JMedRoBERTa: 日本語の医学論文にもとづいた事前学習済み言語モデルの構築と評価

杉本 海人 [1,3]　壹岐 太一 [2,3*]　知田 悠生 [1,3]　金沢 輝一 [3]　相澤 彰子 [1,3]
[1] 東京大学大学院　[2] 総合研究大学院大学
[3] 国立情報学研究所
{kaito_sugimoto, iki, chida, tkana, aizawa}@nii.ac.jp

概要

英語圏において，医学論文で事前学習を行った言語モデルが，医療ドメインの様々なタスクで有効であることが示されている．日本語では診療記録で事前学習を行った UTH-BERT [1] が公開されているものの，医学論文を用いたものは存在しない．そこで本研究では，日本語の医学論文で事前学習した言語モデルである JMedRoBERTa を提案する．評価実験の結果，JMedRoBERTa は医療ドメインタスクにおいて UTH-BERT と同等またはより高い性能を発揮するほか，一般ドメインタスクにおいても比較的高い精度を示すことを確認した．また，複数トークナイザの比較から，トークナイザの差によりモデルが得意とするタスクが異なる傾向が示唆された．

1　はじめに

近年，医療ドメインに特化した事前学習済み言語モデルを構築する研究が盛んである．英語圏では，医学論文データベース PubMed[1)] や診療記録データベース MIMIC-III [2] にもとづき，BioBERT [3]，ClinicalBERT [4]，BlueBERT [5]，PubMedBERT [6] などのモデルが提案されている．こうしたモデルは，医学用語の自動抽出 [7] や表記揺れ解消 [8]，医学論文の検索 [9] など，幅広いタスクに応用されている．また，日本語のモデルとしては，診療記録を用いて事前学習を行った UTH-BERT [1] が存在する．

診療記録が患者情報の記録に主眼を置くのに対し，医学論文は基礎研究から症例報告まで幅広い内容を扱うのが特徴である．したがって，医学論文で事前学習を行うことにより，診療記録からは得られない知識を獲得できる可能性がある．また，英語のモデルの研究では，医学論文で事前学習された PubMedBERT が診療記録で事前学習された ClinicalBERT よりも下流タスクでの性能が高い例が見られる [6, 7]．こうした背景から，日本語の医学論文を用いたモデルの事前学習が有益だと考える．

本研究では約 1.8GB （約 1,100 万文）の日本語の医学論文データを用いて RoBERTa [10] のフルスクラッチ学習を行う．モデルのトークナイザに関しては，分かち書きを前処理として行う場合と行わない場合の双方を試す．学習したモデルを 4 つの医療ドメインタスク（2 つが医学論文に関するもの，2 つが診療記録に関するもの），および JGLUE [11] に含まれる 5 つの一般ドメインタスクで評価する．

評価実験の結果，提案モデルが医学論文に関するタスクで UTH-BERT よりも高いスコアを示したほか，診療記録に関するタスクでも UTH-BERT に並ぶスコアを示した．また，分かち書きを行わないモデルは行うモデルよりも概ね高いスコアを示したが，一部のタスクにおいては逆の結果が見られた．さらに，複数の一般ドメインタスクにおいて，提案モデルは一般ドメインの言語モデルに匹敵するスコアを示した．提案モデルを我々は **JMedRoBERTa** と名付け，付録 A に示したリンクで公開している．

2　JMedRoBERTa の構築

2.1　フルスクラッチ学習・追加学習

あるドメインに特化した言語モデルを作成したい場合，既存の一般ドメインの言語モデルに追加で事前学習を行う流儀と，フルスクラッチで事前学習を行う流儀の 2 種類が知られている [12, 13, 14]．膨大な医学用語を含む医療ドメインにおいては，フルスクラッチ学習モデルである PubMedBERT [6] が追加学習モデルである BioBERT [3] に比べて多くの医学用語を語彙に含み，複数の医療ドメインのタスクで

* 現在は NTT 人間情報研究所に所属している．
1） https://pubmed.ncbi.nlm.nih.gov/

表 1　東北大 BERT，UTH-BERT，JMedRoBERTa に含まれる語彙の比較．当該の医学用語が語彙に含まれる場合は ○ で示し，そうでない場合は，トークナイザによって分割された後のトークン群を示す．なお，[UNK] は対応するトークンが存在しないことを表し，##は前のトークンと結合して単語をなすことを表す．

