# BERT による参考文献書誌情報抽出の誤り分析

Error analysis of bibliographic information extraction from reference strings using BERT

中山 竣平
Shunpei Nakayama
岡山大学 太田研究室
Ohta Laboratory, Okayama University

**概要** 学術論文の参考文献文字列には著者名やタイトルなどの有用な書誌情報が含まれており，学術論文を取り扱う電子図書館では，これらの書誌情報は検索や文書館リンク等の機能を実現する上で不可欠である．そのため，荒川らは，BERT を利用し参考文献文字列から著者や論文名などの書誌情報を抽出し，93.37%の書誌情報抽出精度を実現した．本稿では，この抽出精度の改善などのため，荒川らの方法による参考文献書誌情報抽出の誤りを調査して分析する．

## 1 はじめに

学術論文は日々増え続けている．多くの学術論文を取り扱う電子図書館では，検索や文書間リンク等の機能の実現のために，著者名やタイトルといった書誌情報のデータベースを必要とする．例えば参考文献の書誌情報が分かれば，文献を同定してその文献へのリンクが生成できる．しかし，日々増え続ける学術論文から正確な書誌情報抽出を人手で抽出することは難しいため，書誌情報を自動で抽出する技術が求められている．そこで荒川ら[1]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した．本稿では，荒川らの参考文献書誌情報抽出法で実際に参考文献書誌情報抽出を行い，その抽出結果の誤りを分析する．

## 2 BERT

BERT は 2018 年に Google が発表した自然言語処理モデルで，11 の自然言語処理タスクで最高スコアを達成した[2]．それ以前の自然言語処理モデルは文章を文頭からの事前学習するものが多かったが，BERT は文頭と文末の双方向から事前学習するように設計されている．また，Wikipedia や BookCorpus などから得た大量の教師なし文章データを事前学習したモデルが公開されている[3]．また，BERT は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様々なタスクに応用できる．BERT の事前学習は Masked Language Model (MLM)と Next Sentence Prediction (NSP)の 2 つのタスクより行われる．これらのタスクの詳細を 2.1 節，2.2 節で述べる．

### 2.1 Masked Language Model (MLM)

MLM[2]は図 1 のように文章から特定のトークン(文字列)を 15%ランダムに選び，MASK というトークンに置き換え，元のトークンを当てるタスクである．図 1 の例では，”Multiuser”と”University”が MASK トークンに置換されている．このタスクでは単語間の関係を BERT の双方向性

元の参考文献文字列
P. Stoica, Multiuser Detection, Kyoto University, 1977.

MASK 処理後
P. Stoica, [MASK] Detection, Kyoto [MASK], 1977.

**図 1：MLM における参考文献文字列のマスク処理**

を利用し，予測を間違えた単語を減らすよう学習する．さらに，MASK するトークンに対して以下の 3 つの条件を設けている[2].

① 80%の確率で[MASK]トークンに置き換える
② 10%の確率でランダムな別のトークンに置き換える
③ 10%の確率で置き換えない

### 2.2 Next Sentence Prediction (NSP)

NSP[2]は，文同士の関係性を考慮するために，ランダムに 2 文を選択し，それらが連続した文であるかどうかを当てる 2 クラス分類タスクである．BERT の入力表現図を図 2 に示す．Input が入力シーケンスを表しており，先頭に[CLS]，2 文の末尾に[SEP]というトークンを挿入して連結させた文を入力としている．また，Token Embeddings はトークンの ID を，Segment Embeddings は 2 文の区切れを，Position Embeddings は単語の入力文内での位置を表している．また，入力される文は図 2 より“dog”，“is”，“cute”とのように単語に分割される． 図 2 の例で NSP は，“dog is cute”と“he likes playing”が連続した文かどうかを判定する．図 3 に実際の参考文献文字列の分割の例を示す．NSP の入力は 2 文をつなげた 1 文であるので，本研究では 1 つの参考文献文字列の中で，ワード数が最も均等になるような位置にあるデリミタを探索し，そこで参考文献文字列を分割し 2 文にする．ここでワードとは，参考文献の区切りであるカンマやコロン，ピリオドなどの 24 種類のデリミタを用いて，参考文献文字列を分割して得られる文字列のことである．図 3 では Multiuser Detection の後の半角カンマで参考文献文字列が分割されている．

