Vol.2010-DBS-151 No.18
2010/11/13

学術論文文書画像からのページレイアウトに依存しない自動書誌要素抽出

井上諒平[†1]　太田学[†1]　高須淳宏[†2]

国立情報学研究所の電子図書館 NII-ELS は，国内の主要な学術論文を網羅しており，その蔵書検索には著者名等の書誌情報が利用される．NII-ELS では論文文書画像を蓄積しているため，書誌情報は文書画像からなるべく自動で抽出することが望ましい．現在の文書画像処理技術により一定の抽出精度は達成しているが，学習のため人手で書誌要素を抽出した論文データを学術雑誌ごとに用意する必要がある．しかし 1000 雑誌以上を所蔵する NII-ELS では，各雑誌ごとにこの学習データを用意するコストは無視できない．そこで本研究では，書誌要素抽出対象の雑誌とは異なる雑誌を学習データに用いて効率的に書誌要素を抽出する手法を提案する．提案手法は，論文タイトルページの各行に対して，雑誌のレイアウトに依存しない文字列等の情報を利用して書誌ラベルを付与する．

Automatic Extraction of Bibliographic Elements from Scanned Academic Articles without Using Page Layout

Ryohei Inoue,[†1] Manabu Ohta[†1]
and Atsuhiro Takasu[†2]

NII-ELS developed by the National Institute of Informatics is a digital library which stores scanned document images of a wide variety of academic journals in Japan. Bibliographic information is indispensable for searching such a digital library, hence, automatic extraction of bibliographic data from the images is very important. Therefore, Yakushi et al. proposed an automatic method of extracting bibliographies for academic articles scanned with OCR markup. Although they achieved excellent extraction accuracies for some journals, they needed a substantial amount of training data obtained through costly manual extraction of bibliographies from document images. We cannot ignore this cost because NII-ELS stores more than a thousand journals. This paper, therefore, proposes an automatic bibliography extraction method to use training data collected from journals different from a target journal. The proposed method labels each text line on an article's title page as appropriate bibliographic names by using linguistic information which is independent of page layout varying by journal.

1. はじめに

国立情報学研究所がサービスを提供する電子図書館 NII-ELS には，日本国内の主要な学術論文の文書画像が蓄積されている．電子図書館の蔵書検索には，表題や著者名などの書誌情報を利用することが通例だが，これらの書誌情報をデータベースへ人手で入力しようとすると膨大なコストがかかる．よって，論文の書誌情報を可能な限り自動で抽出する文書解析技術が必要とされている．光学文字認識（OCR）[1] などの技術により，文書画像をテキストデータに変換することは可能だが，得られたテキストからの自動書誌要素抽出は容易ではない．

先行研究となる薬師らの自動書誌要素抽出法[11),12)] では，あらかじめ人手で書誌要素を抽出した論文データをトレーニングデータとして学習を行い，学術雑誌毎に差はあるものの 93 %から 98 %の高い抽出精度を達成している．しかし彼らの方法では，学習のためのトレーニングデータを個々の学術雑誌ごとに数百件用意する必要があった．NII-ELS には 1000 を超える学術雑誌が収録されており，その全ての雑誌に対して何百件もの学習データを作成するとなると，その時間的・金銭的コストは無視できない．

そこで本研究では，学術雑誌毎に異なる論文のレイアウトではなく，記述されている文字情報を利用して書誌要素を自動抽出する手法を提案する．提案手法は，書誌要素抽出対象の雑誌とは異なる雑誌を用いて学習した場合に，一定の精度で書誌要素を抽出することを目標とする．書誌要素のラベル付けには，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）[5] を利用する．

本研究の目的は，正解データが用意されていない未知の論文誌に遭遇した際のラベル付け

†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

コストを軽減することにある．よって提案手法は薬師らの手法より精度が悪くなっても，既に正解データが用意されている既存の論文誌を学習に用い，未知の論文誌のデータから自動で書誌要素を抽出する．自動抽出不可能な部分については人手で抽出し，これにより十分な量の正解データが確保できれば，それ以降は高い精度を達成している薬師らの手法に切り替えて書誌要素抽出を行えばよい．

