学術ドメインに特化した日本語事前訓練モデルの構築

山内 洋輝 [1]　梶原 智之 [1]　桂井 麻里衣 [2]　大向 一輝 [3,4]　二宮 崇 [1]
[1] 愛媛大学　[2] 同志社大学　[3] 東京大学　[4] 国立情報学研究所
yamauchi@ai.cs.ehime-u.ac.jp　{kajiwara,ninomiya}@cs.ehime-u.ac.jp
katsurai@mm.doshisha.ac.jp　i2k@l.u-tokyo.ac.jp

概要

本研究では、日本語の学術文書を用いて RoBERTa および BART を事前訓練し、公開する。近年の自然言語処理では、事前訓練モデルの転移学習によって、様々なタスクの性能が向上している。特に、専門用語の扱いが重要となる医療や学術などのドメインにおいては、目的ドメインに特化した事前訓練の有効性が報告されている。我々は、学術データベース CiNii Articles の論文抄録を用いて、日本語の学術ドメインにおける事前訓練モデルを構築する。科研費の研究課題名を対象とする評価実験の結果、汎用モデルに対する提案モデルの有効性を確認した。

1　はじめに

様々な分野と言語において、研究成果が学術論文として日々 Web 上に蓄積されている。例えば、自然言語処理の分野の英語論文は ACL Anthology[1] 上に現在およそ 8.1 万件が公開されている。これらの大量の学術論文から人手で漏れなく知識を得るには大きなコストがかかるため、自然言語処理を用いた情報抽出 [1] および知識獲得 [2, 3] が期待されている。

現在主流の深層学習に基づく自然言語処理では、高品質なモデルを得るために、大規模なラベル付きコーパスを用いた教師あり学習が有効である。しかし、学術分野などの専門用語の扱いが重要なドメインにおいては、専門知識を持つアノテータが必要であり、アノテーションのコストが非常に高い。そのため、様々なドメイン・タスク・言語において、大規模なラベル付きコーパスが得られない少資源問題が大きな課題となっている。

近年の自然言語処理では、少資源問題の対策として、BERT [4] や BART [5] のような大規模な生コーパスを用いた事前訓練モデルの転移学習が広く利用されている。特に、医療 [6]・学術 [7, 8]・SNS [9] などのドメインでは、目的ドメインに特化した事前訓練の有効性が報告されている。日本語では、Wikipedia[2] や CC100 [10] を用いて事前訓練された汎用モデル[3)4)5)] とともに、医療 [11] や SNS[6)] のドメインに特化した事前訓練モデルが公開されている。しかし、SciBERT [7] のような、学術ドメインに特化した日本語の事前訓練モデルは公開されていない。

本研究では、日本語における学術ドメインの自然言語処理の性能改善のために、学術データベース CiNii Articles[7)] の論文抄録を用いて事前訓練モデルを構築した。我々は、CiNii Articles に所蔵されている日本語論文のうち約 127 万件から抽出した約 628 万文の論文抄録テキストを用いて、RoBERTa [12] および BART [5] に相当する事前訓練を行った。前者は、テキスト分類などの利用を想定した事前訓練済みエンコーダモデルであり、以降は Academic RoBERTa[8)] と呼ぶ。後者は、テキスト生成などの利用を想定した事前訓練済みエンコーダ・デコーダモデルであり、以降は Academic BART[9)] と呼ぶ。

学術ドメインにおける日本語の文分類および文対分類の評価実験の結果、汎用モデルである東北大 BERT[3)] および早大 RoBERTa[4)] と比較して、学術ドメインに特化した Academic RoBERTa の有効性を確認できた。また、学術ドメインにおける日本語のヘッドライン生成の評価実験の結果、汎用モデルである京大 BART[5)] と比較して、学術ドメインに特化した Academic BART の有効性を確認できた。

