情報処理学会論文誌　データベース　Vol.8　No.2　18–29 (June 2015)

推薦論文

少量学習データによる参考文献書誌情報抽出精度の向上

川上 尚慶[1,†1,a)]　太田 学[1,b)]　高須 淳宏[2,c)]　安達 淳[2,d)]

受付日 2014年12月21日, 採録日 2015年4月6日

概要：電子図書館の運用には，書誌情報データベースの整備が必須である．特に学術論文の参考文献欄には有用な書誌情報が集約されている．そこで我々は，Conditional Random Field（CRF）を用いて参考文献文字列から書誌情報を自動抽出する手法を提案した．しかし，書誌情報を高精度に抽出するには雑誌ごとに一定量の学習データを用意する必要があり，その生成コストが問題だった．本稿では，学習データが少ない場合に，能動サンプリングと擬似学習データ，転移学習を利用して抽出精度を改善する方法を提案する．実験では，抽出精度と必要とする学習データ件数を評価し，提案手法の有効性について考察した．

キーワード：情報抽出，参考文献文字列，能動サンプリング，擬似学習データ，転移学習

Improvement in Accuracy for Bibliography Extraction from Reference Strings in Academic Papers Using a Small Amount of Training Data

Naomichi Kawakami[1,†1,a)]　Manabu Ohta[1,b)]　Atsuhiro Takasu[2,c)]　Jun Adachi[2,d)]

Received: December 21, 2014, Accepted: April 6, 2015

Abstract: The effective use of digital libraries demands maintenance of bibliographic databases. Especially, the reference fields of academic papers are full of useful bibliographic information. We, therefore, proposed a method of automatically extracting bibliographic information from reference strings using a conditional random field (CRF). However, at least a few hundred reference strings are necessary for training the CRF to achieve high extraction accuracies and the preparation of such human-labeled data for training is usually expensive. As described herein, we propose the use of active sampling, pseudo-training data and transfer learning to improve extraction accuracies with a small amount of training data. Then we evaluate the extraction accuracies and the associated training costs by experimentation and discuss the effectiveness of the proposed approach.

Keywords: information extraction, reference string, active sampling, pseudo-training data, transfer learning

1. はじめに

多数の学術論文を蓄積する電子図書館のサービスを快適に利用するには，検索やソート，文書間リンクなどの機能が必須である．これらの機能を利用するには，著者名やタイトルといった書誌情報が必要となる．しかし，人手でこれらの書誌情報をデータベースに登録するコストは膨大なため，その作業を可能な限り自動化する文書解析技術が求められている．特に学術論文の参考文献欄には，関連分野の文献が集約されており，その書誌情報は有用である．

そこで我々のグループは，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 1 つである Conditional Random Field（CRF）を利用して，論文中の参考文献文字

[1] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University, Okayama 700–8530, Japan
[2] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[†1] 現在，三菱電機インフォメーションシステムズ株式会社
Presently with Mitsubishi Electric Information Systems Corporation
[a)] kawakami@de.cs.okayama-u.ac.jp
[b)] ohta@de.cs.okayama-u.ac.jp
[c)] takasu@nii.ac.jp
[d)] adachi@nii.ac.jp

本稿の内容は 2014 年 11 月の WebDB フォーラム 2014 にて発表され，同シンポジウムプログラム委員会により情報処理学会論文誌データベースへの掲載が推薦された論文である．

列から書誌情報を自動で抽出する手法を提案した [1]．この研究では，参考文献書誌情報抽出精度を評価し，その抽出誤りの詳細な分析を示したが，高い抽出精度を得るには，参考文献文字列の書式が異なる学術雑誌ごとに，少なくとも数百件の参考文献文字列を学習データとして用意する必要があり，その生成コストは無視できない．そこで本稿では，能動サンプリングと擬似学習データ，他雑誌のデータにおいて学習した書誌情報抽出器の推定結果を利用して，少量学習データでの書誌情報抽出精度の向上を図る[*1]．

本稿の構成は次のとおりである．まず，2 章で学術論文からの書誌情報抽出に関する研究を紹介し，3 章で提案している CRF による参考文献書誌情報の自動抽出について説明する．続く 4 章で少量学習データによる書誌情報抽出の方法について説明し，5 章で提案手法の評価実験について述べる．最後に，6 章で本稿をまとめる．

2. 関連研究

学術論文からの書誌情報抽出には，機械学習を用いる手法のほかにルールによる手法が用いられる場合がある．しかし，通常学術雑誌が異なれば，それぞれ参考文献文字列の書式も異なる．**図 1** に示した電子情報通信学会論文誌と情報処理学会論文誌の 2 つの参考文献文字列は，含む書誌要素は同じだが，著者名やタイトル，発行年の書式が異なっている．ルールの場合，書式の異なる参考文献文字列から正確に書誌情報を抽出するには，その書式ごとにルールを定義する必要があり，増大する学術雑誌をかかえる電子図書館では，このようなルールを定義し，管理することは今後ますます困難となることが予想される．そのため，学習データさえ用意すれば利用可能な機械学習による書誌情報抽出手法が多く提案されている．たとえば，CRF [4] を用いた書誌情報抽出に関する研究に，Peng ら [5]，Councill ら [6] の研究がある．Peng らは，英語論文のタイトルページと参考文献欄の単語ごとに書誌要素ラベルを付与し，書誌情報を抽出した．実験では，タイトルページから著者名や所属など 13 項目の書誌情報を抽出し，その 13 項目の平均 F 値は 0.939 だった．また，参考文献欄からも著者名や論文誌名など 13 項目の書誌情報を抽出し，その 13 項目の平均 F 値は 0.915 であった．一方，Councill らは，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である ParsCit を開発した．空白文字をデリミタとして参考文献文字列をトークン列に変換し，英文の単語トークン列に書誌要素ラベルを付与して，書誌情報を抽出した．Cora データセット [7] を対象に，著者名やタイトルなど 13 項目の書誌情報を抽出し，13 項目の平均 F 値が 0.950 であったと報告している．しかし，これらの研究は，英語論文を抽出対象とするため，日本語の参考文献文字列からは書誌情報をうまく抽出できない．また，書誌情報抽出精度のみを評価し，書誌情報抽出にかかるコストの評価は行っていない．本研究では，日英の両言語で書かれた参考文献文字列から書誌情報を抽出し，その抽出精度とコストの両面を評価する．

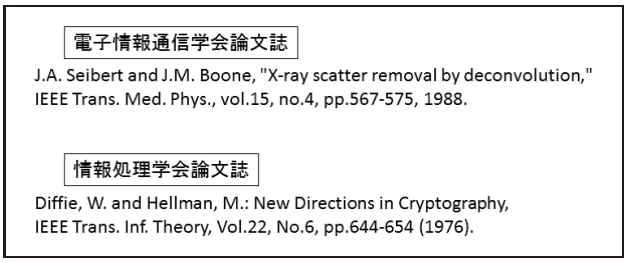


図 1　学術雑誌による参考文献文字列の書式の違い
Fig. 1　Differences in reference string formats of different academic journals.