医学用語	カテゴリ	東北大 BERT	UTH-BERT	JMedRoBERTa（万病 WordPiece）	JMedRoBERTa（SentencePiece）
網膜剥離	病気	網 ##膜 剥 ##離	○	○	○
角化症	病気	角 化 症	角化 ##症	○	○
リドカイン	医薬品	リ ##ド ##カイ ##ン	リ ##ド ##カイン	○	○
エゼチミブ	医薬品	エ ##ゼ ##チ ##ミ ##ブ	エ ##ゼ ##チ ##ミ ##ブ	○	○
自己抜去	臨床	自己 抜 ##去	○	○	自己 抜去
軽度倦怠感	臨床	軽 ##度 [UNK] 感	○	軽度 ##倦 ##怠感	軽度 倦怠感
肺動脈	人体部位	肺 動脈	肺 動脈	肺 動脈	○
上顎前歯部	人体部位	上顎 前 ##歯 部	上顎 前歯 部	上顎 前歯 部	○

より高い性能を発揮している．この結果を踏まえ，本研究でもフルスクラッチ学習を行う．

2.2　学習用データ

本研究では，大規模な日本語の医学論文データを用いる．各データには論文の抄録（アブストラクト）のテキストのほか，その論文が属する分野や索引語（キーワード）等のラベルが付与されている．一部のデータは論文の本文のテキストを含む．それらのうち，JMedRoBERTa の事前学習には抄録および本文のテキストを用いる．また，その他のデータは評価タスクに使用する．詳細は付録 B に示す．

モデルの事前学習で用いるテキストのサイズは，合計で約 1.8GB （約 1,100 万文）である．なお，テキストは事前に半角文字を全角文字に正規化する．

2.3　トークナイザの訓練

英語と異なり日本語では単語間の空白文字が存在しないため，トークナイザの構築において，分かち書きを前処理として行うべきか否かが焦点の一つとなる．この違いは，モデルの下流タスクの性能へも一定の影響を与えることが確認されている [15, 16]．

本研究では，分かち書きを行うトークナイザとして**万病 WordPiece トークナイザ**を，分かち書きを行わないトークナイザとして **SentencePiece トークナイザ**を訓練する．両者ともに語彙数は 30,000 とする．万病 WordPiece トークナイザは，MeCab[2] を用いて入力文を単語列に分かち書きした後，WordPiece [17] によりサブワード化を行うものである．MeCab の辞書選択においては，通常用いられる ipadic に加えて，大規模な病名辞書である万病辞書 [18] を使う．これにより，病名が分かち書きの時点で複数の単語列に分割されることを防ぐ．

一方，SentencePiece トークナイザは，SentencePiece [19]（Unigram モデル）により，文から直接サブワード化を行うものである．出現頻度の高い文字列はひとかたまりのトークンとして処理されるため，医学論文において頻度の高い医学用語は，分割されずに 1 トークンとして扱われることが期待される．

表 1 はトークナイザの訓練によって得られた JMedRoBERTa の語彙を，日本語 Wikipedia で事前学習された東北大 BERT[3]，および日本語の診療記録で事前学習された UTH-BERT の語彙と比較したものである．JMedRoBERTa が多様な医学用語を語彙に含むことを裏付けている．

2.4　モデルの事前学習

Liu ら [10] は，Next Sentence Prediction タスクの削除やマスク位置の動的な変更などにより BERT [20] の事前学習がさらに効果的になることを指摘し，改良された BERT 型言語モデルを RoBERTa と名付けた．本研究では RoBERTa の事前学習手法を採用する．2.3 節で挙げた 2 種類のトークナイザそれぞれに関してモデルの事前学習を行い，「万病 WordPiece モデル」「SentencePiece モデル」を作成する．訓練設定の詳細は付録 C に記載する．

