参考文献書誌情報抽出における確信度によるCRF学習データの削減

浪越　大貴† 　太田　学†† 　高須　淳宏††† 　安達　淳†††

† 岡山大学工学部情報系学科　〒 700–8530 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: †pajl4dyr@s.okayama-u.ac.jp, ††ohta@de.cs.okayama-u.ac.jp, †††{takasu, adachi}@nii.ac.jp

あらまし　膨大な文書が格納されている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著者名やタイトルなどの有用な書誌情報が集約されているため，参考文献文字列から書誌情報を自動抽出する研究が行われている．機械学習により書誌情報を高精度に抽出するには，一定量の学習データが必要である．そこで川上らは，Conditional Random Field（CRF）による参考文献書誌情報抽出において, 抽出結果に確信度を定義し，能動学習により学習データが少ない場合の抽出精度を改善する手法を提案した．本稿では，川上らとは異なる確信度と転移学習を利用した学習データの削減を提案し，その削減効果を実験により確かめる．

キーワード　情報抽出，CRF，参考文献文字列，確信度，能動学習，転移学習

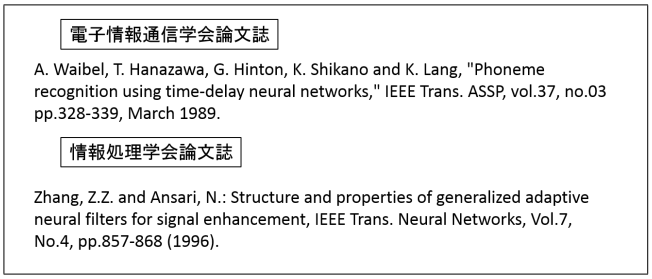
1. はじめに

多数の学術論文を蓄積する電子図書館のサービスを利用する際，検索や文書間リンク等の機能は必須であり，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，これらの書誌情報を人手でデータベースに入力するコストは膨大なため，その作業を可能な限り自動で行う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，関連のあ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．そのため，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，参考文献文字列から書誌情報を自動抽出する研究が行われている [1], [2].

しかし，これまでの研究 [1] から，機械学習を用いて高い抽出精度を得るには，学術雑誌ごとに少なくとも数百件の参考文献文字列が学習データとして必要だった．これは，通常学術雑誌が異なれば，そこに掲載される論文の参考文献文字列の書式が異なるためで，この作成コストは無視できない．そこで川上らは，書誌情報抽出結果に確信度を定義し，学習データを選別する能動学習を用いて，少ない学習データでも抽出精度を改善する手法を提案した [2].

本稿では，川上らとは異なる確信度を利用し能動学習を行う．また，他雑誌の学習データにおいて学習した書誌情報抽出器を用いて，対象雑誌の初期学習データを選出する転移学習を提案する．それにより，書誌情報抽出精度を維持しながら学習データの削減を図る．

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し，3 節で本研究に用いる CRF による自動書誌情報抽出法について説明する．続く 4 節で確信度と確信度を利用した能動学習，転移学習について説明する．5 節で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．


電子情報通信学会論文誌

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, "Phoneme recognition using time-delay neural networks," IEEE Trans. ASSP, vol.37, no.03 pp.328-339, March 1989.

情報処理学会論文誌

Zhang, Z.Z. and Ansari, N.: Structure and properties of generalized adaptive neural filters for signal enhancement, IEEE Trans. Neural Networks, Vol.7, No.4, pp.857-868 (1996).

図 1　学術論文誌による参考文献文字列の書式の違い

2. 関連研究

多数の学術論文を格納する電子図書館において，書誌情報の管理は必須であるため，学術論文から書誌情報を自動抽出する研究が行われている．ルールにより参考文献文字列から書誌情報を抽出する場合，著者名，タイトル，発行年などの書式が異なる論文誌ごとに書誌情報抽出のためのルールを設定する必要がある．例えば図 1 の電子情報通信学会論文誌ではタイトルを二重引用符で囲んでいるが，情報処理学会論文誌では二重引用符で囲んでいない．また，情報処理学会論文誌は発行年を括弧で囲んでいるが，電子情報通信学会論文誌では括弧で囲んでいない．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，管理していく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，学習データを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．例えば，CRF [3] を用いた書誌情報抽出に関する研究に，Peng ら [4]，Councill ら [5] の研究がある．Peng らは，学術論文のタイトルページと参考文献欄から書誌情報を抽出した．タイトルページからの書誌情報抽出では，英語論文 935 件，参考文献欄からの書誌情報抽出では，英語論文 500 件を対象に，著者名や論文誌名など 13 項目の書誌情報を抽出する実験を行った．タイトルペー

