CRF による学術論文からの参考文献文字列の抽出

石本　茜† 太田　学†† 高須　淳宏††† 安達　淳†††

† 岡山大学工学部情報工学科　〒 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: †,††{ishimoto, ohta}@de.cs.okayama-u.ac.jp, †††{takasu, adachi}@nii.ac.jp

あらまし　電子図書館に収録された学術論文の中から目的の論文を探すには，論文の表題や著者名等の書誌情報が必須である．とりわけ論文中の参考文献欄に記述された書誌情報は，文書間リンクの生成等にも利用できるため有用である．しかし，これらの書誌情報データベースを整備するには膨大なコストがかかる．そのため本研究では，論文の参考文献欄から個々の参考文献を表す参考文献文字列を，自動で抽出する手法を提案する．論文 PDF ファイルを変換した XML ファイルを入力とし，参考文献文字列の視覚的特徴と言語的特徴を CRF で学習し抽出する．DEIM2013 の論文 139 編を対象に評価実験を行い，最終的な参考文献文字列の抽出精度が 92%〜96%となることを確認した．

キーワード　情報抽出，CRF，電子図書館，参考文献

Reference String Extraction from Research Papers Using Conditional Random Fields

Akane ISHIMOTO†, Manabu OHTA††, Atsuhiro TAKASU†††, and Jun ADACHI†††

† Department of Information Technology, Faculty of Engineering, Okayama University
3–1–1 Tsushimanaka, Kita-ku, Okayama 700–8530 Japan
†† Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushimanaka, Kita-ku, Okayama 700–8530 Japan
††† National Institute of Informatics
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430 Japan
E-mail: †,††{ishimoto, ohta}@de.cs.okayama-u.ac.jp, †††{takasu, adachi}@nii.ac.jp

Key words　Information Extraction, Conditional Random Field, Digital Library, Reference

1. はじめに

学術論文を収録する電子図書館では，多数の論文の中から目的の論文を見つけるために，論文の表題や著者名等の書誌情報が必要となる．また，この書誌情報は文書間リンクの生成等にも利用することができる．しかし，書誌情報データベースを構築するには膨大なコストがかかるため，論文から自動で書誌情報を抽出する研究が行われている [5] [6] [7].

特に，論文の参考文献欄には関連する文献の書誌情報がまとめられているため，参考文献欄から書誌情報を抽出する研究は重要である．荒内ら [6] は，自然言語処理など様々な分野で使われている識別モデルの一つである Conditional Random Fieild (CRF) [1] を用いて参考文献の書誌情報の抽出を行った．一方，川上ら [7] は，荒内らの参考文献書誌情報抽出器を文献種類別に用意して抽出を行った．どちらの研究でも参考文献文字列のテキストから 90%以上の精度で，書誌情報を自動で抽出できることが報告されている．しかし，これらの研究では 1 件ごとに分割された参考文献文字列を入力とし，そこから書誌情報を抽出する．そこで本研究では，CRF を用いて学術論文の PDF からこの参考文献文字列を抽出する手法を提案する．

本稿の構成は次の通りである．まず 2 節で，書誌情報の抽出と参考文献文字列の抽出について関連する研究を紹介する．3 節で参考文献文字列の抽出の手順について詳しく説明する．4 節では提案手法の実験と評価について述べ，5 節で本稿をまとめる．

2. 関 連 研 究

2.1 書誌情報抽出

阿部川ら [5] は電子化された学術論文を対象に，論文のヘッダ部分からその論文自体の書誌情報を，また参考文献文字列から関連論文の書誌情報を抽出した．対象としたのは英語及び日本語で書かれた様々な論文の PostScript（PS）ファイルと PDF ファイルであり，PS ファイルは PDF ファイルに変換してから処理をした．PDF ファイルは pdftohtml を用いて XML ファイルに変換するが，出力した XML ファイルのテキストは，人間が読む順序で並んでいない場合がある．また，PDF ファイルの中に“ffi”や“fl”のような合字が含まれると，正しくテキストを出力できないという問題もあった．そのため，独自のプログラムを用いて，テキストを正しい順番に並べ変え，合字をそれに対応する文字列に変換するといった処理を行った．XML ファイルに変換した後，Support Vector Machine（SVM）[8] と Hidden Markov Model（HMM）[9] を用いて書誌情報を抽出した．書誌情報がすべて正しく抽出できた論文を正解とすると，論文のヘッダ部分の正解率は 69.2%であった．また，参考文献部分では日本語と英語に参考文献を分け，それぞれに対して抽出実験を行った．正解率は SVM を用いた最も高いもので，日本語の参考文献が 74.8%，英語の参考文献が 81.6%であった．

