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

川上　尚慶† 　太田　学† 　高須　淳宏†† 　安達　淳††

† 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, ohta}@de.cs.okayama-u.ac.jp, † †{takasu, adachi}@nii.ac.jp

あらまし　膨大な電子化文書が格納されている電子図書館を快適に利用するには，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，論文題目名などの有用な書誌情報が集約されている．我々は，CRF を用いて参考文献文字列から自動で書誌情報を抽出する方法を提案したが，高い抽出精度を得るには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った．そこで本稿では，能動学習と擬似データを用いて CRF の学習データ量を削減する方法を提案し，実験により評価する．一年分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を対象にした実験では，能動学習により，高い抽出精度を維持したまま，CRF の学習に必要な学習データ量を百件程度に削減できることがわかった．また，擬似学習データの利用により，さらに書誌情報抽出精度を向上させることに成功した．

キーワード　情報抽出，CRF，参考文献文字列，能動学習

Learning cost reduction for CRF-based bibliography extraction from reference strings

Naomichi KAWAKAMI†, Manabu OHTA†, Atsuhiro TAKASU††, and Jun ADACHI††

† 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, ohta}@de.cs.okayama-u.ac.jp, † †{takasu, adachi}@nii.ac.jp

Key words　Information extraction, CRF, Reference string, Active 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．

そこで我々は，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．

しかし，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，そのコストは無視できない．そこで本稿では，書誌情報抽出結果の確信度を利用し，学習データを選別することで必要な学習データの量を削減する能動学習法と、学習データに人工的に自動生成した擬似データを加えることで，利用可能な学習データ量を底上げする方法を提案する．

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 節で本研究で用いる CRF による自動書誌情報抽出法について説明する．学習コストを削減する手法として、4 節で本研究で用いる能動学習について，続く 5 節で学習データに加える擬似データについて説明する．6 節では提案手法の評価実験について述べる．最後に，7 節で本稿をまとめる．

2. 関 連 研 究

参考文献文字列からの書誌情報抽出に，阿辺川ら [1]，Okada ら [2]，Councill ら [3] の研究がある．

阿辺川らは，光学文字認識 (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．阿辺川らは，日本語及び英語で書かれた様々な論文を対象に，Support Vector Machine (SVM) [4] と Hidden Markov Model (HMM) [5] 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て，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，参考文献文字列を和文と英文に分類して，それぞれの言語に対して実験を行った．実験において，和文で 74.8%，英文で 81.6% 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．

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ン列に変換し，SVM と HMM を用いて各トークンに書誌要素ラベルを付与して，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からその全ての書誌情報を過不足なく抽出した．

Councill らは，CRF [6] 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である ParsCit を開発している．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] の参考文献文字列を対象に，著者名やタイトルなど 13 項目の書誌情報抽出実験を行い，13 項目の平均の適合率，再現率がともに 0.957，F 値が 0.950 であったと報告されている．

書誌情報抽出における学習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，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，CRF の学習に有効なデータを効率よく探す方法である．まず，Ohta 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 を用いて認識された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．そして，この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する能動学習を行った．実験において，CRF による自動書誌情報抽出の精度を維持したまま，学習データ量を三分の一以下に削減できたと報告されている．

3. 書誌情報抽出

本研究では，我々の提案した CRF を用いた書誌情報抽出器 [9] を使用する．この書誌情報抽出器では，参考文献欄に挙げられた文献の参考文献文字列を，そこに含まれる書誌情報の特徴から書籍，会議録論文，ジャーナル論文の三種類とその他の計四種類の文献種類に分類する．本稿では，参考文献文字列をこれらの文献種類に人手で分類した．そして，それぞれの文献種類について用意した書誌情報抽出器で書誌情報を抽出する．

