Vol.2015-DBS-162 No.3
2015/11/26

CRFによる参考文献書誌情報抽出のための有効な素性の検討と拡充

松岡 大樹[1,a)]　太田 学[2,b)]　高須 淳宏[3,c)]　安達 淳[3,d)]

概要：膨大な文書が格納されている電子図書館を快適に運用するためには，書誌情報データベースの整備が必要である．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．本研究では，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が，その際，CRF で利用する素性が書誌情報の抽出精度を左右する．そこで実験により，使用する素性を変えて書誌情報の抽出精度を比較し，参考文献文字列のトークン化においては文字列素性と Bigram 素性が有効であり，トークンへの書誌要素ラベル付与においては辞書素性が有効であることを確認した．そして，本研究では書誌要素ラベル付与に有効であった辞書素性を拡充し，適当な素性を選択することによって抽出精度が向上することを確認した．

Examination and Enhancement of effective features for CRF-based bibliography extraction from reference strings

Daiki Matsuoka[1,a)]　Manabu Ohta[2,b)]　Atsuhiro Takasu[3,c)]　Jun Adachi[3,d)]

1. はじめに

多数の学術論文を蓄積している電子図書館のサービスを快適に運用するためには，検索やソート，文書間リンク等の機能が必須である．しかし，人手でそのため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，本研究では川上ら [1] と同様に，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) を用いて参考文献文字列から書誌情報を自動抽出する．

CRF を用いた参考文献書誌情報抽出においては，利用する素性が書誌情報の抽出精度を左右する．そこで我々は，どのような素性が書誌情報の高精度抽出に有効であるのか確かめるために，使用する素性を変えて実験を行い，抽出精度を比較した．また，実験の結果，書誌要素ラベル付与に有効であった辞書素性の拡充により抽出精度の改善を試みた．

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書誌情報抽出に関する研究を紹介し，3 節で本研究で行う 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 節で実験の評価と考察を行う．最後に 5 節で本稿をまとめる．

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他にルールによる手法もある．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

[1] 岡山大学工学部情報系学科
Department of Information Technology, Faculty of Engineering, Okayama University
[2] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
[3] 国立情報学研究所
National Institute of Informatics
[a)] pobp52cw@okayama-u.ac.jp
[b)] ohta@de.cs.okayama-u.ac.jp
[c)] takasu@nii.ac.jp
[d)] adachi@nii.ac.jp

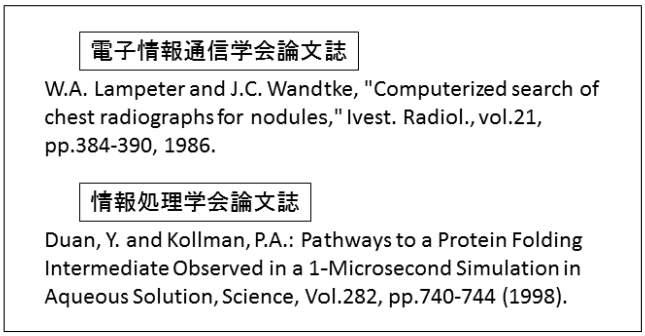


図 1 学術雑誌による参考文献文字列の書式の違い

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二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る．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．例えば，CRF[2] を用いた書誌情報抽出に関する研究に，Peng ら [3] や Councill ら [4] 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 [3]．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 [4]．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して，書誌情報を抽出した．Cora データセット [5] 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 値は 0.950 であった．

また，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 [6] 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．なお能動サンプリングは，学習に有効なデータを効率よ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，学習データ量を三分の一以下に削減できたと報告している．さらに Ohta らは [7] 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ている．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，本研究では川上ら [1] の研究をもとに，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．

