Vol.2012-DBS-156 No.1
2012/12/12

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

荒内 大貴[1]　太田 学[1]　高須 淳宏[2]　安達 淳[2]

概要：学術論文の参考文献に記述されている著者名や論文題目名といった書誌要素は，検索等で利用されるため非常に重要である．本稿では，この参考文献の文字列から自動で書誌要素を抽出する手法を提案する．提案手法では，参考文献文字列の様々な特徴を利用して，Conditional Random Fields (CRF) により，その文字列を分割したトークン列に対して，書誌要素ラベルを付与する．書誌要素抽出実験の結果，主要な書誌要素が和文で 96 %以上，英文で 93 %以上の精度で抽出できることが分かった．

1. はじめに

昨今整備されている電子図書館の中には，多数の学術論文のデータベースを構築しているものがある．そのデータベースを利用し，目的の文書を探すためには，著者名や論文題目名などの書誌情報が必須である．しかし，これらの書誌情報をデータベースに入力するには膨大な人的コストがかかるため，その作業を可能な限り自動で行う文書解析技術が必要とされている．一方で，学術論文にはそれと関連する多くの参考文献が記述され，そこには重要な書誌情報が集中する．この書誌情報を利用して参考文献エンティティの同定ができれば，文書間リンクの自動生成など様々な応用が考えられる．国内で発表されている学術論文の多くは日本語または英語で記述されているため，本稿では学術論文中の和英文の参考文献文字列のテキストから，自動でその書誌要素を抽出する手法を提案する．

本稿の構成は次の通りである．2 節で，学術論文からの自動書誌要素抽出に関する研究を紹介し，続く 3 節で提案手法について説明する．4 節では提案手法の評価実験について述べ，関連研究 [1][2] と比較する．5 節で本稿をまとめる．

また，本稿の主な貢献を以下にまとめる．

(1) BIO ラベルを用いた汎用性の高い参考文献文字列のトークナイズ手法の提案
(2) 和英文の電子情報通信会論文誌を用いて提案手法の有効性を実験的に検証

[1] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
[2] 国立情報学研究所
National Institute of Informatics

2. 参考文献文字列からの書誌要素抽出

参考文献文字列からの書誌要素抽出には，阿辺川ら [3], Peng ら [4], Okada ら [5], Councill ら [2] の研究がある．本稿では Conditional Random Fields (CRF)[6] を用いて書誌要素を抽出するが，阿辺川らと Okada らは Support Vector Machine (SVM)[7] と Hidden Marlov Model (HMM)[8], Peng らは HMM と CRF の二つのモデルを利用して書誌要素を抽出した．Councill らは，CRF のみを利用して書誌要素を抽出している．

阿辺川らは，OCR によって解析された学術論文のタイトルページと参考文献文字列から書誌要素を抽出する手法を提案している．彼らは，日本語及び英語で書かれた様々な論文を対象に，SVM と HMM 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た．また彼らは，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たため，参考文献文字列を和文と英文に分類し，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要素を過不足無く抽出できたと報告している．

Peng らは，英語で書かれた論文を対象に，HMM と CRF を用いて論文タイトルページと参考文献文字列に対して単語単位で書誌要素ラベルを付与した．実験により，参考文献文字列では 77.3 %の精度で，参考文献文字列中の全ての書誌要素を過不足無く抽出できたと報告している．

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ナイズし，各トークンに SVM と HMM を用いて書

<Author> 荒内大貴 </Author>,
<Author> 太田　学 </Author>,
<Author> 高須淳宏 </Author>,
<Author> 安達　淳 </Author>,“
<Title> CRFによる参考文献文字列からの書誌要素抽出の一手法 </Title>,“
<Conference> WebDB Forum 2011 </Conference>,
<Location> 東京 </Location>,
<Month> Nov. </Month>
<Year> 2011. </Year>.

