DEIM Forum 2013 C10-3

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

川上　尚慶† 　荒内　大貴†† 　太田　学†† 　高須　淳宏††† 　安達　淳†††

† 岡山大学工学部情報工学科　〒 700–8530 岡山市北区津島中 3 丁目 1 番 1 号
†† 岡山大学大学院自然科学研究科　〒 700–8530 岡山市北区津島中 3 丁目 1 番 1 号
††† 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: †,††{kawakami,arauchi,ohta}@de.cs.okayama-u.ac.jp,†††{takasu,adachi}@nii.ac.jp

あらまし　タブレット端末等の普及により，インターネットを介して時間と場所を問わずにアクセスできる電子図書館の利用機会が増えている．電子図書館を快適に利用するためには書誌情報のデータベースの整備が必須であるが，そのコストは膨大である．一方で，学術論文の参考文献欄に記述されている著者名やタイトルなどの書誌情報は検索等に利用できるため，その有効活用を図りたいという要求がある．そこで本稿では，参考文献文字列から自動で書誌情報を高精度に抽出する手法を提案する．提案手法は，予め参考文献文字列をジャーナル論文，会議録論文等の文献種類別に分類して，それぞれから CRF を用いて書誌情報を自動抽出する．電子情報通信学会英文論文誌の 1 年分の論文から収集した参考文献文字列コーパスを用いた分類実験では，90.4% の参考文献文字列を正しく分類できた．また，書誌情報の抽出実験では，分類を行わない場合 91.5% の抽出精度が，分類した文献種類別の抽出では 92.5% となった．
キーワード　情報抽出　CRF　参考文献文字列

A method of extracting bibliographic information from reference strings classified by literature type

Naomichi KAWAKAMI†, Daiki ARAUCHI††, 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: †,††{kawakami,arauchi,ohta}@de.cs.okayama-u.ac.jp,†††{takasu,adachi}@nii.ac.jp

Key words　information extraction, CRF, reference string

1. はじめに

多数の学術論文を蓄積する電子図書館において，目的の論文を検索するためには，著者名や論文題目名といった書誌情報が必要となる．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．特に学術論文の参考文献欄には，それと関連する多くの文献が記述されており，その書誌情報は重要である．この書誌情報を自動抽出し，参考文献エンティティを同定できれば，既存のデータベースを利用した文書間リンク自動生成など様々なサービスを提供できるようになる．

そこで荒内ら [1] は，自然言語処理などの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 Field (CRF) を利用して，論文中の参考文献文字列のテキストから書誌情報を自動で抽出する手法を提案した．

荒内らは，全ての参考文献文字列から同一の抽出器を用いて書誌情報を抽出した．しかし，参考文献文字列には，文献の種類ごとに記載される書誌情報に特徴があるため，それぞれに適した学習モデルは異なると予想できる．そこで，本稿では，こ

の書誌情報の特徴に基づいて参考文献文字列を分類し，分類した文献種類別の書誌情報抽出器を用意して，書誌情報を抽出することで，抽出精度の向上を図る．

本稿の構成は次の通りである．まず，2 節で，学術論文からの自動書誌情報抽出に関する研究を紹介する．3 節で，参考文献文字列を文献種類別に分類する方法について説明し，続く 4 節で，本研究で用いた CRF による自動書誌情報抽出法について説明する．5 節では提案手法の評価実験について述べる．最後に，6 節で本稿をまとめる．

2. 関 連 研 究

2.1 機械学習を用いた書誌情報抽出

まず，Support Vector Machine (SVM) [2] や Hidden Markov Model (HMM) [3] を用いた書誌情報抽出には，阿辺川ら [4] や Okada ら [5] の研究がある．

阿辺川らは，OCR 処理された学術論文のタイトルページや参考文献文字列から書誌情報を抽出する手法を提案している．彼らは，日本語及び英語で書かれた様々な論文を対象に，SVM と HMM を用いて論文タイトルページでは行単位，参考文献文字列では文字単位で書誌要素ラベルを付与し，書誌情報を抽出した．また，日本語と英語では学習されるモデルが大きく異なると予想して，参考文献文字列を和文と英文に分類し，それぞれの言語に対して実験を行った．実験において，論文タイトルページでは，論文単位の抽出精度が 69.2%，参考文献文字列では，和文で 74.8%，英文で 81.6% の精度で全ての書誌情報を過不足なく抽出した．