本稿の構成は次の通りである．2 節で，学術論文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[3]．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．我々は文書画像を対象に書誌要素を抽出するが，阿辺川らの研究では論文のテキスト情報を持つ PDF ファイルを XML 形式に変換したものを，入力データとして使用する．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．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．

藤尾ら[2] は，正準判別分析とレイアウト型の DP マッチングを用いた書誌情報抽出を提案している．彼らの手法では，抽出した各文字行に対して書誌情報らしさのスコアを正準判別分析を用いて計算し，全体のコストが最小になるような書誌要素の組み合わせを DP マッチングにより計算する．彼らも PDF データを対象に書誌情報の抽出実験を行っているため本研究との単純比較はできないが，三つの論文誌に対して文書単位での抽出精度が 75 %〜95 %となっている．ただし行単位での精度では，再現率，適合率共にいずれの論文誌でも 98 %以上を達成している．また，この手法では必要とするトレーニングデータ量が少なく，あらかじめ抽出対象と同じ論文誌の学習サンプルを 10 文書程度用意すれば十分だとされている．

2.2 学術論文文書画像からの参考文献抽出

高須らの研究に，OCR 処理された学術論文文書画像の参考文献領域から参考文献を抽出するものがある[4]．彼らはまず論文全体から参考文献領域を抽出し，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 による文字認識誤りも考慮されている．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，論文表題など詳細な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）[5] 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[7),8)] 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[9]．そのため CRF は広範な分野で利用実績がある[6),7),10)]．

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を以下のように与える．

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

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 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．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，本研究では論文の各行である．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．例えば論文のレイアウト情報は視覚的素性，文字情報は言語的素性と言え，CRF はこれらの有用な情報を全て素性として利用してラベル付けを行うことができる．これが例えば通常の HMM では，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 入出力データ

本稿では，学術論文のタイトルページから書誌要素を抽出する．なぜなら，学術論文では主要な書誌要素の多くがタイトルページに集中して出現するからである．例えば表題，著者名といった書誌要素はほとんど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．それに加えて，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．

このような学術論文のタイトルページの文書画像に対して OCR 処理を施したテキストデータを，提案手法の入力データとして扱う．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，文書中の行（line），単語（word），文字（char）にもタグが位置情報とともに与えられている．ただし，本研究ではこれらのレイアウトに関わる位置情報については敢えて考慮しない．論文の書式はそれぞれの学術雑誌ごとに個別に定められており，当然ながら文字の位置や大きさなどのレイアウト情報も論文によって様々である．そのため，抽出対象とは異なる学術雑誌をトレーニングデータとする場合，そうした視覚的な情報が抽出の妨げになると考えられるからである．

視覚的な情報ではなく，本研究では論文に記述された文字情報に着目する．具体的には，論文の各行に含まれる文字そのものや文字の種類，書誌要素の判別に有用な特徴的な文字列が存在するかどうか，などの情報を利用する．出現する文字のような言語的情報は文書レイアウトの影響を受けないので，異なる学術雑誌を用いた学習にも有効と考えられる．こうした言語的な情報を素性として活用し，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 形式データとする．その際，書誌要素のラベル付けは入力テキストの行（line）単位で行う．本稿で抽出対象とした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，単語などより小さな要素に対してラベル付けを行う必要はなかった．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 1 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あるが，図 1 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付与された書誌要素ラベルのことであり，これらはそれぞれ和文表題，和文著者名を表している．

3.3 抽出する書誌要素

本稿では，情報処理学会論文誌の論文（IPSJ），電子情報通信学会論文誌の和文論文（IEICE-J），そして人工知能学会誌の論文（JSAI）の三種類を対象に書誌要素抽出実験を行った．これらの論文誌に含まれる主な書誌要素の一覧と，それに対応する書誌要素ラベルを表 1 に示す．表 1 から分かるように，論文中に出現する書誌要素は，抽出対象とする論文誌の種類によって異なる．例えば IPSJ には和文英文両方の表題，著者名，概要といった書誌要素が含まれているが，IEICE-J には英文概要が含まれず，代わりに和文キーワードが追加されている．本研究では抽出対象とは異なる論文誌をトレーニングデータとして利用するため，実験に用いる全ての論文誌に共通して現れる書誌要素のみを抽出対象とする．

