最大エントロピーモデルに基づく形態素解析と辞書による影響*

内元 清貴† 関根 聡‡ 井佐原 均†
†郵政省通信総合研究所 ‡ニューヨーク大学
{uchimoto|isahara}@crl.go.jp sekine@cs.nyu.edu

1 はじめに

形態素解析は日本語解析の重要な基本技術の一つとして認識されている。形態素解析の形態素とは、単語や接辞など、文法上、最小の単位となる要素のことであり、形態素解析とは、与えられた文を形態素の並びに分解し、それぞれの形態素に対し文法的属性 (品詞や活用など) を決定する処理のことである。

近年、形態素解析において重要な課題となっているのは、辞書に登録されていない、あるいは学習コーパスに現れないが形態素となり得る単語 (未知語) をどのように扱うかということである。この未知語の問題に対処するため、これまで大きく二つの方法がとられてきた。一つは未知語を自動獲得し辞書に登録する方法 (例えば [1] など) であり、もう一つは未知語でも解析できるようなモデルを作成する方法 (例えば [2, 3] など) である。ここで、前者の方法で獲得した単語を辞書に登録し、後者のモデルにその辞書を利用できるような仕組みを取り入れることができれば、両者の利点を生かすことができると考えられる。森らは n-gram モデルに外部辞書を追加する方法を提案している [4]。ある文字列が辞書に登録されている場合にその文字列が形態素となる確率を割り増しするような方法である。しかし、わずかな精度向上に留まっていることから、n-gram モデルでは辞書の情報を利用する仕組みを容易に組み込むのは難しいのではないかと考えられる。本論文では、最大エントロピー (ME) モデルに基づく形態素解析の手法を提案する。この手法では、辞書の情報を学習する機構を容易に組み込めるだけでなく、字種や字種変化などの情報を用いてコーパスから未知語の性質を学習することもできる。ここで辞書の情報とは、辞書に登録されている語が複数の品詞をとり得る場合にどの品詞を選択するべきかといった情報を意味する。京大コーパスを用いた実験では、再現率 95.44%、適合率 94.94% の精度が得られた。本論文では、辞書の情報を用いない場合、未知語の性質を学習しない場合についても実験し、それぞれの精度に及ぼす影響についても考察する。

2 形態素モデル

この章では形態素としての尤もらしさを計算するモデルについて述べる。我々はこのモデルを ME モデルとして実装した。

テストコーパスが与えられたとき、そのコーパスの各文を形態素解析するという問題は文を構成する各文字列に二つのタグのうち一つ、つまり、形態素であるかないかを示す「1」か「0」を割り当てる問題に置き換えることができる。さらに、形態素である場合には文法的属性を付与するために「1」を文法的属性の数だけ分割する。すると、文法的属性の数が n 個のとき、各文字列に「0」から「n」までのうちいずれかのタグを割り当てる問題に置き換えることになる。形態素解析の問題において、この $n+1$ 個のタグは ME モデルを定式化するときに「未来 (futures)」空間を形成する。ここで、未来空間とは学習モデルにおける分類先に対応する。ME モデルでは他の類似したモデルと同様に、可能性のある未来空間 F における任意の f と可能性のある履歴空間 H におけるすべての h に対して確率分布 $P(f|h)$ を計算することができる。ここで、ME モデルにおける「履歴 (history)」とは未来空間においてどこに分類するかという判断を下す根拠となるデータのことである。形態素解析の問題における確率分布は次の式で表すことができる。

$$P(f|h_t) = P(f|\text{テストコーパスから関係}t\text{に関して導出可能な情報})$$

これは、テストコーパスからある関係 t に関して導出可能な情報が得られたときに f の確率が求まることを示している。ME モデルにおける確率分布 $P(f|h)$ の計算は「素性 (features)」の集合、つまり、未来を予測する助けとなる情報に依存する。この情報は素性関数として定義され、近年の計算言語学の研究で用いられてきた他の多くの ME モデルと同様に我々のモデルでも、履歴と未来を引き数とし 0 か 1 を返す 2 値関数として定義する。以下にその一例をあげる。

$$g(h,f) = \begin{cases} 1 : \text{if has}(h,x) = \text{true}, \\ \qquad x = \text{“POS(0)(Major) : 動詞”} \\ \qquad \& \ f = 1 \\ 0 : \text{otherwise.} \end{cases} \quad (1)$$