Okada らは，カンマや“ vol. ”, “ no. ”, “ pp. ”, “ ed. ”といった特定の文字列をデリミタとして参考文献文字列をトークナイズし，各トークンに SVM と HMM を用いて書誌要素ラベルを付与することで，書誌情報を抽出した．電子情報通信学会論文誌 Vol.J83-DII の No.1 から No.12 に掲載されている論文の参考文献文字列を対象に実験を行い，97.6% の精度で参考文献文字列中の全ての書誌情報を過不足なく抽出した．

一方，CRF [6] を用いた書誌情報抽出には，薬師ら [7] [8] や Councill ら [9] の研究がある．

薬師らは，学術論文文書画像のレイアウト解析と文字認識を行い，CRF を用いて学術論文文書画像の矩形テキスト領域や文字へ書誌要素ラベルを付与した．情報処理学会論文誌の論文を対象とした実験では，矩形テキスト領域への書誌要素ラベル付けによる書誌情報抽出と，文字へのラベル付けによる著者名抽出を行った．矩形テキスト領域の書誌情報抽出精度が 98.04%，論文タイトルページから全ての著者名を過不足なく抽出する精度が 99.07% だった．

Councill らは，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 を開発している．Cora データセット [10] を対象に，著者名や論文題目名など 13 項目の書誌情報について抽出実験を行い，13 項目の平均の適合率，再現率がともに 0.957，F 値が 0.950 と報告している．

2.2 CRF

本研究の書誌情報抽出 [1] では，標準的なチェーンモデルの

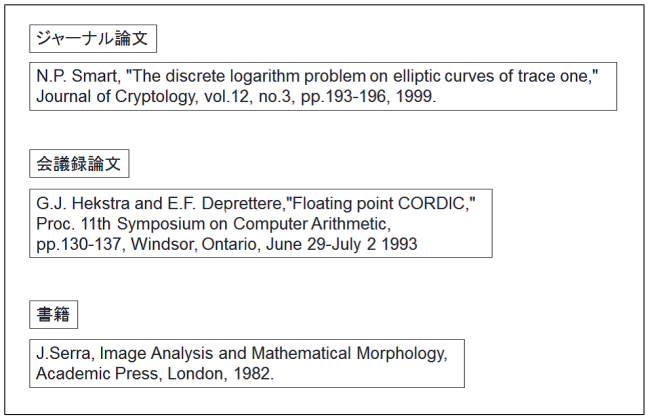

図 1　参考文献文字列の例

CRF [6] の定義を用いて，参考文献文字列をトークナイズして得られるトークン列に書誌情報ラベルを付与する．すなわち，入力系列 $\mathbf{x} = x_1, \ldots, x_n$ が与えられたとき，出力ラベル系列が $\mathbf{y} = y_1, \ldots, y_n$ となる条件付き確率は以下のように与えられる．

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

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，

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

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．

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

4 節で説明する書誌情報抽出では，ラベル付与の対象である入力 x_i は，参考文献文字列をデリミタで区切った各トークンである．一方，ラベル y_i は，著者名，論文題目名といった書誌要素である．

3. 参考文献文字列の分類

3.1 文献の種類

本稿では，参考文献に挙げられた文献の文字列を，文献の種類ごとに分類する．本研究では，文献の種類として，学術論文誌に掲載されたジャーナル論文，国内外の学術会議や学術大会で発表された会議録論文，本として出版されている書籍の 3 種類を定めた．これらが参考文献に記載されている例を図 1 に示す．

各種類の参考文献を示す文字列に含まれる書誌要素についてみると，ジャーナル論文には論文誌名が記載されており，巻

表 1　特定の書誌要素を示唆する特徴的な文字列

特徴的な文字列	対応する書誌要素
Proc., Workshop, Conference, Conf., Symposium, Symp. 他 7 件	Conference
Journal, Magazine, Technical Report, Ph.D., Thesis 他 21 件	Journal
ed., Ed., eds., Eds.	Editor
Publisher, Pub., Verlag, Shuppan, Co., Inc., Ltd. 他 6 件	Publisher
vol., Vol. 他 5 件	Volume
no., nos. 他 5 件	Number
pp., p.	Page
1800〜2013	Year
January, Jan., Janvier 他 33 件	Month
1〜31	Day
http, ftp	URL

