日本語話し言葉コーパスの形態素解析*

内元 清貴† 野畑 周† 山田 篤† 関根 聡‡ 井佐原 均†

†独立行政法人通信総合研究所 ‡ニューヨーク大学

{uchimoto,nova,ark,isahara}@crl.go.jp sekine@cs.nyu.edu

1 はじめに

開放的融合研究「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトでは主に講演などのモノローグを対象とした自発的な話し言葉の大規模コーパス、*Corpus of Spontaneous Japanese (CSJ)* [1]を作成している。このコーパスには音声データだけでなく書き起こしも含まれ、書き起こしには、国立国語研究所で定義された短い単位と長い単位の二種類の形態素に関する情報が付与される予定である。短い単位は**短単位**と呼ばれ、その定義は一般的な辞書の見出しに近い。長い単位は**長単位**と呼ばれ、その定義には様々な複合語が含まれる。二つの単位の違いは長さと品詞体系であり、長単位が短単位を包含するように定義されている。コーパス中のすべての短単位が特定されたと仮定すると全体の大きさは延べ約700万形態素になる。現在までに、その約1/10が人手で特定され、品詞や活用型、活用形などの情報が付与された。現在、その約1/10における形態素区切りと品詞情報の精度は短単位、長単位についてそれぞれ99.9%、97%を超えている。精度をこのレベルまで良くするのに二年以上かかっているため、残りのコーパスに対しても人手でしかも同程度の精度で情報を付与しようとすると20年近くかかってしまう。したがって、残りの約9/10については、自動あるいは半自動で情報を付与する必要がある。

本稿では、短い単位、長い単位の二種類の形態素に関して、それらの区切りと品詞情報を特定する方法、および、大規模な話し言葉コーパスCSJに精度良く形態素の情報を付与するための方法について述べる。以降で、CSJにおける二種類の形態素をそれぞれ短単位と長単位、あるいは総称して形態素と呼ぶ。また、与えられた文を形態素列に分割し、各々の形態素に品詞などの文法的属性を付与する処理のことを一般に呼ばれているように**形態素解析**と呼ぶ。

2 問題点とその解決策

1節で述べたように、CSJのすべてを人手で形態素解析するのは困難である。したがって、半自動的に形態素解析を行なう。この節では、大規模な話し言葉コーパスを精度良く形態素解析するにあたっての問題点とその解決策について述べる。

一般に形態素解析においては、未知語つまり辞書にも学習コーパスにも現れない形態素の存在が最も問題となる。この問題に対処するために、これまで大きく二つの方法がとられてきた。ひとつは未知語を自動獲得し辞書に登録する方法(例えば[2]など)であり、もうひと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[3, 4]など)である。内元らは両者の利点を生かした、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した[5]。この手法で用いられるモデルは、任意の文字列について、その文字列が形態素であるときの尤もらしさを確率値として推定することができるため、未知語の問題を解決できる可能性が高い。したがって、CSJの形態素解析にもこの手法を用いることにした。しかし、CSJのうち、すでに人手で形態素解析済みの約1/10を用いて、形態素の区切りと品詞の自動推定精度を調べたところ、F-measureで94程度に留まることが分かった[6]。その原因としてはいくつか考えられる。以下に、その原因と、コーパス全体の精度を良くするために我々がとった対処方法について述べる。

- フィラーや言いよどみの存在

  話し言葉に特有な現象であるフィラーや言いよどみは任意の位置に出現する可能性があるため特定するのが難しい。CSJではフィラーや言いよどみには人手でタグが付与されているため、これらを削除して形態素解析し、後で挿入することにした。

- 未知語に対する精度

  我々が採用したモデルは、未知語に対しても形態素の区切りと品詞を推定することができる。しかし、その精度は既知語に対する精度に比べると低い。既存の辞書を用いて未知語を減らしても、精度向上はわずかであった[6]。その理由は、区切りや品詞などの定義が単語ごとに異なるためであると考えられる。したがって、同じ定義の辞書エントリを増やす必要がある。そこで、3節に述べるモデルにより、辞書には登録されていなかったが形態素と判断されたもの、および、モデルにより推定された確率値が小さいものを人手でチェックして辞書に追加することにした。未知語の自動推定精度や未知語を登録することによりどの程度精度向上が見込めるかについては4節で述べる。