ここで、「has(h, x)」は履歴 h に素性 x が観測されるときに真を返す 2 値関数である。我々の場合、素性としては辞書の情報†とともに、未知語の性質を学習できるように、着目している文字列の長さや文字種、その文字列が辞書にあるかどうか、連接する形態素の文法的属性、文字種の変化などを用いる。詳しくは 3 章で述べる。

素性集合と学習データが与えられたとき、エントロピーを最大にするという操作によりモデルが生成される。このモデルではすべての素性 g_i に対しパラメータ α_i が関係付けられ、モデルは次のような条件付き確率として表される [5]。

$$P(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\lambda(h)} \quad (2)$$

$$Z_\lambda(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)} \quad (3)$$

パラメータを推定する際には、学習コーパスにおけるすべての素性 g_i に対し、ME モデルから計算される g_i の

*Morphological Analysis Based on A Maximum Entropy Model and Influence of A Dictionary on Its Accuracy
Kiyotaka Uchimoto†, Satoshi Sekine‡, and Hitoshi Isahara†
†Communications Research Laboratory, M. P. T.
‡New York University

†今回の実験では既存の辞書の情報のみを用いたが、自動獲得した辞書の情報も利用可能であると考えている。

期待値が g_i の経験的期待値と等しくなること、つまり、以下の式が成り立つことを保証している。

$$\sum_{h,f} \tilde{P}(h,f) \cdot g_i(h,f) = \sum_{h} \tilde{P}(h) \cdot \sum_{f} P_{ME}(f|h) \cdot g_i(h,f) \quad (4)$$

ここで、$\tilde{P}$ は経験的確率分布であり、P_{ME} は ME モデルとして推定される確率分布である。

形態素に付与するべき文法的属性が n 個あると仮定する。文法的属性としては品詞と文節区切りを考える。品詞が m 個の場合、その各々についてその品詞を付与した形態素の左側が文節区切りであるかないかを考慮し、文法的属性の数は $n = 2 \times m$ とする。文字列が与えられたとき、その文字列が形態素であり、かつ $i(1 \leq i \leq n)$ 番目の文法的属性を持つとしたときの尤もらしさを確率値として求めるモデルを形態素モデルと呼ぶ。このモデルは式 (2) を用いて表される。ここで、f は 0 から n までの値をとる。

一文が与えられたとき、一文全体で確率の積が最大になるよう形態素に分割し文法的属性を付与する。最適解の探索にはビタビアルゴリズムを用いる。N-best 解の探索には文献 [6] の方法を用いる。

3 実験と考察

3.1 実験の条件

品詞体系は JUMAN[7] のものを仮定した。品詞は細分類まで分類すると全部で 53 種類ある。これに文節区切りを考慮すると推定するべき文法的属性の数は倍の 106 種類となる。活用型、活用形は品詞が決まれば表記からほぼ一意に決めることができるので、モデルから確率的に推定することはしない。したがって、式 (2) の f は 0 から 106 までの 107 個の値をとるものとする。

実験には、京大コーパス (Version 2) [8] を用いた。学習には 1 月 1 日と 1 月 3 日から 8 日までの 7 日分 (7,958 文)、試験には 1 月 9 日の 1 日分 (1,246 文) を用いた。

一文が与えられると、5 文字以下のすべての文字列および 5 文字を越えるが辞書に登録されている文字列に対して、その文字列が形態素であるかないか、形態素である場合にはその文法的属性が何かを推定する。5 文字以下のすべての文字列としたのは、5 文字を越えるような形態素は大抵、複合語あるいはカタカナ語であり、辞書に登録されていなければ、ほとんどの場合形態素ではないためである。複合語は辞書に登録されているもの以外は 5 文字以下の文字列に分割できると仮定する。また、カタカナ連続は辞書に登録されていない場合、ひとまとまりにして「未定義語 (大分類)、カタカナ (細分類)」という品詞を持つものとして辞書に登録されていたものとして扱う。ビタビアルゴリズムを用いて最適解を探索する際には、JUMAN で定義されている連接規則を満たさなければならないという制約を加えた。

2 章に述べたモデルでは、各文字列に対し品詞を付与する際、すべての品詞候補 (53 種類) のうち一文全体の確率を最大にするものが選ばれる。このとき、必ずしも辞書に記述されている品詞が選ばれるとは限らない。そこで、辞書に登録されている文字列については、その文字列に付与可能な品詞がすべて辞書に記述されていると仮定し、各文字列に対し品詞を付与する際には、辞書に記述されている品詞の中から選択するという制約を加える。このように制約を加えた場合を本手法 2 と呼び、制約を加えない場合を本手法 1 と呼ぶ。

