DBSJ Journal, Vol.11, No.2
October 2012

CRFによる学術論文タイトルページからの書誌情報抽出における誤り検出

Error Detection of CRF-Based Bibliography Extraction from Title Pages of Research Papers

太田 学♥　井上 諒平♦
高須 淳宏♠

Manabu OHTA　Ryohei INOUE
Atsuhiro TAKASU

電子図書館の利便性向上には，そこに蓄積する文書の様々なメタデータが必要で，特に詳細かつ正確な書誌情報は不可欠である．我々は，Conditional Random Field(CRF)を用いて学術論文文書画像から書誌情報を自動抽出する手法を提案し，一定の抽出精度を確認した．しかし，自動抽出では精度に限界があるため，現実的なコストで人がその抽出誤りに対処しなければならない．本稿では CRF による書誌情報抽出結果に確信度を定義し，それに基づいて抽出誤りを含む論文を自動検出する手法を提案する．その結果，実験で利用した二つの論文誌において CRF による抽出精度がそれぞれ約 94%，96%となり，このとき全体の約 1 割の論文を確信度に基づき検出して人手で処理すれば，最終的に 99%の精度が実現できることを示した．

Digital libraries need various metadata of the stored documents for providing usability. Especially, accurate bibliographic information is indispensable. We proposed an automatic bibliography extraction method from research papers scanned with OCR markup, using the formalism of conditional random fields (CRFs), which achieved good extraction accuracies for some Japanese academic journals. However, there needs to be some human intervention to correct extraction errors because such errors are inevitable. Therefore, this paper proposes confidence measures for the CRF-based bibliography extraction to detect papers with such extraction errors. Our experiments revealed that using the proposed confidence measures improved the accuracy to 99% by manually checking the detected papers, about a tenth of the total papers, when we applied our CRF-based bibliography extraction to two journals used for the experiments and their automatic extraction accuracies were about 94% and 96%, respectively.

♥ 正会員　岡山大学大学院自然科学研究科
ohta@de.cs.okayama-u.ac.jp
♦ 非会員　株式会社四国日立システムズ
innowait@gmail.com
♠ 正会員　国立情報学研究所コンテンツ科学研究系
takasu@nii.ac.jp

1. はじめに

電子書籍閲覧端末の普及が急速に進んでいるが，その背景には社会の隅々まで文書の電子化が浸透したことが挙げられる．電子図書館のみならず，大学等でも機関リポジトリの構築が進むなど，インターネットアクセス可能な情報アーカイブが組織的に整備されるようになった．電子文書へ効率よくアクセスするには，書誌情報等のメタデータの整備が不可欠であるが，良質のメタ情報が付与された電子文書の作成技術はまだ成熟していない．ゆえに書誌情報を含む様々なメタ情報を文書から自動抽出する技術は，知的資産としての情報アーカイブ実現のための核となる技術といえる．

学術論文の場合，表題，著者名，雑誌名などの要素から構成される書誌情報が最も重要なメタ情報といえる．我々は，論文文書画像のページのレイアウトを解析し，Conditional Random Field(CRF)を利用して，タイトルページに記載されている書誌情報を自動抽出する手法を提案した[13]．また複数の学術論文誌を対象とした実験により，各論文誌に対して93%～98%の抽出精度を確認している[14]．

国内では，阿辺川らが日英の様々な雑誌論文を対象としたサポートベクトルマシン(SVM)による書誌情報抽出法を提案している[1]．彼らは，論文タイトルページから12種類の書誌要素を69%の精度で抽出し，参考文献欄からは和文，英文でそれぞれ75%，82%の精度で6種類の書誌要素を抽出した．また藤尾らは，正準判別分析とレイアウトのDPマッチングを用いた書誌情報抽出法を提案し，テキスト情報を持つPDFデータを対象に書誌情報の抽出実験を行い，各論文誌に対して85%～94%の抽出精度を報告している[3]．

