BERT による参考文献書誌情報抽出の誤り検出の評価

中山　竣平† 　金澤　輝一†† 　高須　淳宏†† 　上野　史††† 　太田　学†††

† 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
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2-1-2
††† 岡山大学学術研究院自然科学学域　〒 700–8530 岡山市北区津島中 3-1-1
E-mail: †pb8y7087@s.okayama-u.ac.jp, ††{tkana, takasu}@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし　電子図書館の文書間リンク等の実現において，学術論文の参考文献欄の著者名やタイトルなどの書誌情報は重要である．そのため，参考文献文字列からその書誌情報を自動抽出する研究が行われている．荒川らは Bidirectional Encoder Representations from Transformers (BERT) を利用して参考文献文字列から書誌情報を抽出する手法を提案し，実験によりその精度が 93.37%であることを示した．しかし自動抽出では誤りを完全になくすことは困難なため，書誌情報の抽出誤りを事後に人手で修正する必要があり，それは一般に高コストである．そこで本稿では，その修正コストの削減のため，参考文献書誌情報抽出の誤りの自動検出を試みる．実験では，荒川らの手法により参考文献文字列から書誌情報を抽出した後に，その抽出誤りを検出し，検出精度と見込まれる修正コストなどを評価する．

キーワード　書誌情報抽出，参考文献文字列，BERT

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る．特に参考文献へのリンクを付与するには，参考文献の著者名やタイトルといった書誌情報が必要となる．これらの書誌情報を人手で抽出するコストは膨大なため，機械学習により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．近年では，浪越ら [1] が Bi-directional LSTM-CNN-CRF，荒川ら [2] は Bidirectional Encoder Representations from Transformers (BERT) [3] を利用して英語論文の参考文献文字列から書誌情報を自動抽出する方法を提案した. [1] の参考文献書誌情報抽出精度は 92.77%, [2] の BERT による抽出精度は 93.37%となっている．しかし，抽出誤りを完全になくすことは困難であるため，最後は人手で誤りを修正する必要があるが，この作業は高コストである．そこで本研究は，この修正のための確認作業を削減するために，BERT による書誌情報抽出結果の誤りの自動検出を試みる．実験では抽出誤りの自動検出の精度を評価し，修正によって得られる書誌情報の精度とそれにかかる修正コストについて考察する．

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参考文献書誌情報抽出法について説明する．つづく 4 節で本研究で行う参考文献書誌情報抽出の誤りの自動検出について説明する．5 節で本研究で行う実験について説明し，その結果と考察を述べる．最後に 6 節でまとめる．

2 関連研究

多数の学術論文を収蔵する電子図書館では，書誌情報の管理は必須である．しかし，論文から人手で書誌情報を抽出をすることは高コストであるため，例えば，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．

機械学習による参考文献書誌情報抽出には，CRF [4] を用いた書誌情報抽出に関する研究が多く，例えば Peng ら [5], Councill ら [6]，Do ら [7]，Cuong ら [9] の研究がある．Peng ら [5] は学術論文のタイトルページと参考文献欄からそれぞれ Title，Author 等 13 項目の書誌情報を抽出し，実験では抽出の平均 F 値は 0.915 であった．Councill ら [6] は，大規模な特徴量の集合を利用するため訓練済みの CRF モデルを利用し，参考文献文字列から書誌情報を抽出する“ ParsCit ”を開発した．ParsCit は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．実験では 3 つのデータセットの参考文献文字列から書誌情報を抽出しており，Cora [8] データセットを対象にした抽出実験では，各書誌要素についての抽出の平均 F 値は 0.950 であった．Do ら [7] は CRF により著者とその所属機関を識別する情報抽出システムである Enlil を開発した．3 つのデータセットを対象に著者名を抽出し，その所属機関を照合する実験を行った．ACM Digital Library を対象とした著者名の抽出では F 値 0.946 であり，著者とその所属機関の照合における F 値は 0.889 であった．また，Cuong ら [9] は CRF を拡張した higher order semi-Markov CRFs (HO-SCRFs) を用いて参考文献文字列から書誌情報を抽出し，合計 1,384 件の参考文献文字列を対象に，Councill らと同様に 13 項目の書誌情報を抽出する実験を行ったところ，平均 F 値は 0.943 であった．

