DEIM Forum 2010 F7-2

# 学術論文文書画像からの自動書誌要素抽出

井上　諒平† 　太田　学†† 　高須　淳宏†††

† 岡山大学工学部　〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 岡山大学大学院自然科学研究科　〒 700–8530 岡山県岡山市北区津島中 3–1–1
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †inouer@de.cs.okayama-u.ac.jp, ††ohta@de.cs.okayama-u.ca.jp, †††takasu@nii.ac.jp

**あらまし**　今日では国内の主要な学術論文を網羅した電子図書館が構築されており，その蔵書検索の際には著者名等の書誌情報を利用することが通例となっている．しかし文書画像を扱う電子図書館では、データベースへの書誌情報の入力に大きな人的コストがかかる．そこで本研究では，学術論文の文書画像を OCR 認識したテキストデータから，自動で書誌要素を抽出する手法を提案する．提案手法は，OCR の画像処理によって得られた各行に対して，文書レイアウト等の視覚的な情報，および文字列などの言語的な情報から書誌要素の判別を行い，書誌ラベルを付与する．その際のラベル付けには Conditional Random Fields（CRF）を使用する．

**キーワード**　電子図書館, 文書画像, 書誌情報, 情報抽出, OCR, CRF

# Automatic Extraction of Bibliographic Elements from Scanned Academic Articles

Ryohei INOUE†, Manabu OHTA††, and Atsuhiro TAKASU†††

† Faculty of Engineering, Okayama University
Tsushima-naka 3–1–1, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
†† Graduate School of Natural Science and Technology, Okayama University
Tsushima-naka 3–1–1, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
††† National Institute of Informatics
Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †inouer@de.cs.okayama-u.ac.jp, ††ohta@de.cs.okayama-u.ca.jp, †††takasu@nii.ac.jp

**Abstract**　Bibliographic information is indispensable for searching a digital library covering many academic journals. However, inputting bibliographic information into bibliographic databases requires a lot of human intervention for such a digital library as storing document images. This paper, therefore, proposes an automatic bibliographic element extraction method for academic articles scanned with OCR markups. The proposed method labels each text line of OCRed articles as a bibliographic element by using linguistic information such as an existence of characteristic words in the line in addition to visual information such as its layout. The proposed method also uses conditional random fields (CRF) for its labeling.

**Key words**　digital library, document image, bibliographic information, information extraction, OCR, CRF

## 1. はじめに

昨今整備されている電子図書館の中には，多数の文書画像を保存してデータベースを構築しているものがある．その中を検索し目的の文書を探し当てるためには，論文タイトルや著者名などの書誌情報が必須となる．しかしこれらの書誌情報を人手でデータベースに入力するには膨大なコストがかかるため，その作業を可能な限り自動で行う文書解析技術が必要とされている．これまでに，光学文字認識（OCR）[1] などの技術により紙媒体の論文文書画像をテキストデータに変換することは可能となっているが，得られたテキストがどの書誌要素に対応するかを自動で判別することは未だ容易ではない．そこで本研究では，学術論文の文書画像を OCR により認識したテキストデータから，自動で書誌要素を抽出する手法を提案する．

薬師らの提案する自動書誌要素抽出法 [10], [11] では，抽出精度は高いが，事前に一定量のトレーニングデータを個々の論文誌ごとに用意しなければならない．しかし，収録するすべての学術雑誌について，このトレーニングデータを人手で作成するにはそれなりのコストがかかる．そこで本研究では，論文誌の種類によらない論文文書の特徴に注目することにより，学術雑誌の種類によらない高精度な自動抽出法を提案する．

書誌要素のラベル付与に，本研究では自然言語処理など様々な分野で利用されている識別モデルの一つである Conditional Random Fields（CRF）[4] を利用する．

本稿の構成は次の通りである．２節で，OCR で認識したテキストデータからの情報抽出に関する研究について説明する．続く３節で，提案手法について詳しく解説する．４節で提案手法の評価実験について述べ，５節で本稿をまとめる．

