自然言語処理 126－8
(1998. 7. 23)

書き換え規則と文脈情報を用いた形態素解析後処理

久光 徹，丹羽 芳樹
日立製作所基礎研究所
埼玉県比企郡鳩山町赤沼2520
{hisamitu, yniwa}@harl.hitachi.co.jp

あらまし

本報では，日本語形態素解析における誤りを，書き換え規則により修正する後処理方法について述べる．書き換え規則は，誤り主導型教師付き学習により生成する．すなわち，誤りを含む解析結果と正解の差分から，語彙的なルールから，字種や品詞だけを用いるルールまでを含む，さまざまな抽象度のルールを生成し，正解データに適用してそれらの信頼度を評価する．後処理は複数の文集合ごとに行い，上記ルールと，集合内の各文の解析結果の相互参照により，出力結果を修正する．実験では，後処理により解析精度を3%程度向上させることが確認できた．

キーワード: 形態素解析，後処理，誤り駆動教師付き学習

Post-processing of Japanese Morphological Analysis Using Transformation Rules and Contextual Information

Toru Hisamitsu and Yoshiki Niwa
Advanced Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-0395, JAPAN
{hisamitu, yniwa}@harl.hitachi.co.jp

Abstract

A method is proposed for the post-processing of Japanese morphological analysis using transformation rules and contextual information. The method corrects both segmentation errors and part-of-speech tagging errors.

The transformation rules are acquired automatically by error-driven supervised learning. The rules consist of various types, such as lexicalized rules and schematic rules. Each rule is assigned a value for reliability. The rules are not specifically tailored for detecting unregistered words, but can correct errors caused by unregistered words.

In addition, we propose the use of contextual information obtained from the result of analysis of neighboring sentences. The information reinforces unregistered word detection and disambiguation.

The post-processing improved the precision of the analysis of an open corpus by 3%.

keywords: morphological analysis, post-processing, error-driven supervised learning

1 はじめに

1.1 背景

形態素解析器の高速化に伴い，新聞や特許等の情報検索のインデキシングに形態素解析がしばしば利用されるようになっている．昨今，形態素解析器の速度向上は著しいが，精度に関してはいまだに改善の余地がある．

形態素解析の精度低化の要因は未登録語の存在と文法や尤度基準の不備である．未登録語に起因する誤り対策としては，オフラインでの未登録語獲得と辞書登録が基本的な方法であるが，これのみでは，未知文書に現れる未登録語や，文法や尤度基準の不備に起因する誤りには無論対処できない．なんらかの誤り低減技術の研究が不可欠である．

1.2 従来の手法

形態素解析の手法は，人手により先験的に与えたルールに基づく手法と，確率に基づく手法に大別される．ルールに基づく形態素解析は古くから研究されており，公開プログラム（代表例としてはJUMAN[1]，茶筌[2]）の数や，コンパクトさ等の理由から，ルールに基づく形態素解析システムが数多く利用されている．

ルールに基づく形態素解析システムの利点は，人間の直感を利用するため，小規模な実験データを参照するだけで比較的高精度なシステムを構成できることと，解選択の根拠となるルール集合がコンパクトかつ人間に可読であること，コンパクトさのため形態素解析器の速度の点で有利であること，等である．

ルールに基づく形態素解析システムの一層の精度向上を目指しては，後処理的な考え方，すなわち，誤り例の人手による分析から，修正用のテンプレートやルールを生成する研究が行われており[3][4][5]，特に未登録語による誤り修正に力点を置くものが多い．

このとき問題になると考えられるのは，様々な誤りに対処するためにルールが増加してゆくと，ルール間の整合性の維持や，副作用の評価が極めて困難となることであり，この点に留意してルールの生成・評価の自動化を図る必要がある．

確率に基づく枠組みでは，未登録語による誤りを考慮して，確率的語構成モデルを組み込む試みがなされている(例として[6]，[7]，[8])．そして，例えば[6]では，未登録語の存在するテキストに対して，recall, precision 共に95%程度の精度を実現している．また，文法や尤度基準の不備に対しては，文脈木を用いたvariable-gramのマルコフモデルを用いた学習を行う研究[9]，正解と誤りの比較を利用して，人手で付与した品詞体系を自動的に調整し，精度の改善を図る研究[10]等が行われている．