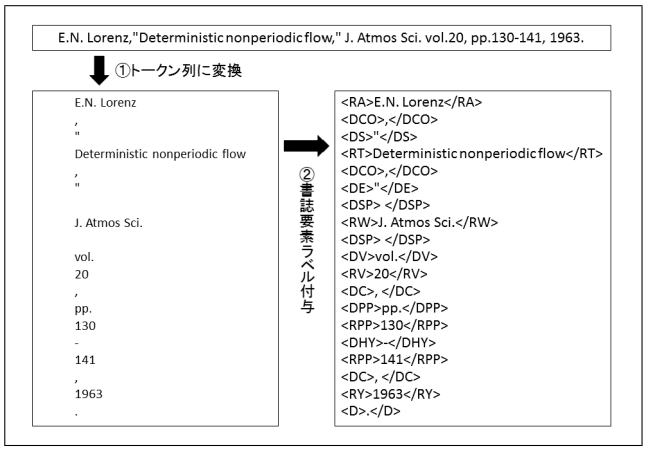


図 2 参考文献書誌情報抽出の例

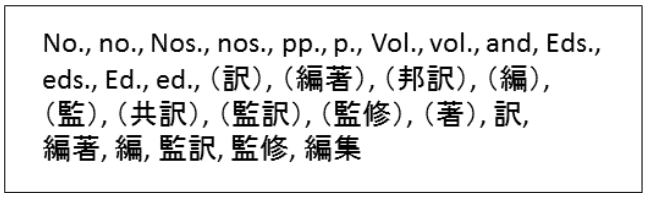


図 3 デリミタとする文字列 [9]

3. CRF による書誌情報抽出

3.1 書誌情報抽出

本研究では，学術論文の参考文献文字列から書誌情報を抽出する．そのため，図 2 のように参考文献文字列をまずトークン列に変換し，その後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．参考文献文字列のトークン化は BIO ラベル [8] と CRF を用いて行う．また，トークン列に書誌要素ラベル付与する際にも CRF を用いる．ゆえに，本研究においては CRF を 2 回使用する．

3.2 BIO ラベルによるトークン化

BIO ラベル [8] によるトークン化は，荒内ら [9] の方法を採用する．BIO ラベルを付与するために，まず，参考文献文字列をデリミタを用いてワードに分割する．トークン化に利用するデリミタは，図 3 に示す文字列とカンマ，ピリオド，空白文字，二重引用符，コロン，セミコロン，スラッシュ，ハイフン，丸括弧，鍵括弧，角括弧，波括弧とする．次に，各ワードに対して，書誌要素の先頭に該当すれ

表 1　抽出する書誌情報 [1]

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

表 2　素性テンプレート [1]

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)>	3	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 TB というラベルを付与し，先頭以外であれば TI というラベルを付与する．また，ワードがデリミタであれば，その先頭のワードには DB，それ以外であれば DI というラベルを付与する．このようなラベルが付与されたデータを CRF の学習データとし，未知の参考文献文字列のワード列に BIO ラベルを付与する．その後，ラベル付与されたデータの TB から TI，DB から DI のワードを結合してトークン列を得る．

3.3　書誌要素ラベル付与

書誌要素ラベル付与は川上ら [1] の方法を採用する．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，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 [9]．

3.4　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[2] の定義を用いて，参考文献文字列を BIO ラベル [8] により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．また CRF では，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 $(i-1)$ 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，トークン化においてはワード，書誌要素ラベル付与においては参考文献文字列をトークン化して得られるトークンである．一方，ラベル y_i は，トークン化においては BIO ラベル，書誌要素ラベル付与においては書誌要素またはデリミタのラベルである．

3.5　素性テンプレート

本研究では工藤が作成した CRF++ ver.0.58[*1] を利用して書誌情報を抽出する．CRF++で用いる素性テンプレートは川上ら [1] の素性テンプレートを使用する．これを表 2 にまとめる．この素性テンプレートは 48 種類の Unigram 素性と 1 種類の Bigram 素性の合計 49 種類の素性で構成されている．これらは全て言語的な素性で，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