一方海外では，PengらがCRFによる学術論文のPDFファイルからの書誌情報抽出を提案した[10]．彼らは，隠れマルコフモデル(HMM)やSVMと比較して，CRFの優位性を実験により示すとともに，論文タイトルページと参考文献欄の双方から，それぞれ事前に定めた13種類の書誌要素を抽出し，73%，77%の抽出精度を報告している．近年では，CRFにより参考文献文字列の書誌情報を解析するオープンソースのツールであるParsCit[2]がCouncillらによって公開され，参考文献欄の解析ではPengらの精度を上回ったことが報告されている．また論文には，表題や著者名等が冒頭にあり，梗概，本文とつづき，結論と参考文献で終わるというような論理構造がある．Luongらはこの論理構造を，参考文献情報やタイトルページの書誌情報も含めて，論文PDFファイルからまるごと復元(抽出)する方法を提案している[7]．

しかしながら，書誌情報抽出システムに求められる抽出精度が非常に高い場合は，実用上自動抽出後に人手での検査・修正処理が必要となる．電子図書館には通常膨大な学術論文誌が収録されておりそのコストは無視できないため，この人手の処理にかかるコストはなるべく低減しなければならない．そこで本稿では，CRFによる書誌情報抽出の結果に確信度を定義し，確信度が低い論文を自動抽出が困難な論文として検出する方法を提案する[1]．本提案は，従来の書誌情報抽出研究でほとんど顧みられなかった誤りへの対処方法を示し，人手による介入が可能な実用的なメタ情報編集環境を実現する基礎となるものである．

[1] 本稿は[4]，[9]を拡張し，誤り検出と後処理コストの関係を示す実験と分析を追加したものである．

2. CRF による書誌要素抽出

2.1 CRF

CRFは，Lafferty ら[6]によって提案された観測系列のラベル付けに統計的な枠組みを与える識別モデルであり，形態素解析[5]や固有表現抽出などにおいて広く利用されている．CRFはラベル付与問題において，事実上利用可能な学習データが十分でない場合においても，しばしばHMMのような生成モデルよりも良い結果を示しており，広範な分野で利用実績がある[12]，[15]．

本研究では標準的なチェーンモデルのCRFを用いて，論文タイトルページの各テキスト行に対して書誌要素ラベルを付与することで書誌要素を抽出する．すなわち，入力トークン系列 $\mathbf{x}=x_1,...,x_n$ 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．

$$p(\mathbf{y}|\mathbf{x})=\frac{1}{Z(\mathbf{x})}\exp(\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1},y_i,\mathbf{x})) \quad (1)$$

ここで $Z(\mathbf{x})$ は，全てのラベル系列を考慮したときに確率の和を1とするための正規化項，$f_k(y_{i-1},y_i,\mathbf{x})$ は $i-1$ 番目の出力ラベル y_{i-1} と i 番目の出力ラベル y_i 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．そして，入力系列 $\mathbf{x}$ に対する最適な出力ラベル系列 $\mathbf{y}^*$ が次式で与えられる．

$$\mathbf{y}^*=\underset{\mathbf{y}}{\operatorname{argmax}}p(\mathbf{y}|\mathbf{x}) \quad (2)$$

ここでラベル付与の対象である入力トークン x_i は，具体的には論文タイトルページの各テキスト行に対応する．一方出力ラベル y_i は，表題，著者名，梗概といった書誌要素に対応する．

CRFでは相関のある特徴を素性として柔軟に扱えるため，論文タイトルページのレイアウト情報を視覚的素性，文字情報を言語的素性として，有用な情報を全て有効に利用して書誌要素抽出を行うことができる．

また，実際の書誌要素抽出にはCRFのオープンソースの実装であるCRF++[16]を利用した．

2.2 書誌要素抽出

本研究では論文タイトルページ内の書誌要素を自動抽出する[13]．具体的には，CRFを用いて各テキスト行に表題，著者名，梗概といった書誌要素ラベルを付与する．前提として，1行は必ず一つの書誌要素だけからなり複数の書誌要素が混在することはないとし，また梗概のように複数行で一つの書誌要素を構成する場合もある．

入力データは，学術論文の文書画像をOCRで認識して得られるXMLファイルである．我々がOCRベンダと開発したOCRは，文書画像のレイアウトを解析してそのテキスト領域を文字認識する．よってこのXMLファイルの中には，認識された文字情報に加えて，文字，語，行，テキストブロックの文書中の配置を示すレイアウト情報が含まれる．具体的にはそれらを囲む矩形領域のx，y座標と幅，高さが属性として与えられている．書誌要素抽出では，その中の各行がそれぞれどの書誌要素に該当するか推定して書誌要素ラベルを付与する．