しかし，一般に確率に基づく枠組みでは，(1)高精度なモデルを構成するために解析された大量のデータが必要であるが，システム毎に異なる品詞セットについてこれを用意することは困難である，(2) 確率モデルは精緻化するとデータが巨大化し，一般に解析速度に悪影響が生じる，(3)確率モデルとして得られた知識は人間にとって可読性に欠け，ルール化できるような誤りがあっても，人間がルールの形で明示的に追加することは難しい，等の問題がある．また，確率モデルの形で得られた知見を，実際に数多く稼働しているルールに基づく形態素解析システム直接応用することは難しい．

1.3 本論文の目的

我々は，1.2に挙げた利点から，ルールに基づく形態素解析を開発しており[11]，改良の結果，現在速度の点では，十分実用的なレベルに到達しており(60,000文字/秒，DEC Alpha Server, 300MHz)，精度向上を今後の重要な目標としている．

本報の目的は，ルールに基づく形態素解析器に対し，後処理を用いてさまざまなタイプの解析誤りの修正を行うことにより，形態素解析の精度を向上させる手法を提案することである．

人手により与えられた比較的少量の正解データに基づき，形態素解析の誤り修正ルールを自動的に獲得する手法は，誤り駆動型教師付き学習としてBrillにより提案されている[12]が，語境界が変化する誤りを含む日本語形態素解析には直接適用できない．本論文では，日本語形態素解析への適用を目指した一つの応用方法を提案する[注1]．

更に，後処理は「窓」と呼ぶ文集合ごとに行う．後処理ルールと，窓内の文の解析結果の相互参照を組み合わせることにより，局所的なパターンマッチのみを用いる修正ルールでは困難な修正が可能となることが示される．

以下，2では，誤り駆動型・教師付き学習を説明し，我々の用いたルール獲得の手法について述べ，3では，後処理のアルゴリズムについて，4では実験結果について述べる．

2 書き換え規則に基づくアプローチ

2.1 誤り主導学習

既に述べたように，ルールに基づく形態素解析の枠組みでは，人手により数多くのルールを整合性を保って導入することは困難である．Brillは，ルールに基づく英語のPOS tagger（品詞付けプログラム）を対象として，解析結果と正解を比較することにより，誤りを低減するためのルールを自動的に抽出・追加する手法を提案した[12]．その枠組みは，「誤り主導型の変形ルール獲得」によるものであり，以下の3つの要素に基づいて構成される：

[注1] 日本語，トルコ語についての適用例はあるが，語境界の変更が扱われていないか[13]，もしくはBrillの枠組みで統一的に扱われてはいない[14]．

1 任意のPOS tagger．initial state annotatorと呼ばれ，初期値としての品詞付けを行う．
2 許容される変形ルール集合(変形ルールは，タグの書き換えと適用条件からなる)．
3 各ルールの学習適用前と適用後の比較し，その改善効果を評価する評価関数．

ここで，initial state annotatorは，理論上は与えられた語の可能な品詞をランダムに付与するものでも良いが，人手で構成したPOS taggerを用いるのが実際的である．許容される変形ルールの集合は一般に巨大な集合であるため，実際にはそれらを生成するテンプレート集合として与える．例えば，

Change tag **a** to **b** when
1. The preceding (following) word is tagged z.
2. The word two before (after) is tagged z.
where **a**, **b**, and z are variables.

は，**a**, **b**, zをinstantiationすることにより，前後一単語の品詞を見てそれに挟まれる語の品詞を書き換える変形ルール全体を生成する．また評価関数は，正解データが与えられている場合，正解との食い違いを評価するものとなる．

以下では，正しく解析された結果付きの学習用コーパスの存在を前提とする，教師付き学習の枠組みについて説明する．この場合，initial state annotatorの出力結果を書き換えてゆく手順が，変形ルールの順序付きリストとして与えられる：

step 1) initial state annotatorにより学習用コーパスにタグ付けし，これをTとする．
step 2) 許容される変形ルールすべてを個別にTに適用し，その結果が最も正解に近づくルールを評価関数により選び出し，順序付きリストOLに記録する．
いかなるルールでも結果が向上しないときは，終了してOLを返す．
step 3) このルールに基づきTを書き直したものを再びTとし，step 2にもどる．