図 1　トークンに付与された書誌要素

誌要素ラベルを付与している．実験において，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，97.6 %の精度で全ての書誌要素が過不足無く抽出できたことを報告している．

Councill らは，CRF により参考文献文字列の書誌情報を抽出できるオープンソースのツールである ParsCit を用いて，英文の参考文献文字列から書誌要素を抽出した．参考文献文字列を空白文字をデリミタとしてトークナイズし，英文の単語単位で書誌要素ラベルを付与した．Cora データセット [9] の参考文献文献文字列を対象にした実験において，著者名や論文題目名といった重要な書誌要素抽出で 98 %以上の精度を達成した．

3. 提案手法

3.1 概要

本稿では学術論文の参考文献文字列から書誌要素を抽出する手法を提案する．例えば以下のような参考文献文字列があるとする．

- 荒内 大貴，太田 学，高須 淳宏，安達 淳，“ CRF による参考文献文字列からの書誌要素抽出の一手法，” WebDB Forum 2011，東京，Nov. 2011.

本手法の目的は，このような参考文献文字列をパージングして，著者名や論文題目名といった重要な書誌要素を抽出することである．本手法では，まず参考文献文字列を BIO ラベルと CRF を用いてトークナイズする．その後それぞれのトークンに CRF を用いて書誌要素ラベルを付与する．よって上記の文字列からは，図 1 のようにそれぞれのトークンに書誌ラベルが付与されたデータが得られれば良い．

3.2 BIO ラベルによるトークナイズ

本節では，BIO ラベル [10] を用いたトークナイズ処理について説明する．BIO ラベル付与には CRF を利用する．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．本稿でデリミタは，図 2 に定める文字列とカンマ，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，波括弧と定義する．

No., no., Nos., nos., pp., p., Vol., vol., and, Eds., eds., Ed., ed.,（訳），（編著），（邦訳），（編），（監），（共訳），（監訳），（監修），（著），訳, 編著,編,監訳, 監修, 編集

図 2　デリミタとする文字列

参考文献文字列

I. Zoghlami, O. Faugeras, and R. Derche, "Using geometric corners to build a 2D mosaic from a set of images," CVPR '97, pp.420-425, 1997.

BIOラベルが付与されたワード

<TB> I </TB>
<TI> . </TI>
<TI> _ </TI>
<TI> Zoghlami </TI>
<DB> ,_ </DB>
<TB> O </TB>
<TI> . </TI>
<TI> _ </TI>
<TI> Faugeras </TI>
<DB> ,_ </DB>
<DB> and_ </DB>
<TB> R </TB>
<TI> . </TI>
<TI> _ </TI>
<TI> Derche </TI>
<DB> ,_ </DB>
<DB> " </DB>
<TB> Using </TB>
<TI> _ </TI>
<TI> geometric </TI>
<TI> _ </TI>
<TI> corners </TI>
<TI> _ </TI>
<TI> to </TI>
<TI> _ </TI>
<TI> build </TI>
…

図 3　BIO ラベルが付与されたワードの例

次に各ワードに対して，ワードが書誌要素の先頭に該当すれば TB というラベルを付与し，先頭以外にあれば TI というラベルを付与する．またワードがデリミタを構成するなら，その先頭のワードには DB，それ以外のワードには DI というラベルを付与する．その例を図 3 に示す．

このようにラベル付与されたデータを学習データとして CRF を学習し，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とする．その際，一つの書誌要素が複数のトークンに分割される過剰分割と一つのトークン中に複数の書誌要素が存在する過少分割が発生する場合がある．過剰分割と過少分割それぞれの例を図 4 と図 5 に示す．

図 4 では 8 行目から 10 行目にかけて会議名が複数のトークンに分割されており，過剰分割が発生している．図 5 では 9 行目において本のタイトルとそれに続く出版社が分割できておらず，過少分割が発生している．このようなトークナイズの誤りは，その後のトークンへの書誌要素ラベル付与の精度を下げる原因になるため，精度の良いトークナイズが要求される．BIO ラベルを用いたトークナイズ処理

参考文献文字列

B.-H. Juang, "From speech recognition to understanding : Shifting paradigm to achieve natural human-machine communication," Proc. 16th ICA and 135th Meeting ASA, pp.617-618, 1998.

トークナイズ処理の結果

1	B.-H._Juang
2	,_
3	"
4	From_speech_recognition_to_understanding_:_Shifting_ paradigm_to_achieve_natural_human-machine_communication
5	,
6	"
7	_
8	Proc._16th_ICA
9	_and_
10	135th_Meeting_ASA
11	,_
12	pp.
13	617
14	-
15	618
16	,_
17	1998
18	.