## 2. 関 連 研 究

### 2.1 学術論文タイトルページからの書誌要素抽出

OCR で認識した論文タイトルページからの書誌要素抽出には，阿辺川らの研究がある [2]．彼らの研究は，サポートベクトルマシン（SVM）を用いて論文から書誌要素抽出を行うものである．本研究では文書画像を対象に書誌要素の抽出を行っているが，阿辺川らの研究ではテキスト情報を持つ PDF ファイルの論文を pdftohtml を用いて XML 形式に変換したものを入力データとして使用している．このデータにはテキスト情報，フォントサイズやフォント属性といった情報があらかじめ含まれているため，その点では本研究よりも条件は易しいと言える．ただし阿辺川らはトレーニングデータを雑誌ごとに分けずに複数種類の論文を対象とする学習を行っており，この点では本研究よりも条件は厳しい．こうした相違点から本研究と単純に比較はできないが，論文全体の抽出精度は最も高いもので 69.2 %と報告されている．また彼らは参考文献中に含まれる書誌要素の抽出も行っており，和文で 74.8 %，英文で 81.6 %の精度を達成している．

### 2.2 参考文献抽出

高須らの行った研究に，OCR 処理された学術論文の参考文献領域から参考文献を抽出するものがある [3]．彼らの手法ではまず論文全体から参考文献領域を抽出したのち，そこからさらに個々の参考文献を抽出する．彼らの提案は隠れマルコフモデル（HMM）に基づいており，OCR による文字認識に誤りがある場合も考慮されている．情報処理学会論文誌の論文を対象に実験を行い，OCR の認識精度が 97.85 %のとき，参考文献の最終的な抽出精度が 89.99 %であると報告されている．ただし，抽出した参考文献からさらに著者名，論文表題などの細かい書誌要素を抽出することはしていない．

## 3. CRF による書誌要素ラベル付け

### 3.1 Conditional Random Fields

まず，提案手法で利用する Conditional Random Fields（CRF）[4] について説明する．CRF とは，Lafferty らによって提案された観測系列のラベル付けに統計的な枠組みを与える識別モデルであり，形態素解析 [6], [7] や固有表現抽出などにおいて広く利用されている．CRF はラベル付与問題において，事実上利用可能なトレーニングデータが十分でない場合においても，しばしば HMM のような生成モデルよりも良い結果を示している [8]．そのため CRF は広範な分野で利用実績がある [5], [6], [9].

本研究の書誌要素ラベル付与問題では，チェーンモデルである標準的な CRF の定義を用いる．すなわち，入力系列 $\mathbf{x} = x_1, ..., x_n$ が与えられたとき，出力ラベル系列が $\mathbf{y} = l_1, ..., l_n$ となる条件付き確率は以下のように与えられる．

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(l_{i-1}, l_i, \mathbf{x})) \quad (1)$$

ただし $Z_x$ は，全てのラベル系列を考慮したときに確率の和が 1 となるための正規化項で，

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}' \in Y(\mathbf{x})} \exp(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(l'_{i-1}, l'_i, \mathbf{x})) \quad (2)$$

である．ここで，$f_k(l_{i-1}, l_i, \mathbf{x})$ は $i$ 番目と $i-1$ 番目の出力ラベルと入力系列 $\mathbf{x}$ に依存する任意の素性関数である．また $\lambda_k$ は素性関数 $f_k$ の重みを表すパラメータで学習により定める．また $Y(\mathbf{x})$ は入力系列 $\mathbf{x}$ に対する出力ラベル系列の集合である．そして，入力系列 $\mathbf{x}$ に対する最適な出力ラベル系列 $\hat{\mathbf{y}}$ は次式で与えられる．

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in Y(\mathbf{x})}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) \quad (3)$$

