モンテカルロ法を用いた隠れマルコフモデルに基づく分かち書き

Word Segmentation Based on Hidden Markov Model Using Markov Chain Monte Carlo Method

福田 拓也♠ 三浦 孝夫♠
Takuya FUKUDA Takao MIURA

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追うところが大きい。筆者らはこれまでCRF（条件確率場）を用いた高性能な方式を提案してきた。しかし，学習に大量の計算時間を必要とし，また領域固有の大規模なテストコーパスを必要とした。本稿では，マルコフ過程モンテカルロ法(MCMC) に基づいた領域依存コーパスを自動的に生成し，マルコフ過程モデルを用いた分かち書きを高性能に学習させることができることを示す。

It is well-known that Japanese has no word boundary, so that we should think about how to separate each sentence into words by means of morphological analysis or some other word segmentation analysis. It is said, however, that the separation depends on domain specific rules. The author have proposed a sophisticated word separation method based on Conditional Random Fields (CRF). Unfortunately we need a huge amount of test corpus in application domains as well as computation time for learning. In this investigation, we propose a new approach to obtain test corpus based on Markov Chain Monte Carlo (MCMC) method, by which we can obtain efficient Markov model for segmentation.

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。

これまで数多くの研究が提案されており，これらはおおよそ連接表アプローチと確率過程アプローチに大別できる。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。正確で実体に即した経験則を表現することができるが，知識それ自体の抽出が容易ではなく，また確立した手順も無い。確率過程アプローチで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，隠れマルコフモデルなどにより形態素同士の結びつきからなる状態の並びを推定する。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。反面，前提となる確率過程モデルが学習データと対応するとは限らず，適用範囲が限られる可能性がある。

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を提案する。本稿では，隠れマルコフモデル(HMM) による確率過程アプローチを提案する。ここで,HMMのモデルを得るために学習コーパスを用いる。多くのコーパスを用いるほどよりよいモデルとなる。しかし，巨大なコーパスを作るには多くの時間がかかる。問題の解決にBaum-Welchアルゴリズムがあるが，結果はコーパスに依存してしまう。そこで，マルコフ連鎖モンテカルロ法(MCMC) に基づくHMMによるアプローチを提案する。コーパスと同様の分布から乱数を発生させることにより，巨大な学習コーパスを発生させ，発生させた学習コーパスを用いて頻度を求めることにより，HMMでの遷移確率，シンボル出現確率とすることで，領域依存性を有する分かち書き機構を提案する。

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HMM を要約したあと，5 章でMCMC，6 章でMCMC による分かち書きを述べる。これによる実験結果を7 章で述べ,8 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。

自然言語の文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(単語) 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。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。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。

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。

♠ 学生会員 法政大学大学院工学研究科修士課程 takuya.fukuda.2u@gs-eng.hosei.ac.jp
♠ 正会員 法政大学工学部情報電気電子工学科 miurat@k.hosei.ac.jp

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，隠れマルコフモデル(HMM)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する。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。

ここには大きく2つのアプローチがある。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。

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。連接表には，”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，隠れマルコフモデル(HMM) などの確率過程手法を用いる。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。確率アプローチとは確率判定(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とである。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。

分かち書きに一貫性はないことが多い．”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。しかし，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後処理されるほうが多い。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,CRF などよりも計算が複雑にならず速度が速いため，確率過程の中でもHMM を用いる。

4. 隠れマルコフモデル

本研究で用いる隠れマルコフモデル[5] とは，状態遷移確率とシンボル出力確率を用いた単純マルコフに基づくオートマトンである。HMM は状態遷移確率，シンボル出力確率，初期状確率の分布，状態の有限集合，シンボルの有限集合によって定義される。これら5つをまとめてモデルMとよぶ。

状態遷移確率 a_{ij} とは単純マルコフ過程において，ある状態 i からある状態 j へ移る確率であり，$\sum_j a_{ij} = 1$ をみたす。また，シンボル出力確率 $b_i(o_t)$ とはある状態 i において，あるシンボル o_t を出力する確率であり，$\sum_t b_i(o_t) = 1$ をみたす。さらに，初期状態が i である確率を初期状態確率 π_i とよぶ。