近年では，ニューラルネットワークによる参考文献書誌情報抽出の研究も行われている．Prased らの開発した Neural ParsCit [10] は，単語の分散表現と単語を構成する文字の分散表現を Bi-directional LSTM [14] へ入力し，その出力を CRF に入力することで，書誌要素を予測する．Cora [8] データセットに対する実験の結果，その平均 F 値は 0.914 であった．浪

越ら [1] は Bi-directional LSTM-CNN-CRF を利用して論文中の参考文献文字列から書誌情報を抽出する実験を行い，書誌情報抽出精度は 0.928 であった．荒川ら [2] は BERT を利用して同様に参考文献書誌情報を抽出し，その書誌情報抽出精度は 0.934 であった．

参考文献書誌情報抽出誤りの自動検出の研究では，荒内ら [11] が CRF で参考文献書誌情報を抽出した結果から，トークンに付与される書誌要素ラベルの周辺確率を用いた 3 種類の確信度を利用して，抽出誤りを自動検出する手法を提案した．実験により情報処理学会論文誌の参考文献文字列では，自動抽出後にその 35.7%を人手で確認して修正することで，99%の抽出精度となることを示した．

3 BERT による参考文献書誌情報抽出

3.1 参考文献書誌情報抽出

本研究の参考文献書誌情報抽出では，図 1 に示す処理によって参考文献文字列から著者名やタイトルといった書誌情報を抽出する．図 1 に示すように，参考文献文字列が書誌情報抽出器に入力されると，参考文献文字列をデリミタで分割して得られるワード列に書誌情報 BI ラベルを付与する．この書誌情報 BI ラベルとは，先頭とそれ以外を分ける BI ラベルと書誌情報の種類を組み合わせたものである．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「書誌要素 B」，先頭以外にあれば「書誌要素 I」というラベルを付与する．これらを書誌要素 BI ラベルと呼ぶ．同様に，ワードがデリミタの先頭に該当すれば「デリミタ B」，先頭以外にあれば「デリミタ I」というラベルを付与する．これらをデリミタ BI ラベルと呼ぶ．この書誌要素 BI ラベルとデリミタ BI ラベルを合わせたものが書誌情報 BI ラベルである．

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 に示す．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．また，抽出するデリミタの一覧と，それに対応するデリミタラベルを表 2 にまとめる．本研究の参考文献書誌情報抽出は，図 1 の結合処理によって書誌情報 BI ラベルを結合し，書誌情報を抽出する．すなわち，個々の書誌要素やデリミタの塊を表すチャンクとその書誌情報の種類を同時に推定している．

3.2 BERT

BERT [3] は大量の教師なしデータで事前学習し，少量の教師ありデータでファインチューニングすることで多様なタスクに対応できる自然言語処理モデルである．本研究では，英語 Wikipedia と Bookcorpus [13] で事前学習した $\text{BERT}_{\text{BASE}}$ を，教師データとして書誌情報 BI ラベルが付与された参考文献文字列によりファインチューニングして，参考文献文字列から書誌情報を抽出する．$\text{BERT}_{\text{BASE}}$ は Multi-Head Attention を利用した Transformer [12] のエンコーダが 12 層重なり構成されている．Transformer では入力を前後の双方向から読み込むことで入力間の関係を重みとして計算し，ベクトルを更新する．

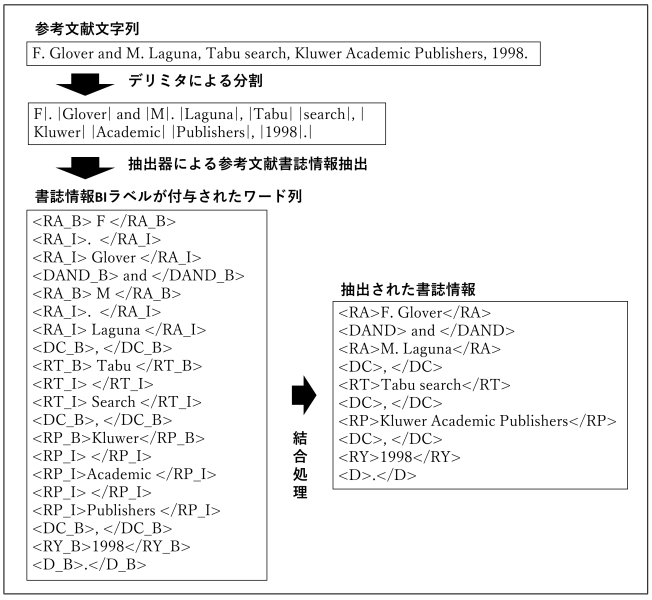

図 1　本研究における参考文献書誌情報抽出の処理

表 1　抽出する書誌要素とそのラベル [2]