ここでラベル付与の対象である入力 $x_i$ は，本研究の場合は各行の識別番号である．一方ラベル $l_i$ は，表題，著者名，概要といった書誌要素名である．

本研究で CRF を利用した理由としては，CRF では相関のある特徴を素性として柔軟に扱えるということが挙げられる．本研究での書誌要素ラベル付けには，レイアウト情報や文字情報を素性として利用している．レイアウト情報は視覚的素性，文字情報は言語的素性といえ，CRF ではこれらの有用な情報を全て素性として利用しラベル付けを行う．これが例えば通常の HMM では，設計者は状態と出力シンボルにしかラベルや特徴を割り当てることができず，多数の相関のある特徴をそのまま利用することができない．

### 3.2 提 案 手 法

#### 3.2.1 学術論文タイトルページ

本研究では，学術論文のタイトルページから書誌要素を抽出する．ラベル付けの対象とする学術論文タイトルページの一例として，本稿のタイトルページのレイアウトを図 1 に示す．この図のように，学術論文では主要な書誌要素がタイトルページに集中して出現すると考えてよい．例えば表題，著者名といった書誌要素は大抵の論文に必須の書誌要素であり，様々な論文のタイトルページにほぼ確実に現れる．これらの情報は文書を検索する際に有用な情報であるため，利用価値も高い．それに加えて，タイトルページには著者の所属やメールアドレス，論文のキーワードなどの書誌要素が記述されることもあり，抽出対象として適当である．

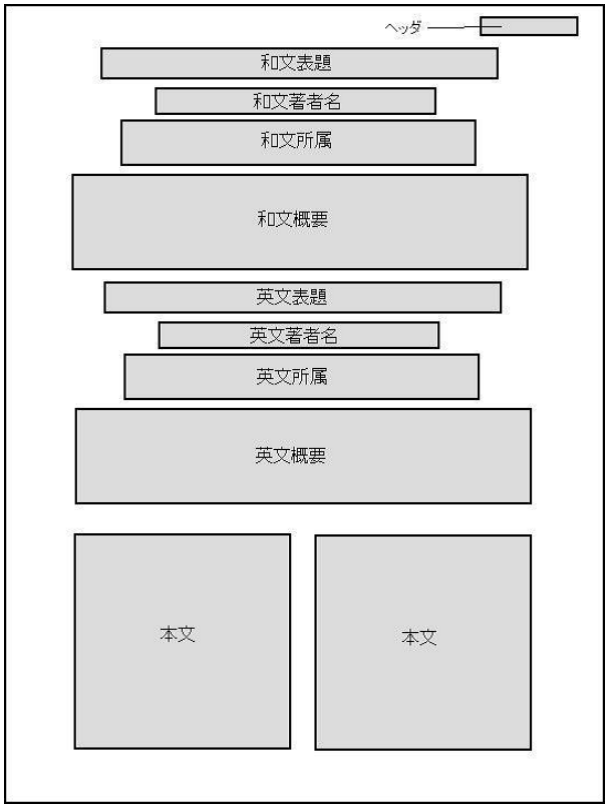


図 1　学術論文タイトルページのレイアウト例

```
<?xml version="1.0" encoding="euc-jp" ?>
<paper>
    <page number="1">
        <block x="666" y="479" width="1814" height="87" sec="">
            <line x="666" y="479" width="1814" height="87">
                <word x="666" y="479" width="1814" height="87">
                    <char  x=" 666"  y=" 486"  width="80"  height="80"> 学 </char>
                    <char  x=" 756"  y=" 484"  width="84"  height="82"> 術 </char>
                    <char  x=" 863"  y=" 493"  width="52"  height="64"> 論 </char>
                    <char  x=" 944"  y=" 493"  width="70"  height="65"> 文 </char>
                    <char  x="1048"  y=" 490"  width="47"  height="72"> 文 </char>
                    <char  x="1128"  y=" 501"  width="70"  height="53"> 書 </char>
                    <char  x="1216"  y=" 493"  width="78"  height="61"> 画 </char>
                    <char  x="1310"  y=" 488"  width="71"  height="71"> 像 </char>
                    <char  x="1399"  y=" 479"  width="77"  height="79"> か </char>
                    <char  x="1499"  y=" 496"  width="61"  height="58"> ら </char>
                    <char  x="1586"  y=" 489"  width="69"  height="66"> の </char>
                    <char  x="1687"  y=" 512"  width="52"  height="50"> 自 </char>
                    <char  x="1769"  y=" 486"  width="73"  height="72"> 動 </char>
                    <char  x="1950"  y=" 482"  width="81"  height="80"> 書 </char>
                    <char  x="2040"  y=" 479"  width="83"  height="82"> 誌 </char>
                    <char  x="2140"  y=" 484"  width="63"  height="72"> 要 </char>
                    <char  x="2239"  y=" 483"  width="48"  height="74"> 素 </char>
                    <char  x="2325"  y=" 514"  width="61"  height="41"> 抽 </char>
                    <char  x="2416"  y=" 489"  width="61"  height="62"> 出 </char>
                </word>
            </line>
        </block>
        <block x="666" y="602" width="943" height="83" sec="">
            <line x="666" y="602" width="943" height="83">
                <word x="666" y="602" width="943" height="83">
                                        ⋮
```