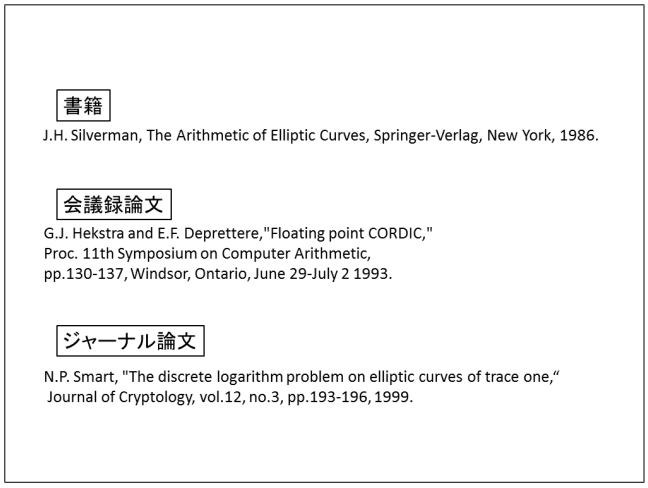

図 1　参考文献文字列の例

3.1 文献の種類

本研究では，参考文献欄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．本研究における文献の種類として，本として出版されている書籍，国内外の学術会議や学術大会で発表された会議録論文，学術論文誌に掲載されたジャーナル論文の 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情報についてみると，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．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．ジャーナル論文には論文誌名が記載されており，巻 (Volume) や号 (Number) が書かれることが多い．このように，参考文献文字列に記述される書誌情報には，文献種類ごとに特徴があることがわかる．

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 [6] の定義を用いて，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．

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

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．

表 1　抽出する書誌情報

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

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各トークンである．一方，ラベル y_i は，著者名，タイトルといった書誌要素である．

3.3　抽出する書誌情報

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，これには所属機関などが含まれる．

3.4　素性テンプレート

本研究では工藤が作成した CRF++(注1) を利用して参考文献文字列に書誌要素ラベルを付与する．CRF++による書誌情報抽出で用いる素性テンプレートを表 2 に示す．これらの素性は，言語的な素性のみで構成されており，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含まない．

表 2 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，例えば“ 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と予想できる．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 3 にまとめる．

(注1)：http://crfpp.sourceforge.net/

表 2　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
	<num_word(0)>	トークン内の単語数
	<num_period(0)>	トークン内のピリオドの数
	<zen_alp(0)>	トークン内の全角アルファベット数の割合
	<zen_digit(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_digit(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
	<hyphen(0)>	トークン中のハイフンの有無
	<apos(0)>	トークン中のアポストロフィーの有無
	<editor>	参考文献文字列における editor に関する記述の有無
	<URL>	参考文献文字列における URL 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 3　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	予想される書誌要素
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

また，辞書は，人名(注2)，論文誌名(注3)，出版社名(注4)，地名(注5)，月名(注6)の辞書を用意した．人名辞書には 97,942 件，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，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．

(注2)：http://www.census.gov/genealogy/names/ など
(注3)：http://science.thomsonreuters.com
(注4)：http://www.narosa.com/nbd/PublisherDistributed.asp など
(注5)：http://www.fallingrain.com/world/index.html など
(注6)：英語の月名とその省略形，仏語の月名

4. 能動学習

3 節で説明した書誌情報抽出では，高い抽出精度を得るために，参考文献の書式が異なる文献ごとに少なくとも数百件の学習データを事前に人手で用意する必要がある．そこで本研究では，その学習データの作成コストの削減を図る．本稿ではまず，CRF の学習に有効な参考文献文字列を優先的に学習に利用する能動学習 [10] による学習データ量の削減方法を提案する．

提案する能動学習では，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．この時，書誌情報抽出の困難さを表す尺度として，確信度を 4.1 節で定義する．この確信度に基づいて学習データを選択することで，ランダムに選択した場合に比べて効率的な CRF の学習を実現する．

4.1 書誌情報抽出の確信度

CRF によって書誌要素ラベルが付与された参考文献文字列の中から，書誌情報抽出の困難な参考文献文字列を見つけるために，太田らの研究 [11] を参考にして，3 種類の確信度を定義する．この書誌情報抽出の確信度とは，各参考文献文字列の書誌情報抽出の確からしさを表したもので，CRF によって参考文献文字列中の各トークンに割り当てられる書誌要素ラベルの周辺確率などに基づいて定める．この確信度が低い参考文献文字列ほど，書誌情報抽出が困難であると判断することができる．以下で本研究で利用する 3 種類の確信度について述べ，4.2 節でそれらの確信度を利用した能動学習について説明する．

4.1.1 Normalized Likelihood

一つ目の確信度は 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み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．

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

ここで，$|\mathbf{x}|$ は入力系列 $\mathbf{x}$ の長さであり，書誌情報抽出対象の参考文献文字列中のトークン数を表す．以後，この確信度を NLH と呼ぶ．

