

誤り駆動による品詞タグづけ統計モデルの拡張

浅原 正幸 松本 裕治

奈良先端科学技術大学院大学 情報科学研究科

{masayu-a,matsu}@is.aist-nara.ac.jp

統計的品詞タグづけにおいて、マルコフモデルは普及した手法の一つである。しかし、単純なマルコフモデルでは例外的な言語現象に対処することができない。この問題に対処するため、本論文では単語レベルの統計値の利用と選択的 tri-gram の二つの拡張を提示する。しかし、これらの拡張のための、単語や tri-gram 文脈の選択を手で設定することは困難である。そこで、本稿では素性選択に誤り駆動の手法を導入した。英語と中国語について実験を行い、これらの拡張の有効性を検証した。

キーワード: 統計的形態素解析、機械学習、マルコフモデル、選択的 tri-gram、誤り駆動による手法

Error-driven extensions of Statistical Learning Models for POS tagging

Masayuki Asahara Yuji Matsumoto

Graduate School of Information Science, Nara Institute Science and Technology

{masayu-a,matsu}@is.aist-nara.ac.jp

Markov Model is a popular instrument for a statistical Part-of-Speech tagging. However, the normal model cannot cope with exceptional phenomena. To remedy the weakness, we introduce two types of extension to the model. One is to use word-level statistics, and the other is selective tri-gram. We introduce an error-driven method for the selection of words and tri-gram contexts. We show how our extension is effective through various experiments.

Keywords: Statistic Morphological Analysis, Machine Learning, Markov Model, Selective tri-gram, Error-driven methods

1 はじめに

品詞タグづけ (Part-of-Speech tagging) の問題に対して、多くのコーパスに基づく手法が提案されている。統計的手法の一つとして、品詞 tri-gram マルコフモデルを用いる手法がある。しかしながら、通常の tri-gram モデルではデータ量の不足から、学習データに対して過学習する恐れがある。さらに、tri-gram モデルには、例外的な現象に対応できないという問題がある。これらの問題に対処するために、多くの拡張モデルが開発されてきた。Ron [5] は可変長マルコフモデルを提案した。Cutting [3] は連接確率行列上へのグループ化を導入した。Asahara [1] は日本語

形態素解析の問題に対し、単語レベルの統計値の利用や、階層構造を持った品詞体系に対する各件でのグループ化を導入し、さらに Ron の拡張と異なる手法による選択的 tri-gram を提案した。

このようにマルコフモデルに複雑な拡張を採用した際、各拡張に対し適切な素性選択を決定する必要がある。選択すべき素性の例として、個別視する単語や、選択的に利用する tri-gram などがある。これらの素性選択は人手では扱いにくく、特に母国語でない言語に対しては決定が難しい。北内ら [6] は、階層構造を持つ品詞タグ集合上のグループ化の自動決定に誤り駆動の手法を提示している。この手法では、グループ化は品詞の階層構造上の深さとして定義さ

れ、その深さ決定に誤り駆動の手法を採用している。

本敵では、品詞タグづけのための拡張統計モデルについて述べる。我々のモデルでは、単語レベルの統計値と選択的 tri-gram の 2 つの拡張を導入した。さらに、これらの拡張のための素性選択に誤り駆動の手法を導入した。これにより、例外的な現象に対応し、tri-gram 規則を減らした統計モデルを構成することができた。英語と中国語の品詞タグづけに対し評価実験を行い、その有効性を検証した。

2 品詞タグづけのための拡張モデル

まず、品詞タグづけ統計モデルのための二種類の拡張について述べる。一つは単語レベルの統計値で、もう一つは選択的 tri-gram である。

単語レベルの統計値の利用とは、ある単語を異なる品詞として扱うことを意味する。本モデルでは bi-gram 文脈において品詞タグ接続確率を計算する際に、前件と後件とで異なる品詞タグ集合を設定することを許す。これは、単語レベルの統計値を利用する単語について、各件で異なる集合を定義することを意味する。この手法により、例外的なふるまいをする単語に対応できる統計モデルを構築することができる。

選択的 tri-gram は、通常の bi-gram モデルをベースとしている。本モデルでは、tri-gram 文脈を bi-gram モデルの例外的規則として追加する。通常の tri-gram モデルの場合、学習モデルに対して過学習してしまう危険性がある。選択的 tri-gram モデルを採用することにより、過学習を抑制しながら bi-gram 文脈では対応できない現象にも対応することができる。さらに、選択的 tri-gram は統計モデル中の tri-gram 接続規則を削減することができ、解析速度を改善することができる。