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

これにより複雑な文脈や単語間の関係を考慮することができる．参考文献文字列では，大文字と小文字の違いも書誌情報抽出の手がかりとなるため，本研究では大文字と小文字を区別する bert-base-cased を使用する．BERT による参考文献書誌情報抽出の様子を図 2 に示す．BERT への入力は参考文献文字列を表 2 のデリミタで区切ったワード列，出力は表 1 と表 2 で示したラベルに BI ラベルを付加した書誌情報 BI ラベルである．BERT の入力の先頭にある [CLS] はワード列の先頭につける特殊なトークンである．[CLS] トークンの出力は分類タスクの集約シーケンス表現として使用されるが，本研究で行う参考文献書誌情報抽出では必要ないため取り出さない．なお，図 2 の中の Trm は Transfomer を表している．

表 2　抽出するデリミタ [2]

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., editors, 訳, 編, 編著, 監修, 監訳, 編集, （訳）, （編）, （編著）, （監）, （監修）, （監訳）, （邦訳）, （共訳）	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.	DN
Nos., nos.	DNS
pp.	DPP
p.	DP
:, ;（コロン，セミコロン）	DCL
/（スラッシュ）	DSL
-（ハイフン）	DHY
(, [, {（各種半角括弧・始）	DLBR
((全角括弧・始）	DZLBR
),], }（各種半角括弧・終）	DRBR
)（全角括弧・終）	DZRBR
その他	DUN

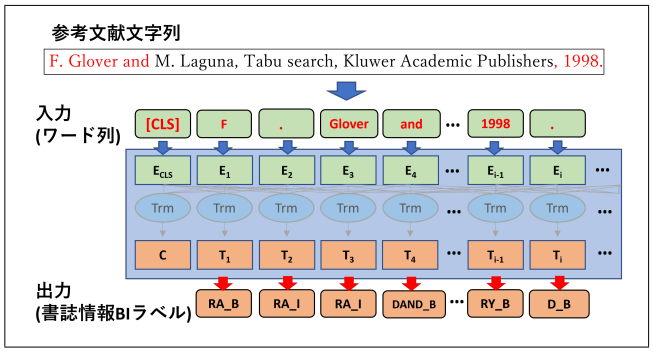

図 2　BERT による参考文献書誌情報抽出の概略

4　参考文献書誌情報抽出の誤り検出

図 2 では T_i が BERT の最終層を表している．3.2 節で説明したようにこの BERT は書誌情報 BI ラベルを出力するが，最終層は書誌情報 BI ラベルの予測確率を保持している．そのため，その確率が最大である書誌情報 BI ラベルを推定結果として出力する．図 3 を使って BERT による参考文献書誌情報抽出誤りの自動検出について説明する．図 3 は本稿で用いる BERT の最終層の例である．図 3 に示すように，BERT の最終層には参考文献文字列の各ワードに対する書誌情報 BI ラベルとその予測確率がある．例えば，先頭のワードである“ F ”では，Author の先頭を表す書誌情報 BI ラベルである RA_B の予測