[*1] 参考文献書誌情報抽出における能動サンプリングと擬似学習データの利用については，文献 [2], [3] において途中経過を報告している．

本研究のような書誌情報抽出にかかるコストには，学習データ生成コストと書誌情報抽出誤りの訂正コストの 2 つがある．このうち，書誌情報抽出における学習データ生成コストの削減に関する研究に，Ohta ら [8] の研究がある．Ohta らは，論文タイトルページからの CRF による書誌情報抽出において，能動サンプリングにより学習データを削減する方法を提案した．能動サンプリングとは，学習に有効なデータを効率良く選択する方法である．Ohta らの書誌情報抽出は，学術論文文書画像のタイトルページに対して，OCR によりレイアウト解析と文字認識を行い，CRF を用いて矩形テキスト領域に対して書誌要素ラベルを付与して，書誌情報を抽出する．このとき，書誌情報抽出結果に確信度を定義し，ある時点の学習モデルで判別が困難なサンプルを優先的に次回の学習データとし，逐次学習モデルを更新した．実験において，書誌情報抽出精度を維持したまま，学習データ量を 3 分の 1 以下に削減できたと報告している．さらに，Ohta らは文献 [9] において，論文タイトルページから CRF により抽出した書誌情報の誤り検出を確信度に基づいて行うことで，人手による後処理のコストを抑えながら，高品質な書誌情報が得られることを示した．しかし，これらの研究がいずれも論文タイトルページからの書誌情報抽出であるのに対して，本研究ではレイアウト情報を持たない参考文献文字列から書誌情報を抽出する．

3. CRF による書誌情報抽出

3.1 書誌情報抽出

本研究では，参考文献文字列を，まずトークン列に変換し，そのトークン列から著者名やタイトルといった主要な書誌情報を抽出する（**図 2**）．参考文献文字列から抽出する書誌情報の一覧と，それに対応する書誌要素ラベルを**表 1** にまとめる．表 1 の Other は他のどの書誌要素にも分類されない書誌要素であり，具体的には所属機関などが含まれる．

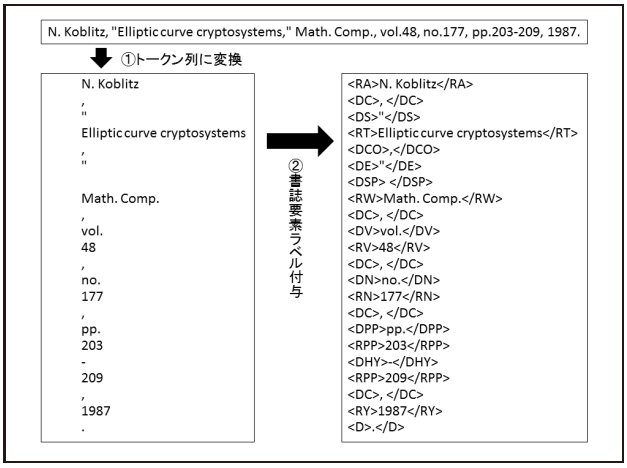


図 2　参考文献書誌情報抽出
Fig. 2　Reference string parsing for bibliography extraction.

表 1　抽出する書誌情報
Table 1　Bibliographic information to be extracted.

書誌要素	書誌要素ラベル
Author	RA
Editor	RE
Translator	RTR
Author Other	RAOT
Title	RT
Booktitle	RBT
Journal	RW
Conference	RC
Volume	RV
Number	RN
Page	RPP
Publisher	RP
Day	RD
Month	RM
Year	RY
Location	RL
URL	RURL
Other	ROT

本研究では図 2 に示すように，トークン列の各トークンに対して <RA> や <RT> などの書誌要素ラベル，または <DC> などのデリミタラベルを付与する．なお，図 2 で D から始まるラベルはデリミタラベルを表し，<DC>（カンマ + 空白）などが定義されている [1]．また文献 [1] では，デリミタによる参考文献文字列の自動トークン化の方法も示されているが，本稿の実験では CRF による書誌要素ラベル付与の精度とコストを評価するため，人手で変換したトークン列を使用する．

3.2　CRF

本研究の書誌情報抽出では，標準的なチェーンモデルの CRF [4] の定義を用い，参考文献文字列を変換して得られるトークン列に書誌要素ラベルを付与する．すなわち，入力トークン系列 $\mathbf{x} = x_1, \ldots, x_n$ 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．

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

ただし，$Z_{\mathbf{x}}$ 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，

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

である．ここで，$f_k(y_{i-1}, y_i, \mathbf{x})$ は $(i-1)$ 番目と i 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．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．

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

本研究の書誌情報抽出では，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るなどして得られるトークンとなる．一方，ラベル y_i は，書誌要素またはデリミタである．

3.3　素性テンプレート

本研究では工藤が作成した CRF++ [*2] [10] を利用して書誌要素ラベルを付与する．CRF++で用いる素性テンプレートを**表 2** にまとめる．この素性テンプレートは 47 個の Unigram 素性と 1 個の Bigram 素性の計 48 個の素性で構成されている．これらは言語的な素性のみで，ページ内での位置情報などのレイアウトに関する素性はない．Unigram 素性には，トークンのトークン列における出現位置や文字数，トークンを構成する文字種とその割合，トークンの先頭・末尾から 4 文字目までの文字列，大文字などの特定の文字や特徴的な文字列の有無，各種辞書のエントリの有無などを用いる．ここで，特徴的な文字列とは，たとえば “Proc.” のことで，これがあれば，そのトークンは Conference を表す書誌要素である可能性が高い．また，辞書としては，人名[*3]，論文誌名[*4]，会議名[*5]，出版社名[*6]，地名[*7]，月名の辞書を用意した．表 2 の各素性の括弧内の数字はトークンの相対位置を表し，0 が現在のトークン，また $i \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ である．なお，表 2 で，“数” はその素性に関する実際の素性数を表す．たとえ

[*2] http://crfpp.googlecode.com/svn/trunk/doc/index.html
[*3] http://www.census.gov/genealogy/names/など
[*4] http://science.thomsonreuters.com など
[*5] http://www.allconferences.com/など
[*6] http://www.narosa.com/nbd/PublisherDistributed.asp など
[*7] http://www.fallingrain.com/world/index.html など

表 2　素性テンプレート
Table 2　Feature template.

種類	素性	数	内容
Unigram	<token_ab_pos(0)>	1	トークン列における絶対的な出現位置
	<token_re_pos(0)>	1	トークン列における相対的な出現位置
	<num_char(0)>	1	トークンの文字数
	<num_word(0)>	4	トークン内の単語数
	<num_period(0)>	4	トークン内のピリオド数
	<f_kanji(0)>	1	トークン内の漢字数の割合
	<f_hiragana(0)>	1	トークン内のひらがな数の割合
	<f_katakana(0)>	1	トークン内のカタカナ数の割合
	<f_alphabet(0)>	1	トークン内の全角アルファベット数の割合
	<f_digit(0)>	1	トークン内の全角数字数の割合
	<h_alphabet(0)>	1	トークン内の半角アルファベット数の割合
	<h_digit(0)>	1	トークン内の半角数字数の割合
	<h_symbol(0)>	1	トークン内の記号数の割合
	<first_1-4_string(0)>	4	トークンの先頭から四文字目までの文字列
	<last_1-4_string(0)>	4	トークンの末尾から四文字目までの文字列
	<token(0)>	1	トークン自身
	<last_char(i)>	1	トークンの最後の文字種
	<token_lc(i)>	1	トークンを小文字にした文字列
	<capital(i)>	1	トークン中の大文字の有無
	<digit(i)>	1	トークン中の数字の有無
	<symbol(i)>	2	トークン中の記号の有無
	<keyword(i)>	2	トークン中の特徴的な文字列の有無
	<dictionary(i)>	8	辞書的素性
	<num_token(0)>	1	参考文献文字列のトークン数
	<editor(0)>	1	参考文献文字列中の Editor に関する記述の有無
	<URL(0)>	1	参考文献文字列中の URL に関する記述の有無
Bigram	< y(-1), y(0)>	1	ラベルの遷移

