社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－NL－162 (28)
2004／7／16

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
沖電気工業株式会社 研究開発本部
nakagawa378@oki.com

松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
matsu@is.naist.jp

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，一般に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精度が低い．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．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
Corporate Research and Development Center
Oki Electric Industry Co., Ltd.
nakagawa378@oki.com

Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
matsu@is.naist.jp

In this paper, we propose a hybrid method for Chinese and Japanese word segmentation which combines the Markov model-based method and the character tagging method. In general, word-based Markov models have difficulties in handling of unknown words and the character-based character tagging method performs worse than other methods for known words. In order to solve these problems, we combine the two methods. Experimental results of Chinese and Japanese word segmentation with multiple corpora showed that the method achieves high accuracy.

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一つである．このような分かち書きされない言語に対して単語分割を行う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．この単語分割をさらに難しくする要因として，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が無いため，正しく分割するのは非常に難しい．この未知語に対処する方法として，自動的な語彙獲得法により辞書の登録語を増やす試みが行われている [14, 13]．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て辞書に追加することで，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なうことが可能となる．しかしながら，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り，解析中に出現する未知語に対しても頑健に処理できる必要がある．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解析精度に優れており，文字ベースの単語分割手法は未知語の解析精度に優れているが既知語の解析精度が低いことが観測されている．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案する．これにより，既知語と未知語に対してバランスのとれた精度の高い解析を行える可能性がある．

以下，2節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節では提案手法の説明を行い，4節で実験結果を報告する．5節では従来研究との比較を行い，6節で結論を述べる．

2 従来研究

我々の提案手法は，既存の二つの単語分割手法に基づいているため，この節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 [18, 19]．コスト最小法は，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一つ前の品詞にのみ依存し，また単語が出現する確率はその単語の品詞にのみ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:

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

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):

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

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	一文字で単語を構成する文字

表 1: '**B**, **I**, **E**, **S**' タグ

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ー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みなすと，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，コスト最小法では次のように形態素解析が行われる:

1) 入力文に対して，単語辞書を使用して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) に対しては高い精度で解析を行うことができ，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一つの単語候補とし，漢字やひらがなは一文字だけで単語の候補にする」というようなルールを用いる方法である [18]．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 [9, 13]．しかしながらこのような方法は，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．

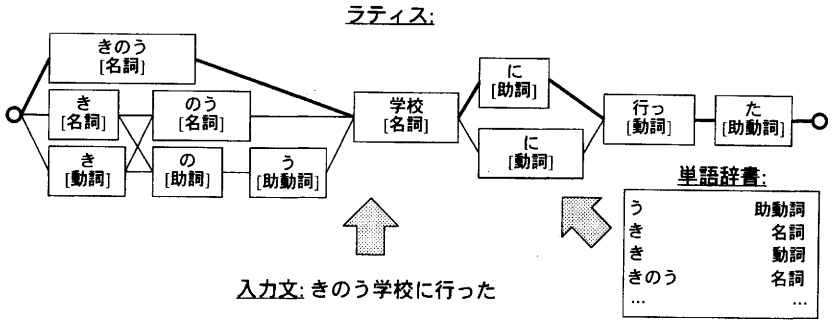

図 1: 解析結果の候補を表すラティス

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における位置を表すタグ (position-of-character タグ; POC タグ [10]) を付与する問題として解く．このような POC タグ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．そして，そのようなタグ付けは一般的な機械学習アルゴリズムを用いた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 [11, 22]，PPM*モデル [21]，決定リスト [12]，最大エントロピー (Maximum Entropy; ME) モデル [10]，サポートベクターマシーン [17, 15, 1] など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．

文字タグ付け法による単語分割は，既知語も未知語も区別せずに扱えるため，未知語に対する特別な処理を必要としない利点がある．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 [6] や固有表現抽出 [7] 等で広く利用されている．

3 単語レベルと文字レベルの情報を用いた単語分割

前節では，既存の二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精度は比較的高いものの未知語に対する精度が低く，また文字タグ付け法は未知語に対する精度は高いものの既知語に対する精度が低い事が実験結果から観測されている [17, 10, 8]．このような性質は，各手法の特徴を考えると妥当と思われる．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 bigram 確率等) を扱うことができ，また文中の広範囲の情報を利用して解析を行うことができる．しかしながら，未知語を直接処理することはできず，特に単純な規則等によって未知語処理を行った場合は十分な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，常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困難である．

以上のことを踏まえ，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

本手法はコスト最小法をベースにしているが，POC タグと品詞タグを同等に扱うことにより，既知語と未知語に対する単語分割を同時に行う．

図 3 に例を示す．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

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

図 2: 文字タグ付けによる単語分割

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のうちいくつかのタグ間の遷移 (**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.

