情報処理学会論文誌　データベース　Vol. 2　No. 2　126–136 (June 2009)

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

薬師貴之[†1,*1]　太田　学[†1]　高須淳宏[†2]

文献データベースは学術論文を所蔵する電子図書館では不可欠である．しかし紙媒体の論文からの書誌要素抽出は，OCR などの画像処理技術を利用してもその抽出コストは高い．そこで本稿では，OCR 処理された学術論文から書誌要素を自動的に抽出する手法を提案する．提案手法では，まず OCR の文書画像処理によって得られた矩形テキスト領域に対して，あらかじめ定義した書誌要素を表すラベルを付与する．さらに，必要に応じて矩形テキスト領域内の各文字に対してもラベル付けを行う．この文字へのラベル付けによって，複数の著者名が記述された矩形テキスト領域から各著者の名前を抽出することができる．提案手法では，矩形テキスト領域や文字へのラベル付けに Conditional Random Fields（CRF）を使用する．言語の異なる 2 種類の論文誌を用いて実験を行ったところ，矩形領域へのラベル付けは，和文誌で 97.56%，英文誌で 97.27%の精度であった．また文字へのラベル付けによる和文誌の和文著書名領域からの各著者名の抽出精度は 99%以上を達成した．

Automatic Bibliographic Element Extraction from OCRed Academic Articles Using Conditional Random Fields

Takayuki Yakushi,[†1,*1] Manabu Ohta[†1]
and Atsuhiro Takasu[†2]

Bibliographic databases are indispensable to digital libraries of academic articles. However, extracting bibliographic elements from printed documents requires a lot of human intervention; it is not cost-effective, even when using various document image-processing techniques such as optical character recognition (OCR). In this paper, we propose an automatic bibliographic element extraction method for academic articles scanned with OCR markup. The proposed method first labels text blocks as predetermined bibliographic elements and then further labels the characters in each labeled text block if necessary. The second labeling enables us to extract each author's name from the authors' text block. The method uses conditional random fields (CRF) for labeling both text blocks and the characters in them. We applied the method to Japanese and English academic articles. The experiments showed that the proposed method correctly extracted all the predefined bibliographic text blocks from 97.56% of the Japanese articles and 97.27% of English ones, respectively. The proposed method also correctly extracted all the author name strings from more than 99% of the Japanese authors' text blocks in the Japanese articles.

1. はじめに

文書画像を扱う電子図書館では，紙媒体の学術論文から書誌情報をなるべく自動で抽出するための文書解析技術[14),17)]が求められている．これまでに，光学文字認識（OCR）[3)]をはじめとする，レイアウト解析，論理解析[8)]などの研究進展により，近年では印刷文書を電子データに変換する自動文書画像処理システムも提案されている．本研究で対象とする学術論文では，検索に利用するため，表題や著者名，参考文献といった書誌要素の抽出が特に重要となる．すでに参考文献の有用性に着目し，文書画像の OCR 出力から参考文献を抽出する技術がいくつか提案されている[12),15)]．一方，学術論文のタイトルページに存在する表題や著者名，キーワードといった有用な書誌要素抽出も電子図書館では重要であり，レイアウトなどを手がかりにこれらを抽出する研究もある[8),14)]．本研究の目的はこの後者で，本稿では学術論文のタイトルページから必要な書誌要素を自動で獲得するための手法を提案する．従来研究では抽出精度が不十分なため，結局後処理を人手で行う必要があった．しかし本研究ではこの後処理を行わなくても実用上問題がないように，高精度な自動書誌要素抽出法を提案する．これにより，現在の人手による書誌データの入力作業を大幅に削減できる．本研究における書誌要素抽出は，主としてレイアウト解析，文字認識，情報抽出の 3 段階に分かれるが，文字認識誤りに対してロバストな情報抽出を目指している．

本稿では OCR 処理された文書から自動的に書誌要素を抽出するため，Conditional Random Fields（CRF）[7)]に基づいたラベル付け手法を提案する．まず実験に使用した情報処理学会論文誌 44 巻に収録されている論文のタイトルページのレイアウト例を**図 1** に示す．

†1 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
†2 国立情報学研究所
National Institute of Informatics
∗1 現在，NEC システムテクノロジー株式会社
Presently with NEC System Technologies, Ltd.

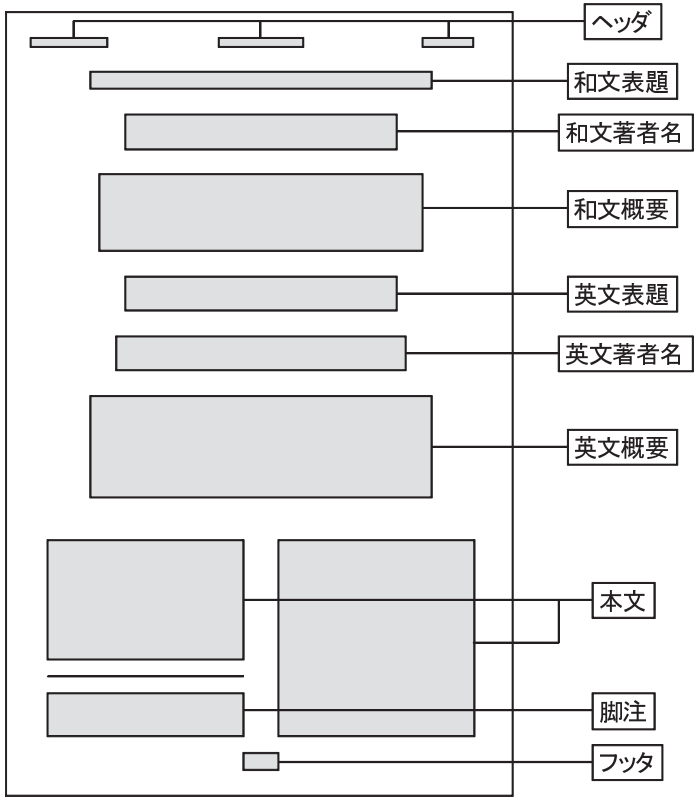

図 1　情報処理学会論文誌のタイトルページのレイアウト
Fig. 1　Layout of a title page of IPSJ Journal.