## 3 参考文献書誌情報抽出

参考文献書誌情報抽出は，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ことである．本研究では，参考文献文字列内の各ワードに書誌要素ラベルを付

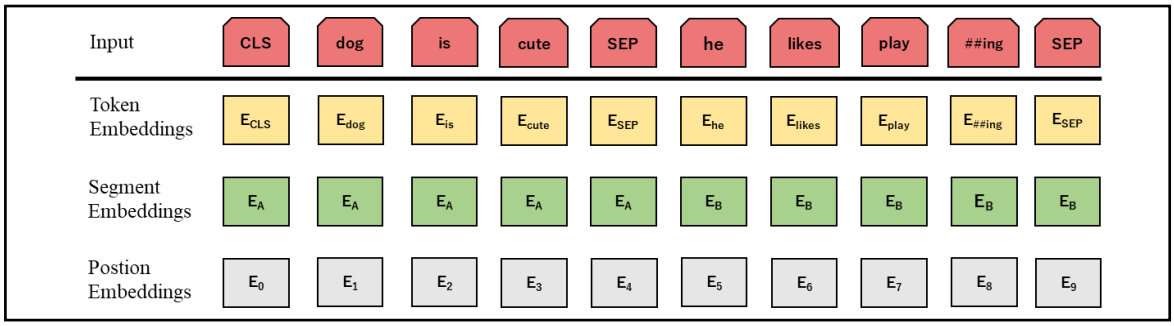


**図 2：BERT の入力表現図**

元の参考文献文字列
P. Stoica, Multiuser Detection, Kyoto University, 1977.

NSP タスクへ入力するため分割した参考文献文字列
P. Stoica, Multiuser Detection,
Kyoto University, 1977.

**図 3：NSP タスクにおける参考文献文字列分割**

与することで抽出する．書誌要素の抽出精度の算出は，荒川ら[1]と同様に，18 種類の書誌要素の似ている書誌要素をまとめ，9 種類に再分類して行う．抽出する書誌要素と評価のための再分類を表 1 にまとめる．表 1 で“Other”は他の書誌要素には分類できない書誌要素を表しており，具体的には著者の所属機関などが含まれる．さらに，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“書誌要素 B”，先頭以外であれば“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．連続する同じ書誌要素の B と I のラベルがついたワードを組み合わせて書誌要素を抽出する．なお，書誌要素の抽出精度での算出では，デリミタの正誤は無視する．

## 4　参考文献書誌情報抽出実験

### 4.1　実験概要

Book Corpus(8 億語)[4]と English Wikipedia(25 億語)により事前学習された BERT を，書誌要素とデリミタの正解ラベル付き参考文献文字列でファインチューニングして，以下の参考文献文字列から書誌情報を抽出した．

**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

なお，書誌情報抽出精度を 5 分割交差検定で算出するため，IEICE-E 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．ファインチューニングの epoch 数は 3 とする．

### 4.2　実験結果と誤りの分析

まず，抽出した書誌要素毎の再現率と適合率は表 2 のようになった．表 2 から AUTHOR の再現率と適合率は最も高く 0.99 以上であるが，PUBLISHER の適合率は 0.8922 で最も低いことがわかる．また，再現率では OTHER が 0.9292 で最も低いことがわかる．

**表 1：書誌要素と評価におけるその分類**

| 書誌要素 | 評価における分類 |
|---|---|
| Author | AUTHOR |
| Editor | |
| Translator | |
| Author Other | |
| Title | TITLE |
| Booktitle | |
| Journal | JOURNAL |
| Conference | |
| Volume | VOLUME |
| Number | |
| Page | |
| Publisher | PUBLISHER |
| Day | DAY |
| Month | MONTH |
| Year | YEAR |
| Location | OTHER |
| URL | |
| Other | |

次に書誌情報抽出結果の混同行列を表 3 に示す．縦は正解の書誌要素，横は BERT が推定した書誌要素を表す．表 3 に示す書誌要素の抽出誤りは計 214 件あり，特に，JOURNAL を TITLE と間違えたのが 32 件，JOURNAL を PUBLISHER と間違えたのが 30 件と多かった．次に TITLE を JOURNAL と間違えたのが 24 件で，JOURNAL が他の書誌要素に比べて推定が困難であることが分かる．