(1) 式で実現される基本的なモデル (品詞 bigram モデル) はモデルの表現力が不十分なため, 次の式によりラティス中のパスの生成される確率をより精密に計算する:

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

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. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ の値については, 文献 [3] で使われている削除補間法により計算した. 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そのものから作成される. このような場合, 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. そこで上式のモデルのパラメータを学習させる際には, 訓練データ中に 1 回だけ出現した単語は未知語とみなして POC タグの付与された文字に分解して扱った.

様々な文字レベルの素性を利用するために, POC タグで条件づけられた単語 (文字) 出力確率に関してはベイズの定理により次の式で計算する:

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

ここで, w_i は文字で t_i は POC タグである. 上の式において, $P(t_i)$ 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 の文字種を表すとし, $w_i = c_{i'}$ とする. また, 使用した文字種は表 2 の通りである):

1	アルファベット
2	数字 (アラビア数字, 漢数字)
3	記号
4	漢字
5	ひらがな
6	カタカナ

表 2: 素性に使用した文字種

(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, 訓練データ中の全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.

本手法が未知語に対して行っている処理は, 次のように解釈することができる: この手法では, 文中のあらゆる部分文字列が未知語として出現する可能性を調べる. そして, 長さが k の文字列 $w_i = c_j \cdots c_{j+k-1}$ が未知語として出現する確率を次のように計算する:

$$\begin{aligned} & P(w_i t_i|h) \\ &= \begin{cases} P(c_j\mathbf{S}|h) & (k=1), \\ P(c_j\mathbf{B}|h)\prod_{l=j+1}^{j+k-2} P(c_l\mathbf{I}|h) P(c_{j+k-1}\mathbf{E}|h) & \\ & (k>1). \end{cases} \end{aligned} \tag{5}$$

ここで, h はマルコフ過程の履歴である. つまり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で近似している. これは, 文字 trigram により未知語のモデル化を行った Nagata[5] の方法と似ているが, 提案手法ではこの計算をコスト最小法の枠組みの中で文字単位で行っている点が大きく異なる.

4 実験

提案手法の有効性を評価するために, 中国語と日本語単語分割の実験を行った. 単語分割精度の評価

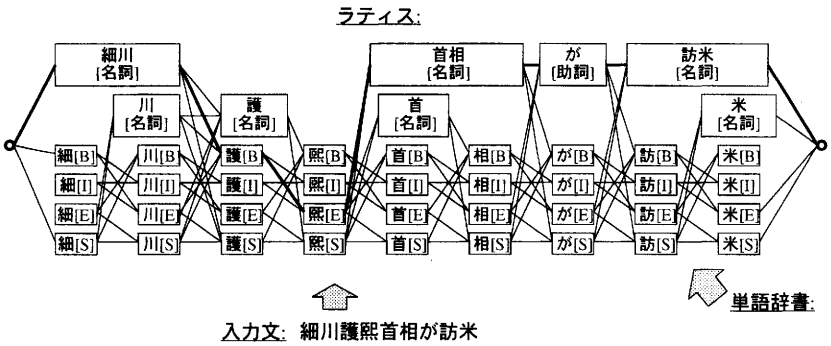

図 3: 提案手法による単語分割

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⟩ $\div$ ⟨正解データ中の単語数⟩)

P : 精度 ($P =$ ⟨解析結果中の正解単語数⟩ $\div$ ⟨解析結果中の単語数⟩)

F : F 値 ($F = 2 \times R \times P \div (R + P)$)

R_{known} : 既知語に対する再現率

$R_{unknown}$: 未知語に対する再現率

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．これらは全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[8] 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 アルゴリズムを利用して，品詞タグの代わりになる状態の付与を教師無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状態の数は 64 とした．