このようにして，OLが得られたとき（このプロセスは必ず終了する），initial annotatorの出力を，OL中の変形ルールを順次用いて書き換えることにより，POS taggerの精度が改良できる．

2.2 日本語形態素解析のための書き換え規則の学習

ルールに基づく形態素解析の精度向上の方法として，起こりやすい誤りパターンを抽出し，そこから人手でルールを抽出する手法は既に提案されてきたが[3][4]，これを自動化するために，2.1で紹介した教師付きの誤り主導型書き換えルール獲得手法を応用する．

Brillの手法は，単語境界が与えられている英語を想定しているため，単語毎のタグの書き換えだけを前提としているが，日本語の場合は単語境界が与えられていないため，書き換え規則として単語境界の変更も考慮する必要があり，許容されるルール集合は格段に大きくなる．実際，日本語形態素解析のエラーには，次の4種類がある：

(A) 分割の誤り．以下の3種類に分かれる：
(A-1) 過分割
正：今日/の/金/相場/は，...
誤：今日/の/金/相/場/は，...
(A-2) 分割不足
正：ユニックス /ワークステーション
誤：ユニックスワークステーション
(A-3) その他の誤り（語境界交差型）
正：病気 / が / まん延 / .
誤：病気 / がまん / 延 / .
(B) 品詞のみの誤り
正：...」/と(引用助詞)/い/う
誤：...」/と(並立助詞)/い/う

このため，変形ルールのテンプレートをあらかじめ人手で用意することは困難であり，学習方法の検討が必要である．また，ルールの適用についても，順序付きリストに登録された順序に従い，初期形態素解析結果を何度も走査して書き直してゆくことは，速度の点で好ましくない．我々はこれらの点を考慮し，ルール獲得をデータ駆動型にし，ルール適用も形態素解析結果を1回走査するだけで可能なように簡略化した．本節と次節で，これらについて述べる

2.2.1 書き換え規則の生成

規則の生成には，正しく解析されたデータの存在を仮定する．書き換え規則は，解析結果と正解データの比較により以下の手順で自動的に生成する：

【第1段階】
K，Lを自然数とし，すべての誤り部分に対し，誤りの前K語，後L語を環境とし，誤り部分を正しく書き直す次のような語彙化ルールを生成する：

$$\mathbf{a}_1 \ldots \mathbf{a}_K W_1 W_2 .. W_n \mathbf{b}_1 \ldots \mathbf{b}_L \Rightarrow \mathbf{a}_1 \ldots \mathbf{a}_K W_1' W_2' .. W_m' \mathbf{b}_1 \ldots \mathbf{b}_L$$

ここで $W_1W_2..W_n$，$W_1'W_2'..W_m'$，$\mathbf{a}_1 \ldots \mathbf{a}_K$，$\mathbf{b}_1 \ldots \mathbf{b}_L$ は単語列であり，$W_1W_2 \ldots W_n$ は誤り部分，$W_1'W_2'..W_m'$ は正解，$\mathbf{a}_1 \ldots \mathbf{a}_K$，$\mathbf{b}_1 \ldots \mathbf{b}_L$ は書き換えを適用する環境である．以下，$W_1W_2..W_n$を第1パターン，$W_1'W_2'..W_m'$ を第2パターン，$\mathbf{a}_1 \ldots \mathbf{a}_K$を前方環境，$\mathbf{b}_1 \ldots \mathbf{b}_L$を後方環境と言う．実際には，K，Lがあまり大きいとデータスパースネスが生じるため，K=L=1とした．

例えば，誤り(A-1)からは，次の語彙化ルールが得られる：

(R1) 金：普通名詞／相：名詞接辞／
場：名詞接辞／は：副助詞
=>
金：普通名詞／ 相場：普通名詞／
は：副助詞

【第２段階】

第１段階で得られたすべての語彙化ルールを，次の一般化規則の組み合わせを用いて一般化する：

(G1)第１，第２パターンの品詞だけに注目する[2]．
(G2) 第１，第２パターンの文字を，文字クラスに置き換える．
(G3) 前方環境の品詞だけに注目する．
(G4)前方環境の文字を，文字クラスに置き換える．
(G5) 後方環境を無視する．
(G6) 後方環境の品詞だけに注目する．
(G7)後方環境の文字を，文字クラスに置き換える．
(G8) 後方環境を無視する．