4.1.2 Minimum Probability of Token Assignment

二つ目の確信度は，参考文献文字列中の各トークンに付与された書誌要素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 を付与できる書誌要素ラベルの集合とすると，$P(Y_i = l)$ は書誌要素ラベル $l \in L$ 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 $max_{l \in L} P(Y_i = l)$ は i 番目のトークンに注目したラベル付与の確信度とみなすことができる．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．

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

以後，この確信度を MP と呼ぶ．

4.1.3 Average Token Entropy

三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，すなわち，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，エントロピー $\sum_{l \in L} -P(Y_i = l) \log P(Y_i = l)$ を算出する．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．

$$c_{ATE}(\mathbf{x}) = -\frac{\sum_{x_i \in |\mathbf{x}|} \sum_{l \in L} -P(Y_i = l) \log P(Y_i = l)}{|\mathbf{x}|} \tag{6}$$

ここで，$|\mathbf{x}|$ は入力系列 $\mathbf{x}$ の長さであり，書誌情報抽出対象の参考文献文字列中のトークン数を表す．以後，この確信度を ATE と呼ぶ．

4.2 確信度を用いた能動学習

学習データの選択方法について説明する．まず，書誌要素ラベルの付与されていない参考文献文字列集合を S とする．$(t-1)$ 回目の学習において，人手により正しい書誌要素ラベルを付与された参考文献文字列 $S_{t-1} \subset S$ で CRF を学習し，CRF 抽出器 M_{t-1} を作る．続く t 回目の学習では，$S - S_{t-1}$ の中から n 個の参考文献文字列集合 C_t を以下に述べる方法で選出する．この C_t に人手で書誌要素ラベルを付与し，学習データに追加する．つまり，t 回目の学習によって得られる CRF 抽出器 M_t は，$S_t := S_{t-1} \cup C_t$ の参考文献文字列を学習に利用する．

C_t の選出方法は比較のため 2 種類定めた．

一つ目は $S - S_{t-1}$ の中から n 個の参考文献文字列を無作為に選出する．これは，t 回目の学習において能動学習の有無による抽出精度の比較を行うために定義した．以後この方法を RAND と呼ぶ．

二つ目は $S - S_{t-1}$ に含まれる参考文献文字列に対して，式 (4)，(5)，(6) で定義した確信度 $c(\mathbf{x})$ を算出し，それに従って昇順にランキングした参考文献文字列のランク上位 n 件を選出する．このとき，ランク上位の論文ほど確信度が低いため書誌情報抽出が困難であるとみなし，優先的に学習データに取り込む．この方法は，採用する確信度に応じて，それぞれ MP，NLH，ATE と呼ぶ．

5. 擬似学習データ

本節では，擬似学習データを用いて利用可能な学習データの量を増やすことで抽出精度の向上を図る方法を提案する．本稿では，擬似データの生成方法として書誌情報抽出結果の確信度が高い参考文献文字列を用いる方法と，確信度が低い参考文献文字列と書誌情報データベースを用いる方法の 2 種類を提案する．

5.1 確信度の高い書誌情報抽出結果を用いた擬似学習データ

一つ目の擬似学習データは，ある時点の学習モデルによって自動的に書誌情報を抽出した参考文献文字列の中で，確信度が高いものを利用する．これを次回の学習データに追加する．書誌情報の抽出結果をそのまま用いるので，擬似学習データの作成コストは無視できるが，若干誤りを含む可能性がある．

具体的には，$(t-1)$ 回目に $S_{t-1} \subset S$ で学習して CRF 抽出器 M_{t-1} を作る．M_{t-1} の出力した書誌要素ラベル系列 $\mathbf{y}^*$ の確信度が高ければ，その参考文献文字列の書誌情報は正しく抽出できたとみなす．そこで，続く t 回目の学習では，$S - S_{t-1}$ に含まれる参考文献文字列に対して，式 (4)，(5)，(6) で定義した確信度 $c(\mathbf{x})$ を算出し，それに従って降順にランキングした参考文献文字列のランク上位 l 件を選出する．この l 件の参考文献文字列に M_{t-1} が書誌要素ラベルを付与したものを t 回目の学習における擬似学習データ C'_t とする．そして，人手で書誌要素ラベルを付与した学習データ C_t と擬似学習データ C'_t を学習データに追加する．つまり，t 回目の学習によって得られる CRF 抽出器 M_t は，$S_t := S_{t-1} \cup C_t \cup C'_t$ の参考文献文字列を学習に利用する．ただし，C'_t のうちいくつかの参考文献文字列は，M_{t-1} によって誤った書誌要素ラベルを付与されている可能性がある．よって，この擬似学習データは次回の学習データへは持ち越さない．したがって，t 回目の学習データ件数は以下の式で表せる．

$$\text{学習データ件数} = n_0 + nt + l \tag{7}$$

ただし，n_0 は最初の学習データ件数，n は 1 回の能動学習において人手で書誌情報ラベルを付与する参考文献文字列の件数，t は繰り返し回数，l は擬似学習データとする確信度上位の件数である．

