DEIM2022 G43-5(day4 p15)

BERT による和文の参考文献文字列からの書誌情報抽出の評価

高橋　春成† 　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: †ps4v8vmm@s.okayama-u.ac.jp,††{tkana, takasu}@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし　膨大な文書が格納されている電子図書館の運用には，書誌情報データベースが必須である．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ため，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．また荒川らは，Bidirectional Encoder Representations from Transformers（BERT）を利用して，英文論文を対象に参考文献書誌情報抽出を行い，Conditional Random Field（CRF）など他のモデルの抽出精度を上回った．そこで本研究では，Huggingface が提供する訓練済み日本語 BERT モデルを利用して，和文論文の和文の参考文献文字列から書誌情報を抽出し，実験により評価する．3 種類のトークナイザを使用した実験の結果，BertJapaneseTokenizer を使用してから荒内らのトークナイザを使用したとき最も書誌情報抽出精度が高くなった．

キーワード　書誌情報抽出，ニューラルネットワーク，BERT，参考文献文字列

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．一方で，学術論文の参考文献欄には，多くの文献の著者やタイトルといった書誌情報が記述されている．この参考文献の書誌情報を利用すれば，文書間リンクなどが可能となるため，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．

荒内ら [1] は Conditional Random Field（CRF）を利用して，和文と英文の参考文献文字列から書誌情報を自動で抽出した．CRF は学習データを用意すれば利用できるが，有効な素性を見つけるのは難しい．そこで，荒川ら [2] はニューラルネットワークである BERT を利用して，英文論文中の参考文献文字列から書誌情報を自動で抽出した．本研究では和文論文に対応するため，BERT を利用して和文の参考文献文字列から書誌情報を自動で抽出する手法を提案する．本研究では，Huggingface が提供する訓練済み日本語 BERT モデルを利用する．この BERT への入力は，参考文献文字列を変換したトークン列である．本研究ではこの参考文献文字列のトークン列への変換に，Huggingface に付属するトークナイザ，荒内らのトークナイザ [1] を用いる．

本稿の構成は次の通りである．まず，2 節で関連研究を紹介し，3 節で本研究で提案する和文の参考文献文字列からの自動書誌情報抽出法について説明する．つづく 4 節で評価実験を示し，その結果について考察する．最後に 5 節で本稿をまとめる．

2 関連研究

2.1 BERT

Devlin ら [3] は自然言語処理モデルである Bidirectional Encoder Representations from Transformers（BERT）を提案した．BERT の特徴は，文章を文頭と文末の双方向から学習することで，前後の文脈を考慮することができることである．BERT は公開時，2 つの文が意味的に同じかを判定する Microsoft Research Paraphrase Corpus（MRPC）タスクのような 11 の自然言語処理タスクにおいて最高精度を記録した．

BERT の学習は大量の教師なしデータによる事前学習と，比較的少量の教師ありデータでのファインチューニングに分けられる．BERT は，Masked Language Model（MLM）と Next Sentence Prediction（NSP）の 2 つのタスクで事前学習を行う．MLM は，入力された文章からランダムに選ばれた 15%のトークンを [MASK] という特殊トークンに置き換え，[MASK] トークンに置き換えられたトークンを予測するタスクである．NSP は，2 つの文を受け取り，受け取った文が連続した文であるかを予測するタスクである．事前学習を行った後，比較的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，多様なタスクを扱うことができる．

2.2 書誌情報抽出に関する研究

Peng ら [4] は CRF を用いて学術論文の参考文献欄とタイトルページから書誌情報を抽出した．参考文献欄から書誌情報を抽出する実験では，500 件の学術論文を対象に実験を行った．タイトルページから書誌情報を抽出する実験では，935 件の学術論文を対象に実験を行った．参考文献欄からの書誌情報抽出における各書誌要素の平均 F 値は 0.915 であった．タイトルページからの書誌情報抽出における各書誌要素の平均 F 値は