例えば，G2，G3，G6を適用すると，R1 は，次のように書き換えられる：

(R1) 普通名詞／"C_1"：名詞接辞／
"C_2"：名詞接辞／助詞 =>
普通名詞／"C_1C_2"：普通名詞／助詞

ここで，"C_i"は単漢字を表す．実際には，予備実験により，一般化パターンとして22通りの組み合わせに限定した．

2.2.2 書き換え規則の信頼度

Brill の方法では，変形ルールは順序付きリストに格納され，initial state annotator の出力を順次書き換えるため適用される．これに対し，我々は，形態素解析の出力結果を1回走査するだけで後処理を終わらせることを意図し，ルール適用は，ルールの信頼度と，書き換えパターンの長さを用いた"Greedy method"(3.2)で行い，順序付きリストの作成は行わなかった．

ルールへの信頼度の付与は，次のようにして行う．まず，ルールRに対し，自然数の三つ組$\{T, CE, EC\}$を，以下で定義する：

T： ルールにより書き換えられた単語数．
CE：書き換え前に正しく解析されていた単語のうち，書き換えられてしまう単語数．
EC：書き換え前に誤って解析されていた単語のうち，正しく書き換えられる単語数．

三つ組$\{T, CE, EC\}$より，Rの信頼度rを以下で定義する：

$$r = \frac{EC - CE}{T}.$$

このままでは，低頻度のルールの信頼度が大きく評価される傾向があるため，「＋１ルール」と呼ばれる簡易手法で補正した．すなわち，仮想的に，3種類の書き換え(正→誤の書き換え，誤→正の書き換え，誤→誤の書き換え)が1回ずつ余計に生じたと考え，$\{T, CE, EC\}$を計算し直す．このとき，ルールにより書き換えられる単語数をnとすれば，書き換えられた３項組$\{T', CE', EC'\}$は，

$$EC' = EC + n,\quad CE' = CE + n,\quad T' = T + 3n$$

となり，これらを用いて，

$$r' = \frac{EC' - CE'}{T'}$$

により，信頼度を定義し直す．負の信頼度を持つルールは排除した．

以下は，得られたルールと信頼度の例である．ここで，G5(G8)の適用により，前(後)環境を無視した場合，これを*(wild card)におきかえた：

語彙化ルールの例：
*／と：並立助詞／い：子音動詞語幹 =>
*／と：引用助詞／い：子音動詞語幹
…… 信頼度：0.947

非語彙化ルールの例：
助詞／"C_1"：普通名詞／"C_2"：未登録語／助詞
=> 助詞／"C_1C_2"：普通名詞／助詞
…… 信頼度：0.667

廃棄されたルールの例：
*／"$C_1C_2C_3$"：固有名詞／
"C_4"：普通名詞／*
=> *／"C_1C_2"：普通名詞／
"C_3C_4"：普通名詞／*
…… 信頼度：-1.58

3 後処理アルゴリズム

本節では，以上に得られたルールと，文脈情報を組み合わせた後処理の方法について述べる．

3.1 考え方

initial state annotator として，我々はルール主導型の作表型形態素解析器を用いた[15]．この形態素解析器は，先験的に与えた原則に従って人手で付与した，品詞バイグラムに対するコスト関数を用いている．

後処理は，一文ごとでなく，「窓」と呼ぶ文集合を対象として行う．「窓」は，論理的な単位(記事や，章）を用いても，単なる数量的な単位を用いても定義しうるが，内容的につながった文の集合を含むことを意図しており，我々の実験では一記事を「窓」とした．

「窓」を利用する理由は，一定の範囲の文に繰り返し現れる表層文字列に注目すれば，単純なルールを用いて一文毎に処理していたのでは得られない効果が期待できるためである．ここで我々は，「窓内の文集合には，内容の関連性から，特定の内容語が繰り返して使われる傾向が高い」という

