分かち書き学習データのための確率的処理の適用

福田　拓也† 　三浦　孝夫†

† 法政大学 工学研究科 電気工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †takuya.fukuda.2u@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　日本語には明示的な単語境界がない。そのため形態素解析や分かち書きなどによる処理が必要となる。しかし統一的な規則が存在せず慣習に負うところが大きく，あらかじめ規則を決定しておくことが容易ではない。筆者らはこれまで CRF (条件付確率場) を用いた高性能な機械学習方式を提案したが，ここでは領域固有の大規模なテストコーパスを必要とするという欠点がある。本稿では, マルコフ過程モンテカルロ法 (MCMC) に基づいて，領域依存コーパスを自動的に生成して CRF 学習に供し，高性能な分かち書きが構築できることを示す。

キーワード　分かち書き，テストコーパス，マルコフ過程モンテカルロ法

Probability Technique of Processing Training Data for Word Segmentation

Takuya FUKUDA† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †takuya.fukuda.2u@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　Unlike western languages, there exists no word boundary in Japanese. This is why we face to hard problems to analyze documents in Japanese very often. More difficulty arises in expertised domains such as medical, mechanical, computer science documents. Although there have been proposed morphological analysis and word segmentation so far, there is no rule in advance and the results depend on customs in each domain. The authors have proposed sophisticated techniques using Conditional Random Fields (CRF), a stochastic approach to obtain excellent results, but we should have much amount of training data (test corpus) for learning in advance in each domain. In this work, we discuss how to obtain pseudo test corpus based on Markov process Monte Carlo Method (MCMC), given small amount of test data. In this environment we show nice results using CRF based approach.

Key words　Word Segmentation, Conditional Random Fields, Markov Chain Monte Carlo (MCMC) method

1. 前　書　き

近年インターネットの急速な普及により, タグやコマンドなどの構造的手がかりを有さないテキストデータを検索解析する必要が増えている. テキストマイニングに対する近年の活発な研究では, アンケート等の自由記述データや営業日報・会議録等の，多種多様で膨大なテキストデータをどのように処理するかという問題を扱う。

テキストデータの解析で問題になるのが，日本語処理を行う際の分かち書き，即ち単語分割 (word segmentation) である。しかし, 英語やフランス語など言語とは異なって，単語間に空白を入れる合意がなく，単語の認識そのものが難しい。

これらの問題を解決するための数多くの研究が提案されており，おおよそ連接表アプローチと確率過程アプローチ大別できる。前者では，形態素間の関連を，品詞あるいは個別の語彙間の関連規則を連接表としてまとめ，実際の処理ではこれを参照する。後者は, 統計的手法により学習データを形態素解析して頻度を調べ，確率的に状態の並びを推定していく。しかしいずれの場合も，言語が一般的に利用される状況を想定しており，特定の問題領域に固有の状況にどのように対応するか，という問題に答えていない。状況に依存した解析手法についてはほとんど知られていない。

本研究では，日本語の領域依存分かち書き処理に関する実験的なアプローチとして, 条件付確率場 (CRF) による確率過程アプローチを用いる。これまでの多くの確率過程アプローチと同様に, 巨大な学習コーパスを用いることにより, よりよい CRF のモデルを得ることができる。しかし, 領域に固有で巨大なコーパスを構築するには, 時間やコストがかかり, 人手による作業のため主観に依存してしまう, エラーが発生する, という問題点がある。そこで, 本稿ではマルコフ連鎖モンテカルロ法

(MCMC) に基づく確率的処理により巨大なコーパスを生成させ, 生成したコーパスを分かち書きの学習コーパスとすることで特定の問題領域に固有の状況に対応した確率過程アプローチの精度向上を提案する。

筆者らはこれまで MCMC に基づくコーパス生成手法を提案している [9] が、ここでは隠れマルコフモデル（HMM）を想定した。本稿では CRF を用いることにより大幅な精度向上を得ることを示す。本論文の構成は次の通りである。第 2 章では形態素解析と分かち書きの特徴と処理を述べ，本研究で扱う問題の意義を明らかにする。第 3 章では計算機による分かち書きの処理と, 問題領域に依存して分かち書きが存在することを示す。続く 4 章で CRF を要約したあと，5 章で MCMC を述べる。これによる実験結果を 6 章で述べ,7 章で結びとする。