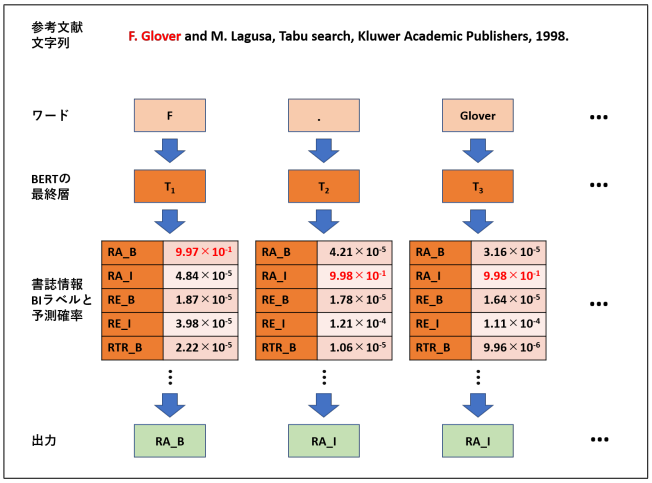

図 3　参考文献書誌情報抽出を行う BERT の最終層とその出力

確率が 0.997 で最大となっているため，BERT の出力は RA_B となっている．本研究では，この確率が低いワードは誤りの書誌情報 BI ラベルが付与されている可能性が高いと考え，この最大の確率が閾値より低ければ，推定を誤っている可能性が高い参考文献文字列として抽出する．具体的には書誌要素 BI ラベルが付与されたワードのその予測確率が 1 つでも閾値よりも低ければ，その参考文献文字列を検出する．ただし，表 2 のデリミタ BI ラベルが付与されたワードとその予測確率は誤りの検出に利用しない．

5　評 価 実 験

5.1　実験の諸条件

4 節で説明した方法による参考文献書誌情報抽出の誤りの自動検出精度を検証するため，評価実験を行う．実験データとして，2000 年の電子情報通信学会英文論文誌 (IEICE-E) に含まれる参考文献文字列 4,497 件，1952 年から 2012 年までの IEEE Trans. Computers (IEEE-CS) に含まれる参考文献文字列の引用回数上位 4,770 件の 2 つを利用する．これらは荒川らが [2] で使用したデータセットであるが，IEICE-E では参考文献文字列 36 件の書誌情報に誤りがあったため，本稿ではそれを修正して利用する．

実験ではまず，BERT による参考文献書誌情報抽出を行う．次に，その抽出誤りを 4 節で説明した方法で自動検出する．本研究では参考文献文字列中の書誌要素を構成する全てのトークンに正しい書誌要素ラベルを付与した場合，その参考文献文字列からの書誌情報抽出に成功したと判定する．また，抽出した書誌要素の正解判定は類似している書誌要素をまとめた表 3 の大分類で行う．つまり，この大分類が同じであれば正解と判定する．なおデリミタのトークンの誤りは無視する．参考文献書誌情報抽出誤りの自動検出では，書誌要素と推定されたワードの一つでも閾値以下であれば抽出誤りを含む可能性が高い参考文献文字列として検出する．

5.2　BERT による参考文献書誌情報抽出実験

ここでは 3.2 節で説明した荒川らの BERT を用いた手法に

表 3　抽出する書誌要素とその大分類 [2]

書誌要素 (書誌要素ラベル)	評価実験における大分類
Author (RA)，Editor (RE) Translator (RTR)，Author Other (RAOT)	AUTHOR
Title (RT)，Booktitle (RBT)	TITLE
Journal (RW)，Conference (RC)	JOURNAL
Volume (RV)，Number (RN)，Page (RPP)	VOLUME
Publisher (RP)	PUBLISHER
Day (RD)	DAY
Month (RM)	MONTH
Year (RY)	YEAR
Location (RL)，URL (RURL) , Other (ROT)	OTHER

表 4　bert-base-cased のハイパーパラメータ

パラメータ	値
max sequence length	300
training batch size	4
evaluation batch size	64
warmup propotion	0.1
training rate	5.0×10^{-4}
weight decay	1.0×10^{-2}

表 5　BERT の抽出精度及び誤りの総数

データセット	IEICE-E	IEEE-CS
書誌情報抽出精度	0.9508	0.8935
書誌要素の推定誤りを含む 参考文献文字列数	221	508
書誌要素の推定誤りを含まない 参考文献文字列数	4,276	4,262

より参考文献文字列から書誌情報を抽出する．参考文献書誌情報抽出に利用する BERT のハイパーパラメータを表 4 にまとめる．なお，書誌情報抽出精度を 5 分割交差検証で算出するため，各データセットの参考文献文字列を 5 つに分割し，そのうち 4 つをファインチューニングデータ，残りの 1 つをテストデータとする．

表 5 に IEICE-E と IEEE-CS からの BERT による参考文献書誌情報抽出の精度と，書誌要素の推定誤りを含む参考文献文字列数と含まない参考文献文字列数を示す．表 5 に示す通り，IEICE-E では書誌情報抽出精度が 0.9508，IEEE-CS で 0.8935 となったため，IEICE-E のほうが 5.73 ポイント高い．