ば，<keyword(i)> の場合，トークンに含まれる特徴的な文字列の種類を区別する素性と，特徴的な文字列とトークンのマッチングの照合が，“完全一致”，“前方一致”，“後方一致”，または “部分一致” のいずれかという照合の種類を区別する素性の 2 つからなる．

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する．

4. 少量学習データによる書誌情報抽出

4.1 確信度

少量学習データによる参考文献書誌情報抽出において高い抽出精度を得るには，効率良く学習を進める必要がある．そこで本研究では，少量学習データでなるべく高い精度を得るため，能動サンプリングを行う．さらに，擬似学習データを利用したり，他雑誌で学習した CRF の書誌情報抽出器の推定ラベルを素性に加えたりすることで，さらなる精度向上を図る．

能動サンプリングは，ある時点の学習モデルで書誌情報抽出が困難な参考文献文字列を，優先的に選択して次回の学習データとし，逐次学習モデルを更新する．そのため，文献 [11] を参考に書誌情報抽出の困難さを表す尺度として，以下の 3 つの確信度を定義する．これらの確信度の一部は，CRF による論文タイトルページからの書誌情報抽出における能動サンプリング [8] やその書誌情報抽出の誤り検出 [9] において有効性が確認されているため，参考文献書誌情報抽出においても有望であることが期待できる．しかし，タスクによって適切な確信度は異なる可能性があるため，本稿では参考文献書誌情報抽出の能動サンプリングにおいてこれらの確信度を比較する．

4.1.1 Normalized Likelihood（NLH）

1 つ目の確信度は CRF の出力する条件付き確率に基づいて定める．CRF は式 (1) に示す入力系列に対する条件付き確率が最大になるような出力ラベル系列 $\mathbf{y}^*$ を導出する．よって，$P(\mathbf{y}^*|\mathbf{x})$ の値が小さければ，その参考文献文字列に対するラベル付与は困難であると見なすことができるので，この $P(\mathbf{y}^*|\mathbf{x})$ の値を確信度として利用する．ただし，この条件付き確率は入力系列 $\mathbf{x}$ の長さの影響を受けるため，入力系列の長さで正規化した以下の式で確信度を定義する．

$$c_{NLH}(\mathbf{x}) = \frac{\log(P(\mathbf{y}^* \mid \mathbf{x}))}{|\mathbf{x}|} \tag{4}$$

ここで，$|\mathbf{x}|$ は入力系列 $\mathbf{x}$ の長さであり，参考文献文字列を構成するトークンの数を表す．

なお, この確信度は, 文献 [11] において “least confidence” とされるサンプルを優先的に学習データに選ぶのと同じ考えに基づく．またこの確信度は，CRF が付与したラベル系列の確率に基づく．

4.1.2 Mininum Probability of Token Assignment（MP）

2 つ目の確信度は，参考文献文字列中の各トークンに付与されたラベルの周辺確率そのものを利用する．入力系列 $\mathbf{x}$ に対して，Y_i を参考文献文字列中の i 番目のトークン x_i に対して付与されうるラベルを表す確率変数とする．L を付与できるラベルの集合とすると，$P(Y_i = l|\mathbf{x})$ はラベル $l \in L$ が x_i に割り当てられる周辺確率を表している．よって，確率 $max_{l \in L} P(Y_i = l|\mathbf{x})$ は i 番目のトークンに注目したラベル付与の確信度と見なすことができる．そして，参考文献文字列中の各トークンに対するこのラベル付与の確信度の中で最小のものを，その参考文献文字列の書誌情報抽出の確信度とする．具体的には以下の式で定義する．

$$c_{MP}(\mathbf{x}) = \min_{1 \le i \le |\mathbf{x}|} \max_{l \in L} P(Y_i = l|\mathbf{x}) \tag{5}$$

この確信度は，トークン系列中の 1 トークンへのラベル付与確率に基づく．

4.1.3 Average Token Entropy（ATE）

3 つ目の確信度は全ラベル候補の周辺確率のエントロピーに基づいて定める．この確信度では，各トークンに付与された周辺確率が最大のラベルだけでなく，その他のラベルも考慮できる．エントロピーの値が大きいほど，より多くの書誌要素ラベルに確率が分散しているため，ラベル付与が困難であると判断する．まず，参考文献文字列中の各トークンに付与されうるすべての書誌要素ラベルの確率を計算し，以下の式でエントロピーを算出する．

$$H(Y; \mathbf{x}) = \sum_{l \in L} -P(Y = l|\mathbf{x}) \log P(Y = l|\mathbf{x}) \tag{6}$$

これを全トークンで平均し，値を正負逆転させたものを確信度とする．具体的には以下の式で定義する．

$$c_{ATE}(\mathbf{x}) = -\frac{\sum_{1\le i\le|\mathbf{x}|} H(Y_i;\mathbf{x})}{|\mathbf{x}|} \quad (7)$$

なお，この確信度は文献 [11] に示された “token entropy” の値の正負を逆転させたものと等しい．またこの確信度は，トークン系列全体のすべての付与ラベル候補の確率に基づく．

4.2　能動サンプリング

4.1 節で定義した確信度を利用し，本研究では，以下の手順で能動サンプリングを行う．これは，確信度の低い，つまり書誌情報抽出が困難なサンプルは学習に有効であるという考え方に基づいている．

(1) ラベル未付与の参考文献文字列 S を大量に収集する．
(2) 少量の参考文献文字列 $S_0 \subset S$ を選出して，ラベルを付与し，これを初期学習データとして CRF M_0 を学習する．
(3) 以下の手順を書誌情報抽出精度の向上が収束するまで繰り返す．
  (a) CRF M_{t-1} を用いて，参考文献文字列 $S - \cup_{i=0}^{t-1} S_i$ の確信度 $c.(\mathbf{x})$ を式 (4)，(5)，または (7) により算出する．
  (b) 各確信度を用いて参考文献文字列を昇順にランキングする．
  (c) 上位 n 件の参考文献文字列 $S_t \subseteq S - \cup_{i=0}^{t-1} S_i$ を選出する．
  (d) 参考文献文字列 S_t に人手でラベルを付与する．
  (e) ラベル付与した参考文献文字列 $\cup_{i=0}^{t} S_i$ を学習データとして CRF M_t を学習する．

したがって，t 回目の学習に利用される学習データ件数は

$$\text{学習データ件数} = n_0 + nt \quad (8)$$

となる．ただし，$n_0 = |S_0|$，$n = |S_t|$ とする．

4.3　擬似学習データ

本稿では，書誌要素ラベル付与において擬似学習データを学習データに加えて抽出精度の向上を図る方法を提案する．

擬似学習データは，能動サンプリングによって人手でラベルを付与する確信度が低い参考文献文字列を基に生成する．この参考文献文字列の各トークンに割り当てられた書誌要素ラベルと同じ書誌要素の文字列を書誌情報データベースから無作為に選出し，元のトークンと置換して生成した参考文献文字列を擬似学習データとする．ここで，置換する書誌要素の種類は表 1 に示した書誌情報すべてである．また，この文字列の置換は，参考文献文字列の書誌要素ラベルを割り当てられたすべてのトークンに対して行い，デリミタラベルを割り当てられたトークンについては変更しない．したがって，擬似学習データを構成する各書誌要素は元の参考文献文字列と異なるが，デリミタは同一である．この方法で，1 つの参考文献文字列から任意の数の擬似学習データを生成する．また，人手でラベル付与した参考文献文字列を基に自動生成するので，擬似学習データの作成コストは無視できる．