図 4 過剰分割の例

参考文献文字列

竹林 滋, 渡辺未耶子, 清水あつ子, 斉藤弘子, 初級英語音声学, 大修館書店, 東京, 1991.

トークナイズ処理の結果

1	竹林 滋
2	,
3	渡辺未耶子
4	,
5	清水あつ子
6	,
7	斉藤弘子
8	,
9	初級英語音声学, 大修館書店
10	,
11	東京
12	,
13	1991
14	.

図 5 過少分割の例

は，対象の学術雑誌の種類を問わず適用できるので汎用性が高い．

3.3 CRF を用いたトークンへのラベル付与

3.3.1 Conditional Random Fields

本研究では，SVM や HMM と並び，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 CRF を利用する．CRF とは系列ラベル付与のために設計された識別モデルであり, 正しい系列ラベリングを他の全ラベリング候補と弁別するような学習を行う．また CRF は条件付き確率場とも呼ばれ，形態素解析 [11] や固有表現抽出 [10] などの自然言語処理のような分野で広く利用されている．

本研究の書誌要素ラベル付与問題では, 標準的なチェーンモデルの CRF を用いる．すなわち, 入力系列 $\mathbf{x} = x_1, ..., x_n$ が与えられた時，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．

表 1 抽出する書誌要素

書誌要素	書誌要素ラベル
Author	RA
Editor	RE
Title	RT
Booktitle	RB
Journal	RW
Conference	RC
Volume	RV
Number	RN
Page	RPP
Publisher	RP
Day	RD
Month	RM
Year	RY
Etc	ETC

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

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以下のようになる．

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

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．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 $\mathbf{x}$ に対する可能な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 $\hat{\mathbf{y}}$ は次式で与えられる．

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in Y(\mathbf{x})}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) \qquad (3)$$

本稿の場合，ラベル付与の対象である入力 x_i は，参考文献文字列をトークナイズして得たトークンである．一方ラベル y_i は，著者名，論文題目名といった書誌要素である．

3.3.2 書誌要素とデリミタ

4 節の実験で用いる和英の電子情報通信学会論文誌の論文の参考文献文字列から抽出する書誌要素の一覧と，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 Etc は他のどの書誌要素にも分類されない書誌要素である．

また参考文献文字列中のデリミタの種類が書誌要素判定の手掛かりになるので，デリミタも抽出する．抽出するデリミタの一覧を表 2 に示す．

よって出力は，参考文献のトークン列に書誌要素ラベルとデリミタラベルを付与したデータとなる．出力データの例を図 6 に示す．各行の <RA> や <RT> などのラベルが本手法によって各トークンに割り当てられた書誌要素ラベルを表す．一方，<DZC> や <DSP> など D で始まるラベルはデリミタである．

表 2 抽出するデリミタ

デリミタ	ラベル
.（ピリオド）	D
,␣（半角カンマ+空白文字）	DC
,（半角カンマ）	DCO
，（全角カンマ）	DZC
␣（空白文字）	DSP
␣and␣ , and␣	DAND
Eds., eds., Ed., ed.,（訳）,（編）,（編著）,（監）,（監修）,（監訳）,（邦訳）,（共訳）, 訳, 編, 編著, 監修, 監訳, 編集, editors	DED
”(半角二重引用符)	DS
,”␣(カンマ+半角二重引用符+空白文字)	DE
“ (全角二重引用符・始)	DZS
，” (全角カンマ+全角二重引用符・終)	DZE
Vol., vol.	DV
No., no., Nos., nos.	DN
pp., p.	DPP
:, ;（コロン，セミコロン）	DCL
/（スラッシュ）	DSL
-（ハイフン）	DHY
(, [, 「, {（各種括弧・始）	DLBR
),], 」, }（各種括弧・終）	DRBR

```
<RA> 荒内大貴 </RA>
<DZC>，</DZC>
<RA> 太田 学 </RA>
<DZC>，</DZC>
<RA> 高須淳宏 </RA>
<DZC>，</DZC>
<RA> 安達淳 </RA>
<DZC>，</DZC>
<DZS>“ </DZS>
<RT> CRF による参考文献文字列からの書誌要素抽出の一手法 </RT>
<DZE>，”</DZE>
<RC> WebDB Forum 2011 </RC>
<DZC>，</DZC>
<ETC> 東京 </ETC>
<DZC>，</DZC>
<RM> Nov. </RM>
<DSP>  </DSP>
<RY> 2011 </RY>
<D> . </D>
```