- 素性の不足

  現在のモデルではひとつ前に連接する形態素の情報を考慮しているが、より良いモデルとするためには、さらに前に連接する形態素の情報も考慮するべきである。しかし、多くの情報を考慮しすぎると過学習に陥ったり、学習のためのデータが大きくなり過ぎて学習が困難になることがある。限られた時間内にコーパスの精度を良くするためには、可能な限りモデルを改良するとともに、自動解析結果の誤りを人手でチェックするのが最善であると考えられる。そこで、モデルにより推定された確率値が小さいほど誤りの可能性が高いと考え、確率値の小さいものから順に人手修正することによりコーパス全体の精度を高めることにした。これにより、どの程度精度向上が見込めるかについては4節で述べる。

- 読み

  音声認識のための言語モデルを作成するためには、形態素に関する情報のひとつとして読みの情報が欠かせない。しかし、辞書情報を用いて読みの情報を補うのは無理がある。そこで、CSJの読みのフィールドと形態素解析の結果との対応をとることにより実際の読みを付与する。

---

*Morphological Analysis of The Corpus of Spontaneous Japanese
Kiyotaka Uchimoto†, Chikashi Nobata†, Atsushi Yamada†, Satoshi Sekine‡, and Hitoshi Isahara†
†Communications Research Laboratory
‡New York University

3 モデルとアルゴリズム

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ひとつ、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られる。

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。我々は、このモデルを ME モデルにより実装した。このモデルは式 (1) を用いて表わされる。

$$p_\lambda(a|b) = \frac{\exp\left(\sum_{i,j} \lambda_{i,j} g_{i,j}(a,b)\right)}{Z_\lambda(b)} \quad (1)$$

$$Z_\lambda(b) = \sum_a \exp\left(\sum_{i,j} \lambda_{i,j} g_{i,j}(a,b)\right) \quad (2)$$

ここで、a は、「future」と呼ばれ、クラス分類問題におけるカテゴリを表わす。a は 0 から n までの $n+1$ 個の値をとる。b は、「history」と呼ばれ、どの「future」を選択するかを決定するための文脈的あるいは条件的な情報を表わす。$Z_\lambda(b)$ は正規化定数で、すべての b に対し $\sum_a p_\lambda(a|b) = 1$ を満たすように定められる。ME モデルにおいて、条件付き確率分布 $p_\lambda(a|b)$ の計算は**素性**の集合に依存する。この素性は式 (1) では素性関数 $g_{i,j}(a,b)$ として表わされる。これは、a と b を引き数とし 0 か 1 を返す 2 値関数として定義される。以下にその一例をあげる。

$$g_{i,j}(a,b) = \begin{cases} 1: \text{ if } has(b,f_j) = 1 \ \&\ a = a_i \\ \qquad f_j = \text{“POS}(-1)\text{(Major) : 動詞,”} \\ 0: \text{ otherwise.} \end{cases} \quad (3)$$

ここで、「$has(b,f_j)$」は b に素性 f_j が観測されるときに 1 を返す 2 値関数である。実験に用いた素性については、4 節で詳しく述べる。

一文が与えられたとき、その文中の任意の長さのすべての文字列に対し、式 (1) により、1 から n までの n 個の文法的属性に対する確率値を計算する。与えられた文を分割して得られる形態素列の候補 (各形態素には文法的属性がひとつずつ付与されている) のうち、一文全体での確率値の積が最大になるようなものを最適解とする。最適解の探索にはビタビアルゴリズムを用いる。