[*1] http://taku910.github.io/crfpp/

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・末尾から四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，各種辞書のエントリの有無などを用いている．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 *2, 論文誌名 *3，会議名 *4，出版社名 *5，地名 *6，月名の辞書と，学会誌名など分類困難なものをまとめた辞書となっている．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，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要素数を表す．例えば，<first_1-4_string(0)> の場合，トークンの先頭の文字，先頭から二文字目までの文字，先頭から三文字目までの文字，先頭から四文字目までの文字という 4 つの要素を持つ．Bigram 素性は付与される書誌要素ラベルの連接に関する情報を表し，これにより書誌要素の出現順などに関する制約を考慮する．また，表 2 に示す素性テンプレートは書誌要素ラベル付与に対応したものであるため，ここに書かれているトークンは，トークン化においてはワードとなる．

4. 評価実験

4.1 実験概要

BIO ラベルによるトークン化，および，書誌要素ラベル付与においてどのような素性が有効であるか実験する．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，精度は 5 分割交差検定を用いて算出する．評価指標として，トークン化においては，正しくトークン列に分割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もの，書誌要素ラベル付与においては，参考文献文字列に含まれる書誌要素が過不足なく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ものを用いる．ただし，書誌要素ラベル付与の評価においては，表 1 を先行研究 [9] に倣って表 3 のように集約し，表 3 において同じ分類のものは正解判定において区別しない．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た．素性を比較するため，まず表 2 の素性を表 4 のように分類する．そして，表 4 の ALL 以外のカテゴリの素性を 1 カテゴリずつ抜いて実験を行い，精度を比較する．よって，精度が大きく低下したカテゴリほどトークン化，あるいは書誌要素ラベル付与への寄与が大きい素性といえる．本研究では，トークン化における有効な素性と書誌要素ラベル付与における有効な素性をそれぞれ検討するため，実験を分けて行う．また，書誌要素ラベル付与における実験では，人手で変換したトークン列を使用する．

*2 http://www.census.gov/genealogy/names/など
*3 http://science.thomsonreuters.com など
*4 http://www.allconferences.com/など
*5 http://www.narosa.com/nbd/PublisherDistributed.asp など
*6 http://www.fallingrain.com/world/index.html など

表 3 書誌要素ラベルの再分類 [9]

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

表 4 素性分類表

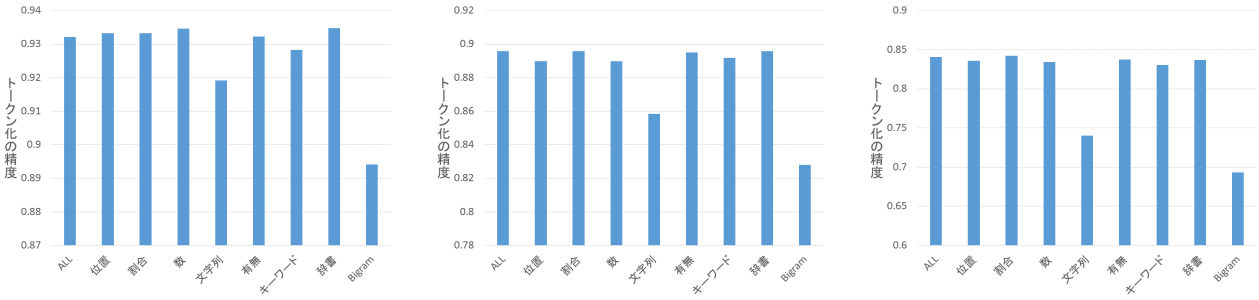
カテゴリ	素性
ALL	全ての素性
位置	ワード/トークンの絶対位置，相対位置
割合	各文字種の割合
数	全ワード/トークンの数，単語数，ピリオド数
文字列	ワード/トークン自身，ワード/トークンを小文字にした文字列と前後 4 文字までの文字列
有無	大文字，数字，記号の有無
キーワード	キーワードの有無，Editor や URL を表す語との照合
辞書	辞書的素性
Bigram	Bigram 的素性

