより，PA 規則の数を増加させることが可能である．実際，出現頻度のスレショルドを満たさないために除外された多くの PA 規則候補の中にも非常に多くの正しい PA 規則を確認することができる．また，コーパスサイズの拡張は，品詞コンテキストの範囲をより広くとる，より多くの機能語を表層レベルのタグとするなど，品詞コンテキスト，すなわち，規則の適用条件をより詳細なものにするなどの拡張を行うことを可能とする．詳細な分類を十分に有効にするためには，それぞれの具体的な例における最大の差を反映するようなレベルに特定の規則を抽象化する必要がある．

5. おわりに

本論文では，プレインテキストコーパスから自動的に言語知識（規則）を獲得する新しい手法について述べ，そのフィージビリティを示すために実証実験を行った．実験では，タガーとパーサを含み，開発が進んだ自然言語処理システムである機械翻訳システムを利用し，高い精度で正しい規則が抽出できることを示した．また，本方式は，抽出された規則が誤っている場合にも，全体としての結果に対しては，悪影響を与えることが少ないという意味で，ロバストであるといえる．本論文では，語彙依存の知識を獲得する最初のステップとして，プレインテキストからの教師なし学習による規則獲得を試みたが，今後は，本方式におけるトレーニングコーパスサイズの拡大と PA 規則の抑制的な適用の実験を行うとともに，開発蓄積された知識間の相互作用と大規模コーパスを利用して，知識の改善を行うというフレームワークをタガーとパーサという品詞の判定の枠組み以外にも適用することを検討していきたい．

参考文献

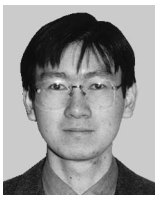
1) Church, K.: A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, *Proc. 2nd Conference on Applied Natural Language Processing*, Austin, Texas, pp.126–143 (1988).
2) Kupiec, J.: Robust Part-of-Speech Tagging Using a Hidden Markov Model, *Computer Speech & Language*, Vol.6, No.3, pp.225–242 (1992).
3) Brill, E.: A Simple Rule-Based Part of Speech Tagger, *Proc. 3rd Conference on Applied Natural Language Processing*, pp.152–155 (1992).
4) Voutilainen, A., Heikkila, J. and Anttila, A.: CONSTRAINT GRAMMAR OF ENGLISH — A Performance-Oriented Introduction, *Publications of the Department of General Linguistics*, University of Helsinki, No.21, (1992).
5) Dermatas, E. and Kokkinakis, G.: Automatic Stochastic Tagging of Natural Language

Texts, *Computational Linguistics*, Vol.21, No.4 (1995).
6) Mikheev, A.: Unsupervised Learning of Word-Category Guessing Rules, *Proc. 34th Annual Meeting of the Association of Computational Linguistics*, Santa Cruz, California (1996).
7) Haruno, M. and Matsumoto, Y.: Mistake-Driven Mixture of Hierarchical Tag Context Trees, *Proc. 35th Annual Meeting of the Association of Computational Linguistics*, Madrid, Spain (1997).
8) Brill, E.: Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging, *3rd Workshop on Very Large Corpora* (1995).
9) Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, *Computational Linguistics*, Vol.21, No.4 (1995).
10) Hirakawa, H., Nogami, H. and Amano, S.: EJ/JE Machine Translation System ASTRANSAC-Extensions toward Personalization, *Proc. MT SUMMIT-III*, Washington, D.C., pp.73–80 (1991).
11) van Halteren, H., Zavrel, J. and Daelemans, W.: Improving Data Driven Wordclass Tagging by System Combination, *Proc. 17th COLING* (1998).
12) 吉村裕美子：構文解析情報を利用した英語品詞列選定，情報処理学会第 50 回全国大会，2R-7, pp.65–66 (1995).

(平成 12 年 6 月 21 日受付)
(平成 13 年 9 月 12 日採録)

平川　秀樹（正会員）

昭和 31 年生．昭和 55 年京都大学大学院工学研究科電気工学専攻修士課程修了．同年（株）東京芝浦電機入社．機械翻訳システムの研究開発に従事．昭和 57 年〜59 年新世代コンピュータ開発機構（ICOT）研究員，平成 6 年〜7 年 MIT Media Lab. 派遣研究員，平成 8 年より（株）東芝研究開発センター知識メディアラボラトリ所属，自然言語処理，知識処理，ヒューマンインタフェースに関する研究に従事．電子情報通信学会，人工知能学会，言語処理学会，ACL，ヒューマンインタフェース学会各会員．



小野　顕司（正会員）

昭和 38 年生．昭和 62 年東京大学工学部計数工学科卒業．同年（株）東芝入社．以降自然言語処理システムの研究開発に従事．現在（株）東芝研究開発センター知識メディアラボラトリ所属．言語処理学会会員．

吉村裕美子（正会員）

昭和 37 年生．昭和 60 年京都大学文学部文学科言語学専攻卒業．同年，（株）東芝入社．現在，（株）東芝研究開発センター知識メディアラボラトリに所属し，機械翻訳を主とする自然言語処理システムの研究開発に従事．