2. 文章と形態素解析

文書情報の大半は文章や図表などであり，文章は語の並びとして構成される。英語では (空白などの) 特殊文字で区切られた文字列を単語と呼ぶが, 複合語 (New York 等のように複数の単語からなる語) や共起性の強い語 (連語や慣用句) 等を考慮するかどうかは, 分析結果に大きな影響を与える。

自然言語において, 文章は形態素で構成されている。形態素とは，これ以上に細かくすると意味を失う最小の文字列に関する情報を言う。自然言語の文章に対して，単語 (トークン)，その語形変化 (語幹抽出や語尾変化) 品詞 (動詞, 名詞など) を持つ形態素の列に分解することができる。これを形態素解析と言う。形態素解析は，構文解析，意味解析，文脈理解などともに自然言語処理の基礎となる要素技術である。本研究では, 形態素とは単語と品詞の属性のペアを意味する。

形態素解析処理では，隣り合った形態素間の結合に関する規則を含む形態素辞書と，形態素に関する文法の知識を用いて，文を単語単位に分かち書きし，それぞれの構文上の役割を決定する。

英語やフランス語などは，語順や語形変化によって，性，数・格などの文法関係を表す言語を屈折語と言う。これに対して，中国語や日本語は助詞や助動詞などの付属語によって文法関係を示し，これを膠着語と言う。このため日本語形態素解析では, 形態素にとどまらず，形態素の格，数や性までを含めた単語の同定まで，すなわち文を文節まで含めて分割すること (文節分かち書き) が重要な課題となる。

実用上の観点から言えば，複合語への対応も大きな課題である。例えば複合名詞 (「法政大学」を「法政」と「大学」に分割) や複合動詞 (「乗り継ぐ」を「乗り」と「継ぐ」に分割) などの分割は領域固有に依存する問題である。例えば，著者らの所属する大学では,「情報電気電子工学科」は「情報電気電子」「工学科」としてのみ分かち書き可能である。

本研究では, 確率過程に基づく領域依存分かち書きの効率的なアプローチを提案する。本研究では, 日本語に N-グラムモデルを適用し, 分かち書きに対して, CRF によって形態素間の関係を調査する。

3. 分かち書きと領域規則

文節分かち書きと形態素解析は屈折語系では基本技術が共通しており, 文節ごとに開始タグや終了タグを挿入するため, タグ付け操作として位置付けることができる。

すでに述べたようにこれらの問題を解決するための数多くの研究が提案されており，おおよそ連接表アプローチと確率過程アプローチ大別できる。

連接表アプローチとは，語・品詞とその前後に生じる語・品詞とのパターンを手作業で検出し, これを規則 (連接表) として検出する手法である。連接表とは隣り合う形態素の結合に関する規則を列挙したものである。連接表は,"定冠詞のあとに動詞は生じない" などの規則からなる。うまく作られた連接表を利用したとき, 確率過程アプローチに匹敵することが知られている。しかし, 規則の生成には全域的な無矛盾性を必要とし, 自明な作業ではない。必要な規則を抽出できるが，手作業による規則抽出のため，手間がかかり正確性に欠けるという欠点がある。

これに対して確率アプローチでは確率判定 (Bayes) や確率過程 (HMM，MEMM，CRF) などを用いて確率的に規則を抽出する。ここでは状態をタグあるいはタグ列で表し, 観測された語の並びから状態列を推定する。学習データの分布情報から特徴量を抽出するため処理の汎用性が期待できる。連接表で欠点であった手間がかかり正確性に欠けるという欠点を確率アプローチを用いて自動化することにより改善する。

分かち書きに矛盾を含むことがある。"成田空港" は複合語として単一語になるが，"宮崎空港" は"宮崎" と "空港" に分割されることが多い。即ち, 現実の場では分かち書きは慣習的に用いられることが多い。問題領域によっては，該当分野に強く依存した分かち書き規則が用いられる。例えば農業分野の特許情報には "浮動位置" や"昇降アーム" といった農業器具に対する語が共起しており単一語とみなされることが多くなる。同様に器械分野の特許情報では"風洞実験" や "温度分布" など当該分野を象徴するものが多い。この問題は，Bayes 統計アプローチなどでは認識されているが，確率過程アプローチでは議論が無い。

これらの知識を学習データから自動抽出し，確率過程モデルで処理することにより，連接表生成と確率過程手法を組み合わせ，分かち書きによる推定を大幅に改善できるであろう。