5.2 書誌情報データベースを用いた擬似学習データ

二つ目の擬似学習データは，能動学習によって人手でラベル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．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．

能動学習では $(t-1)$ 回目に $S_{t-1} \subset S$ で学習して CRF 抽出器 M_{t-1} を作り，M_{t-1} の出力した書誌要素ラベル系列 $\mathbf{y}^*$ の確信度を利用して学習データに加える参考文献文字列を選出する．つまり，続く t 回目の学習では，$S - S_{t-1}$ に含まれる参考文献文字列に対して，式 (4)，(5)，(6) で定義した確信度 $c(\mathbf{x})$ を算出し，それに従って昇順にランキングした参考文献文字列のランク上位 n 件に人手で書誌要素ラベルを付与し学習データ C_t とする．この n 件の参考文献文字列の各トークンを，書誌情報データベースの同じ書誌要素の文字列と置換し，擬似データ C'_t を作る．従って，t 回目の学習によって得られる CRF 抽出器 M_t は，$S_t := S_{t-1} \cup C_t \cup C'_t$ の参考文献文字列を学習に利用する．ただし，C'_t の参考文献文字列に付与されている書誌要素ラベルは正しいとみなして，次回の学習データへと持ち越す．1 件の参考文献文字列につき m 件の擬似データを作成した場合，t 回目の学習データ件数は以下の式で表せる．

$$\text{学習データ件数} = n_0 + nt + ntm \tag{8}$$

ただし，n_0 は最初の学習データ件数，n は 1 回の能動学習において人手で書誌情報ラベルを付与する参考文献文字列の件数，t は繰り返し回数，m は能動学習で利用する参考文献文字列 1 件につき生成する擬似学習データの件数である．

6. 評価実験

提案する学習データの削減手法の有効性を調べるために評価実験を行った．実験に利用するデータとして，2000 年の電子情報通信学会英文論文誌 Vol.E83-A No.1 から No.12 の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 件のトークンに分割された参考文献文字列を使用した．この参考文献文字列を文献種類別の内訳は，書籍が 624 件，会議録論文が 1,247 件，ジャーナル論文が 2,231 件，その他が 395 件である．本実験における書誌情報抽出精度は，表 1 に示した参考文献文字列中の全ての書誌情報が過不足なく正確に抽出された参考文献文字列の割合とする．

6.1 能動学習の有効性評価

本節では，4 節で説明した能動学習により，書誌情報抽出精度を維持したまま、必要な学習データ量がどの程度削減できるかを実験により検証する．5 分割交差検定を行うため，各文献種類の参考文献文字列データを五つに分割し，そのうち四つを学習用データ，残りの一つをテストデータとする．まず，学習用データの中から 10 件の参考文献文字列を無作為に選出して最初の学習データとし，CRF を学習する．次に，学習した CRF を用いてその 10 件を除いた未使用の学習用データの参考文献文字列の書誌要素を推定して確信度を算出し，参考文献文字列を確信度に基づいて昇順にランキングする．そして，その上位 10 件の参考文献文字列を新たな学習データに選出し，先ほど利用した 10 件の学習データに追加して CRF を再学習する．このようにして学習データを 10 件ずつ増やし，最初に用意した学習用データを全て使い切るまで実験を繰り返して書誌情報抽出精度を算出する．なお，本実験では 5 分割交差検定を行ったので，実験結果は全て 5 回の平均値である．

4 節で説明した各確信度を用いた能動学習による学習データ件数と書誌情報抽出精度の関係を文献種類ごとにまとめたものを図 2，3，4，5 に示す．

これらの図より，いずれの文献種類においても安定した性能を示した MP を RAND と比較する．書誌情報抽出精度が収束するとみなせる点の学習データ件数は，書籍では RAND が 300 件であるのに対して MP が 100 件，会議録論文では RAND が 740 件であるのに対して MP が 130 件，ジャーナル論文では RAND が 1,500 件であるのに対して MP が 120 件であった．つまり，書誌情報抽出精度を維持したまま，書籍では学習データ量を約三分の一，会議録論文では学習データ量を約五分の一，ジャーナル論文では学習データ量を約十分の一に削減できることがわかる．この結果より，MP を利用してデータを選別すれば，各