2.2 参考文献文字列抽出

Takasu ら [2] の研究では，電子情報通信学会（IEICE）や情報処理学会（IPSJ），人工知能学会（JSAI）の OCR 処理された論文文書画像を対象に，参考文献領域の行を抽出し，そこから個々の参考文献文字列を抽出している．参考文献領域の開始は“REFERENCES”のような特定のパターンを手掛かりに検出し，それ以降の行を抽出している．その際，HMM を用いて OCR の認識誤りも考慮している．次に，“[1]”のような参考文献文字列の先頭を表す文字列をもとに，抽出した複数の行を連結して個々の参考文献文字列とした．IEICE では OCR の誤りが 1.29%で参考文献文字列の抽出誤りが 3.84%，JSAI では OCR の誤りが 1.88%で参考文献文字列の抽出誤りが 9.86%，IPSJ では OCR の誤りが 2.42%で参考文献文字列の抽出誤りが 10.01%であった．

Councill ら [3] は，CRF を用いて参考文献文字列から書誌情報を抽出するオープンソースのツール ParsCit を開発した．彼らは参考文献文字列から書誌情報を抽出する前処理として，参考文献領域の検出を行った．Cora データセット [4] の参考文献文字列を対象に，著者名，タイトル等 13 項目の書誌要素を抽出し，それらの平均の抽出の F 値は 0.95 であった．

3. 参考文献文字列の抽出

3.1 概　　要

ここで，提案する参考文献文字列の抽出手順を詳しく説明する．提案手法ではまず，3.2 節の方法で論文 PDF ファイルを図 1 に示すような XML ファイルに変換する．次に 3.3 節の方法で，XML ファイルから“文献”等の文字列を手掛かりに参考文献領域の開始位置を検出する．さらに，3.4 節の方法で，検出した開始位置以降の行から，視覚的情報や言語的情報を抽出する．抽出したこれらの情報を素性として CRF で学習し，参考文献文字列である行に“Ref”等のラベルを付与する．個々の参考文献の開始行を表す“SRef”のラベルや“[1]”のような文字列を手掛かりに 3.5 節の方法で，個々の参考文献文字列を抽出しする．その結果，図 2 に示すような個々の参考文献文字列が得られる．


```
<BLOCK id="p4_b1">
  <TEXT width="8.628" height="8.628" id="p4_t1" x="142.8" y="43.2185">
    <TOKEN sid="p4_s794" id="p4_w1" font-name="MS" … >文</TOKEN>
    </TEXT>
  </BLOCK>
  <BLOCK id="p4_b2">
    <TEXT width="8.628" height="8.628" id="p4_t2" x="177.31" y="43.2185">
      <TOKEN sid="p4_s795" id="p4_w2" font-name="ms" … >献</TOKEN>
    </TEXT>
  </BLOCK>
  <BLOCK id="p4_b3">
    <TEXT width="235.041" height="7.07736" x="50.05" y="56.8523" id="p4_t3">
      <TOKEN sid="p4_s796" id="p4_w3" font-name="CMR8" … >[1]</TOKEN>
      <TOKEN sid="p4_s797" id="p4_w4" font-name="MS" … >石本</TOKEN>
      <TOKEN sid="p4_s798" id="p4_w5" font-name="ms" … >茜,</TOKEN>
      <TOKEN sid="p4_s799" id="p4_w6" font-name="ms" … >太田</TOKEN>
      <TOKEN sid="p4_s800" id="p4_w7" font-name="ms" … >学,</TOKEN>
      <TOKEN sid="p4_s801" id="p4_w8" font-name="ms" … >高須</TOKEN>
      <TOKEN sid="p4_s802" id="p4_w9" font-name="ms" … >淳宏,</TOKEN>
      <TOKEN sid="p4_s803" id="p4_w10" font-name="ms" … >安達</TOKEN>
      <TOKEN sid="p4_s804" id="p4_w11" font-name="ms" … >淳：</TOKEN>
      <TOKEN sid="p4_s805" id="p4_w12" font-name="cmr8" … >CRF</TOKEN>
      <TOKEN sid="p4_s806" id="p4_w13" font-name="ms" … >による学術論文か</TOKEN>
    </TEXT>
    <TEXT width="218.435" height="8.13385" x="66.66" y="67.4823" id="p4_t4">
      <TOKEN sid="p4_s807" id="p4_w14" font-name="ms" … >らの参考文献文字列の抽出,</TOKEN>
      <TOKEN sid="p4_s808" id="p4_w15" font-name="ms" … >第</TOKEN>
      <TOKEN sid="p4_s809" id="p4_w16" font-name="cmr8" … >6</TOKEN>
      <TOKEN sid="p4_s810" id="p4_w17" font-name="ms" … >回データ工学と情報マネジメン</TOKEN>
    </TEXT>
    <TEXT width="181.355" height="8.13385" id="p4_t5" x="66.66" y="78.1123">
      <TOKEN sid="p4_s811" id="p4_w18" font-name="ms" … >トに関するフォーラム</TOKEN>
      <TOKEN sid="p4_s812" id="p4_w19" font-name="cmr8" … >(DEIM2014),</TOKEN>
      <TOKEN sid="p4_s813" id="p4_w20" font-name="cmr8" … >C5-3</TOKEN>
      <TOKEN sid="p4_s814" id="p4_w21" font-name="cmr8" … >(2014).</TOKEN>
    </TEXT>
    <TEXT width="235.029" height="7.07736" id="p4_t6" x="50.05" y="89.7988">
      <TOKEN sid="p4_s815" id="p4_w22" font-name="cmr8" … >[2]</TOKEN>
```