次に、実験に用いた素性を表 1 にあげる。このうち、学習には学習コーパスで 3 回以上観測された素性 20,701 個を用いた。表 1 の素性名で「(0)」「(-1)」はそれぞれ、着目している文字列、その文字列の左に連接する一形態素を意味する。以下で、表 1 の各素性について説明する。

(文字列) 学習コーパスに形態素として現れた文字列のうち、頻度 5 以上のもの

(長さ) 文字列の長さ

(文字種) 文字の種類。「(頭)」「(末尾)」はそれぞれ文字列の先頭と末尾の文字を表す。文字列ではなく一文字の場合はともに同じ文字を指すものとする。「文字種 (0)(変化)」は先頭と末尾の文字の変化を表す。「文字種 (-1)(変化)」は左に連接する一形態素の末尾文字の文字種から着目している文字列の先頭文字の文字種への変化を表す。例えば、左に連接する一形態素が「先生」、着目している文字が「に」の場合、素性値は「漢字 → 平仮名」と表す。

(辞書) JUMAN の辞書を用いる。この辞書に登録されている異なり形態素数は約 20 万個である。Major、Minor はそれぞれ JUMAN の品詞大分類と細分類に対応する。Major&Minor は Major と Minor の可能な組み合わせである。着目している文字列が辞書に登録されている場合、辞書に記述されている品詞の情報を素性として利用する。複数の品詞を持つものとして登録されている場合にはそれぞれを素性として用いたときに形態素モデルから推定される確率が一文全体で最大となるものを採用する。その文字列が、連語辞書に登録されている形態素列の一番左の形態素の文字列である場合には、その文字列が連語の先頭の形態素であるという情報を付加したものを素性として利用する。この場合、素性値としては「連語」という表記が付加されているものを用いる。連語については文献 [9] に詳しい説明がある。

未知語の性質を学習するために、学習コーパスにおいて各文字列に対し辞書引きをしたときに一回しか引かれなかったものは辞書になかったものとして学習する。今回の実験ではそのような語の数は 20,317 個であった。ちなみに、辞書引きされた語の延べ数は 1,964,829 個、異なり語の総数は 60,908 個であった。このような学習方法をとることによって、辞書が充実すればその情報を反映できるとともに、辞書に依存し過ぎることなく未知語にも対処できると考えている。

(品詞) Major、Minor はそれぞれ JUMAN の品詞大分類と細分類に対応

(活用) Major、Minor はそれぞれ JUMAN の活用型、活用形に対応

(文節区切り) 形態素の左側に文節区切りがあるかないか

3.2 実験結果

形態素解析の結果を表 2 に示す。ここで、再現率はコーパス中の全形態素に対して区切りと品詞 (大分類のみ) を正しく推定できたものの割合を、適合率はシステムが推定した全形態素に対して区切りと品詞 (大分類のみ) を正しく推定できたものの割合を求めたものである。表中の F というのは F-measure のことで、以下の定義式により計算した。