4.2 節の能動サンプリングでは，t 回目の学習において，人手でラベル付与した参考文献文字列 S_t を学習データに追加し，参考文献文字列 $\cup_{i=0}^{t} S_i$ を学習データとして CRF M_t を学習する．擬似学習データを追加する場合，参考文献文字列 S_t から擬似学習データ P_t を生成し，これを追加した $\cup_{i=0}^{t} S_i \cup P_i$ を学習データとして CRF M_t を学習する．ただし，$P_0 = \emptyset$ とする．

1 件の参考文献文字列につき m 件の擬似学習データを生成した場合，t 回目の学習データ件数は以下の式で表せる．

$$\text{学習データ件数} = n_0 + n(1+m)t \quad (9)$$

ただし，$n_0 = |S_0|$, $n = |S_t|$, m は参考文献文字列 1 件につき生成する擬似学習データ件数である．よって，$|P_t| = m|S_t|$ が成り立つ．

4.4　他雑誌用書誌情報抽出器の利用

雑誌ごとに参考文献文字列の書式は異なるため，書誌情報を高精度に抽出するには，通常対象雑誌の学習データで CRF を学習する．たとえ，他雑誌のラベル付き学習データが大量にあり，それで学習した書誌情報抽出器があっても，それらは利用されない．しかし，抽出する書誌情報や参考文献文字列に表れる特徴には雑誌の種類によらない共通点があるため，これらの情報を対象雑誌の CRF でも利用できれば，書誌情報抽出精度の向上が期待できる．そこで本稿では，他雑誌について学習した書誌情報抽出器を利用して，対象雑誌の書誌情報抽出器の書誌情報抽出精度を向上させる手法を提案する．

具体的には，対象雑誌のラベル付き学習データが少量の場合に，対象雑誌の書誌情報抽出器の CRF の素性として，他雑誌の書誌情報抽出器が推定した書誌要素ラベルを追加する．すなわち，表 2 に示した対象雑誌の素性テンプレートに，Unigram 素性として，他雑誌 J の学習データで学習した CRF $M^{[J]}$ の出力したラベル <label$M^{[J]}(0)$> を追加する．4.2 節で示した能動サンプリングの手順においては，(2) や (3) の (e) の学習結果が変わるので，結果的に学習データとして選択する参考文献文字列も変わることになる．この提案手法は，基本的には，文献 [12] で転移学習の一分類として示されている “transductive transfer learning” に位置付けられるが，目標ドメインのラベルを能動的に取得することによって精度の高い CRF を学習することを試み

ている．

5. 評価実験

5.1 実験環境

提案手法の有効性を検証するため，評価実験を行う．実験データとして，以下の参考文献文字列コーパスを利用する．

IEICE-J　2000 年の電子情報通信学会和文論文誌に含まれる参考文献文字列 4,787 件（うち，和文 2,193 件）

IEICE-E　2000 年の電子情報通信学会英文論文誌に含まれる参考文献文字列 4,497 件（うち，和文 0 件）

IPSJ　2000 年の情報処理学会論文誌に含まれる参考文献文字列 4,574 件（うち，和文 1,537 件）

また，評価指標として，1 つの参考文献文字列に含まれるすべての書誌要素が過不足なく正確に抽出された参考文献文字列数を，全参考文献文字列数で割った書誌情報抽出精度を用いる．ただし，表 1 の書誌情報を，先行研究 [1] に倣って**表 3** のように集約し，同じ分類のものは正解判定において区別しない．そして，表 3 の分類に基づいて，CRF が参考文献文字列に含まれる書誌要素を構成するすべてのトークンに，正しい書誌要素ラベルを付与した場合を成功と見なす．つまり，1 つでも間違った書誌要素ラベルを付与した場合，残りすべての書誌要素ラベルが正しく付与されていても失敗とする．ただし，デリミタの種類の誤りは無視する．実験において，CRF++の学習パラメータはデフォルトの値を利用し，5 分割交差検定で書誌情報抽出精度を算出した．

表 3　書誌要素ラベルの再分類
Table 3　Reclassification of bibliographic element labels.

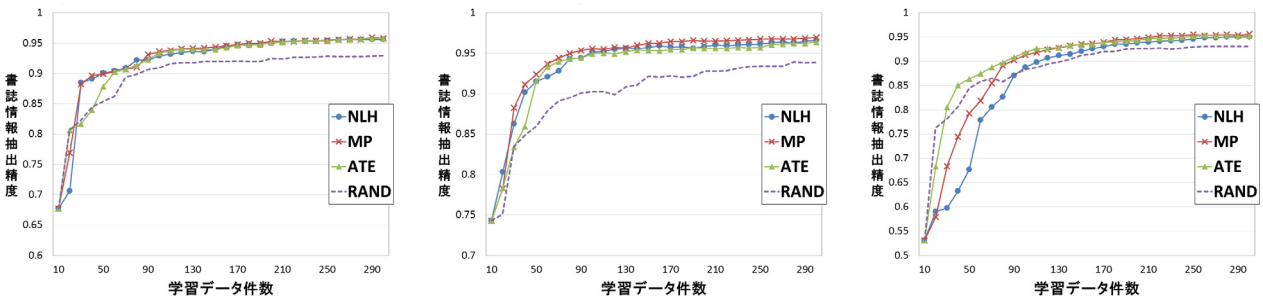
書誌要素ラベル	分類名
RA, RE, RTR, RAOT	AUTHOR
RT, RBT	TITLE
RW, RC	JOURNAL
RV, RN, RPP	VOLUME
RP	PUBLISHER
RD	DAY
RM	MONTH
RY	YEAR
RL, RURL, ROT	OTHER

5.2 能動サンプリングの効果

まず，能動サンプリングの効果を**図 3** に示す．5 分割交差検定を行うため，各雑誌の参考文献文字列を 5 つに分割し，そのうち 4 つを学習用データ，残りの 1 つをテストデータとする．ここで，4.2 節において，手順 (2) の初期学習データは学習用データ S から無作為に 10 件選出し ($n_0 = 10$)，手順 (3) の (c) では $n = 10$ とする．また，能動サンプリングの有無による抽出精度の比較のため，ラベルを付与する参考文献文字列を無作為に選出した場合を RAND と記し，ベースラインとする．図 3 の縦軸は書誌情報抽出精度，横軸は学習データ件数，凡例はサンプリングに使用した確信度の種類，またはその方法を表す．図 3 に示すように，確信度の種類によって差はあるものの，おおむね RAND より少ない学習データ件数で高い抽出精度が得られた．特に，図 3 (b) の IEICE-E では，その効果が顕著だった．また，確信度の種類で比較すると，IEICE-J と IEICE-E では MP，IPSJ では ATE が効果的であった．

