1S-06

BERT による参考文献書誌情報抽出の精度向上

荒川 瞭平† 金澤 輝一‡ 高須 淳宏‡ 上野 史† 太田 学†
岡山大学 大学院自然科学研究科† 国立情報学研究所‡

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．浪越ら[1]は，学術論文の参考文献文字列に着目し，Bi-directional LSTM-CNN-CRF[2]を利用して，論文中の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自動で抽出する方法を提案した．本研究では，自然言語処理の多くのタスクで良い性能を示している，Bidirectional Encoder Representations from Transformers(BERT)[3]を利用して書誌情報を抽出する．実験では，ファインチューニングおよび事前学習データの書誌情報抽出精度への影響を評価する．

2. 参考文献書誌情報抽出

2.1 問題定義とアプローチ

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主要な書誌情報を抽出することである．浪越らは参考文献文字列中のトークンと書誌要素を同時推定することで，参考文献文字列から高精度に書誌情報を抽出した[1]．本研究もそれにならう．

2.2 トークンと書誌要素の同時推定

本研究では，あらかじめ定義したデリミタを用いて，まず参考文献文字列をワード列に分解する．次にデリミタを含む各ワードに対して、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．これらを書誌情報 BI ラベルと呼ぶ．その後，連続する同じ種類の“B”から“I”を結合して，書誌要素トークンまたはデリミタトークンを得る．図 1 は実際に参考文献文字列のワード列に書誌情報 BI ラベルを付与し，ワードを結合して書誌要素とデリミタを抽出する例である．なお，書誌要素 BI ラベルは Author，Year 等の書誌要素ラベルを 18 種類，デリミタ BI ラベルはピリオド(.)，スラッシュ（/）等のデリミタに対応するラベルを

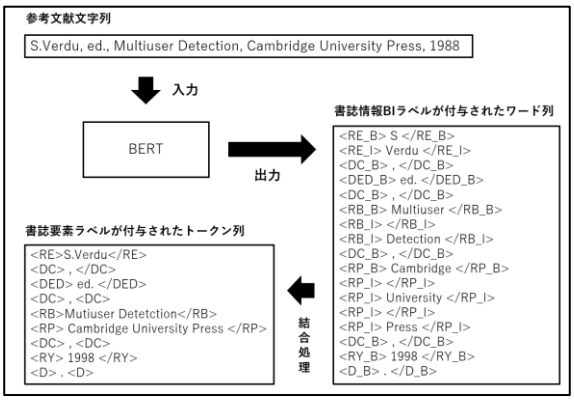

図 1 トークンと書誌要素の同時推定

24 種類定義している[1]．

2.3 BERT

本研究で用いる$BERT_{base}$は，約 1.1 億のパラメータを利用し，大規模コーパスにより事前学習された言語モデルであり，事前学習のタスクは Next Sentence Prediction(NSP)とマスク化言語モデル(MLM)である[3]．NSP では 2 つの文章が連続する文章であるか否かの 2 クラス分類を行う．MLM では入力トークンの一部をランダムにマスクし，マスクされたトークンを予測する．

3. BERT による書誌情報抽出

$BERT_{base}$は大量の教師無しデータで事前学習されており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，多様なタスクを扱える．本研究では，正しい書誌情報 BI ラベルが付与された抽出対象雑誌の参考文献文字列により，$BERT_{base}$をファインチューニングし，書誌情報を抽出する．

$BERT_{base}$ モデルは English Wikipedia[4] と BookCorpus[5]により事前学習された言語モデルである．一方参考文献文字列には，通常の文章には無い固有の特徴がある．そこで本研究では上記の大規模コーパスで事前学習した$BERT_{base}$モデルに加えて，Web で集めた参考文献文字列で事前学習して得られたモデルでも書誌情報を抽出する．

4. 評価実験

4.1 実験内容

BERT による書誌情報抽出の精度を検証する．実験データは，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 件である．また，18 種類の書誌要素は，

Improvement in Accuracy of Bibliography Extraction from Reference Strings in Academic Papers using BERT
Ryohei Arakawa† Teruhito Kanazawa‡ Atsuhiro Takasu‡ Fumito Uwano† Manabu Ohta†
† Graduate School of Natural Science and Technology, Okayama University
‡ National Institute of Informatics