次に、これらの拡張のためには、拡張する単語や、選択される tri-gram 文脈の決定が必要になる。しかし、この素性選択を手で行なうことは難しい。そこで、我々は誤り駆動による手法を導入する。これは、多くの誤りを生成する単語や tri-gram 文脈こそが拡張すべきであるという仮定に基づく。本手法では、素性を誤りの多いものから順に追加していく手

法を取る。選択された素性が統計モデルを改善した場合にのみ、その素性を拡張する。これを繰り返すことにより、統計モデルを漸進的に変形していく。

本節では、これらの手法の詳細について提示する。

2.1 単語レベルの統計値

接続確率を見た場合、いくつかの単語は同じ品詞である他の単語とは別のふるまいをする。これらの単語を扱うために、単語レベルの統計値を利用する。これは、いくつかの単語を個別の品詞としてみなすことを意味する。さらに、この単語の選択について、前件と後件とで異なる単語を選択するようにした。

最初のタグ集合 \mathcal{T} に対し、いくつかの単語を拡張した二種類のタグ集合に拡張する。新しく定義される後件のタグ集合を \mathcal{T}^c とし、前件のタグ集合を \mathcal{T}^p とする。語彙化されたタグに対する確率式の修正を次に示す。

前件の単語 w_{i-1} が拡張された場合、単語生起確率は変更しない。品詞接続確率は次のようになる：

$$\begin{aligned} P(t_i|t_{i-1}) &= P(t_i|w_{i-1}) \\ &= \frac{F(w_{i-1}, t_i)}{F(w_{i-1})} \end{aligned}$$

後件の単語 w_i が拡張された場合、単語生起確率は次のようになる：

$$\begin{aligned} P(w_i|t_i) &= P(w_i|w_i) \\ &= \frac{F(w_i)}{F(w_i)} \\ &= 1 \end{aligned}$$

また、品詞接続確率は次のようになる：

$$\begin{aligned} P(t_i|t_{i-1}) &= P(w_i|t_{i-1}) \\ &= \frac{F(t_{i-1}, w_i)}{F(t_{i-1})} \end{aligned}$$

注意すべき点として、品詞中の単語が語彙化された場合に、品詞レベルの統計値が変更される点である。品詞タグ集合 \mathcal{T} 中の品詞タグ A と B について考える。後件の品詞タグ集合 \mathcal{T}^c について、単

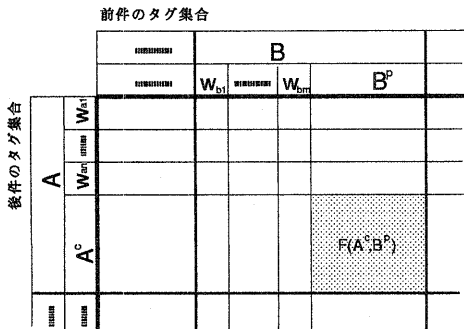


図 1: 単語レベルの統計値の利用

語 $w_{a_1}, \dots, w_{a_n} \in A$ が語彙化された場合、品詞タグ $A^c \in \mathcal{T}^c$ は次のようになる:

$$A^c = A \setminus \{w_{a_1}, \dots, w_{a_n}\}$$

同様に、前件の品詞タグ集合 \mathcal{T}^p について単語 $w_{b_1}, \dots, w_{b_m} \in B$ が語彙化された場合、品詞タグ $B^p \in \mathcal{T}^p$ は次のようになる:

$$B^p = B \setminus \{w_{b_1}, \dots, w_{b_m}\}$$

連接 $A-B$ の確率を推定するために、頻度 $F(A, B)$ ではなく $F(A^c, B^p)$ を利用する。図 1 に、単語レベルの統計値を利用した際の頻度推定を示す。

2.2 選択的 tri-gram

選択的 tri-gram とは、通常の bi-gram モデルに tri-gram 文脈を部分的に追加するモデルである。実際、品詞決定に tri-gram の長さの文脈を必要とする場合は少ない。さらに本手法では、語彙化した品詞の追加により品詞タグの数が増えるためこの手法は非常に効果的である。