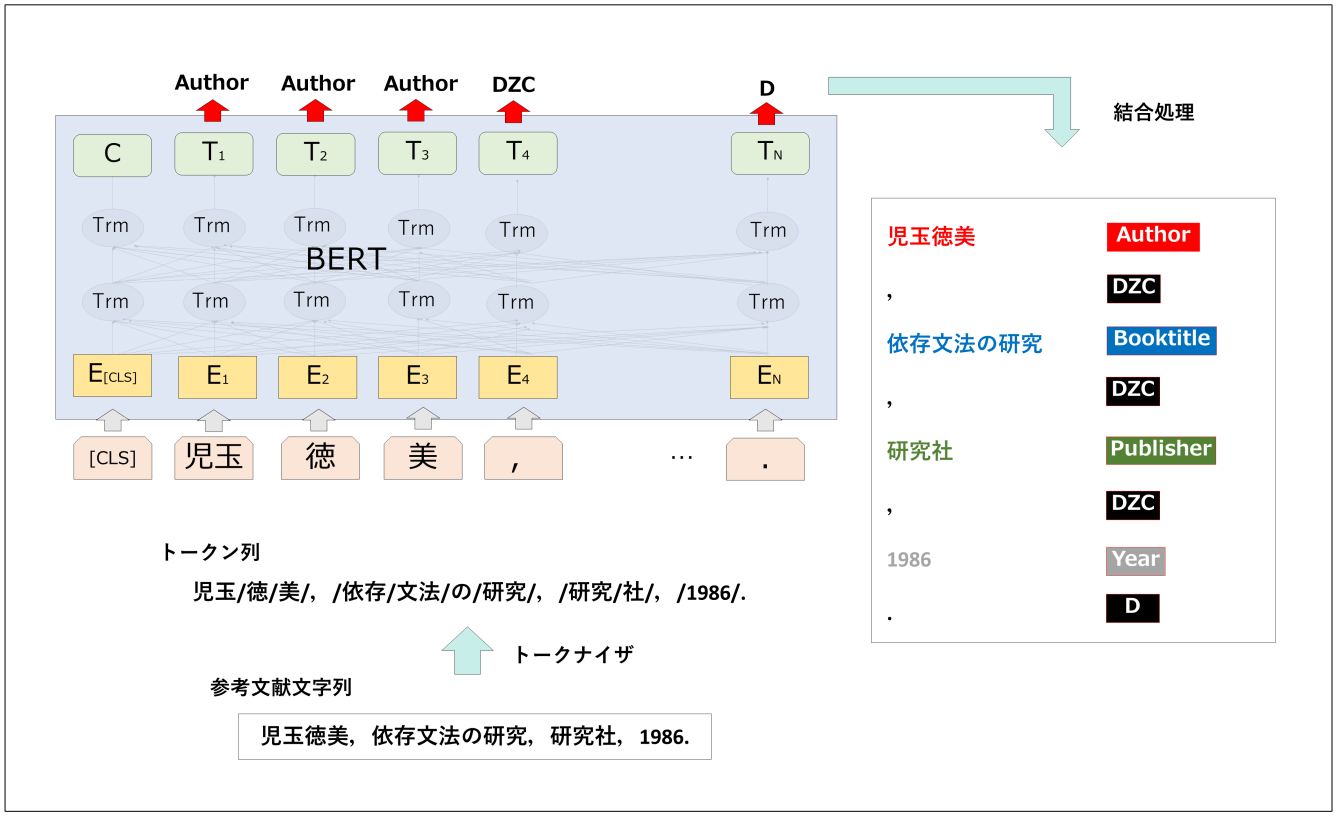

図 1　BERT による書誌情報抽出

0.939 であった．

Council ら [5] は学術論文の参考文献欄を認識し，CRF を用いて参考文献文字列から書誌情報を抽出するツールである“ ParsCit ”を開発した．Cora データセット [6] を対象にした実験では，各書誌要素抽出の平均 F 値は 0.950 だった．

Do ら [7] は CRF を用いて学術論文の PDF から著者とその所属機関の書誌情報を抽出するツールである“ Enlil ”を開発した．ACL Anthology，ACM Digital Library，Cross Disciplinary Corpus を対象に実験を行った．著者名抽出の F 値は ACL Anthology で 0.918，ACM Digital Library で 0.946，Cross Disciplinary Corpus で 0.916 であった．所属機関抽出の F 値は ACL Anthology で 0.836，ACM Digital Library で 0.889，Cross Disciplinary Corpus で 0.870 であった．