3.4 素性テンプレート

3.4.1 薬師らの素性テンプレート

CRF によるラベル付けに利用する論文文書の特徴の一覧を素性テンプレートと呼ぶ．本研究で比較対象とする，薬師らの手法で用いられた素性テンプレート[12] 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しており，例えば x(0) はある行のラベル付けにその行自身の X 座標を考慮するという意味になる．これが例えば x(-1) や x(1) であれば前後の行の X 座標を考慮するという意味である．薬師らの研究や本研究では，このような素性テンプレートに基づいて素性関数が生成される．

薬師らの手法は，素性テンプレートに行の XY 座標や文字のサイズなどを含め，OCR テキスト中に含まれるレイアウト情報を素性として利用した．この方法は，書誌要素抽出対象

```
(省略)
<block>
  <j-title>
    <line> 学術論文文書画像からの < /line>
    <line> ページレイアウトに依存しない自動書誌要素抽出 < /line>
  </j-title>
< /block>
<block>
  <j-authors>
    <line> 井上諒平 † 1 太田学 † 1 高須淳宏 † 2< /line>
  </j-authors>
< /block>
<block>
  <j-abstract>
    <line> 国立情報学研究所の電子図書館 NII-ELS は，国内の....< /line>
    <line> り，その蔵書検索には著者名等の書誌情報が利用さ....< /line>
    <line> 像を蓄積しているため，書誌情報は文書画像からな....< /line>
    <line> しい．現在の文書画像処理技術により一定の抽出精....< /line>
    <line> 人手で書誌要素を抽出した論文データを学術雑誌ご....< /line>
    <line>1000 雑誌以上を所蔵する NII-ELS では，各雑誌ごと....< /line>
    <line> ストは無視できない．そこで本研究では，書誌要素....< /line>
    <line> 学習データに用いて効率的に書誌要素を抽出する手....< /line>
    <line> タイトルページの各行に対して，雑誌のレイアウト....< /line>
    <line> 用して書誌ラベルを付与する．< /line>
  </j-abstract>
< /block>
<block>
  <e-title>
    <line>Automatic Extraction of Bibliographic Elements< /line>
    <line>from Scanned Academic Articles< /line>
    <line>without Using Page Layout< /line>
  </e-title>
< /block>
<block>
  <e-authors>
    <line>RYOHEI INOUE, † 1MANABU OHTA † 1< /line>
    <line>and ATSUHIRO TAKASU † 2< /line>
  </e-authors>
< /block>
(省略)
```

図 1 出力データの例

の論文誌とトレーニングデータの論文誌が同じ場合は有効であるが，トレーニングデータとテストデータの論文誌が異なる場合に，書誌要素の抽出精度が大幅に低下する．これは論文のレイアウトが各論文誌ごとに個別に定められており，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，論文誌が異なれば収録されている書誌要素の種類そのものが異なることも珍しくない．異なる種類の論文誌の論文データを学習に利用するにはこれらの点を考慮しなければならない．

3.4.2 提案する素性テンプレート

本研究では，抽出対象と異なる論文誌の論文データをトレーニングデータとして，高精度な書誌要素抽出を目指す．そのため薬師らの利用した視覚的素性は利用せず，言語的素性のみを利用する．提案手法で用いる素性テンプレートを表 3 に示し，それぞれの素性について以下で詳しく説明する．

表 1 書誌要素ラベル

書誌要素	ラベル	IPSJ	IEICE-J	JSAI
論文種別	j-class_JSAI	-	-	○
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	-	○	-
和文所属	j-affiliation	-	-	○
英文表題	e-title	○	○	○
英文著者名	e-authors	○	○	○
英文概要	e-abstract	○	-	-
英文キーワード	e-keywords	-	-	○
英文所属	e-affiliation	-	-	○
email と URL	Email&URL	-	-	○
その他	other	○	○	○

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

文字の種類

学術論文では，例えば日本語の著者名はほぼ漢字のみで記述されており，英語の概要などは英数字のみで記述されているなどの特徴が存在する．そうした特徴を考慮するために，英数字，漢字，平仮名片仮名，記号が使用されている割合をそれぞれ算出して利用する．具体的には，まず各行中でこれらの文字種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する．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．