図 6 書誌要素ラベルが付与された出力データの例

3.3.3 素性テンプレート

CRF による書誌要素のラベル付与に利用するトークン等の特徴をまとめたものを素性テンプレートと呼ぶ．本手法で用いる素性テンプレートを表 3 に示す．表 3 で各素性を表す文字列の括弧内の数字はトークンの相対位置を表している．ここで i は-4，-3，-2，-1，0，1，2，3，4 を表し，前後 4 つまでのトークンを素性にすることを表している．本手法で使用する素性テンプレートはレイアウト情報を用いておらず，言語的素性のみで構成されている．

本研究の書誌要素抽出では CRF を 2 回利用する．すなわち，参考文献文字列を BIO ラベルを用いてトークナイズする時と，そのトークンに書誌要素ラベルを付与するときの 2 回である．表 3 に示す素性テンプレートは，それらの両方で用いる．

表 3 ラベル付与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
	<zen_kan(0)>	トークン内の漢字数の割合
	<zen_hir(0)>	トークン内の平仮名数の割合
	<zen_kat(0)>	トークン内の片仮名数の割合
	<zen_alp(0)>	トークン内の全角アルファベット数の割合
	<zen_fig(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_fig(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
	<editor(0)>	参考文献文字列における editor 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(i)>	トークン自身
Bigram	< y(-1),y(0)>	ラベルの遷移

表 3 より本手法では素性として，トークンをトークン列の先頭から数えた時の絶対的な出現位置と，1 から 10 までの相対的な出現位置を用いる．またトークンの文字数，その文字数に対するトークン内の各文字種の文字数の割合，トークンの最後の文字，トークンの先頭から 4 文字目までの文字列と末尾から 4 文字目までの文字列，トークン内のアルファベットを全て小文字にした文字列，トークン内の大文字アルファベットの有無，トークン内の数字の有無，参考文献文字列中の “Eds.” や “編著” といったような Editor に関係する文字列の有無を用いる．

ここでトークンの先頭から四文字目までの文字列と末尾から四文字目までの文字列について説明する．例えば “Charlotte” というトークンの先頭から四文字目までの文字列というのは，“C”，“Ch”，“Cha”，“Char” の 4 つの文字列のことを意味する．同様に，末尾から四文字目までの文字列というのは，“e”，“te”，“tte”，“otte” の 4 つの文字列のことを意味する．

さらに書誌要素の判別に有用な辞書的素性と特徴的な文字列の種類の有無も用いる．辞書的素性は各種の辞書との照合状況を示すものである．人名[*1][*2][*3]，月名，地名[*4]，出版社名[*5][*6] の辞書を用意した．人名辞書には 97,942 件，月名辞書には 25 件，地名辞書には 49,885 件，出版社名辞書には 727 件の語を収録している．また，ParsCit[2] にならい，辞書 ID として人名辞書に “1”，月名辞書に “2”，地名辞書に “4”，出版社名辞書に “8” を与える．あるトークン

*1 ftp://ftp.funet.fi/pub/doc/dictionaries/DanKlein/
*2 http://www.census.gov/genealogy/names/
*3 http://www.geocities.com/Tokyo/3919/atoz.html
*4 http://www.akatsukinishisu.net/kanji/chimei.html
*5 http://www.narosa.com/nbd/PublisherDistributed.asp
*6 http://en.wikipedia.org/wiki/List_of_publishers

表 4　特徴的な文字列の例

特徴的な文字列	予想される書誌要素
ブック, Handbook	Booktitle
～論文誌	Journal
Proc., シンポジウム	Conference
～社, ～出版	Publisher
Jan., Feb., March	Month
et al., http	Etc

がこれらの辞書にヒットした時，ヒットした辞書の ID の和を辞書的素性とする．例えば，“Charlotte” というトークンは人名の辞書と地名の辞書にヒットするため，人名辞書の ID である “1” と地名辞書の ID である “4” の和の “5”，がこのトークンの辞書的素性となる．これらの辞書の内容は全て英語で構成されている．