ジからの書誌情報抽出の平均 F 値は 0.939，参考文献欄からの書誌情報抽出の平均 F 値は 0.915 であった．一方，Councill らは，CRF に基いて参考文献欄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．ParsCit は，空白文字をデリミタとして，英文の参考文献文字列をトークン列に変換し，そのトークン列に書誌要素ラベルを付与する．Cora データセット [6] 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する実験を行ったところ，その平均 F 値は 0.950 であった．しかし，これらの研究は英語論文のみを抽出対象としており，日英の両言語を含む参考文献文字列から抽出していない．また，実験では書誌情報抽出精度のみを評価し，抽出コストである学習データ量の評価は行っていない．本研究では，日英の両言語を含む参考文献文字列から書誌情報を抽出し，その抽出コストを実験により評価する．

本研究のような書誌情報抽出における学習データの削減に関する研究に，Ohta ら [7] の研究がある．Ohta らは，論文タイトルページからの書誌情報抽出において，学習に有効なデータを効率的に選出する能動サンプリングを用いて，学習データを削減する手法を提案した．Ohta らの書誌情報抽出は，学術論文の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 を用いて短形テキスト領域に対して書誌要素ラベルを付与する．このとき，書誌情報抽出の困難さを表す確信度を定義し，確信度が低いサンプルは学習に有効であるとみなし，優先的に学習データに加え，逐次学習モデルを更新した．実験において学習データ件数を三分の一以下に削減しても書誌情報抽出精度を維持できたと報告されている．しかし，Ohta らは論文タイトルページから書誌情報を抽出したが，本研究ではレイアウト情報を持たない参考文献文字列のみ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，参考文献文字列を図 2 のように，まずトークン列に変換し，そのトークン列から著者名などの主要な書誌情報を抽出する．本研究では [2] と同様にトークン列への変換は人手で行い，トークン列への書誌要素ラベル付与は，CRF を用いて行う．これは，CRF 学習データの削減効果をトークン列への書誌要素ラベル付与の問題において評価するためで，参考文献文字列のトークン列への自動変換については Ohta ら [1] が CRF を利用する方法を提案している．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．本研究では，変換されたトークン列に対し，<RA> や <RT> などの書誌要素ラベル，<DC>（カンマ＋空白）などのデリミタラベルを各トークンに付与する．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，24 種類が定義されている [2]．

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

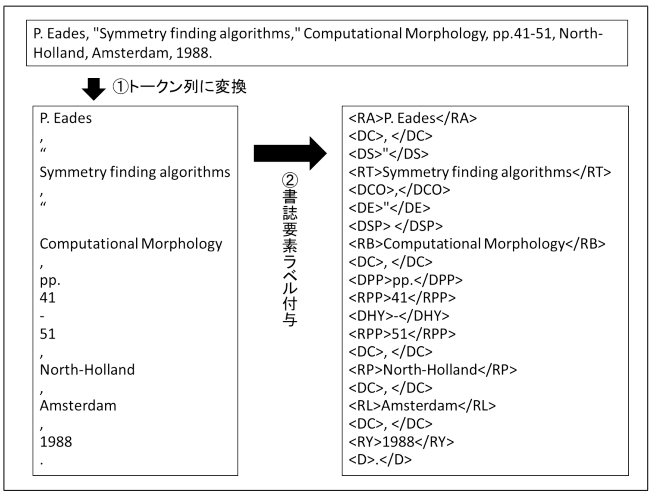


図 2　参考文献書誌情報抽出

表 1　抽出する書誌情報 [2]

書誌要素	書誌要素ラベル
Author	RA
Editor	RE
Translator	RTR
Author Other	RTR
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

CRF [3] の定義を用いて，参考文献文字列の各トークンに書誌要素ラベルを付与する．CRF では，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，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，

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

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．

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

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っ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++[(注1)] [8] を利用して書誌要素ラベル付与を行う．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 48 の Unigram 素性と，一つの Bigram 素性の計 49 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，文字数，トークンの文字種とその割合，トークンの先頭・末尾から四文字目までの文字列，大文字，数字や記号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，例えば“ Academic ”のことで，この文字列を含むトークンは Publisher 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 [(注2)], 論文誌名 [(注3)]，会議名 [(注4)]，出版社名 [(注5)]，地名 [(注6)]，月名の辞書と，学会誌名などの分類困難なものをまとめた辞書の 7 種類を使用する．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ており，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．例えば，<num_word(0)> の場合，“ 大文字のみ ”，“ 先頭のみ大文字 ”，“ 小文字のみ ”の各単語数と総単語数という四つの素性を持つ．さらに，付与されるラベルの連接に関する情報を表す Bigram 素性により，書誌要素の出現順などによる制約を考慮する．

表 2　素性テンプレート [2]

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

4. 確信度による学習データの削減

4.1 確　信　度

参考文献書誌情報抽出において，少ない学習データで高い抽出精度を得るには，効率よく学習を進めていく必要がある．そこで本研究では川上ら [2] と同様に能動学習を行う．さらに，他雑誌で学習した CRF の書誌情報抽出器を用いて初期学習データを選別することで，さらなる精度向上を図る．