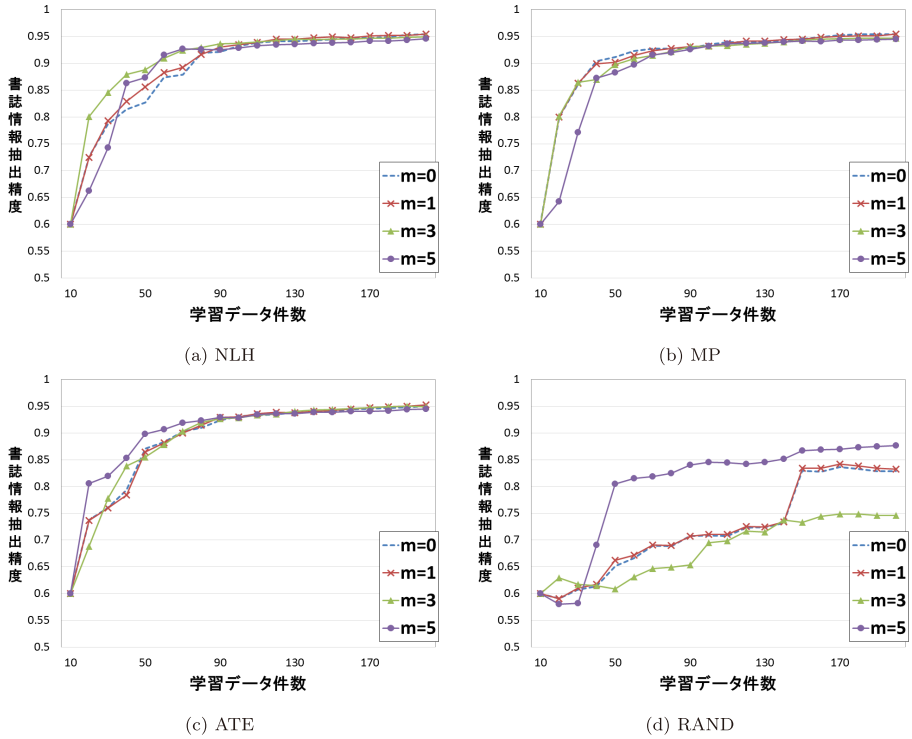
5.3 擬似学習データの効果

ここでは，4.3 節で説明した擬似学習データの利用により，書誌情報抽出精度がどの程度向上するかを実験により確認する．本実験は，5.2 節と同様の条件で行うが，学習データを追加する際，4.3 節で説明した擬似学習データも追加して CRF を再学習する．ここで，擬似学習データ生成に利用する書誌情報データベースは，抽出対象雑誌以外の雑誌 2 誌分の参考文献文字列コーパスから生成した．なお，5.2 節の実験において，いずれの文献種類でも学習データが 200 件程度で書誌情報抽出精度の上昇がほぼ収束しているため，本実験では，人手で書誌要素ラベルを付与した学習データが 200 件までの書誌情報抽出精度を示す．



(a) IEICE-J　(b) IEICE-E　(c) IPSJ

図 3　能動サンプリングにおける学習データ件数と書誌情報抽出精度
Fig. 3　Numbers of training samples and bibliographic information extraction accuracies when using active sampling.

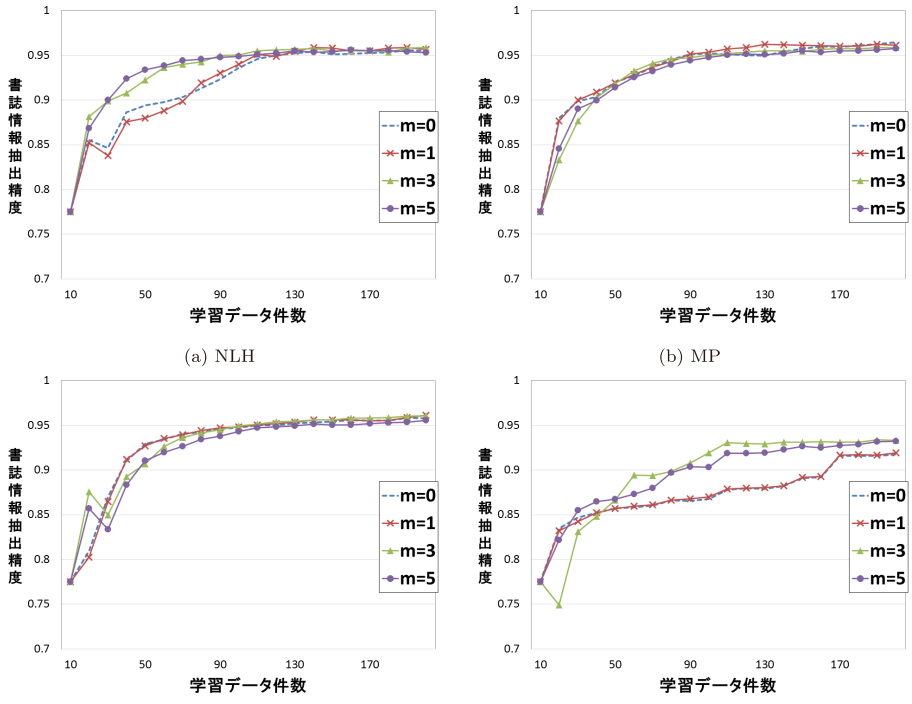


(a) NLH　(b) MP

(c) ATE　(d) RAND

図 4　擬似学習データを追加した場合の書誌情報抽出精度（IEICE-J）

Fig. 4　Bibliographic information extraction accuracies obtained by using pseudo training data (IEICE-J).



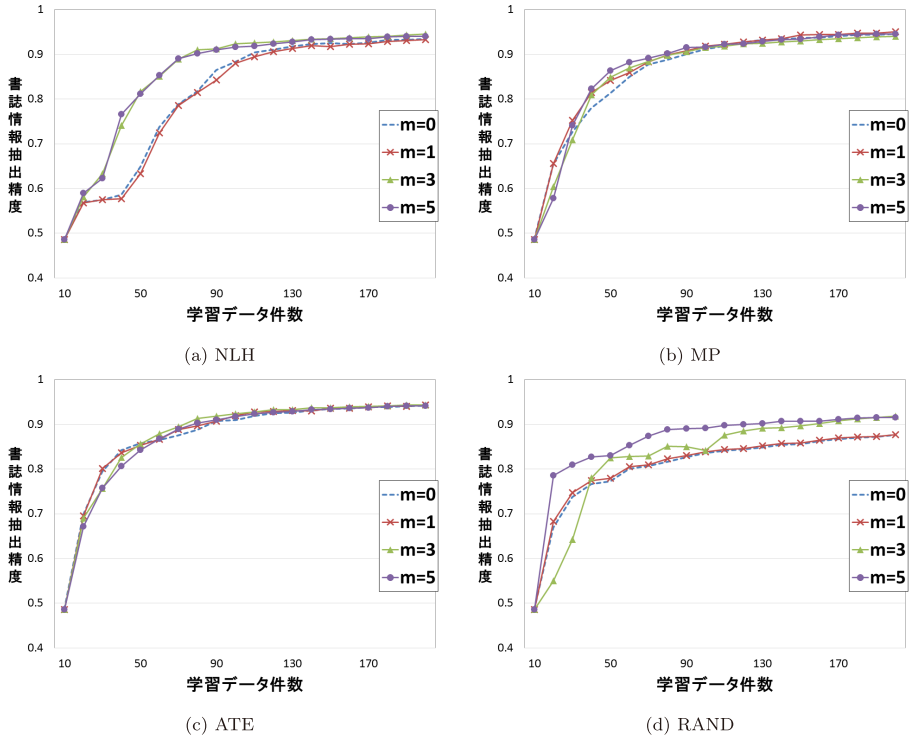
(a) NLH　(b) MP

(c) ATE　(d) RAND

図 5　擬似学習データを追加した場合の書誌情報抽出精度（IEICE-E）

Fig. 5　Bibliographic information extraction accuracies obtained by using pseudo training data (IEICE-E).