紙面の上からヘッダ，和文表題，和文著者，和文概要，英文表題，英文著者，英文概要，本文，脚注，フッタと並んでいる．また本文は 2 段組である．提案手法の入力データは学術論文文書画像を OCR によって処理した XML 形式のデータであり，文字，行，さらにそれらを含む矩形テキスト領域に，それぞれ char，line，block というレイアウトを示すタグが OCR によって付与される．図 1 に示す各書誌要素に 1 つの block タグが付与されることが理想であるが，1 つの書誌要素が複数の block 領域を含むことも多い．また，複数の書誌要素に 1 組の block タグが付与される可能性もある．

一方出力は，入力データに書誌要素を示すタグを追加した XML ファイルとなる．本研究で扱う XML ファイルの主要部の例を**図 2** に示す．斜体で記述している j-title や j-authors などのタグは本稿で提案する自動書誌要素抽出によって付与するタグであり，これらはそれぞれ和文表題，和文著者という書誌要素を表す．図 2 では省略したが実際の入力 XML ファイルでは，各文字に char タグが付けられている．また，レイアウトを表す block や line，char のタグは，X 座標，Y 座標，幅，高さの情報を属性として持つ．これによりその文字や領域がページのどこに位置しており，どの程度の大きさなのかを把握することができる．

提案手法は 2 段構成である．まず矩形テキスト領域に対して，表題，著者名，概要などのあらかじめ定めた書誌要素を表すラベルを付与する．次に必要に応じて矩形テキスト領域内の各文字に対して書誌要素ラベルを付与する．これはより細分化された書誌情報を抽出するためで，この文字へのラベル付けを利用して，抽出された著者名領域からさらに各々の著者名の抽出を試みる．よって，提案手法による書誌要素抽出手順は，以下のようになる．

(1)　OCR によるレイアウト解析と文字認識
(2)　CRF を用いた矩形テキスト領域へのラベル付け
(3)　CRF を用いた文字へのラベル付け

本稿では上記 (2) および (3) について詳細に述べる．

2. Conditional Random Fields

Conditional random fields（CRF）は，Lafferty ら[7] 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，形態素解析[5),6)] や固有表現抽出などによく利用されている．CRF はラベル付与問題において，実用上利用できるトレーニングデータが十分でないような場合でも，しばしば隠れマルコフモデル（HMM）のような生成モデルより良い結果を示している[16]．そのため CRF は，自然言語処理からバイオインフォマティクスなど広範な分野で利用実績がある[2),5),18)]．

本研究の矩形テキスト領域および文字へのラベル付与問題では，標準的な CRF の定義を用いた．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（タグ）系列が $\mathbf{y} = t_1, ..., t_n$ となる条件付き確率は以下のように与えられる．

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

ただし Z_x は，すべ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，

```
<j-title>
<block>
  <line>CRF を用いた学術論文 OCR テキスト. ...< /line>
  <line> 自動書誌要素抽出 < /line>
< /block>
</j-title>
<j-authors>
<block>
  <line> 薬師貴之† 1,*1 太田学† 1 高須淳宏† 2< /line>
< /block>
</j-authors>
<j-abstract>
<block>
  <line> 文献データベースは学術論文を所蔵する. ..< /line>
        （省略）
  <line> 領域からの各著者名の抽出精度は 99%.....< /line>
< /block>
</j-abstract>
<e-title>
<block>
  <line>Automatic Bibliographic Element .....< /line>
  <line>OCRed Academic Articles Using C....< /line>
  <line>Random Fields< /line>
< /block>
</e-title>
<e-authors>
<block>
  <line>TAKAYUKI YAKUSHI, † 1,*1MA...< /line>
< /block>
</e-authors>
<e-abstract>
<block>
  <line>Bibliographic databases are indispe...< /line>
        （省略）
  <line>99% of the Japanese authors' text ....< /line>
< /block>
</e-abstract>
<block>
  <line>1. はじめに < /line>
```

図 2　レイアウトタグと書誌要素タグ
Fig. 2　Layout and bibliographic element tags.

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}' \in Y(\mathbf{x})} \exp\Big(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(t'_{i-1}, t'_i, \mathbf{x})\Big) \tag{2}$$

である．ここで，$f_k(t_{i-1}, t_i, \mathbf{x})$ 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 λ_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．$Y(\mathbf{x})$ は入力系列 $\mathbf{x}$ に対する出力ラベル系列の集合である．また式の記述を簡単にするため，Lafferty らと同様にここでは仮想的な t_0 の存在を仮定している[7]．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．

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in Y(\mathbf{x})} P(\mathbf{y}|\mathbf{x}) \tag{3}$$

ここでラベル付与の対象である入力 x_i は，矩形テキスト領域へのラベル付与の場合はそのテキスト領域であり，文字へのラベル付与の場合は特定の矩形領域内，たとえば著者名領域内，の文字である．一方ラベル t_i は，矩形テキスト領域へのラベル付与の場合は表題，(複数）著者，概要といった書誌要素名であり，文字へのラベル付与の場合は各著者名といったより細分化された書誌要素となる．

本研究で CRF を利用した大きな理由の 1 つは，相関のある特徴を素性として柔軟に扱えることである．本研究では，矩形テキスト領域へのラベル付与では主にレイアウト情報を，文字へのラベル付与ではそれに加えて文字の情報そのものを素性として利用する．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，両者ともラベル付けに有用と考えられる．CRF ではこのような特徴をすべて利用してラベル付けが行える．これがたとえば通常の HMM であれば，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．さらに，文字へのラベル付け問題では，我々が提案した CRF に基づく手法は HMM に基づく手法[11] を上回る精度を示している[9]．本稿では矩形テキスト領域へのラベル付与に用いた素性については 3 章で，文字へのラベル付与に用いた素性は 4 章で詳しく説明する．

3. 矩形テキスト領域へのラベル付け

矩形テキスト領域へのラベル付けによって，学術論文の書誌要素抽出を行う[10]．本研究では，通常論文のタイトルページに記載されている，表題や著者名，概要など主要な書誌要素をすべて抽出対象とする．またここでラベル付けの対象とする矩形テキスト領域は，1 つ以上の line（行）を含む block タグでくくられた領域のことで，使用する OCR により順序

表 1　矩形テキスト領域用書誌要素タグ
Table 1　Bibliographic element tags for text blocks.