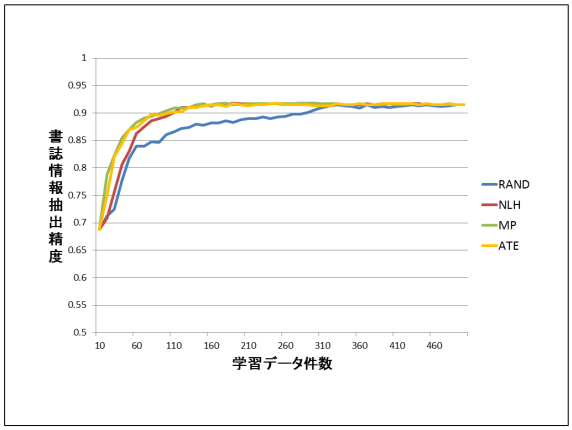

図 2　能動学習における学習データ数と書誌情報抽出精度 (書籍)

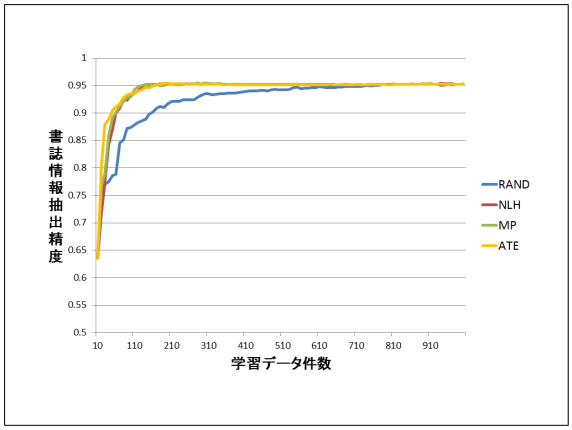

図 3　能動学習における学習データ数と書誌情報抽出精度 (会議録論文)

文献種類ごとに百件程度の学習データを用意すれば，高い精度で書誌情報を抽出できることがわかる．一方，図 5 の「その他」では RAND と三つの確信度の間に明確な差が見られなかった．この原因として,「その他」に含まれる参考文献文字列は書籍，会議録論文，ジャーナル論文のいずれにも分類されなかった参考文献文字列の集合であるため，類似点が少ない雑多な集合であることが考えられる.「その他」では，ある学習データを用いて学習しても，残りのデータとの類似点がもともと少ないため，うまく学習が行えず，他の三つの文献種類に比べて書誌情報抽出精度が悪い．従って，新たな文献種類を定義するなど妥当な文献種類の検討も必要である．

また，確信度を比較すると，NLH，MP，ATE のいずれの確信度を用いて能動学習を行った場合においても，RAND と比較して学習データ量を大きく削減できることがわかった．

6.2　擬似学習データ有効性評価

本節では，5 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検証する．本実験では，6.1 節と同様の条件で実験を行うが，学習データを追加する際，5 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．なお，6.1 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．

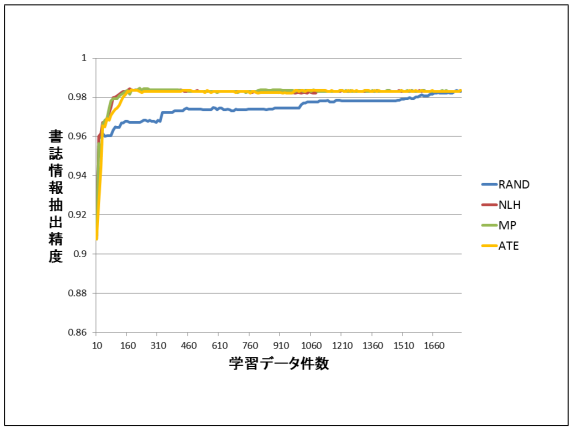

図 4　能動学習における学習データ数と書誌情報抽出精度 (ジャーナル論文)