本稿で実験に用いる三つの学術論文誌の論文から抽出する書誌要素を表1にまとめる．ここで情報処理学会論文誌(IPSJ)，電子情報通信学会英文論文誌(IEICE-E)，電子情報通信学会和文論文誌(IEICE-J)である．また「その他」は他のどの書誌要素にも該当しない行を表す．表1に示すように，論文誌毎にタイトルページに現れる書誌要素は多少異なる．

CRFで書誌要素抽出を行うため，CRF++の素性テンプレートとして考慮した文書の特徴を表2にまとめる．まず文書のレイアウト情報を表す視覚的素性として，各テキスト行のx座標，y座標，幅，高さに加えて，前行との隔たり，各行のベースとなる文字の大きさ，行を構成する文字数を採用した．一方言語的素性として，テキストの各行に含まれる英数字や漢字などの文字種別の割合，特定の書誌要素との関連が強い特徴的な文字列の有無を用いる．さらに付与する書誌要素ラベルの連接に関する情報(bigram素性)も加えて，書誌要素の出現順に関する制約を考慮した．

表 1　抽出する書誌要素
Table 1　Bibliographic Elements for Extraction

書誌要素	IPSJ	IEICE-E	IEICE-J
論文種別	－	○	－
和文表題	○	－	○
和文著者名	○	－	○
和文梗概	○	－	○
和文キーワード	－	－	○
英文表題	○	○	○
英文著者名	○	○	○
英文梗概	○	○	－
その他	○	○	○

表 2　書誌要素抽出に用いる素性テンプレート
Table 2　Feature Template for Extracting Bibliographic Elements

種類	素性	内容
unigram	<i(0)>	行の ID
	<x(0)>	行の x 座標
	<y(0)>	行の y 座標
	<w(0)>	行の幅
	<h(0)>	行の高さ
	<g(0)>	前の行との隔たり
	<cw(0)>	行内文字の幅の中央値
	<ch(0)>	行内文字の高さの中央値
	<#c(0)>	行の文字数
	<ec(0)>	行内の英数字の割合
	<kc(0)>	行内の漢字の割合
	<jc(0)>	行内の平仮名・片仮名の割合
	<s(0)>	行内の記号の割合
	<kw(0)>	行内の特徴的な文字列の有無
bigram	<y(-1),y(0)>	ラベル遷移

3. 書誌要素抽出誤りの検出

3.1 書誌要素抽出の確信度

CRF によって書誌要素ラベルを付与した論文の中から誤りを含む論文を検出するために，本稿では書誌要素抽出の確信度を三つ提案する[4]，[9]．これらの確信度を書誌要素抽出の確からしさの指標とするため，実際に書誌要素抽出の正誤と高い相関がなければならない．さらに，提案した確信度による抽出誤りの検出方法を示し，実験により確信度と書誌要素抽出精度との関係を示す．

以下では提案する3種類の確信度について述べ，その後それらの確信度を利用した抽出誤りの検出について説明する．また最初に説明する二つの確信度は，我々が学習データの能動サンプリングのために提案して，学習データ量の削減に有効であることを確認したものである[8]．また三つ目の確信度は，各行の書誌要素ラベルを推定した結果のエントロピーに基づいている[11]．

3.1.1 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)に示す入力トークン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 $p(\mathbf{y}^* \mid \mathbf{x})$ の値が大きければそのサンプルに対するラベル付けは確信度が高いとみなすことができ，反対にその値が小さければそのサンプルに対するラベル付けは困難であると判断できる．そこでこの $p(\mathbf{y}^* \mid \mathbf{x})$ 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次の式で確信度を定義する．

$$c_{NLH}(\mathbf{x}) := \frac{\log(p(\mathbf{y}^* \mid \mathbf{x}))}{|\mathbf{x}|} \qquad (3)$$

ここで $|\mathbf{x}|$ は入力トークン系列 $\mathbf{x}$ の長さであり，書誌要素抽出対象の論文タイトルページの行数を表す．以後この確信度を NLH と呼ぶ．