生成する擬似学習データを，能動サンプリングに用いる人手でラベル付与した参考文献文字列 1 件につき 0 件，1 件，3 件，5 件として実験し，それぞれの書誌情報抽出精度を比較する．ここで，0 件の場合は能動サンプリングのみ適用した場合の抽出精度であり，これがベースラインとなる．**図 4**，**図 5**，**図 6** に擬似学習データを追加した場合の



(a) NLH　(b) MP

(c) ATE　(d) RAND

図 6　擬似学習データを追加した場合の書誌情報抽出精度（IPSJ）

Fig. 6　Bibliographic information extraction accuracies obtained by using pseudo training data (IPSJ).

書誌情報抽出精度を示す．これらの図の縦軸は書誌情報抽出精度，横軸は人手で書誌要素ラベルを付与した学習データの件数，凡例は人手でラベル付与した参考文献文字列 1 件につき生成する擬似学習データの件数（m）である．

図 4，図 5，図 6 より，NLH と RAND の場合は，いずれの雑誌においても擬似学習データの追加により書誌情報抽出精度に向上が見られたが，MP と ATE の場合は雑誌によっては効果が見られなかった．確信度に NLH を用いた場合，人手でラベル付与した学習データ件数が 50 件で，$m = 5$ のとき，書誌情報抽出精度は $m = 0$，つまり擬似学習データを利用しない場合に比べ，IEICE-J では約 5 ポイント，IEICE-E では約 4 ポイント，IPSJ では約 17 ポイント向上した．擬似学習データの追加により，NLH でも能動サンプリングにおいて性能の良かった MP と同等の書誌情報抽出精度を達成することができたが，それ以上の向上は見られなかった．

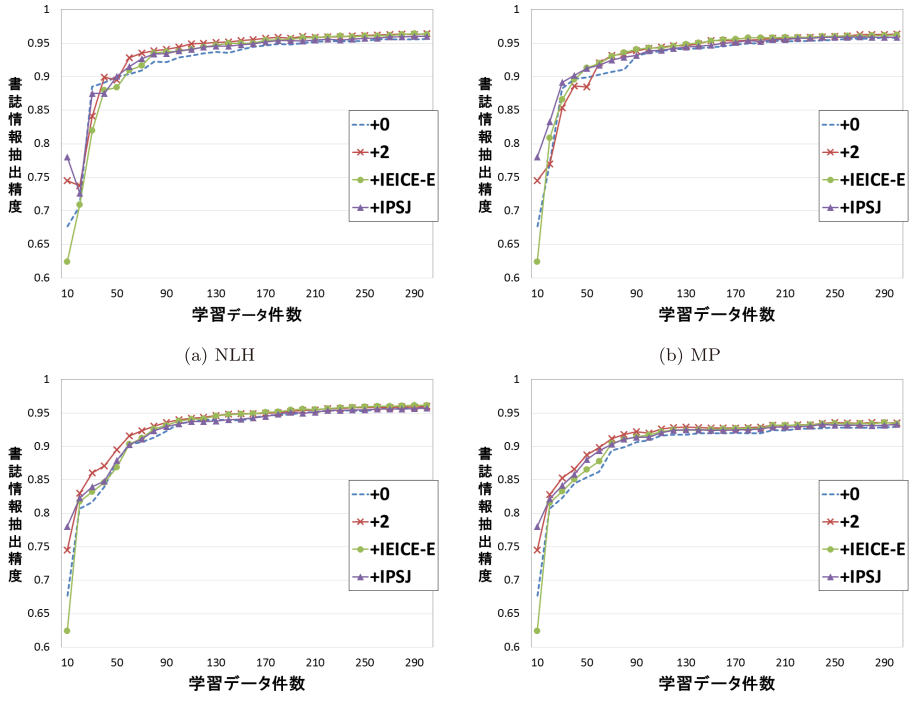
生成する擬似学習データ件数は多いほど抽出精度が良くなる傾向が見られるが，人手でラベル付与した学習データ件数が少ないときはかえって悪くなる場合がある．擬似学習データは，ある時点の学習モデルで書誌情報抽出が困難な参考文献文字列の書誌要素を置換して生成するため，このような参考文献文字列は，それぞれの学術雑誌論文の典型的な参考文献文字列とは異なる特徴を持っている可能性がある．したがって，生成する擬似学習データが多すぎると，特異なパターンを持つデータを大量に学習してしまい，抽出精度が悪くなる可能性がある．

5.4　他雑誌用書誌情報抽出器の効果

ここでは，4.4 節で説明したように，他雑誌で学習した書誌情報抽出器によるラベル推定結果を，対象雑誌の抽出器の CRF の素性に加えることで，対象雑誌の抽出器の性能が向上するか，実験により評価する．なお，他雑誌の書誌情報抽出器は，それぞれ全参考文献文字列数の 5 分の 4 をすべて学習データとして学習した CRF を利用する．また，本実験でも 5 分割交差検定を行うので，実験結果はすべて 5 回の平均値である．いずれの場合も 5.2 節の実験と同様に，能動サンプリングによって学習データを逐次追加する．

この素性の追加の有無による，学習データ件数と書誌情報抽出精度の関係を雑誌ごとにまとめたものが**図 7**，**図 8**，**図 9** である．これらの図の縦軸は書誌情報抽出精度，横軸は学習データ件数を表す．また，凡例は素性として追加した推定ラベルを出力した CRF の学習データの雑誌の種類を表し，+0 が何も素性を加えない能動サンプリングの結果，+2 は他の 2 雑誌とも追加した場合の結果である．

図 7，図 8，図 9 より，提案した素性の追加によって，雑誌によって効果に差はあるが，対象雑誌の学習データが 300 件未満では書誌情報抽出精度が向上した．3 種類すべての確信度でおおむね書誌情報抽出精度が向上しており，雑誌で比較すると，特に IEICE-E で効果が大きかった（図 8）．