能動学習は，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．川上ら [2] は，[9] を参考に書誌情報抽出の困難さを表す尺度として，三つの確信度を定義した．しかし，論文誌によって有効な確信度が異なるため，本研究では [9] を参考に新たな確信度を三つ定義し，参考文献書誌情報抽出の能動学習においてこれら六つの確信度を比較する．まず，4.1.1 項から 4.1.3 項で川上らの確信度を説明する．次に，4.1.4 項から 4.1.6 項で本研究で定義した確信度を説明する．

4.1.1 Normalized Likelihood（NLH）

この確信度は 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，ラベル付与が困難であると見なし，この値を確信度として利用する．入力系列の長さで正規化した以下の式で確信度を定義される．

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

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．

4.1.2 Mininum Probability of Token Assignment（MP）

この確信度は，参考文献文字列中の各トークンに付与されるラベルの周辺確率を利用している．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 を付与されう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 $max_{l\in L} P(Y_i = l|\mathbf{x})$ を i 番目のトークンに着目したラベル付与の確信度と見なし，参考文献文字列中の各トークンに対する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される．

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

4.1.3 Average Token Entropy（ATE）

こ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．エントロピーの値が大きいほど，より多くの書誌

(注1)：http://crfpp.googlecode.com/svn/trunk/doc/index.html
(注2)：http://www.census.gov/genealogy/names/など
(注3)：http://science.thomsonreuters.com など
(注4)：http://www.allconferences.com/など
(注5)：http://www.narosa.com/nbd/PublisherDistributed.asp など
(注6)：http://www.fallingrain.com/world/index.html など

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全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．

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

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される．

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

なお，この確信度は [9] の“ token entropy ”の値を正負逆転させたものと等しい．

4.1.4 Non-normalized Likelihood（NNLH）

この確信度は，川上ららが定義した確信度 NLH において，入力系列の長さによる正規化を行わないものである．入力系列が長い場合，そもそもラベル付与が困難である [9] ため，入力系列の長さによる正規化を行わない．具体的には以下の式で定義する．

$$c_{NNLH}(\mathbf{x}) = \log(P(\mathbf{y}^* \mid \mathbf{x})) \tag{8}$$

4.1.5 Token Margin（TM）

この確信度は各トークンに付与された，上位二つのラベルの周辺確率を用いる．まず，入力系列 $\mathbf{x}$ の各トークンに対する上位二つのラベルの周辺確率の差を求める．この周辺確率の差が小さいほど，ラベル付与が困難であると判断し，トークン列中の最小の差を確信度とする．具体的には以下の式で定義する．

$$c_{TM}(\mathbf{x}) = \min_{1\leqq i\leqq|\mathbf{x}|}\left(P(Y_i = l_{i1}|\mathbf{x}) - P(Y_i = l_{i2}|\mathbf{x})\right) \tag{9}$$

ここで l_{i1}，l_{i2} はそれぞれ i 番目のトークンに対して周辺確率が 1，2 番目に大きいラベルを表す．

4.1.6 Sequence Margin（SM）

この確信度は入力系列 $\mathbf{x}$ に対する条件付き確率 $P(\mathbf{y}|\mathbf{x})$ の上位二つの値を用いる．まず，入力系列に対する条件付き確率の上位二つの値の差を求める．この条件付き確率の差が小さいほど，ラベル付与が困難であると判断し，その値を確信度とする．具体的には以下の式で定義する．

$$c_{SM}(\mathbf{x}) = P(\mathbf{y_1} \mid \mathbf{x}) - P(\mathbf{y_2} \mid \mathbf{x}) \tag{10}$$

ここで $\mathbf{y_1}$，$\mathbf{y_2}$ は，入力系列 $\mathbf{x}$ に対する上位二つの出力ラベル系列を表し，$\mathbf{y_1} = \mathbf{y}^*$ である．なお，先に説明した確信度 TM とこの SM は, [9] における“ margin ”と同じ考えに基づく．

4.2 能 動 学 習

4.1 節で示した確信度を用いて，本研究では，川上らと同様に，以下の手順で能動学習を行う [2]．まず，ラベル未付与の参考文献文字列 S を一定量収集する．次に，少量の参考文献文字列 $S_0 \subset S$ を無作為に選出して，人手でラベルを付与し，これを初期学習データとして CRF M_0 を学習する．その後，以下の手順を繰り返す．

（1） 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 $c.(\mathbf{x})$ をそれぞれ算出する．

（2） 参考文献文字列を算出した各確信度の昇順にランキングする．

（3） 上位 n 件の参考文献文字列 $S_t \subseteqq S - \cup_{i=0}^{t-1} S_i$ を学習データとして選出する．

（4） 選出した参考文献文字列 S_t に人手でラベルを付与する．

（5） ラベル付与した参考文献文字列 $\cup_{i=0}^{t} S_i$ を用いて CRF M_t を学習する．

これは，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く．なお，5 節では $n = 10$ として実験を行う．