図 4 に参考文献書誌情報抽出誤りの例を示す．黒字の部分が正しく推定できた部分で，赤字の部分が推定を誤った部分である．抽出結果にある D から始まるラベルはデリミタであり，D はピリオド，DC は半角カンマ+空白文字，DS は二重引用符，DV は”vol.”，DPP は”pp.”を表す．この例では，赤字の部分は Title の一部であるのに，Conference と推定されたが，タイトル内の“one million”が論文タイトルの区

表 2：各書誌要素の再現率と適合率

| 書誌要素 | AUTHOR | TITLE | JOURNAL | VOLUME | PUBLISHER | DAY | MONTH | YEAR | OTHER |
|---|---|---|---|---|---|---|---|---|---|
| 再現率 | 0.9965 | 0.9941 | 0.9792 | 0.9943 | 0.9733 | 0.9523 | 0.9891 | 0.9921 | 0.9292 |
| 適合率 | 0.9981 | 0.9939 | 0.9856 | 0.9943 | 0.8922 | 0.9090 | 0.9836 | 0.9954 | 0.9510 |

表 3：書誌情報抽出結果の混同行列

| | AUTHOR | TITLE | JOURNAL | VOLUME | PUBLISHER | DAY | MONTH | YEAR | OTHER |
|---|---|---|---|---|---|---|---|---|---|
| AUTHOR | 4982 | 12 | 0 | 0 | 5 | 0 | 0 | 0 | 0 |
| TITLE | 5 | 8276 | 24 | 1 | 9 | 0 | 0 | 1 | 9 |
| JOURNAL | 2 | 32 | 3440 | 2 | 30 | 0 | 0 | 0 | 7 |
| VOLUME | 1 | 0 | 5 | 2301 | 1 | 2 | 0 | 1 | 3 |
| PUBLISHER | 0 | 0 | 8 | 2 | 439 | 0 | 0 | 0 | 2 |
| DAY | 0 | 0 | 0 | 1 | 0 | 40 | 1 | 0 | 0 |
| MONTH | 0 | 0 | 0 | 1 | 1 | 0 | 360 | 1 | 1 |
| YEAR | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 883 | 3 |
| OTHER | 1 | 5 | 13 | 5 | 7 | 1 | 4 | 1 | 486 |

参考文献文字列
K. Sugihara and M. Iri, "Construction of the Voronoi diagram for "one million" generators in single-precision arithmetic," Proc. IEEE, vol.80, pp.1471-1484, 1992.

書誌情報抽出結果
<Author> K. Sugihara and M. Iri</Author>
<DC>, </DC>
<DS>"</DS>
<Title> Construction of the Voronoi diagram for "one million</Title>
<DS>"</DS>
<Conference> generators in single-precision arithmetic," Proc. IEEE</Conference>
<DC>, </DC>
<DV>vol.</DV>
<Volume>80</Volume>
<DC>, </DC>
<DPP>pp.</DPP>
<Page>1471-1484</Page>
<DC>, </DC>
<Year>1992</Year>
<D>.</D>

図 4：参考文献書誌情報抽出誤りの例

切りによく使われる二重引用符で囲まれている．この二重引用符が原因で BERT が Conference と推定した可能性がある．このようなデリミタに関わる推定誤りは多かった．

## 5　まとめ

本稿では，荒川らの BERT による参考文献書誌情報抽出法で電子情報通信学会英文論文誌の参考部文献文字列を解析し，その抽出誤りについて分析した．実験結果より，参考文献文字列内の Journal に該当する書誌情報がほかの書誌情報と比べて誤りが多いことが分かった．また具体的な誤り事例の分析から，書誌要素の区切りではなくその一部となるようなデリミタがあると推定を誤ることが分かった．

今後の課題としては，BERT などによる書誌情報抽出誤りの自動検出に取り組みたい．

## 参考文献


[1] 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疑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. 25-28，2019.

[2] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K., “BERT: pre-training of deep bidirectional transformers for language understanding,” In Proc. of NAACL-HLT, pp. 4171-4186, 2019.

[3] Huggingface, 2022, "BERT – Huggingface," (Retrieved July 18, 2022. https://huggingface.co/docs/transformers/model_doc/bert

[4] Y. zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, " Alignings book and movies; Towards Story-Like Vosual Explanations by watching movies and reading books," In Proc. of IEEE-CV, pp. 19-27, 2015.