$$\text{F} - \text{measure} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

表の各行にはそれぞれ、3.1 節で述べた本手法 1、本手

表 1: 学習に利用した素性

素性番号	素性名	素性値	削除した時の精度					
			本手法 1			本手法 2		
			再現率	適合率	F	再現率	適合率	F
1	文字列 (0)	(2,279 個)	85.44% (−5.75%)	86.74% (−4.72%)	86.08 (−5.24)	93.87% (−1.57%)	94.04% (−0.90%)	93.95 (−1.24)
2	文字列 (-1)	(2,279 個)						
3	辞書 (0)(Major)	動詞 動詞 & 連語 形容詞 形容詞 & 連語 . (28 個)	83.97% (−7.22%)	85.05% (−6.41%)	84.51 (−6.81)	94.06% (−1.38%)	92.22% (−2.72%)	93.13 (−2.06)
4	辞書 (0)(Minor)	普通名詞 普通名詞 & 連語 副助詞 (90 個)						
5	辞書 (0)(Major&Minor)	名詞 & 普通名詞 名詞 & 普通名詞 & 連語 . (103 個)						
6	長さ (0)	1 2 3 4 5 6 以上 (6 個)	89.29% (−1.90%)	89.67% (−1.79%)	89.48 (−1.84)	95.07% (−0.37%)	94.02% (−0.92%)	94.54 (−0.65)
7	長さ (-1)	1 2 3 4 5 6 以上 (6 個)						
8	文字種 (0)(頭)	漢字 平仮名 記号 数字 カタカナ アルファベット (6 個)	90.18% (−1.01%)	89.60% (−1.86%)	89.89 (−1.43)	94.23% (−1.21%)	93.23% (−1.71%)	93.73 (−1.46)
9	文字種 (0)(末尾)	漢字 平仮名 記号 数字 カタカナ アルファベット (6 個)						
10	文字種 (0)(変化)	漢字 → 平仮名 数字 → 漢字 カタカナ → 漢字 . (30 個)						
11	文字種 (-1)(末尾)	漢字 平仮名 記号 数字 カタカナ アルファベット						
12	文字種 (-1)(変化)	漢字 → 平仮名 数字 → 漢字 カタカナ → 漢字 (30 個)						
13	品詞 (-1)(Major)	動詞 形容詞 名詞 助動詞 接続詞 未定義語 . (15 個)	90.20% (−0.99%)	91.93% (+0.47%)	91.06 (−0.26)	95.14% (−0.30%)	95.22% (+0.28%)	95.18 (−0.01)
14	品詞 (-1)(Minor)	普通名詞 サ変名詞 数詞 程度副詞 .. (45 個)						
15	品詞 (-1)(Major&Minor)	無 名詞 & 普通名詞 名詞 & 普通名詞 & 連語 (54 個)						
16	活用 (-1)(Major)	母音動詞 子音動詞カ行 ... (33 個)	90.70% (−0.49%)	91.07% (−0.39%)	90.89 (−0.33)	95.28% (−0.16%)	94.83% (−0.11%)	95.05 (−0.14)
17	活用 (-1)(Minor)	語幹 基本形 未然形 意志形 命令形 ... (60 個)						
18	文節区切り (-1)	無 有 (2 個)	90.83% (−0.36%)	91.59% (+0.13%)	91.21 (−0.11)	95.51% (+0.07%)	95.16% (+0.22%)	95.33 (+0.14)
19	文節区切り (-1) & 品詞 (-1)(Major&Minor)	名詞 & 普通名詞 & 区切り 名詞 & 普通名詞 & 区切りではない .. (106 個)						

法 2 および JUMAN による精度をあげた。JUMAN は単独では辞書に登録されていないカタカナ語に対し「未定義語」という品詞を付与するため、それによる誤りが多くなる。ルールベースの構文解析システム KNP[10] は、JUMAN に複数解の出力を許しその出力を入力とすると、構文解析の過程で品詞の曖昧性を解消し、未定義語も何らかの品詞に置き換えることができる。そこで、JUMAN と KNP で解析した結果も評価した。表には +KNP と表記した。

表 2: 解析結果 (形態素区切りと品詞大分類)

	再現率	適合率	F
本手法 1	91.19% (28,543/31,302)	91.46% (29,500/31,209)	91.32
本手法 2	95.44% (29,875/31,302)	94.94% (29,875/31,467)	95.19
JUMAN	95.25% (29,814/31,302)	94.90% (29,814/31,417)	95.07
+KNP	98.49% (30,830/31,302)	98.13% (30,830/31,417)	98.31

制約を加えない本手法 1 は制約を加えた本手法 2 に比べると 4% 程度精度が低い。これは 53 種類という多くの品詞候補のなかから適切なものを選ぶには学習データがスパースであったためと考えられる。本手法 1 が本手法 2 と同程度の精度を得るためには、もっと多くの学習コーパスが必要であると考えている。

文節認定の精度は手法 1 で再現率 76.04%、適合率 67.84%、手法 2 で再現率 76.97%、適合率 65.19% であった。ここで、再現率はコーパス中の全文節に対して区切りを正しく推定できたものの割合を、適合率はシステムが推定した全文節に対して区切りを正しく推定できたものの割合をそれぞれ求めたものである。この結果は、形態素解析の誤りが影響しているとは言え、良い結果とは言えない。先行研究 [11] では、左に連接する一形態素だけでなく二形態素以上の情報を用いることにより、正しい形態素列を入力とした場合に 99% 程度の文節認定の精度を得ている。文節を高精度で認定するには連接する二形態素以上の情報を用いて学習する必要がありそうである。