書誌要素	タグ	情報処理学会論文誌（和文）	電子情報通信学会論文誌（英文）
論文種別	type	-	○
和文表題	j-title	○	-
和文著者	j-authors	○	-
和文概要	j-abstract	○	-
英文表題	e-title	○	○
英文著者	e-authors	○	○
英文概要	e-abstract	○	○
その他	other	○	○

付けられたうえでシリアライズされている（図 2 参照）．よってこのような block 系列にラベル付けを行うため，使用する CRF は linear chain model となる．またここで注意すべきは，各書誌要素に 1 つの block のみが対応するとは限らず，複数の block が 1 つの書誌要素を構成する場合があるということである．その場合，その書誌要素を構成するすべての block に正しくラベル付けできなければ，正しく抽出したことにはならない．

本稿の実験では，情報処理学会論文誌（和文誌）と電子情報通信学会論文誌（英文誌）という 2 種類の論文誌を対象に書誌要素抽出を行った．これらのために用意した書誌要素タグを**表 1** にまとめる．表に示すように，和文誌では 7 種類，英文誌では 5 種類のタグを用いて書誌要素抽出を行った．これについては，論文誌ごとに記載されている書誌要素が異なるので，基本的にタグも論文誌ごとに用意する必要がある．表題や著者など，どの論文誌にも必須と思われる書誌情報のみを定義して抽出する方法も考えられるが，本研究では表 1 に示した “論文種別” のように論文誌に固有の書誌要素も抽出対象に含めた．

3.1　素性の展開

表 2 に矩形テキスト領域へのラベル付けに用いた素性テンプレートを示す．ここで素性を表す文字列の括弧内の数字は相対位置を表す．提案手法では，この素性テンプレートに基づいて素性関数が生成され利用される．表 2 の素性はタグの遷移以外はすべて視覚的素性といえる．矩形テキスト領域の座標や大きさだけではなく，直前の矩形テキスト領域との距離や，矩形テキスト領域中の文字の大きさや数も考慮している．また，我々の OCR は，論文文書画像の二次元のレイアウトをシリアライズして図 2 のような一次元の XML ファイルを出力するが，素性として X，Y 座標の両方向の特徴を考慮して書誌要素抽出を行う．一方，論文誌によっては概要のはじめに “概要” や “要約” といった文字列が必ず記載されて

表 2　矩形テキスト領域へのラベル付けに用いる素性テンプレート
Table 2　Feature templates for text block labeling.

種類	素性	内容
unigram	<bi(0)>	block 識別番号
	<bx(0)>	block の X 座標
	<by(0)>	block の Y 座標
	<bw(0)>	block の幅
	<bh(0)>	block の高さ
	<bg(0)>	前の block との間隔
	<b#c(0)>	block 内の文字数
	<bcw(0)>	block 内文字の幅の中央値
	<bch(0)>	block 内文字の高さの中央値
bigram	<t(-1),t(0)>	タグの遷移

1	500	400	1100	80	460	10	80	70
2	500	500	2000	80	30	20	70	70
3	700	700	1800	60	130	10	60	60
4	400	900	2300	300	130	220	40	40
5	400	1400	2400	70	140	50	40	50
6	600	1600	2000	60	100	40	30	30
7	400	1800	2300	520	110	640	20	20
8	1600	2500	1400	1790	160	490	40	40
9	200	2500	400	50	-1750	0	40	40
10	100	2600	1300	1270	70	360	40	40
				⋮				

図 3　素性の具体例
Fig. 3　Examples of the features.

いるなど，決まった定型文が存在することがある．その場合はラベル付けの助けとなりうるので文字情報を言語的素性として利用すれば有効と考えられる．

表 2 の素性テンプレートに対応する素性の具体例を**図 3** に示す．図 3 は 5.2.2 項の実験で用いた情報処理学会論文誌のデータの一部で，ここでは 1 行が 1 つの矩形テキスト領域（block）に対応しており，表 2 に示したタグの遷移以外の素性の値が左から順に並んでいる．よって図 3 から，識別番号が 1 の block の X，Y 座標はそれぞれ 500，400 で，幅，高さがそれぞれ 1100，80 であることが分かる．ただし，これらの素性の数値は元の値そのものではなく代表値を用いている．すなわち表 2 の unigram 素性のうち，<bi(0)> は元の値そのものであるが，<bx(0)>，<by(0)>，<bw(0)> は 100 単位，その他は 10 単位で区

切っており，データスパースネスの問題にある程度対処できると考える．よって，図 3 で識別番号が 9 の block の block 内文字数が 0 となっているのは，文字数が 10 未満であることを示している．また，たとえば 5.2.2 項の実験で用いた情報処理学会論文誌では，<bx(0)> の異なり数は 33，<bcw(0)> のそれは 9 となった．

一方，bigram 素性テンプレート <t(-1),t(0)> からは，たとえば以下のような素性関数が生成される．

$$f_k(t_{i-1}, t_i, \mathbf{x}) = \begin{cases} 1 & \text{if } t_{i-1} = \text{j-title}, t_i = \text{j-authors} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