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[8] 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 [10] とほぼ同じであり，素性には 3.1 節で記述した (1)–(4) と次の (5) を使用して，ME モデルにより文字 $c_{i'}$ 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 $(t_{i'-1}, t_{i'-2}t_{i'-1})$

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，最長一致法や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全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 3 に示す．また，式 (3) 中の λ_i の値は，学習の結果表 4 のような値となった．

実験結果を表 5 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．既知語に対する再現率は他のシステムと比べてやや低いが，既知語と未知語の精度は比較的バランスががとれていた．最長一致法と文字タグ付け法の結果は，3 節で論じたトレードオフを反映していると思われる．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 と HK と 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 と HK と 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，京大コーパス version 2.0(KY)，RWCP コーパス (RWCP) の 2 つのコーパスを使用した．どちらも，単語分割済み

コーパス	訓練データの語数	テストデータの語数 (既知語/未知語)	辞書の登録語数	未知語の割合
AS	5,806,611	11,985 (11,727/ 258)	146,212	0.0215
CTB	250,841	39,922 (32,706/7,216)	43,544	0.1808
HK	239,852	34,955 (32,463/2,492)	23,747	0.0713
PK	1,121,017	17,194 (16,005/1,189)	55,226	0.0692
KY	198,514	31,302 (29,926/1,376)	1,983,156	0.0440
RWCP	840,879	93,155 (93,085/ 70)	315,602	0.0008

表 3: 使用したコーパス

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
KY	0.080	0.126	0.237	0.556
RWCP	0.073	0.105	0.252	0.571

表 4: 計算された λ_i の値

コーパス		R	P	F	R_{known}	$R_{unknown}$
AS	提案手法	**0.973**	**0.971**	**0.972**	0.979	0.717
	Bakeoff-1	0.966	0.956	0.961	**0.980**	0.364
	Bakeoff-2	0.961	0.958	0.959	0.966	0.729
	Bakeoff-3	0.944	0.945	0.945	0.952	0.574
	最長一致法	0.917	0.912	0.915	0.938	0.000
	文字タグ付け法	0.962	0.959	0.960	0.966	**0.744**
CTB	提案手法	0.877	0.872	0.875	0.928	0.647
	Bakeoff-1	0.886	**0.875**	**0.881**	0.927	**0.705**
	Bakeoff-2	**0.892**	0.856	0.874	0.947	0.644
	Bakeoff-3	0.867	0.797	0.831	**0.963**	0.431
	最長一致法	0.800	0.663	0.725	**0.963**	0.063
	文字タグ付け法	0.832	0.836	0.834	0.872	0.651
HK	提案手法	**0.951**	**0.948**	**0.950**	0.969	0.715
	Bakeoff-1	0.947	0.934	0.940	0.972	0.625
	Bakeoff-2	0.940	0.908	0.924	**0.980**	0.415
	Bakeoff-3	0.917	0.915	0.916	0.936	0.670
	最長一致法	0.908	0.830	0.867	0.975	0.037
	文字タグ付け法	0.917	0.917	0.917	0.932	**0.728**
PK	提案手法	0.957	**0.952**	**0.954**	0.970	0.774
	Bakeoff-1	**0.962**	0.940	0.951	**0.979**	0.724
	Bakeoff-2	0.955	0.938	0.947	0.976	0.680
	Bakeoff-3	0.955	0.938	0.946	0.977	0.647
	最長一致法	0.930	0.883	0.906	0.974	0.020
	文字タグ付け法	0.932	0.931	0.931	0.943	**0.786**

表 5: 中国語単語分割の実験結果

コーパス		R	P	F	R_{known}	$R_{unknown}$
KY	提案手法	0.986	0.984	0.985	0.989	0.9237
	JUMAN	**0.989**	**0.985**	**0.987**	**0.993**	0.8888
	最長一致法	0.806	0.753	0.779	0.843	0.0044
	文字タグ付け法	0.946	0.942	0.944	0.947	**0.9397**
RWCP	提案手法	**0.993**	**0.994**	**0.993**	**0.993**	0.586
	茶筌	0.991	0.992	0.991	0.991	0.243
	最長一致法	0.880	0.918	0.898	0.880	0.100
	文字タグ付け法	0.972	0.968	0.970	0.972	**0.629**