表 5 文字列素性

素性	素性の説明
First_1-4_string	ワードの前から最大 4 文字までとった文字列
Last_1-4_string	ワードの後ろから最大 4 文字までとった文字列
Last_char	ワードの最後の文字
Token_lc	小文字にしたワード
Token	ワードそのもの

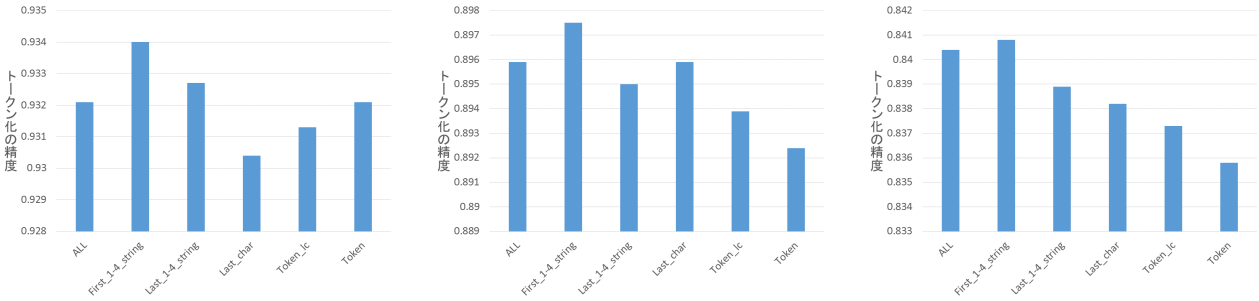
4.2 トークン化における有効な素性

CRF の素性テンプレートから表 4 の分類に従って 1 カテゴリずつ素性を抜き，BIO ラベルを用いてトークン化を行った．トークン化の精度を図 4 に示す．図 4 より文字列素性を用いなかった場合の抽出精度が大きく低下していることが確認できる．そこで，文字列素性の細分類素性を一つずつ抜き，抽出精度を比較し，文字列素性の中で特にどの素性が有効か検証する．文字列の分類に含まれる素性を表 5 に示す．表 5 の素性は表 2 のそれであるが，素性の説明はトークン化の実験のため，トークンではなくワードとした．これらの文字列素性をそれぞれ抜いた比較実験の結果を図 5 に示す．図 5 の (b)，(c) おいて Token の抽出精度が大きく低下しているため，トークン化において



(a) IEICE-J　(b) IEICE-E　(c) IPSJ

図 4　BIO ラベルによるトークン化の精度



(a) IEICE-J　(b) IEICE-E　(c) IPSJ

図 5　BIO ラベルによるトークン化の精度 (文字列の各素性を抜いた場合)

有効な素性であることがわかる．図 5 の (a) においては，Last_char を抜いた場合の抽出精度が大きく低下しているので，IEICE-J においては Last_char の方が Token よりも有効であるといえる．

また，図 4 をみると Bigram 素性を抜いた場合は文字列素性を抜いた場合より抽出精度が低下していることから，文字列素性よりもさらにトークン化において有効であることがわかる．BIO ラベルの遷移の具体例を見ると，前後のワードに TI のラベルが付与されると，そのワードも TI のラベルが付与されることが多い．よって，このような傾向を反映する Bigram 素性が，トークン化において有効だったといえる．

4.3　書誌要素ラベル付与における有効な素性

CRF の素性テンプレートから表 4 の分類に従って 1 カテゴリずつ素性を抜き，書誌情報抽出精度を比較した．結果を図 6 に示す．図 6 より，書誌情報抽出においては，辞書素性が特に有効であることがわかる．そこで，辞書素性の細分類素性を一つずつ抜き，書誌情報抽出精度を比較した．辞書素性の細分類素性を表 6 にまとめる．ここで，Dict は使用した辞書のビットを 1 とし，2 進表現したものを 10 進数に直した素性である．例えば図 7 のように，あるトークンが人名辞書と地名辞書と出版社名辞書のエントリに一致するとすると，このとき素性 Dict の値は 7 になる．辞書素性の比較実験の結果を図 8 に示す．図 8 より，辞書素性においては，全ての論文誌において Publisher を