Cuong ら [8] 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た．1,384 件の参考文献文字列を対象にした実験では，各書誌要素抽出の平均 F 値は 0.943 であった．

荒内ら [1] は CRF を用いて学術論文の参考文献文字列を書誌要素やデリミタに対応するトークンの列に分割し，各トークンに書誌要素ラベルまたはデリミタラベルを付与した．2000 年の電子情報通信学会和文論文誌の参考文献文字列を対象にした実験では，和文のトークン化の精度が 0.9302，英文のトークン化の精度が 0.9037，分割されたトークンに書誌要素ラベルを付与する精度は 0.9410 だった．

浪越ら [9] は Bi-directional LSTM-CNN-CRF を用いて参考文献文字列中のトークンとその書誌要素ラベルを同時に推定した．2000 年の電子情報通信学会英文論文誌を対象にした実験では，92.77%の参考文献文字列から全ての書誌要素を正しく抽出し，デリミタを書誌要素として抽出しなかった．

荒川ら [2] は BERT を用いて参考文献文字列中のトークンとそれに対応する書誌要素を同時に推定した．2000 年の電子情報通信学会英文論文誌を対象にした実験では，93.37%の参考文献文字列から全ての書誌要素を正しく抽出した．

3　和文の参考文献文字列からの書誌情報抽出

3.1　和文論文の参考文献文字列

和文論文の参考文献欄には，通常和文の参考文献文字列と英文の参考文献文字列が混在している．本研究で使用する BERT は，利用する事前学習済みモデルによって対応する言語が異なる．そのため，和文の参考文献文字列と英文の参考文献文字列を区別し，本研究では和文の参考文献文字列のみを扱う．

和文の参考文献文字列と英文の参考文献文字列の例を以下に示す．

- 児玉徳美，依存文法の研究，研究社，1986.
- H.G. Ruck, ”The Tate pairing on elliptic curves,” ECC’98, Waterloo, 1998.

和文の参考文献文字列は，英文の参考文献文字列と異なる特徴を持つ．例えば，英文の参考文献文字列は，書誌要素を区切るカンマ等の記号の他にも，空白文字で単語に分割されている．一方和文の参考文献文字列は，空白文字によって単語に分割されていないため，単語に分割するには形態素解析などを行う必要がある．

3.2 BERT による参考文献書誌情報抽出

BERT を用いて参考文献文字列から書誌情報を抽出する例を図 1 に示す．図 1 の E_i が入力系列，T_i が出力系列，Trm が Transformer [10] を表す．

提案手法では，まず参考文献文字列をトークナイザを用いて分割し，トークン列を得る．そして，このトークン列を BERT へ入力する．図 1 の例は，BertJapaneseTokenizer を使用している．

BERT は各トークンに対して，トークンが書誌要素に該当すれば書誌要素ラベルを付与し，トークンがデリミタに該当すればデリミタラベルを付与する．その後提案手法は，同じラベルが付与された連続するトークンを結合し，書誌要素またはデリミタを得る．

本研究で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 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名などが含まれる．デリミタとそれに対応するデリミタラベルは表 2 の通りである．

本研究では，書誌要素ラベル，デリミタラベルが付与された和文の参考文献文字列により BERT をファインチューニングし，書誌情報を抽出する．また本研究で用いる BERT モデルは，Huggingface が提供する訓練済み日本語 BERT モデルである [12]．これは日本語 Wikipedia によって事前学習されている．

3.3 参考文献文字列のトークナイザ

参考文献文字列のトークナイザとして，BertTokenizer，BertJapaneseTokenizer，荒内らのトークナイザ [1] を使用する．これらのトークナイザが参考文献文字列を分割する例を図 2 に示す．

BertTokenizer は，Huggingface の BERT で文章をトークン列に変換するために用意されているライブラリである．BertTokenizer により和文の参考文献文字列をトークン化すると，日本語の多くが文字に分割される．これは過分割ともいえる．