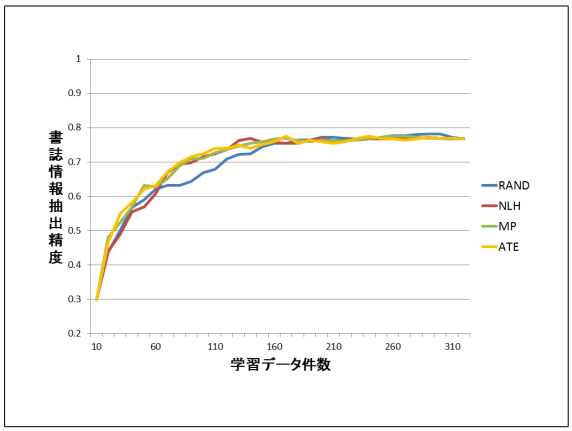

図 5　能動学習における学習データ数と書誌情報抽出精度 (その他)

6.2.1　高確信度の擬似学習データ

追加する擬似学習データを，確信度の高いものから 0 件，10 件，30 件，50 件，100 件とした場合の実験を行い，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 6.1 節の書誌情報抽出精度と同一である．

擬似学習データが与える効果について，能動学習で効果のあった確信度 MP を使用した場合の結果を図 6，7，8，9 に示す．ただし，これらの図中の横軸の学習データ件数は人手で書誌要素ラベルを付与した学習データ件数を表し，凡例は追加した擬似学習データの件数を表す．

図 6，7，8，9 より，高確信度の書誌情報抽出結果をそのまま利用した擬似学習データはいずれの文献種類でも書誌情報抽出精度の向上に効果が見られなかった．また，MP 以外の確信度を用いた実験でも同様の結果となった．学習データ件数が少ないとき，擬似学習データを加えるとかえって書誌情報抽出精度が下がるのは，擬似学習データに誤りが含まれるためと考えられる．また，ある程度学習データ件数が増えても書誌情報抽出精度に向上が見られなかった原因として，確信度の高い擬似学習データでは，書誌情報抽出の難しい参考文献文字列には対応できなかったことが考えられる．

6.2.2　書誌情報データベースを利用した低確信度の擬似学習データ

本実験で使用する書誌情報データベースは，2000 年の電子情

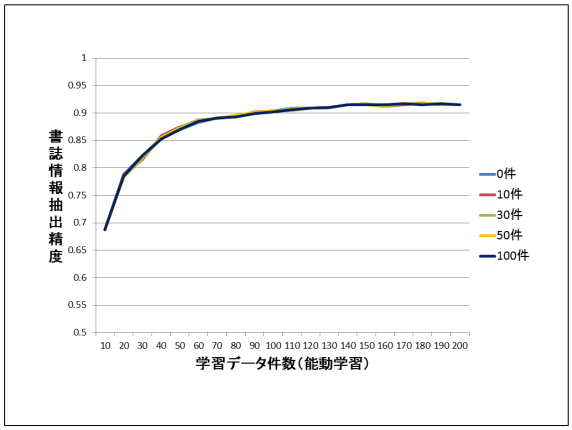

図 6　擬似学習データと書誌情報抽出精度 (書籍・MP)

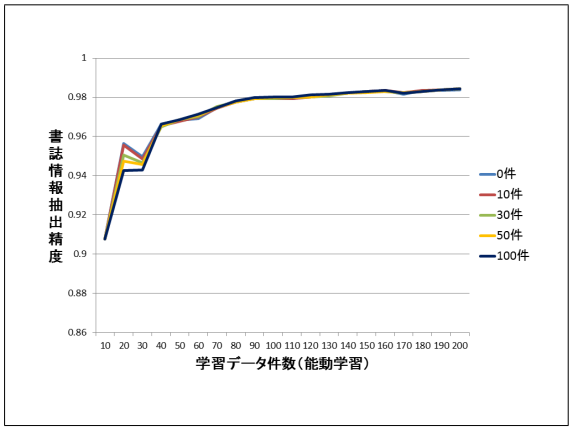

図 8　擬似学習データと書誌情報抽出精度 (ジャーナル論文・MP)

