単語レベルと文字レベルの情報を用いた中国語・日本語単語分割

中川哲治[†,††] 松本裕治[††]

本稿では，中国語と日本語の単語分割を行うために，コスト最小法と文字タグ付け法を組み合わせた単語分割手法を提案する．単語分割に関してこれまでに多くの研究が行われているが，単語単位で処理を行うコスト最小法は未知語の扱いが困難であり，文字単位で処理を行う文字タグ付け法は既知語に対する解析精度が低い傾向が見られる．そこで，2つの手法を組み合わせることでこれらの問題を解決することを試みる．提案手法では品詞タグと文字位置タグを同等に扱うことにより，単語単位の解候補と文字単位の解候補を統一的に扱い，既知語と未知語を同時に処理する．複数のコーパスを使用して中国語と日本語の単語分割実験を行った結果，高い解析精度が得られることを確認した．

Chinese and Japanese Word Segmentation Using Word-level and Character-level Information

Tetsuji Nakagawa[†,††] and Yuji Matsumoto[††]

In this paper, we propose a hybrid method for Chinese and Japanese word segmentation, which combines a Markov model-based method and a character tagging method. While word-based Markov models have difficulties in handling unknown words, character-based tagging methods perform worse in handling known words compared with other methods. To compensate the weaknesses of these approaches, we propose a combined method of those two. By handling part-of-speech tags and position-of-character tags equally, word-level hypotheses and character-level hypotheses are treated uniformly, and known words and unknown words are processed simultaneously. Experiments of word segmentation are conducted on multiple Chinese and Japanese corpora, showing that the proposed method achieves higher performance than most of previous methods.

1. はじめに

中国語や日本語の単語分割処理は，最も基本的な言語解析処理の1つである．このような分かち書きされない言語の文を単語分割する際には曖昧性が存在するため，それを正しく解消する必要がある．これらの言語での単語分割を難しくする1つの要因として，未知語の存在がある．未知語とは，単語分割システムの辞書中に存在しない単語と定義される．そのような未知語に対しては単語に関する知識をシステムが持っていないため，正しく単語分割を行うのは非常に難しい．この未知語に対処する方法として，自動的な語彙獲得法を用いて辞書の登録語を増やす試みがいくつか行われている[19),20)]．これは，あらかじめ大量のテキストデータから統計情報等を用いて未知語を抽出し，単語分割システムの辞書に追加しておくことで，解析中に出現する未知語を減らすというアプローチである．このような手法を用いて解析対象の文書に出現する語彙を登録しておくことができれば，精度の高い解析を行うことが可能となる．しかしながら，特に固有表現等の未知語は日々増え続けており，あらゆる単語を前もって辞書に登録しておくことは不可能である．高精度な単語分割システムを実現するうえで未知語は避けることのできない問題であり，解析中に出現する未知語に対して頑健に処理できることが重要である．そこで，本稿では解析中に未知語を処理する方法に焦点をあてる．

単語分割に関する従来研究の中で，未知語の扱いが難しい単語ベースの単語分割手法は既知語の再現率が高く，文字ベースの単語分割手法は未知語の再現率が高いが既知語の再現率が低いことが確認されている．そこで本稿では，中国語や日本語の単語分割のために，単語単位で解析を行うコスト最小法と文字単位で解析を行う文字タグ付け法を結合させた単語分割手法を提

† 沖電気工業株式会社研究開発本部
Corporate Research and Development Center, Oki Electric Industry Co., Ltd.
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

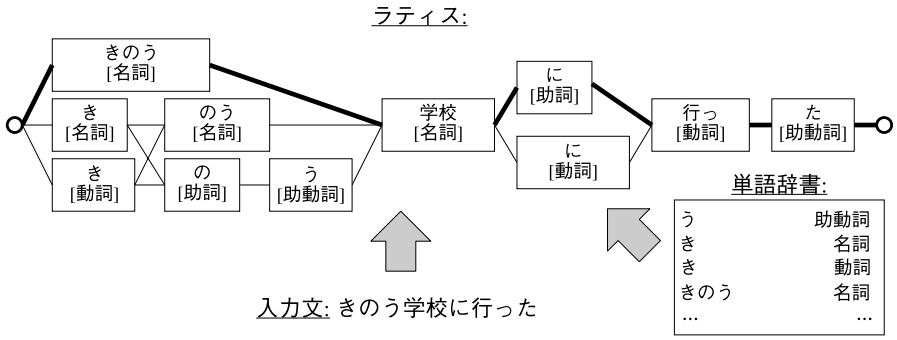

図 1 解析結果の候補を表すラティス
Fig. 1 Example of a lattice representing possible hypotheses.

案する．これにより，既知語に対しても未知語に対しても高い解析精度が得られる可能性がある．なお，本稿では中国語と日本語の 2 つの言語を対象とするが，その理由は，単語分割を必要とする言語の中でもこれらの 2 言語は重要なものであり，また実験を行ううえで必要となるコーパス等の言語資源が比較的整備されているからである．

以下，2 章では本手法が参考とした既存の単語分割手法について説明する．3 章では提案手法の説明を行い，4 章で実験結果を報告する．5 章では従来研究との比較と考察を行い，6 章で結論を述べる．

2. 従 来 研 究

我々の提案手法は，既存の 2 つの単語分割手法（コスト最小法と文字タグ付け法）に基づいているため，この章ではそれらの従来手法について説明する．

2.1 コスト最小法

コスト最小法は，実用的な日本語形態素解析（単語分割）システムにおいて，広く用いられている方法である[24),25)]．コスト最小法は，英語の品詞タグ付けで用いられるマルコフモデルと本質的に同じであるため，以下ではマルコフモデルをもとにしてコスト最小法の説明を行う．マルコフモデルに基づく形態素解析では，入力された文 S に対して，それを単語分割した結果である単語列 $W = w_1, \cdots, w_n$ と，各単語の品詞を推定した結果の品詞列 $T = t_1, \cdots, t_n$ を確率に基づいて同時に決定する．この際に，品詞が出現する確率はその 1 つ前の品詞のみに依存し，また単語が出現する確率はその単語の品詞のみに依存すると仮定する．このような仮定のもとで，品詞列 T を持つ単語列 W の出現確率は次のように計算される：

$$\begin{aligned} P(W,T) &= \prod_{i=1}^{n} P(w_i t_i | w_0 t_0 \dots w_{i-1} t_{i-1}), \\ &= \prod_{i=1}^{n} P(w_i | w_0 t_0 \dots w_{i-1} t_{i-1} t_i) \\ &\quad \times P(t_i | w_0 t_0 \dots w_{i-1} t_{i-1}), \\ &\simeq \prod_{i=1}^{n} P(w_i | t_i) P(t_i | t_{i-1}). \end{aligned} \tag{1}$$

ここで，w_0（t_0）は文頭を表す特殊な単語（品詞）である．入力文 S が与えられた場合，その形態素解析結果 $\hat{W}$ と $\hat{T}$ は次のように計算される（ここで，W は S の可能な単語分割候補であり，$w_1 \cdots w_n = S$ である）：

$$\begin{aligned} (\hat{W}, \hat{T}) &= \operatorname*{argmax}_{W,T} P(W,T|S), \\ &= \operatorname*{argmax}_{W,T} P(W,T), \\ &\simeq \operatorname*{argmax}_{W,T} \prod_{i=1}^{n} P(w_i|t_i) P(t_i|t_{i-1}). \end{aligned} \tag{2}$$

上式の解は，Viterbi アルゴリズムを用いて効率的に求めることができる．実際に上式の計算を行う際には，小さな数の乗算により計算機上でアンダフローが起こるのを防ぐため，確率値の逆数の対数をとった値が使用される．このような値をコストと見なすと，確率の積を最大にする問題はコストの和を最小にする問題となり，上述のマルコフモデルに基づく形態素解析手法は，従来から日本語の形態素解析手法として知られてきたコスト最小法と等しくなる．また，入力文に対する可能な単語分割候補は，**図 1** に示されるようなラティスを用いて表現することができる．

以上の説明をまとめると，コスト最小法による形態素解析は次のように行われる：

1) 入力文に対して単語辞書を使用して，解析結果の候補を表すラティスを作成する．
2) ラティス中から式 (2) に基づいて文の生成確率を最大にする（生成コストを最小にする）最適なパ

表 1 'B, I, E, S' タグ
Table 1 The 'B, I, E, S' tag set.