3.1.2 Minimum Probability of Token Assignment

NLH は系列 $\mathbf{x}$ を成すトークン全体への書誌要素ラベル付け結果に基づく確信度といえる．それに対して二つ目の確信度は，論文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に基づいて定める．Y_i を入力系列 $\mathbf{x}$ の i 番目のトークン x_i に付与されるラベルを表す確率変数とする．L を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 i 番目のトークン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けの確信度とみなすことができる．そしてトークン系列中の各トークンに対するこのラベル付け確信度の中で，最も小さい値をその入力系列に対する書誌要素抽出の確信度とする．具体的には以下の式で定義する．

$$c_{MP}(\mathbf{x}) := \min_{i \le |\mathbf{x}|} \max_{l \in L} p(Y_i = l) \qquad (4)$$

この確信度を MP と呼ぶ．

3.1.3 Maximum Token Entropy

NLH は入力系列に対して事後確率が最大のラベル系列 $\mathbf{y}^*$ に注目し，MP は各トークンに対して割り当てられた周辺確率が最大のラベルに注目した．しかしこれらの確信度では，各トークンに最終的に割り当てられなかった他の多くの書誌要素ラベルの確率を考慮することができない．そこで，各トークンに付与された確率が最大のラベルだけでなく，その他のラベルとその確率も考慮する確信度を提案する．

この三つ目の確信度は各トークンへのラベル付けのエントロピーに基づいており，エントロピー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，すなわちラベル付けが困難であると判断する．反対に，一つの書誌要素ラベルの確率が高く，他が低ければラベル付けに自信があると判断する．そこでこのエントロピーに基づく第三の確信度を以下の式で定義する．

$$c_{MTE}(\mathbf{x}) := -\max_{i \le |\mathbf{x}|} \sum_{l \in L} -p(Y_i = l) \log p(Y_i = l) \qquad (5)$$

この式の冒頭の負号は，エントロピーが大きいほど書誌要素抽出が困難，よって確信度は小さいと考えるためである．この確信度を MTE と呼ぶ．

3.2 抽出誤りの検出方法

提案した書誌要素抽出の確信度を利用して，CRF が書誌要素抽出を誤っている論文を検出する．本研究では，確信度が低い論文は確信度が高い論文よりも誤りを含んでいる可能性が高いと考え，以下の方法で確信度の低い論文を検出する．

1. CRF によって書誌要素を抽出した各入力系列(論文)の確信度 $c.(\mathbf{x})$ を算出する．
2. 求めた確信度 $c.(\mathbf{x})$ に従って論文を昇順にランキングする．この時，ランク上位の論文ほど確信度が低いので，誤りが含まれる可能性が高いと判断する．
3. ランク上位の論文を誤りとして検出する．

このようにして検出した論文だけを人手で確認して，必要に応じて抽出された書誌情報を修正すれば，検出を行わない場合に比べて確認のコストを大幅に軽減できる．

4. 実験

4.1 概要

実験では，2.2 節で説明した学術論文のタイトルページの XML ファイルを入力として，CRF++によって各行の書誌要素ラベル付けを行うことでその書誌要素を抽出する．そこでまずこの書誌要素抽出精度を評価し，その後確信度による抽出誤りの検出性能について説明する．なお，この書誌要素抽出と確信度による誤り検出実験では，以下の 3 種類の学術論文誌の論文データを利用した．

- 情報処理学会論文誌(IPSJ)：2003 年，479 件
- 電子情報通信学会英文論文誌(IEICE-E)：2003 年，473 件
- 電子情報通信学会和文論文誌(IEICE-J)：2003, 2004 年，174 件

これらの論文のタイトルページの文書画像を OCR でレイアウト解析および文字認識して得られる XML テキストが，書誌要素を抽出する CRF への入力となる．また，このデータ作成に用いた OCR の文字認識精度は梗概の部分で 99%，参考文献の部分で 97%であった．参考文献の部分の認識精度が低いのは，この部分では日本語と英語の文字が混在しており，さらに様々なフォントや記号が使われていたことが主な理由である．なお書誌要素の抽出精度および誤り検出の精度の算出のための実験は全て 5 分割交差検定で行った．