1）https://aclanthology.org/
2）https://ja.wikipedia.org/
3）https://huggingface.co/cl-tohoku/bert-base-japanese
4）https://huggingface.co/nlp-waseda/roberta-base-japanese
5）https://nlp.ist.i.kyoto-u.ac.jp/?BART 日本語 Pretrained モデル
6）https://github.com/hottolink/hottoSNS-bert
7）https://ci.nii.ac.jp/
8）https://github.com/hirokiyamauch/AcademicRoBERTa
9）https://github.com/hirokiyamauch/AcademicBART

2 日本語学術文書を用いた事前訓練

本研究では、日本語における学術ドメインの自然言語処理の性能改善のために、学術ドメインに特化した日本語事前訓練モデルを構築し、公開する。まず 2.1 節において、学術論文情報データベース CiNii Articles[7]から論文抄録を抽出し、事前訓練用の日本語コーパスを作成する。そして 2.2 節において、本コーパス上で単語穴埋め [5, 12] の事前訓練を行う。

2.1 コーパスの作成

国立情報学研究所が運用する学術論文情報データベース CiNii Articles を用いて、学術ドメインに特化した日本語コーパスを作成した。2022 年 3 月時点で CiNii Articles に収録されていた学術論文のうち、平仮名または片仮名を含む約 127 万件の論文抄録を抽出し、以下の 5 段階の前処理を施して約 628 万文（約 1.8 億語）のコーパスを作成した。

1. 定型表現の削除
2. 文分割
3. 日本語文の抽出
4. 重複文の削除
5. 文字数制限

それぞれの前処理ステップにおけるコーパスサイズの変化を表 1 に示す。

定型表現の削除　対象とした論文抄録には、「論文タイプ || 研究ノート」や「特集 乱流の数値シミュレーション（NST） その 3」など、自動的な情報抽出によるノイズが含まれる。これらの定型表現をコーパスから除外するために、同一文書が一定回数以上出現する際に、それらの文書を削除する。なお、論文抄録のテキストとしては適切なものの、ID などの登録ミスの影響で同一文書が 5〜6 回出現するという例が見られたため、定型表現として判定する際の閾値を 7 回以上と設定した。

文分割　上記の前処理によって得られた約 115 万文書に対して、文への分割を行う。ルールベースの文分割[10]によって、約 730 万文が得られた。

日本語文の抽出　日本語以外の言語で書かれた文を削除し、日本語文のみを得る。ただし、専門用語は他言語で表現される場合が少なくないため、文字単位で数えて閾値以上の割合が日本語（平仮名または片仮名または漢字）である文を抽出する。本研究では、この閾値を 0.5 に設定することで約 62 万文を削除し、約 668 万文の日本語文を得た。

表 1　前処理によるコーパスサイズの変化

前処理	コーパスサイズ
対象の論文抄録	1,269,361 文書
1. 定型表現の削除	1,145,812 文書
2. 文分割	7,305,893 文
3. 日本語文の抽出	6,683,983 文
4. 重複文の削除	6,333,833 文
5. 文字数制限	6,275,756 文

重複文の削除　高頻度な表現によるバイアスを防ぐために、「下腹部痛を主訴に来院」などの特定の分野において頻出する文や「その結果を以下に示す」などの学術論文における定型文を削除する。文単位での重複がある場合に、その文を 1 回だけコーパスに含めることで約 35 万文を削除し、約 633 万文の重複のない日本語文を得た。

文字数制限　定型表現の検出漏れや文分割のエラーを取り除くために、極端な短文および長文を削除する。10 文字未満の文には「(編集委員会作成)」などの実際の論文抄録には含まれないと考えられる表現が多く見られた。そこで本研究では、10 文字以上 200 文字以下の文を抽出することで、最終的に約 628 万文のコーパスを作成した。