(Volume) や号 (Number) が書かれることが多い．会議録論文には会議名が記載されており，会議の開催地や開催年月日が書かれることが多い．書籍には出版社名が記載されており，ページに関する記述がないものが多い．このように，参考文献文字列に記述される書誌要素には，文献種類ごとに特徴があることがわかる．そこで，これらの特徴に基づく参考文献文字列の分類方法を 3.2 節で述べる．

3.2　参考文献文字列の分類手法

3.2.1　概　　要

本研究では，参考文献文字列を次のように分類する．まず，カンマをデリミタとして参考文献文字列を分割する．次に，各分割文字列に対して，特徴的な文字列および辞書のエントリを包含するか照合し，包含すればそれに対応する書誌要素があると判定する．そして，その参考文献文字列に含まれると判定された書誌要素のリストを作成する．そのリストの書誌要素の集合から文献の種類を決定し，参考文献文字列を分類する．

3.2.2　書誌要素リストの作成

本研究では，参考文献文字列に現れる特徴的な文字列から，特定の書誌要素の有無を推定し，その有無を表す書誌要素リストを作成する．

ここで，特徴的な文字列とは，例えば“ Proc. ”のことで，これがあれば，その参考文献文字列は Conference を表す書誌要素を含むと判定する．このような特徴的な文字列とそれに対応する書誌要素の例を表 1 にまとめる．

また，各書誌要素に頻出する固有名詞も存在する．そのような固有名詞に対処するため，論文誌名(注1)，出版社名(注2)，地名(注3)の辞書を用意した．論文誌名辞書には 8,576 件，出版社名辞書には 727 件，地名辞書には 4,371 件の語を収録している．

(注1)：http://science.thomsonreuters.com
(注2)：http://www.narosa.com/nbd/PublisherDistributed.asp など
(注3)：http://www.fallingrain.com/world/index.html など

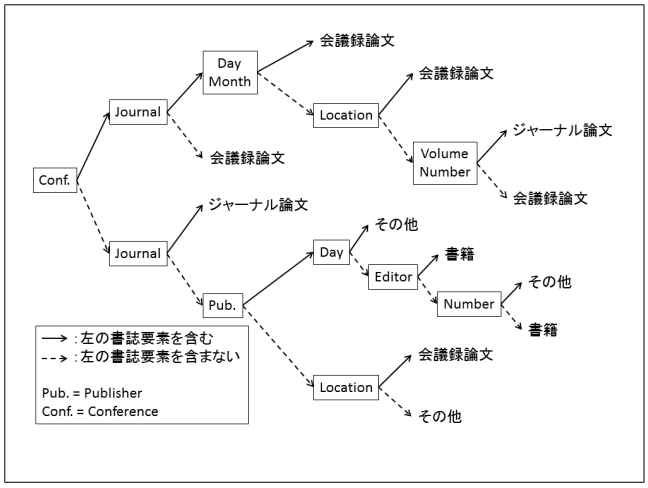

図 2　文献種類判定のための決定木

3.2.3　ルールに基づく分類

ルールに基づく文献種類の分類方法について説明する．3.2.2 項で説明した書誌要素リストを用いた文献種類の判定のための決定木を図 2 に示す．

各参考文献文字列に含まれると判定された書誌要素のリストを入力とする．まず，そのリストに Conference が含まれるか，次に，Journal が含まれるかをみる．ここで，Conference のみが含まれる場合は会議録論文，Journal のみが含まれる場合はジャーナル論文に分類する．Conference，Journal が両方とも含まれる場合，Month または Day が含まれるかをみる．そのいずれかが含まれる場合は会議録論文に分類する．Month，Day がともに含まれない場合は，Location が含まれるかをみる．Location が含まれる場合は会議録論文に分類する．Location が含まれない場合は，Volume または Number が含まれるかをみる．そのいずれかが含まれる場合はジャーナル論文，含まれない場合は会議録論文に分類する．一方，Conference，Journal がともに含まれない場合，Publisher が含まれるかをみる．Publisher が含まれる場合は，Day が含まれるかをみる．Day が含まれる場合は，その他に分類する．Day が含まれない場合は，Editor が含まれるかをみる．Editor が含まれる場合は，書籍に分類する．Editor が含まれない場合は，Number が含まれるかをみる．Number が含まれる場合はその他，含まれない場合は書籍に分類する．Publisher が含まれない場合，Location が含まれるかをみる．これが含まれる場合は会議録論文，含まれない場合はその他に分類する．本研究では，その他を含め，以上 4 種類に参考文献文字列を分類する．