この bigram 素性によって，論文誌ごとにレイアウトが違っても，たとえば表題の後に著者が続き次に概要が記述される，などといった書誌要素のレイアウトに関する制約が考慮される．

生成される素性関数の数は，出力タグの種類数によって異なる．情報処理学会論文誌では，<bi(0)> などの unigram 素性テンプレートによって生成される素性関数の数は $7 \times N$ で，7 は表 1 に示した情報処理学会論文誌に存在する出力タグの種類数，N は <bi(0)> の場合 block 識別番号の異なり数である．本研究では，OCR によってシリアライズされた block 列に，便宜上先頭から順にこの block 識別番号を割り当てている．よって実際には N は，トレーニングデータとした論文における block 数の最大値となる．また bigram 素性テンプレート <t(-1),t(0)> から生成される素性関数の数は，7×7 となる．一方電子情報通信学会論文誌では，unigram 素性テンプレート <bi(0)> によって生成される素性関数の数は $5 \times N$ で，bigram 素性テンプレート <t(-1),t(0)> から生成されるその数は，5×5 となる．

4. 文字へのラベル付け

紙面上の近い位置にあってしかも同じくらいの大きさのフォントで印刷されていると，本来分離したい書誌情報でも 1 つの矩形テキスト領域（block）として認識されることが多い．たとえば実験で使用した情報処理学会論文誌では，著者名 block に複数の著者名が含まれる．書式によっては，著者名のすぐ下に所属や email アドレスなどが連続して記載されるものもあり，これらが互いにレイアウト上で近い位置にあれば OCR の文書画像解析によって同じ block として認識される可能性が高い．我々は著者名などをアンカとした論文文書画像のハイパーテキスト化を検討しており，そのためには複数の著者名を含む著者名 block から各著者名を抽出しなければならない．そこで，CRF を利用した文字へのラベル付与による詳細書誌情報抽出を提案する．本章では，これを各著者名抽出の問題として説明する．

表 3　文字用タグ
Table 3　Character tag sets.

	タグ	意味
2-tag 集合	a	著者名の文字
	d	デリミタの文字
2+pos-tag 集合	ai	著者名の i 番目の文字
	dj	デリミタの j 番目の文字

4.1 著者名抽出

著者名抽出は，著者名について記述された領域のすべての文字に対してラベル付けを行い，著者名とデリミタ（区切り文字）を区別することで実現できる[9]．図 2 の例のように，通常著者名領域は 1 人目の著者名から始まりデリミタ，2 人目の著者名と続く．デリミタは具体的には ‘†’ などの記号で，著者の所属を示すために使われる．図 2 の例では，デリミタと区別して 3 人の著者名を抽出する必要がある．また本稿では著者名領域に存在する，著者名の文字ではない文字はすべてデリミタとして扱っている．

提案手法では，著者名の文字とデリミタの文字を区別するために，**表 3** に示すような 2 種類のタグを用意した．2-tag は単に著者名またはデリミタを示すタグからなり，一方 2+pos-tag は，著者名またはデリミタという情報に加えて，それぞれの文字列中の何文字目であるかを示すタグである．また $1 \le i \le x$ および $1 \le j \le y$ が成り立ち，x および y の値はそれぞれ，トレーニングデータに現れる著者名およびデリミタの文字列の最長の長さと一致する．

4.2 素性の展開

著者名抽出に用いる素性は，文字そのものの情報と文字の幅である（**表 4**）．文字は言語的素性，文字幅は視覚的素性で，この両者を利用した．

著者名抽出では表 4 にまとめた素性テンプレートから素性関数が生成される．たとえば 2-tag を使用した場合，unigram 素性テンプレート <c(-1)> からは以下のような素性関数が生成され 2 値を返す．

$$f_k(t_{i-1}, t_i, \mathbf{x}) = \begin{cases} 1 & \text{if } x_{i-1} = '\dagger', t_i = d \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

一方，たとえば 2+pos-tag を使用した場合，bigram 素性テンプレート <t(-1),t(0)> からは以下のような素性関数が生成される．

表 4　文字へのラベル付けに用いる素性テンプレート
Table 4　Feature templates for character labeling.

種類	素性	内容
unigram	<c(-1/0/1)>	文字情報（数字は相対位置を表す）
	<w(-1/0/1)>	文字幅（数字は相対位置を表す）
bigram	<t(-1),t(0)>	タグの遷移

$$f_k(t_{i-1}, t_i, \mathbf{x}) = \begin{cases} 1 & \text{if } t_{i-1} = \text{a1}, t_i = \text{a2} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

素性テンプレート <c(j)> から生成される素性関数の数は $L \times N$ で，L は表 3 に示した出力タグの種類数，N は異なり文字数である．素性テンプレート <w(j)> から生成される素性関数の数は $L \times M$ で，M は文字幅の異なり数である．bigram 素性テンプレート <t(-1),t(0)> から生成されるその数は，$L \times L$ となる．

5. 評 価 実 験

提案手法に対して評価実験を行う．実験は，矩形テキスト領域へのラベル付与による書誌要素抽出，文字へのラベル付与による著者名抽出の 2 つに大別できる．

実験には工藤が開発して公開している CRF++[⋆1]を利用した．CRF++の学習では，精度を重視して L1 正則化モデルではなく L2 正則化モデルを用いた．また実際小規模な予備実験においてこの両者を比較したところ，L2 正則化モデルの方がかなり良かった．また素性の出現頻度のカットオフは特に指定せず，1 度でも出現する素性はすべて学習に利用した．CRF++では，学習データへの fitting を調節するパラメータ c が設定可能で，デフォルトは 1 である．この c の値を大きくすると overfitting になり，小さければ underfitting となる．そこで 5.2.3 項の実験において，トレーニングデータおよびテストデータと異なるデータを用いて精度を評価した．c を 1 から 10 まで 1 ずつ変化させても，10 から 100 まで 10 ずつ変化させても精度に変化がなかったため，0.2 から 2 まで 0.2 ずつ変化させたところ 0.8 以下では精度が低下し，0.2 が最低となった．よって，c はデフォルトの 1 として本章の実験を行った．