2.2 事前訓練

Academic RoBERTa　2.1 節のコーパスを用いて、マスク言語モデリング [12] の事前訓練を行う。前処理として、SentencePiece[11] [13] によるサブワード分割を実施し、語彙サイズは 32,000 とした。fairseq[12] [14] の実装を用いて、roberta-base と同じ構造（隠れ層が 12 層、次元数が 768 次元、自己注意のヘッド数が 12 個）の Transformer [15] を訓練した。最大入力トークン数は 512、バッチサイズは 64 文、Dropout 率は 0.1 とし、最適化には Adam [16] を用いた。学習率スケジューリングには polynomial decay を用い、最大の学習率は 0.0001、Warmup ステップは 1 万とした。先行研究[4]との公平な比較のために、訓練ステップ数は 70 万ステップとした。

Academic BART　2.1 節のコーパスを用いて、雑音除去自己符号化 [5] の事前訓練を行う。fairseq[12] [14] の実装を用いて、bart-base と同じ構造

10） https://github.com/wwwcojp/ja_sentence_segmenter

11） https://github.com/google/sentencepiece

12） https://github.com/facebookresearch/fairseq

(encoder-decoder 層が 6 層、次元数が 768 次元、自己注意のヘッド数が 12 個）の Transformer [15] を訓練した。最大入力トークン数は 512、バッチサイズは 16 文（8 回の勾配蓄積)、Dropout 率は 0.1 とし、最適化には Adam [16] を用いた。学習率スケジューリングには polynomial decay を用い、最大の学習率は 0.0005、Warmup ステップは 1 万とした。先行研究[5]との公平な比較のために、訓練ステップ数は 50 万ステップとした。なお、事前訓練には 4 枚の GPU（RTX A6000 48GB）を使用した。

3 評価実験

学術ドメインに特化した我々の事前訓練モデルの有効性を評価するために、科学研究費補助金（科研費）の研究課題名を対象とする評価実験を通して、既存の汎用的な日本語事前訓練モデルと比較する。

3.1 比較手法

日本語の汎用的な事前訓練済みエンコーダモデルである東北大 BERT[3]および早大 RoBERTa[4]との比較によって、学術ドメインに特化した Academic RoBERTa の有効性を評価する。これらの比較手法は、どちらも Academic RoBERTa と同じ構造の Transformer [15] であり、同じサイズの語彙を持っている。ただし、事前訓練に用いるコーパスやその前処理、訓練時のハイパーパラメータが異なる。

東北大 BERT は、日本語 Wikipedia の約 1,700 万文で事前訓練された BERT [4] である。前処理として、MeCab（IPADIC）[17] による形態素解析および WordPiece [18] によるサブワード分割を行っている。最大入力トークン数は 512、バッチサイズは 256 文であり、100 万ステップの訓練を実施している。

早大 RoBERTa は、日本語 Wikipedia および CC100 [10] の日本語約 40 億文で事前訓練された RoBERTa [12] である。前処理として、Juman++ [19] による形態素解析および SentencePiace [13] によるサブワード分割を行っている。最大入力トークン数は 128、バッチサイズは 256 文（× 8 GPU）であり、70 万ステップの訓練を実施している。

Academic BART との比較には、日本語の汎用的な事前訓練済みエンコーダ・デコーダモデルである京大 BART[5]を用いる。京大 BART は、日本語 Wikipedia の約 1,800 万文で事前訓練された BART [5] である。前処理として、Juman++ [19] による形態素解析および SentencePiace [13] によるサブワード分割を行っている。最大入力トークン数は 512、バッチサイズは 128 文（× 4 GPU）であり、50 万ステップの訓練を実施している。

表 2 評価用データセットの統計

タスク	単位	訓練用	検証用	評価用
著者同定	文対	100,000	10,000	10,000
カテゴリ分類	文	70,000	1,500	1,500
ヘッドライン生成	文	68,000	1,500	1,500