図 1　論文 PDF を変換して得られる XML ファイルの例

文　　献

[1] 石本 茜, 太田 学, 高須 淳宏, 安達 淳： CRF による学術論文からの参考文献文字列の抽出, 第6 回データ工学と情報マネジメントに関するフォーラム(DEIM2014), C5-3 (2014).

[2] 荒内大貴, 太田学, 高須淳宏, 安達淳： CRF による和英文の参考文献文字列からの自動書誌情報抽出, 情報処理学会研究報告, 2012-DBS-156(1), pp. 1-8 (2012).

[3] 川上尚慶, 荒内大貴, 太田学, 高須淳宏, 安達淳： 文献種類別に分類した参考文献文字列からの書誌情報抽出の一手法, 第5 回データ工学と情報マネジメントに関するフォーラム（DEIM2013）, C10-3 (2013).

⋮

図 2　抽出した参考文献文字列の例

3.2 論文 XML ファイルの変換

論文 PDF ファイルを pdf2xml(注1) を用いて XML ファイルへと変換する．pdf2xml は，PDF ファイルに含まれるテキストに，図 1 に示すような TOKEN，TEXT，BLOCK といったタグを付与し，属性値としてページ内の座標やフォントサイズ等の情報を出力する．論文中の表には，本文と同様にこれらのタグが付与される．図には基本的に本文とは異なる IMAGE タグが付与され，ページの最後に出力される．ただし，図中の文字列に本文と同様に TOKEN，TEXT，BLOCK といったタグが付与される場合もある．

TEXT タグが付与される文字列は，論文中の行と一致するものが多く，本研究では TEXT を単位として参考文献文字列を抽出する．ただし，Microsoft Word 等で作成された PDF ファ

(注1)：http://sourceforge.jp/projects/sfnet_pdf2xml/

イルを変換した XML ファイルの中には，一つの単語に対して TEXT タグが付与されるものもある．

また，pdf2xml はフォントによってはテキスト全体や“ fi ”，“ fl ”といった合字が文字化けするが，本稿では対応していない．

3.3 参考文献領域の開始位置検出

本研究では，参考文献領域の開始位置を検出し，それ以降の TEXT の情報を CRF への入力として参考文献文字列を抽出する．この開始位置は，“ 文献 ”や“ Reference ”といった参考文献領域の始まりを示すキーワードを手掛かりに検出する．同じ行にこれらのキーワード以外の文字がある場合は，論文の本文中の記述と判断し，単独で出現した場合にのみ開始位置と判断する．