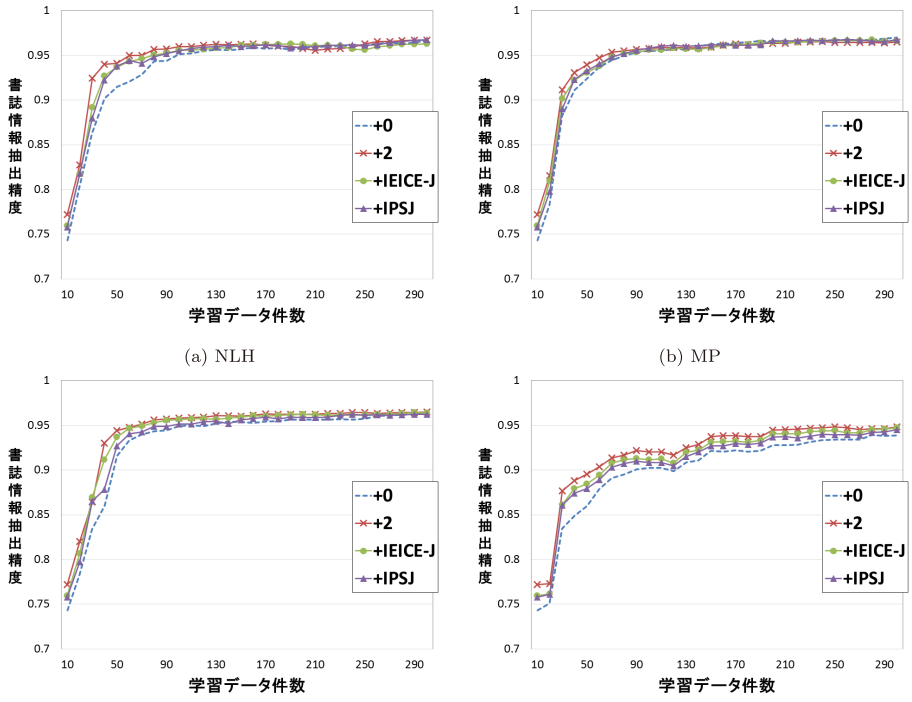


(a) NLH　(b) MP

(c) ATE　(d) RAND

図 7　推定ラベルを素性に追加した場合の書誌情報抽出精度（IEICE-J）

Fig. 7　Bibliographic information extraction accuracies obtained by adding estimated labels as a feature (IEICE-J).



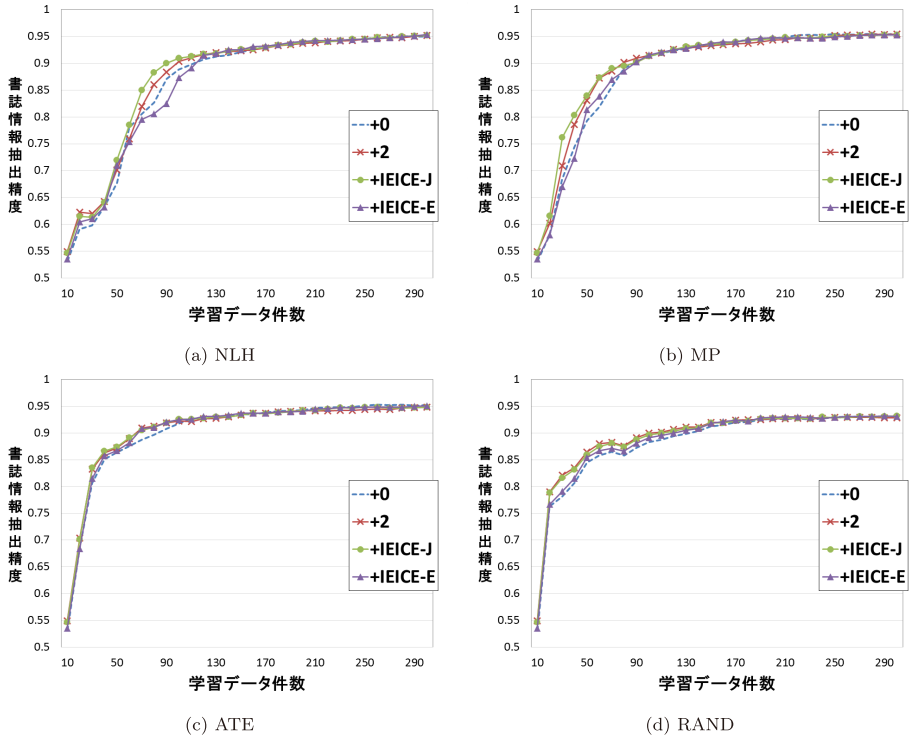
(a) NLH　(b) MP

(c) ATE　(d) RAND

図 8　推定ラベルを素性に追加した場合の書誌情報抽出精度（IEICE-E）

Fig. 8　Bibliographic information extraction accuracies obtained by adding estimated labels as a feature (IEICE-E).

IEICE-J からの書誌情報抽出において，学習データ件数が 10 件のときに大きな改善が見られたことは特筆に値する．たとえば，IPSJ で学習した書誌情報抽出器による推定結果を素性に追加した場合，およそ 11 ポイント向上した（図 7 (a)，(b)，(c)，(d)）．図 7 の (a)，(b)，(c)，(d) において，この 10 件は初期学習データであるため，その参



(a) NLH　(b) MP

(c) ATE　(d) RAND

図 9　推定ラベルを素性に追加した場合の書誌情報抽出精度（IPSJ）

Fig. 9　Bibliographic information extraction accuracies obtained by adding estimated labels as a feature (IPSJ).

考文献文字列は同一であることから，少量学習データでの学習に提案した素性は有効であるといえる．

次に，各雑誌について，追加した雑誌の種類による効果を比較する．まず，図 7 の (a)，(b)，(c)，(d) すべてにおいて，IEICE-E を追加した場合，初期学習データで抽出精度が悪くなっている．この原因として，IEICE-J には和英両言語の論文が含まれるが，IEICE-E には英語論文しかないため，IEICE-E の書誌情報抽出器による日本語の参考文献文字列に対するラベル付与に誤りが多いことがあげられる．次に，図 8 では，いずれの雑誌を素性に追加した場合も学習データ 10 件から 300 件にかけて抽出精度が向上しているが，IEICE-J と IPSJ を比較すると，特に図 8 (c) では，学習データ 30 件から 150 件にかけて IEICE-J の方が効果が大きかった．これは，IEICE-E と IEICE-J は同じ学会の論文誌であり，書式などに類似点が多いことが理由にあげられる．最後に図 9 を見ると，追加素性の効果はあったが，その大きさは最も小さかった．また，たとえば図 9 (a) のように IEICE-E を素性に追加した場合，抽出精度が悪化することがある原因は，IEICE-J の場合と同様に，IEICE-E は英文のみであることがあげられる．一方，言語の種類が同じ IEICE-J の追加効果が小さいのは，IPSJ と IEICE-J は学会が異なるため，書式などに類似点が少ないことが理由にあげられる．以上より，ある程度特徴の類似する論文誌で学習した抽出器の書誌要素推定結果を素性に追加すれば，抽出精度の向上が見込めるが，類似点の少ない場合は効果がなかったり，逆効果となったりすることが分かる．よって，参考文献文字列の書式などの類似点を精査して，たとえば，特定の書誌要素の推定結果のみを素性に追加することなども検討したい．

表 4　他雑誌用抽出器による書誌情報抽出精度

Table 4　Bibliographic information extraction accuracies of extractors for a different journal.

抽出対象データ	書誌情報抽出器の学習データ	精度
IEICE-J	IEICE-J	0.9637
	IEICE-E	0.5567
	IPSJ	0.8901
IEICE-E	IEICE-J	0.9475
	IEICE-E	0.9706
	IPSJ	0.9199
IPSJ	IEICE-J	0.8266
	IEICE-E	0.7298
	IPSJ	0.9652

5.5　考察

他雑誌用書誌情報抽出器による推定ラベルがどの程度正しいかを検証する実験を行った．5 分割交差検定を行うため，まず 5.1 節で説明した実験データの 5 分の 4 を学習データとした．よって，各雑誌の学習データ件数はいずれも 4,000 件弱となる．この学習データを用いて各書誌情報抽出器を学習し，3 雑誌すべてを対象に書誌情報を抽出した．その書誌情報抽出結果を**表 4** に示す．表 4 より，

表 5　全データ学習時の書誌情報抽出精度
Table 5　Bibliographic information extraction accuracies when training the CRFs with all their training data.