本モデルでは tri-gram 文脈を例外的なものとして考える。Bi-gram 文脈と tri-gram 文脈が交わりを持つ場合、tri-gram 文脈は bi-gram 文脈の例外的規則としてみなし、全ての文脈は互いに交わりを持つことはない。Bi-gram 文脈が tri-gram 文脈と重なったとき、bi-gram 文脈の統計量として tri-gram 文脈を除外したものを用いる。

Tri-gram 文脈 $A-C-B$ をモデルに利用するとする場合、bi-gram 文脈 $C-B$ の統計値は次のよう

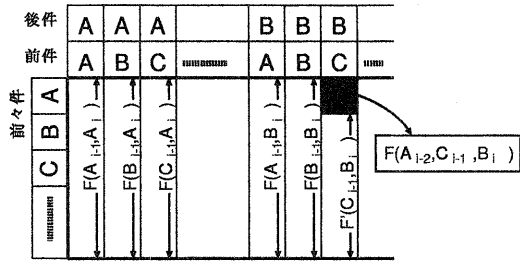


図 2: 選択的 tri-gram

になる (F はコーパスから学習される真の統計値を表し、 F' は解析に利用される推定頻度を表す):

$$F'(C_{i-1}, B_i) = F(C_{i-1}, B_i) - F(A_{i-2}, C_{i-1}, B_i)$$

図 2 に、選択的 tri-gram における頻度推定を示す。

2.3 誤り駆動による素性選択

単語レベルの統計量を利用するためには、どの単語を語彙化した品詞として定義するかを決定する必要がある。また、選択的 tri-gram は、どの tri-gram 文脈を選択するか決定する必要がある。しかし、これらの決定を人手で行うのは非常に困難である。この決定を自動化するために、誤り駆動による手法を導入する。

この方法は、多くの誤りを生成する素性は拡張すべきであるという仮定に基づいている。もし、単語が多くの誤りを生成する場合、その単語の拡張は精度を改善できると推測される。また、多くの誤りを生成する tri-gram 文脈は、その tri-gram 文脈を追加することによりモデルを改善できると推測される。

誤り駆動による素性選択の方法を具体的に示す。3種類のコーパスを利用する。最初のコーパスは、マルコフモデルのパラメータ推定に用いる。このモデルを用いて、2番目のコーパスを評価する。2番目のコーパスの解析結果から、誤りの多い素性を選択する。最後に、選択された素性を拡張したモデル(マルコフモデルのパラメータ推定には最初のコーパスを用いる)を用いて、3番目のコーパスを評価する。3番目のコーパスの精度をも改善する場合に、その新しい素性を統計モデルに導入する。改善しない場合

には、その素性は破棄する。この手順を繰り返すことにより、統計モデルを漸進的に強化していく。

3 評価

本手法を評価するために、英語と中国語の2種類の言語のコーパスを用いた評価実験を行なった。英語のコーパスとして PennTreebank tagged corpus(52725 sentences) [4] を利用し、中国語のコーパスとして Academia Sinica Balanced Corpus (284888 sentences)¹を利用した。

3.1 実験手順

実験手順は、英語も中国語も共通している。まず最初に、コーパスを5つのコーパス(A, B, C, D, E)に分割する。コーパスA, B, Cを素性選択に用い、コーパスD, Eを評価に用いる。ここで素性とは、語彙化した品詞として定義する単語や、選択的に導入するtri-gram 接続を意味する。

● 素性選択用データ

- A: マルコフモデルパラメータ推定用コーパス
- B: 素性選択用コーパス
- C: 素性決定用コーパス

● 評価用データ

- D: マルコフモデルパラメータ推定用コーパス
- E: 評価用コーパス

手順を以下に示す。以下の手順を繰り返す。

1. 初期化

まず、通常の bi-gram モデルをコーパス A を用いて作成する。その後、このモデルを用いてコーパス B を評価する。

2. 素性選択

最も多くのエラーを出す素性をコーパス B の解析結果から選択する。

¹It is obtainable through Computational Linguistic Society of R.O.C.