3.2.4　CRF を用いた分類

本研究では比較のため，CRF を用いた文献種類の分類も行った．2.2 節の式で，ラベル付与の対象である入力 x_i は，各参考文献文字列である．一方，ラベル y_i は，ジャーナル論文，書籍，会議録論文，その他の 4 種類の文献種類である．CRF によるラベル付与に利用する，参考文献文字列に現れる特徴をまとめたものを素性テンプレートと呼ぶ．分類器で用いる素性テンプレートを表 2 に示す．

表 2 に示すように，本分類器では素性として，含まれる可能

表 2　参考文献文字列の分類に用いる素性テンプレート

種類	素性	内容
Unigram	<RC>	書誌要素リストの Conference の有無
	<RW>	書誌要素リストの Journal の有無
	<RV>	書誌要素リストの Volume の有無
	<RN>	書誌要素リストの Number の有無
	<RPP>	書誌要素リストの Page の有無
	<RE>	書誌要素リストの Editor の有無
	<RP>	書誌要素リストの Publisher の有無
	<RD>	書誌要素リストの Day の有無
	<RM>	書誌要素リストの Month の有無
	<RY>	書誌要素リストの Year の有無
	<RL>	書誌要素リストの Location の有無
	<RURL>	書誌要素リストの URL の有無

表 3　抽出する書誌情報

書誌要素	書誌要素ラベル
Author	RA
Editor	RE
Title	RT
Booktitle	RB
Journal	RW
Conference	RC
Volume	RV
Number	RN
Page	RPP
Publisher	RP
Day	RD
Month	RM
Year	RY
Etc	ETC

性の高い書誌要素リストを用いる．

4. 書誌情報抽出器

本研究では，荒内ら [1] らの提案する CRF を用いた書誌情報抽出器を使用する．荒内らは，参考文献文字列を文献種類別に分類することなく，一つの抽出器で全ての参考文献文字列から書誌情報を抽出した．一方，本研究では，文献種類別に分類し，その文献種類別に用意した抽出器を用いて書誌情報を抽出する．

4.1 抽出する書誌情報

参考文献文字列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 3 にまとめる．これは，実験で使用する電子情報通信学会英文論文誌にあわせて用意した．なお，ETC のラベルは所属機関や会議の開催国，出版社のある国等に付与される．

4.2 素性テンプレート

荒内らの書誌情報抽出器で用いる素性テンプレートを表 4 に示す．この素性テンプレートは，言語的な素性のみで構成されている．

表 4 よりこの抽出器では素性として，トークンの出現位置や文字数，トークンを構成する文字種とその割合，特徴的な文字列や各種辞書のエントリの有無を用いる．特徴的な文字列は，表 1 のような文字列である．また，辞書は，3.2.2 項で示した

表 4　荒内らの書誌情報抽出に用いる素性テンプレート

種類	素性	内容
Unigram	<token_ab_pos(0)>	トークン列における絶対的なトークン出現位置
	<token_re_pos(0)>	トークン列における相対的なトークン出現位置
	<num_token(0)>	トークンの文字数
	<zen_alp(0)>	トークン内の全角アルファベット数の割合
	<zen_fig(0)>	トークン内の全角数字数の割合
	<han_alp(0)>	トークン内の半角アルファベット数の割合
	<han_fig(0)>	トークン内の半角数字数の割合
	<han_etc(0)>	トークン内の記号の文字数の割合
	<last_chara(*i*)>	トークンの最後の文字
	<front_1-4_string(0)>	トークンの先頭から四文字目までの文字列
	<back_1-4_string(0)>	トークンの末尾から四文字目までの文字列
	<token_lc(*i*)>	トークンを小文字にした文字列
	<capital(*i*)>	トークン中の大文字の有無
	<digit(*i*)>	トークン中の数字の有無
	<editor>	参考文献文字列における editor に関する記述の有無
	<dictionary(*i*)>	辞書的素性
	<feature_term(*i*)>	トークン内の特徴的な文字列の種類
	<token(0)>	トークン自身
Bigram	< *y*(-1),*y*(0)>	ラベルの遷移

表 5　追加した素性テンプレート

種類	素性	内容	抽出器の文献種類
Unigram	<num_period(0)>	トークン内のピリオドの数	ジャーナル論文
	<apos(0)>	トークン中のアポストロフィーの有無	会議録論文
	<URL>	参考文献文字列における URL に関する記述の有無	その他
	<hyphen(0)>	トークン中のハイフンの有無	書籍，会議録論文
	<num_word(0)>	トークン内の単語数	全体