5.3　書誌情報抽出結果の誤りの分析

IEICE-E と IEEE-CS の参考文献文字列データセットからの書誌情報抽出結果をそれぞれ表 6 と表 7 にまとめる．表 6 は IEICE-E の結果で，表 7 が IEEE-CS の結果である．これらの表では第 1 列が正解の大分類，第 1 行が推定した大分類の書誌要素で，ワード数は 5 分割交差検証で使用した 5 つのテストデータの合計である．IEICE-E で最も多い誤りは JOURNAL を TITLE と間違える誤りで 198 ワードあり，次に TITLE を JOURNAL と間違える誤りが 155 ワードあった．表 7 の IEEE-CS では最も多い誤りは TITLE を JOURNAL と間違える誤りで 410 ワードあり，次に JOURNAL を TITLE と間違える誤りが 378 ワードあった．このように，TITLE と JOURNAL に関

J. Lilius	Author
High-level nets, and linear logic	Title
Lecture Notes in Computer Science	Conference
616	Volume
310-327	Page
June	Month
1992	Year

M. Jantzen and R. Valk	Author
Formal properties of place/transition nets	Title
Lecture Notes in Computer Science	Journal
84	Volume
168-205	Page
1980	Year

A. Valmari	Author
Stubborn sets for reduced state space generation	Title
Lecture Notes in Computer Sience	Booktitle
483	Volume
491-515	Page
Springer-Verlag	Publisher
1990	Year

図 4　IEICE-E の正解の書誌要素に揺れがある参考文献文字列の例

連する誤りが多い．この一因として学習データの正解ラベルに揺れがあることがあげられる．図 4 に IEICE-E のデータの一部を示す．図 4 では“Lecture Notes in Computer Science”という同じ文字列に対して Conference，Journal，Booktitle の 3 つの書誌要素ラベルが付与されている．BERT はこれらのデータでファインチューニングするため，このような正解の揺れがあると推定を間違えやすい．なお，表 3 に示したように Conference と Journal の大分類が JOURNAL であり，Booktitle の大分類が TITLE である．

5.4　参考文献書誌情報抽出の誤り検出実験

4 節で説明した方法で，参考文献書誌情報抽出の誤りを含む可能性が高いと判断する参考文献文字列を自動検出する．5.1 節で説明したデータセットにおいて，書誌要素と推定されたワードの予測確率の閾値を 0.55〜0.95 の間で 0.05 ずつで変えて抽出誤りを検出する．IEICE-E の検出結果を図 5 に，IEEE-CS の検出結果を図 6 に示す．図 5 と図 6 には以下に定義する抽出誤り検出の再現率と適合率，その調和平均である F 値を示している．

再現率

参考文献書誌情報抽出誤りを含む参考文献文字列のうち検出できた参考文献文字列の数

適合率

検出した参考文献文字列のうち参考文献書誌情報抽出誤りを含む参考文献文字列の数

IEICE-E の結果を示す図 5 からわかるように，閾値を上げればあげるほど，参考文献書誌情報抽出誤りの参考文献文字列の検出数が増えるため，再現率は単調増加する．一方で，適合

表 6　IEICE-E の各書誌要素の推定結果のワード数 (第 1 列: 正解ラベル，第 1 行: 推定ラベル)

	AUTHOR	TITLE	JOURNAL	VOLUME	PUBLISHER	DAY	MONTH	YEAR	OTHER
AUTHOR	50,702	23	6	0	29	1	0	0	36
TITLE	23	76,630	198	6	34	0	0	1	59
JOURNAL	1	155	33,800	17	78	0	0	3	31
VOLUME	0	16	32	12,141	0	2	1	2	36
PUBLISHER	31	0	52	0	3,859	0	0	0	26
DAY	0	0	0	2	0	226	0	0	0
MONTH	4	1	0	2	2	0	2,826	0	1
YEAR	1	0	3	4	0	0	0	4,443	1
OTHER	14	22	82	26	39	0	2	1	3,466

表 7　IEEE-CS の各書誌要素の推定結果のワード数 (第 1 列: 正解ラベル，第 1 行: 推定ラベル)

	AUTHOR	TITLE	JOURNAL	VOLUME	PUBLISHER	DAY	MONTH	YEAR	OTHER
AUTHOR	60,266	80	20	1	77	0	0	0	8
TITLE	38	72,453	378	13	29	0	0	0	38
JOURNAL	12	410	42,892	127	148	0	0	5	42
VOLUME	0	33	128	12,066	0	8	0	4	34
PUBLISHER	30	25	145	2	6,086	0	0	0	48
DAY	0	0	0	28	0	27	0	0	1
MONTH	0	0	1	0	0	0	3,097	0	2
YEAR	0	0	4	7	0	1	1	4,720	0
OTHER	2	57	69	91	60	0	6	2	5,510