これらを与えることにより，あるシンボル列Oが観測されたとき，最適な状態遷移列を確率的に求めることができる。HMM ではモデルが非観測であるため，シンボル列Oが観測されたとき，モデルMからOが出力される確率 $P(M \mid O)$ が最大になるようなMを探さなければいけない．しかし,HMMのモデルM中の状態遷移確率 a_{ij} , シンボル出力確率 $b_i(o_t)$, 初期状態確率 π_i を明確に決定するのは難しい。この問題をHMMのモデル計算と呼ぶ。通常はこの問題に対し機械学習を用いる。

アプローチのうちの１つは教師あり学習である。このアプローチではモデルを計算するために学習データを用いる。しかし，特徴的な規則を抽出するためには，データを手作業で正確に作る必要がある。もう一方のアプローチは教師なし学習である。学習データではなく，多くの未分類のデータを用い，一旦分類されたデータと未分類のデータの間の類似点を得ることができれば，期待値最大化(EM) を用いて学習データを拡張可能である。

教師なし学習で有名なアプローチの１つにBaum-Welch アルゴリズムがある。教師なし学習データの生成確率を最大にするため，パラメータを何度も調節する。この過程はEM 計算と同様である。Baum-Welch アルゴリズムは次のようなアルゴリムを繰り返し行うことによってモデルを推定する。ここで状態 i から状態 j に時刻 t で遷移する確率を $\gamma_t(i,j)$ とする。時刻 t で状態 i にとどまる確率 $\gamma_t(i)$ は

$$\gamma_t(i) = \sum_{j=1}^{N} \gamma_t(i,j)$$

となる。Nは全状態数である。これら2つの確率を用いて

$$\pi_i = \gamma_1(i)$$

$$a_{ij} = \frac{\text{状態}i\text{から状態}j\text{へ遷移する回数の期待値}}{\text{状態}i\text{から遷移する回数の期待値}}$$

$$b_i(o_t) = \frac{\text{状態}i\text{にとどまりシンボル}o_t\text{を出力する回数の期待値}}{\text{状態}i\text{にとどまる回数の期待値}}$$

を使いモデルを再推定する。これらは変化が起こらなくなるまで繰り返す。

また，シンボル列と最適なモデルが与えられたとき，Viterbi アルゴリズムを用いて最適な状態遷移列を求めることができる。

Baum-Welch アルゴリズムは少量の学習データでも適用が容易である。しかし，初期値に依存しているため実情とは不一致になってしまし精度が悪くなる場合がある。

一方，学習データを用いることによりモデルを推定可能

である。このアプローチではモデルを計算し，かつ状態と出力シンボル両方の列を調査するために十分な学習データを用いる。さらに，学習データの頻度を数え，絶対値を確率と見なす。

π_i ＝シンボル列の始まりが状態iの回数

$$a_{ij} = \frac{\text{状態}i\text{から状態}j\text{へ遷移する回数}}{\text{状態}i\text{から遷移する回数}}$$

$$b_i(o_t) = \frac{\text{状態}i\text{にとどまりシンボル}o_t\text{を出力する回数}}{\text{状態}i\text{にとどまる回数}}$$

をモデルとする。 例えば状態 i から j へ遷移する回数を学習データから検出し，その相対頻度を遷移確率として与える。このことにより実態の分布を反映でき，高精度が期待できる。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，状態列が隠れ状態であるため，正しい状態列を作る必要がある。これは自明な作業ではないため人手と時間をかける必要がある。

本研究では，この問題を解決するために次章のマルコフ連鎖モンテカルロ法を使用する。

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 $P(X_n \mid X_0=s_0)$ は $p^nP(X_0=s_0)$ 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 $P(x_i^{(k+1)} \mid x_1^{(k+1)},...,x_{i-1}^{(k+1)},x_1^{(k)},...,x_m^{(k)})$ 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 $(x_1^{(0)},x_2^{(0)},...,x_m^{(0)})$ を適当に発生させる。