辞書に加えて，人名(注4)，月名(注5)の辞書を用意した．人名辞書には 97,942 件，月名辞書には 36 件の語を収録している．

さらに，付与される書誌要素ラベルの連接に関する情報を表す Bigram 素性を使用し，書誌要素の出現順に関する制約を考慮している．

本研究ではさらに，各文献種類の参考文献文字列に現れる特徴から，それぞれ個別の抽出器に表 5 の素性を追加した．

Journal では，略語が使われることが多いため，略語表記に使われるピリオドの数をジャーナル論文の抽出器の素性に追加した．Conference には，例えば“ Proc. Crypto’85 ”のように，会議名と開催年をアポストロフィーでつなげて表記するものが多数存在する．このため，アポストロフィーの有無を会議録論文の抽出器の素性に追加した．その他に分類された参考文献文字列の中に，URL を含むという特徴を持つものがあるため，URL の有無をその他の抽出器の素性に追加した．書籍と会議録論文の抽出器の素性にハイフンの有無を追加したのは，Conference, Publisher の表記や，複数日にわたる Day の表記にハイフンが使われることが多いためである．

5. 評 価 実 験

参考文献文字列の分類実験と書誌情報抽出実験を行い，分類精度と抽出精度をそれぞれ算出した．分類実験では，正解と決定木による分類結果，CRF による分類結果を比較する．書誌情報抽出実験では，文献種類別に分類した場合と分類しない場合の抽出精度を比較する．

5.1 実験データ

電子情報通信学会英文論文誌 Vol.E83-A No.1 から No.12 の

(注4)：http://www.census.gov/genealogy/names/ など
(注5)：英語の月名とその省略形，仏語の月名

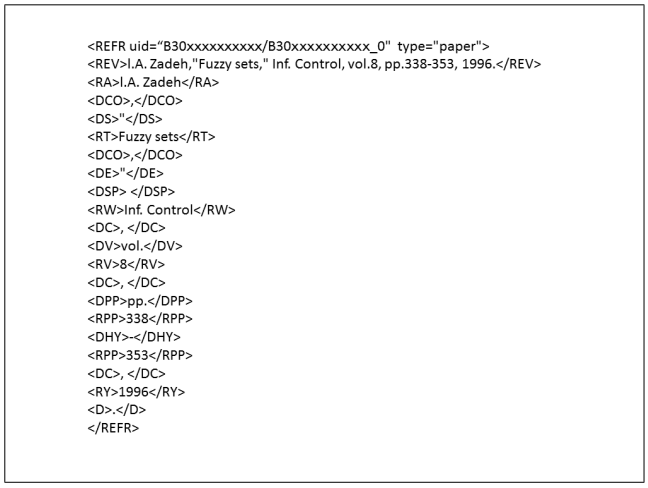

図 3　参考文献コーパスの例

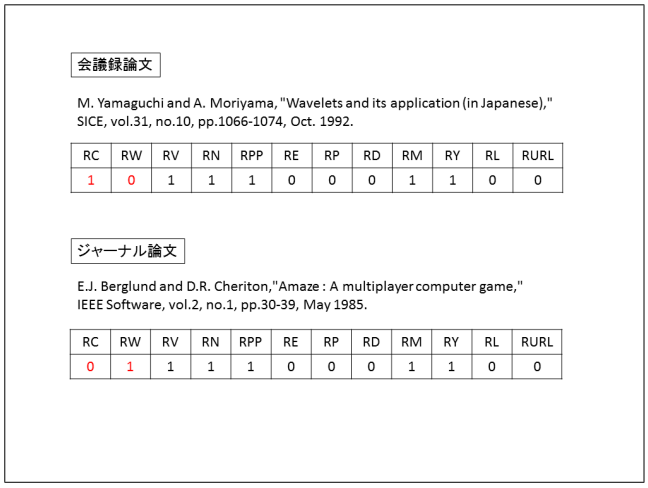

図 4　特徴の類似した参考文献文字列の例

表 6　ルールに基づく分類の分類状況

	ジャーナル論文	書籍	会議録論文	その他
ジャーナル論文	2,132	7	14	201
書籍	2	580	3	27
会議録論文	15	12	1,192	6
その他	82	18	45	161

表 7　CRF を用いた分類の分類状況

	ジャーナル論文	書籍	会議録論文	その他
ジャーナル論文	2,109	11	243	18
書籍	8	553	134	175
会議録論文	74	57	851	54
その他	40	3	19	148