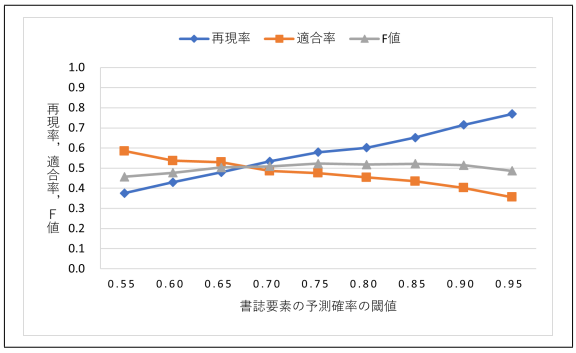

図 5　IEICE-E における抽出誤りの検出精度

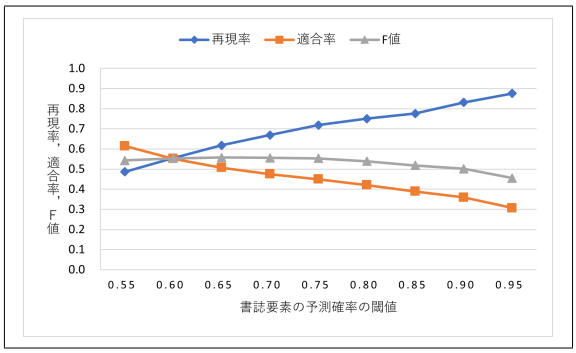

図 6　IEEE-CS における抽出誤りの検出精度

率は閾値を上げると単調減少しているため，抽出誤りを含まない参考文献文字列の誤検出も増加することがわかる．F 値は閾値 0.75 の時が最大で，0.5223 だった．

IEEE-CS の結果を示す図 6 からは，IEICE-E と同様に再現率は単調増加で，適合率は単調減少であることがわかる．F 値は閾値 0.65 の時が最大で 0.5571 だった．

次に，このようにして検出した参考文献文字列を人手で確認し，その中に参考文献書誌情報抽出誤りを含む参考文献文字列があればそれを修正したと仮定して参考文献書誌情報抽出精度を算出する．IEICE-E のその修正後の抽出精度と修正コストに相等する検出された参考文献文字列数を図 7 に，IEEE-CS の修正後の抽出精度と修正コストを図 8 に示す．これらの図にお