特徴的な文字列とは，例えば，トークン中に “Proc.” といった文字列があれば，そのトークンは会議名を表す書誌要素であると予想できる．したがって，“Proc.” という文字列の有無を素性とする．このような特徴的な文字列の例を表 4 にまとめる．

またトークンへのラベル付与では Bigram 素性も使用する．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，例えば著者名の後ろに論文題目名が記述され，続いて会議名が記述されるといった書誌要素の出現順に関する制約を考慮することができる．

4. 評価実験

提案手法の有効性を調べるため，評価実験を行った．本研究では工藤が作成した CRF++[*7] を利用して参考文献文字列データに書誌要素ラベルを付与する．CRF++は，系列データに対して分割やラベル付与などの処理を行うオープンソースのソフトウェアである．実験では CRF++の学習パラメータはデフォルトの値を利用した．

また実験のため，2000 年の和文の電子情報通信学会論文誌 Vol.J83-DII No.1 から No.12 と，同年の英文の電子情報通信学会論文誌 Vol.E83-A No.1 から No.12 の，それぞれ一年分に相当する論文の参考文献コーパスを用意した．実験では，そこに記述されている，和文 4,787 件，英文 4,497 件の参考文献文字列を使用した．和文の電子情報通信学会論文誌には日本語と英語の参考文献文字列が混在しているが，英文の電子情報通信学会論文誌には英語の参考文献文字列のみが記述されている．

正解ラベル付き参考文献コーパスを用いて，BIO ラベルを用いたトークナイズ精度を算出した．また，参考文献文字列中の各書誌要素の抽出精度と一つの参考文献の文字列から全ての書誌要素を過不足なく正確に抽出できたかどうかの全体的な抽出精度を算出した．書誌要素の抽出精度を評価する際には，表 1 の書誌要素ラベルを表 5 のようにまとめ，5 分割交差検定で精度を算出した．和文誌では我々の WebDB Forum 2011 で提案した手法 [1] と，また英文誌では 2 章で紹介した ParsCit[2] と，抽出精度を比較した．

[*7] http://crfpp.sourceforge.net/

表 5　書誌要素ラベルの再分類

書誌要素ラベル	分類名
RA，RE	AUTHOR
RT，RB	TITLE
RW，RC	JOURNAL
RV，RN，RPP	VOLUME
RP	PUBLISHER
RD	DAY
RM	MONTH
RY	YEAR
ETC	ETC

4.1 BIO ラベルを用いたトークナイズ精度評価

和英文論文誌の参考文献文字列を正しくトークナイズできるかどうか実験し，5 分割交差検定で精度を算出した．その結果，和文で 93.02 %，英文で 90.37 %の参考文献文字列で過不足なくトークンに分割することができた．

4.2 和文論文誌からの参考文献書誌要素抽出

3.2 で説明した BIO ラベルを用いて参考文献文字列をトークン列に変換して書誌要素ラベルを付与した．その結果の書誌要素抽出精度を表 6 に「BIO」として示す．さらに，比較対象として我々の WebDB Forum 2011 で提案した手法 [1] の精度も表 6 に「BEST-WebDBF2011」及び「BIO-WebDBF2011」として示す．「BEST-WebDBF2011」とは電子情報通信学会論文誌の参考文献文字列のヒューリスティクスを利用してトークナイズを行った場合の実験結果で，WebDB Forum 2011 で提案した手法の中で最も精度が高かった．この手法はトークナイズ精度が本手法よりも優れ，それにより高精度の書誌要素抽出を実現した．しかし，この手法はトークナイズにヒューリスティクスを利用しているため，他雑誌に適用する際にコストがかかる．一方「BIO-WebDBF2011」は WebDB Forum 2011 で BIO ラベルを用いた手法の精度である．本手法とは違い，この手法ではコロン類，スラッシュ，ハイフン，括弧類をデリミタとして扱っていない．さらに，用いている素性テンプレートが異なる．