本研究では,HMM よりも品詞情報を柔軟に取り込めるため, CRF 手法を用いる。

4. 条件付確率場

CRF は, マルコフ過程を拡張した識別モデルである。

マルコフモデルを用いて，ある観測系列 $X = \{x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n\}$ が与えられたとき，それに付随するラベル系列 $Y = \{y_0, ..., y_{i-1}, y_i, y_{i+1}, ..., y_n\}$ の条件付確率

$$P(Y|X) = \frac{1}{Z(X)} \exp\Big(\sum_i^n (\sum_a \lambda_a f_a(X_i, Y_i)) + (\sum_b \nu_b g_b(Y_{i-1}, Y_i))\Big)$$

$$Z(X) = \sum_{Y \in Tn} \exp\left(\sum_i^n (\sum_a \lambda_a f_a(X_i, Y_i)) + (\sum_b \nu_b g_b(Y_{i-1}, Y_i))\right)$$

を与えることにより, 最適なラベル系列を確率的に求めることができる。ここで, $Z(X)$ は全系列を考慮した正規化項, $f_a(X_i, Y_i)$ は二値の出力を持つ素性関数で, CRF ではこの素性関数を予め素性定義という形で定義することにより自動的に生成される。CRF は, 各単語のラベルを決定する際, 現在考えている単語の周辺の単語, もしくはその他の観測系列, たとえば品詞や細品詞などの出現有無を考慮しラベルを決定する。また, λ_a は素性関数に対する重みである。

CRF の素性関数に対する重み集合 $\Lambda = \lambda_1, ..., \lambda_a$ は, 学習データに対する対数尤度を最大にすることにより求められる。

$$\mathcal{L}_{\Lambda \cdot N} = \sum_u \log(P(Y^u | X^u))$$

$$= \sum_u \left[\log\left(\sum_Y \exp(\Lambda \cdot [F(Y^u, X^u) - F(Y, X^u)] + N \cdot [G(Y^u, X^u) - G(Y, X^u)])\right)\right]$$

$$= \sum_u [\Lambda \cdot F(Y^u, X^u) + N \cdot G(Y^u, X^u) - \log(Z(X^u))]$$

$$\hat{\Lambda} \cdot \hat{N} = \arg \max_{\Lambda \in R^A, N \in R^B} \mathcal{L}_{\Lambda \cdot N}$$

この問題を解く手法として, 一般的には *Generalized Iterative Scaling* (GIS), *Limited-memory quasi-Newton method* (L-BFGS) and *Preconditioned conjugate-gradient* (Precond.CG). などが用いられる。

また, 最適なラベルパスの探索には Viterbi アルゴリズムなどの動的プログラミング手法を用いて, 条件付確率 $P(Y|X)$ を最大化するようなラベル系列を一意に決定することができる。

これより, CRF ではパラメータ推定やラベルパスの探索において上記のアルゴリズムを用いて解決することができることから, CRF を使用するユーザ側からは, 素性の定義および学習データの作成のみを考慮することにより, 適切なラベル付けを行うことができる。本稿での観測系列は単語列, 品詞列, 細品詞列の 3 つをあわせたもの, 素性はそれぞれの列の前後 3gram, 品詞列, 細品詞列の前後 3gram と単語を見たものを定義している。

しかし, すべての系列に対する十分なデータが必要となるが, 十分に大きいデータを作るには人手, 時間がかかってしまう。つまり, 正しいラベル系列を全ての学習データに作る必要がある。本研究では, 次章の MCMC を用いることによりこの問題を解決する。

5. マルコフ連鎖モンテカルロ法

モンテカルロ法はさまざまな乱数を生成するアルゴリズムの総称である。一様乱数や正規乱数などは多くのソフトウェアで生成できるが, これらを用いて, 一般の確率分布 (確率密度関数 ρ) に従う乱数を生成する。代表的に, 棄却サンプリング (rejection sampling) を用いることが多い。しかし状態空間 A 上の確率分布が事前に判明することはまれであり, 出現頻度や経験値としてノンパラメトリックに与えられている場合, この手法を利用できない。

マルコフ連鎖モンテカルロ法 (Markov Chain Monte Carlo, MCMC) は所与分布に従う乱数を近似的に生成する手法である [8]。MCMC は実行時性能は悪いが, ギブスサンプラは簡易で高速に実行できることから頻繁に利用される。

状態空間 $A = \{1, .., N\}$ 上の長さ n の確率変数列 $X_1, ..., X_n$ に対して, これが状態列 $s_1...s_n$ となる確率 $P(X_1 = s_1, .., X_n = s_n)$ が定まるとき, 定常的 (stationary) という。一般には, 同じ状態列 $s_1...s_n$ であっても, これを状態値とする確率事象 $X_1 = s_1, .., X_n = s_n$ は, 一定の確率である (定常的) とは限らない。