⋆1 http://crfpp.sourceforge.net/

5.1　OCR

ここではまず本研究で使用している OCR について説明する．我々は OCR ベンダとともに，レイアウト解析と文字認識機能を持つ OCR システムを開発した．和文の論文では日本語と英語の両方が用いられるため，OCR システムは両言語の OCR エンジンを持ち，それらを自動的に切り替えて認識を行う．さらにこの OCR システムは，通常は出力しないレイアウト解析結果，すなわち矩形テキスト領域，行，文字の位置情報を，認識結果の XML ファイルに属性として明示的に出力する．

この OCR の認識精度は，概要の部分では 99.00%，参考文献の部分では 97.01%とおおむね良好であった．誤りは主に，参考文献における日本語と英語の文字の混在や，様々なフォントや句読点などの記号が使われていることに起因している．

5.2　矩形テキスト領域へのラベル付け精度

本節では，矩形テキスト領域へのラベル付け精度を評価する実験について述べる．提案手法との比較のため，表 2 に示したすべての unigram 素性を利用した Support Vector Machine（SVM）によるラベル付け精度についても評価した．

5.2.1　SVM

SVM によるラベル付け実験には，Chang らが公開している LIBSVM [4] を用い，各矩形テキスト領域を書誌要素に対応するクラスへ分類した．SVM の学習には，図 3 の素性の値を正規化した値を用い，RBF カーネルを選択した．パラメータは，トレーニングデータの five-fold cross-validation により決定したが，精度評価の基準は我々のものとは若干異なる．すなわち，LIBSVM では単純にどれだけの矩形テキスト領域を正しくラベル付けできたかを精度としているが，我々は次項で述べるように書誌要素を構成するすべての矩形テキスト領域に過不足なくラベル付けできたものを正解とし，そのように正しくラベル付けできた書誌要素の割合を精度としている．

5.2.2　block への書誌要素ラベル付け

言語およびレイアウトの異なる 2 種類の論文誌を用いて矩形テキスト領域へのラベル付け精度を評価する．1 つは情報処理学会論文誌の和文誌で，タイトルページに日本語の表題，著者名，概要と英語のそれらがこの順に記載されている．実験ではトレーニングセットに 44 巻 280 ファイル，テストセットに 45 巻 409 ファイルを使用した．もう 1 つは，電子情報通信学会論文誌の英文誌であり，トレーニングセットに E87-B 巻 431 ファイル，テストセットに E88-B 巻 587 ファイルを使用した．また，ラベル付与の対象となる矩形テキスト領域は block であり，表 2 の素性テンプレートを用いた．ここでいうラベル付け精度とは，

表 5　矩形テキスト領域へのラベル付け精度
Table 5　Text block labeling accuracy.

	情報処理学会論文誌（block 単位・言語的素性なし）		電子情報通信学会論文誌（block 単位・言語的素性なし）		電子情報通信学会論文誌（block 単位・言語的素性あり）		電子情報通信学会論文誌（line 単位・言語的素性あり）	
	SVM	CRF	SVM	CRF	SVM	CRF	SVM	CRF
論文種別	-	-	100.00	99.49	100.00	99.66	99.84	99.37
和文表題	93.89	99.27	-	-	-	-	-	-
和文著者	91.93	98.53	-	-	-	-	-	-
和文概要	97.31	99.27	-	-	-	-	-	-
英文表題	92.67	99.27	99.66	99.32	99.66	99.32	99.05	99.05
英文著者	93.89	99.02	97.44	99.15	98.13	98.98	91.27	99.21
英文概要	96.82	99.51	90.63	96.93	98.98	98.98	71.43	97.62
論文	81.42	97.56	87.73	95.23	96.76	97.27	64.44	96.03

1 つの書誌要素を構成する 1 つ以上の矩形テキスト領域に過不足なく正しくラベル付けができたものを正解とし，書誌要素ごとにこの正解の割合を求めたものである．たとえば，英文概要が複数の矩形テキスト領域として認識されていた場合，そのすべての領域にのみ英文概要のラベルが付与されてはじめて正解となる．

テストセットに対する抽出結果を**表 5** の左 2 列に示す．表 5 には，書誌要素ごとのラベル付け精度と，“論文” として全書誌要素の抽出に成功した論文の割合を示している．また SVM とあるのは，5.2.1 項で述べた SVM によるラベル付け精度である．これは主に視覚的素性で構成された表 2 の素性を用いた実験であるが，CRF では情報処理学会論文誌で 97.56%，電子情報通信学会論文誌の英文誌で 95.23%の論文を正しくラベル付けできており，これらは SVM よりかなり良い結果といえる．各書誌要素の精度についてもおおむね良好といえるが，なかには電子情報通信学会論文誌の英文概要のように相対的に若干悪いものもある．

次に，表 2 の素性に加えて，言語的素性を用いることで精度の向上を図る．情報処理学会論文誌には特に目印となる文字列は存在しないが，電子情報通信学会論文誌には概要が “SUMMARY” ではじまり “key words:” に続くキーワードの並びで終わるという定められた書式が存在する．これらの文字列の有無を素性に加えることで，相対的に精度の悪かった英文概要のラベル付け精度の改善を図る．その抽出結果を表 5 の左から 3 列目に示す．比較すると英文著者の精度が若干下がったものの，狙いどおり英文概要の精度が向上し，論文全体としては約 2%精度が向上し 97.27%となった．言語的素性の追加によるこの精度向上は符号検定の結果，有意水準 5%で有意であった．一方言語的素性の追加により SVM の精度は，87.73%から 96.76%へと大きく向上した．SVM の精度の 96.76%は，CRF の 97.27%には及ばないものの，符号検定ではこの両者の差は有意とはならなかった．