3　医療ドメインタスクによる評価

3.1　タスク・データセットの概要

医学論文の分野分類　抄録テキストから分野を予測するタスクである．2.2 節で示した学習用データのうち事前学習に用いていないものから，分野が 1 つだけラベル付けされた医学論文のデータをランダムに 100,000 件抽出し，8:1:1 で訓練用・バリデー

2）https://taku910.github.io/mecab/

3）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

表 2 医療ドメインタスクにおける評価結果．各モデルのスコアは，異なるランダムシードを用いた 5 回の実験結果の平均値である．太字は全モデルにおける最も高いスコア，下線は 2 種類の JMedRoBERTa のうちより高いスコアを示す．

タスク	評価指標	東北大 BERT	UTH-BERT	JMedRoBERTa（万病 WordPiece）	JMedRoBERTa（SentencePiece）
医学論文の分野分類	macro-F1	0.505	0.493	0.552	**0.566**
	micro-F1	0.622	0.627	0.671	**0.678**
医学論文の索引語への副標目の付与	macro-F1	0.451	0.456	0.487	**0.514**
	micro-F1	0.553	0.566	0.584	**0.604**
診療記録の固有表現抽出（MedNLP-2）	F1（病状）	0.855	**0.862**	0.862	0.833
	F1（時間表現）	**0.852**	0.834	0.817	0.739
診療記録への病名コード付与（MedNLPDoc）	F1（LV4 / SURE）	0.029	**0.038**	0.026	0.034
	F1（LV4 / MAJOR）	0.036	**0.067**	0.043	0.050
	F1（LV4 / POSSIBLE）	0.045	**0.078**	0.055	0.069
	F1（LV3 / SURE）	0.043	**0.059**	0.043	0.047
	F1（LV3 / MAJOR）	0.062	**0.103**	0.070	0.076
	F1（LV3 / POSSIBLE）	0.075	**0.121**	0.085	0.112
	F1（LV0 / SURE）	0.287	0.306	0.278	**0.324**
	F1（LV0 / MAJOR）	0.367	0.414	0.382	**0.432**
	F1（LV0 / POSSIBLE）	0.399	**0.448**	0.389	0.448

ション用・テスト用とする．訓練におけるラベルの種類数は 111 である．

医学論文の索引語への副標目の付与　論文に付与された索引語に対して副標目（サブヘディング）[4] を付与するタスクである．論文の抄録と索引語のペアを入力し，適切な副標目を候補の中から複数予測するマルチラベル分類として定式化する．データ作成の詳細は付録 D に示す．なお，予測候補の副標目は訓練用データに出現する 26 語とする．

診療記録の固有表現抽出　NTCIR-11 MedNLP-2 [21] の Task 1 を用いた評価を行う．このタスクでは，入力である診療記録から病状および時間表現を抽出する．訓練データ，テストデータにおける診療記録の数はそれぞれ 102 件，49 件である．

診療記録への病名コード付与　NTCIR-12 MedNLPDoc [22] の Task 1 を用いた評価を行う．このタスクでは，診療記録から対応する複数の ICD-10 病名コード [5] を予測する．訓練データ，テストデータにおける診療記録の数はそれぞれ 200 件，78 件である．なお，テストデータは 3 人のアノテータによって病名コードが付与されており，3 人全員が付与したコードのみを含むデータは SURE，これに加えて 2 人以上が付与したコードも含むデータは MAJOR，これに加えて誰か 1 人が付与したコードも含むデータは POSSIBLE と名づけられている．

4）副標目とはある索引語がどのような文脈で使われているかを示す副次的なキーワードである．例えば「精神障害」という索引語に対し「薬物療法」という副標目を加えて検索することで，薬物を用いた精神障害の治療法に関する文献のみを調べることができる．

5）https://www.cdc.gov/nchs/icd/icd10.htm

3.2 実験設定

医学論文の分野分類と診療記録の固有表現抽出の 2 つのタスクは，それぞれ通常の分類問題，系列ラベリング問題として処理する．

副標目付与タスクは一度に各ラベルの有無を二値分類するマルチラベル分類として扱う．評価の際，macro-F1 はテストデータにおいて出現しない副標目を無視して 20 副標目を対象に算出する．