一年分に相当する論文の参考文献コーパスを用意し，そこに記述されている 4,497 件の参考文献文字列を使用して評価実験を行った．参考文献コーパスの例を図 3 に示す．

1 行目の type はその参考文献文字列の文献種類を表し，分類実験の正解分類とする．2 行目の REV は参考文献文字列の原文を表す．3 行目以降は 2 行目の参考文献文字列を分割し，書誌要素ラベルを付与したもので，書誌情報抽出実験の正解とする．また，参考文献文字列を文献種類別に分けると，ジャーナル論文が 2,231 件，書籍が 624 件，会議録論文が 1,247 件，その他が 395 件であった．

5.2　参考文献文字列の分類実験

図 3 のような参考文献コーパスを用いて，参考文献文字列の分類実験を行った．ただし，3.2.4 項の CRF を用いた分類実験では，5 分割交差検定で精度を算出した．4,497 件の参考文献文字列の内，3.2.3 項の決定木による分類では 4,065 件，3.2.4 項の CRF による分類では 3,661 件の参考文献文字列を正しく分類でき，分類精度はそれぞれ 90.4%，81.4% だった．表 6，表 7 に各手法の参考文献文字列の分類状況をまとめる．表 6，表 7 において，縦は提案手法により分類した文献種類，横は正解の文献種類である．

ルールに基づく分類では，表 6 より，その他をジャーナル論文と誤る場合が多いとわかる．これは，参考文献コーパスで“ Technical Report ”や“ Thesis ”といった文字列は Journal の書誌要素に対応付けているが，ジャーナル論文の文献種類には含まれていないことが主な原因である．これに対処するには,「研究報告」という文献種類を用意するなど，文献種類の定義を見直す必要がある．また，ジャーナル論文，会議録論文をその他と誤る場合も多いとわかる．これは，参考文献文字列が含む書誌要素リストの作成において，Journal，Conference の書誌要素を含むと判定できない場合が多いことが主な原因である．これには，詳細なエラー解析を行い，新たに特徴的な文字列や辞書を拡充するといった対策がある．

CRF を用いた分類では，表 7 より，会議録論文をジャーナル論文や書籍と誤る場合が多い．これは，会議録論文は他の文献種類に比べて多くの種類の書誌要素を含むことが多いため，書誌要素リストの作成において，ある書誌要素が含まれないと誤って判定された場合，他の文献種類と似た特徴になってしまうことが原因に挙げられる．このような，文献種類は異なるが，文字列に現れる特徴が類似した参考文献文字列の例を図 4 に示す．

図 4 より，この会議録論文とジャーナル論文の参考文献文字列の書誌要素リストから作られる素性には，赤字で示した Conference の有無と Journal の有無の違いしかないため，これらの書誌要素の有無の判定を誤ると他方の文献種類に分類されてしまう可能性がある．

また，会議録論文やその他を書籍と誤る場合が多い．書籍に含まれる特徴的な書誌情報として Booktitle がある．しかし，Booktitle には共通して現れる表現がほとんどない．さらに，Booktitle と Title の違いを判別するのは難しい．したがって，参考文献文字列中に Booktitle を表す文字列だけから，それが本当に Booktitle かどうか判定することは困難である．このため，参考文献文字列の分類に用いる書誌要素リストに Booktitle を含められないことが一因に挙げられる．これらの対策としては，CRF に用いる素性に，書誌情報抽出器で用いる参考文献文字列に含まれる文字や記号の特徴を組み込むことが挙げられる．