表 6　辞書素性

素性	素性の説明	エントリ数
Dict	どの辞書に一致したか	—
Name	人名辞書	97,941
Month	月名辞書	50
Place	地名辞書	44,925
Publisher	出版社名辞書	3,718
Journal	論文誌名辞書	97,646
Conference	会議名辞書	3,204
Society	分類困難なものをまとめた辞書	442

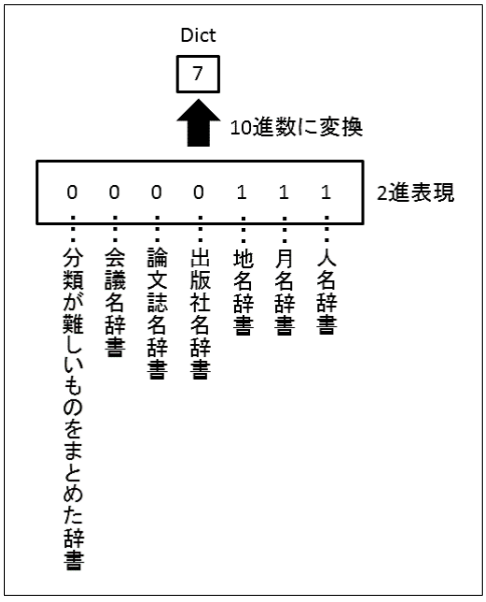
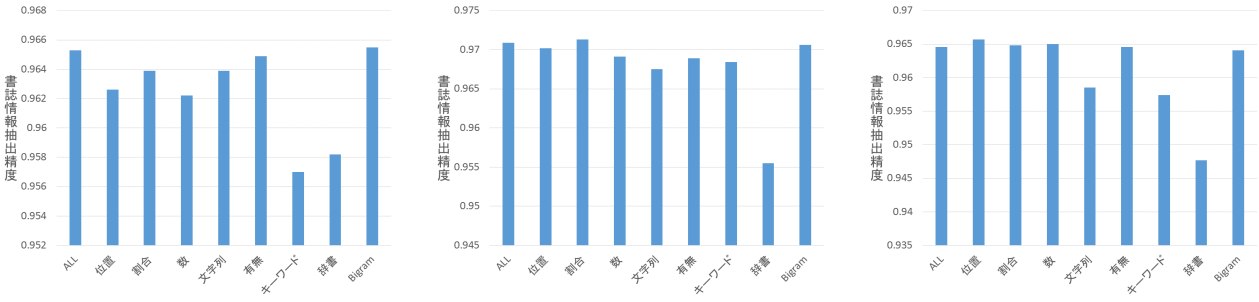
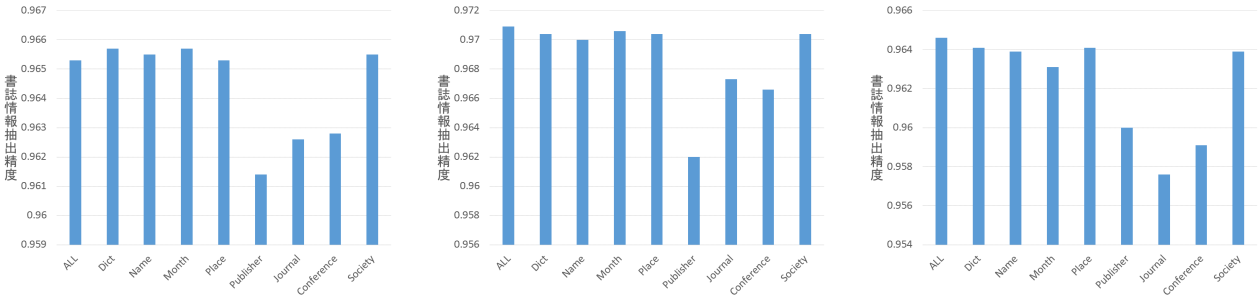