タグ	意味
B	単語の先頭にある文字
I	単語の中間にある文字
E	単語の末尾にある文字
S	1 文字で単語を構成する文字

文: き の う | 学 校 | に | 行 っ | た
POCタグ: **B I E B E S B E S**

図 2 文字タグ付け法による単語分割
Fig. 2 Example of the character tagging method.

スを探索し，解析結果を得る．

この手法は，既知語（未知語とは逆に，システムの辞書中に存在する単語）に対しては高い解析精度を持ち，計算時間も速いという特徴がある．しかしながらそのままでは未知語を処理することはできない．そこで，ラティスを構築する際には辞書中に存在する単語だけではなく，未知語の候補が何らかの処理によって追加される．しばしば用いられる手法として，文字種を利用したヒューリスティックな規則により未知語候補を生成する方法がある．これは，「連続するカタカナやアルファベットはまとめて 1 つの単語候補とし，漢字やひらがなは 1 文字だけで単語の候補にする」というような規則を用いる方法である[24),25)]．また，任意の文字列に対してそれが未知語として出現する確率を計算するようなモデル（単語モデル）を用意する方法も試みられている[7),13)]．しかしながらこれらの方法は，対象とする言語に依存した規則やモデルをあらかじめ注意深く設計する必要があり，多様な未知語を幅広く扱うのは難しい．

2.2 文字タグ付け法

文字タグ付け法は，単語分割の問題を，文を構成する各文字に対してその文字の単語中での位置を表すタグ（position-of-character タグ；POC タグ[15)]）を付与する問題として解く．POC タグ集合はいくつか提案されているが，本稿では**表 1** のような **B**, **I**, **E**, **S** の 4 つのタグを使用する☆．

図 2 に POC タグ付けの例を示す．このように，POC タグを使うことで任意の文に対して単語分割の情報を表現できるため，単語分割の問題は POC タグ付けの問題に変換することができる．そのようなタグ付けの問題は，一般的な機械学習アルゴリズムを用いた多値分類問題として解くことができる．これまでに，マルコフモデル[16),28)]，PPM*モデル[27)]，決定リスト[18)]，最大エントロピー（Maximum Entropy; ME）モデル[15)]，AdaBoost[17)]，サポートベクタマシーン[1),2),23)] 等の機械学習アルゴリズムを用いた方法が試みられている．このように 1 つのまとまり（この場合は単語）を構成する要素に対して，その要素の位置を表現するタグを付与することによりまとまりの同定を行う手法は，単名詞句同定[10),14)] や固有表現抽出[11)] 等で広く利用されている．

文字タグ付け法による単語分割は単語辞書を利用しないため，既知語も未知語も区別することなく扱うことができ，未知語に対する特別な処理を必要としない利点がある．しかしながら，単語辞書が存在する場合でもその情報を利用するのは困難であり，単語分割だけではなく品詞の推定も行う場合にはコスト最小法と比較して探索空間が大きく計算量が増大するという性質を持つこと等から広く利用されてはいない．

3. 単語レベルと文字レベルの情報を用いた単語分割

前章では，既存の 2 つの単語分割手法について説明した．従来研究において，コスト最小法は全体的な解析精度は比較的高いものの未知語に対する再現率が低く，また文字タグ付け法は未知語に対する再現率は高いものの既知語に対する再現率が低いことが実験結果から観測されている[12),15),23)]．このような性質は，各手法の特徴を考えると妥当であると思われる．つまり，コスト最小法では単語を単位として処理が行われるため，既知語に対しては単語に関する豊富な情報（品詞や単語の出現確率等）を扱うことができ，また文字単位で処理する場合と比較して文中の広範囲の情報を利用することができる．しかしながら，コスト最小法のような単語ベースの方法では未知語を直接処理することはできず，また単純な規則等を用いた未知語処理では十分な解析精度を得るのは難しい．一方で，文字タグ付け法では文字を単位として処理が行われる．文字の種類数は一般に有限であり，つねに増加し続けている単語の種類数に比べるとはるかに少ない．そのため，文字タグ付け法は単語単位の処理と比べて頑健であると思われるが，文字レベル以上の細かい情報や広い範囲の情報を扱うのは難しい．単語辞書等の単語レベルの情報は，単語分割を行ううえで非常に有用であると

☆ 予備実験において，別の POC タグ集合である IOB2 モデル[14)] の利用も試みたが，**B**, **I**, **E**, **S** タグを利用した場合と比較して解析精度が若干低下した．

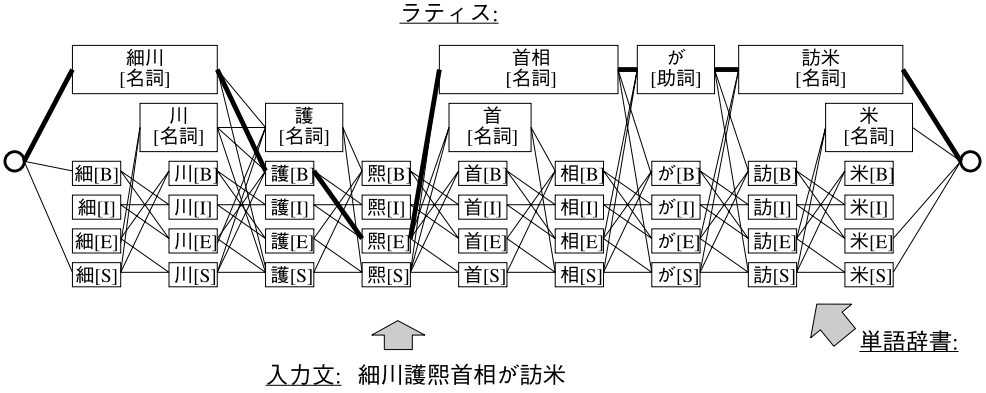

図 3 提案手法による単語分割
Fig. 3 Example of the proposed method.

思われるが☆，それを利用できないのはこの手法の問題点であると思われる．

以上のことをふまえ，単語単位と文字単位の両方の情報を利用することで高精度で頑健な解析を行うことをねらいとして，コスト最小法と文字タグ付け法を組み合わせた単語分割手法を提案する．

3.1 提案手法

本稿で提案する手法は，コスト最小法をベースにしているが，POC タグを用いて未知語を構成する文字の候補を考慮することにより未知語の同定を行う．POC タグと品詞タグを同等に扱って計算を行うことで，単語単位と文字単位の解候補を統一的に扱い，既知語と未知語に対する単語分割を同時に行う．

提案手法により単語分割を行う際の，ラティスの例を**図 3** に示す．ここでは,「細川護熙首相が訪米」という文が入力され,「護熙」という単語が未知語であったと仮定する．このような文が与えられると，まず初めに通常のコスト最小法と同様にして既知語に対するラティスのノードを作成する．次に，文中の各文字に対して POC タグのノード（1 つの文字に対して計 4 つのノード）を作成する．そして，このラティスの中から最も尤度の高いパスを探索する（図 3 の例では正解のパスは太線で示されている）．ラティス中のノード間の遷移について，いくつかのタグ間での遷移（**I** タグから **B** タグ，あるいは任意の品詞タグから **E** タグ，等）は許されないため，そのような遷移は計算時に無視する．