5.3　書誌情報抽出実験

書誌情報の抽出精度を評価する際に，荒内らの方法と比較するため，表 3 の書誌情報ラベルを表 8 のようにまとめた．

表 8　書誌情報抽出精度評価のための書誌情報の再分類

書誌情報ラベル	分類名
RA, RE	AUTHOR
RT, RB	TITLE
RW, RC	JOURNAL
RV, RN, RPP	VOLUME
RP	PUBLISHER
RD	DAY
RM	MONTH
RY	YEAR
ETC	ETC

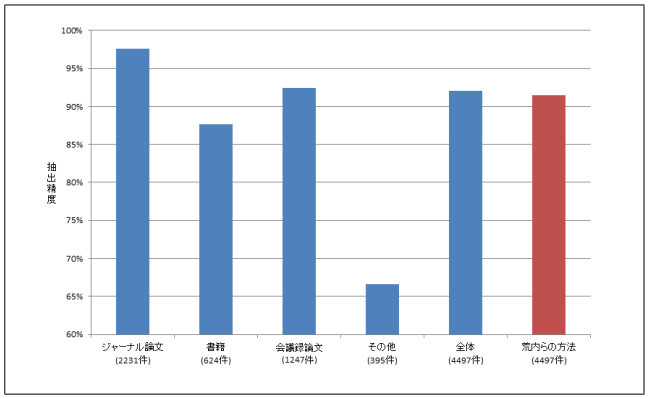

図 5　分類による抽出精度の差

本実験では，5 分割交差検定により，参考文献文字列中の各書誌情報の抽出精度と，各参考文献文字列から全ての書誌情報を過不足なく正確に抽出できるかどうかの全体的な抽出精度を算出した．

5.3.1　分類による抽出精度の変化

参考文献文字列の分類による抽出精度の差を調べるため，文献種類別の抽出器全てに荒内らの書誌情報抽出器を用いて抽出実験を行い，分類しない場合 (荒内らの方法) と正しく分類された場合の抽出結果を比較した．図 5 に文献種類別に，各参考文献文字列から全ての書誌情報を過不足なく抽出できた精度を示す．また，図 5 の「全体」は文献種類別の抽出結果を集約したものである．

図 5 より，分類しない場合 91.5% の抽出精度が，文献種類別に分類することによって全体として 92.1% の抽出精度となった．

5.3.2　素性による抽出精度の変化

素性による抽出精度の差を調べるため，正しく分類された参考文献文字列に対し，表 4 の素性テンプレートを全ての文献種類別の抽出器に用いた場合と，表 4 に表 5 の素性を加えて実験した場合の抽出結果を比較した結果を表 9，10，11，12 に示す．

これらの表より，この素性の変更によって全文献種類で抽出精度が向上したことがわかる．

5.3.3　提案手法による抽出精度の変化

3.2.3 項の方法で参考文献文字列を文献種類別に分類し，文献種類ごとに素性テンプレートを変更した時の書誌情報の抽出結果と，正解データに基づく誤りのない分類結果に対して，文献種類ごとに素性テンプレートを変更した時の書誌情報の抽出結果，荒内らの手法の書誌情報の抽出結果を比較した．なお，

表 9　「ジャーナル論文」の抽出精度

書誌情報	素性追加前	素性追加後
AUTHOR	0.999	0.999
TITLE	0.993	0.995
JOURNAL	0.992	0.993
VOLUME	0.995	0.995
PUBLISHER	0.538	0.615
DAY	0	0
MONTH	0.999	0.999
YEAR	0.998	0.998
ETC	0.560	0.560
ALL	0.976	0.978

表 10　「書籍」の抽出精度

書誌情報	素性追加前	素性追加後
AUTHOR	0.988	0.988
TITLE	0.953	0.979
JOURNAL	0	0
VOLUME	0.897	0.913
PUBLISHER	0.956	0.960
DAY	-	-
MONTH	1.000	1.000
YEAR	0.995	0.996
ETC	0.944	0.954
ALL	0.876	0.894

表 11　「会議録論文」の抽出精度

書誌情報	素性追加前	素性追加後
AUTHOR	0.996	0.996
TITLE	0.984	0.993
JOURNAL	0.973	0.979
VOLUME	0.984	0.984
PUBLISHER	0.940	0.977
DAY	0.866	0.866
MONTH	0.985	0.989
YEAR	0.997	0.998
ETC	0.944	0.944
ALL	0.923	0.931

表 12　「その他」の抽出精度

書誌情報	素性追加前	素性追加後
AUTHOR	0.975	0.975
TITLE	0.840	0.924
JOURNAL	0.828	0.892
VOLUME	0.868	0.874
PUBLISHER	0.657	0.764
DAY	0.200	0.500
MONTH	0.950	0.968
YEAR	0.984	0.985
ETC	0.761	0.771
ALL	0.665	0.683

3.2.3 項の決定木による分類方法で，ジャーナル論文に 2,354 件，書籍に 612 件，会議録論文に 1,225 件，その他に 306 件の参考文献文字列が分類される．また，誤りのない分類では，