4.3 転 移 学 習

4.2 節で説明した能動学習では，初期学習データを選出する際，参考文献文字列を無作為に選出する．しかし，無作為に学習データを選出するため，当然ながら選出される学習データによって書誌情報抽出精度が変動する．そのため例えば，選出される学習データが URL とタイトルのみで構成される参考文献文字列のみであったり，日英両言語を含む論文にも関わらず，選出される学習データが英語の参考文献文字列のみであったりした場合，初期学習データで学習した M_0 の書誌情報抽出精度は低い．また，最初の書誌情報抽出精度が低いと，能動学習を行っても，なかなか精度が向上しないことがある．そこでこの初期学習データを選出する際に，書誌情報の抽出対象とは異なる雑誌 J の学習データで学習した CRF $M^{[J]}$ を用いて，対象雑誌の参考文献文字列の確信度 $c.(\mathbf{x})$ を算出する．その確信度の値に基づき初期学習データを選出することで，無作為に選出する場合よりも書誌情報抽出精度の向上を図る．本研究では，この雑誌 J を転移雑誌と呼ぶ．

この転移学習が有効であれば，電子図書館に新しい学術雑誌が加わった際などに，それまでに得られた他雑誌の書誌情報抽出器を利用して，効率よく新雑誌から書誌情報を抽出できる．

5. 評 価 実 験

5.1 実 験 環 境

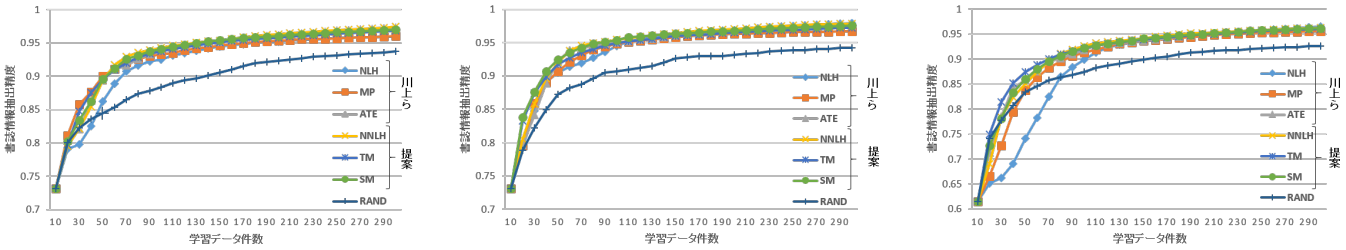
提案手法による学習データ削減効果を検証するため，評価実験を行う．実験データとして，以下の参考文献文字列コーパスを利用する．

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（内，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（内，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（内，和文 1,537 件）

また，表 1 に示した，Author，Title 等の 18 種類の書誌要素ラベルは，評価の際には川上らの研究 [2] に倣い RA，RE，RTR，RAOT を AUTHOR，RT，RBT を TITLE，RW，RC を JOURNAL，RV，RN，RPP を VOLUME，RP を PUBLISHER，RD を DAY，RM を MONTH，RY を YEAR，RL，RURL，ROT を OTHER と再分類する．すなわち書誌要素ラ



(a) IEICE-J　　(b) IEICE-E　　(c) IPSJ

図 3　能動学習における学習データ件数と書誌情報抽出精度

ベルが同じ分類のものは正解判定において区別しない．また，デリミタラベルの種類の誤りは無視する．評価指標として，参考文献文字列を構成する全てのトークンに正しく書誌要素ラベルを付与でき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．また，5 分割交差検定で書誌情報抽出精度を算出する．

5.2　能動学習の効果

4.1 節で説明した確信度を用いた能動学習の効果を図 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五つに分割し，そのうち四つを学習データ，残りの一つをテストデータとする．また，初期学習データを無作為に選出するため，5 分割交差検定を各 10 回行い，書誌情報抽出精度はその平均を取った．ここで，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，その後，確信度が低い参考文献文字列を 10 件ずつ学習データに追加する．また，能動学習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常に無作為に選出した場合を RAND と記しベースラインとする．RAND も能動学習と同様に，5 分割交差検定を 10 回行い，書誌情報抽出精度はその平均である．図 3 の縦軸は書誌情報抽出精度，横軸は学習データ件数，凡例は能動学習に使用した各確信度を表す．