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

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

(5). $x_m^{(1)}$ まで発生させ $x^{(1)}$ とする。

(6). 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 $\rho(s \mid Y)$ 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。

6. MCMC に基づく分かち書き

この章では,MCMC に基づく分かち書き用学習データの生成方法について述べる。HMM では学習データの状態遷移および出力シンボルの頻度を計算することにより，モデルを直接計算するため，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初期学習データを用いて,HMM モデルを生成する。

7. 実験

本章では，本研究で提案するMCMC を用いたHMM 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。

7.1 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,3 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,Baum-Welch アルゴリズムとの精度比較を行いMCMCの有用性を示す。さらに，条件付確率場との比較で本研究の有用性を示す。

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，

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

を分かち書きの精度とする。

7.1 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，ギブスサンプラによって生成される学習データの頻度を使用し，HMMモデルを計算する。

表1 農業領域学習データ
Table 1 Ratios in Agriculture Data

学習データ(形態素)	MCMC(%)	BW(%)
997	71.23	49.69
2016	71.08	48.79
3019	70.84	50.81
4002	70.51	50.64
5007	71.02	50.15
6004	74.70	50.59
7022	74.60	50.49

表2 器械領域学習データ
Table 2 Ratios in Instrument Data

学習データ(形態素)	MCMC(%)	BW(%)
1004	77.95	49.23
2002	77.58	48.42
2999	77.50	50.43
4010	78.17	47.69
4996	77.58	49.17
6019	77.56	50.83
6999	77.61	48.98

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74.70%で最大になっている。5000 から6000 の間で精度が3.68%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。約4000形態素のあたりで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と5000 形態素では「ハウジング」や「腔」など農業領域に関する語句と思われるものが生じる。また，不正解形態素の種類数でも5000 形態素では254 種類なのに対し，6000 形態素では206 種類と少なくなっている。器械領域での不正解の詳細にはほぼ変化が見られない。

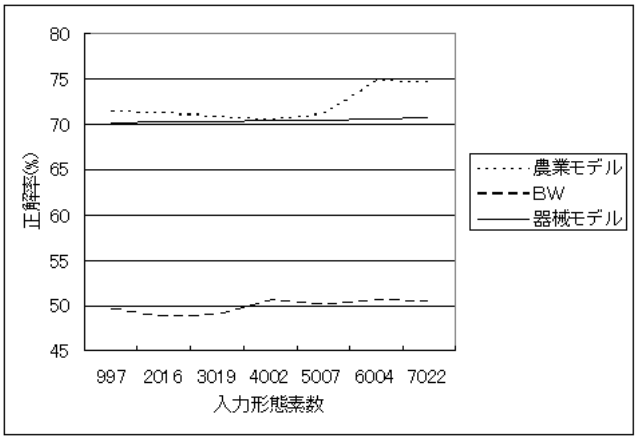

図1 農業領域テストデータ
Fig.1 Ratios in Agriculture Data

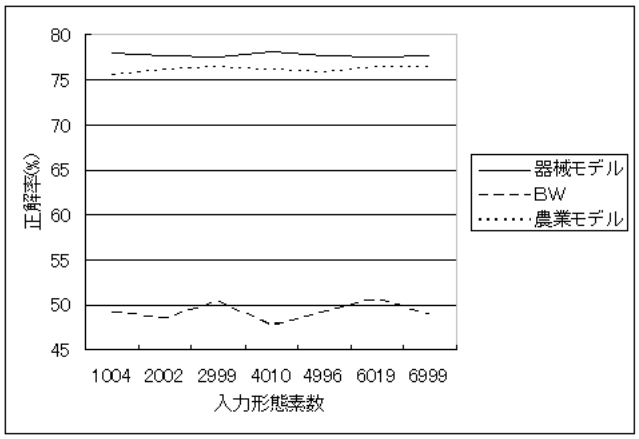

図2 器械領域テストデータ
Fig.2 Ratios in Instrument Data

さらに，表1，表2 中にBaum-Welch アルゴリズムによる結果を示す。

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。Baum Welch アルゴリズムでは4000 形態素を用いて50.64% なのに対し，MCMC アプローチでは6000 形態素を用いて74.70% と，24.04% 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。Baum-Welch アルゴリズムでは6000 形態素を用いて50.83% なのに対し，MCMC アプローチでは4000 形態素を用いて78.17% と,27.34% 精度がよい。