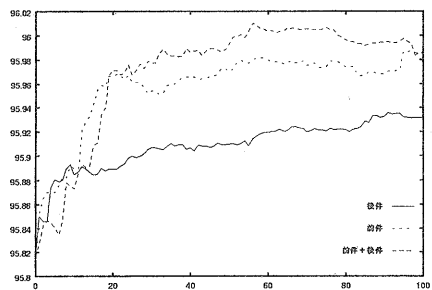


図 3: 語彙化する単語選択 (英語)

3. 素性決定

選択された素性をモデルに一時的に追加する。この素性選択に基づき、マルコフモデルのパラメータをコーパス A から推定する。その後、コーパス C を新しいモデルで評価する。もしコーパス C が改善された場合、その素性を採用する。改善されなかった場合、その素性を破棄する。破棄された素性は、再び選択されることはない。

4. 評価

決定された素性を基にして、マルコフモデルのパラメータをコーパス D から推定する。このモデルを用いてコーパス E を評価する。

以下に示す評価は、コーパス E によるものである。

3.2 単語レベルの統計値

誤り駆動による語彙化する単語の選択を評価するために3種類の実験を行なった。

- 前件についてのみ単語を拡張する実験
- 後件についてのみ単語を拡張する実験
- 前件と後件を同時に単語を拡張する実験

図 3 に英語コーパスによる実験結果、図 4 に中国語コーパスによる実験結果を示す。

英語の実験において、選択された単語を表 1 に示す。

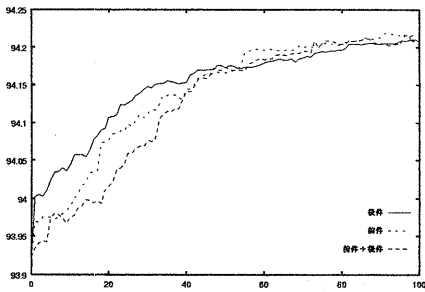


図 4: 語彙化する単語選択 (中国語)

表 1: 選択された単語 (英語)

前件		後件	
品詞	単語	品詞	単語
DT	the	VBD	called
IN	as	JJ	much
VBN	been	JJ	clear
NNP	Hong	NNP	Hong
VBZ	is	RB	up
WDT	which	RB	off
VB	be	IN	though
RB	not	JJ	hard
RB	also	IN	about
DT	a	IN	ago
PRP	it	NN	yen
VBZ	has	IN	that
RB	just	NNP	Los
WDT	that	VBN	called
:	-	NN	stock-index
IN	of	RB	out
CC	or	NNP	Wall
NNP	Wall	RBR	lower
RB	even	VBN	said
JJ	chief	NNP	GNP
DT	an	JJ	due
RB	as	RB	so
VBZ	's	RB	as
PRP	them	VBZ	's
POS	's		
DT	The		

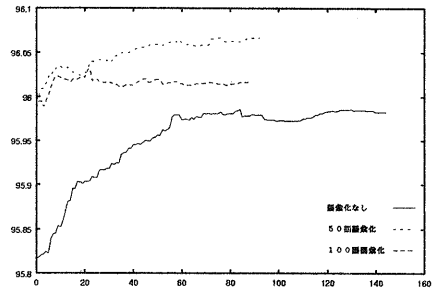


図 5: 選択的 tri-gram 単位 $P' - P$ (英語)

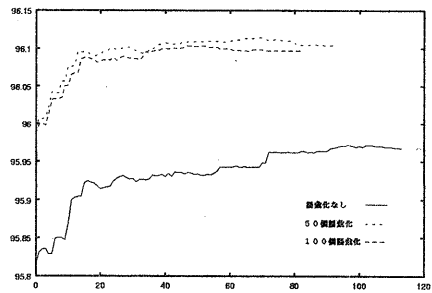


図 6: 選択的 tri-gram 単位 $P - C$ (英語)

3.3 選択的 tri-gram

まず、個別の tri-gram 文脈の単位での追加で実験を行なった。しかしこの単位では、精度の変化が小さく、有用な素性選択をすることができなかった。そこで、我々は 2 つの単位を定義した。1 つは前件と後件を共有する tri-gram 接続の集合 ($P'_1 - P - C, P'_2 - P - C, \dots, P'_n - P - C$) (単位 $P - C$ と呼ぶ) であり、もう一つは前々件と前件を共有する tri-gram 接続の集合 ($P' - P - C_1, P' - P - C_2, \dots, P' - P - C_n$) (単位 $P' - P$ と呼ぶ) である。

これらの文脈は、エラーを生成する文脈のみを追加した。エラーを生成しない場合、その tri-gram 文脈は追加されない。このため、エラーを生成しない文脈は、bi-gram 文脈のルールとして定義される。