図 7　Dict 素性



(a) IEICE-J　(b) IEICE-E　(c) IPSJ

図 6　書誌情報抽出精度



(a) IEICE-J　(b) IEICE-E　(c) IPSJ

図 8　書誌情報抽出精度 (辞書の各素性を抜いた場合)

抜いたときの抽出精度が大きく低下しているため，この辞書はどの論文誌においても特に有効であることがわかる．また，(a)，(c) においては Journal と Conference の抽出精度が大きく低下しており，和文論文誌ではこれらの辞書も有効であることが確認できる．

一方，図 8 の (a) をみると，Dict や Name や Month は素性を抜くことによって抽出精度が向上している．これは Dict や Name や Month が CRF による書誌要素ラベル付与にあまり有効でないことを示唆する．しかしこれらの辞書は，図 8 の (b)，(c) においては素性としての有効性を示しているので，論文誌の違いによる辞書素性の影響をさらに精査する予定である．

4.4　書誌要素ラベル付与における辞書素性の拡充

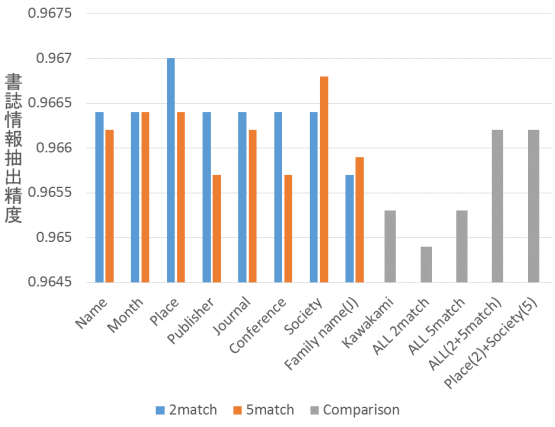
4.3 節の実験から書誌要素ラベル付与に有効であった辞書素性を拡充してさらに実験を行った．追加した素性は，表 2 の <keyword(i)>1 種類，<dictionary(i)>7 種類である．よって，素性の数は 56 種類の Unigram 素性と 1 種類の Bigram 素性の合計 57 種類となる．

追加した素性について説明する．川上ら [1] は，<keyword(i)> に含まれる日本語姓辞書の素性においては，完全一致，前方一致，後方一致，部分一致，不一致の 5 段階の判定を用いた．一方，<dictionary(i)> に含まれる辞書素性においては，各辞書のエントリに一致したか一致してないかの 2 段階判定を用いた．追加した素性は，日本語姓辞書との照合を 2 段階判定した素性が 1 種類と，辞書との照合を 5 段階判定した素性が 7 種類である．これにより，どちらの辞書素性においても 2 段階と 5 段階の両方で判定する．