病名コード付与タスクは，診療記録と病名の類似度学習を行うことで解く．詳細は付録 E に示す．評価の際は，シェアードタスクに倣い本研究でも，予測と正解ラベルのコードが完全に一致する場合に正解とみなす LV4 (Exact match)，コードの前 3 文字が一致する場合に正解とみなす LV3 (Rough match)，コードの大分類が一致する場合に正解とみなす LV0 (Category match) の 3 種類でそれぞれ評価を行う．

3.3 結果

表 2 に結果を示す．医学論文に関する 2 つのタスクにおいては，JMedRoBERTa の 2 つのモデルが東北大 BERT や UTH-BERT よりも高いスコアを示している．さらに，診療記録の固有表現抽出においては万病 WordPiece モデルが，診療記録への病名コード付与においては SentencePiece モデルが UTH-BERT に匹敵するスコアを示している．これらから，医学論文を用いた事前学習が医療ドメインの幅広いタスクにおいて有効であることが示唆される．

JMedRoBERTa の 2 つのモデルの間では，固有表現

表 3 JGLUE の dev セットにおける評価結果．JMedRoBERTa のスコアは，異なるランダムシードを用いた 3 回の実験結果の平均値である．JMedRoBERTa 以外のモデルのスコアは，JGLUE の GitHub リポジトリに掲載された base モデルの値である．下線は 2 種類の JMedRoBERTa のうちより高いスコアを示す．

データセット	評価指標	(Human)	東北大 BERT	NICT BERT	早大 RoBERTa	JMedRoBERTa（万病 WordPiece）	JMedRoBERTa（SentencePiece）
MARC-ja	accuracy	(0.989)	0.958	0.958	0.962	0.855	0.855
JSTS	Pearson	(0.899)	0.909	0.910	0.913	0.880	0.878
	Spearman	(0.861)	0.868	0.871	0.873	0.839	0.834
JNLI	accuracy	(0.925)	0.899	0.902	0.895	0.879	0.874
JSQuAD	EM	(0.871)	0.871	0.897	0.864	0.859	0.798
	F1	(0.944)	0.941	0.947	0.927	0.925	0.800
JCommonsenseQA	accuracy	(0.986)	0.808	0.823	0.840	0.654	0.669

抽出を除いて SentencePiece モデルが万病 WordPiece モデルよりも高いスコアを示している．これは，分かち書きを行わない SentencePiece モデルは日本語の単語境界を越えたトークンを語彙に多く含み[6]，同一の文書をより少ないトークン数で表現できることが影響している可能性がある．一方で，固有表現抽出に関しては，固有表現のアノテーションが単語境界に従って行われる場合が多く，分かち書きを行う万病 WordPiece モデルに有利だと考えられる．

4 一般ドメインタスクによる評価

4.1 タスク・データセットの概要

日本語言語理解ベンチマーク JGLUE [11] に含まれる，以下の 5 つのデータセットを用いる：MARC-ja（ネガポジ分類による感情分析タスク），JSTS（意味的類似度計算タスク），JNLI（自然言語推論タスク），JSQuAD（抽出型機械読解タスク），JCommonsenseQA（常識推論を要求する選択肢型機械読解タスク）．

4.2 実験設定

JGLUE の GitHub リポジトリ[7]に公開されているデータ（v1.1）をもとに実験を行う．データ内のテキストは事前に全角文字に正規化する．

4.3 結果

表 3 に結果を示す．JSTS や JNLI において，JMedRoBERTa は一般ドメインで事前学習を行った他の言語モデルに匹敵する高いスコアを出している．[8] これは，意味的類似度計算や自然言語推論が言語理解において特に基本的なタスクであり，ドメインにかかわらず学習可能であることを示唆している．一方で MARC-ja において，JMedRoBERTa は予測を全て positive と返した場合のスコア（0.855）になっており，分類そのものに失敗している．また，JCommonsenseQA においても一般ドメインの言語モデルに比べるとスコアを大きく下げている．これらは医学論文に記述されない日本語の性質や常識を問うものであるため，自然な結果である．

JSQuAD については，万病 WordPiece モデルに限って見ると，JMedRoBERTa は他の言語モデルに匹敵するスコアを出している．すなわち，JSQuAD は一般ドメインでの読解評価を目指して構築されたデータセットであるが，特定のドメインで事前学習されたモデルでも容易に解けることを意味する．[9]