選択的 tri-gram の評価について、単語レベルの統計値の利用との関係性を調べるために複数の評価を行なった。各言語に対し、語彙化を用いないモデルと、単語 50 個を語彙化した品詞として利用するモデルについて評価実験を行なった。英語については、単語

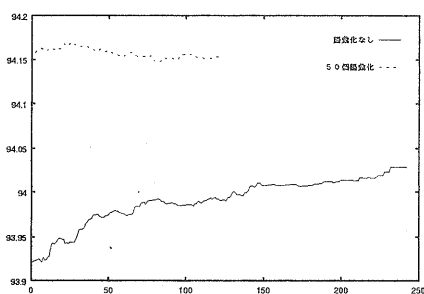


図 7: 選択的 tri-gram 単位 $P' - P$ (中国語)

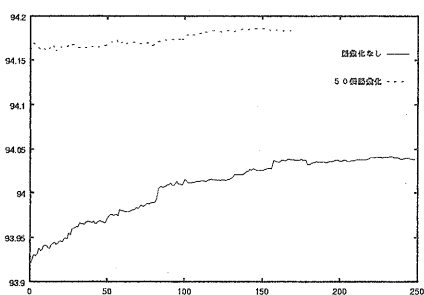


図 8: 選択的 tri-gram 単位 $P - C$ (中国語)

100 個を語彙化した品詞として利用するモデルについても評価を行なった。この単語の選択には、前節の実験で得られたものを利用した。

図 5, 6 に英語コーパスの実験で得られた結果を示す。

図 7, 8 に中国語コーパスの実験で得られた結果を示す。

4 考察

4.1 各コーパスの性質

英語コーパスについて以下のことが言える。図 3 の結果から、前件の品詞を語彙化の方が有効であることがわかる。実際、前件と後件を同時の拡張した際に、前件の方が多く選択される。また語彙化すべき単語数は 60 個前後で上限に達している。図 5, 6 の結果から、前々件と前件が同じものでグループ化したもの (単位 $P' - P$) よりも、前件と後件が同じも

のでグループ化したもの (単位 $P - C$) の方が精度が良いことがわかる。

また、中国語コーパスについて以下のことが言える。図 4 の結果から、選択する単語 100 個以下では、前件、後件、前件 + 後件ともにほぼ単調に上昇している。また、図 7, 8 の結果から、語彙化の方が有効なのに比べて、tri-gram 連接への延長がさほど効果がないことがわかる。特に 50 個語彙化したものについては、tri-gram 連接を追加していくことによる改善があまり見られない。これは、中国語コーパスの方が、同じ品詞に属する単語間で曖昧性が大きいことを示している。このことから、中国語形態素解析器の作成のためには、品詞体系の詳細化によって、より大きい改善を達成できると考える。

4.2 連接数の増減と精度

表 2, 3 に、各拡張と連接数の増減を示す。

品詞の語彙化は、精度改善に有効であることがわかる。品詞を語彙化するにつれて、全 tri-gram を利用するモデルの場合、学習データへの過学習により、評価データの精度が落ちることがわかる。選択的 tri-gram により、通常の tri-gram モデルと比べて、連接数を 5 分の 1 以下に削減することができる。また、品詞の語彙化の精度を維持したまま、精度改善に必要な tri-gram 連接のみを追加することができる。

5 まとめ

本論文では、品詞タグづけのための拡張統計モデルを提示した。単語レベルの統計値の利用と選択的 tri-gram の 2 種類の拡張を導入した。

単語レベルの統計値の利用により、例外的な言語現象にも対応できるようになった。しかし、語彙化された品詞が増えることにより、通常の tri-gram モデルだと、tri-gram 規則が増え、学習データに対し過学習してしまう。そこで、選択的に tri-gram 規則を利用することにより、連接規則を減少させ、過学習を回避することができた。

語彙化した品詞として定義する単語の選択や、利用する tri-gram の選択は人手で行うのは困難である。この素性選択の問題に誤り駆動の手法を導入し、自

表 2: 各拡張と tri-gram 接続数 (英語)

dataset	語彙化	精度	tri-gram 接続数	全接続数
単純 bi-gram	なし	95.816	0	1309
全 tri-gram	なし	96.040	11550	12859
tri-gram(P ¹ -P)	なし	95.982	780	2089
tri-gram(P-C)	なし	95.967	750	2059
bi-gram	50 語	95.992	0	2125
全 tri-gram	50 語	95.079	14637	16762
tri-gram(P ¹ -P)	50 語	96.066	286	2411
tri-gram(P-C)	50 語	96.102	516	2634