4.2 書誌要素の抽出精度

実験では，CRF がテストデータである論文タイトルページの全ての行の書誌要素を過不足なく抽出できた論文のみを正解とした．よって本稿でいう書誌要素の抽出精度は，全ての書誌要素が正しく抽出された論文の全テストデータの論文に占める割合のことである．これはトークン(行)ではなく，トークン系列(論文)単位で算出した精度といえる．

また CRF の学習では，交差検定で学習用データに分類された論文の中から 20，100，300 件をそれぞれ 30 回無作為抽出し，学習データとした．このように学習データの件数を変えるのは，その結果として書誌要素抽出精度を変えて，4.3 節

において誤り検出の性能を評価するためである．

テストデータにおける書誌要素の抽出精度，およびCRFが書誌要素抽出を誤った論文の件数を表3に示す．なお各論文誌のテストデータの論文数は平均でIPSJが95.8件，IEICE-Eが94.6件，IEICE-Jが34.8件である．学習時に各件数で学習データを無作為に30回選択して5分割交差検定を行っているので，表3に示した結果は150回の実験の平均である．

表3から明らかに，学習データ量が増加するにつれて抽出誤りを含む論文件数は減少し，抽出精度は向上している．なおIEICE-Jについて学習データが300件の実験結果がないのは，IEICE-Jには全体で174件しか論文がないからである．4.3節では，書誌要素の抽出誤りを含む論文の検出実験について述べ，論文誌毎に検出性能と書誌要素抽出精度との関係を示す．

表3　書誌要素抽出精度(抽出誤り件数)
Table 3　Bibliographic Element Extraction Accuracies (%) and # of Erroneously Labeled Test Sequences (in parentheses)

学習データ件数	20	100	300
IPSJ	83.4% (15.9)	91.9% (7.8)	93.8% (6.0)
IEICE-E	69.5% (28.8)	89.7% (9.8)	95.9% (3.9)
IEICE-J	65.7% (12.0)	79.8% (7.0)	—

4.3 書誌要素抽出誤りの検出性能

CRFによって書誌要素抽出を行った論文の中からその誤りを含む論文を検出するため，提案した3種類の確信度に基づいて論文をランキングする．ランキング上位n件の論文についてCRFによる書誌要素抽出の正誤を確認し，これらを誤りとして検出したときの再現率と適合率を計算した．すなわち抽出誤りの検出を，抽出誤りを含む論文の検索としてとらえ，再現率と適合率をそれぞれ以下のように定義する．

$$\text{再現率} = \frac{\text{書誌要素抽出誤りを含む論文の検出数}}{\text{書誌要素抽出誤りを含む論文数}}$$

$$\text{適合率} = \frac{\text{書誌要素抽出誤りを含む論文の検出数}}{\text{検出した論文数}}$$

ここで適合率の分母の「検出した論文数」はnに等しい．

学習データを100件として，各確信度により誤りとして検出する論文の件数nを変えて，誤り検出の再現率及び適合率を算出した(図1)．また，そのときの再現率・適合率曲線を図2に示す．これらの図においてグラフ(a)，(b)，(c)は，それぞれ論文誌IPSJ，IEICE-E，IEICE-Jにおける実験結果を示している．

図1，2から，誤り検出の検索効率は，IEICE-E，IEICE-Jに比べてIPSJの方が良いことが分かる．またIEICE-Jは，図2(c)に示すように再現率が70%程度まで増加しても約60%の適合率を保っているため，IEICE-Eと比較すると若干良いといえる．しかし3種類の確信度の比較では，どの確信度が誤り検出に最も有効であるか断定するのは難しい．これは，それぞれの性能が論文誌毎に異なり，さらに論文誌が同じ場合でも再現率の水準によって異なっているからである．例えば，図2(a)においてNLHは再現率が約60%以下であれば最もよい適合率を示したが，再現率がこの水準を超えると3種類の確信度の中で適合率が最も悪くなっている．