3.4 CRF による参考文献文字列抽出

参考文献領域の開始位置以降の TEXT は，参考文献文字列であるものが多い．しかし，参考文献文字列の間や後ろに，表やページ番号といった参考文献文字列ではない要素が出現する場合がある．そこで，参考文献文字列だけを抽出するために，TEXT のレイアウト情報や言語情報を素性として利用し，CRF によるラベル付けをする．本研究では荒内ら [6] や川上ら [7] と同様に，CRF++(注2) を利用して，TEXT にラベルを付与する．

3.4.1 Conditional Random Field

本研究では，標準的なチェーンモデルの CRF の定義を用いて，論文中の TEXT の列に参考文献と他を区別するラベルを付与する．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 $\mathbf{y} = y_1, ..., y_n$ となる条件付き確率を以下のように与える．

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

ただし，$Z_{\mathbf{x}}$ 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，

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

である．ここで，$f_k(y_{i-1}, y_i, \mathbf{x})$ 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の系列 $\mathbf{y}^*$ は次式で与えられる．

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in Y(\mathbf{x})} P(\mathbf{y}|\mathbf{x}) \quad (3)$$

今回ラベル付与の対象である入力 x_i は，論文 XML ファイル中の TEXT であり，ラベル y_i は以下の二つのラベル集合のいずれかのラベルとする．すなわち，“ Ref ”，“ Page ”，“ ETC ”の 3 種類のラベルと“ SRef ”，“ NRef ”，“ Page ”，“ ETC ”の 4 種類のラベルを定義した．3 種類のラベルのうち，“ Ref ”は参考文献文字列，“ Page ”はページ番号，“ ETC ”はその他の要素を表す．一方，4 種類のラベルでは，[1] のような文字列ではじまる参考文献文字列の開始行を表す TEXT を“ SRef ”とし，開始行以外の参考文献文字列は“ NRef ”とする．“ Page ”と“ ETC ”は 3 種類のラベルで定義したそれらと同じものを用いる．実験ではこれら 2 種類の参考文献文字列のラベルを比較する．

(注2)：http://crfpp.googlecode.com/svn/trunk/doc/index.html

表 1　素性テンプレート

種類	素性	内容
Unigram 視覚的素性	<i(0)>	TEXT の識別番号
	<x(0)>	TEXT の X 座標
	<y(0)>	TEXT の Y 座標
	<w(0)>	TEXT の幅
	<h(0)>	TEXT の高さ
	<gapx(0)>	前の TEXT との X 方向の間隔
	<gapy(0)>	前の TEXT との Y 方向の間隔
Unigram 言語的素性	<c(0)>	TEXT 内の文字数
	<ec(0)>	TEXT 内の英字の割合
	<nc(0)>	TEXT 内の数字の割合
	<kc(0)>	TEXT 内の漢字の割合
	<jc(0)>	TEXT 内の平仮名・片仮名の割合
	<sc(0)>	TEXT 内の記号の割合
	<ekw(0)>	TEXT 内に現れる参考文献以外の要素に特徴的な文字列の種類
	<rkw(0)>	TEXT 内に現れる参考文献に特徴的な文字列の種類
Bigram	<y(-1), y(0)>	ラベルの遷移

3.4.2 素性テンプレート

CRF++によるラベル付けに利用する素性テンプレートを表 1 にまとめる．

- 視覚的素性

視覚的素性として，ページ上での TEXT の位置やその中の文字の大きさ，前の TEXT との間隔等の特徴が，XML ファイルから TEXT の属性値として得られる．ただし，属性値として得られる数値は小数であるため，一の位以下を切り捨てた数値を素性とする．

- 言語的素性

言語的素性には TEXT 中の文字列を解析して得られる文字の種類や特徴的な文字列の有無等が該当する．本研究では，TEXT 内に含まれる文字数の一の位を切り捨てた値と，文字の種類ごとの割合を素性とする．また，表や図等を参考文献文字列と区別するために，表 2，表 3 に示す特徴的な文字列の有無を素性とする．ここで示した参考文献に特徴的な文字列は，川上ら [7] が書誌要素の分類，抽出に使用した特徴的な文字列を参考にして定めた．