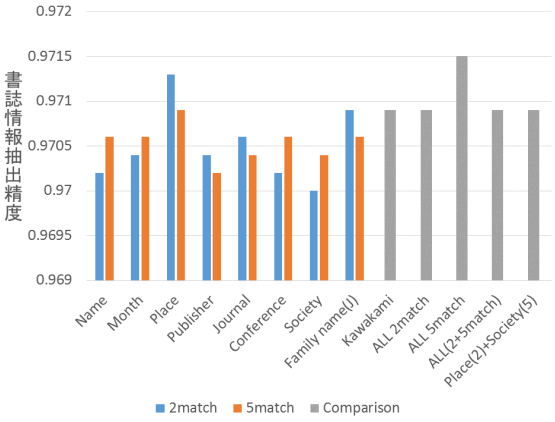
追加した辞書素性を用いてどちらの判定方法が適切であるか実験により評価した．実験で使用した辞書素性の照合判定を表 7 にまとめる．表 7 の Family name(J) は表 2 の <keyword(i)> に含まれる日本語姓辞書であり，エントリ数は 10,425 である．表 7 で○がついている素性が実験において使用した素性である．川上ら [1] の方法は，Name, Month, Place, Publisher, Journal, Conference, Society においては 2 段階の判定を使用し，Family name(J) においては 5 段階の判定を使用する．また，ALL 2match では全ての素性において 2 段階の判定を使用し，ALL 5match では全ての素性において 5 段階の判定を使用する．ALL(2+5match) では 2 段階と 5 段階のどちらの判定も使用する．照合判定を比較する実験では，比較対象の辞書のみ 2 段階または 5 段階のどちらかの判定方法を選択し，その他の辞書素性は 2 段階と 5 段階のどちら素性も使用している．例えば，人名辞書 (Name) の 2 段階の判定の実験では，人名辞書の 2 段階判定の素性およびその他の辞書素性の 2 段階判定と 5 段階の素性を使用した．この素性比較実験の結果を図 9 に示す．図 9 は，各辞書の 2 段階判定のみを使用したときと 5 段階判定のみを使用したときの比較結果および表 7 に示す素性による実験の結果である．論文誌別にみると，図 9 の (a) では，Month，Soceity，Family name(J) 以外の辞書において 2 段階判定の素性のみを使用したほう

表 7　辞書素性の照合判定

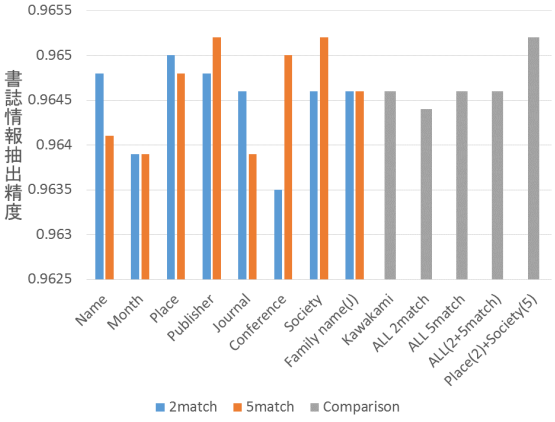
	kawakami		ALL 2match		ALL 5match		ALL(2+5match)		Place(2)+Society(5)	
	2	5	2	5	2	5	2	5	2	5
Name	○		○			○	○	○	○	○
Month	○		○			○	○	○	○	○
Place	○		○			○	○	○	○	
Publisher	○		○			○	○	○	○	○
Journal	○		○			○	○	○	○	○
Conference	○		○			○	○	○	○	○
Society	○		○			○	○	○		○
Family name(J)		○	○			○	○	○	○	○



(a) IEICE-J



(b) IEICE-E



(c) IPSJ

図 9　追加した辞書素性を用いた比較実験

表 8　各辞書における有効な照合方法の比較

辞書	IEICE-J	IEICE-E	IPSJ
Name	2 段階判定	5 段階判定	2 段階判定
Month	同精度	5 段階判定	同精度
Place	2 段階判定	2 段階判定	2 段階判定
Publisher	2 段階判定	2 段階判定	5 段階判定
Journal	2 段階判定	2 段階判定	2 段階判定
Conference	2 段階判定	2 段階判定	5 段階判定
Society	5 段階判定	5 段階判定	5 段階判定
Family name(J)	5 段階判定	2 段階判定	同精度