☆ 最も単純な単語ベースの解析手法である最長一致法を用いて中国語の単語分割を行った場合，テストデータ中の単語がすべて辞書に登録されて未知語が存在しない状況では，98%以上の高い精度で解析できることが報告されている[12]．

3.1.1 確率モデル

提案手法では，POC タグを品詞タグと同等に扱う．これにより，図 3 のラティス中での未知語を構成する文字のノードと既知語のノードはまったく同じように扱うことができ，通常のマルコフモデルに基づく形態素解析と同じ方法を用いて尤度の高いパスを探索することができる．ただし，式 (1) で実現される基本的な確率モデル（品詞 bigram モデル）はモデルの表現力が十分ではない．そのため，英語の品詞タグ付けや日本語の形態素解析ではしばしば品詞 trigram モデルや語彙情報の利用が行われる．本稿では，品詞 unigram，品詞 bigram，品詞 trigram，単語 bigram 確率☆☆を線形補間した次の式により，ラティス中のパスの生成確率を計算する：

$$\begin{aligned} P(W,T) &= \prod_{i=1}^{n} P(w_i t_i | w_0 t_0 \ldots w_{i-1} t_{i-1}), \\ &\simeq \prod_{i=1}^{n} \{\lambda_1 P(w_i|t_i) P(t_i) \\ &\quad +\lambda_2 P(w_i|t_i) P(t_i|t_{i-1}) \\ &\quad +\lambda_3 P(w_i|t_i) P(t_i|t_{i-2} t_{i-1}) \\ &\quad +\lambda_4 P(w_i t_i | w_{i-1} t_{i-1})\}, \\ &\quad (\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1). \end{aligned} \tag{3}$$

上式の各確率の値は，訓練データ（品詞タグ付きコーパス）から最尤推定により求める．ただし，辞書中に含まれるにもかかわらず訓練データ中に出現しない単語に対処するために，出現頻度が 0 回だった単語は 0.5 回出現したものとして扱った．λ_1，λ_2，λ_3，λ_4 の値は，**図 4** に示される leave-one-out 法により計算した[3]．上式の各パラメータを推定する際には，訓練

☆☆ ここでは，単語と品詞の組の bigram 確率を単語 bigram 確率と呼ぶことにする．

```
# W:  コーパス中に出現する単語の集合,
# T:  コーパス中に出現する品詞の集合,
# f(w_i, t_i, ...):  単語・品詞列 w_i, t_i, ... が訓練コーパス中に出現する回数,
# N:  コーパス中の単語数.

# 初期化
for i := 1 to 4
  λ_i := 0
# λ_i の計算
foreach t_{-2}, t_{-1}, t_0 ∈ T, w_{-1}, w_0 ∈ W, f(t_{-2}, w_{-1}, t_{-1}, w_0, t_0) ≥ 1
begin
  # 以下の p_1, p_2, p_3, p_4 の計算において，除数が 0 ならば除算の結果は 0 と定義する
  p_1 := (f(w_0,t_0)-1)/(f(t_0)-1) * (f(t_0)-1)/(N-1)
  p_2 := (f(w_0,t_0)-1)/(f(t_0)-1) * (f(t_{-1},t_0)-1)/(f(t_{-1})-1)
  p_3 := (f(w_0,t_0)-1)/(f(t_0)-1) * (f(t_{-2},t_{-1},t_0)-1)/(f(t_{-2},t_{-1})-1)
  p_4 := (f(w_{-1},t_{-1},w_0,t_0)-1)/(f(w_{-1},t_{-1})-1)
  for i := 1 to 4
    if p_i = max_j p_j
      λ_i := λ_i + f(t_{-2}, w_{-1}, t_{-1}, w_0, t_0)
end
# 正規化
Z := Σ_i λ_i
for i := 1 to 4
  λ_i := λ_i / Z
```

図 4 λ_i を計算するためのアルゴリズム
Fig. 4 Algorithm for calculating λ_i.

データ中に存在するにもかかわらず辞書中には存在しない単語と品詞の組を，POC タグの付与された文字に分解して扱った．コスト最小法で使用される単語辞書はしばしば訓練データから作成されるが，その場合訓練データ中には未知語が存在しないため，未知語に関する情報を学習させることができない．そこでこのような場合には訓練データ中に 1 回だけ出現した単語を未知語と見なし，POC タグの付与された文字に分解して扱った．

3.1.2 文字レベルの素性の利用

未知語の単語境界を決定する際に，境界付近の文字種等の情報は有用であると思われる．そこで様々な文字レベルの素性を利用するために，式 (3) における POC タグで条件付けられた単語（文字）出力確率はベイズの定理により次の式で計算することにする：

$$P(w_i|t_i) = \frac{P(t_i|w_i)P(w_i)}{P(t_i)}. \qquad (4)$$

ここで，w_i は文字で t_i は POC タグである．上の式の右辺において，$P(t_i)$ と $P(w_i)$ は最尤推定により求めるが，文字 w_i が与えられた場合の POC タグ t_i の確率 $P(t_i|w_i)$ は，ME モデルにより計算する．ME モデルの素性としては次のものを用いた（ただし，c_x は文頭から x 番目の文字を，y_x は c_x の文字種を表すものとし，$w_i = c_{i'}$ とする．また，文字種としては，アルファベット，数字，記号，漢字，ひらがな，カタカナの 6 つを用いる）：

(1) 文字（$c_{i'-2}$, $c_{i'-1}$, $c_{i'}$, $c_{i'+1}$, $c_{i'+2}$）
(2) 文字の対（$c_{i'-2}c_{i'-1}$, $c_{i'-1}c_{i'}$, $c_{i'-1}c_{i'+1}$, $c_{i'}c_{i'+1}$, $c_{i'+1}c_{i'+2}$）
(3) 文字種（$y_{i'-2}$, $y_{i'-1}$, $y_{i'}$, $y_{i'+1}$, $y_{i'+2}$）
(4) 文字種の対（$y_{i'-2}y_{i'-1}$, $y_{i'-1}y_{i'}$, $y_{i'-1}y_{i'+1}$, $y_{i'}y_{i'+1}$, $y_{i'+1}y_{i'+2}$）

ME モデルのパラメータの学習は，訓練データ中のすべての単語を用いて行った．パラメータの推定には GIS アルゴリズム[4] を使用し，出現回数が 10 回以下の素性は過学習を避けるために削除した．

4. 実　　験

提案手法の有効性を評価するために，中国語と日本語の単語分割実験を行った．提案手法では単語の分割だけではなく既知語に対する品詞のタグ付けも行うが，未知語に対する品詞の付与は行われないため，本稿では単語分割の解析精度のみを評価することとする．単語分割精度の評価には，以下の評価尺度を使用した：

R：再現率（$R =$ ⟨解析結果中の正解単語数⟩ ÷ ⟨テストデータ中の単語数⟩）
P：精度（$P =$ ⟨解析結果中の正解単語数⟩ ÷ ⟨解析結果中の単語数⟩）

表 2 使用したコーパス
Table 2 Statistical information of corpora.