また，WebDB Forum 2011 で用いた実験データは本稿と同じ電子情報通信学会論文誌 Vol.J83-DII No.1 から No.12 であるが，両者のコーパスは正解ラベルが若干異なっている．本稿のコーパスの方がより詳しく分類された正解ラベルになっている．

表 6 より書誌要素 DAY の抽出精度が全体的に悪い結果となっているが，これは 4,787 件の参考文献文字列に 9 件しか出現しない要素であるため，参考文献文字列全体の抽

出精度に与える影響は大きくない．

表 6 より提案手法と「BEST-WebDBF2011」と比較すると，AUTHOR と JOURNAL の二つの書誌要素で勝っていることが分かる．また，「BIO-WebDBF2011」と比較すると DAY 以外の書誌要素と全体的な精度である ALL の抽出精度で勝っていることが分かる．本手法では「BIO-WebDBF2011」で用いた素性テンプレートを拡充したものを用いているので精度が向上したと考える．

4.3 和文論文誌におけるエラー解析

抽出誤りの原因は，過剰分割による誤りと過少分割による誤り，そして単純なラベル付与誤りの大きく 3 つに分けられる．単純なラベル付与誤りとは，トークナイズ処理には過不足が無いにも関わらずラベル付与を誤ることである．

4.2 の実験では 4,787 件の参考文献文字列中，335 件の参考文献文字列で抽出誤りが発生した．図 7 にその抽出誤りの原因別の分類を示す．図 7 より，約半数の 174 件の参考文献文字列で単純なラベル付与誤りが発生していた．

また，表 7 にトークナイズ処理の過不足が無い場合について，トークン毎の書誌要素ラベル付与の状況をまとめる．表 7 の縦は正解ラベル，横は CRF による付与ラベルを表す．また，空欄は 0 件である．表 7 より TITLE を AUTHOR または JOURNAL，JOURNAL を TITLE，PUBLISHER を JOURNAL，ETC を JOURNAL と誤る場合が多いことが分かる．

そこでジャーナル名の日本語と英語の辞書と PUBLISHER[*8]の日本語の辞書を追加し，同様の実験を行った．その結果が表 6 の「BIO(add dictionary)」である．この辞書の追加によって AUTHOR，TITLE，VOLUME，PUBLISHER，YEAR，ETC の 6 つの書誌要素で抽出精度の向上が見られた．また，全体的な精度である ALL も 94 %を超え，「BEST-WebDBF2011」の精度を上回った．

4.4 英文論文誌からの参考文献書誌要素抽出

4.2 と同様に英文の参考文献文字列から切り出されたトークンに書誌要素ラベルを付与した．その結果の書誌要素抽出精度を表 8 に「BIO」として示す．また比較対象である ParsCit[2] の抽出精度を「ParsCit」として示す．

ただし，公開されている ParsCit とは書誌要素体系の細部が異なり，また ParsCit をそのまま用いると一部の書誌要素，具体的には DAY と MONTH が出力されなかった．そこで提案手法の素性テンプレートのみを ParsCit のそれに置き換えて実験した．

表 8 より「ParsCit」と比較すると YEAR 以外の書誌要素と ALL で「ParsCit」を上回っている．

4.5 英文論文誌におけるエラー解析

4.4 の実験では 4,497 件の参考文献文字列中，339 件の参考文献文字列で抽出誤りが発生した．図 8 にその抽出誤りの原因別の分類を示す．図 8 より，単純なラベル付与誤りは全体の約 3 割程度であり，抽出誤りの多くはトークナイズ誤りが原因であることが分かる．

また，表 9 にトークナイズ処理の過不足が無い場合について，トークン毎の書誌要素ラベル付与の状況をまとめる．表 9 より，TITLE を AUTHOR や JOURNAL，JOURNAL を TITLE，PUBLISHER を JOURNAL と誤る場合が多いことが分かる．その中でも PUBLISHER を JOURNAL と誤る割合が大きいことが読み取れる．この結果より，出版社名辞書の拡充が精度の向上に繋がると考えている．

4.6 考察

4.1 で示したトークナイズ精度より，和文論文誌より英文論文誌の方がトークナイズ誤りが多く発生していて，また図 7 と図 8 より，書誌要素抽出においてもその影響が英文論文誌の方が大きいことが分かる．英単語を区切る空白文字は，多くは書誌要素の一部であるが，デリミタとしても用いられる．そのため，空白文字が書誌要素の一部なのかデリミタなのかが曖昧になり，トークナイズ誤りを引き起こすのではないかと考える．