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	<lalphabet(0)>	line 内の英数字の割合
	<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
	<lcuni1(0)>～<lcuniK(0)>	line 内の文字 unigram
	<lcbi1(0)>～<lcbiK(0)>	line 内の文字 bigram
bigram	<l(-1),l(0)>	ラベルの遷移

表 4 文字 N-gram 素性の具体例

原文	学術論文文書画像からの ページレイアウトに依存しない自動書誌要素抽出
文字 unigram	学 術 論 文 文 書 画 像 か ら の nil nil nil nil ペ ー ジ レ イ ア ウ ト に 依 存 し な い 自
文字 bigram	学術 術論 論文 文文 文書 書画 画像 像か から らの の nil nil nil nil nil ペー ージ ジレ レイ イア アウ ウト トに に依 依存 存し しな ない い自 自動

特徴的な文字列

文書中の文字列を解析すると，それだけで書誌要素の判別に利用できそうな特徴的な文字列が含まれていることがある．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．そのような特徴的な文字列をいくつか列挙し，それらの文字列が行の中に出現するかどうかを判定する．薬師らは「あらまし」や「キーワード」など 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 18 種類の文字列を利用する．

文字 N-gram 素性

前述の「文字の種類」素性では行中の文字種のみを考慮しているが，さらに論文中に記述された文字情報そのものも素性として扱う．その中でも，各行に含まれる文字を一文字ごとに区切った文字 unigram，二文字ごとに区切った文字 bigram を素性とする．ただし各行に含まれる文字数は様々であるため，一行の文字数により素性の数が変化しないよう次のように処理する．

- 各行の先頭から数えて K 個の N-gram を素性とし，残りは破棄する．
- N-gram 素性の数が K に満たない場合，空いた文字スペースを「nil」で補填する．

本稿の表題部分の 2 行を例として，K=15 とした場合の文字 unigram 及び文字 bigram を表 4 に示す．

ラベル bigram 素性

この素性は，各行に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ある．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．

以上，提案手法では大別して 4 種類の素性を利用して学習及びテストデータへのラベル付けを行う．この 4 種類の素性の中で文字の種類，特徴的な文字列，ラベル bigram といった素性はこれまでの視覚的素性と言語的素性を両方利用した実験（薬師らの手法や著者らの研究[13]）の中で既に利用されており，一定の有効性は確認されている．よって，4 節では主に文字 N-gram 素性の効果を確認するための実験を行う．

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，工藤が作成した CRF++[⋆1] を利用して書誌要素ラベル付けを実行した．CRF++は，トレーニングデータによる CRF の学習，及びテストデータのラベル付けなどの処理を行うことができるオープンソースのソフトウェアである．このラベル付け結果から，それぞれの書誌要素毎のラベル付け正解率，及び論文全体に対するラベル付け正解率をそれぞれ算出する．

また，実験に利用するデータとして以下の三種類の論文誌の論文データを用意した．

- 情報処理学会論文誌（IPSJ）：2003 年分　479 件
- 電子情報通信学会論文誌-和文（IEICE-J）：2003-2005 年分　324 件
- 人工知能学会誌（JSAI）：1997,1999,2000 年分　219 件

4.1 提案手法の有効性評価

まず，3.4 節で説明した文字 N-gram 素性について，定数 K を定めるための予備実験を行った．文字 unigram 素性と文字 bigram 素性のうち，ここでは unigram 素性のみを使用し，unigram 素性の数 K を 10 から 90 まで変化させて書誌要素の抽出精度を求めた．トレーニングデータを IPSJ（479 件），テストデータを IEICE-J（324 件）としたときの抽出精度が表 5 であり，逆にトレーニングデータを IEICE-J，テストデータを IPSJ とした実験での抽出精度が表 6 である．なお，ここでの抽出精度は論文中に含まれている全ての書誌要素が正確に抽出できた割合を指す．表 5 の結果では，K が小さい場合には効果が少なく，K が概ね 50 を超えるとそれ以降は抽出精度に大きな変化はない．一方，表 6 の結果で

⋆1 http://crfpp.sourceforge.net/

表 5 素性とする文字 N-gram の数と抽出精度（1）

K	10	30	50	70	90
抽出精度（%）	0.31	13.58	28.70	27.78	31.48

表 6 素性とする文字 N-gram の数と抽出精度（2）

K	10	30	50	70	90
抽出精度（%）	2.71	8.77	1.46	0.84	0.63

は K が 30 の場合に最も精度が高い．しかし表 6 に示す抽出精度は後の実験結果の考察に示す理由により全体的に大変低いので，$K = 50$ として他の評価実験を行った．