5 まとめ

本稿では，日本語の医学論文にもとづく事前学習済み言語モデル JMedRoBERTa の構築と評価について述べた．医療言語処理の様々な領域において JMedRoBERTa が広く活用されることを期待する．

今後の方向性として，医学論文の引用関係を事前学習で用いた BioLinkBERT [23] や，医学オントロジーの知識を事前学習で用いた DRAGON [24] のようなモデルを日本語でも作成することが挙げられる．また，言語モデルのフルスクラッチ学習には大量の計算リソースが必要になるのが課題である．そこで，計算コストを抑えつつ，モデルを追加学習よりも効果的にドメイン特化させる exBERT [25] のようなアプローチを日本語で試すことも考えられる．

6） 例えば「可能性が示唆された」や「有意差は認められなかった」といったフレーズが語彙に含まれている．

7） https://github.com/yahoojapan/JGLUE

8） 例えば JNLI における majority baseline は 0.553 であり，JMedRoBERTa のスコアはそれよりも十分に高い．

9） SentencePiece モデルは万病 WordPiece モデルに比べてスコアを大きく落としている．これは，3.3 節の固有表現抽出タスクにおいて SentencePiece モデルがスコアを落とした原因と同様，トークンの区切りと解答のスパンの開始終了位置が一致しないことによるものである．なお，JGLUE の著者らも XLM-RoBERTa に対して同様の報告をしている．

謝辞

本研究は，JST の AIP 日独仏 AI 研究（課題番号 JPMJCR20G9）および学際大規模情報基盤共同利用・共同研究拠点（JHPCN）（課題番号：jh221004) の支援を受けた．本研究成果の一部は，データ活用社会創成プラットフォーム mdx を利用して得られた．

参考文献

[1] 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, 11 2021.

[2] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. **Scientific data**, Vol. 3, No. 1, pp. 1–9, 2016.

[3] 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. **Bioinform.**, Vol. 36, No. 4, pp. 1234–1240, 2020.

[4] 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, Minneapolis, Minnesota, USA, June 2019.

[5] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In **Proceedings of the 18th BioNLP Workshop and Shared Task**, pp. 58–65, Florence, Italy, August 2019.

[6] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **ACM Trans. Comput. Heal.**, Vol. 3, No. 1, pp. 2:1–2:23, 2022.

[7] Yongwei Zhang, Rui Cheng, Lu Luo, Haifeng Gao, Shanshan Jiang, and Bin Dong. SRCB at the NTCIR-16 RealMedNLP Task. In **NTCIR-16**, 2022.

[8] Shogo Ujiie, Hayate Iso, and Eiji Aramaki. Biomedical entity linking with contrastive context matching. **CoRR**, Vol. abs/2106.07583, , 2021.

[9] 吉井瑞貴, 竹下昌志, ジェプカ・ラファウ, 荒木健治. 論文の構造とタイトルを独立して考慮した医学論文検索モデルの提案. ことば工学研究会 : 人工知能学会第 2 種研究会ことば工学研究会資料, Vol. 67, pp. 1–8, 12 2021.

[10] 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. **CoRR**, Vol. abs/1907.11692, , 2019.

[11] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **LREC 2022**, pp. 2957–2966, Marseille, France, June 2022.

[12] 壹岐太一, 金沢輝一, 相澤彰子. 学術分野に特化した事前学習済み日本語言語モデルの構築. 第 139 回情報基礎とアクセス技術研究発表会, 2020.

[13] Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, and Kiyoshi Izumi. Construction and Validation of a Pre-Training and Additional Pre-Training Financial Language Model. In **SIG-FIN 28**, pp. 132–137, 2022.

[14] Keisuke Miyazaki, Hiroaki Yamada, and Takenobu Tokunaga. Cross-domain analysis on Japanese legal pretrained language models. In **AACL-IJCNLP 2022 (Findings)**, pp. 274–281, Online only, November 2022.

[15] 築地俊平, 新納浩幸. Tokenizer の違いによる日本語 BERT モデルの性能評価. 言語処理学会第 27 回年次大会, 2021.