BertJapaneseTokenizer は，Huggingface の BERT で日本語の文章をトークン列に変換するために用意されているライブラリである．BertJapaneseTokenizer による和文の参考文献文字列のトークン化では，日本語はおおよそ意味のある単語に分割される．

荒内らのトークナイザは，表 2 に定義したデリミタを用いてトークンに分割するトークナイザである．図 2 に示すように，荒内らのトークナイザによる和文の参考文献文字列のトークン化では，おおよそ各書誌要素やデリミタがトークンに対応している．

また，BertJapaneseTokenizer による和文の参考文献文字列のトークン化では，書誌要素とデリミタの境界で分割されないことがある．そこで本研究では，BertJapaneseTokenizer を使用した後に，荒内らのトークナイザを使用する．つまり，BertJapaneseTokenizer でのトークン分割に加えて，さらに表 2 のデリミタでも分割する．

表 1　書誌要素と書誌要素ラベル [11]

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　デリミタとデリミタラベル [11]

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+空白文字）， and （空白文字+and+空白文字）	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

4 評価実験

4.1 実験環境

提案手法を用いて和文の参考文献文字列から書誌情報を抽出

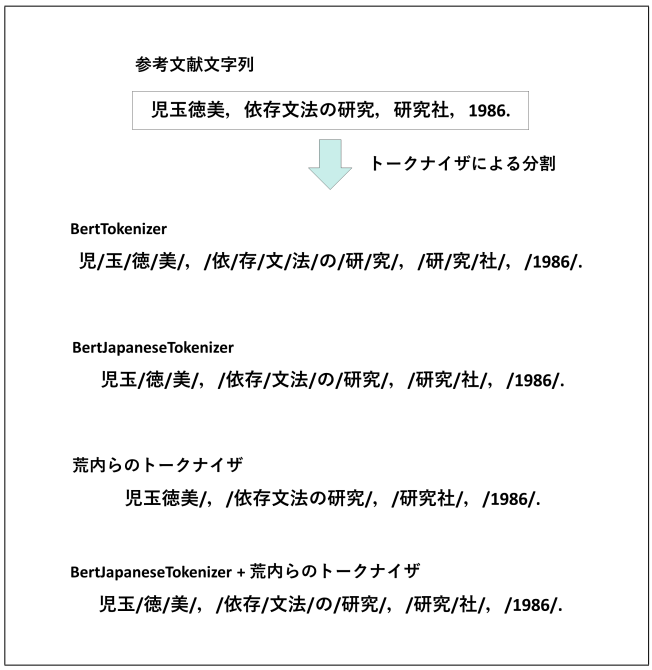

図 2　トークナイザによる参考文献文字列の分割

し，その抽出精度を検証する．

実験データは，2000 年の電子情報通信学会和文論文誌（IEICE-J）に含まれる参考文献文字列のうち，和文の参考文献文字列 2,193 件である．

本研究では参考文献文字列の全ての書誌要素を過不足なく正しく推定した場合，その参考文献文字列の書誌要素推定に成功したと判定する．評価の際，表 1 の 18 種類の書誌要素は，表 3 に示す書誌要素の大分類にまとめ，この大分類で区別する．つまり，大分類が同じ書誌要素は正解判定において区別しないため，例えば“Author”を“Editor”と推定しても正解とする．なおデリミタは正解判定において区別しない．

正解判定の例を図 3 に示す．この図の例 1 は書誌要素推定に成功した例である．書誌要素“Author”が書誌要素“Editor”に，書誌要素“Booktitle”が書誌要素“Title”になっているが，大分類が同じ書誌要素は，正解判定において区別しない．また，デリミタの“DZC”がデリミタの“DC”，“DED”，“DV”と推定されているが，デリミタは正解判定において区別しない．例 2 は書誌要素推定に失敗した例である．デリミタ“DZC”を書誌要素“Journal”と推定している．この場合，推定結果に余分な書誌要素が生じる．そのため，書誌要素推定に失敗したと判定される．実験結果に示す書誌情報抽出精度は，推定に成功した参考文献文字列の全体に対する割合である．また，書誌情報抽出精度は 5 分割交差検定で算出し，学習データのうちの 4 分の 1 を検証データとしてファインチューニングを行う．