実験では，文字 N-gram 素性の有効性を確認するために，文字 N-gram 素性を利用しない場合，文字 unigram を利用する場合，文字 bigram を利用する場合，そしてその両方を利用する場合の抽出精度をそれぞれ求めて比較する．また，薬師らの手法を用いた場合の書誌要素抽出実験も同様に行い，それと比較して提案手法全体の有効性を評価する．

なお，本稿の 3.2 節で説明したように，提案手法と薬師らの手法では素性として利用する「特徴的な文字列」に若干の違いがある．今回の実験の目的は文字 N-gram 素性の有効性を確認することであるので，それ以外の素性の内容に差が生じるのは好ましくない．そのためこの点については薬師らの手法を修正して，提案手法と同様に 18 種類の特徴文字列をラベル付けに利用することとした．

4.1.1 二種類の論文誌を用いた実験

まず，IPSJ と IEICE-J の二種類の論文誌を利用し，一方をトレーニングデータ，もう一方をテストデータとして書誌要素の抽出精度を比較する．この二種類の論文誌に共通する書誌要素は和文表題，和文著者名，和文概要，英文表題，英文著者名であるので，これら 5 種類の書誌要素を抽出対象とし，それ以外を「その他」として扱う．IPSJ をトレーニングデータ，IEICE-J をテストデータとした場合の抽出精度比較を表 7 に示す．表中の「和文表題」から「その他」までの項目は各書誌要素の抽出精度を，All は論文全体の抽出精度を表している．

薬師らの手法と比較した場合，視覚的素性を除外し，文字 N-gram も利用しない場合は全体的に精度が悪化している．ただし，本研究で提案した文字 N-gram を利用する手法ではいずれも精度が向上しており，提案手法の有効性が確認できた．特に英文表題や英文著者名の抽出精度は 10 %前後から 90 %前後へと，論文全体の抽出精度は 1 %未満から 30 %前後へと改善されている．また，素性とする文字 N-gram の組み合わせを比較すると，文字 bigram よりも文字 unigram を素性として利用したほうが有効であることが分かる．さらに文字 unigram と文字 bigram を両方使用した場合，文字 unigram を利用した手法と同等かそれ以上の抽出精度を達成している．

表 7 提案手法の抽出精度（%）の比較（トレーニングデータ:IPSJ, テストデータ:IEICE-J）

書誌要素	薬師らの手法	提案手法（文字 N-gram なし）	提案手法（文字 unigram）	提案手法（文字 bigram）	提案手法（文字 unigram +文字 bigram）
和文表題	55.25	66.36	80.86	79.63	82.41
和文著者名	42.90	74.69	91.67	87.65	91.36
和文概要	6.79	0.00	40.12	38.27	42.90
英文表題	12.96	26.85	96.60	91.98	95.06
英文著者名	9.65	26.23	89.51	83.95	88.58
その他	5.56	0.00	31.17	28.40	33.95
All	0.62	0.00	28.70	26.23	32.10

一方，IEICE-J をトレーニングデータ，IPSJ をテストデータとした際の精度比較を表 8 に示す．表 8 では，表 7 とは異なり，薬師らの手法と比較して本稿で提案した手法はいずれも抽出精度が悪化していることが分かる．英文著者名のように文字 N-gram を用いることで抽出精度が大きく向上している書誌要素もあるが，英文表題や「その他」の精度が低すぎるために全体の抽出精度も悪くなっている．

表 8 のラベル付けの結果を詳しく調べた結果，IPSJ のみに出現する英文概要を表す部分に誤って Eti ラベルが付与されている例が非常に多く見られた．トレーニングデータである IEICE-J には「英文表題」と「英文著者名」の書誌要素しか含まれないため，英語で記述された部分はその二つの書誌要素のどちらかであると判断され，結果としてそのような誤りが頻発したものと考えられる．本実験では，英文概要の箇所には「その他」ラベルを付与すれば正解としているが，IEICE-J における「その他」部分は基本的に日本語で記述されているため，英文概要に etc ラベルを付与するのは難しいと考えられる．