3.2 評価タスク

学術ドメインにおける評価実験として、科研費の研究課題名に関するテキスト分類およびヘッドライン生成を行う。2013 年から 2017 年までの科研費の採択課題[13]を 7.3 万件収集し、データセットを構築した。テキスト分類として、研究課題名から研究分野を推定するカテゴリ分類と、研究課題名の対から研究代表者が同一か否かを推定する著者同定の 2 つの評価タスクを設計した。両タスクとも、分類正解率で自動評価する。また、概要から研究課題名を生成するヘッドライン生成の評価タスクも設計した。本タスクは ROUGE [20] で自動評価する。

著者同定 本タスクは、2 つの研究課題について、それらの研究代表者が同一か否かを 2 値分類する文対分類タスクである。本実験では、研究代表者が同一である正例を 5 万組と同一ではない負例を 7 万組の合計 12 万組を収集し、表 2 のように無作為に分割して使用した。モデルには、[SEP] の特殊トークンを挟んで 2 文を同時に入力した。

カテゴリ分類 本タスクは、科研費の研究課題名から研究分野を推定する文分類タスクである。科研費では、4 段階の階層構造を持つ研究分野の分類を採用しており、大区分から順に、4・14・77・318 種類のカテゴリが含まれる。本実験では、区分ごとに 4 種類のカテゴリ分類を実施した。

ヘッドライン生成 本タスクは、科研費における研究課題の概要から研究課題名を生成する要約タスクである。平均 270 文字で構成される概要を入力として、平均 31 文字の研究課題名を推定する。

3.3 再訓練

テキスト分類 3.2 節のコーパスを用いて、事前訓練済みエンコーダモデルを再訓練した。前処理として、それぞれ事前訓練と同じ設定でサブワード分

13） https://kaken.nii.ac.jp/

表 3 テキスト分類の評価実験

	著者同定	カテゴリ分類			
クラス数	2	4	14	77	318
東北大 BERT	95.1	83.7	69.6	53.3	40.3
早大 RoBERTa	97.1	83.9	71.9	55.4	42.7
Academic RoBERTa	**98.7**	**84.7**	**72.9**	**58.8**	**44.6**

表 4 ヘッドライン生成の評価実験

	ROUGE-1	ROUGE-2	ROUGE-L
京大 BART (Base)	95.1	83.7	69.6
京大 BART (Large)	97.1	83.9	71.9
Academic BART	**98.7**	**84.7**	**72.9**

割を行った。バッチサイズは 256 文、Dropout 率は 0.1 とし、最適化には Adam [16] を用い、最大の学習率を 5e-5 とした。検証用データの正解率を用いて 10 エポックの early stopping にて再訓練を終了した。

ヘッドライン生成 3.2 節のコーパスを用いて、事前訓練済みエンコーダ・デコーダモデルを再訓練した。前処理として、モデルごとに事前訓練と同じ設定でサブワード分割を行った。バッチサイズは 32 文、Dropout 率は 0.1 とし、最適化には Adam を用い、最大の学習率を 3e-5 とした。検証用データにおける ROUGE-L の評価値を用いて、5 エポックの early stopping にて再訓練を終了した。

3.4 実験結果

テキスト分類の実験結果を表 3 に示す。RoBERTa は一貫して BERT よりも高い性能を示し、学術ドメインに特化した Academic RoBERTa は全てのタスクにおいて最高性能を達成した。特に、詳細な専門知識が要求される小区分の分類において、提案手法では大区分よりも大きな性能の改善が見られた。

ヘッドライン生成の実験結果を表 4 に示す。本タスクでも、学術ドメインに特化した Academic BART が全ての評価指標において最高性能を示した。

早大 RoBERTa と Academic RoBERTa の間や、京大 BART (Base) と Academic BART の間には、モデル構造や学習ステップ数の相違はない。また、東北大 BERT は約 1,700 万文、早大 RoBERTa は約 40 億文、京大 BART は約 1,800 万文のコーパスを事前訓練に使用しており、我々の約 628 万文はコーパスサイズの点でも優位性を持たない。そのため、Academic RoBERTa および Academic BART の性能改善は、学術ドメインへの特化にのみ起因すると考えられる。