定常の確率分布を持っていると仮定すると, 初期状態 s_0 から続く状態列 $s_0 s_1...s_n$ を得る定常確率 $P(X_0 = s_0, X_1 = s_1, .., X_n = s_n)$ が, 直前状態だけに依存した確率 $P(X_{n-1} = s_{n-1}, X_n = s_n)$ に一致するときマルコフ連鎖 (Markov Chain) 性を有するという。このとき, 状態 s_i から状態 s_j への遷移確率を p_{ij} とすれば, 状態遷移確率 $p = ((p_{ij}))$ を用いて $P(X_n|X_0 = s_0)$ は $p^n P(X_0 = s_0)$ で表せる。さらに, ある条件の下では, 確率変数の極限分布 $\lim_{n \to \infty} P(X_n)$ が存在することが知られる。

MCMC は, 状態遷移にマルコフ性を仮定し, 学習データなどの出現頻度を手がかりに, 極限分布に近似された乱数を生成するアルゴリズムであり, 十分に正確な定常確率を算出するために乱数 X_n を X_{n-1} から生成する手間を必要とする。

実際に MCMC 法を用いて乱数を生成する場合, 各状態は $\{1, .., N\}$ 上のベクトル $x_k = (x_1^{(k)}, .., x_m^{(k)})$ であることが多く, MCMC 法を素直に適用するためには膨大な時間と記憶域を必要とする [8]。そのために MCMC 法を更に準用し, ベクトル要素 $x_i^{(k+1)}$ を個々に生成するとき $P(x_i^{(k)}|x_1^{(k+1)}, .., x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, .., x_m^{(k)})$ を用いる手法をギブスサンプラという。この手法によって生成された乱数状態列もマルコフ性が保証される。

以下はギブスサンプラーのアルゴリズムである未知の母数 $x = (x_1, ..., x_m)$, 観測されたデータ Y とする。

（1） 乱数 m を選ぶ

（2） 初期値 $(x_1^{(0)}, x_2^{(0)}, ..., x_m^{(0)}, Y)$ 適当に発生させる。

（3） $x_1^{(1)}$ を分布 $\rho(x_1|x_2^{(0)}, ..., x_m^{(0)}, Y)$ から発生させる。

（4） $x_2^{(1)}$ を分布 $\rho(x_2|x_1^{(1)}, x_3^{(0)}, ..., x_m^{(0)}, Y)$ から発生させる。

（5） 上記を繰り返し $x^{(k)}$ を得る。

ここで $x^{(k)}$ の極限値 $\lim_{k \to \infty} x^{(k)}$ は極限分布 $\rho(s|Y)$ のサンプルとなる。この過程は定常状態になるまで繰り返されるべきであるが, 実際には k は有限である。

CRF では学習データを用いてパラメータの対数尤度の最大化を行うため, MCMC に基づく多量の学習データを生成する。

まず, 観測データとして少量のコーパスを形態素解析する。各形態素は単語とタグで構成される。ここでタグとは B(分かち書き開始), または I(分かち書き途中) である。次に形態素の

頻度を調査する。例えば,

”私は犬を見る”

という文に形態素解析を適用し

”私”(代名詞), ”は”(係助詞), ”犬”(名詞), ”を”(格助詞), ”見る”(動詞)

を得る。ここで分かち書き結果を 2 種類のタグ B, I を用いて

”私 (B)”, ”は (I)”, ”犬 (B)” ”を (I)”, ”見る (B)”

とする。これは

/私 は/犬 を/見る/

を表している。

ここに出現する単語とタグを組として乱数を対応させ, 各組の出現頻度の分布を得る。このようにして初期学習データの分布を得たあと, 一様乱数で得た初期値から開始して, 長さ m の文 s を単語列ベクトル $s = (s_1, .., s_m)$ を生成する。また, この方法を用いることで繰り返し回数を調整することにより, 必要な分だけの学習データが生成可能となる。ただし定常状態になるまで生成前の繰り返しが必要となる。

ギブスサンプラによって生成される乱数列が表す文は, 何ら意味を有するものではない。しかし, 初期学習データには, 各語の出現頻度だけではなくて, 連続する語の共起性も表現している。このため, サンプラが生成した文は部分的には整合した内容を表すことが多い。