[2] 正確には，簡易化した品詞に置き換える．例えば，助詞は区別しない．

仮定に立っている．「ある一定の範囲の文集合の表層情報を利用する」ことの効果は，これまでにも，複合名詞解析[16]や，構文解析[17]において示唆されており，今回の実験では，形態素解析の修正における，未登録語の検出や曖昧性の解消に，文脈情報を反映させることを試みた．

形態素解析器は，各文に対して，一つの最小コスト解を出力すると同時に，最小コスト解が複数個ある場合には複数の可能性がある部分を，解析表に用いられるデータ構造であるlattice形式で保持できると仮定する[15]．

図1にlatticeの模式図を示した．この図では，"製品の価格差"の"価格差"の部分に，{"価", "格差"}と{"価格", "差"}の二通りの曖昧性があり，lattice形で保存されるる．この曖昧性は既存のルールだけでは解消されないため，例えば"価", "格差"が出力される．

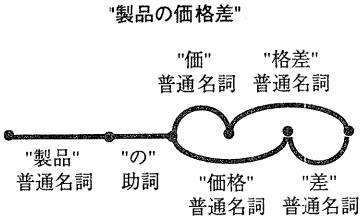

図1
曖昧さを保持するlatticeの例

出力された解は，1回走査され，3.2で説明する一種の"greedy method"に従って，変形ルールを適用する（変形ルールは，実際に出力される解に対して学習されたものであるから，この手順は自然である）．

ルールの適用可否については，信頼度の値がある閾値より大きいことと，書き換え結果の合文法性に関する条件の成立により判断される(3.2)．語境界を変更しない型のルールについては，単純に二つの条件を検査するだけであるが，語境界を変更するルール(例えば2.2.2の2番目のルール)が，名詞，動詞，形容動詞等の内容語を生成し，かつある場所における最も信頼度の高いルールであった場合，信頼性が予め設定した閾値より低いだけならば，適用箇所が他のルールによって書き換えられない限り，そのルールと適用位置を，「潜在的に適用されうるルール，潜在的に書き換えられる可能性がある箇所」の対として記録する．

これは，特定の単語が繰り返し出現し，かつその単語が正しく解析されない場合（未知語が出現する時に典型的に生じる)を想定したためである．すなわち，語境界を変えるルールは，解析結果と異なるなんらかの単語の存在を示唆するが，同じ「窓」の中の複数箇所で，異なるコンテクストのもとで同一の単語の存在が示唆されるならば，一つ一つのルールの信頼度は低くても，それらを総合すれば，その単語が実際に出現すると判断する妥当性が高まる考える．

この考え方に基づき，窓内の解析結果を通常の方法で書き換え終えた後，同一単語の存在を示唆する潜在的な適用可能ルールの信頼度の再計算を行い(3.3.1)，最終的な書き換えを決定する．実際，語境界を変える型のルールは単独では信頼度が低いことが多く，窓の概念を用いることによりはじめて適用可能となる場合がある．

「窓」のもう一つの利用法として，書き換えがすべて終わった後に，曖昧性解消手続きにおいて，「窓」内の非曖昧箇所における出現単語を利用することが考えられる．その詳細は3.3.2で述べる．

これらの効果は，4節で述べる．

3.2 ルールの適用

本節では，"greedy method"の一種と見なせる，ルールの適用方法について述べる．

解析結果は先頭から走査され，各位置毎に適用可能なルールをルール辞書の検索により抽出する．ルールが複数個ある場合，最も信頼度の高いルールを選び出し，そのようなルールが複数個存在する場合，書き換え対象語数がもっとも多いルールを選ぶ．

更に，そのルールが，予め定めた条件を満たす場合，そのルールを適用し，書き換え後の位置から，単語列とルールの照合を再会する．ルールの適用がなかった場合，ルール照合開始単語位置を1語分進める．

ルールは，この段階では，次の条件を満たす場合にのみ適用される：

(1) 信頼度が，あらかじめ定めた閾値より高い．

(2) 書き換えの結果，非文法的な接続を含む単語列が生成されない．

前節で述べたように，語境界を変更し，新たに語を生成するルールが，ある場所における最も信頼度の高いルールであり，かつ上記の条件(1)のみを満たさない場合，そのルールとその箇所を，「潜在的に適用されうるルール，潜在的に書き換えられる可能性がある箇所」の対として記録する．