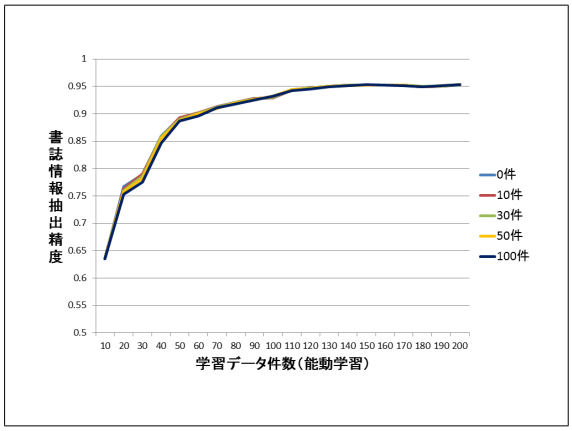

図 7　擬似学習データと書誌情報抽出精度 (会議録論文・MP)

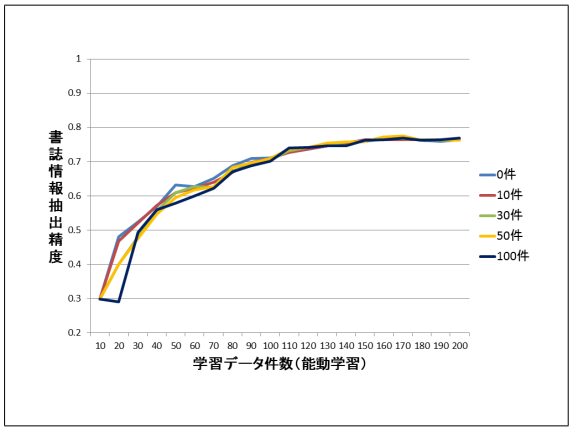

図 9　擬似学習データと書誌情報抽出精度 (その他・MP)

報通信学会和文論文誌 Vol.J83-DII の No.1 から No.12 の一年分に相当する論文の参考文献コーパスに記述されている参考文献文字列の内，英文のみで記述されているもの 2,594 件から人手で作成した．

生成する擬似学習データを，能動学習のため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，10 件とした場合の実験を行い，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 6.1 節の書誌情報抽出精度と同一である．

擬似学習データが与える効果について，確信度に MP を使用した場合の結果を図 10，11，12，13 に示す．ただし，これらの図中の横軸の学習データ件数は人手で書誌要素ラベルを付与した学習データ件数を表し，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である．

図 10，11，12，13 より，書誌情報データベースを利用した擬似学習データは全文献種類で書誌情報抽出精度の向上に効果があった．能動学習による学習データ件数が 30 件のとき，その参考文献文字列 1 件につき 5 件の擬似データを生成した場合，書誌情報抽出精度は，書籍では 79.2% から 84.3%，会議録論文では 78.3% から 83.0%，ジャーナル論文では 95.0% から 96.8%，その他では 49.9% から 55.9% に改善された．また，NLH，ATE を用いた実験でも同様の結果となった．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が，人手でラベル付与した学習データ件数が少ないとき悪くなる場合がある．この擬似学習データは，ある時点の学習モデルで書誌情報抽出が困難と思われる参考文献文字列の書誌情報を置換して生成する．この参考文献文字列は，それぞれの文献種類の一般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．また，この擬似学習データ生成法では書誌情報の出現順は変化しない．したがって，生成する擬似学習データが多いとき，特異なパターンをもつデータを大量に学習してしまい，抽出精度の悪くなる可能性がある．

7. ま と め

本稿では，能動学習と擬似学習データの利用によって，CRF による参考文献書誌情報抽出に必要な学習コストを削減する手法を提案した．能動学習では，参考文献文字列の書誌情報抽出結果に確信度を定義して，確信度の低い参考文献文字列ほど書誌情報抽出が困難であると判断し，そのようなサンプルを優先的に学習することで学習データ量を削減する．また，擬似学習データの利用では，同じ確信度を用いて，2 種類の擬似学習データを検討した．一つ目は，確信度の高い参考文献文字列はおおむね正しい書誌情報抽出ができているとみなし，そのまま擬似学習データとした．二つ目は，能動学習において人手でラベル付与した参考文献文字列を基に，各トークンの文字列を書誌情報データベースの文字列と置換して擬似学習データを生成した．