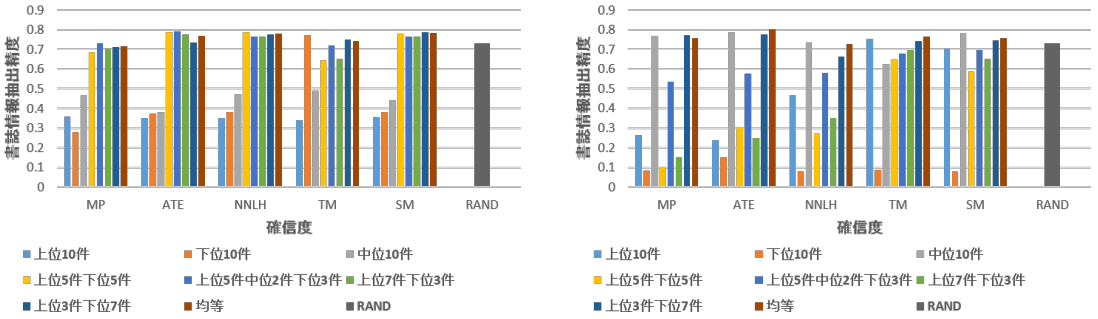
図 3 より，学習データ件数が少ない場合，IEICE-J では MP，TM，IEICE-E では SM，IPSJ では TM の書誌情報抽出精度が高く有効であることがわかる．また，全ての学術論文誌において，NNLH は NLH に比べ高い抽出精度が得られた．また，図 3（c）の IPSJ では，NLH の抽出精度が特に低かった．

5.3　転移学習の効果

転移学習の効果を図 4，図 5，図 6 に示す．評価方法として [10] に示された“ Jumpstart ”を参考に，初期学習データを無作為に選出した場合の書誌情報抽出精度と比較する．なお，転移雑誌の書誌情報抽出器は，その転移雑誌の全参考文献文字列を学習データとして使用した CRF 抽出器である．本実験でも 5 分割交差検定によって書誌情報抽出精度を算出する．また，5.2 節で述べた通り，NLH は NNLH に対し全ての論文誌において抽出精度が劣ったため，NLH は本実験から除いた．図 4，図 5，図 6 の縦軸は書誌情報抽出精度，横軸は初期学習データの選出に使用した確信度，凡例は確信度による初期学習データの選出方法を表す．また，RAND は 5.2 節と同様に，初期学習データ 10 件を無作為に 10 回選出した場合の書誌情報抽出精度の平均である．本研究の転移学習では，転移雑誌 J の学習データで学習した CRF $M^{[J]}$ を用いて対象雑誌の参考文献文字列の確信度を算出し昇順にランキングする．その後，凡例に示した方法で初期学習データを選出する．ここで，凡例の中位 10 件は，確信度の昇順にランキングした学習データの中央の参考文献文字列とその直前の参考文献文字列 5 件，その直後の参考文献文字列 4 件を選ぶことを表す．また，均等は，確信度の昇順にランキングした学習データを 10 等分し，その 10 等分した学習データの先頭からそれぞれ参考文献文字列を 1 件ずつ選ぶ方法である．

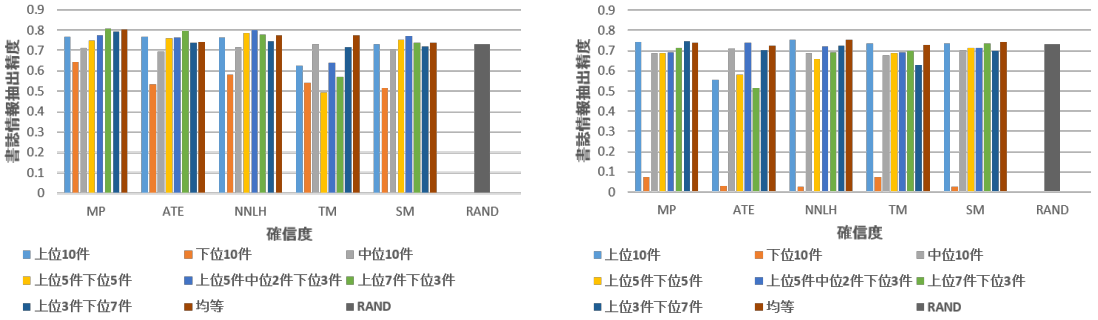
図 4（b）より，IEICE-J では，転移雑誌を IPSJ とすると，確信度が ATE，選出方法が均等の時に最も抽出精度が高く，RAND を約 7 ポイント上回った．また，図 5（a）より，IEICE-E では，転移雑誌を IEICE-J とすると，確信度が MP，選出方法が上位 7 件下位 3 件の時に最も抽出精度が高く，RAND を約 7 ポイント上回った．次に，図 6（a）より，情報処理学会論文誌では，転移雑誌を IEICE-J とすると，確信度が MP，選出方法が均等の時に最も抽出精度が高く，RAND を約 10 ポイント上回った．