- Bigram 素性

付与されるラベルの連接に関する情報を表す Bigram 素性を使用し，ラベルの出現順に関する制約を考慮する．

3.5 参考文献文字列抽出

3 種類のラベルセットによりラベル付けした場合は，“ Ref ”

表 2　参考文献以外の構成要素の特徴的な文字列の例

特徴的な文字列の例	種類
図, Figure, Fig.	FG
表, Table	TB
付録, Appendix	AP
注	FT

表 3　参考文献文字列を示す特徴的な文字列の例

特徴的な文字列の例	予想される内容	種類
[1]，[2]	参考文献文字列の始まり	SR
シンポジウム，Conf	Conference	RE
～論文誌，Journal	Journal	
ブック，Handbook	Book title	
（編），編集，Ed.	Editor	
～社，～出版，Publisher	Publisher	
Volume，Vol.	Volume	
No.	Number	
pp.，p.	Page	
January，Jan.	Month	
www，http	URL	
研究報告，大学	Other	

ラベルが付与された TEXT を参考文献文字列として出力する．その後，[1] のような文字列を手掛かりに個々の参考文献文字列に切り分ける．一方，4 種類のラベルセットによるラベル付けでは，“SRef”，“NRef”ラベルが付与された TEXT を参考文献文字列として出力する．

4. 実　　験

参考文献領域の開始位置検出と CRF による参考文献文字列抽出の実験を行い，開始位置の検出精度と参考文献文字列の抽出精度をそれぞれ算出した．

4.1 実験データ

第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013) の 142 編の論文 PDF ファイルを実験データとし，pdf2xml で XML ファイルへ変換した．しかし，142 の論文のうち 3 編の論文では論文全体で文字化けが発生したため，この論文を除く 139 の論文で，開始位置検出と CRF によるラベル付与の実験を行った．DEIM2013 の論文はほとんどが日本語だが，実験データの中には 5 編の英語の論文を含む．本研究では，日本語と英語の論文は区別せずに扱った．なお，データには人手で正解ラベルをつけて学習と評価に利用する．

CRF によるラベル付与の実験では上記のデータの中からデータセットを 2 種類用意した．まず，1 つ目のデータセットは DEIM2013 で論文 PDF がある 56 セッションのうち，22 セッションの論文計 118 編（データセット 1）である．この中には，参考文献領域の開始位置以降に，参考文献文字列ではない要素が出現する論文が 20 含まれる．その具体的な内訳を表 4 に示す．1 件の論文の中に，表 4 に示した要素が 2 つ以上含まれるものもある．なお，参考文献領域の開始位置以降に参考文献文字列ではない要素が出現する論文を増やすため，データセット 2 を用意した．

表 4　参考文献領域の開始位置以降に含まれる論文要素とその論文数 (データセット 1)

要素名	論文数
ページ番号	7
表	4
図中の文字	4
図キャプション	6
脚注	2
付録	1

表 5　参考文献領域の開始位置以降に含まれる論文要素とその論文数 (データセット 2)

要素名	論文数
ページ番号	13
表	13
図中の文字	7
図キャプション	15
脚注	2
付録	2

データセット 2 は，データセット 1 に，参考文献領域の開始位置以降に参考文献文字列ではない要素が出現する DEIM2013 の論文を 21 追加した計 139 編の論文である．よってデータセット 2 には，参考文献領域の開始位置以降に参考文献文字列ではない要素が出現する論文が 41 含まれる．その具体的な内訳を表 5 に示す．なお，1 件の論文の中に表 5 に示した要素が 2 つ以上含まれるものもある．

4.2 開始位置検出実験

参考文献領域の開始位置検出は 139 の論文全てで成功し，検出率は 100%であった．本実験のデータセットではうまく検出できたが，参考文献領域の始まりを示すキーワードが論文中に複数存在する，もしくは存在しない論文では正しく検出が行えない可能性がある．

4.3 CRF による参考文献文字列抽出実験

参考文献文字列の開始位置以降の TEXT に CRF を用いてラベルを付与する実験を，データセット 1 の 118 論文とデータセット 2 の 139 論文に対しそれぞれ 5 分割交差検定で行った．TEXT ごとのラベル付与精度を算出し，論文全体の正解率を次の式で計算した．

$$\text{論文正解率} = \frac{\text{個々の参考文献文字列を全て正しく抽出できた論文数}}{\text{テストデータの論文数}} \quad (4)$$