なお，ルールはTRIE構造を持ったルール辞書により保持され，形態素解析結果の各位置に適用可能なルールを効率的に検索することができる．

3.3 周辺情報の利用

本節では，3.1で述べた，周辺情報の利用方法の詳細について述べる．

3.3.1 ルールの信頼度の窓内での再評価

3.2で述べた方法にもとづき，窓内の解析結果について，最初の書き換えが終了したとする．書き換えの中には，

助詞／"C_1"：普通名詞／"C_2"：未登録語／助詞
=> 助詞／"C_1C_2"：普通名詞／助詞

のように，単語境界を変えるもの（未登録語を生成する場合がある）や，

*／と：並立助詞／い：子音動詞語幹

=> *／と：引用助詞／い：子音動詞語幹

のように，品詞を書き換えるだけのものがある（*はwild card）．

単語境界を書き換えたすべてのルールと，実際には適用されていないが単語境界を変える可能性があるとして記録されたすべてのルールを，単語境界の変更の結果として得られる単語wにより分類する．特定の窓における，単語wを生成する（可能性のある）ルールの信頼度の再評価を，次の方法で行う．

単語wを生成する（可能性のある）ルールの集合を$\{R_i,..,R_k\}$とし，R_i の信頼度をr_i $(i = 1,..., k)$とする．このとき，「単語wが窓内に現れる」という事象の信頼度を，

$$r(w) = 1 - (1 - r_1) \times \cdots \times (1 - r_k).$$

で定義する．ここで，$\{R_i,..,R_k\}$中で，ルールとして等しく，かつ適用箇所の字面上のコンテクストが共通なものがある場合，$r(w)$の定義における乗算には，それらのうち任意の一個のみを反映させる．

$r(w)$がある閾値を越えた場合，すべても場所にwが実際に存在するとして，$\{R_i,..,R_k\}$のうち，最初の書き換えにおいても適用されなかったルールを適用し，解析結果を書き換える．wが未登録語であれば，未登録語が検出されたことになる．

3.3.2 曖昧性解消

3.1で述べたように，曖昧な箇所はlatticeの形式で保存されているが，ルールの適用後も曖昧な箇所が書き直されなかった場合，曖昧性解消を行う．

窓の中の大部分は，曖昧性無く解析されているため，各曖昧箇所について，窓内に曖昧性無く現出現している内容語を多く含むものを最適パスとして決定する．この方法で唯一の最適パスが定まらないときは，書き直しを行わない．

例の場合，図2のように，窓内の非曖昧箇所に，"価格"が多く出現しているならば，{"価格"，"差"}が選択される．

…工業製品に対しては，内外での価格が著しく異なる……
……
……
製品の価格差が生じる原因については，……
……
……このように価格の多重化が……
……
……

図2
窓内の文における単語の出現例

この方法は単純であるが，実験の結果，解析結果の精度向上に有効であることがわかった（4参照）．

ここで注意すべきことは，ルールは固定されていても，簡単な文脈情報と組み合わせることにより，場合に応じて異なる結果を返すことであり，ルールだけを用いた方法の限界を越えうる可能性を示唆している．

3.3.3 後処理の時間計算量

窓中の文を形態素解析する時間計算量は，窓内の文字数をnとして$O(n)$である．ルールを用いた通常の書き換えは，"greedy method"の定義から，明らかに$O(n)$で実行可能である．信頼度の再計算と，二回目の書き換えも，明らかに$O(n)$で実行可能である．

窓中の曖昧箇所の検出，非曖昧箇所の単語の検出，曖昧性箇所の再解析もすべて，時間計算量$O(n)$で実行できる（個別の証明は容易である）．

これらから，後処理は文字列の長さをnとして，$O(n)$で実現可能である．

4 実験結果

4.1 実験対象

日経新聞から抽出した142記事，1500文を，形態素解析後人手で修正した正解データを作成し，750文を用いてルールの抽出と評価を行い，残りの750文を用いて後処理の効果を調べた．