精度の良い書誌要素抽出を実現するためには，精度の良いトークナイズが必要不可欠ではあるが，BIO ラベルを用いたトークナイズの汎用性を維持したまま抽出精度を上げることは困難であると考えている．そこで，抽出結果に確信度を定義し，その確信度を用いて抽出困難な参考文献文字列を自動検出して，人手で修正するような書誌情報抽出を考えている．

5. おわりに

本稿では，学術論文の参考文献文字列のテキストから，CRF を用いて書誌要素を自動抽出する手法を提案した．提案手法は CRF により，参考文献文字列のテキストをまずトークン列に変換し，各トークンに著者名，論文題目名，会議名などの書誌要素ラベルを付与する．書誌要素抽出実験において，和文論文誌では著者名や論文題目名，論文誌名といった重要な書誌要素と全体的な精度で，我々が昨年提案した手法の抽出精度を上回った．また英文の論文誌でも，汎用的な参考文献文字列解析ツールである ParsCit の抽出精度を全体的に上回った．今後は，確信度等を用いて書誌要素抽出が困難な参考文献文字列を自動検出し，検出された参考文献文字列を人手で修正することで最終的な精度向上を図る手法について検討していきたい．

[*8] http://ja.wikipedia.org/wiki/日本の出版社一覧

参考文献

[1] 荒内 大貴，太田 学，高須 淳宏，安達 淳：CRF による参考文献文字列からの書誌要素抽出の一手法，WebDB Forum 2011 4G-2 (2011).

[2] Councill, Isaac G., Giles, C. L. and Kan, Min-yen : ParsCit: An open-source CRF reference string parsing package, In *Proceedings of language resources and evaluation conference* (2008).

[3] 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).

[4] Peng, F. and McCallum, A. : Accurate Information Extraction from Research Papers using Conditional Random Fields, *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics* (2004).

[5] Okada, T., Takasu, A., and Adachi, J. : Bibliographic Component Extraction Using Support Vector Machines and Hidden Markov Models, *ECDL 2004*, *LNCS 3232*, pp.501-512 (2004).

[6] Lafferty, J., McCallum, A. and Pereira, F. : 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] Cortes, C. and Vapnik, V. : Support-Vector Networks, *Machine Learning*, pp.273-297 (1995).

[8] Seymore, K., McCallum, A., and Rosenfeld, R. : Learning hidden Markov model structure for information extraction, In *AAAI 99 Workshop on Machine Learning for Information Extraction* (1999).

[9] McCallum, A., Nigam, K., Rennie, J., Seymore, K. : Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval Journal*, volume 3, pages 127-163, Kluwer. (2000).

[10] Tjong Kim Sang, E. and Buchholz, S. : 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*, Vol.7, pp.127-132, Association for Computational Linguistics (2000).

[11] 工藤 拓, 山本 薫, 松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告，Vol.2004-NL-161，pp.89-96 (2004).

表 6 電子情報通信学会和文論文誌の書誌要素抽出精度

	AUTHOR	TITLE	JOURNAL	VOLUME	PUBLISHER	DAY	MONTH	YEAR	ETC	ALL
BEST-WebDBF2011	0.987	0.980	0.974	0.988	**0.967**	0.188	**0.999**	**0.998**	**0.916**	0.935
BIO-WebDBF2011	0.987	0.952	0.959	0.985	0.933	**0.300**	0.995	0.995	0.770	0.884
BIO	0.995	0.978	**0.979**	0.988	0.941	0.111	0.996	0.997	0.844	0.929
BIO(add dictionary)	**0.996**	**0.983**	0.978	**0.989**	0.966	0.000	0.996	**0.998**	0.887	**0.941**