4.3.1 データセット 1

データセット 1 では，検出した参考文献領域の開始位置以降の TEXT 5,426 件にラベルを付与した．ラベル別の付与精度を表 6，表 7 に示す．“Ref”，“Page”，“ETC”の 3 種類のラベルセットによる実験では，正しくラベルを付与できた論文は 112 あり，誤ったラベルを付与した論文が 6 あったため，論文正解率は 94.92%であった．“SRef”，“NRef”，“Page”，“ETC”の 4 種類のラベルセットによる実験では，正しくラベルを付与できた論文は 113 あり，誤ったラベルを付与した論文が 5 あったため，論文正解率は 95.76%であった．2 つの実験で誤った論

表 6　ラベル別の付与精度（データセット 1・ラベル 3 種）

	正解 TEXT 数	再現率	適合率	F 値
Ref	4,616	1.000	0.991	0.995
Page	10	1.000	1.000	1.000
ETC	800	0.948	1.000	0.973
全 TEXT	5,426	0.992	0.992	0.992

表 7　ラベル別の付与精度（データセット 1・ラベル 4 種）

	正解 TEXT 数	再現率	適合率	F 値
SRef	1,401	1.000	0.999	0.999
NRef	3,215	1.000	0.981	0.991
Page	10	1.000	1.000	1.000
ETC	800	0.921	1.000	0.959
全 TEXT	5,426	0.988	0.988	0.988

表 8　誤ったラベルを付与した TEXT（データセット 1・ラベル 3 種）

実際の構成要素名	付与したラベル	TEXT 数	論文数
表	Ref	17	1
図中の文字	Ref	16	1
図キャプション	Ref	7	2
脚注	Ref	2	2

表 9　誤ったラベルを付与した TEXT（データセット 1・ラベル 4 種）

実際の構成要素名	付与したラベル	TEXT 数	論文数
図中の文字	NRef	54	1
図キャプション	NRef	7	2
脚注	SRef	2	2

文のうち 4 編は同じ論文であった．誤ったラベルを付与した論文には，すべて参考文献以外の要素が含まれていた．その具体的な内訳を，表 8 と表 9 に示す．

4.3.2　データセット 2

データセット 2 では，検出した参考文献領域の開始位置以降の TEXT 7,873 件にラベルを付与した．ラベル別の付与精度を表 10，表 11 に示す．3 種類のラベルセットによる実験では，正しくラベルを付与できた論文は 129 あり，誤ったラベルを付与した論文が 10 あったため，論文正解率は 92.81%であった．4 種類のラベルセットによる実験では，正しくラベルを付与できた論文は 128 あり，誤ったラベルを付与した論文が 11 あったため，論文正解率は 92.09%であった．2 つの実験で誤った論文のうち 10 編は同じ論文であった．誤ったラベルを付与した論文には，すべて参考文献以外の要素が含まれていた．具体的な内訳を，表 12 と表 13 に示す．

4.4　考　　察

データセット 1 では参考文献文字列全てを抽出できたが，脚注や図，表も参考文献文字列と判断してしまった．この原因は，参考文献領域を分断するように出現する，参考文献文字列以外の論文構成要素を区別できなかったためである．

一方，参考文献文字列以外の論文構成要素が参考文献領域を分断している論文を増やしたデータセット 2 に対する実験では，データセット 1 で付与するラベルを誤った論文のうち，図や表に対して正しくラベル付与できたものもあったが，脚注に対してはデータセット 1 との差分のデータにもそのような事例がなかったため全く解決できなかった．また，データセット 1 との差分のデータでも，図や表等の要素は境界部分のラベル付与が正しくできていないものがいくつかあったため，さらに素性などを検討する必要がある．

表 10　ラベル別の付与精度（データセット 2・ラベル 3 種）

	正解 TEXT 数	再現率	適合率	F 値
Ref	5,508	0.998	0.969	0.984
Page	18	1.000	1.000	1.000
ETC	2,347	0.926	0.995	0.959
全 TEXT	7,873	0.977	0.977	0.977