表 5 提案モデルの語彙中の特徴的なトークンの例

論文表現	専門用語
本研究の目的は	ニューラルネットワーク
する手法を提案する	オープンソース
であることが確認された	ヘモグロビン
について考察を行った	肝障害
以下の結論を得た	リン酸カルシウム

3.5 考察

先行研究 [7] と同様に、語彙について分析する。提案モデルの語彙のうち、49.4%は既存の事前訓練モデルの語彙[14] に含まれないことがわかった。提案モデルにのみ含まれる特徴的なトークンの例を表 5 に示す。「であることが確認された」のような分野を問わず学術論文に頻出するフレーズや、「ニューラルネットワーク」のような特定の分野において頻出する専門用語が、提案モデルの語彙のみに見られた。このようなドメインに特化したトークンを多く語彙に含むことで、学術ドメインにおける高性能な処理を実現できたと考えられる。

4 おわりに

本研究では、CiNii Articles の論文抄録から作成した日本語コーパスを用いて、学術ドメインに特化した事前訓練モデル Academic RoBERTa[8] および Academic BART[9] を構築し、公開した。科研費の研究課題名を対象とする著者同定・カテゴリ分類・ヘッドライン生成の評価実験の結果、学術ドメインに特化したモデルが既存の汎用モデルを上回る性能を達成した。分析の結果、学術ドメインに特有の表現を多く語彙に含むことや、詳細な専門知識が要求される小区分ほどテキスト分類の正解率が大きく向上していることから、ドメインに特化した事前訓練の有効性を確認できた。

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP20H04484）および 2020 年度国立情報学研究所公募型共同研究（20S0405）の助成を受けたものです。本稿の内容の一部は Third Workshop on Scholarly Document Processing [21] において報告したものです。

14) 以下の和集合：東北大 BERT の語彙、早大 RoBERTa の語彙、日本語 Wikipedia に対して語彙サイズ 32,000 で SentencePiece のサブワード分割を訓練した際の語彙

参考文献

[1] Arman Cohan and Nazli Goharian. Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 390–400, 2015.

[2] Mayank Singh, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, Rajarshi Haldar, Pawan Goyal, and Animesh Mukherjee. CL Scholar: The ACL Anthology Knowledge Graph Miner. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations**, pp. 16–20, 2018.

[3] Saif M. Mohammad. NLP Scholar: A Dataset for Examining the State of NLP Research. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 868–877, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, 2020.

[6] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In **Proceedings of the 2nd Clinical Natural Language Processing Workshop**, pp. 72–78, 2019.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 3615–3620, 2019.

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. **Bioinformatics**, Vol. 36, No. 4, pp. 1234–1240, 2020.

[9] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A Pre-trained Language Model for English Tweets. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 9–14, 2020.

[10] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 4003–4012, 2020.

[11] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A Clinical Specific BERT Developed Using a Huge Japanese Clinical Text Corpus. **PLOS ONE**, Vol. 16, No. 11, pp. 1–11, 2021.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692**, 2019.

[13] Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, 2018.

[14] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, 2019.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems**, pp. 5998–6008, 2017.

[16] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In **Proceedings of the 3rd International Conference for Learning Representations**, 2015.

[17] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, 2004.

[18] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. **arXiv:1609.08144**, 2016.

[19] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Design and Structure of The Juman++ Morphological Analyzer Toolkit. **Journal of Natural Language Processing**, Vol. 27, No. 1, pp. 89–132, 2020.

[20] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81, 2004.

[21] Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. A Japanese Masked Language Model for Academic Domain. In **Proceedings of the Third Workshop on Scholarly Document Processing**, pp. 152–157, 2022.