評価の際には浪越らにならい，Title や BookTitle 等の似た種類のラベルをまとめて，AUTHOR, TITLE, JOURNAL, VOLUME, PUBLISHER, DAY, MONTH, YEAR, OTHER と再分類したものを用いる[1]．つまり再分類ラベルが同じものは正解判定において区別しない．本研究では参考文献文字列の書誌要素を構成する全てのワード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の書誌要素推定に成功したと判定し，その際デリミタの正解判定は行わない．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割合を表す．また，抽出精度を 5 分割交差検定で算出するため，雑誌中の参考文献文字列データセットを 5 つに分割し，そのうち 4 つをファインチューニングに，残りの 1 つをテストデータに用いる．

4.2 BERT による書誌情報抽出の有効性評価

3 節で説明した大規模コーパスにより事前学習された $\mathrm{BERT}_{\mathrm{base}}$ による抽出精度を，浪越らの Bi-directional LSTM-CNN-CRF と比較する．表 1 に書誌情報抽出の結果を示す．表 1 より IEICE-E で約 0.6 ポイント，IEEE-CS では約 2 ポイント Bi-LSTM の抽出精度を上回り，BERT の有効性が確認できた．

4.3 事前学習データの有効性評価

本節では，3 節で説明したように事前学習データを変更した際の，書誌情報抽出精度の変化を実験により検証する．事前学習に用いる参考文献文字列データは，dblp[6]から収集した参考文献文字列 460,754 件である．また，事前学習用データを 9,297 件，126,966 件とした場合の抽出精度も算出する．結果は表 2 のようになった．両雑誌とも事前学習データが多いほど抽出精度が向上した．また，両雑誌ともに書誌要素 VOLUME, YEAR の抽出精度が表 1 の BERT の結果に比べ大幅に低かった．

表 2 の実験から，dblp と IEEE では参考文献文字列の書式が異なっていることが精度に影響を与えているのではないかとの仮説を得た．そこで、事前学習用参考文献文字列(9,297 件)の書式を，IEEE-CS で多く用いられている書式に変換することで精度を改善できるか検証した．図 2 に変換前の参考文献文字列データと変換後の参考文献文字列の例を示す．変換したデータにより事前学習を行い，その後抽出対象雑誌でファインチューニングした際の書誌情報抽出精度は IEICE-E で 66.38%,

表 1　各モデルごとの書誌情報抽出精度(%)

	IEICE-E	IEEE-CS
BERT	93.37	89.44
Bi-LSTM	92.77	87.57

表 2　各事前学習データ件数ごとの書誌情報抽出精度(%)

	IEICE-E	IEEE-CS
9,297 件	58.35	42.31
126,966 件	63.51	42.39
460,754 件	64.58	43.34

図 2　参考文献文字列変換の例

IEEE-CS で 44.99%という結果になった．表 2 の実験と比較して両雑誌ともに最高だった 460,754 件の抽出精度を上回るという結果になった．また，9,297 件の書式を変換しない場合と比較して，IEICE-E については書誌要素 VOLUME で約 3 ポイント，YEAR で約 4 ポイント抽出精度が向上し，また IEEE-CS においても VOLUME, YEAR の両方で約 2 ポイントの向上がみられた．両書誌要素に該当する文字列は，変換前後で特に書式が異なっており，本手法は有効といえる．

5. 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，ファインチューニングおよび事前学習データの影響を実験により評価した．大規模コーパスにより事前学習された $\mathrm{BERT}_{\mathrm{base}}$ をファインチューニングした場合の書誌情報抽出精度は，浪越らの手法よりも高かった．また，大量の事前学習データと，その書式の変換が参考文献書誌情報抽出に有効であることが分かった．今後の課題としては，大規模コーパスにより事前学習されたモデルに対して，書式を変換した参考文献文字列を用いて追加で事前学習を行う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(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2020 年度国立情報学研究所公募型共同研究(20FC07)の援助による．

参考文献

[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," 信学技法, vol. 118, no. 377, pp. 17-22, 2018.
[2] X. Ma, and E. Hovy, ``End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in Proc. of 54th Annual Meeting of the Association for Computational Linguistics （ACL 2016）, pp. 1064-1074, Association for Computational Linguistics, 2016.
[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] WIKIPEDIA, https://en.wikipedia.org/wiki/Main_Page, (参照 2020-12-16)
[5] Y. Zhu, R. Kiros, R.Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, ``Aligning books and movies: To wards story-like visual explanations by watching movies and reading books," in Proc. of ICCV, pp. 19-27, 2015.
[6] dblp, https://dblp.uni-trier.de/ , (参照 2020-12-16)