図 2　入力する OCR テキストの例

表 1　書誌要素ラベル

| 書誌要素 | ラベル | TIPSJ | TIEICE-E | TIEICE-J |
|---|---|---|---|---|
| 論文種別 | e-class-*journal* | - | ○ | - |
| 和文表題 | j-title | ○ | - | ○ |
| 和文著者名 | j-authors | ○ | - | ○ |
| 和文概要 | j-abstract | ○ | - | ○ |
| 和文キーワード | j-keywords | - | - | ○ |
| 英文表題 | e-title | ○ | ○ | ○ |
| 英文著者名 | e-authors | ○ | ○ | ○ |
| 英文概要 | e-abstract | ○ | ○ | - |
| その他 | other | ○ | ○ | ○ |

### 3.2.2　入出力データ

提案手法の入力データは，学術論文の文書画像を OCR 認識したテキストデータである．このテキストデータは，文書画像に OCR によるレイアウト解析と文字認識を施した結果得られるもので，XML 形式で記述されている．この XML 形式の入力データの例を図 2 に示す．このデータには，文書画像から読み取った文字情報に加えて，それぞれの文字がどのような配置で文書に出現しているかを表すレイアウトタグが記述されている．例えば文書の各ページには page タグが，その中の各テキストブロックには block タグが付与されており，そのそれぞれに x 座標，y 座標，幅，高さなどの位置情報が与えられている．同様に，文書中の各行，単語，文字にもタグが位置情報とともに与えられており，本研究ではこれらの情報を視覚的素性として利用する．またさらに，文書中に出現する特徴的な文字列を言語的素性として利用する．例えば文中に「○○研究所」や「△△大学」といった文字列が確認できれば，その箇所は著者の所属機関を表す書誌要素であると予想できる．こうした言語的な情報も素性として活用し，CRF を用いて書誌要素のラベル付けを行う．

一方出力データは，入力データに書誌要素を示すタグを追加した XML 形式データとなる．例えば本稿のタイトルページに対して提案手法による書誌要素抽出を行った場合，図 3 のような出力ファイルが期待される．ただし実際の出力ファイルにはさらにそれぞれの単語や文字を示すタグが割り振られているが，図 3 では省略している．図中に存在する block や line といったタグが OCR 認識によって元から付与されているタグであり，これらは文書中の各テキストブロック，あるいは各行といったレイアウト情報を示している．それに対して，斜体で記述されている *j-title* や *j-authors* などのタグが提案手法によって追加する書誌要素タグであり，これらはそれぞれ和文表題，和文著者名を表している．

### 3.2.3　抽出する書誌要素

抽出する書誌要素は，抽出対象となる論文の種類によって異なる．これは論文誌毎に，記述されている書誌要素に違いがあるためである．本研究では，情報処理理学会論文誌の論文 (TIPSJ)，電子情報通信学会論文誌における英語の論文 (TIEICE-E) および日本語の論文 (TIEICE-J) の三種類を対象に評価実験を行っている．これらの論文誌から抽出する書誌要素の一覧と，それに対応する書誌要素ラベルを表 1 に示す．