ここで同じ学会の論文誌で参考文献文字列の言語が異なる場合を考える．図 4（a）より，電子情報通信学会の英文誌 IEICE-E を転移雑誌として日英両言語を含むその和文誌 IEICE-J の参考文献文字列の確信度を算出した場合，転移雑誌には日本語の参考文献文字列が存在しないため，日本語の参考文献文字列の確信度が低くなり，英語の参考文献文字列の確信度が高くなる．よって上位 10 件，下位 10 件を選出する場合，それぞれ日英片方の言語の参考文献文字列しか含まれないため，これらで学習しても書誌情報抽出精度は高くならない．そのため，参考文献文字列の書式が同じ学会で言語の違いがある場合は，ランキングした対象雑誌の学習データの上位，下位，中位の参考文献文字列をバランスよく，あるいは均等に選出するのが良い．また，図 5（a）より，日英両言語を含む論文誌 IEICE-J を転移雑誌として同学会の英文誌 IEICE-E の参考文献文字列の確信度を算出した場合，既に IEICE-J で英語の参考文献文字列を学習しているため，確信度は全体的に高くなる．そのため，TM 以外の確信度では，上位 10 件，上位 5 件中位 2 件下位 3 件，上位 7 件下位 3 件で参考文献文字列を初期学習データとした場合，無作為に選出した RAND よりも書誌情報抽出精度が高くなる．また，全ての確信度において，均等に参考文献文字列を



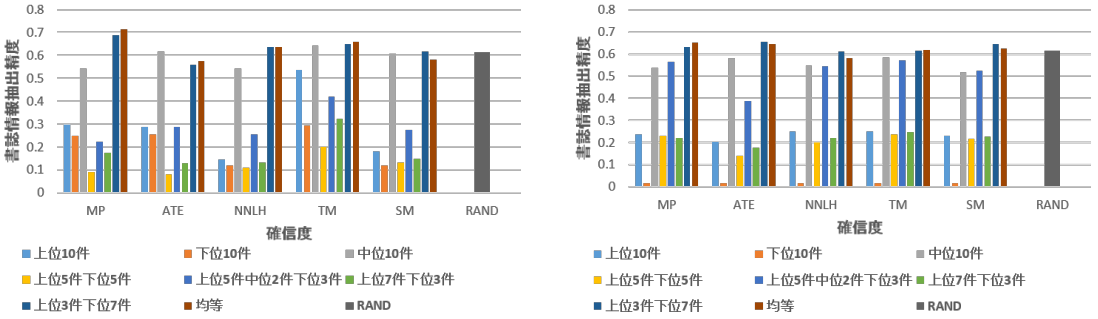
(a) 転移雑誌（IEICE-E）　(b) 転移雑誌（IPSJ）

図 4　転移学習による初期学習データの書誌情報抽出精度 (IEICE-J)



(a) 転移雑誌（IEICE-J）　(b) 転移雑誌（IPSJ）

図 5　転移学習による初期学習データの書誌情報抽出精度 (IEICE-E)



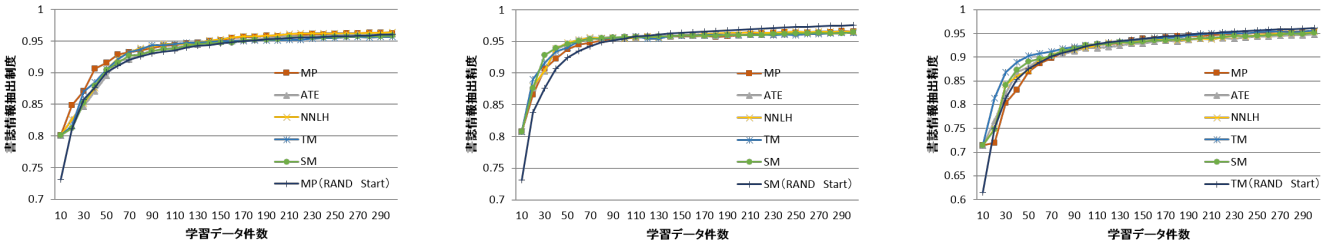
(a) 転移雑誌（IEICE-J）　(b) 転移雑誌（IEICE-E）

図 6　転移学習による初期学習データの書誌情報抽出精度 (IPSJ)

選出するのが良い．特に，確信度 MP の上位 7 件下位 3 件で学習データを選出した場合，最も抽出精度が高くなった．また，均等に学習データを選出した場合でも，ほぼ同等の抽出精度を得られた．よって，書式が似ていて使用言語が異なる場合は参考文献文字列の選出方法を工夫することで，無作為に選出する場合に比べて書誌情報抽出精度が向上することがわかる．

次に，使用言語は日英で同じだが，学会が異なる論文誌を転移させる場合を考える．図 4（b），図 6（a）より，下位 10 件の参考文献文字列を学習データとした場合の抽出精度が非常に低くなっている．書式が異なる論文誌を扱う場合，デリミタも論文誌ごとに異なるため参考文献文字列のデリミタのトークンの確信度が低くなる．しかし，URL のみで構成される参考文献文字列はデリミタを持たず，書式の違いを考慮する必要がない．そのため，URL のみで構成される参考文献文字列は確信度が高くなり，下位 10 件に多く含まれる．そのため抽出精度が非常に低くなる．よって，初期学習データによる書誌情報抽出精度を向上させるためには，上位と下位を適当に混ぜる，あるいは均等に参考文献文字列を初期学習データとして学習するのが良い．