抽出対象データ	追加素性の雑誌	精度
IEICE-J	+0	0.9637
	+IEICE-E	0.9664
	+IPSJ	0.9647
	+2(IEICE-E, IPSJ)	0.9666
IEICE-E	+0	0.9706
	+IEICE-J	0.9686
	+IPSJ	0.9709
	+2(IEICE-J, IPSJ)	0.9695
IPSJ	+0	0.9652
	+IEICE-J	0.9635
	+IEICE-E	0.9641
	+2(IEICE-J, IEICE-E)	0.9620

IEICE-E で学習した書誌情報抽出器は，5.4 節で述べたとおり，日本語の参考文献文字列を学習していないので，IEICE-J，IPSJ のどちらからの抽出でも特に精度が悪かった．また，IEICE-E の抽出対象データに対して，IEICE-J と IPSJ で学習した書誌情報抽出器でラベルを付与すると，IEICE-J で学習した書誌情報抽出器の方が精度が高くなっている．IEICE-E と IEICE-J はいずれも電子情報通信学会の論文誌で，参考文献文字列の書式も似ているため，類似点の多い雑誌の方が追加素性としての効果が大きいことが分かる．

次に，十分な量のラベル付き学習データがある場合に，提案した追加素性が書誌情報抽出精度の向上に効果があるか検証する実験を行った．5.1 節で説明した実験データのすべてに表 4 の実験で使用した他雑誌の書誌情報抽出器による付与ラベルを素性として追加し，5 分割交差検定を行うためそれらの 5 分の 4 を学習データとした．よって，各雑誌の学習データ件数はそれぞれ 4,000 件弱となる．この学習データを用いて対象雑誌の書誌情報抽出器を学習し，書誌情報を抽出した結果を**表 5** に示す．表 5 の “追加素性の雑誌” は追加素性のラベルを推定した抽出器の学習データの雑誌の種類を表し，+0 は素性を追加しない場合，+2 は他の 2 雑誌とも追加した場合を表す．表 5 より，素性を追加した場合の精度が +0 を上回る組合せが 4 つ，下回る組合せが 5 つあり，いずれも差はほとんどなかった．したがって，目標ドメインのラベル付き学習データが十分ある場合は，本稿で提案した素性を追加しても書誌情報抽出精度に大きな影響は与えないといえる．

6. まとめ

本稿では，能動サンプリングと擬似学習データ，他雑誌において学習した書誌情報抽出器の CRF の推定結果を利用して，少量学習データによる書誌情報抽出の精度を改善する手法を提案した．提案手法では，擬似学習データを自動生成して学習データに加えた．また，他雑誌用の書誌情報抽出器の書誌要素推定結果を，対象雑誌の書誌情報抽出器の CRF の素性に追加した．これらを能動サンプリングと組み合わせることで，少量学習データにおける書誌情報抽出精度の改善を図った．

国内の 3 種類の学術論文誌の参考文献文字列に対して実験を行い，能動サンプリングと擬似学習データ，他雑誌用抽出器の書誌要素推定結果が書誌情報抽出精度に与える効果を評価した．その結果，能動サンプリングが少量データでの学習にきわめて有効であること，擬似学習データを追加したり，また，他雑誌の学習データで学習した CRF が出力したラベルを素性に追加したりすることで，少量学習データによる書誌情報抽出精度が向上することを確認した．少量学習データにおいて，さらに効果的に書誌情報抽出精度を向上させるには，各雑誌の参考文献文字列の書式などの類似点を精査し，転移させる素性のより詳細な検討が必要であると考えている．

本稿に示した知見から，多様な学術論文を扱う電子図書館において，同じ体裁を持つ学術雑誌ごとに，能動サンプリングによって比較的少量の学習データから参考文献書誌情報を整備できることが分かる．さらに，このようにしていくつかの雑誌の参考文献書誌情報が一定量整備されれば，それを学習データとして高精度の書誌情報抽出器を獲得し，その抽出器に未整備の学術雑誌の参考文献書誌情報推定を行わせることで，その学術雑誌の参考文献書誌情報抽出のさらなる省力化が図れる見通しが得られたといえる．

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），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．

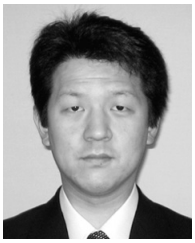
参考文献

[1] Ohta, M., Arauchi, D., Takasu, A. and Adachi, J.: Empirical Evaluation of CRF-Based Bibliography Extraction from Reference Strings, *Proc. IAPR DAS 2014*, pp.287–292 (2014).
[2] 川上尚慶，太田　学，高須淳宏，安達　淳：CRF による参考文献書誌情報抽出のための学習コストの削減，*DEIM Forum 2014*, C5-3 (2014).
[3] Kawakami, N., Ohta, M., Takasu, A. and Adachi, J.: Cost Evaluation of CRF-Based Bibliography Extraction from Reference Strings, *Proc. ICADL 2014*, LNCS 8839, pp.268–278 (2014).
[4] 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).
[5] Peng, F. and McCallum, A.: Accurate Information Extraction from Research Papers Using Conditional Random Fields, *HLT-NAACL 2004*, pp.329–336 (2004).
[6] Councill, I.G., Giles, C.L. and Kan, M.-Y.: ParsCit: An Open-Source CRF Reference String Parsing Pack-

age, *Proc. 6th International Conference on Language Resources and Evaluation*, pp.661–667 (2008).
[7] McCallum, A., Nigam, K., Rennie, J. and Seymore, K.: Automating the Construction of Internet Portals with Machine Learning, *Information Retrieval*, Vol.3, No.2, pp.127–163 (2000).
[8] Ohta, M., Inoue, R. and Takasu, A.: Empirical Evaluation of Active Sampling for CRF-Based Analysis of Pages, *Proc. IEEE IRI 2010*, pp.13–18 (2010).
[9] Ohta, M., Inoue, R. and Takasu, A.: Empirical Evaluation of CRF-Based Bibliography Extraction from Research Papers, *IADIS International Journal on Computer Science and Information Systems*, Vol.7, No.2, pp.18–31 (2012).
[10] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. EMNLP 2004*, pp.230–237 (2004).
[11] Settles, B. and Craven, M.: An Analysis of Active Learning Strategies for Sequence Labeling Tasks, *Proc. EMNLP 2008*, pp.1070–1079 (2008).
[12] Pan, S.J. and Yang, Q.: A Survey on Transfer Learning, *IEEE Trans. Knowledge and Engineering*, Vol.22, No.10, pp.1345–1359 (2010).

川上 尚慶

2013 年岡山大学工学部情報工学科卒業．2015 年同大学大学院自然科学研究科電子情報システム工学専攻博士前期課程修了．同年三菱電機インフォメーションシステムズ株式会社入社．大学院在学中に学術論文からの書誌情報抽出に関する研究に従事．



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）．東京都立大学大学院工学研究科助手，岡山大学大学院自然科学研究科助教授，准教授を経て，2013 年より岡山大学大学院自然科学研究科教授．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．

安達 淳　（フェロー）

1981 年東京大学大学院工学系研究科博士課程修了．工学博士．東京大学大型計算機センター助手，文部省学術情報センター研究開発部助教授，教授を経て，現在，国立情報学研究所教授，副所長．東京大学大学院情報理工学研究科教授を併任．データベースシステム，情報検索等の開発研究に従事．電子情報通信学会，日本データベース学会，IEEE，ACM 各会員．

（担当編集委員　清水 敏之）