なお，本研究では書誌要素のラベル付けを入力テキストの line（行）単位で行っている．本研究で抽出対象とした学術雑誌論文では，同じ行に複数の書誌要素が記述されていることはほとんどなく，line の子要素の word（単語）やさらにその子要素の char（文字）単位でラベル付けを行う必要はないと判断した．

## 3.3　素性テンプレート

CRF によるラベル付けに利用する，行の位置やその中の文字のサイズなど論文文書の特徴をまとめたものを素性テンプレートと呼ぶ．本研究で比較対象とする，先行研究となる薬師らの手法で用いられていた素性テンプレート [11] を表 2 に示す．この表で素性を表す文字列の括弧内の数字は相対位置を表

```
（省略）
<block>
 <j-title>
    <line> 学術論文文書画像からの自動書誌要素抽出 < /line>
 </j-title>
< /block>
<block>
 <j-authors>
    <line> 井上諒平 † 太田学 † † 高須淳宏 † † † < /line>
 </j-authors>
< /block>
<block>
 <j-affiliation>
    <line> † 岡山大学工学部 〒 700-8530 岡山県岡山市北区....< /line>
    <line> † † 岡山大学大学院自然科学研究科 〒 700-8530....< /line>
    <line> † † † 国立情報学研究所 〒 101-8430 東京都千代田....< /line>
    <line>E-mail: † inouer@de.cs.okayama-u.ac.jp, † † ....< /line>
 </j-affiliation>
<block>
<block>
 <j-abstract>
    <line> あらまし　今日では国内の主要な学術論文を網羅し....< /line>
    <line> の書誌情報を利用することが通例となっている．し....< /line>
    <line> の入力に大きな人的コストがかかる．そこで本研究....< /line>
    <line> 自動で書誌要素を抽出する手法を提案する．提案手....< /line>
    <line> アウト等の視覚的な情報，および文字列などの言語....< /line>
    <line> のラベル付けには Conditional Random Fields（CR....< /line>
 </j-abstract>
 <j-keywords>
    <line> キーワード　電子図書館, 文書画像, 書誌情報....< /line>
 </j-keywords>
< /block>
<block>
 <e-title>
    <line>Automatic Extraction of Bibliographic Elements< /line>
    <line>from Scanned Academic Articles< /line>
 </e-title>
< /block>
<block>
 <e-authors>
    <line>Ryohei INOUE †, Manabu OHTA † †, and...< /line>
 </e-authors>
< /block>
（省略）
```

図 3　出力データの例

しており，薬師らの研究や本研究では，この素性テンプレートに基づいて素性関数が生成される．薬師らの手法では，素性テンプレートには行の XY 座標や文字のサイズなど，OCR テキスト中に含まれるあらゆるレイアウト情報の絶対値を素性として利用していた．しかしこの方法では，トレーニングデータとテストデータの論文誌が異なると，書誌要素の抽出精度が大幅に低下する．これは，論文の書式が各論文誌ごとに個別に定められているため，同じ書誌要素でも論文誌によってサイズや出現する順序などが異なるためである．また，収録されている書誌要素の種類そのものが異なる場合もある．このように異なる種類の論文誌の論文データを学習に利用することは難しい．

この問題点を解決するため，本研究では薬師らの提案した素性テンプレートを大幅に見直し，複数の論文誌のデータを混合したトレーニングデータを用いて，高精度な書誌要素抽出を目指す．本研究で使用する素性テンプレートを表 3 に示し，薬師らの素性テンプレートとの違いについて以下で説明する．

### 3.3.1　視覚的素性

視覚的素性は，論文文書の文書画像解析によって得られる．例えば，各行の位置や文字の大きさ，行間の間隔などがこの視覚的素性にあたる．本研究では，行の XY 座標や文字サイズなどはその数値そのものは利用せず，まずそれらの値とその論文の平均値との比を算出する．そうして得られた比の値が一定の範囲にあるものをまとめた代表値を素性として利用する．