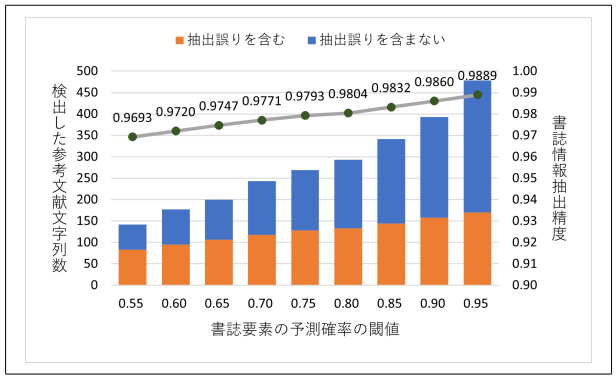

図 7　IEICE-E における検出した参考文献文字列数とその誤りを修正した場合の書誌情報抽出精度

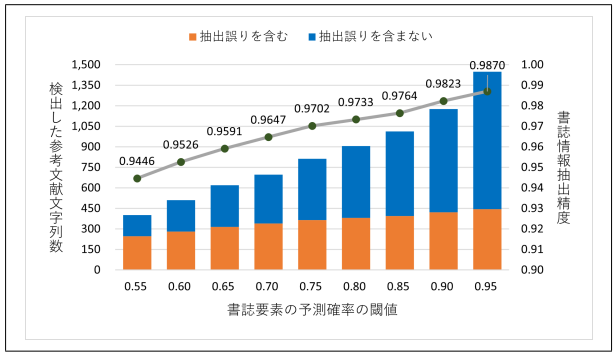

図 8　IEEE-CS における検出した参考文献文字列数とその誤りを修正した場合の書誌情報抽出精度

ワード	正解ラベル	推定ラベル	予測確率
vol.	DV_B	DV_B	0.9968
33	RV_B	RV_B	0.9959
,_	DC_B	DC_B	0.9987
no.	DN_B	DN_B	0.9966
4	RN_B	RN_B	0.9956
_	DSP_B	DSP_B	0.9965
550	RN_B	RN_B	0.2975
_	RN_I	RN_I	0.4120
284	RN_I	RN_B	0.3778
,_	DC_B	DC_B	0.9148
Oct	RM_B	RM_B	0.9958
.	RM_I	RM_I	0.9950
_	DSP_B	DSP_B	0.9957
1985	RY_B	RY_B	0.9965
.	D_B	D_B	0.9961

ワード	正解ラベル	推定ラベル	予測確率
Technical	RW_B	RW_B	0.9743
_	RW_I	RW_I	0.9792
Report	RW_I	RW_I	0.9800
_	DSP_B	DSP_B	0.9793
97	RV_B	RV_B	0.3294
/	DSL_B	DSL_B	0.9823
12	RN_B	RN_B	0.4552
_	DSP_B	DSP_B	0.9641
of	DUN_B	DUN_B	0.9624
_	DSP_B	DSP_B	0.9483
IRIDIA	RP_B	RP_B	0.9244
,_	RP_I	DC_B	0.9468
Universite	RP_I	RP_B	0.9535
_	RP_I	RP_I	0.9449
Libre	RP_I	RP_I	0.9428
_	RP_I	RP_I	0.9503
de	RP_I	RP_I	0.9529
_	RP_I	RP_I	0.9506
Bruxelles	RP_I	RP_I	0.9783

図 9　書誌情報抽出における予測確率が低いが正しく書誌情報が抽出されている参考文献文字列の例

いて，棒グラフは検出した参考文献文字列数を表しており，抽出誤りを含む参考文献文字列数と抽出誤りを含まない参考文献文字列数を色分けしている．一方，折れ線グラフは検出した抽出誤りを修正したとみなしたときの書誌情報抽出精度である．また，書誌情報抽出精度は 5 分割交差検証で求めた値で，参考文献文字列数は，5 つのテストデータの合計である．

図 7 と図 8 から，予測確率の閾値が上がるにつれて検出した参考文献文字列数は増えるが，青色である書誌情報抽出の誤りを含まない参考文献文字列の割合も増加していることがわかる．IEICE-E で抽出誤り検出の F 値が最大であった閾値 0.75 では，参考文献文字列 269 件を抽出誤りとして検出した．これは参考文献文字列全体の 5.9%である．それを人手で確認し，抽出誤りを修正したとみなしたときの書誌情報抽出精度は 0.9793 となる．IEEE-CS で F 値が最大であった閾値 0.65 では，参考文献文字列 614 件を抽出誤りとして検出した．これは参考文献文字列全体の 12.9%である．それを人手で確認し，抽出誤りを修正したとみなしたときの書誌情報抽出精度が 0.9591 となる．再現率では，2 雑誌とも閾値 0.95 のとき最大となった．その時見込まれる修正コストは検出した参考文献文字列数で，IEICE-E は参考文献文字列 478 件，IEEE-CS は参考文献文字列 1,449 件であり，これらは参考文献文字列全体のそれぞれ 10.6%，30.4%であった．その時検出した参考文献文字列に含まれる抽出誤りを修正したとみなしたときの書誌情報抽出精度は，IEICE-E で 0.9889，IEEE-CS で 0.9870 であった．

5.5 考　　察

5.4 節で説明したように抽出誤りの検出数は，書誌要素の予測確率の閾値を上げれば増加するが，それに対して抽出誤りを含まない参考文献文字列の検出数の増加の割合が大きい．そのため，高い閾値を設定するならば，抽出誤りを含まない参考文献文字列を減らすことが重要である．そこで，閾値を高くした場合に抽出誤りを含まない参考文献文字列が多く検出された理由について考察する．図 9 に誤り検出結果の一部を示す．図 9 には参考文献文字列のワードと正解ラベル，予測ラベル，そのラベルの予測確率を示している．黄色に表示しているワードがほかの書誌要素に比べて予測確率が低い．これらは，Number や Volume の書誌要素だが，近くにデリミタの“ No.”や“ Vol.”などがない．このような数字のみからなるワードは，正しく推定されていても予測確率が閾値より低くなることが多かった．