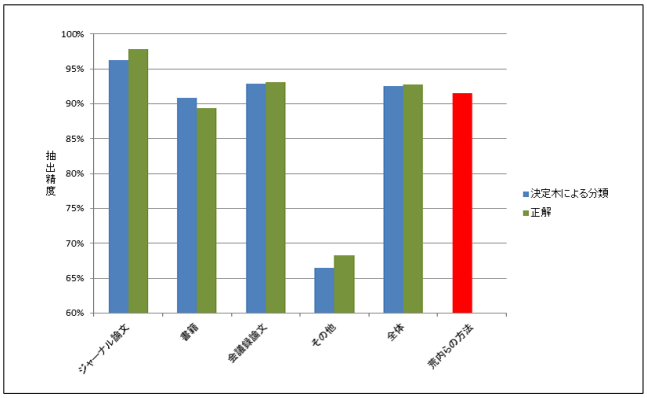

図 6　書誌情報の抽出精度の比較

ジャーナル論文に 2,231 件，書籍に 624 件，会議録論文に 1,247 件，その他に 395 件の参考文献文字列が分類される．各条件で各参考文献文字列から全ての書誌情報を過不足なく抽出できた割合を図 6 に示す．

図 6 より，抽出精度は荒内らの方法の 91.5% に対して，決定木による分類で書誌情報を抽出した場合 92.5%，誤りのない分類で書誌情報を抽出した場合 92.7% であった．

5.3.4　考　　察

図 5 より，参考文献文字列を文献種類別に分類した場合，分類しない場合の抽出精度を上回っている．文献種類ごとにみると，「ジャーナル論文」では抽出精度が大きく向上している．しかし，「書籍」，特に「その他」の抽出精度は大きく下回っている．この原因の一つは，これらの文献種類の参考文献文字列が少ないことである．これには，能動学習を行うことで学習の質を高めるといった対策などがありえる．表 9，10，11，12 より，素性の変更によって各文献種類の抽出精度に向上が見られたが，改善幅は大きくなかった．表中の抽出精度 0.600 未満の書誌要素を含む参考文献文字列は，それぞれの文献種類に 10 件程度しかないため，学習が不十分であった可能性が高い．また，表 12 の「その他」の各書誌要素の抽出精度は，他の文献種類のものと比べて悪い．この「その他」は，今回定義した文献種類に分類できなかった参考文献文字列の集合なので，その参考文献文字列には類似点が少ない．したがって，新たな文献種類の定義を含め，妥当な文献種類についての検討も必要である．

6.　ま　と　め

本稿では，学術論文の参考文献欄の参考文献文字列を文献種類別に分類し，分類した文献種類別の書誌情報抽出器を用いて書誌情報を抽出する方法を提案した．

実験では，まず，参考文献文字列に記載される書誌要素の特徴に基づいて文献種類別に参考文献文字列を分類した．その結果，90.4% の精度で参考文献文字列を文献種類別に分類できた．また，文献種類別に分類した参考文献文字列から書誌情報を抽出した．分類した場合，一部の文献種類では高精度に書誌情報を抽出し，全体でも分類しない場合の抽出精度を上回った．

今後，参考文献文字列の分類については，特徴的な文字列と辞書の拡充などにより書誌要素の有無の判定精度を改善し，分類精度の向上を図りたい．また，抽出精度がより向上するような文献種類の定義について検討していきたい．書誌情報抽出については，文献種類別の参考文献文字列に現れる特徴をさらに詳しく分析し，有効な素性を検討して，抽出精度の向上を図りたい．

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] 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要素抽出，”情報処理学会研究報告，2012-DBS-156(1)，pp.1-8，2012.

[2] C.Cortes and V.Vapnik, ”Support-Vector Networks,” Machine Learning, vol.20, no,3, pp.273-297, 1995.

[3] K.Seymore, A.McCallum and R.Rosenfeld, ”Learning hidden Markov model structure for information extraction,” In AAAI 99 Workshop on Machine Learning for Information Extraction, 1999.

[4] 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.

[5] Takashi Okada, Atsuhiro Takasu, and Jun Adachi, ”Bibliographic Component Extraction Using Support Vector Machines and Hidden Markov Models,” ECDL 2004, LNCS 3332, pp.501-512, 2004.

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

[7] 薬師孝之，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，June. 2009.

[8] 薬師孝之，“ 学術論文 OCR テキストからの書誌情報抽出に関する研究，”岡山大学大学院自然科学研究科修士論文，2009.

[9] Issac G.Councill, C. L. Giles and Min-yen Kan, ”ParsCit: An open-source CRF reference string parsing package,” In Proceedings of language resource and evaluation conference, 2008.

[10] 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.