図3に，学習データを300件としたときの，検出する論文の件数nと再現率および適合率の関係を示す．また図4にはこのときの再現率・適合率曲線を示す．これらの図にIEICE-Jのグラフがないのは，IEICE-Jの論文は全部で174件しかないからである．図4から，IPSJの結果がIEICE-Eの結果より良いことが分かる．図4では，論文誌の種類によらず，また広範な再現率の水準(約80%以下)においてNLHの適合率が最も良くなっている．またIPSJの結果について図4(a)を図2(a)と比べてもあまり差がない．一方，IEICE-Eの結果について比べると，図4(b)は図2(b)よりもかなり悪い．この理由の一つには，表3から明らかなように，学習データが増えたことによる抽出精度の改善が挙げられる．すなわち，学習データ件数が100の場合は検出すべき抽出誤りを含む論文が9.8件あったのに対して，学習データ件数が300ではそれが3.9件まで減っている．これにより誤り検出が困難になったと考えられる．

4.4 書誌情報の品質と人的コスト

最後に，本稿で提案した誤り検出によって，どの程度人手のコストをかければ，どの程度の品質の書誌情報が得られるかを検討する．具体的には，CRFによる自動抽出とその後の人手による確認と修正の後処理によって，最終的な書誌情報の精度として99%を実現することを考える．なお99%という数字自体に特に根拠はないが，書誌情報の精度としてはかなり高いと考えられる．

例えば論文誌をIPSJ，学習データ件数を300とすると，表3より書誌要素の抽出精度は93.8%で6.0件の論文に誤りが含まれている．この精度を後処理により99%にするには，抽出誤りを含む論文6.0件のうちの84%を検出すればよい．図3(a)から，NLHによる誤り検出で再現率が初めて84%を超えるのは検出件数が10件のときなので，この10件が後処理コストとみなせる．同様にして，表3に示した全ての書誌要素抽出結果に対して，確信度NLHによる誤り検出で精度99%を実現するために必要な後処理コスト，すなわち人が確認すべき論文数をまとめたのが表4である．

表3，4から，学習データ件数が少なく，その結果CRFの抽出精度自体が低い場合は，半分以上のデータを人が確認しなければならない場合が多く，99%という精度を実現するのは現実的ではない．しかし，CRFの学習データが300件で，そのときの抽出精度が93.8%のIPSJと95.9%のIEICE-Eでは，CRFによる抽出後にテストデータの約1割を人手で確認すれば，99%という高い精度が実現可能であることが分かる．

表4　後処理コスト(データ全体に占める割合)
Table 4　# of Manually Checked Articles for Post-processing and Its Ratio to the Total (in parentheses)

学習データ件数	20	100	300
IPSJ	43 (44.9%)	17 (17.7%)	10 (10.4%)
IEICE-E	76 (80.3%)	49 (51.8%)	10 (10.6%)
IEICE-J	34 (97.7%)	18 (51.7%)	—

5. まとめ

本稿では，学術論文のタイトルページから書誌情報を自動抽出する際に不可避な抽出誤りを検出する方法を提案した．

書誌情報の自動抽出は，特に英語論文を対象にいくつか提案されているが，抽出精度の比較に留まっており，本稿に示したような自動抽出の後処理の議論は聞かない．また現状の抽出精度も後処理なしで書誌情報抽出に利用できるような水準にはない．そこで我々は，CRF による論文タイトルページからの書誌情報抽出結果に確信度を定義して，確信度の低い論文を抽出誤りが含まれる可能性が高いとして検出した．これにより検出した確信度の低い論文だけを人手による後処理に回すことで，効率的に書誌情報の品質が改善できる見通しを得た．

実験では，提案した 3 種類の確信度がいずれも抽出誤り検出に有効であることを，複数の論文誌の論文データにおいて確認した．さらに，後処理のコストと後処理を含めて実現可能な書誌情報の精度について検討し，提案した誤り検出法により，一定のコストで高品質の書誌情報が得られることを確認した．すなわち，CRF による書誌情報抽出精度が約 94%の情報処理学会論文誌と約 96%の電子情報通信学会英文論文誌において，全体の約 1 割に相当する検出した論文を人が確認すれば，最終的に 99%の精度が実現可能であることを確認した．一方，実験では確信度による誤り検出の性能が雑誌毎にばらついたため，今後実験対象とする学術雑誌を増やすなどして，提案した確信度の有効性や利用法についてさらに検討していきたい．

【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] 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] Councill, I. G., Giles, C. L., and Kan, M. Y.: “Parscit: An Open-source CRF Reference String Parsing Package”, Proc. of Language Resources and Evaluation Conference (LREC 08), pp. 661-667 (2008).