コーパス	訓練データの単語数	テストデータの単語数 (既知語/未知語)	辞書の登録語数	未知語の割合	品詞の種類数
AS	5,806,611	11,985 (11,727/ 258)	146,212	0.0215	(64)
CTB	250,841	39,922 (32,706/ 7,216)	19,730	0.1808	(64)
HK	239,852	34,955 (32,463/ 2,492)	23,747	0.0713	(64)
PK	1,121,017	17,194 (16,005/ 1,189)	55,226	0.0692	(64)
EDR	2,452,891	2,652,156 (2,600,051/52,105)	82,410	0.0196	15
KY	198,514	31,302 (29,926/ 1,376)	1,870,461	0.0440	42
RWCP	840,879	93,155 (93,085/ 70)	315,602	0.0008	69

F：F 値（$F = 2 \times R \times P \div (R + P)$）
R_{known}：既知語に対する再現率
（R_{known} = ⟨解析結果中の正解既知語数⟩ ÷ ⟨テストデータ中の既知語数⟩）
$R_{unknown}$：未知語に対する再現率
（$R_{unknown}$ = ⟨解析結果中の正解未知語数⟩ ÷ ⟨テストデータ中の未知語数⟩）
P_{known}：既知語に対する精度
（P_{known} = ⟨解析結果中の正解既知語数⟩ ÷ ⟨解析結果中の既知語数⟩）
$P_{unknown}$：未知語に対する精度
（$P_{unknown}$ = ⟨解析結果中の正解未知語数⟩ ÷ ⟨解析結果中の未知語数⟩）

4.1 中国語単語分割

中国語単語分割の性能を評価するために，Academia Sinica コーパス（AS），Penn Chinese Treebank コーパス（CTB），Hong Kong City University コーパス（HK），Beijing University コーパス（PK）の 4 つのコーパスを使用した☆．AS，HK は繁体字のコーパスであり，CTB，PK は簡体字のコーパスである．これらはすべて単語分割済みのコーパスで，ACL-SIGHAN 2003 における First International Chinese Word Segmentation Bakeoff[12] で使用されたものである．これらのコーパスは，単語分割はされているが品詞タグが付与されていないため，コスト最小法に基づく解析に使用するためには各単語に品詞タグを付与しなければならない．そこで，隠れマルコフモデルの学習に用いられる Baum-Welch アルゴリズム[5] を利用して，品詞タグの代わりになるクラスの付与を教師なし学習により行った．その際に，初期状態はランダムに与えクラスの数は 64 とした☆☆．

☆ AS, HK, PK の各データは SIGHAN Bakeoff の Web ページ（http://www.sighan.org/bakeoff2003/alldata.html）より入手可能である．また CTB のデータは，LDC（http://www.ldc.upenn.edu/）より入手可能である（カタログ番号：LDC2003E16）．

表 3 計算された λ_i の値
Table 3 Calculated values of λ_i.

コーパス	λ_1	λ_2	λ_3	λ_4
AS	0.037	0.178	0.257	0.528
CTB	0.044	0.219	0.251	0.486
HK	0.048	0.251	0.313	0.388
PK	0.055	0.207	0.242	0.495
EDR	0.047	0.085	0.116	0.752
KY	0.080	0.126	0.237	0.556
RWCP	0.073	0.105	0.252	0.571

解析精度の比較を行うために，次のシステムを使用した：

Bakeoff-1, 2, 3 SIGHAN Bakeoff[12] に参加した上位 3 つのシステム．

最長一致法 最長一致法に基づく単語分割システム．

文字タグ付け法 文字タグ付け法に基づく単語分割システム．これは Xue の方法[15] とほぼ同じであり，素性には 3.1.2 項で記述した (1)–(4) と次の (5) を使用して，ME モデルにより文中の各文字 $c_{i'}$ の POC タグを推定した（ここで，t_x は文字 c_x の POC タグである）：

(5) POC タグと POC タグの対 $(t_{i'-1}, t_{i'-2}t_{i'-1})$ 解の探索は Viterbi アルゴリズムを用いて行う．

上記のシステムと提案手法によるシステムは，訓練データ以外の言語資源等は一切用いていない．この実験では，最長一致法と提案手法で使用される単語辞書は，訓練データ中に含まれるすべての単語を取り出して作成した☆☆☆．これらのデータに関する統計情報を**表 2** に示す．また，式 (3) 中の λ_i の値は，学習の結果**表 3** のような値となった．

☆☆ コーパスや辞書で使用される品詞の種類数は 40〜70 程度であることが多いため（英語の Penn Treebank コーパスは 45，JUMAN の辞書（細分類）は 42，IPADIC（細分類）は 69 である），それに近い 64 という数を使用した．

☆☆☆ PK データ中のアラビア数字は，使用されている文字コードが訓練データ中とテストデータ中で異なっているため，訓練データ中のすべてのアラビア数字を ASCII コードに変換して使用した．

表 4 中国語単語分割の実験結果
Table 4 Performance of Chinese word segmentation.

コーパス		R	P	F	R_{known}	$R_{unknown}$	P_{known}	$P_{unknown}$
AS	提案手法	**0.973**	**0.970**	**0.972**	0.979	0.717	**0.974**	**0.804**
	Bakeoff-1	0.966	0.956	0.961	**0.980**	0.364	0.961	0.584
	Bakeoff-2	0.961	0.958	0.959	0.966	0.729	0.967	0.614
	Bakeoff-3	0.944	0.945	0.945	0.952	0.574	0.957	0.490
	最長一致法	0.917	0.912	0.915	0.938	0.000	0.914	0.000
	文字タグ付け法	0.961	0.959	0.960	0.966	**0.744**	0.968	0.610
CTB	提案手法	0.877	0.872	0.874	0.927	0.647	**0.919**	0.655
	Bakeoff-1	0.886	**0.875**	**0.881**	0.927	**0.705**	0.913	**0.701**
	Bakeoff-2	**0.892**	0.856	0.874	0.947	0.644	N/A	N/A
	Bakeoff-3	0.867	0.797	0.831	**0.963**	0.431	0.834	0.551
	最長一致法	0.800	0.663	0.725	**0.963**	0.063	0.736	0.084
	文字タグ付け法	0.832	0.836	0.834	0.872	0.650	0.913	0.552
HK	提案手法	**0.951**	**0.948**	**0.950**	0.969	0.715	**0.965**	**0.718**
	Bakeoff-1	0.947	0.934	0.940	0.972	0.625	N/A	N/A
	Bakeoff-2	0.940	0.908	0.924	**0.980**	0.415	0.922	0.627
	Bakeoff-3	0.917	0.915	0.916	0.936	0.670	0.953	0.537
	最長一致法	0.908	0.830	0.867	0.974	0.037	0.867	0.052
	文字タグ付け法	0.917	0.917	0.917	0.932	**0.728**	0.957	0.545
PK	提案手法	0.957	**0.951**	**0.954**	0.970	0.774	**0.961**	0.811
	Bakeoff-1	**0.962**	0.940	0.951	**0.979**	0.724	0.943	**0.904**
	Bakeoff-2	0.955	0.938	0.947	0.976	0.680	0.942	0.867
	Bakeoff-3	0.955	0.938	0.946	0.977	0.647	0.946	0.815
	最長一致法	0.930	0.883	0.906	0.973	0.347	0.896	0.579
	文字タグ付け法	0.932	0.931	0.931	0.943	**0.786**	0.958	0.639

実験結果を**表 4** に示す．提案手法は，AS，HK，PK コーパスで最も高い F 値を得た．再現率（R）と精度（P）について統計的仮説検定（危険率 5%）を行ったところ，AS コーパスについては再現率と精度が，HK と PK コーパスについては精度のみが，他のシステムと比較して有意差があった☆．提案手法では既知語に対する再現率が他のシステムと比べてやや低いが，既知語と未知語の再現率は比較的バランスがとれている．最長一致法と文字タグ付け法の結果には，3 章で論じたトレードオフの関係が見られる．最長一致法は最も単純な単語ベースの手法といえるが，CTB，HK，PK コーパスで既知語に対する再現率が文字タグ付け法と比較して高い．文字タグ付け法は文字ベースの手法だが，AS，HK，PK コーパスで未知語に対する再現率は最も高い（なお，AS の Bakeoff-2 と，HK の Bakeoff-3 も文字タグ付け法に基づくシステムである）．