が抽出精度が高かった．また，Month においては，どちらの判定も同精度であった．(b) では，Place，Publisher，Journal，Family name(J) の 4 つの辞書においては 2 段階判定の素性を使用したほうが抽出精度が高かった．一方，Name，Month，Conference，Society においては 5 段階判定の素性を使用したほうが抽出精度が高かった．(c) では Name，Place，Journal においては 2 段階判定の素性を使用したほうが抽出精度が高かった．Month，Family name(J) においてはどちらの判定も同精度であったが，Publisher，Conference，Society においては 5 段階判定の素性を使用したほうが抽出精度が高かった．各辞書において有効な照合方法を表 8 にまとめる．表 8 より，いずれの論文誌でも Place は 2 段階判定，Society は 5 段階判定の方が有効であることが確認できた．その他の辞書については，論文誌によって優劣が逆転した．そこで，Place を 2 段階判定のみ，Society を 5 段階判定のみ，その他の辞書は 2 段階と 5 段階のどちらの判定も利用した実験結果が図 9 の Place(2)+Society(5) である．このとき使用した素性は表 7 の Place(2)+Society(5) である．この実験の結果，(b) においては川上ら [1] の方法と比べて抽出精度は向上しなかった．(a) においては，川上ら [1] の方法より抽出精度は向上したものの，ALL(2+5match) と同等であり，最も高い抽出精度ではなかった．(c) においては，Publisher の 5 段階判定の素性のみを使用したときと Society の 5 段階判定の素性のみを使用したときと同じ，最も高い抽出精度を示した．

結局図 9 の (a) においては Place の 2 段階判定のみを使

用したときの 0.967，(b) においては全ての辞書素性において 5 段階判定の素性を使用したとき (ALL 5match) の 0.9715，(c) においては Place の 2 段階判定の素性と Society の 5 段階判定の素性を使用したときなど，先に述べた三つの場合の抽出精度が 0.9652 で最も高かった．

4.5 辞書素性に関する考察

4.4 節で辞書との照合判定を変えることによって先行研究より書誌情報抽出精度が向上することを確認した．また，実験結果より，辞書ごとに 2 段階判定の素性と 5 段階判定の素性のどちらが書誌情報抽出に有効であるかを確認したが，有効な判定方法は辞書および論文誌によって異なることが多いとわかった．また，素性の判定方法を変更するだけでは抽出精度は大きく向上しなかったため，今後は辞書のエントリ数を増やすこと検討したい．本実験において使用した辞書は，日本語姓辞書と分類困難なものをまとめた辞書以外全て英語の辞書であった．相対的に日本語が少ないため，英語のエントリを持つ各辞書において，日本語のエントリを追加すれば，和文論文誌において抽出精度の向上が期待できる．

5. まとめ

本稿では，CRF による参考文献書誌情報抽出に有効な素性を検討し，書誌要素ラベル付与において有効だった辞書素性の拡充を試みた．実験の結果，BIO ラベルによるトークン化においては，ワード自身の素性や，ラベルの遷移を表す Bigram 素性が有効であることを確認した．また，書誌要素ラベル付与においては，辞書素性，その中でも出版社名辞書や論文誌名辞書，会議名辞書が有効であることを確認した．この実験結果に基づいて，書誌要素ラベル付与で利用する辞書素性を拡充し，各辞書との照合において 2 段階判定と 5 段階判定のどちらが有効であるかを確認した．その結果，地名辞書は 2 段階判定，学会誌名など分類困難なものをまとめた辞書は 5 段階判定が有効であることがわかった．しかしその他の辞書素姓は，辞書および論文誌によって有効な判定方法は異なっていた．今後さらに抽出精度を向上させるために，有効な辞書のエントリを増やしていくことを検討する．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (C)(課題番号 25330384，15H02789) 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] 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.
[2] 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.
[3] F. Peng, A. McCallum, “Accurate Information Extraction from Research Papers Using Conditional Random Fields”, HLT-NAACL 2004, pp. 329–336, 2004.
[4] 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.
[5] 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.
[6] 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.
[7] M. Ohta, R. Inoue, A. Takasu, “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.
[8] E. Tjong Kim Sang, S. Buchholz, “Introduction to the CoNLL-2000 Shared Task: Chunking.”, Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning, Association for Computational Linguistics, vol. 7, pp. 127–132, 2000.
[9] 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, vol. 2012-DBS-156, no. 1，pp. 1–8, 2012.