[3] 藤尾正和，永崎健，高橋寿一: “正準判別分析とレイアウト型DPマッチングによる学術文献からの書誌情報抽出”，Proc. of DEIM Forum 2010, A9-3 (2010).

[4] 井上諒平，太田学，高須淳宏: “CRFによる論文文書画像の書誌要素推定における自動誤り検出”，Proc. of WebDB Forum 2011, 4G-2-3 (2011).

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

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

[7] Luong, M. T., Nguyen, T. D., and Kan, M. Y.: “Logical Structure Recovery in Scholarly Articles with Rich Document Features”, Int. J. Digital Library Systems, vol. 1, pp. 1-23 (2010).

[8] Ohta, M., Inoue, R., and Takasu, A.: “Empirical Evaluation of Active Sampling for CRF-based Analysis of Pages”, Proc. of 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”, Proc. of IADIS IS 2012, pp. 18–26 (2012).

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

[11] Settles, B. and Craven, M.: “An Analysis of Active Learning Strategies for Sequence Labeling Tasks”, Proc. of EMNLP 2008, pp. 1070-1079 (2008).

[12] Takechi, M., Tokunaga, T. and Matsumoto, Y.: “Chunking-Based Question Type Identification for Multi-sentence Queries”, In Proc. of SIGIR 2007 Workshop on Focused Retrieval (2007).

[13] 薬師貴之，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 (2009).

[14] 薬師貴之，太田学，高須淳宏: “様々な学術論文誌 OCR テキストからの書誌要素抽出”，2009 年電子情報通信学会総合大会講演論文集，情報・システム 2，D-12-48，p. 157 (2009).

[15] Zhao, H., Huang, C. N. and Li, M.: “An Improved Chinese Word Segmentation System with Conditional Random Field”, Proc. of Fifth SIGHAN Workshop on Chinese Language Processing, pp. 162-165 (2006).

[16] CRF++: Yet Another CRF toolkit, http://crfpp.sourceforge.net/

太田 学 Manabu OHTA

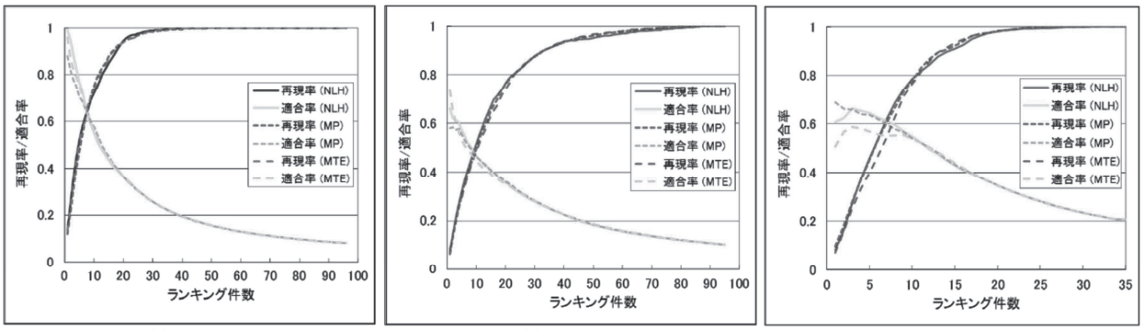
岡山大学大学院自然科学研究科准教授．1999 東京大学大学院工学系研究科電気工学専攻博士課程修了，博士(工学)．東京都立大学大学院工学研究科助手をへて 2005 岡山大学大学院自然科学研究科助教授．2007 より現職．Web 情報検索ならびに電子図書館に関する研究に従事．電子情報通信学会，情報処理学会，日本データベース学会，IEEE 各会員．