4 実験と考察

4.1 実験の条件

実験には、CSJ のうち人手による形態素解析が済んでいる 338 講演 (全体の約 1/10) を学習とテストに分けて用いた。学習には、319 講演 (短単位で 744,204 語、長単位で 618,538 語。ただしフィラーと言いよどみを除く。) を、テストには、19 講演 (短単位で 63,037 語、長単位で 51,796 語。ただしフィラーと言いよどみを除く。) を用いた。

書き起こしは図 1 のように基本形と発音形の部分からなる。ここで、発音形は発話者の発声を忠実に書き起こしたものであり、基本形はその仮名漢字混じり表記である。数字から始まる行はタイムスタンプを表わし、次のタイムスタンプまでの文字列が発表開始後何秒から何秒の間に発話されたかを示している。タイムスタンプ以外

基本形	発音形
0017 00051.425-00052.869 L:	
(F えー)	(F エー)
形態素解析	ケータイソカイセキ
0018 00053.073-00054.503 L:	
について	ニツイテ
0019 00054.707-00056.341 L:	
お話しいたします	オハナシイタシマス

図 1: 書き起こしの例

の各行は文節からなる。実験では発音形は用いず基本形のみを用いた。ここには括弧付きでラベルが付与されている部分がある。これには、表 1 にあげたようなタイプの違いがあり、それぞれ表の右欄のような規則で置き換えて用いた。

表 1: ラベルのタイプと整形規則

ラベルのタイプ	例	整形規則
フィラー、感情表出系感動詞	(F あの)	全部削除
言い直し	(D こ) これ、これ (D2 は) が	全部削除
聞きとり、語彙同定、漢字表記に自信なし	(? タオングー)	候補を残す
全く分からない	(?)	削除
複数候補あり	(? あのー, あんのー)	前の候補を残す
音や言葉に関する引用	(M わ) は (M は) と表記	候補を残す
外国語や古語、方言など	(O ザッツファイン)	候補を残す
個人名、差別語、誹謗中傷、など	○○研の (R △△) さんが	候補を残す
基本形で漢字仮名以外の文字を使用する場合	(A イーユー;EU)	前者の候補のみ残す
何らかの原因で漢字表記できなくなった場合	(K い (F んー) ずみ; 泉)	後者の候補のみ残す

CSJ を含めて話し言葉には明確な文の境界が存在しないため、500ms 以上のポーズがある位置を文境界とした。また、ショートポーズ (50ms 以上 200ms 未満のポーズ) で末尾が文末形式の場合も文境界とした。

実験では、フィラーと言いよどみを削除したが、フィラーと言いよどみがあった位置の情報は素性として用いた。また、文節境界や、(F) や (D) 以外のタグの情報も素性として用いた。したがって、システムの入力は、フィラーと言いよどみを除いた文字列に各種の境界の情報、タグの情報が付与されたものである。形態素は文節境界をまたぐことはないため、文節を越える文字列は最適解探索の範囲から除外した。システムの出力は図 2 に示すような文法的属性が付与された形態素列である。文法的属性としては、CSJ の品詞体系に基づく品詞を用いた。短単位で 14 品詞、長単位で 15 品詞定義されている。

次に、実験に用いた素性を表 2 にあげる。ここで素性とは、各素性名に対し素性値を展開したもののことである。各々の素性は式 (1) の素性関数 $g_{i,j}(a,b)$ の j に対応する。素性の番号は便宜上設けたものであり、各素性名に対応している。表 2 で素性名に使われている「(0)」「(-1)」という表記はそれぞれ、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。素性関数としては、素性と future の組が学習コーパスで 3 回以上観測されたもののみを用いた。以下で、表 2 の各素性名、素性値について説明する。

文字列: 学習コーパスに形態素として現れた文字列。

部分文字列: 学習コーパスに形態素として現れた文字列の、先頭の 1 文字と 2 文字 (それぞれ (左 1)(左 2) と表記)、末尾の 1 文字と 2 文字 (それぞれ (右 1)(右 2) と表記)。

辞書: 辞書情報。辞書情報は、学習データからフィラーと言いよどみを除くすべての形態素を抽出し、辞書に登録した。カタカナ連続について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」という品詞と「カタカナ」