4.2 日本語単語分割

日本語単語分割の性能を評価するために，EDR コーパス version 1.0（EDR），京大コーパス version 2.0（KY），RWCP コーパス（RWCP）の 3 つのコーパスを使用した．いずれも，単語分割済みの品詞タグ付きコーパスである．

解析精度の比較を行うために，次のシステムを使用した：

茶筌（version 2.2.8） 拡張統計モデル[21] に基づく日本語形態素解析システム[25]．未知語は文字種を利用した規則により処理する．

JUMAN（version 3.61） コスト最小法に基づく日本語形態素解析システム[24]．未知語は文字種を利用した規則により処理する．

最長一致法 中国語の実験で使用したものと同じ．

文字タグ付け法 中国語の実験で使用したものと同じ．

上記のシステムの中で，JUMAN は人手により設定されたパラメータを使用するが，その他のシステムでは訓練データと単語辞書以外のデータは使用していない☆☆．また茶筌は RWCP コーパスの形態素体系を利用するシステムであり，JUMAN は京大コーパスの形

☆ テストデータ中の単語数が n，再現率が R である場合，再現率の 95%信頼区間は $R \pm 2\sqrt{R(1-R)/n}$ であり，異なるシステム間の信頼区間に重なりがなければ有意差があるといえる[12]．精度（P）についても同様に検定を行った．

☆☆ 京大コーパスを用いた実験と RWCP コーパスを用いた実験において，最長一致法では訓練データの情報を利用することができないため辞書のみを使用し，文字タグ付け法では辞書の情報を利用することができないため訓練データのみを使用している．そのため，これらのシステムにとっては不利な条件となっている．EDR コーパスを用いた実験と前節の中国語の実験では，辞書は訓練データから作成したため，各システムが利用している情報にこのような違いは存在しない．

表 5 日本語単語分割の実験結果
Table 5 Performance of Japanese word segmentation.

コーパス		R	P	F	R_{known}	$R_{unknown}$	P_{known}	$P_{unknown}$
EDR	提案手法	**0.952**	**0.948**	**0.950**	**0.962**	0.446	0.951	**0.693**
	最長一致法	0.757	0.824	0.789	0.772	0.009	0.828	0.039
	文字タグ付け法	0.946	0.946	0.946	0.953	**0.591**	**0.956**	0.501
KY	提案手法	0.986	0.984	0.985	0.989	0.924	0.984	0.968
	JUMAN	**0.989**	**0.985**	**0.987**	**0.993**	0.889	**0.985**	**0.985**
	最長一致法	0.806	0.753	0.779	0.843	0.004	0.829	0.002
	文字タグ付け法	0.946	0.942	0.944	0.947	**0.940**	0.959	0.685
RWCP	提案手法	**0.993**	**0.994**	**0.993**	**0.993**	0.586	**0.994**	**0.820**
	茶筌	0.991	0.992	0.991	0.991	0.243	0.992	0.515
	最長一致法	0.879	0.918	0.898	0.880	0.100	0.919	0.103
	文字タグ付け法	0.972	0.968	0.970	0.972	**0.629**	0.977	0.051

態素体系を利用するシステムであるが，これらのシステムを他の形態素体系に適用するのは容易ではないため，それぞれ対応したコーパスでのみ評価を行った．

EDR コーパスを用いた実験では，コーパスの先頭から 100,000 文を訓練データ，残りの 107,802 文をテストデータとして，最長一致法と提案手法で使用する辞書は訓練データから作成した．京大コーパスを用いた実験では，1 月 1 日と 1 月 3 日から 8 日までの 7 日分のデータを訓練データ，1 月 9 日の 1 日分のデータをテストデータとして，JUMAN，最長一致法，提案手法で使用する辞書には JUMAN version 3.61 に付属の辞書を使用した．RWCP コーパスを用いた実験では，訓練データとテストデータはランダムに抽出し，茶筌，最長一致法，提案手法で使用する辞書には茶筌とともに配布されている IPADIC version 2.4.4 [26] を用いた．これらのデータの統計情報を表 2 に示す．式 (3) 中の λ_i の値は，学習の結果表 3 のような値となった．

実験結果を，**表 5** に示す☆．提案手法の結果を茶筌や JUMAN の結果と比較した場合，F 値にあまり大きな差はなかったが，未知語に対する再現率については高い値を得た．文字タグ付け法は，中国語での実験の場合と同様に未知語に対する再現率が高かった．統計的仮説検定（危険率 5%）を行ったところ，提案手法は，EDR と RWCP コーパスについては再現率（R）・精度（P）ともに他のシステムと比較して有意差があった．

KY コーパスでの実験について提案手法での解析に失敗した単語を調べてみると，全体の誤りのうちの約 1/4 が未知語に対する誤りであり，未知語は依然として解析誤りの大きな原因となっていた．残りの誤りのうち約 1/5 は辞書には存在するが訓練コーパスには出現しない単語であった．また，内元ら [22] によって指摘されているようなコーパス中での揺れによると思われる解析誤りも見られた．たとえば，「やさしさ」という言葉はコーパス中では「やさしさ」という 1 語として扱われていたり「やさし/さ」という 2 語として扱われていたりするため，正しく解析できない場合があった．

4.3 条件を変えた実験

4.3.1 確率モデルを変えた実験

提案手法ではラティス中のパスの生成確率を計算する際に，式 (3) によって品詞 unigram，品詞 bigram，品詞 trigram，単語 bigram を混合して使用した．これらの各 n-gram モデルが解析精度に及ぼす影響を調べるため，ラティス中のパスの生成確率を計算するモデルを変えて実験を行った．実験の結果を**表 6** に示す．この表中で，**品詞 bigram**，**品詞 trigram**，**単語 bigram** は，ラティス中のパスの生成確率（式 (2) の $P(W,T)$）にそれぞれ式 (5)，(6)，(3) を使用した場合の結果を表す．

$$P(W,T) \simeq \prod_{i=1}^{n} \{\lambda_1 P(w_i|t_i)P(t_i) + \lambda_2 P(w_i|t_i)P(t_i|t_{i-1})\}. \quad (5)$$

$$P(W,T) \simeq \prod_{i=1}^{n} \{\lambda_1 P(w_i|t_i)P(t_i) + \lambda_2 P(w_i|t_i)P(t_i|t_{i-1}) + \lambda_3 P(w_i|t_i)P(t_i|t_{i-2}t_{i-1})\}. \quad (6)$$

この実験結果から，品詞 trigram よりも，語彙情報を利用する単語 bigram の方が精度の向上に寄与していたことが分かる．また中国語のコーパス（AS，CTB，

☆ この評価において，$R_{known}, R_{unknown}, P_{known}, P_{unknown}$ は，辞書中に存在した単語を既知語として計算しており，訓練コーパス中に含まれていても辞書中に含まれない単語は未知語としている．表 2 における KY，RWCP コーパスの既知語/未知語の数も同じ方法で計算されている．なお，出現形が同じで品詞が異なった単語が辞書中に登録されていた場合，その単語は既知語として扱った．

表 6 確率モデルを変えた場合の実験結果
Table 6 Performance for different probabilistic models.