CRF で言語的素性を考慮してもラベル付けに失敗した論文は 587 件中 16 件で，文字認識の誤りが原因であるものが 3 件あった．たとえば，書誌要素の論文種別は枠で囲まれているが，この枠の右辺がアルファベットの ‘i’ や ‘I’ として認識され，そこに誤って書誌要素ラベルを付与する例があった．一方レイアウト解析の誤りが原因で，英文概要の一部に正しくラベル付けできない例が多く見られた．電子情報通信学会論文誌のレイアウトでは，まず論文種別，英文表題，英文著者名が 1 段組で記述され，そのあとに英文概要と本文が 2 段組で記述されている．用いた OCR は基本的に紙面の上の方にあるものから出力するため，2 段組になっている部分では左の段と右の段が交互に出力される場合がある．たとえば，英文概要が複数の block に分割されて認識されると，英文概要（左の段），本文（右の段），英文概要（左の段）という順番で出力されることがある．このとき本文の後にくる英文概要の block にうまくラベルを付与できない例が多かった．また，1 つの単語だけで構成された小さい block に対するラベル付与を誤る例もあった．

5.2.3　line への書誌要素ラベル付け

5.2.2 項の実験では block をラベル付けの対象としたが，電子情報通信学会論文誌には 1 つの block 内に複数の書誌要素を含む論文がある（**表 6**）．付与するラベルは書誌要素と 1 対 1 に対応するため，そのような block には適切なラベル付けができず，これらの論文は 5.2.2 項の実験では除いていた．そこで本節では，電子情報通信学会論文誌について，ラベル付けの対象を block から line に変更した実験について述べる．図 2 に示したように line は block の下位のレイアウトタグである．line をラベル付けの対象とすることで，表 6 に

表 6　実験で使用した電子情報通信学会論文誌の論文
Table 6　The articles of IEICE Trans. Commun. used in the experiments.

	block 単位でラベル付けできる論文数	複数の書誌要素を含む block がある論文数	全論文数
トレーニングデータ	431	82	513
テストデータ	587	43	630

示した電子情報通信学会論文誌のすべての論文にラベル付けが可能となる．その結果トレーニングセットの論文数は 82 増え E87-B 巻の 513 件，テストセットのそれは 43 増え E88-B 巻の 630 件となった．また，5.2.2 項の実験で精度の向上に有効であった言語的素性も同様に使用した．実験の結果を表 5 の一番右の列に示す．

5.2.2 項の block に対するラベル付けと比較すると精度は下がるものの，表 6 に示した全論文の 96%以上の論文について正しくラベル付けができた．一方 SVM の論文全体のラベル付け精度は 64.44%と CRF に比べてかなり悪く，言語的素性を考慮しても line へのラベル付けで高精度を達成するのは難しいことが分かる．

表 5 の各列の実験条件で CRF と SVM を比較すると，差が有意でないものもあるがすべて CRF の方が良い結果となった．その理由の 1 つには，SVM では加えられなかったが，CRF では容易に考慮できたタグの bigram という素性の影響が考えられる．Peng らによると，CRF においてもこの bigram 素性の有無が精度に大きく影響することが分かっている[13]．

5.2.4　トレーニングデータのサイズと抽出精度

トレーニングデータのサイズが精度に及ぼす影響を評価するための実験を行った．これにより，トレーニングデータの妥当な大きさが分かれば実用上有用である．5.2.3 項の電子情報通信学会論文誌の実験と同じ条件で実験を行ったところ，**図 4** に示すトレーニングデータのサイズと精度の関係が得られた．これはトレーニングデータの論文数を 50 ずつ増加させたときの，それにともなう抽出精度の変化を示している．

このグラフから，トレーニングデータの論文数が 150 よりも多ければ論文単位でも 90%以上の精度が出ていることが分かる．

5.3　著者名の抽出精度

4 章で提案した文字へのラベル付け手法を用いて著者名領域から各々の著者名を抽出し，その抽出精度を求めた．実験には，情報処理学会論文誌 44 巻の 361 ファイルをトレーニングデータ，45 巻の 323 ファイルをテストデータとして使用した．ただし，実験に使用した

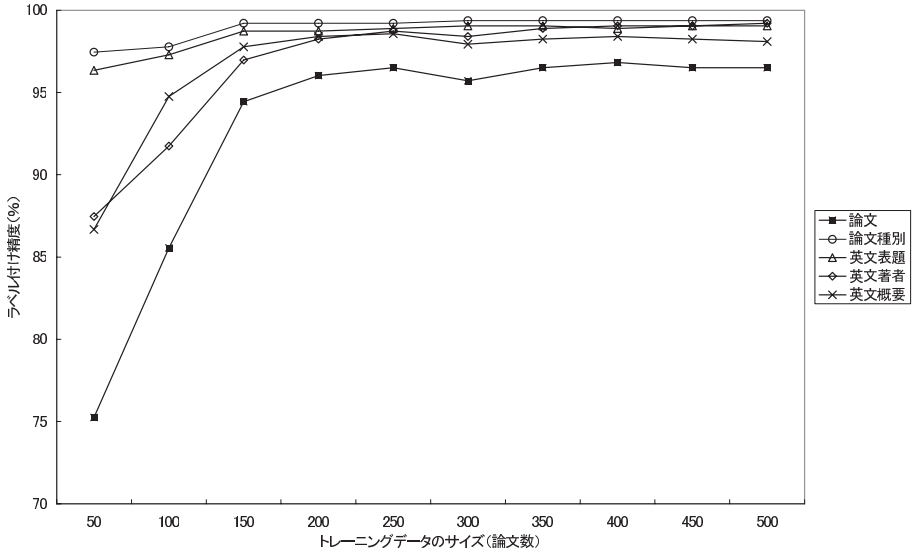

図 4　トレーニングデータサイズと書誌要素抽出精度
Fig. 4　Training data size vs. bibliography extraction accuracy.