実験では，電子情報通信学会英文論文誌の参考文献文字列に

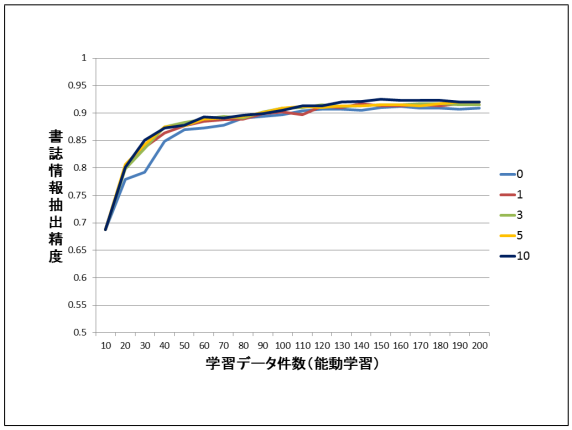

図 10　擬似学習データと書誌情報抽出精度 (書籍・MP)

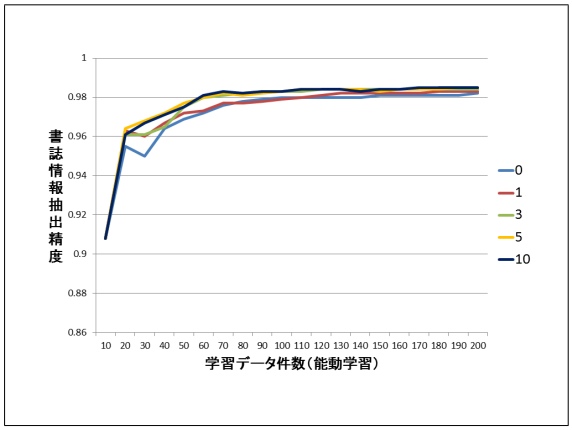

図 12　擬似学習データと書誌情報抽出精度 (ジャーナル論文・MP)

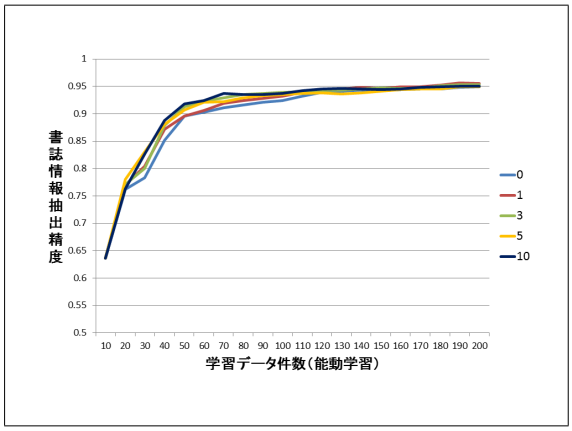

図 11　擬似学習データと書誌情報抽出精度 (会議録論文・MP)

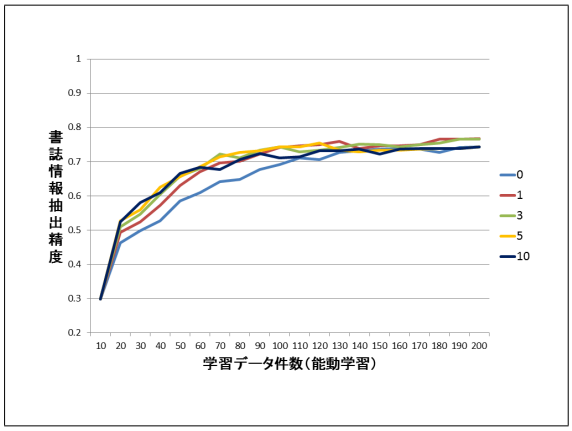

図 13　擬似学習データと書誌情報抽出精度 (その他・MP)

対し，確信度による能動学習と擬似学習データの有効性を検証した．その結果，高い精度を維持したまま，能動学習により CRF の学習に必要な学習データ量を百件前後まで削減できることを示した．一方，擬似学習データは，高確信度の書誌情報抽出結果を利用したものでは書誌情報抽出精度の向上に効果が見られなかったが，書誌情報データベースを利用したものでは書誌情報抽出精度が最大 6 ポイント向上した．

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] 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.

[2] 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.

[3] 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.

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

[5] 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, pp. 37-42, 1999.

[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] 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.

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

[9] 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，”DEIM Forum 2013，C10-3，2013.

[10] B.Settles, M.Craven, “An Analysis of Active Learning Strategies for Sequence Labeling Tasks,” Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1070-1079, 2008.

[11] 太田学，井上諒平，高須淳宏，“ CRF による学術論文タイトルページからの書誌要素抽出における誤り検出，”日本データベース学会論文誌，Vol. 11，No. 2，pp. 37-42，2012.