3.3 辞書と未知語

辞書の情報、未知語の性質は、我々が実験で用いた素性に反映されている。表 1 にあげた素性のうち、「文字列」「辞書」の素性が辞書の情報を‡、「長さ」「文字種」の素性が未知語の性質を反映する。表 1 の右欄には、それぞれの素性を削除したときの解析精度と削除したことによる精度の増減を示した。ほとんどの素性が精度向上に貢献しており、特に辞書情報の貢献度が高いことが分かる。

逆に辞書が解析結果に悪い影響を及ぼす例もある。例えば、「／海／に／かけた／ロマンは／、／」「／荒波／に／負け／ない心／と／」(「／」は形態素区切り) といった形態素区切りが出力として得られることがある。これは、漢字を使った表記「ロマン派」「内心」に加えて平仮名を使った表記「ロマンは」と「ない心」も名詞として辞書に登録されていたために生じた誤りである。このような間違いをなくすためには、不自然な表記を辞書に登録しないようにする、あるいは、辞書の表記に使われる文字種の性質を学習する必要がある。

学習の際、一回しか辞書引きされなかった語は辞書に登録されていなかったものとして扱った。このようにしたのは、テストコーパスを解析するときには未知語が多くなると予想されるため、学習の際にもそれと同じ状況に少しでも近付けようとしたためである。ところが、実験後、学習コーパス、テストコーパスにおける未知語の割合を調べたところ、辞書に登録されていなかった語の数 (見出し語の異なり数) の異なり形態素数に対する割合は、学習コーパスで 26.6%(3,859/14,493)、テストコーパスで 17.7%(901/5,093) であり、テストコーパスにおける未知語の割合の方が学習コーパスにおける割合より少ないことが分かった。ちなみに、未知語の大部分は数詞およびカタカナで表記された名詞が占めていた。そこで、辞書に登録されていた場合には辞書引きの頻度に関わりなくその情報をすべて学習に用いることにすると、精度は、本手法 1 で再現率 91.34%、適合率 91.80%、F-measure 91.57、本手法 2 で再現率 95.32%、適合率 95.12%、F-measure 95.22 となった。これは表 1 にあげた精度よりわずかに良い結果である。今回の実験では学習コーパスより未知語の割合が少ないコーパスに対して実験したためこのような結果となったが、本手法を学習コーパスよりも未知語の割合が多い分野に適用するときには我々がとった学習手法は有効ではないかと考えている。その有効性を調べることは今後の課題である。

‡「文字列」は学習コーパスに 5 回以上出現した形態素の文字列であり、これを素性として用いることは、学習コーパスから辞書的な情報を得て利用していることに相当する。

3.4 JUMANとの比較

JUMANはルールベースのシステムであり、形態素に品詞を付与するときにかかるコスト(品詞コスト)と形態素を連接するときにかかるコスト(連接コスト)の和が一文全体で最小となるように形態素区切りと品詞を決める。それぞれのコストは予め人手により設定する必要がある。一方、我々の手法は学習に基づくシステムであり、JUMANの品詞コストと連接コストに相当するものを一つの確率値として表し、その確率値を計算するためのモデルをコーパスから統計的に学習する。大きな違いは、ルールベースと統計ベースという点だけでなく、JUMANが未知語を一文字からなる名詞と既知語に分割して出力するのに対し、我々の手法は、未知語に対しても前後の形態素のつながりから形態素と認定でき、適切な品詞を付与することができる点にもある。例えば、「漱石」や「露伴」はJUMANの辞書には登録されていないため、JUMAN+KNPでは「漱(名詞)石(名詞)」「露(副詞)伴(名詞)」のように解析されるのに対し、我々のシステムではどちらも正しく名詞であると解析される。この場合は、細分類も正しく人名であると解析できた。このような固有名詞などは未知語になることが多い。そこで、未知語(辞書にも素性にもなかった語)に対する再現率を調査した。結果を表3にあげる。表には品詞細分類まで正しい場合に正解とするという基準で求めた再現率もあげた。この基準で求めた我々の手法の精度はJUMAN+KNPに比べて10%以上良かった。この結果は我々のモデルでは未知語、特に固有名詞や人名、組織名、地名に関する語に対する学習が比較的にできていることを示していると考えて良いだろう。

表3: 未知語に対する精度(再現率)

	形態素区切りと品詞大分類	形態素区切りと品詞細分類
本手法1	80.53% (877/1,089)	37.10% (404/1,089)
本手法2	83.56% (910/1,089)	43.34% (472/1,089)
JUMAN+KNP	86.87% (946/1,089)	29.94% (326/1,089)