例えば, 後述する実験データ「公開特許公報全文データ」から次のような文を得る。

/水槽 部 内/【 は /この /軸 に/風洞 た/処理 れる は/

このようにして拡張した学習データを用いて,CRF モデルを生成する。

6. 実　　験

本章では, 本研究で提案する MCMC を用いた CRF に基づく分かち書きの有用性について検証する。本研究では, 観測データとして「公開特許公報全文データ (98, 99)」から農業領域, 器械領域に関する形態素を検証する。

6.1 実験準備

本実験では「公開特許公報全文データ (98, 99)」から，農業領域データ 859 形態素 および器械領域データから 1030 形態素を学習データとして使用する。またテストデータとして，同様に「公開特許公報全文データ (98, 99)」から，農業領域データ 3886 形態素および器械領域データ 3569 形態素を用いる。テストデータの正解, 学習データはそれぞれ人手でタグ付けを行っている。また, 学習データ, テストデータの形態素解析に形態素解析器として「Chasen」を使用する。

ここで,MCMC が有用に働くことを検証するため,2 種類の実験を行う。まず, 拡張した形態素数での精度比較をする。学習データと同じ領域のテストデータを使い, 学習データを約 1000 形態素ずつ増やし比較し, 拡張した学習データの有用性を示す。次に, MCMC で学習データを生成した HMM, MCMC で学習データを生成した CRF の精度比較を行い CRF の有用性を示す。

評価方法として精度 (正解率) を用いる。人手でタグ付けしたテストデータと比較しタグの正解数を出し。

$$\frac{\text{正解数}}{\text{テストデータの全形態素数}} \times 100 \qquad (\%) \qquad (1)$$

を分かち書きの精度とする。

6.2 実験結果

まず, 学習データ量での精度比較をする。ここでは, ギブスサンプラーによって生成される学習データを使用し, CRF と HMM の精度を示す。

表 1 は農業領域学習データを用いて分かち書きした結果である。約 6000 形態素の時に精度が 89.77%で最大になっている。5000 から 6000 の間で精度が 1.80%向上している。

表 2 は器械領域学習データを用いて分かち書きした結果である。若干の精度向上は見られるがほぼ精度の変化は見られない。

また, 農業領域での学習データ 5000 から 6000 での分かち書きの不正解の詳細を比較する。種類数で比較すると, 84 種類から 83 種類とほぼ変化がなかった。不正解数が減った単語を見てみると, 「アーム」,「要素」など, その領域で使われる単語の一部であるようなものが出てきていた。

さらに, 表 1, 表 2 中に HMM による結果を示す。

表 1 は農業領域テストデータを分かち書きしたときの比較である。HMM では, 約 6000 形態素の時に精度が 74.70%で最大になっている。CRF では 6000 形態素を用いて 89.77% と, 15.07% 精度がよい。同様に, 表 2 は器械領域テストデータを分かち書きしたときの比較である。HMM では, 約 4000 形態素を用いて 78.17% なのに対し, CRF では 2000 形態素を用いて 93.85% と, 15.68% 精度がよい。

表 3, 表 4 に不正解となった形態素の詳細を示す。CRF による結果の詳細をみると,HMM に比べ「の」「を」などの頻出形態素の不正解数が少ない。一方で, 学習データに依存する名詞の不正解数にはほとんど変化が見られない。

学習データ (形態素)	CRF(%)	HMM(%)
1003	87.58	71.23
2001	87.22	71.08
3009	87.05	70.84
4019	87.99	70.51
5010	87.97	71.02
6018	89.77	74.70
7018	89.46	74.60

表 1　農業領域学習データ

6.3 考　　察

学習データ量での比較実験では, 器械領域データにはあまり違いは見られなかったが, 農業領域データでは 6000 形態素を用いることで精度の向上を示した。これは領域に依存したよりよい結果を得るために, 十分なデータ量を必要とすることを意味

学習データ (形態素)	CRF (%)	HMM (%)
1004	93.43	77.95
2002	93.85	77.58
2999	93.80	77.50
4010	93.77	78.17
4996	93.80	77.58
6019	93.82	77.56
6999	93.69	77.61

表 2　器械領域学習データ

する。また, 農業領域データでは, 5000 形態素から 6000 形態素へ学習コーパスを増やしたときにほぼ学習コーパスの単語分布に変化は見られなかった。しかし, 「アーム」,「棒」などのような単語に対して B, I のタグの分布が近づいていた。そのことから, 精度向上につながったと考えられる。MCMC で約 1000 形態素から分かち書きに必要な量の学習データを生成した。これにより MCMC が分かち書きにとって有効であることを示した