コーパス	$F(R_{known}/R_{unknown})$		
	品詞 bigram	品詞 trigram	単語 bigram
AS	0.967 (0.976/0.616)	0.967 (0.976/0.624)	0.972 (0.979/0.717)
CTB	0.868 (0.924/0.636)	0.868 (0.923/0.644)	0.874 (0.927/0.647)
HK	0.940 (0.966/0.660)	0.939 (0.964/0.675)	0.950 (0.969/0.715)
PK	0.948 (0.969/0.740)	0.948 (0.968/0.751)	0.954 (0.970/0.774)
EDR	0.922 (0.923/0.319)	0.923 (0.925/0.320)	0.950 (0.962/0.446)
KY	0.980 (0.983/0.898)	0.981 (0.984/0.907)	0.985 (0.989/0.924)
RWCP	0.989 (0.987/0.543)	0.991 (0.989/0.557)	0.993 (0.993/0.586)

表 7 文字レベルの素性の効果
Table 7 Effect of character-level features.

コーパス	$F(R_{known}/R_{unknown})$	
	文字レベルの素性なし	文字レベルの素性あり
AS	0.969 (0.977/0.671)	0.972 (0.979/0.717)
CTB	0.854 (0.904/0.593)	0.874 (0.927/0.647)
HK	0.941 (0.958/0.671)	0.950 (0.969/0.715)
PK	0.952 (0.967/0.750)	0.954 (0.970/0.774)
EDR	0.948 (0.961/0.382)	0.950 (0.962/0.446)
KY	0.982 (0.988/0.885)	0.985 (0.989/0.924)
RWCP	0.993 (0.993/0.557)	0.993 (0.993/0.586)

HK，PK）に対しては，品詞 trigram を使用しても精度は向上していない．この原因としては，教師なし学習によって学習したクラスを品詞の代わりに利用したことも考えられるが，浅原ら[21] が指摘しているように，中国語の単語分割においては品詞 trigram の情報はあまり有用ではないことも考えられる．また，今回使用した確率モデルにおいて補間係数 λ_i は文脈に依存していないが，文脈に依存した係数を使用することにより解析精度が改善される可能性もある．

4.3.2 文字レベルの素性の影響

提案手法では，文字レベルの様々な素性を扱うために，ME モデルを使用して POC タグに対する文字出現確率を計算した（3.1.2 項参照）．このような文字レベルの素性を用いたことによる影響を調べるため，POC タグに対する文字出現確率（式 (3) 中で，t_i が POC タグである場合の $P(w_i|t_i)$）の計算の際に，文字レベルの素性は利用せずに単純な最尤推定を用いた場合と，3.1.2 項で説明した方法により文字レベルの素性を利用して ME モデルを用いた場合の実験を行った．結果を**表 7** に示す．ME モデルを用いて文字レベルの詳細な素性を利用した場合，未知語に対する再現率が大きく向上し，RWCP 以外のコーパスで F 値も向上している．RWCP コーパスにおいても未知語に対する再現率は向上しており，また未知語に対する精度や既知語に対する再現率・精度はいずれもわずかに向上していたが，未知語の割合が極端に低かったため（表 2 参照），F 値はほとんど変わらなかった．

4.3.3 教師なし学習の影響

4.1 節の中国語単語分割の実験では，使用したコーパスに品詞タグが付与されていなかったため，教師なし学習によって付与したクラスを品詞タグの代わりに使用した．このような教師なし学習を利用したことの妥当性を調べるため，人手によって品詞タグが付与されたコーパスを使用して，人手で付与された品詞タグと教師なし学習で付与されたクラスのそれぞれをマルコフモデルの状態として用いた場合の解析精度の比較を行った．コーパスとしては，EDR，KY，RWCP コーパスを利用し，Baum-Welch アルゴリズムを用いて 4.1 節の実験の場合と同様に初期値をランダムに設定して 64 のクラスを付与した．使用した訓練データ・テストデータのサイズは 4.2 節の実験の場合と同一であるが，この実験では辞書は訓練データから作成した．そのようにした理由は，教師なし学習によって品詞を付与した場合，辞書中の品詞と訓練データ中の品詞の整合性がとれなくなるからである．結果を**表 8** に示す．この結果から，人手で付与された品詞を用いた場合も教師なし学習によって付与された品詞を用いた場合も F 値に大きな違いはなく，Baum-Welch アルゴリズムにより付与したクラスを品詞の代わりに用いても十分な解析精度が得られていることが分かる．教師なし学習で付与された品詞を用いた場合に未知語の再現率が高くなっているが，この原因の 1 つとして次のことが考えられる．教師なし学習を用いた場合，同一の単語に対してしばしば異なるクラスが付与される（京大コーパスにおいて 1 つの単語が持つ品詞の数の平均値を調べたところ，人手で付与された場合は 1.07 個，教師なし学習で付与された場合は 2.08 個だった）．結果として，教師なし学習で付与された品詞を用いる場合は，単語と品詞を組として考えた場合の未知語が多くなり，確率モデルのパラメータを求める際の未知語（実際には未知語を構成する文字）に対する訓練事例の数が相対的に大きくなるため，未知語に対する再

表 8 教師なし学習の影響
Table 8 Effect of unsupervised learning.

コーパス	F ($R_{known}/R_{unknown}$)	
	人手で付与された品詞	教師なし学習で付与された品詞
EDR	0.950 (0.962/0.446)	0.950 (0.959/0.520)
KY	0.959 (0.989/0.683)	0.960 (0.981/0.750)
RWCP	0.985 (0.994/0.668)	0.984 (0.991/0.719)

現率が高くなった可能性がある．

5. 関連研究と考察

これまでに，単語分割や未知語処理に関して様々な研究が行われている．Xue[15] は，ME モデルを使用した文字タグ付け法により中国語の単語分割を行った．2 章で述べたように，この方法は既知語も未知語も区別することなく扱うことができる利点がある．4 章の実験結果では，この方法は未知語に対しては非常に高い再現率を持つものの，既知語に対しては他の手法と比較して再現率が低かった．

Nagata[7] は，文字種の違いを考慮した未知語の確率的モデルを定義することにより，日本語の単語分割と品詞付与を行った．EDR コーパスで我々とほぼ同量の訓練データを使用して実験を行った結果，単語分割の再現率/精度/F 値は，テストデータ全体に対して 0.946/0.937/0.941，未知語に対して 0.420/0.664/0.514 であったと報告している．実験の条件が同一ではないため単純には比較できないが，我々の提案手法は，それと同程度以上の解析精度を得ている．

Asahara ら[1),2)] は，コスト最小法とサポートベクタマシンを使用した文字タグ付け法を用いて単語分割を行った．この方法では，入力文をまずコスト最小法に基づく形態素解析器により解析し，その結果を素性に利用して新たに文字タグ付け法により解析を行う．我々とはアプローチが異なるが，文字単位と単語単位の両方の情報を利用して単語分割を行っているともいえる．AS，CTB，HK，PK コーパスにおける既知語/未知語の再現率はそれぞれ 0.952/0.574，0.949/0.412，0.980/0.415，0.975/0.357 であったと報告しており，全体的に既知語に対する再現率が高く未知語に対する再現率が低い結果を得ている．また Ng ら[9] は，Xue[15] と同様の文字タグ付け法により単語分割を行い SIGHAN Bakeoff のデータで評価した結果，AS，HK，PK の 3 つのデータで Bakeoff のどの参加者よりも高い F 値を得たと報告している．彼らは単語レベルの情報も含んだ素性を追加することで 0.4%精度が向上したと報告しており，文字ベースの手法に単語レベルの情報を導入するアプローチも有用であると思われる．