6 ま と め

本稿では，BERT による参考文献書誌情報抽出において，その抽出誤りの自動検出を行い，実験によりその検出の精度ならびに見込まれる修正コストなどを評価した．実験では，IEICE-E と IEEE-CS の 2 雑誌の参考文献文字列から書誌情報を抽出したところ，その書誌情報抽出精度はそれぞれ 0.9508，0.8530 だった．次に，書誌要素と推定された各ワードの予測確率に閾値を設定して，書誌情報抽出誤りを含む可能性が高いと判断する参考文献文字列を自動検出した．検出精度を表す F 値は，IEICE-E では閾値 0.75 のとき最大の 0.5223，IEEE-CS では閾値 0.65 のとき，最大の 0.5571 となった．抽出誤りの検出では，2 雑誌とも閾値 0.95 のとき最大となった．その時見込まれる修正コストは検出した参考文献文字列数で，IEICE-E は参考文献文字列 478 件，IEEE-CS は参考文献文字列 1,449 件であった．これらは参考文献文字列全体のそれぞれ 10.6%，30.4%であった．また，閾値 0.95 のとき，検出した参考文献文字列に含まれる抽出誤りを修正したとみなしたときの書誌情報抽出精度は，IEICE-E で 0.9889，IEEE-CS で 0.9870 であった．

今後の課題としては，参考文献文字列からの書誌情報抽出結果により適切な確信度を付与して誤りを検出することなどがあげられる．

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 22H03904)，同基盤研究 (C)(課題番号 18K11989)，新エネルギー・産業技術総合開発機構 (NEDO) の戦略的イノベーション創造プログラム (SIP) 第二期「ビッグデータ・AI を活用したサイバー空間基盤技術」および 2022 年度国立情報学研究所共同研究 (22FC01) の援助による．

文　　献

[1] 浪越大貴, 太田学, 高須淳宏, 安達淳, “Bi-directional LSTM-CNN-CRF による参考文献書誌情報抽出,” 信学技報, vol. 118, no. 377, pp. 17-22, 2018.

[2] 荒川瞭平，金澤輝一，高須淳宏，上野史，太田学，“BERT による参考文献書誌情報抽出における擬似学習データの有効性の評価，” ARG 第 17 回 WI2 研究会予稿集, pp. 25-28, 2021.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. of NAACL-HLT, pp. 4171-4186, 2019.

[4] J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data, ” in Proc. of ICML 2001, pp. 282-289, 2001.

[5] F. Peng, and A. McCallum, “Accurate Information Extraction from Research Papers Using Conditional Random Fields, ” in Proc. of HLT-NAACL 2004, pp. 329-336, 2014.

[6] I. G. Councill, C. L. Giles, and M. Y. Kan, “ParsCit: an open-source CRF reference string parsing package,” in Proc. of LREC 2008, pp. 661-667, 2008.

[7] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho, and M. Y. Kan, “Extracting and matching authors and affiliations in scholarly documents, ” in Proc. of JCDL 2013, pp. 219-228, 2013.

[8] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, “Automating the construction of internet portals with machine learning,” In Proc. of Information Retrieval, vol. 3, no. 2, pp. 127-163, 2000.

[9] N. V. Cuong, M. K. Chandrasekaran, M. Y. Kan, and W. S. Lee, “Scholarly Document Information Extraction using Extensible Features for Efficient Higher Order Semi-CRFs,” in Proc. of JCDL 2015, pp. 61-64, 2015.

[10] A. Prased, M. Kaur, and M. Y. Kan, “Neural ParsCit: a deep learning based reference string parser, ” International Journal on Digital Libraries (IJDL), vol. 19, no. 4, pp. 323-337, 2018.

[11] 荒内大貴，“CRF による参考文献文字列からの自動書誌情報抽出に関する研究，” 岡山大学大学院自然科学研究科修士論文, 2013.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All You Need,”in Proc. of NeurIPS, vol. 30, pp. 5998-6008, 2017.

[13] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in Proc. of the IEEE international conference on computer vision, pp.19–27. 2015.

[14] X. Ma, and E. Hovy, “End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF,” in Proc. of ACL 2016, Association for Computational Linguistics, pp. 1064-1074, 2016.