これらのデータは，著者名領域の抽出に成功している block のみである．またこのトレーニングデータにより求まった著者名およびデリミタの最長文字列数は，それぞれ 9 と 11 である．これらは 4.1 節で説明した x と y に相当する．情報処理学会論文誌からは，和文著者と英文著者の両方を抽出しているが，著者名の抽出対象は和文著者のみとした．表 4 に示した素性のうち文字情報はデータのそれをそのまま用いたが，文字幅については 10 で割って小数点以下を切り捨てた値を素性とした．また実験では，表 4 に示した素性を次のような 4 つの組合せで使用した．すなわち，

(1)　Current: <c(0)> + <w(0)> + <t(-1),t(0)>
(2)　+Previous: Current + <c(-1)> + <w(-1)>
(3)　+Next: Current + <c(1)> + <w(1)>
(4)　+Both: +Previous + <c(1)> + <w(1)>

の組合せで，+Both が表 4 に示したすべての素性を利用する．

表 7 にテストデータにおける著者名抽出精度をまとめる．本実験では，2+pos-tag と+Both の素性を用いたときに，著者数に基づく算出で 99.82%という最高の抽出精度が得られた．

表 7　和文著者名抽出精度
Table 7　Extraction accuracy of Japanese authors' names.

	Current		+Previous		+Next		+Both		
	2-tag	2+pos-tag	2-tag	2+pos-tag	2-tag	2+pos-tag	2-tag	2+pos-tag	BIO-tag
著者名	99.18	99.18	99.45	99.73	99.27	99.45	99.36	99.82	99.18
論文	97.52	97.52	98.76	99.07	97.83	98.14	98.45	99.07	97.83

これを論文数に基づいてみると，99.07%の論文から正しくすべての著者を抽出できている．

素性の選択という点では，+Both が最高の結果を示したことから，表 4 に示したすべての素性が著者名抽出に有効であることが分かる．また+Previous と+Next の比較から，本実験では直前の文字の方が直後の文字よりもラベル付けに若干有効であった．

表 3 に示した 2 種類のタグ集合を比較すると，文字列中の位置情報を含む 2+pos-tag の方が位置情報を持たない 2-tag よりもおおむね良好な結果を示した．表 7 に示すうち唯一の例外は Current の場合で，この素性の組合せでは両者の抽出精度は同じであった．

最高の抽出精度を示した 2+pos-tag と+Both の組合せでは，延べ 1,092 名の著者名のうち 1,090 名の抽出に成功しており，誤りは 2 名のみであった．1 件は，著者名の最後の文字を誤ってデリミタとラベル付けしており，もう 1 件はデリミタ文字列の最初の文字を誤って著者名の最後の文字と判定していた．どちらの誤りも著者名文字列とデリミタ文字列の境界で発生しており，1 つは OCR の認識誤りを含む文字列であった．一方抽出に成功した論文 50 件を無作為に選んで調べたところ，そのうち 7 件は認識誤りを含んでいた．このうち 5 件では文書画像の汚れなどを文字と誤認識したと思われる '.' や '・' といった文字が誤って挿入されていた．残りの 2 件には置換誤りがあった．よって提案手法は，これらの認識誤りを含んでいてもロバストなラベル付けを行っていることが分かる．

さらに結果が最良であった+Both の素性を用いて，固有表現抽出などで一般的に用いられる BIO-tag によるラベル付け実験を行った．B，I，O のタグはそれぞれ，著者名の先頭の文字，著者名の先頭以外の文字，デリミタの文字に割り当てられる．その結果を表 7 の一番右の列に示す．タグの種類数では，BIO-tag は 2-tag より多く，2+pos-tag より少ないが，抽出精度は 2-tag よりもわずかに劣る結果となった．また+Both の 3 つの結果（"論文"）について符号検定を行ったが，いずれの組合せも有意水準 5%で有意な差とはならなかった．

6. 考　　察

OCR 処理された学術論文のタイトルページから書誌要素を抽出する研究については，比較的近年のものに阿辺川らの研究がある[1]．彼らは，日本語および英語で書かれた様々な論文誌の論文を対象に，SVM を用いて書誌要素を抽出する手法を提案した．本研究との違いは，もともとテキスト情報を持つ PDF ファイルの論文を収集して利用している点と，論文誌ごとに分けず複数の論文誌の論文を対象とする学習を行っている点である．前者は OCR が不要ということを意味しており，本研究より条件が易しく，一方後者は条件が厳しい．これらの条件の違いはあるが，論文単位の抽出精度が 69.2%と報告されている．彼らはまた SVM を用いた参考文献中の書誌情報抽出も提案しており，和文，英文でそれぞれ 74.8%および 81.6%の精度を達成している．

参考文献抽出では，Takasu らは 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手法を提案し，様々な論文誌の論文に適用した[15]．彼らは最初に参考文献領域を抽出し，そこから個々の参考文献を抜き出した．これは我々が提案した 2 段階の抽出法と処理手順が似ている．また彼らの手法は HMM に基づいており，置換や挿入，欠落誤りを考慮している．彼らは情報処理学会論文誌を対象に実験を行い，OCR の認識精度が 97.85%のとき，89.99%という参考文献の抽出精度を示した．なおこの抽出精度は，参考文献領域の抽出精度とそこからの参考文献の抽出精度の積である．参考文献抽出には本稿の扱う問題とは異なる困難さはあるものの，表 5 と表 7 に示した精度は彼らのそれよりも高い．ただし彼らは，抽出した参考文献からの書誌要素抽出は行っていない．

CRF を用いた書誌情報抽出では，Peng らが英語の学術論文のタイトルページおよび個々の参考文献からの書誌要素抽出を行っている[13]．彼らが抽出対象としたデータは文書画像ではなくテキスト情報であるため，本稿のように OCR を必要としないが，レイアウトに関する素性が少なく，そもそも紙面上の位置に関する情報はない．このような点が本稿とは異なるが，CRF の正則化の問題や素性の選択などについて，実験に基づく知見が述べられて