表2、3にあげた形態素区切りと品詞大分類に対する推定精度は、我々の手法ではJUMAN+KNPよりも3%程度低かった。その原因として学習コーパスの量、素性、コーパスにおける形態素の揺れなどが考えられる。今回用いた学習コーパスは約8,000文と少なく、素性については文献[12]などで用いられているような組み合せの素性に相当するものはあまり用いていない。利用可能なマシンのメモリ容量の都合上、今回は学習コーパスの量、素性の数ともにこれ以上増やすのは困難であったが、いずれ可能になるだろう。次に形態素の揺れについてであるが、これは実験に用いた京大コーパスがJUMAN+KNPの解析結果を人手で修正したものであるということに起因していると思われる。このことはJUMAN+KNPの出力の評価に有利に働いている。例えば、最後が「者」で終わる形態素はテストコーパス中に153個あり、すべてJUMAN+KNPの出力と同じであった。このうち我々のシステムの誤りは3個(約2%)であった。コーパスには「生産(名詞)者(接尾辞)」と「消費者(名詞)」の違いなどの揺れがあり、このように区切りに一貫性のない場合、過学習にならないように学習するのは難しい。揺れに関してはコーパス全体を通して他にも同様な例がいくつかある。例えば、「芸術家(名詞)」と「工芸(名詞)家(接尾辞)」、「警視庁(名詞)」と「検察(名詞)庁(名詞)」、「現実的(形容詞)」と「理想(名詞)的(接尾辞)」などがそうである。この揺れの問題を解決するためには、コーパス修正の研究がより活発に行なわれる必要がある。一つの方法として、我々のモデルを用いる方法が考えられる。学習したモデルを用いて学習コーパス中の各形態素の確率を再推定し、確率の低い部分に一貫性を欠いたものがある可能性が高いと推測する方法である。今後、この方法を試してみたい。

4 まとめ

本論文では次の二つの特徴をもつモデルをMEモデルとして実装した形態素解析の手法を提案した。(1)学習コーパスからだけでなく辞書から得られる情報も用いる。(2)形態素となる文字列だけでなく形態素とはならない文字列の性質も学習することによって、未知語も形態素として推定でき、同時にその文法的属性も推定できる。実験により、辞書の精度に及ぼす影響の大きさ、および、我々の手法が、固有名詞、人名、組織名、地名など未知語になりやすいものに対して比較的に推定精度がよいことが分かった。

今後の課題としては以下の三点をあげておきたい。(1)学習に用いる情報について。一つ前の形態素の情報だけでなく、二つから四つくらい前の形態素の情報を利用するとともに、組み合わせの素性を増やす。(2)コーパスについて。コーパスの量を増やすとともに、コーパス修正の研究を活発に進める。また、異なるコーパスについても実験する。(3)辞書について。文法体系が変わったときにその体系に合うように辞書情報を変換する技術を開発する。

謝辞

本研究の評価にあたり、評価ツールを提供して下さった京都大学の黒橋禎夫講師に心から感謝の意を表する。

参考文献

[1] Shinsuke Mori and Makoto Nagao. Word extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*, pp. 1119–1122, 1996.

[2] 柏岡秀紀, Stephen G. Eubank, Ezra W Black. 確率付決定木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, pp. 433–436, 1997.

[3] Masaaki Nagata. A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 277–284, 1999.

[4] 森信介, 長尾眞. 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. 75–103, 1998.

[5] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

[6] Masaaki Nagata. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING94)*, pp. 201–207, 1994.

[7] 黒橋禎夫, 長尾眞. 日本語形態素解析システム JUMAN 使用説明書 Version 3.61. 京都大学大学院情報学研究科, 1999.

[8] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第3回 年次大会, pp. 115–118, 1997.

[9] 山地治, 黒橋禎夫, 長尾眞. 連語登録による形態素解析システム JUMAN の精度向上. 言語処理学会 第2回 年次大会, pp. 73–76, 1996.

[10] 黒橋禎夫. 日本語構文解析システム KNP 使用説明書 Version 2.0b6. 京都大学大学院情報学研究科, 1998.

[11] 村田真樹, 内元清貴, 馬青, 井佐原均. 排反な規則を用いた文節まとめあげ. 情報処理学会論文誌, Vol. 41, No. 1, 2000.

[12] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 196–203, 1999.