短単位				長単位			
出現形	読み	品詞	その他の情報	出現形	読み	品詞	その他の情報
形態	ケータイ	名詞		形態素解析	ケータイソカイセキ	名詞	
素	ソ	接尾辞					
解析	カイセキ	名詞					
に	ニ	助詞	格助詞	について	ニツイテ	助詞	格助詞 連語
つい	ツイ	動詞	カ行五段 連用形 イ音便				
て	テ	助詞	接続助詞				
お	オ	接頭辞		お話しいたし	オハナシシタシ	動詞	サ行五段 連用形
話し	ハナシ	動詞	サ行五段 連用形				
いたし	イタシ	動詞	サ行五段 連用形				
ます	マス	助動詞	終止形	ます	マス	助動詞	終止形

図 2: 形態素解析結果の例

という情報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。Major、Minor はそれぞれ CSJ の品詞と活用形なのどのその他の情報に対応する。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やその他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。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。

品詞: CSJ の品詞を表わす。

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わす。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わす。「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わす。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」と表わす。

境界: 文節境界、タグが付与されている位置を表わす。表で、「(始)」、「(終)」は、着目している文字列の右側、左側がそれぞれ境界であることを表わす。

組: 素性の組み合わせを表わす。

4.2 実験結果と考察

実験結果を表 3 と表 4 に示す。これらの表で OOV は未知語率、つまり、コーパスにも辞書にも存在しなかった形態素の割合を表わす。表 4 では、形態素の代わりに形態素とその品詞のペアを未知語率の計算に用いた。再現率はコーパス中の全形態素に対して区切りと品詞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 3 と表 4 から、未知語がない場合はかなり精度が良くなることが分かる。特に、長単位に関しては現時点でのコーパスの精度に近い。これは、未知語をなくすことができれば、コーパス全体に対してかなり高い精度で自動的に形態素解析ができることを示している。未知語を登録することにより期待できる精度向上は、F-measure で、形態素の区切りについては短単位で約 1.5、長単位で約 2.5、形態素の区切りと品詞については短単位で約 2、長単位で約 3 ポイントである。

表 2: 素性

番号	素性名	素性値 (短単位: 長単位)(個)
1	文字列 (0)	(113,474:117,002)
2	文字列 (-1)	(17,064:32,037)
3	部分文字列 (0)(左 1)	(2,351:2,375)
4	部分文字列 (0)(右 1)	(2,148:2,171)
5	部分文字列 (0)(左 2)	(30,684:31,456)
6	部分文字列 (0)(右 2)	(25,442:25,541)
7	部分文字列 (-1)(左 1)	(2,160:2,088)
8	部分文字列 (-1)(右 1)	(1,820:1,675)
9	部分文字列 (-1)(左 2)	(11,025:12,875)
10	部分文字列 (-1)(右 2)	(10,439:13,364)
11	辞書 (0)(Major)	名詞, 動詞, 形容詞, ... 未定義語 (15:16)
12	辞書 (0)(Minor)	普通名詞, 副助詞, 基本形 ... (75:71)
13	辞書 (0)(Major&Minor)	名詞 & 普通名詞, 動詞 & 基本形, ... (246:227)
14	辞書 (-1)(Minor)	普通名詞, 副助詞, 基本形 ... (16:16)
15	品詞 (-1)	名詞, 動詞, 形容詞, ... (14:15)
16	長さ (0)	1, 2, 3, 4, 5, 6_or_more (6:6)
17	長さ (-1)	1, 2, 3, 4, 5, 6_or_more (6:6)
18	文字種 (0)(頭)	漢字, 平仮名, 数字, カタカナ, アルファベット (5:5)
19	文字種 (0)(末尾)	漢字, 平仮名, 数字, カタカナ, アルファベット (5:5)
20	文字種 (0)(変化)	漢字 → 平仮名, 数字 → 漢字, カタカナ → 漢字, ... (25:25)
21	文字種 (-1)(末尾)	漢字, 平仮名, 数字, カタカナ, アルファベット (5:5)
22	文字種 (-1)(変化)	漢字 → 平仮名, 数字 → 漢字, カタカナ → 漢字, ... (16:15)
23	境界	文節 (始), 文節 (終), タグ (始), タグ (終), (4:4)
24	組 (1,15)	(74,602:59,140)
25	組 (1,2,15)	(141,976:136,334)
26	組 (1,13,15)	(78,821:61,813)
27	組 (1,2,13,15)	(156,187:141,442)
28	組 (11,15)	(209:230)
29	組 (12,15)	(733:682)
30	組 (13,15)	(1,549:1,397)
31	組 (12,14)	(730:675)