4.1.2 三種類の論文誌を用いた実験

次に IPSJ，IEICE-J，JSAI の三種類の論文誌を利用し，そのうち二種類をトレーニングデータ，残りの一種類をテストデータとして書誌要素の抽出実験を行う．この実験で抽出対象とする書誌要素は，4.1.1 節の実験で抽出対象とした 6 種類から「和文概要」を除外した 5 種類の書誌要素である．IPSJ と IEICE-J をトレーニングデータ，JSAI をテストデータとした場合の抽出精度比較を表 9 に示す．同様に，IPSJ と JSAI をトレーニングデータ，

表 8 提案手法の抽出精度（%）の比較（トレーニングデータ:IEICE-J, テストデータ:IPSJ）

書誌要素	薬師らの手法	提案手法 (文字 N-gram なし)	提案手法 (文字 unigram)	提案手法 (文字 bigram)	提案手法 (文字 unigram +文字 bigram)
和文表題	80.38	58.87	85.39	82.67	88.31
和文著者名	78.91	57.20	89.56	82.88	89.14
和文概要	69.52	33.40	88.10	85.59	83.51
英文表題	31.94	6.05	2.71	0.42	1.46
英文著者名	8.56	0.21	96.45	94.15	96.87
その他	7.10	0.21	1.46	0.21	0.63
All	6.89	0.21	1.46	0.21	0.63

IEICE-J をテストデータとした場合の精度比較を表 10 に，IEICE-J と JSAI をトレーニングデータ，IPSJ をテストデータとした場合の精度比較を表 11 に示す．

表 10 及び表 11 の実験では，提案手による精度向上が確認できた．中には精度が悪化している書誌要素もあるが，論文全体の抽出精度は向上している．また，ここでは提案手法の中でも文字 unigram 素性のみを利用する手法が最も高い抽出精度を示している．一方表 9 の実験結果を見ると，ほとんどの書誌要素の抽出精度が大幅に悪化してしまっていることが分かる．

表 9 の実験が極端に悪い理由としては，論文中の書誌要素の出現順序の違いによる影響が挙げられる．トレーニングデータである IPSJ と IEICE-J ではまず和文の表題・著者名が記述された後に英文の表題・著者名が記述されるのに対して，テストデータである JSAI では和文と英文両方の表題が記述されてから和文と英文の著者名が記述される．こうした書誌要素の出現順に関する情報が「ラベル bigram 素性」に反映され，書誌要素抽出に悪影響を与えてしまったものと考えられる．

以上のような特徴により表 9 の JSAI に対するラベル付け精度は著しく低いものとなったが，逆に JSAI をトレーニングデータの一部として用いている表 10，表 11 の実験ではこれほどの大幅な精度低下は見られない．特に表 10 の実験では，トレーニングデータの 3 分の 2 以上を IPSJ が占めていることもあって，論文全体で 7 割を超える抽出精度を達成している．今回の実験では 2 種類の論文誌をトレーニングデータとして利用したが，さらに論文誌の種類を増やして実験を行い，提案手法の有効性について検討したい．

4.2 必要とするトレーニングデータ量

表 7 で最も抽出精度の良かった文字 unigram と文字 bigram 素性を両方利用する手法について，トレーニングデータ量と抽出精度の関係を調べる実験を行った．テストデータについては IEICE-J（324 件）をそのまま用い，トレーニングデータである IPSJ の論文数を 20 件から 470 件まで変化させて抽出精度の変化を見た．その結果，論文全体としてはトレーニングデータを 70 件利用した時に抽出精度が最も高くなることを確認できた．比較的高精度で抽出できている和文表題や和文著者名などの書誌要素については，トレーニングデータ

表 9 提案手法の抽出精度（%）の比較（トレーニングデータ:IPSJ+IEICE-J, テストデータ:JSAI）

書誌要素	薬師らの手法	提案手法 (文字 unigram)	提案手法 (文字 bigram)	提案手法 (文字 unigram +文字 bigram)
和文表題	60.27	0.46	0.46	0.46
和文著者名	6.39	4.11	4.11	4.11
英文表題	63.93	7.31	2.74	15.53
英文著者名	1.83	1.83	1.83	1.83
その他	1.37	0.00	0.00	0.00
All	0.46	0.00	0.00	0.00