また，書誌要素の大分類ごとの抽出精度の評価は再現率（Recall），適合率（Precision），F 値（F-measure）で示す．それぞれ以下の式で定義する．

表 3　抽出する書誌要素の大分類 [13]

書誌要素（書誌要素ラベル）	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

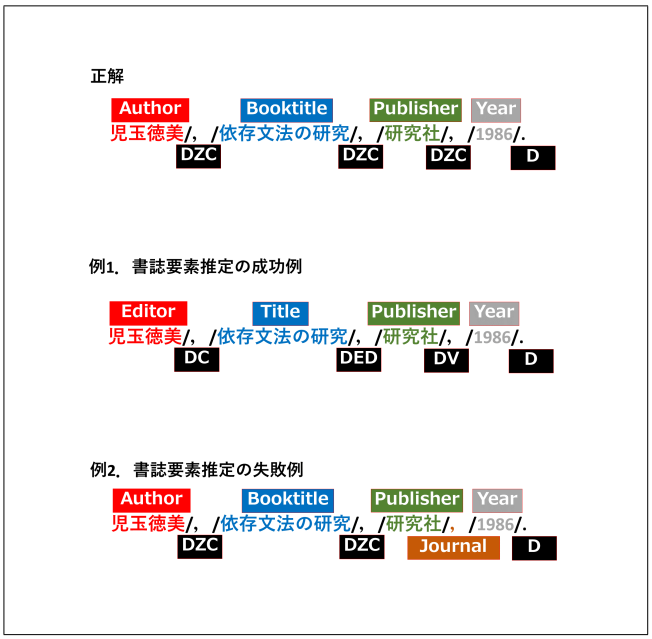

図 3　正解判定の例

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{F-measure} = \frac{2\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3}$$

書誌要素“AUTHOR”を例に説明すると以下のようになる．

- TP は書誌要素“AUTHOR”と推定された書誌要素“AUTHOR”の数
- FP は書誌要素“AUTHOR”と推定された“AUTHOR”以外のものの数
- FN は書誌要素“AUTHOR”と推定されなかった書誌要素“AUTHOR”の数

よって再現率は，全ての書誌要素“AUTHOR”に対し，実際に得ることができた書誌要素“AUTHOR”の割合である．適合率は，書誌要素“AUTHOR”と推定されたものが実際に書誌要素“AUTHOR”であった割合である．F 値は，再現率と

適合率の調和平均である．

4.2 実験結果

4.2.1 書誌情報抽出精度

トークナイザを変えて参考文献文字列から書誌情報を抽出した．その書誌情報抽出精度を表 4 に示す．3 種類のトークナイザを使った抽出精度を見ると，BertTokenizer が 80.08%で最も高かった．さらに，BertJapaneseTokenizer を使用した後に荒内らのトークナイザ [1] を使用すると，その書誌情報抽出精度は 81.26%になった．

4.2.2 書誌要素の大分類ごとの評価

BertTokenizer を使った時の，書誌要素の大分類ごとの抽出精度を表 5 に示す．他の書誌要素と比べて“JOURNAL”，“PUBLISHER”の F 値が低いことがわかる．さらに，“OTHER”の F 値が低くなっている．また，実験データの IEICE-J の和文の参考文献文字列中に“ DAY ”は 1 件だけだったため，“ DAY ”の抽出精度は算出しない．

BertJapaneseTokenizer を使った時の，書誌要素の大分類ごとの抽出精度を表 6 に示す．他の書誌要素と比べて“ JOURNAL ”，“ PUBLISHER ”の F 値が低いことがわかる．さらに，“ OTHER ”の F 値が低くなっている．また“ JOURNAL ”以外は，BertTokenizer の F 値と似た値となった．

荒内らのトークナイザを使った時の，書誌要素の大分類ごとの抽出精度を表 7 に示す．他の書誌要素に比べて，“ TITLE ”，“ JOURNAL ”，“ PUBLISHER ”の F 値が低いことがわかる．また，BertTokenizer，BertJapaneseTokenizer を使った時よりも，“ OTHER ”の F 値が比較的高い結果であった．