学習データは22,280語(異なり数4,082)，テストデータは22,648語(異なり数3,675)からなる．

4.2 形態素解析器

実験で用いた形態素解析器は，[15]で述べたもので，用いた辞書の見出し数は67,443であり，比較的小さく，コーパスに対するチューニングは行っていない．このため，学習データは481語の未登録語を含み，(異なり数344) テストデータは626語の未登録語(異なり数329)を含む．この条件は，比較的未登録語が多く，精度が低下しやすい場合を想定している．

形態素解析器の精度は，recallとprecisionで評価する．recallとprecisionは，で定義される：

$$Recall = \frac{\text{正しく解析された単語数}}{\text{正解中の単語数}},$$

$$Precision = \frac{\text{正しく解析された単語数}}{\text{解析結果中の単語数}}.$$

表1に形態素解析器の後処理前の精度をあげる．精度は，学習データ，テストデータの両方で，未登録語を辞書に追加しない場合と，追加した場合に分けて評価した．

Table 1: 形態素解析器の精度

	未登録語を辞書に追加せず		未登録語を辞書に追加	
	recall	precision	recall	precision
学習データ	94.8%	93.2%	97.6%	96.8%
テストデータ	94.1%	92.7%	97.6%	97.0%

4.3 後処理の効果

2.2.2で述べた信頼度評価により，獲得されたルールのうち，適用の対象として考慮されるのは約3,000 個となった．

後処理の結果は，形態素解析プログラムの本来の品詞セット(420種類)と，情報検索を意図して簡易化した品詞セット(各品詞における機能細分を簡略化した13大品詞)を用いて評価した．後者は，情報検索のためのインデキシングに利用する場合の，実効的な精度を見るためである．

表2に示すごとく，形態素解析の精度は，学習データで5%，テストデータで3%向上した．

Table 2: 後処理の効果

	本来の品詞集合		簡略化した品詞集合	
	recall	precision	recall	precision
学習データ	98.7%	98.3%	99.5%	99.3%
テストデータ	96.6%	95.7%	97.8%	97.1%

4.4 効果の分析

後処理により，誤りのうちの603個が正しく修正され，そのうち246個は，語境界を変更する型の書き換えによる．

後処理により，新たに誤りが生じた件数は23件であり，そのうち実際には無い単語を生成したものが4例，曖昧性解消に失敗して正解を誤りに書き換えたものが2例あった．

未登録語は全部で41個特定され，そのうち「窓」の文脈情報によるものは18であった．「窓」による正しい曖昧性解消は21例であり，修正の約7%は「窓」による効果であった．

以下は，後処理により修正された語分割，後処理により検知された未登録語の例である：

修正された語分割（未登録語含まず）
相場：普通名詞
米国：固有名詞

正しく検出された未登録語：
中村屋：固有名詞
預入：サ変名詞

誤って検出された未登録語
米ロ：普通名詞
英仏：固有名詞

以下は，曖昧な箇所が後処理により正しく解消された例と，誤って書き直された例である：

正しく解消された曖昧性の例
{"価：普通名詞／格差：普通名詞",
"価格：普通名詞／差：普通名詞"}
→
"価格：普通名詞／差：普通名詞"

{"金：普通名詞／外：普通名詞／信託：サ変名詞",
"金：普通名詞／外信：普通名詞／託：サ変名詞"}
→
"金：普通名詞／外：普通名詞／信託：サ変名詞"

曖昧性解消により誤って書き換えられた例
{"評：サ変名詞／議会：普通名詞",
"評議：サ変名詞／会：普通名詞"}
→"評：サ変名詞／議会：普通名詞"

5 おわりに

5.1 結論

本報では，変形ルールと文脈情報を組み合わせて用いる日本語形態素解析の後処理方法を提案した．変形ルールは，正解データと形態素解析結果の差から自動的に生成され，評価される．変形ルールは，品詞の書き換え・形態素境界の移動の双方をカバーし，さまざまな抽象度のルールを含む．後処理により，形態素解析の精度を，テストセットにおいて3%程度向上させることができた．誤り修正のうちの7%は，文脈中の表層情報の参照により達成され，文脈情報利用の効果が確認された．