表 2　薬師らの手法で使用する素性テンプレート

| 種類 | 素性 | 内容 |
|---|---|---|
| unigram | <i(0)> | line の識別番号 |
| | <x(0)> | line の X 座標 |
| | <y(0)> | line の Y 座標 |
| | <w(0)> | line の幅 |
| | <h(0)> | line の高さ |
| | <g(0)> | 前の line との間隔 |
| | <cw(0)> | line 内文字の幅の中央値 |
| | <ch(0)> | line 内文字の高さの中央値 |
| | <#c(0)> | line 内の文字数 |
| | <ec(0)> | line 内の英数字の割合 |
| | <kc(0)> | line 内の漢字の割合 |
| | <jc(0)> | line 内の平仮名・片仮名の割合 |
| | <s(0)> | line 内の記号の割合 |
| | <kw(0)> | line の最初の特徴的な文字列の有無 |
| bigram | <*l*(-1),*l*(0)> | ラベルの遷移 |

表 3　提案手法で使用する素性テンプレート

| 種類 | 素性 | 内容 |
|---|---|---|
| unigram | <li(0)> | line の識別番号 |
| 視覚的素性 | <lx(0)> | line の X 座標の論文全体の平均に対する比 |
| | <ly(0)> | line の Y 座標の論文全体の平均に対する比 |
| | <lw(0)> | line の幅の論文全体の平均に対する比 |
| | <lh(0)> | line の高さの論文全体の平均に対する比 |
| | <lg(0)> | line の間隔の論文全体の平均に対する比 |
| | <lcw(0)> | line 内の文字幅の平均と<br>論文全体の文字幅の平均との比 |
| | <lch(0)> | line 内の文字高さの平均と<br>論文全体の文字高さの平均との比 |
| | <lc(0)> | line の文字数の論文全体の平均に対する比 |
| unigram | <lalphabet(0)> | line 内の英数字の割合 |
| 言語的素性 | <lkanji(0)> | line 内の漢字の割合 |
| | <lkana(0)> | line 内の平仮名・片仮名の割合 |
| | <lsymbol(0)> | line 内の記号の割合 |
| | <lfeature(0)> | line 内に現れる特徴的な文字列の種類 |
| bigram | <*l*(-1),*l*(0)> | ラベルの遷移 |

### 3.3.2　言語的素性

言語的素性は，論文文書のテキストデータを分析して得られる情報である．例えば，各行に含まれる文字の種類，書誌要素の判別に利用可能な特徴的な文字列といった情報がこの言語的素性にあたる．出現する文字の種類などの情報は，文書レイアウトの違いには影響を受けないので，論文誌の種類によらない普遍的な情報だと考えられる．

英数字，漢字など文字の種類の割合については，薬師らの手法とほぼ同様に各行中の出現頻度の割合を百分率で算出したのち，その一の位を切り捨てて素性として使用している．つまり，この素性として使用される数値は，0,10,20,...,100 の 11 種類となる．また，文書中の特徴的な文字列の有無に関する素性については，薬師らは 4 種類の文字列を特徴的な文字列としてラベル付けに利用したが，本研究ではこれにさらに追加して合計 17 種類の文字列を利用する．

### 3.3.3　bigram 素性

薬師らの手法で利用する bigram 素性は，本研究でも同じように使用する．この素性は付与される書誌要素ラベルの連接に関する情報を表したもので，例えば表題の後に著者名が記述され，つづいて概要が記述される，などの書誌要素の出現順に関する制約を考慮することができる．

## 4.　実　　験

提案手法の有効性を調べるため，評価実験を行う．この実験では，工藤らが作成した CRF++[(注1)] を利用して書誌要素ラベ

(注1)：http://crfpp.sourceforge.net/

ル付けを実行した．CRF++は，何らかの連続したデータに対して CRF を用いて分割やラベル付けなどの処理を行うオープンソースのソフトウェアである．この実験結果から，論文全体の書誌要素ラベル付けの正解率を算出した．実験には三種類の論文誌の論文データをそれぞれ 2 年分から 6 年分用意し，トレーニングデータとテストデータに分けて利用した．