最後に使用される言語も書式も異なる論文誌を転移させる場合を考える．図 5（b），6（b）より，使用される言語も学会も異なる場合，本研究の転移学習はそれほど有効ではない．均等に参考文献文字列を選出して初期学習データとする場合を除くと，確信度によるほぼすべての選出方法で書誌情報抽出精度は無作為に選出した場合に比べ低かった．しかし，このような場合でも，いずれかの確信度を利用して上位 3 件下位 7 件で，あるいは均等に初期学習データを選べば RAND に勝ることが多かった．



(a) IEICE-J　　(b) IEICE-E　　(c) IPSJ

図 7　転移学習で選出した初期学習データを用いた能動学習の書誌情報抽出精度

表 3　転移学習による M_0 の書誌情報抽出精度（対象雑誌 IEICE-E，転移雑誌 IEICE-J）

確信度	全トークン	書誌要素のみ	デリミタのみ
MP	**0.7690**	0.7943	**0.7801**
ATE	0.7683	**0.8199**	0.1434
TM	0.6260	0.7549	0.6268
RAND	0.7309	-	-

表 4　転移学習による M_0 の書誌情報抽出精度（対象雑誌 IPSJ，転移雑誌 IEICE-J）

確信度	全トークン	書誌要素のみ	デリミタのみ
MP	0.2986	**0.4209**	0.2879
ATE	0.2857	0.2674	0.1493
TM	0.5376	0.2409	**0.7381**
RAND	**0.6148**	-	-

5.4　考　　察

5.4.1　確信度の算出方法

確信度 MP，ATE，TM は参考文献文字列の各トークンの周辺確率に基づく確信度である．5.2 節および 5.3 節の実験では，参考文献文字列を構成する全トークンに対し，確信度を算出した．しかし，転移学習において，転移雑誌と対象雑誌が異なる言語や異なる書式であった場合，参考文献文字列の全トークンに対し，確信度を算出するのではなく，書誌要素やデリミタのトークンのみで確信度を算出した方が，より有効ではないかと考えた．そこで，転移学習において，書誌要素のみ，あるいはデリミタのみのトークンの周辺確率に基づく確信度を昇順にランキングし，その上位 10 件を学習して書誌情報抽出精度を算出した．ただし，確信度 NLH，NNLH，SM は CRF の入力系列全体に対する条件付き確率に基づくため省略した．実験結果を表 3，表 4 に示す．なお，表 3，表 4 以外の転移雑誌と対象雑誌の組み合わせでも同様の実験を行ったが効果は確認できなかったため結果は省略する．

表 3 より，確信度 ATE において，CRF が書誌要素ラベルと判定したトークンのみで確信度を算出した場合の書誌情報抽出精度が，全トークンで確信度を算出した場合の書誌情報抽出精度を約 5 ポイント，無作為に選出した場合の書誌情報抽出精度を約 8 ポイント上回った．これは転移雑誌 IEICE-J が日英両言語を含み，対象雑誌 IEICE-E と同じ学会でデリミタの書式が同じなため，デリミタラベルよりも書誌要素ラベルの付与が難しいトークンを持つ参考文献文字列が学習に有効だったためと考えられる．

表 5　転移学習における雑誌，確信度，選出方法の最も有効な組み合わせと書誌情報抽出精度

対象雑誌	転移雑誌	確信度	選出方法	抽出精度
IEICE-J	IPSJ	ATE	均等	0.8011
IEICE-E	IEICE-J	MP	上位 7 件下位 3 件	0.8072
IPSJ	IEICE-J	MP	均等	0.7138

また，表 4 より，確信度 TM において，CRF がデリミタラベルと判定したトークンのみで確信度を算出した場合の書誌情報抽出精度が，全トークンで確信度を算出した場合の書誌情報抽出精度を約 20 ポイント，無作為に選出した場合の書誌情報抽出精度を約 12 ポイント上回った．これは，転移雑誌 IEICE-J は対象雑誌 IPSJ と同言語であるが，学会が異なりデリミタの書式も異なるため，書誌要素ラベルよりもデリミタラベルの付与が難しいトークンを持つ参考文献文字列の方が学習に有効だったためだと考えられる．

5.4.2　Jump Start の優位性評価

転移学習によって初期学習データで書誌情報抽出精度が RAND に比べて向上した場合，その後の学習においても抽出精度の優位を維持できるか検証する実験を行った．5.3 節より，各論文誌における転移雑誌，確信度，選出方法の最も書誌情報抽出精度が高かった組み合わせを表 5 に示す．表 5 に示した最も有効な組み合わせの初期学習データを用いて 5.2 節と同様に能動学習を行い，5 分割交差検定によって書誌情報抽出精度を算出した．実験結果を図 7 に示す．図 7 の縦軸は書誌情報抽出精度，横軸は学習データ件数，凡例は各確信度を表す．また，図 3 に示した各論文誌において最も有効であった確信度を“確信度（RAND Start）”と表記し，ベースラインとする．