表 3: 形態素区切りの精度

形態素	再現率	適合率	F	OOV
短単位	97.47% ($\frac{61,444}{63,037}$)	97.62% ($\frac{61,444}{62,945}$)	97.54	1.66%
	99.23% ($\frac{62,553}{63,037}$)	99.11% ($\frac{62,553}{63,114}$)	99.17	0%
長単位	96.72% ($\frac{50,095}{51,796}$)	95.70% ($\frac{50,095}{52,346}$)	96.21	5.81%
	99.05% ($\frac{51,306}{51,796}$)	98.58% ($\frac{51,306}{52,047}$)	98.81	0%

表 4: 形態素区切りと品詞付与の精度

形態素	再現率	適合率	F	OOV
短単位	95.72% ($\frac{60,341}{63,037}$)	95.86% ($\frac{60,341}{62,945}$)	95.79	2.64%
	97.57% ($\frac{61,505}{63,037}$)	97.45% ($\frac{61,505}{63,114}$)	97.51	0%
長単位	94.71% ($\frac{49,058}{51,796}$)	93.72% ($\frac{49,058}{52,346}$)	94.21	6.93%
	97.30% ($\frac{50,396}{51,796}$)	96.83% ($\frac{50,396}{52,047}$)	97.06	0%

次に、未知語がある場合について考察する。長単位の未知語率は短単位の未知語率に比べて4%以上高かった。一般に未知語率が高いほど解析が難しくなると考えられるが、短単位と長単位の精度の差は、再現率で約1%、適合率で約2%であり、未知語率の差に比べると小さい。この結果は、我々の用いた形態素モデルにより未知語に対しても精度良く推定できていることを示している可能性が高い。そこで、テストコーパスにおける未知語に対する再現率を調べたところ、短単位、長単位に対し、区切りのみが一致している場合で、それぞれ、55.7%(928/1,667)と74.1%(2,660/3,590)、区切りと品詞が一致している場合で、それぞれ、47.5%(791/1,667)と67.3%(2,415/3,590)であることが分かった。短単位よりも長単位の方が20%程度良いのは、複合名詞に対し、長単位ではひとつの名詞として、短単位では文脈に依存して複数の形態素として定義されることが多く、短単位の定義の方が難しいからであると考えられる。さらに、システムの出力のうち、辞書に登録されていなかったものを抽出し、テストコーパスにおける未知語と一致あるいは部分的に一致しているものの割合を調べたところ、短単位、長単位に対し、それぞれ、77.3%(1,289/1,667)、80.5%(2,889/3,590)であることが分かった。この方法で検出できなかった約20%のほとんどは複合語である。これについては、モデルにより推定された確率値が低いものから人手でチェックすることにより検出できると考えている。

未知語に対する再現率は既知語に比べると低い。また、自動形態素解析の精度はまだ、人手による形態素解析済みのコーパスに比べて低い。そこで、モデルにより推定された確率値が小さいほど誤りの可能性が高いと考え、確率値の小さいものから順に人手修正することにより、どの程度コーパス全体の精度向上が期待できるかを調べた。まず、モデルにより推定された確率値に閾値を設けて変化させ、システムの出力のうち、閾値を越える形態素の割合とその形態素の適合率との関係を調べた。結果を図 3 にあげる。この図で、