5.1 単語単位での未知語候補生成と文字単位での未知語候補生成

内元ら[22] は，最大エントロピーモデルに基づく形態素解析手法を提案した．彼らの方法は単語を単位として処理を行うが，コスト最小法のように直接辞書を用いて解の候補を生成するのではなく，既知語も未知語も同じ方法で扱っている．この方法では，任意の文字列に対して，それが形態素である確率を ME モデルを用いて推定する．文が入力されると，まず文中のすべての部分文字列に対してその確率を計算する．そして，あらゆる分割結果の候補の中から，分割結果の各部分文字列の確率を掛け合わせて計算される文全体の出現確率を最大化するような単語分割を解とする．彼らの手法は単語単位で未知語を処理しているため，「アルファベットで始まり数字で終わる単語」や「4 文字から構成される単語」というような単語レベルの素性を扱うことができる．一方で，我々の手法は文字単位で未知語候補を処理しているため，未知語に対してこのような素性を利用することは容易ではない．しかしながら，彼らの方法は計算量の問題があるように思われる．

n 文字から構成される文中に含まれる部分文字列の数は $n(n+1)/2$ 個あり，単語単位で未知語候補を生成する場合，原則としてこれらがすべて単語になる可能性を考慮する必要がある（**図 5**（左））．実際には，彼らは文中のあらゆる部分文字列を考慮するのではなく，5 文字以下の部分文字列，5 文字以上で辞書に登録されている語，および連続するカタカナからなる語のみを候補とするように制限を加えている．

これに対して，我々の手法や文字タグ付け法では文字単位で未知語候補の生成を行う．つまり，文中のすべての部分文字列に対してそれが未知語として出現する候補を明示的に作るのではなく，文字単位に分解された独立した候補により未知語を非明示的に扱う．未知語処理のためにラティス中に新たに追加するノードの数は $4n$ 個であり（図 5（右）），未知語の長さに関する制約はない．

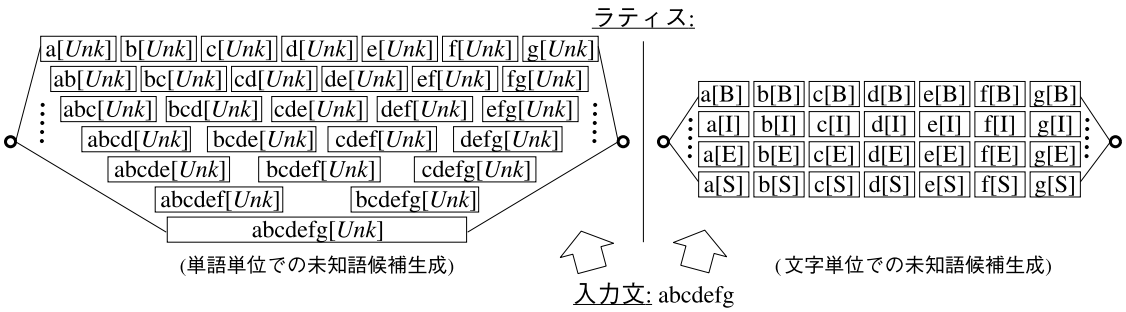

図 5 単語単位と文字単位で生成された未知語候補
Fig. 5 Generated unknown word candidates in a word-based approach and a character-based approach.

ヨチヨチ歩き（名詞），当たりはずれ（名詞），六フッ化ウラン（名詞），獅噛式三累環頭大刀（名詞），単純ヘルペスウイルス 1 型（名詞），低ピンチオフ電圧 FET 回路（名詞），駆けずりまわ（動詞），いたずらっぽ（形容詞），行き当たりばったり（形容動詞），根ほり葉ほり（副詞）

図 6 正しく同定された未知語の例（括弧内は正解コーパス中での品詞を表す）
Fig. 6 Examples of correctly identified unknown words.

ここで未知語の長さの制約に関する影響を調べるため，4.2 節で行った実験の結果を分析した．EDR コーパスを用いた実験において，提案手法により正しく同定された未知語の中で 6 文字以上であったものの例を**図 6** に示す．このような未知語は，未知語候補の長さを 5 文字以下とするような制約を課した場合は扱うことのできないものであり，このような未知語をある程度扱うことができることは提案手法の優位な点であると思われる．

次に，このような未知語の含まれる割合を調べた．京大コーパスのテストデータ中には 1,376 個の未知語があったが，そのうち 6 文字以上からなるカタカナの連続ではない未知語はわずかに 14 個であった（未知語全体の 1.0%）．EDR コーパスのテストデータ中には 52,105 個の未知語があったが，そのうち 6 文字以上からなるカタカナの連続ではない未知語は 8,195 個であり（未知語全体の 15.7%），提案手法ではそのうちの 2,071 個を正しく解析することができた．この 2,071 個の未知語が不正解だったと単純に仮定した場合の解析精度を計算したところ，正しく解析したときと比較して，再現率と精度のどちらも有意に（危険率 5%）低い値となった．京大コーパスの場合，5 文字以下の文字列と連続したカタカナを考慮すればほとんどの未知語をカバーできるが，EDR コーパスの場合はそれでは十分ではないと思われる．どの程度の長さの未知語が存在するかということは，コーパスが用いている形態素体系やコーパスの内容または言語に依存すると思われるが，提案手法や文字タグ付け法は，未知語の長さの上限をあらかじめ仮定する必要がないという点では一般性が高いと思われる．

中国語や日本語の単語分割における未知語処理の困難な点は，文中のどの位置からどれだけの長さの未知語が存在するかを予測するのが困難なことである．あらゆる未知語を単語単位で処理することを試みると計算速度が犠牲となるが，未知語の候補に制約を与えると処理できる未知語が制限される．1 つの極端な例は，しばしば用いられるような文字種に基づくヒューリスティックな規則を使用する場合である．たとえば，1 文字からなる漢字しか未知語の候補にしないという規則を使った場合は，2 文字以上の漢字からなる未知語を扱うことができない．このような点からは，文字タグ付け法や提案手法のように文字単位で未知語を処理する方法が有用であるように思われる．この文字単位の未知語処理は未知語を構成する文字を独立に扱う素朴な方法であり，単語の構成に関する詳しい情報は用いないが，実験結果によれば未知語の解析精度は比較的高いといえる．

5.2 提案手法における未知語のモデル

提案手法では，単語単位の候補と文字単位の候補を混在させて扱っているが，本手法では次のように未知語をモデル化していると解釈することができる．文中のあらゆる部分文字列が特殊な品詞 Unk を持つ単語として出現する可能性を調べて，この品詞 Unk を持つ単語により未知語の区切りを同定することができる．そして本手法では，長さが k の文字列 $w_i = c_j \cdots c_{j+k-1}$ が品詞 Unk を持つ単語として出現する確率を，次の

ように計算していると見なすことができる☆:

$$P(w_i\,Unk|h) = \begin{cases} P(c_j\mathbf{S}|h) & (k=1), \\ P(c_j\mathbf{B}|h)\prod_{l=j+1}^{j+k-2}P(c_l\mathbf{I}|h)P(c_{j+k-1}\mathbf{E}|h) & (k>1). \end{cases} \quad (7)$$

ここで，h はマルコフ過程の履歴である．つまり，品詞 Unk を持つ単語として扱われる未知語の出現確率は，それを構成する各文字の出現確率の積によってモデル化されていると解釈できる．これは，文字 trigram により未知語をモデル化した Nagata[6] のモデルと似ている．しかしながら，彼の方法は単語単位で未知語の候補を生成しており，前節で説明したような，内元らの手法と同様の利点と問題点を持つと思われる．提案手法では POC タグを用いて文字単位で未知語を処理することにより，この式 (7) の計算をコスト最小法の枠組みの中で効率的に行っている点が大きく異なる．

たとえば，「α/名詞 β/(未知語) γ/名詞」という単語/品詞の列が入力されたとして，α，β，γ は単語であり，β は a，b，c という 3 つの文字から構成される未知語だったと仮定する（β = abc）．マルコフモデルに基づく単語分割で，図 5（左）のように単語単位で未知語候補を生成する場合，この入力列の出現確率は，次のように計算することができる．

$$\begin{aligned} &P(\alpha, \text{名詞}, \beta, Unk, \gamma, \text{名詞}) \\ &\simeq P(\alpha, \text{名詞}\,|h)P(\beta, Unk|h)P(\gamma, \text{名詞}\,|h). \end{aligned} \quad (8)$$