CRF		
形態素	タグ	不正解数
要素	B	37
方向	B	29
棒	B	27
環	B	18
ピン	B	15
込	B	15
腔	B	13
アーム	B	11
ヨーク	B	11
位置	B	10

HMM		
形態素	タグ	不正解数
要素	B	62
)	B	41
,	B	37
。	B	35
に	B	32
を	B	31
方向	B	30
内	B	26
の	B	25
棒	B	19

表 3　農業領域データでの詳細

CRF		
形態素	タグ	不正解数
装置	B	34
b	I	20
輻射熱	I	13
性能	I	8
発明	B	6
項	I	6
空気	I	5
供給	I	5
制御	I	5
特性	I	5

HMM		
形態素	タグ	不正解数
装置	B	56
温度	I	33
ヒーター	I	31
成層	B	26
体	B	19
a	B	18
風洞	B	17
形成	I	13
発熱	I	11
輻射熱	I	11

表 4　器械領域データでの詳細

次に, HMM との比較して CRF では 約 15%高い精度が得られた。不正解の詳細を見ると, HMM では農業領域, 器械領域共に領域依存性の高くない (「の」「を」などの) 形態素が多く生じている。また, 農業領域, 器械領域共に領域依存性の高い単語も若干多く発生している。例えばたとえば「を」などは農業テストデータでは 31 回, 器械テストデータでは 10 回, 不正解が生じる。CRF では「を」の不正解数が 農業テストデータ, 器械テストデータ共に生じない。CRF では, HMM より柔軟に素性を取り込めるため, これらの形態素が正しく分かち書きされることにより, HMM より高い精度を実現した。つまり, HMM では前状態から推定状態への遷移確率と推定状態から出力される観測系列への出力確率の積で決定されるが, 今回の手法では出力 (単語) が 200 種以上なのに対し, 状態数は 2 と非常に少ない, そのため遷移確率の比重が大きくなってしまっている。一方 CRF では, 素性 (単語, 品詞, 細品詞の前後 3gram) に依存して状態 (タグ) が決定するため, 頻出単語では CRF のほうが高い精度を示す。また, MCMC で拡張したデータを学習しても 2 つのアプローチともに精度が落ちない。このことより 本論文のモデルに対しては MCMC を用いたデータの拡張を CRF に学習させることは有用である。

7. 結　　論

本研究では, マルコフ連鎖モンテカルロ法用いた分かち書きのための学習データの確率的処理を提案した。また, この手法の有効性を実験によって示した。本研究では,CRF,HMM のための MCMC アプローチについて述べてきた。しかし, 他の確率的なアプローチに対し,MCMC アプローチを適用可能である。さらに, 本研究で用いたタグは B,I の 2 種類だったが, 新しいタグを増やし適用することも可能である。

文　　献

[1] Abney, S.: Part of Speech Tagging and Partial Parsing, In *Corpus-Based Methods in Language and Speech*, Kluwer Academic Publishers, 1996

[2] Fukuda, T., Izumi, M. and Miura, T.: Word Segmentation using Domain Knowledge Based On Conditional Random Fields, proc. *Tools with Artificial Intelligence* (ICTAI), pp.436-439, 2007

[3] Gelfond, A.E. and Smith, A.F.M.: Sampling-based Approaches to Calculating Marginal Densities, *J. of the American Stat. Assoc.* Vol.85, pp.398-409, 1990

[4] Igarashi, H. and Takaoka, Y. Japanese into Braille Translating for the Internet with ChaSen proc.18th *JCMI*, 2K6-2, 1998

[5] Kita, K.: Probabilistic Language Model, Univ. of Tokyo Press, 1999 (in Japanese)

[6] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random Fields to Japanese morphological analysis, proc. *EMNLP*, 2004

[7] Mitchell, T.: Machine Learning, McGraw Hill Companies, 1997

[8] Ohmori, Y.: Recent Trends in Markov Chain Monte Carlo Methods, *J.of the Japan. Stat.Assoc.* , Vol.31, pp.305-344, 2001 (in Japanese)

[9] Fukuda, T. and Miura, T.: Word Segmentation Based on Hidden Markov Model Using Markov Chain Monte Carlo Method, *Journal of the DBSJ* , Vol.7, pp.73-78, 2008 (in Japanese)