また,図1,図2はそれぞれの学習モデルを使い,農業・器械領域のテストデータを分かち書きした結果をグラフにしたものである。

表 3 農業領域データでの詳細
Table 3 Incorrect Data in Agriculture

MCMC		
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

Baum-Welch		
形態素	タグ	不正解数
，	B	190
て	B	91
要素	B	73
）	B	66
を	B	59
。	B	55
】	B	42
【	I	41
が	B	41
ロッド	I	40

表3，表4 に不正解となった形態素の詳細を示す。Baum-Welch アルゴリズムによる結果の詳細をみると，「の」「を」などの頻出形態素が不正解として上位に多く出てきている。一方で,MCMC アプローチでは学習データに依存する名詞が出てきている。

次に条件付確率場(CRF) を用いた分かち書き[2] との比較をする。

表 5，表 6 では約 13000 形態素を用いて学習した CRF と MCMC アプローチの最もよい結果との比較結果を示す。農業領域データでは MCMC アプローチより CRF のほうが 21.90% 精度がよい。器械領域データでは MCMC アプローチより CRF のほうが 19.95%精度がよい。

表 4 器械領域データでの詳細
Table 4 Incorrect Data in Instrument

MCMC		
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

Baum-Welch		
形態素	タグ	不正解数
の	B	145
を	B	84
さ	I	59
ヒーター	I	47
温度	I	44
し	I	42
。	B	34
】	B	32
輻射熱	I	29
が	B	29

表 5 CRF との精度比較（農業領域）
Table 5 Comparison in Agriculture

手法	正解率(%)
MCMC(6000)	74.70
CRF	96.60

表 6 CRF との精度比較（器械領域）
Table 6 Comparison in Instrument

手法	正解率(%)
MCMC(4000)	78.17
CRF	98.12

7.3 考察

まず，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する。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と示した。

次に,Baum-Welch アルゴリズムと比較して，MCMC アプローチでは25% を超える精度が得られた。不正解の詳細でBaum-Welch アルゴリズムの結果には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。たとえば「を」などは農業テストデータでは59 回，器械テストデータでは84回，不正解が生じる。MCMC アプローチでは「を」の不正解数が農業テストデータでは31 回，器械テストデータでは10 回生じる。MCMC アプローチではそれらの形態素を正しく分かち書きすることによってBaum-Welch アルゴリズムより精度向上を実現した。このことより本論文のモデルではBaum-Welchアルゴリズムより MCMC アプローチのほうが有用である。

最後にCRF との結果と比較する。MCMC アプローチはCRF アプローチに比べ20% 以上精度が悪い。しかし，これは約13000 形態素の学習データを用いたときに出された結果である。CRF は多量の正しい学習データを用いなければ十分な精度を得る事ができない[2]。本研究の提案手法では約1000 形態素の学習データを用いている。提案手法では精度はCRF に劣っているが学習データの量，つまり人手と時間を短縮し分かち書きすることができ，非常に効率的である。

8. 結論

本研究では，マルコフ連鎖モンテカルロ法用いた隠れマルコフモデルによる分かち書きを提案した。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,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 Approachesto 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)

福田 拓也　Takuya FUKUDA
法政大学大学院工学研究科修士課程在学中。

三浦 孝夫　Takao MIURA
法政大学工学部情報電気電子工学科教授。データモデル，知識表現，演繹データベース，複合オブジェクトなどの分野の研究に従事。