おり示唆に富む．また彼らの実験のエラー解析には本稿と共通するものがあり，たとえばラベル付けの誤りが異なる書誌要素間の境界で発生することが多いというのは，5.3 節で述べた知見と一致する．なお彼らのタイトルページからの書誌要素抽出精度は，論文単位で 72.4%と報告されている．

7. ま と め

本稿では，学術論文 OCR テキストからその論文ファイルの書誌要素を抽出する手法を提案した．まず矩形テキスト領域に対する書誌要素ラベルの付与により，言語およびレイアウトの異なる 2 種類の論文誌に対して，視覚的素性とタグの bigram 素性を用いて高精度な書誌要素抽出が可能であることを示した．すなわち，情報処理学会論文誌の和文誌で 97.56%，電子情報通信学会論文誌の英文誌で 95.23%の精度で書誌要素を抽出することができた．さらに電子情報通信学会論文誌について言語的素性を追加したところ，抽出精度が 97.27%に改善された．このことから，特徴的な文字列を含む書誌要素については，これらを言語的素性として使用することで精度が向上することを確認できた．また，文字へのラベル付けによる和文著者名抽出では，ラベル付けを行う文字の前後の文字情報を使用することで，99%以上の著者名を正確に抽出することに成功した．

今後の課題としては，本稿であまり用いなかった言語的素性についての検討があげられる．たとえば，平仮名や片仮名，アルファベットといった文字種ごとの出現の有無などは，論文誌ごとに個別に対応する必要がなく素性として扱いやすいと考えられる．またデータの作成コストは高いが実験対象の論文誌を追加して，抽出すべき書誌要素やその抽出に有用な素性について包括的な検討を行い，汎用的で高精度な書誌情報抽出を実現したい．

謝辞　本研究の一部は，科学研究費補助金基盤研究（B）（課題番号：19300032）および国立情報学研究所公募型共同研究の援助による．

参 考 文 献

1) 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).
2) 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).
3) Bunke, H. and Wang, P.: Handbook of Character Recognition and Document Image Analysis, *World Scientific* (1997).
4) Chang, C.-C. and Lin, C.-J.: LIBSVM: A Library for Support Vector Machines (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
5) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. EMNLP 2004* (2004).
6) 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).
7) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data, *Proc. 18th International Conference on Machine Learning*, pp.282–289 (2001).
8) Nagy, G., Seth, S. and Viswanathan, M.: A Prototype Document Image Analysis System for Technical Journals, *IEEE Computer*, Vol.25, No.7, pp.10–22 (1992).
9) Ohta, M. and Takasu, A.: CRF-based Authors' Name Tagging for Scanned Documents, *Proc. JCDL'08*, pp.272–275 (2008).
10) Ohta, M., Yakushi, T. and Takasu, A.: Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields, *Proc. ICDIM 2008*, pp.99–104 (2008).
11) Ohta, M., Yamasaki, S., Yakushi, T. and Takasu, A.: Authors' Names Extraction from Scanned Documents, *Proc. ICDIM 2007*, pp.67–72 (2007).
12) Parmentier, F. and Belaid, A.: Bibliography References Validation Using Emergent Architecture, *Proc. ICDAR'95*, pp.532–535 (1995).
13) Peng, F. and Mccallum, A.: Accurate Information Extraction from Research Papers Using Conditional Random Fields, *Proc. HLT-NAACL 2004*, pp.329–336 (2004).
14) Story, G.A., O'Gorman, L., Fox, D., Schaper, L.L. and Jagadish, H.V.: The Right-Pages Image-based Electronic Library for Alerting and Browsing, *IEEE Computer*, Vol.25, No.9, pp.17–26 (1992).
15) Takasu, A. and Aihara, K.: Quality Enhancement in Information Extraction from Scanned Documents, *Proc. DocEng '06*, pp.122–124 (2006).
16) Takechi, M., Tokunaga, T. and Matsumoto, Y.: Chunking-based Question Type Identification for Multi-Sentence Queries, *Proc. SIGIR 2007 Workshop on Focused Retrieval* (2007).
17) Wong, K.Y., Casey, R.G. and Wahl, F.M.: Document Analysis System, *IBM Journal of Research and Development*, Vol.26, No.6, pp.647–656 (1982).
18) Zhao, H., Huang, C.N. and Li, M.: An Improved Chinese Word Segmentation System with Conditional Random Field, *Proc. 5th SIGHAN Workshop on Chinese Language Processing*, pp.162–165 (2006).

(平成 20 年 12 月 20 日受付)
(平成 21 年 4 月 7 日採録)

(担当編集委員　安形 輝)

薬師　貴之

2007 年岡山大学工学部情報工学科卒業．2009 年同大学大学院自然科学研究科博士前期課程修了．同年 NEC システムテクノロジー株式会社入社．大学院在学中に文献データベース，書誌要素抽出の研究に従事．

太田　　学（正会員）

1994 年東京大学工学部電気工学科卒業．1999 年同大学大学院工学系研究科電気工学専攻博士課程修了．博士（工学）．東京都立大学工学研究科助手を経て 2005 年岡山大学大学院自然科学研究科助教授．2007 年より同研究科准教授．Web 情報検索ならびに電子図書館の研究に従事．電子情報通信学会，日本データベース学会，IEEE 各会員．

高須　淳宏（正会員）

1984 年東京大学工学部航空学科卒業．1989 年同大学大学院工学系研究科博士課程修了．工学博士．同年学術情報センター研究開発部助手．同センター助教授．国立情報学研究所助教授を経て 2003 年より同研究所教授．データ工学，特にデータ解析と解析モデルの学習の研究に従事．電子情報通信学会，人工知能学会，日本データベース学会，ACM，IEEE 各会員．