### 4.1 トレーニングデータとテストデータ

実験では，以下の計 480 件の論文データを均等に混合し，トレーニングデータとする．

- 2003 年：情報処理学会論文誌　160 件
- 2003 年：電子情報通信学会論文誌（英語）　160 件
- 2000-04 年：電子情報通信学会論文誌（日本語）160 件

一方，テストデータとして下記の論文データを用意した．

- 2004 年：情報処理学会論文誌　409 件
- 2005 年：電子情報通信学会論文誌（英語）　630 件
- 2005 年：電子情報通信学会論文誌（日本語）　150 件

この三種類のテストデータに対して，トレーニングデータ量を変えて書誌要素抽出精度の評価実験を行う．

### 4.2 実験結果と考察

4.1 節で説明したデータを用いて，薬師らの手法との比較実験を行った．比較するのは，表題や著者名など個別の書誌要素の抽出精度ではなく，論文中の全ての書誌要素が正確に抽出できるかどうかの総合的な抽出精度である．各論文誌ごとの抽出精度の比較結果を図 4，図 5，図 6 に示す．

実験の結果，提案手法による書誌要素の抽出精度は，薬師らの研究と比較しておおむね同等かそれ以上だった．特に図 5 では，提案手法が薬師らの手法をトレーニングデータの量によらず上回った．利用するトレーニングデータ量が増加するにつれて薬師らの手法との差は小さくなるが，トレーニングデータの量が少ない場合は薬師らの手法を大幅に上回っている．図 6 では，トレーニングデータが増加すると薬師らの手法による抽出精度を下回っている．ただし，トレーニングデータが少ない場合の抽出精度は提案手法のほうが高い．一方，図 4 の実験結果では提案手法の抽出精度は図 5 などとそれほど変わらないが，薬師らの手法も同様に高い抽出精度となったので，二つの手法の間に大きな差はなかった．ただし提案手法は，トレーニング