井上 諒平 Ryohei INOUE

株式会社四国日立システムズ勤務．2012 岡山大学大学院自然科学研究科電子情報システム工学専攻修了．在学中，学術論文からの書誌情報抽出に関する研究に従事．

高須 淳宏 Atsuhiro TAKASU

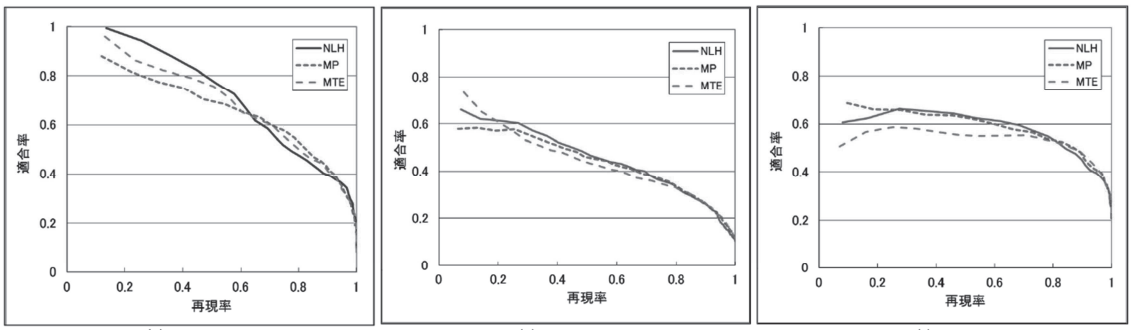
国立情報学研究所コンテンツ科学研究系教授．1989 東京大学大学院工学系研究科博士課程修了．工学博士．同年学術情報センター研究開発部助手．同センター助教授．国立情報学研究所助教授を経て 2003 より同研究所教授．データ工学，特にデータ解析と解析モデルの学習の研究に従事．電子情報通信学会，人工知能学会，日本データベース学会，ACM, IEEE 各会員．



(a) IPSJ　(b) IEICE-E　(c) IEICE-J

図 1　ランキング上位 *n* 件を検出した際の再現率及び適合率(学習データ 100 件)

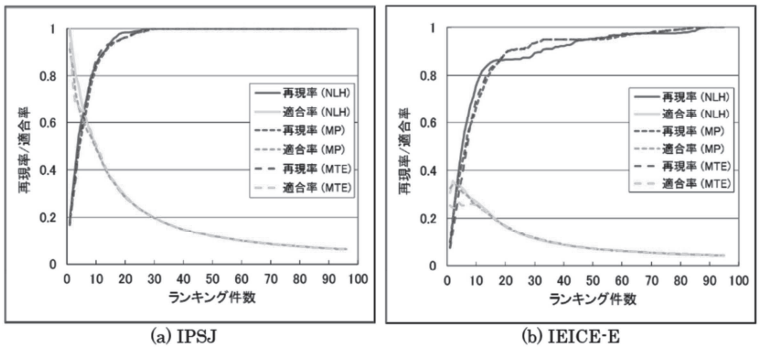
Fig.1　Recall and Precision w.r.t. Rank Cut-off *n* (# of Training Articles = 100)



(a) IPSJ　(b) IEICE-E　(c) IEICE-J

図 2　再現率・適合率曲線(学習データ 100 件)

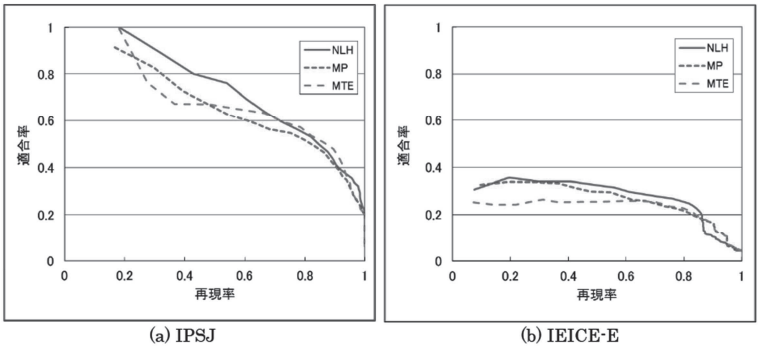
Fig.2　Recall-Precision Curve (# of Training Articles = 100)



(a) IPSJ　(b) IEICE-E

図 3　ランキング上位 *n* 件を検出した際の再現率及び適合率(学習データ 300 件)

Fig.3　Recall and Precision w.r.t. Rank Cut-off *n* (# of Training Articles = 300)



(a) IPSJ　(b) IEICE-E

図 4　再現率・適合率曲線(学習データ 300 件)

Fig.4　Recall-Precision Curve (# of Training Articles = 300)