BertJapaenseTokenizer を使用してから荒内らのトークナイザを使用した時の，書誌要素の大分類ごとの抽出精度を表 8 に示す．他の 3 つのトークナイザを使った時よりも，“ JOURNAL ”の F 値が高い結果となった．また“ JOURNAL ”以外は，BertJapaneseTokenizer による F 値と似た値となった．また“ VOLUME ”，“ PUBLISHER ”，“ OTHER ”以外は，荒内らのトークナイザによる F 値よりも高かった．

4.3 考　　察

4.3.1 BertJapaneseTokenizer のトークン化

BertJapaneseTokenizer の推定誤りの原因として，このトークナイザは参考文献文字列を分割すべき場所で分割しない場合があることが挙げられる．この原因による推定誤りの例を図 4 に示す．図 4 では「信学論 (D-II)」を書誌要素“ Journal ”と推定するべきところを,「信学論 (D-II」を書誌要素“ Journal ”と推定している．これは BertJapaneseTokenizer が「),」を「)」と「,」に分割しなかったことが原因である．

表 4　使用したトークナイザと書誌情報抽出精度

トークナイザ	抽出精度（%）
BertTokenizer	80.08
BertJapaneseTokenizer	55.13
荒内らのトークナイザ	70.18
BertJapaneseTokenizer+荒内らのトークナイザ	81.26

表 5　書誌要素の大分類ごとの抽出精度（BertTokenizer）

	再現率	適合率	F 値
AUTHOR	0.9788	0.9763	0.9775
TITLE	0.9582	0.8891	0.9217
JOURNAL	0.9392	0.8304	0.8805
VOLUME	0.9846	0.9814	0.9830
PUBLISHER	0.9324	0.6622	0.7739
DAY	-	-	-
MONTH	0.9945	0.9706	0.9823
YEAR	0.9986	0.9949	0.9968
OTHER	0.0877	0.2250	0.1148

表 6　書誌要素の大分類ごとの抽出精度（BertJapaneseTokenizer）

	再現率	適合率	F 値
AUTHOR	0.9783	0.9777	0.9780
TITLE	0.9558	0.9092	0.9317
JOURNAL	0.6355	0.5826	0.6077
VOLUME	0.9833	0.9789	0.9811
PUBLISHER	0.9446	0.6436	0.7649
DAY	-	-	-
MONTH	0.9936	0.9720	0.9825
YEAR	0.9986	0.9958	0.9972
OTHER	0.0433	0.3000	0.0706

表 7　書誌要素の大分類ごとの抽出精度（荒内らのトークナイザ）

	再現率	適合率	F 値
AUTHOR	0.9614	0.9665	0.9640
TITLE	0.8882	0.8098	0.8468
JOURNAL	0.8655	0.7992	0.8309
VOLUME	0.9798	0.9833	0.9815
PUBLISHER	0.8773	0.7687	0.8124
DAY	-	-	-
MONTH	0.9849	0.9455	0.9647
YEAR	0.9968	0.9909	0.9938
OTHER	0.4871	0.9378	0.5834

表 8　書誌要素の大分類ごとの抽出精度（BertJapaneseTokenizer+荒内らのトークナイザ）

	再現率	適合率	F 値
AUTHOR	0.9800	0.9800	0.9800
TITLE	0.9627	0.9195	0.9402
JOURNAL	0.9427	0.8631	0.9007
VOLUME	0.9828	0.9762	0.9795
PUBLISHER	0.9522	0.6194	0.7488
DAY	-	-	-
MONTH	0.9961	0.9734	0.9843
YEAR	0.9977	0.9959	0.9968
OTHER	0.0529	0.1867	0.0821

BertJapaneseTokenizer を使った実験では，書誌要素推定に失敗した参考文献文字列のうち 73.27%の参考文献文字列で，書誌要素とデリミタの境界や隣接する異なるデリミタ間の境界で分割していなかった．