図 7（a）より，IEICE-J の場合，確信度 MP 以外では，初期学習データの次の参考文献文字列 10 件を学習した際の書誌情報抽出精度があまり上がらず，抽出精度の優位を維持できていない．しかし，確信度 MP の場合は，学習データ件数が 40 件の書誌情報抽出精度が 0.9 を超えており，学習データが少ない場合に有効であるといえる．一方，図 7（b）の IEICE-E の場合，学習データ件数が少ない場合，どの確信度でも抽出精度の優位を維持できていることがわかる．また，全ての確信度において学習データ件数が 30 件の書誌情報抽出精度が 0.9 を超えており，学習データが少ない場合に有効であるといえる．ま

た，図 7（c）の IPSJ の場合，確信度 TM 以外では，初期学習データの次の参考文献文字列 10 件を学習した際の書誌情報抽出精度が確信度 MP はほぼ上がらず，それ以外の確信度でも RAND とほぼ変わらないため，抽出精度の優位を維持できていない．しかし，確信度 TM の場合，学習データ件数が少ない場合に抽出精度の優位を維持できており，学習データが少ない場合に有効であるといえる．

6. ま と め

本稿では，CRF による参考文献書誌情報抽出において川上らとは異なる確信度を利用し能動学習を行い学習データの削減効果を確かめた．また，他雑誌の学習データにおいて学習した書誌情報抽出器を用いて，対象雑誌の初期学習データを選出する転移学習を提案した．それにより，書誌情報抽出精度を維持しながら学習データの削減を図った．

能動学習の実験の結果，トークンに付与された上位二つの周辺確率の差を用いる確信度（TM）が電子情報通信学会の和文誌と情報処理学会論文誌において有効だった．また，トークン列に対する条件付き確率の上位二つの差を用いる確信度（SM）が電子情報通信学会英文論文誌において有効だった．これらの確信度はいずれも本研究で新たに定義したものである．

転移学習の実験では，電子情報通信学会の和文誌と英文誌をそれぞれ対象雑誌，転移雑誌として初期学習データを選ぶと，無作為に選出した場合の書誌情報抽出精度を上回った．また，転移雑誌と対象雑誌の言語や書式の違いに着目して，参考文献文字列の書誌要素やデリミタのみの確信度を利用すると，全トークンの確信度を利用した場合に比べ，書誌情報抽出精度が向上する雑誌があった．また，転移学習によって選出した初期学習データによって書誌情報抽出精度が向上した場合，その優位をその後の学習においても維持できるかについても実験を行った．電子情報通信学会英文論文誌では，初期の抽出精度の優位を維持することができ，電子情報通信学会和文論文誌と情報処理学会論文誌でも，初期の抽出精度の優位を維持できる確信度があることを示した．

今後の課題として，書式が異なる学術雑誌に対する有効な転移学習の方法の検討が挙げられる．また，転移学習によって Jump Start に成功，すなわち初期学習データの書誌情報抽出精度が向上した際に，その後の学習において抽出精度の優位を維持する方法についても検討したい．

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 15H02789)，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．

文　　献

[1] M. Ohta, D. Arauchi, A. Takasu, and J. Adachi, “Empirical Ecaluation of CRF-Based Bibliography Extraction from Reference Strings”, 11th IAPR International Workshop on Document Analysis Systems (DAS 2014), pp. 287-292, 2014.

[2] 川上尚慶, 太田学, 高須淳宏, 安達淳, “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”, 情報処理学会論文誌データベース, vol. 8, no. 2, pp. 18-29, 2015.

[3] 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.

[4] F. Peng, and A. McCallum, “Accurate Information Extraction from Research Papers Using Conditional Random Fields”, HLT-NAACL 2004, pp. 329-336, 2004.

[5] I.G. Councill, C.L. Giles and M.Y. Kan, “ParsCit: An Open-Source CRF Reference String Parsing Package”, In Proc. of language resource and evaluation conference, 2008.

[6] 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.

[7] M. Ohta, R. Inoue, and A. Takasu, “Empirical Evaluation of Active Sampling for CRF-Based Analysis of Pages”, In Proc. of IEEE IRI 2010, pp. 13-18, 2010.

[8] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, In Proc. of EMNLP 2004, pp. 230-237, 2004.

[9] B. Settles, and M. Craven, “An Analysis of Active Leaning Strategies for Sequence Labeling Tasks”, In Proc. Of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1070-1079, ACL Press, 2008.

[10] M. E. Taylor, and P. Stone, “Transfer learning for reinforcement learning domains: A survey”, J. Mach. Learning Res., vol. 10, no. 1, pp. 1633-1685, 2009.