表 10 提案手法の抽出精度（%）の比較（トレーニングデータ:IPSJ+JSAI, テストデータ:IEICE-J）

書誌要素	薬師らの手法	提案手法 (文字 unigram)	提案手法 (文字 bigram)	提案手法 (文字 unigram +文字 bigram)
和文表題	75.62	79.32	74.38	80.86
和文著者名	78.40	89.81	84.57	91.36
英文表題	2.47	94.14	91.67	95.99
英文著者名	0.93	94.14	85.80	89.81
その他	33.64	81.79	71.91	75.93
All	0.62	75.31	64.51	71.60

表 11 提案手法の抽出精度（%）の比較（トレーニングデータ:IEICE-J+JSAI, テストデータ:IPSJ）

書誌要素	薬師らの手法	提案手法 (文字 unigram)	提案手法 (文字 bigram)	提案手法 (文字 unigram +文字 bigram)
和文表題	92.48	84.76	82.05	88.73
和文著者名	93.95	82.46	79.54	85.59
英文表題	27.35	74.32	18.16	17.33
英文著者名	10.02	86.01	83.30	86.85
その他	21.50	61.17	12.73	13.36
All	8.98	54.49	11.48	12.11

が増加するほど抽出精度が上がった．しかし,「和文概要」や「その他」のように抽出精度が低い書誌要素については，トレーニングデータを 70 件より増やしても精度向上には結びつかなかった．論文全体の抽出精度は最も精度が低い書誌要素によって決まるので，このような書誌要素の抽出精度を改善させない限りは論文全体の精度向上は見込めない．

なお，この実験では学習サンプルは無作為に選択しているが，能動サンプリングの手法を取り入れることでより少ないトレーニングデータで効率的に学習できる見込みがある．能動サンプリングの基本的な考え方は，提案する CRF によるラベル付けが難しい論文データを優先的に選んで人手でラベル付けをし，そのデータを用いて学習を行うというものである．我々は，この「ラベル付けの難しさ」を表す指標をいくつか提案し，それに基づいてトレーニングデータを選択することで，無作為に選んだ場合に比べて半分以下のデータ量で学習可能であることを示した[14]．

5. ま と め

本稿では，学術論文の文書画像から書誌要素を自動抽出する際，学習データの作成コストを軽減するために，抽出対象とは異なる種類の学術雑誌を用いて学習を行い，書誌要素を抽出する手法を提案した．提案手法は，テキストの位置や大きさといった視覚的な情報よりも，そこに記述された文字情報そのものに注目した素性テンプレートを用いて CRF を学習する．特に本稿では文書中の文字 N-gram を素性として利用し，レイアウトの異なる別の雑誌をトレーニングデータとして，どの程度書誌要素抽出ができるか評価した．評価実験の結果，文字 N-gram を利用することで薬師らの手法に比べて多くの書誌要素で抽出精度の向上を確認できた．ただし，学習データとテストデータとなる論文誌の組み合わせによっては，特定の書誌要素の抽出が非常に困難となる場合があった．今後は，学習データとして利用する論文誌をさらに追加して提案手法の有効性を評価していきたい．

参 考 文 献

1) Bunke, H. and Wang, P.: Handbook of Character Recognition and Document Image Analysis, World Scientific (1997).
2) 藤尾 正和, 永崎 健, 高橋 寿一: 正準判別分析とレイアウト DP を用いた学術文献からの書誌情報抽出, DEIM Forum 2010, A9-3 (2010).
3) 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).
4) Takasu, A. and Aihara, K.: Quality Enhancement in Information Extraction from Scanned Documents, In Proc. of DocEng '06, pp.122-124 (2006).
5) 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).
6) 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).
7) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, In Proc. of EMNLP 2004, pp.230-237 (2004).
8) 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).
9) 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).
10) 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).
11) 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).
12) 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).
13) 井上 諒平, 太田 学, 高須 淳宏: 学術論文文書画像からの自動書誌要素抽出，DEIM Forum 2010, F7-2 (2010).
14) Ohta, M., Inoue, R. and Takasu, A.: Empirical Evaluation of Active Sampling for CRF-based Analysis of Pages, In Proc. of IEEE IRI 2010, pp.13-18 (2010).