表 6: 日本語単語分割の実験結果

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) 拡張統計モデルに基づく日本語形態素解析システム [2, 19]．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 [18]．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 はあらかじめ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．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とし，1 月 9 日の 1 日分のデータ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には JUMAN version 3.61[18] 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，最長一致法と提案手法で使用する辞書には茶筌で使用される IPADIC version 2.4.4[20] を用いた．これらのデータの統計情報を表 3 に示す．式 (3) 中の λ_i の値は，学習の結果表 4 のような値となった．

実験結果を，表 6 に示す[1]．提案手法の結果を茶筌や JUMAN と比較した場合，F 値は大きな差が無かったが，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．最長一致法は，中国語での実験の場合に比べ既知語に対する再現率は低かったが，解析された結果を見ると助詞等の連続したひらがなに対する解析の誤りが多かった．

5 関連研究

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[10] 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 節で述べたように，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 節の実験結果では，この方法は未知語に対しては非常に高い精度を持つものの，既知語に対しては他の手法と比較して精度が低かった．

Asahara ら [15, 1] は，コスト最小法とサポートベクター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精度が高く未知語に対する精度が低い結果を得ている．

内元ら [16] は，最大エントロピーモデルに基づく形態素解析を提案した．こ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ことはせず，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，まず文中の全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「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が，我々の手法は文字単位で未知語を処理しているためこのような素性を扱うことはできない．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，原則としてこれらが全て単語になる可能性を考慮する必要がある．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と 5 文字以上で辞書に登録された語のみを候補とするように制限を加えている．我々の手法では，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，未知語の長さに関する制約は無い．また Nagata[5] は，コスト最小法における未知語処理として，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文

[1] この評価において，R_{known} と $R_{unknown}$ は，辞書に存在した単語を既知語として計算している．つまり，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計算し，辞書を用いずに訓練コーパスだけを用いる文字タグ付け法のスコアも辞書中に存在する単語は全て既知語として計算している．表 3 中の 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．

字 trigram を用いて計算しているが，これも単語単位で処理を行っているため内元らの手法と同様の利点と問題を持つと思われる．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が不明なことである．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．一つの極端な例は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，例えば一文字からなる漢字しか未知語の候補にしない場合は 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．そのような点では，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，未知語をより正確に解析するには単語の構成に関する知識の利用が必要不可欠にも感じられる．しかしながら実験結果から見る限りでは，現在のところ文字単位の未知語処理は他の手法と比較して比較的精度が高いといえる．

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した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．中国語と日本語の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 値を得られることを確認した．しかしながら，既知語に対する再現率は他の手法と比べて低い傾向があり，さらに分析が必要である．本手法は，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の推定は今後の課題である．

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. L., Wang, X. and Matsumoto, Y.: Combining Segmenter and Chunker for Chinese Word Segmentation, *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, pp. 144–147 (2003).

[2] Asahara, M. and Matsumoto, Y.: Extended Models and Tools for High-performance Part-of-Speech Tagger, *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 21–27 (2000).

[3] Brants, T.: TnT — A Statistical Part-of-Speech Tagger, *Proceedings of the 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] Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proceedings of the 15th International Conference on Computational Linguistcs*, pp. 201–207 (1994).

[6] Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *Proceedings of the 3rd Workwhop on Very Large Corpora*, pp. 88–94 (1995).

[7] Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 171–177 (1998).

[8] Sproat, R. and Emerson, T.: The First International Chinese Word Segmentation Bakeoff, *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp. 133–143 (2003).

[9] 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).

[10] Xue, N.: Chinese Word Segmentation as Character Tagging, *International Journal of Computational Linguistics and Chinese*, Vol. 8, No. 1, pp. 29–48 (2003).

[11] 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).

[12] 新納浩幸: 日本語形態素解析のクラス分類問題への変換とその解法, 情報処理学会研究報告 2000-NL-135, pp. 149–156 (2000).

[13] 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).

[14] 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).

[15] 浅原正幸, 松本裕治: 形態素解析とチャンキングの組み合わせによる日本語テキスト中の未知語出現箇所同定, 情報処理学会研究報告 2003-NL-154, pp. 47–54 (2003).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 松本裕治, 浅原正幸: IPADIC ユーザーズマニュアル version 2.2.4, 奈良先端科学技術大学院大学 松本研究室 (2001).

[21] 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).

[22] 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).