表 11　ラベル別の付与精度（データセット 2・ラベル 4 種）

	正解 TEXT 数	再現率	適合率	F 値
SRef	1,668	0.999	1.000	0.999
NRef	3,840	0.949	0.998	0.972
Page	18	1.000	1.000	1.000
ETC	2,347	0.996	0.911	0.959
全 TEXT	7,873	0.972	0.972	0.972

表 12　誤ったラベルを付与した TEXT（データセット 2・ラベル 3 種）

実際の構成要素名	付与したラベル	TEXT 数	論文数
表	Ref	128	3
図中の文字	Ref	38	1
図キャプション	Ref	5	2
脚注	Ref	2	2
表前後の参考文献文字列	ETC	6	2
図前後の参考文献文字列	ETC	5	1

表 13　誤ったラベルを付与した TEXT（データセット 2・ラベル 4 種）

実際の構成要素名	付与したラベル	TEXT 数	論文数
表	NRef	137	2
図中の文字	NRef	65	1
図キャプション	NRef	6	3
脚注	SRef	2	2
表前後の参考文献文字列	ETC	3	2
図前後の参考文献文字列	ETC	5	1

樫本らの研究 [10] では，論文 PDF ファイルを変換した XML ファイルから論文構成要素を，ルールを用いた手法と SVM を用いた手法で抽出する．このうち実験結果の良かったルールを用いた手法による参考文献領域の抽出結果と本研究の抽出結果を比較する．彼らは DEIM2013 の 50 論文を使用し，そのうち 45 の論文から正しく抽出したので，正解率は 90.00%であった．同じデータセットに対して，本手法を用い，5 分割交差検定で参考文献文字列抽出の実験をしたところ 50 論文中 46 の論文から正しく抽出できたので正解率は 92.00%となる．

5.　ま　と　め

本稿では，学術論文 PDF ファイルから，参考文献文字列を文献一つごとに分離した形で抽出する手法を提案した．提案手法ではまず，“ 文献 ” のような文字列を手がかりに，論文の参考文献領域の開始位置を検出する．次に，CRF を利用して参考

文献文字列か他の論文構成要素であるかを判別する．CRF で利用するための素性として，視覚的情報や言語的情報を用いた結果，実験では 92%〜96%の正解率で論文から正しく参考文献文字列を抽出できた．今後，さらに性能の向上を図るため素性の検討をしていきたい．また，本稿では DEIM2013 の論文のみを実験対象としため，論文データを拡充し英文誌等の論文においても有効性を確認したい．

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 23300040, 24300097)，科学研究費補助金基盤研究 (C)(課題番号 25330384)，科学研究費補助金若手研究 (B)(課題番号 23700119)，および国立情報学研究所公募型共同研究の援助による．ここに記して深謝する．

文　　献

[1] 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).
[2] Takasu, A. and Aihara, K. : Quality Enhancement in Information Extraction from Scanned Documents, In Proc. of DocEng '06, pp. 122-124 (2006).
[3] Isaac G. Councill, C. Lee Giles and Min-Yen Kan : ParsCit: An open-source CRF reference string parsing package, In Proceedings of Language Resources and Evaluation Conference (2008).
[4] A. McCallum, K. Nigam, J. Rennie and K. Seymore : Automating the Construction of Internet Portals with Machine Learning, Information Retrieval, Vol. 3, No. 2, pp. 127-163 (2000).
[5] 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).
[6] 荒内大貴, 太田学, 高須淳宏, 安達淳 : CRF による和英文の参考文献文字列からの自動書誌情報抽出, 情報処理学会研究報告, 2012-DBS-156(1), pp. 1-8 (2012).
[7] 川上尚慶, 荒内大貴, 太田学, 高須淳宏, 安達淳 : 文献種類別に分類した参考文献文字列からの書誌情報抽出の一手法, 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), C10-3 (2013).
[8] Cortes, C. and Vapnik, V. : Support-Vector Networks, Machine Learning, Vol. 20, No. 3, pp. 273-297 (1995).
[9] Seymore, K., McCallum, A., and Rosenfeld, R. : Learning Hidden Markov Model Structure for Information Extraction, In AAAI 99 Workshop on Machine Learning for Information Extraction (1999).
[10] 樫本達矢, 太田学, 高須淳宏 : 学術論文からの構成要素抽出の一手法, 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), C5-2 (2014).