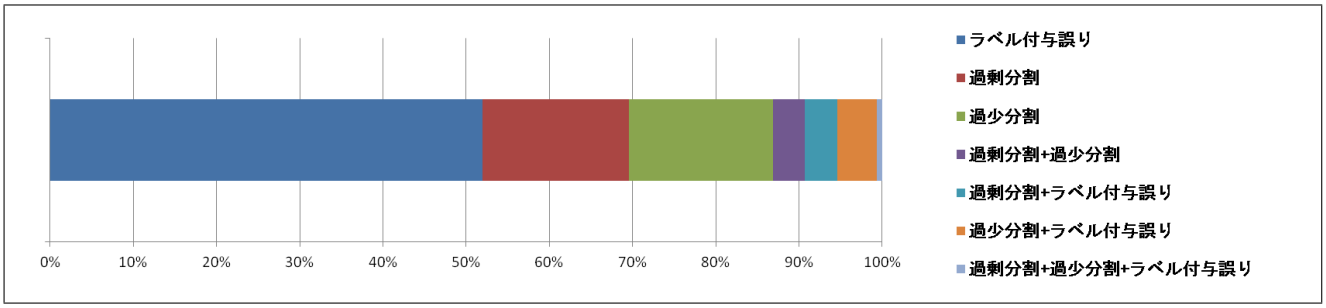

図 7 電子情報通信学会和文論文誌の抽出誤りの原因別の分類

表 7 電子情報通信学会和文論文誌におけるラベル付与状況（トークン数）

	AUTHOR	TITLE	JOURNAL	VOLUME	PUBLISHER	DAY	MONTH	YEAR	ETC
AUTHOR	12,827	6							
TITLE	27	4,620	37	1	5				
JOURNAL	4	22	3,995	3	6				6
VOLUME	1	1	4	13,994	1			2	8
PUBLISHER	5	1	11		696				8
DAY			1	1		1			
MONTH				2	1		1,655		
YEAR	1							4,581	
ETC	4	5	11	8	8				455

表 8 電子情報通信学会英文論文誌の書誌要素抽出精度

	AUTHOR	TITLE	JOURNAL	VOLUME	PUBLISHER	DAY	MONTH	YEAR	ETC	ALL
ParsCit	0.995	0.972	0.971	0.977	0.916	0.660	0.994	**0.998**	0.899	0.895
BIO	**0.996**	**0.980**	**0.981**	**0.984**	**0.935**	**0.809**	**0.998**	**0.998**	**0.907**	**0.925**

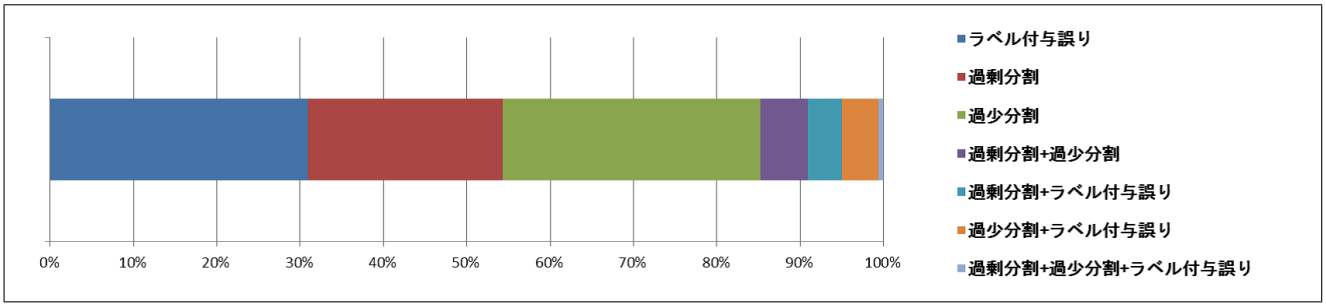

図 8 電子情報通信学会英文論文誌の抽出誤りの原因別の分類

表 9 電子情報通信学会英文論文誌におけるラベル付与状況（トークン数）

	AUTHOR	TITLE	JOURNAL	VOLUME	PUBLISHER	DAY	MONTH	YEAR	ETC
AUTHOR	10,141	5	1		1				1
TITLE	10	4,297	13	1	2				5
JOURNAL	1	8	3,721	4	3				2
VOLUME		1	1	11,531	1			1	8
PUBLISHER	1	3	10		815				10
DAY			1	5		84			
MONTH							1,692		
YEAR			1					4,214	
ETC	4	3	4	1	6		1		1,031