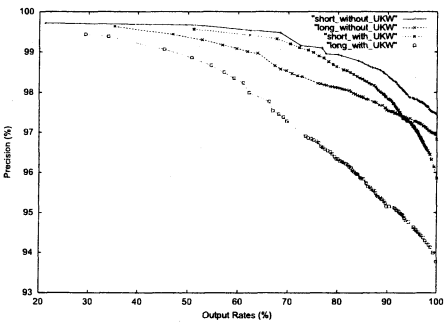

図 3: 部分解析の精度

「short_without_UKW」、「long_without_UKW」、「short_with_UKW」、「long_with_UKW」はそれぞれ、未知語がなかったと仮定したときの、短単位、長単位の適合率、未知語があるときの、短単位、長単位の適合率を表わす。横軸で右側にいくほど確率値の低い形態素がより多く含まれる．いずれのグラフも右下がりであり、確率値の小さいものから順にチェックすると効率良く修正できできそうなことが分かる。

次に、チェックをする割合と修正後の精度との関係を調べた。結果を図 4にあげる。精度はいずれも形態素区切りと品詞付与の精度である。上述した方法で未

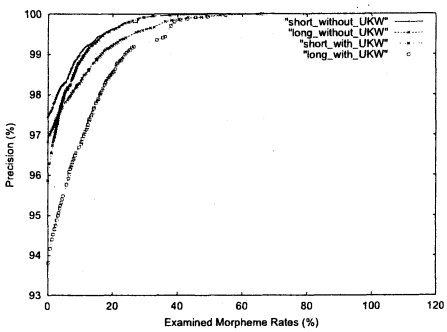

図 4: チェックをする割合と修正後の精度との関係 (前後の形態素もチェックした場合)

知語が登録できれば、グラフは短単位、長単位それぞれ、「short_without_UKW」と「short_with_UKW」の間、「long_without_UKW」と「long_with_UKW」の間になると予想される。これらの図から確率値の低い形態素から順にコーパス全体の約10%をチェックした場合、前後の形態素も合わせてチェックできたと仮定すると、全体の精度は短単位で99%、長単位で97%を越えると期待できる。

最後に、チェックをする割合と修正効率との関係を調べたところ、修正開始直後はチェック対象の約50%に誤りが発見でき、コーパス全体の10%をチェックしたところでも約20%の割合で誤りが発見できることが分かった。未知語を登録すると発見できる誤りの割合は減るが、常に10%以上は誤りが発見できると期待できる。

5 Conclusion

本稿では、短い単位、長い単位の二種類の形態素に関して、それらの区切りと品詞情報を特定する方法、および、大規模な話し言葉コーパス CSJ に精度良く形態素の情報を付与するための方法について述べた。この方法により、約80%の未知語が検出できることが分かった。この未知語を辞書に登録し、さらに、モデルにより推定された確率値の低い形態素から順にコーパス全体の約10%を人手でチェックした場合、全体として短単位で99%、長単位で97%を越える精度が期待できることが分かった。

参考文献

[1] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous Speech Corpus of Japanese. In *Proceedings of LREC2000*, pp. 947–952, 2000.

[2] S. Mori and M. Nagao. Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. In *Proceedings of COLING'96*, pp. 1119–1122, 1996.

[3] H. Kashioka, S. G. Eubank, and E. W. Black. Decision-Tree Morphological Analysis without a Dictionary for Japanese. In *Proceedings of NLPRS'97*, pp. 541–544, 1997.

[4] M. Nagata. A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context. In *Proceedings of ACL'99*, pp. 277–284, 1999.

[5] K. Uchimoto, S. Sekine, and H. Isahara. The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary. In *Proceedings of EMNLP2001*, pp. 91–99, 2001.

[6] K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara. Morphological Analysis of The Spontaneous Speech Corpus. In *Proceedings of COLING2002*, pp. 1298–1302, 2002.

[7] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physcal Review*, Vol. 106, pp. 620–630, 1957.

[8] E. T. Jaynes. Where do we Stand on Maximum Entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, p. 15. M. I. T. Press, 1979.

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