一方で提案手法では，図 5（右）のように文字単位で未知語候補を生成して，次のようにこの確率を計算する．

$$\begin{aligned} &P(\alpha, \text{名詞}, \beta, Unk, \gamma, \text{名詞}) \\ &\simeq P(\alpha, \text{名詞}\,|h)P(\mathrm{a}, \mathbf{B}|h)P(\mathrm{b}, \mathbf{I}|h)P(\mathrm{c}, \mathbf{E}|h) \\ &\quad P(\gamma, \text{名詞}\,|h). \end{aligned} \quad (9)$$

この確率は，式 (3) を用いてコスト最小法の枠組みの中で計算される．つまり，提案手法では図 3 のように単語単位と文字単位の解候補を混在させて扱っており，式 (8) のように未知語に対する確率の計算を明示的に行うことはしていない．しかし，式 (7) で示されるように未知語をモデル化して，未知語に対する確率の計算を非明示的に行っていると解釈することができる．

☆ 提案手法では，w_i が辞書中に登録されているかどうかにかかわらずこの計算は行われることになる．そのため，辞書中に登録されている単語が品詞 Unk を持つ単語（未知語）として出力される可能性がある．

6. 結　　論

本稿では，既知語と未知語を高い精度で解析するために，単語レベルの情報と文字レベルの情報を利用する単語分割手法を提案した．この方法は，コスト最小法と文字タグ付け法を組み合わせて使用する．品詞タグと POC タグを同等に扱うことにより，既知語の候補と未知語を構成する文字の候補を統一的に扱い，既知語と未知語に対する単語分割を同時に行う．従来のコスト最小法では単語単位で解候補を生成しており，従来の文字タグ付け法では文字単位で解候補を生成していたが，提案手法では単語単位の解候補と文字単位の解候補を同時に扱う．中国語と日本語の複数のコーパスで単語分割の実験を行ったところ，既存の手法と比較して，高い F 値が得られることを確認した．本手法はコスト最小法に基づいているため，単語分割だけではなく既知語に対する品詞付与も同時に行えるが，単語区切りが同定された未知語に対する品詞付与は行われない．このような未知語に対して品詞を付与する方法として，英語における未知語の品詞推定で試みられているような多値分類器を用いて品詞の推定を行う方法[8] や，POC タグを文字の位置だけではなく品詞の種類の情報を持つようにして用いる方法[16),18)] 等が考えられるが，今後の課題として残されている．

謝辞　本研究は，情報通信研究機構平成 14 年度民間基盤技術研究促進制度に係る研究開発課題「多言語標準文書処理システムの研究開発」の一環として行われたものである．

参 考 文 献

1) Asahara, M., Goh, C., Wang, X. and Matsumoto, Y.: Combining Segmenter and Chunker for Chinese Word Segmentation, *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, pp.144–147 (2003).
2) Asahara, M. and Matsumoto, Y.: Japanese Unknown Word Identification by Character-based Chunking, *Proc. 20th International Conference on Computational Linguistics*, pp.459–465 (2004).
3) Brants, T.: TnT — A Statistical Part-of-Speech Tagger, *Proc. 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association of Computational Linguistics*, pp.224–231 (2000).
4) Darroch, J. and Ratcliff, D.: Generalized iterative scaling for log-linear models, *The annuals of Mathematical Statistics*, Vol.43, No.5,

pp.1470–1480 (1972).
5) Manning, C. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press (1999).
6) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. 15th International Conference on Computational Linguistcs*, pp.201–207 (1994).
7) Nagata, M.: A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context, *Proc. 27th Annual Meeting of the Association for Computational Linguistics*, pp.277–284 (1999).
8) Nakagawa, T., Kudoh, T. and Matsumoto, Y.: Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines, *Proc. 6th Natural Language Processing Pacific Rim Symposium* (2001).
9) Ng, H. and Low, J.: Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing*, pp.277–284 (2004).
10) Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *Proc. 3rd Workwhop on Very Large Corpora*, pp.88–94 (1995).
11) Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proc. 6th Workshop on Very Large Corpora*, pp.171–177 (1998).
12) Sproat, R. and Emerson, T.: The First International Chinese Word Segmentation Bakeoff, *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, pp.133–143 (2003).
13) Sproat, R., Shih, C., Gale, W. and Chang, N.: A Stochastic Finite-State Word-Segmentation Algorithm for Chinese, *Computational Linguistics*, Vol.22, No.3, pp.377–404 (1996).
14) Tjong Kim Sang, E.F. and Veenstra, J.: Representing Text Chunks, *Proc. 20th International Conference on Computational Linguistics*, pp.173–179 (1999).
15) Xue, N.: Chinese Word Segmentation as Character Tagging, *International Journal of Computational Linguistics and Chinese*, Vol.8, No.1, pp.29–48 (2003).
16) 山本幹雄，増山正和：品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析，言語処理学会第 3 回年次大会発表論文集，pp.421–424 (1997).
17) 新納浩幸：決定リストを弱学習器としたアダブーストによる日本語単語分割，自然言語処理，Vol.8, No.2, pp.3–18 (2001).
18) 新納浩幸：日本語形態素解析の分類問題への変換とその解法，情報処理学会論文誌，Vol.42, No.9, pp.2221–2228 (2001).
19) 永田昌明：未知語の確率モデルと単語の出現頻度の期待値に基づくテキストからの語彙獲得，情報処理学会論文誌，Vol.40, No.9, pp.3373–3386 (1999).
20) 森　信介，長尾　眞：n グラム統計によるコーパスからの未知語抽出，情報処理学会論文誌，Vol.39, No.7, pp.2093–2100 (1998).
21) 浅原正幸，松本裕治：形態素解析のための拡張統計モデル，情報処理学会論文誌，Vol.43, No.3, pp.685–695 (2002).
22) 内元清貴，関根　聡，井佐原均：最大エントロピーモデルに基づく形態素解析—未知語の問題の解決策，自然言語処理，Vol.8, No.1, pp.127–142 (2001).
23) 吉田辰巳，大竹清敬，山本和英：サポートベクトルマシンを用いた中国語解析実験，自然言語処理，Vol.10, No.1, pp.109–131 (2003).
24) 黒橋禎夫，長尾　眞：日本語形態素解析システム JUMAN version 3.61，京都大学大学院情報学研究科 (1998).
25) 松本裕治，北内　啓，山下達雄，平野善隆，松田　寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』version 2.2.8 使用説明書，奈良先端科学技術大学院大学松本研究室 (2001).
26) 松本裕治，浅原正幸：IPADIC ユーザズマニュアル version 2.2.4，奈良先端科学技術大学院大学松本研究室 (2001).
27) 小田裕樹，北　研二：PPM*モデルによる日本語単語分割，情報処理学会研究報告 98-NL-128, pp.9–16 (1998).
28) 小田裕樹，森　信介，北　研二：文字クラスモデルに基づく日本語単語分割，情報処理学会研究報告 99-NL-130，pp.1–8 (1999).

(平成 16 年 12 月 2 日受付)
(平成 17 年 9 月 2 日採録)

中川　哲治（正会員）

2000 年筑波大学第三学群情報学類卒業．2002 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年沖電気工業（株）入社，現在に至る．2004 年より奈良先端科学技術大学院大学情報科学研究科博士後期課程在学．統計的自然言語処理および機械学習に興味を持つ．

松本 裕治（正会員）

1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984〜1985 年英国インペリアルカレッジ客員研究員．1985〜1987 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．