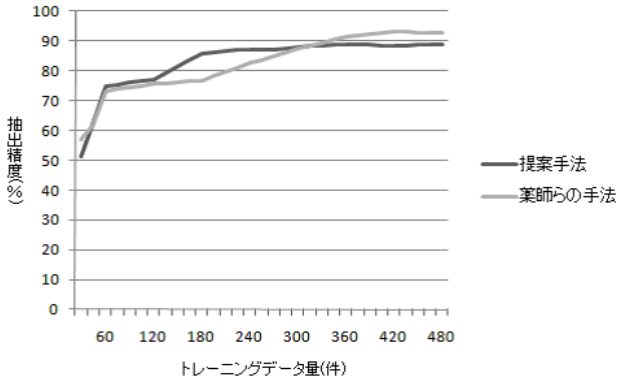


図 4　書誌要素の抽出精度（情報処理学会論文誌）

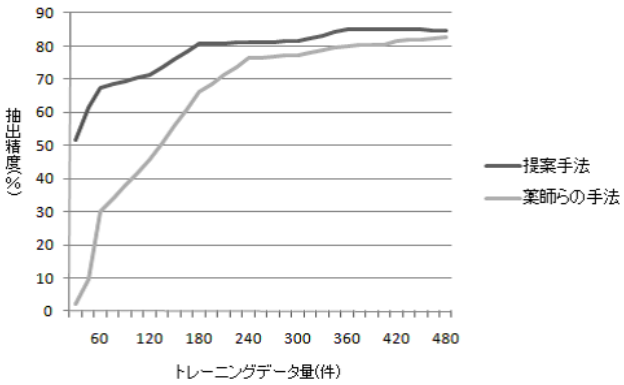


図 5　書誌要素の抽出精度（電子情報通信学会英文論文誌）

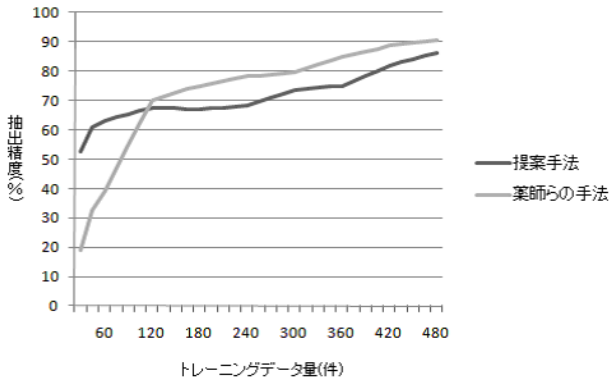


図 6　書誌要素の抽出精度（電子情報通信学会和文論文誌）

データが 15 件しかない場合でも，3 種類の論文誌のいずれに対しても 50 %を超える抽出精度を達成している．これらの実験結果からは，トレーニングデータがおよそ 120 件未満の場合に提案手法が有効であることが分かる．

さらなる精度の向上を実現するためには，OCR テキスト中に含まれるレイアウト情報をさらに分析する必要がある．また，行の位置や文字サイズなどの大きさの出現分布を調べて，それに基づいて素性の値を決定する方法も考えられる．

## 5. ま と め

本稿では，学術論文の文書画像を OCR 認識したテキストデータから，CRF を用いて書誌要素を自動で抽出する手法を提案した．提案手法は，OCR 認識によって得られた学術論文タイトルページのテキストデータに対して，CRF を利用して表題，著者名，概要などの書誌要素の判別を行い，書誌要素ラベルを付与する．書誌要素抽出実験では，三論文誌の論文データを混合したトレーニングデータを用いて各論文誌から書誌要素を抽出した．その結果，提案手法はトレーニングデータが 15 件しかなくても，50 %以上の抽出精度を示すことが確認できた．今後，入力データ中のレイアウト情報をさらに分析し，より有効な素性について検討していきたい．

## 文　　献


[1] Bunke, H. and Wang, P.: Handbook of Character Recognition and Document Image Analysis, World Scientific (1997).

[2] 阿辺川 武, 難波 英嗣, 高村 大也, 奥村 学: 機械学習による科学技術論文からの書誌情報の自動抽出, 情報処理学会研究報告 2003-FI-72/2003-NL-157, pp.83-90 (2003).

[3] Takasu, A. and Aihara, K.: Quality Enhancement in Information Extraction from Scanned Documents, In Proc. of DocEng '06, pp.122-124 (2006).

[4] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data, In Proc. of 18th International Conference on Machine Learning, pp.282-289 (2001).

[5] 東 藍, 浅原 正幸, 松本 裕治: 条件付確率場による日本語未知語処理, 情報処理学会研究報告 2006-NL-173, pp.67-74 (2006).

[6] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, In Proc. of EMNLP 2004, pp.230-237 (2004).

[7] 工藤 拓, 山本 薫, 松本 裕治: Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告 2004-NL-161, pp.89-96 (2004).

[8] Takechi, M., Tokunaga, T. and Matsumoto, Y.: Chunking-based Question Type Identification for Multi-Sentence Queries, In Proc. of SIGIR 2007 Workshop on Focused Retrieval (2007).

[9] Zhao, H., Huang, C. N. and Li, M.: An Improved Chinese Word Segmentation System with Conditional Random Field, In Proc. of Fifth SIGHAN Workshop on Chinese Language Processing, pp.162-165 (2006).

[10] 薬師 貴之, 太田 学, 高須淳宏: CRF を用いた学術論文 OCR テキストからの自動書誌要素抽出, 情報処理学会論文誌：データベース, TOD42, Vol.2 No.2, pp.126-136 (2009).

[11] 薬師 貴之, 太田 学, 高須淳宏: 様々な学術論文誌 OCR テキストからの書誌要素抽出，電子情報通信学会 2009 年総合大会，D-12-48，情報・システム講演論文集 2，p157 (2009).