なお，本報の後処理自体は，形態素解析器本体がルールに基づいている必要はないため，システムの統一性を考慮しなければ，確率モデルに基づく形態素解析器にも適用可能である．

5.2 今後の課題

- コーパスサイズ

本報告で提案した方法は，基本的にあまり多くの正解データが存在しないという前提で適用されるものである．しかし，もう少し多くの正解データがある場合，ルール抽出用，ルール評価用，テスト用に3分割し，後処理の効果を調べることが考えられる．

- 学習方法

本報告では，Brillの手法を簡易化し，"greedy method"によりルールの適用を行ったため，ルールの順序付きリストは生成しなかった．これは，解析結果の走査を繰り返し行うことを避けたためであるが，順序付きリストを利用した場合の効果を調べることは興味深い課題である，

- 窓の定義

本報では，新聞記事を解析の対象としたため，「窓」の定義として一記事を取ることは極めて自然であった．しかし，このような自然な処理単位

が存在しない場合，文数や文字数で「窓」を定義することを検討しなければならない．また，窓と窓のオーヴァーラップにより，情報を直後の窓に伝搬させること，高い確度で獲得された未登録語を暫定的に後の解析で使用すること，などの効果を調べることは今後の課題である．

◦ 実装方法

本研究により，後処理により一定の効果が期待できることがわかった．今後，これを我々の開発した形態素解析器[11]に適用するため，その速度に追随できる実装方法を検討する必要がある．

謝辞

図書館情報大学の藤井敦先生には，Brillの手法を応用した論文について御教示頂きました．感謝致します．

参考文献

[1] JUMAN Ver. 3.5, http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[2] 茶筌 version 1.0，http://cactus.aist-nara.ac.jp/lab/nlt/chasen/manual/manual.html

[3] 西野文人，未登録語テンプレートを用いた日本語形態素解析，情報処理学会第39回全国大会論文集，2F-2, pp.594-595 (1989)

[4] 山田宏忠，統計情報を用いた日本語形態素解析，言語処理学会第3回年次大会論文集，pp.417-420 (1997)

[5] 横尾昭男，白井諭，奥山信輔，河村美砂子，池原悟，日本語形態素解析の誤りの回復について，言語処理学会第3回年次大会論文集，pp.429-432 (1997)

[6] Nagata, M. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. of COLING'94* (1994), pp.201-207

[7] 伊藤伸恭，西村雅史，*N*-gram を用いた日本語テキストの単語単位への分割，NL研資料，NL122-9, pp.57-62 (1997)

[8] 森信介，長尾真，形態素クラスタリングによる形態素解析精度の向上，自然言語処理，Vol.5, No.2, pp.75-98 (1998)

[9] M. Haruno and Y. Matsumoto. Mistake-driven mixture of hierarchical tag context trees, Proc. of 35th Annual Meeting of ACL and 8th Conf. of EACL, pp.230-237 (1997)

[10] 北村啓，宇津呂武仁，松本祐治，誤り駆動型の学習モデルによる日本語形態素解析，NL研資料，NL124-6, pp.41-48 (1998)

[11] 櫻井博文，久光徹，形態素解析プログラムANIMAの実装と評価，情報処理学会第54回全国大会論文集(2)，pp.57-59 (1997)

[12] 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, pp.543-565 (1995)

[13] 斉藤稔，誤りの分析に基づく形態素解析用ルール構築，東京工業大学卒業論文 (1998)

[14] Oflazer K.and Tur, G. Combining Hand-crafted Rules and Unsupervised Learning in Constraint-based Morphological Disambiguation, *Proc. of EMNL* (1996)

[15] 久光 徹 新田義彦，ゆう度付き形態素解析用の汎用アルゴリズムとそれを利用したゆう度基準の比較，電子情報通信学会論文誌 D-II Vol.J77-D-II No.5 (1994), pp.959-969

[16] Hisamitsu, T. and Nitta, Y. Analysis of Japanese Compound Nouns by Direct Text Scanning, *Proc. of COLING'96*, pp.550-555 (1996)

[17] Nasukawa, T. Full-text processing: improving a practical NLP system based on surface information within the context, *Proc. of COLING'96*, pp.824-829 (1996)