4.3.2 未定義語のトークン

荒内らのトークナイザでトークン列に変換すると，4.3.1 項

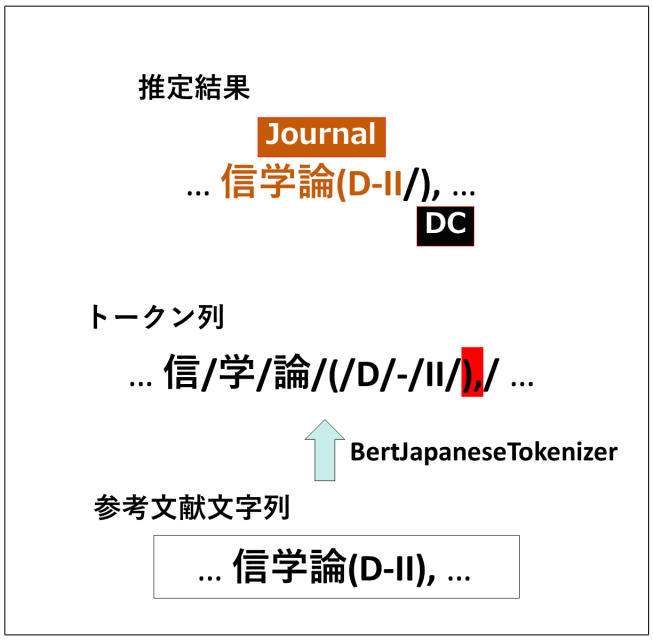

図 4　不適切なトークン化による誤推定の例

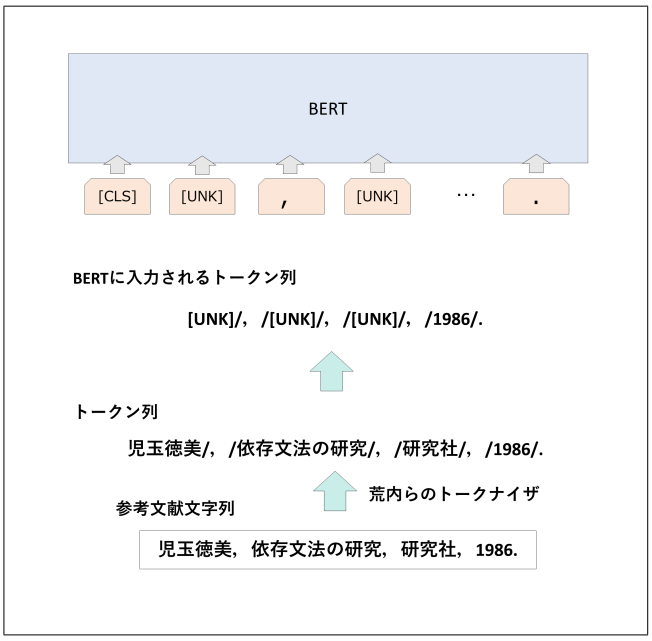

図 5　BERT に未定義語が入力される例

で述べたような不適切なトークン化は見られない．しかし，荒内らのトークナイザは表 2 のデリミタで分割するだけなので，一つのトークンが長い文字列になることがしばしばある．その場合，単語より大きなトークンは BERT に未定義語として入力される．トークンが未定義語として BERT に入力される例を図 5 に示す．

図 5 の [UNK] が未定義語のことである．この例では，「児玉徳美」，「依存文法の研究」，「研究社」の 3 つのトークンが未定義語として BERT に入力されている．これにより，これらの文字列の情報が失われる．

4. 3. 3　出現頻度の低いデリミタ

推定誤りの原因として，出現頻度の低いデリミタが挙げられる．この原因による誤推定は，3 つ全てのトークナイザで，また BertJapaneseTokenizer と荒内らのトークナイザを併用した場合でも発生した．この原因による誤推定の例を図 6 に示す．参考文献文字列に記述されている書誌要素はデリミタで区切られており，その多くは「,」や「.」のような記号，「pp.」のような英語の文字列である．これらのデリミタは使用頻度が高く，多くの参考文献文字列で見られる．しかし図 6 の「編集」のようなデリミタは，使用頻度が低いため，正しくデリミタと推定できなかった．