表 3: 各拡張と tri-gram 接続数 (中国語)

dataset	語彙化	精度	tri-gram 接続数	全接続数
単純 bi-gram	なし	93.921	0	2213
全 tri-gram	なし	93.928	25281	27494
tri-gram(P ¹ -P)	なし	94.028	1826	4039
tri-gram(P-C)	なし	94.038	1861	4074
bi-gram	50 語	94.161	0	3400
全 tri-gram	50 語	92.621	30671	34071
tri-gram(P ¹ -P)	50 語	94.152	505	3950
tri-gram(P-C)	50 語	94.180	982	4382

動的に効率的にモデルを改善することができた。結果、品詞タグづけの精度を向上させ、なおかつ接続規則の数を減らすことができた。

```
I got $ 1,000,000 in New York .
* * * * *
# # # # #
```

図 9: Tokenization と複合語処理

6 今後の課題

6.1 誤り駆動の閾値の設定

今回の実験では、素性を選択する際の、誤りの数や精度の変化などといった閾値については吟味していない。現在、1つでも誤りを犯した素性について、その素性を選択して追加した際に精度を下げなかった場合にのみ追加している。今後、これらの有用な閾値を画策していく必要がある。

6.2 実用的な英語 POS Tagger の開発

現在、本学習モデルを基に、現在英語版「茶筌」を開発している。実用的な POS Tagger を作成するために、以下のような仕様で開発中である。

- Tokenization

公開されている多くの POS Tagger は Tokenization を他のツールを用いて行う必要がある。現在開発中の英語版「茶筌」では、文字 n-gram のマルコフモデルによる Tokenizer を、POS Tagger にカスケード接続して Tokenization も行なっている。

● 階層的な品詞体系

PennTreebank の品詞体系のままでは、前置詞などの機能語の分類がされておらず、あまり実用的でない。現在、一部の品詞を詳細化することにより、階層構造をもった品詞体系を整備中である。

● 複合語処理

PennTreebank では「Hong Kong」のような語が、「Hong」「Kong」の 2 語に分割して登録されている。実際、語彙化する単語 (表 1) には、「Hong」(Hong Kong)、「Los」(Los Angeles)、「Wall」(Wall Street Journal) というような単語が選択されている。開発版の「茶釜」では、Tokenization 後の辞書引き位置をスペース (文頭を含む) の直後として定義をし、複合語辞書を整備することにより、これらの単語に対応している。図 9 に、辞書引き開始位置と登録する辞書の単位を示す。1 行目は Tokenizer の出力である。単語間をスペースによって区切られて出力される。* が辞書引き開始位置、# が単語 (複合語) として辞書に登録されている範囲を示している。

● 未知語処理

非膠着語 (わかち書きされる言語) では、未知語の単語境界は容易である。Part of Speech Tagger TnT [2] では、語頭、語尾や大文字・小文字の情報などから、未知語についての品詞推定を行っている。これらの処理も取り込んでいきたいと考えている。

6.3 中国語形態素解析器の開発

また、中国語形態素解析器の開発も着手している。本敵の実験結果から、現在利用しているコーパスは、同じ品詞に属する単語間の曖昧さが大きいことがわかった。今後、品詞体系を整備し、その品詞体系を基にしたコーパスの整備が必要であると考えている。

参考文献

- [1] M. Asahara and Y. Matsumoto. Extended Models and Tools for High-performance Part-of-Speech Tagger. In *Proceedings of COLING 2000*, 2000.
- [2] T. Brants. TnT - A Statistical Part-of-Speech Tagger. In *6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics *ANLP-NAACL 2000 Proceedings, and Proceedings of the ANLP-NAACL 2000 Student Research Workshop*, pp. 224-231, 2000.
- [3] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.
- [4] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: PennTreebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313-330, 1993.
- [5] D. Ron, Y. Singer, and N. Tishby. Learning Probabilistic Automata with Variable Memory Length. In *COLT-94*, pp. 35-46, 1994.
- [6] 北内啓, 宇津呂武仁, 松本裕治. 誤り駆動型の素性選択による日本語形態素解析の確率モデル学習. *情報処理学会論文誌*, Vol. 40, No. 5, pp. 2325-2337, 5 1999.
- [1] M. Asahara and Y. Matsumoto. Extended Models and Tools for High-performance Part-