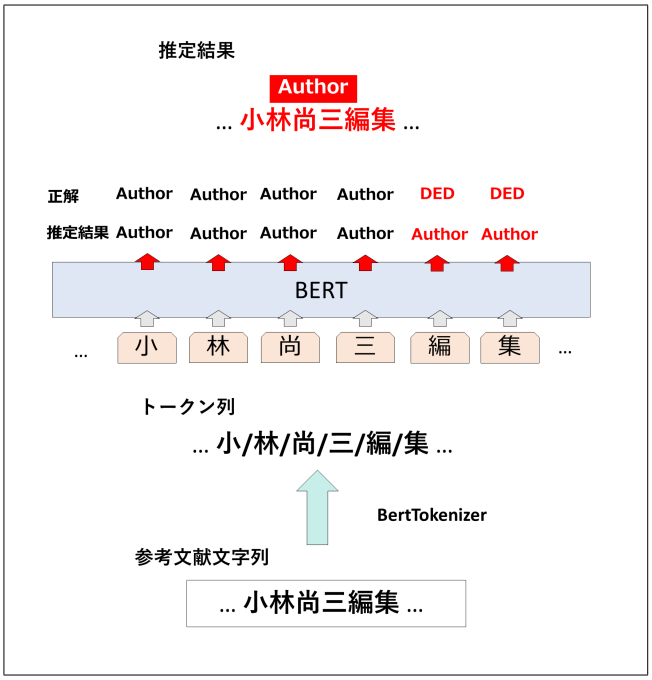

図 6　出現頻度の低いデリミタによる誤推定の例

5　ま　と　め

本稿では，BERT を用いて学術論文の和文の参考文献文字列から書誌情報を抽出する手法を提案し，実験により評価した．提案手法ではまず，参考文献文字列をトークナイザによりトークン列に変換する．そして BERT により，各トークンに対して，トークンが書誌要素に該当すれば 18 種類からなる書誌要素ラベルを付与し，デリミタに該当すれば 24 種類からなるデリミタラベルを付与する．その後，同じラベルが付与された連続するトークンを結合し，書誌要素またはデリミタを得る．

実験では，3 種類のトークナイザを用いて 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のうち和文の参考文献文字列 2,193 件を対象に書誌情報を抽出をした．その結果，3 種類のトークナイザの中では，BertTokenizer を使った場合の書誌情報抽出精度が 80.08%で最も高かった．また，BertJapaneseTokenizer を使用した後に，荒内らのトークナイザを使用すると，書誌情報抽出精度が 81.26%となることを確認した．

なお本研究の実験に用いた和文の参考文献文字列に加えて英文の参考文献文字列を含むデータに対して，荒内らは CRF 抽出器による書誌情報抽出を行い，87.53%の書誌情報抽出精度を

達成している．この抽出精度は本稿で提案した手法よりも高いため，今後和文と英文の参考文献文字列の違いをさらに精査して，提案した BERT による抽出法を改善していきたい．

謝　　辞

本研究の一部は，科学研究費補助金基盤研究 (C)(課題番号 18K11989)，および新エネルギー・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 2021 年度国立情報学研究所公募型共同研究 (21FC04) の援助による．

文　　献

[1] 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データベースシステム（DBS）, vol. 2012-DBS-156, no. 1, pp. 1-8, 2012.

[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の精度向上,” 情報処理学会 第 83 回全国大会, 1S-06, 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] 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.

[5] 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.

[6] A. K. 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.

[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] 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.

[9] 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,” 信学技法, vol. 118, no. 377, pp. 17-22, 2018.

[10] 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.

[11] 川上尚慶, “少量学習データによる参考文献書誌情報の自動抽出に関する研究,” 岡山大学大学院自然科学研究科修士論文, 2015.

[12] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” in Proc. of EMNLP 2020, Association for Computational Linguistics, pp. 38-45, 2020.

[13] 荒川瞭平, 太田学, 金澤輝一, 高須淳宏, “少量学習データと Bi-directional LSTM-CNN-CRF による参考文献書誌情報抽出,” DEIM 2020, F2-1, 2020.