[16] 井上誠一, Nguyen Tung, 中町礼文, 李聖哲, 佐藤敏紀. 日本語 GPT を用いたトークナイザの影響の調査. 言語処理学会第 28 回年次大会, 2022.

[17] Mike Schuster and Kaisuke Nakajima. Japanese and Korean voice search. In **ICASSP 2012**, pp. 5149–5152, 2012.

[18] Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. J-MeDic: A Japanese disease name dictionary based on real clinical usage. In **LREC 2018**, Miyazaki, Japan, May 2018.

[19] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **EMNLP 2018: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT 2019**, pp. 4171–4186, June 2019.

[21] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. Overview of the NTCIR-11 mednlp-2 task. In **NTCIR-11**, 2014.

[22] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. Overview of the NTCIR-12 mednlp-doc task. In **NTCIR-12**, 2016.

[23] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. LinkBERT: Pretraining language models with document links. In **ACL 2022**, pp. 8003–8016, Dublin, Ireland, May 2022.

[24] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In **NeurIPS 2022**, 2022.

[25] Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In **EMNLP 2020 (Findings)**, pp. 1433–1439, Online, November 2020.

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **ICLR 2015**, 2015.

A　モデルの公開リンク

本研究で作成した事前学習済みモデルは以下のリンクで公開されており，Hugging Face の Transformers ライブラリを経由して呼び出すことができる．

- 万病 WordPiece モデル（語彙数 30,000）：
  https://huggingface.co/alabnii/jmedroberta-base-manbyo-wordpiece
- SentencePiece モデル（語彙数 30,000）：
  https://huggingface.co/alabnii/jmedroberta-base-sentencepiece

なお，本文では言及しなかったが，本研究ではモデルの語彙数の影響を調査するため，他の条件を揃えて語彙数 50,000 のモデルも並行して構築し，評価した．これらは以下のリンクで公開されている．

- 万病 WordPiece モデル（語彙数 50,000）：
  https://huggingface.co/alabnii/jmedroberta-base-manbyo-wordpiece-vocab50000
- SentencePiece モデル（語彙数 50,000）：
  https://huggingface.co/alabnii/jmedroberta-base-sentencepiece-vocab50000

ただし，語彙数の差によるタスク性能への影響は（トークナイザの差による影響と比べると）ごくわずかであり，語彙数の多いモデルの方がどのタスクでも性能が高い，または低いというような一貫した影響は見られなかった．

B　学習用データの詳細

本研究では，JST から提供を受けた 2 種類の非公開データを学習に用いる．

1 つは，論文の抄録のみを含むデータである．このデータから，JST 分類コードが「医学」に該当するデータ[10]の和文抄録のみ抽出する．その上で，データをランダムに 8:1:1 に分割し，それぞれ train，validation，test セットとする．train セットはトークナイザの訓練用，およびモデルの事前学習の訓練用とし，validation セットはモデルの事前学習のバリデーション用とする．test セットはモデルの評価データセットの構築にのみ使う．もう 1 つは，論文の抄録に加えて本文を含むデータである．このデータから，同様に医学論文データの和文抄録・本文のみ抽出し，全てトークナイザおよびモデルの訓練に使う．

モデルの事前学習で用いるテキストのサイズは，前者が約 1.6GB（約 1,000 万文），後者が約 0.2GB（約 140 万文）である．

C　事前学習の訓練詳細

C.1　ハイパーパラメータ

表 4 に，JMedRoBERTa の事前学習におけるハイパーパラメータ一覧を示す．

10）https://jdream3.com/service/science/life.html の「大分類：医学 (L0700)」を参照．

表 4　ハイパーパラメータの一覧

最大入力トークン数	512
バッチサイズ	256
Dropout	0.1
Attention Dropout	0.1
訓練ステップ数	2,000,000
ウォームアップステップ数	20,000
ピーク学習率	1e-4
学習スケジューラ	Linear
Weight Decay	0.01
最適化器	Adam [26]
Adam のパラメータ ϵ	1e-8
Adam のパラメータ β_1	0.9
Adam のパラメータ β_2	0.999

C.2　その他の訓練の工夫

万病 WordPiece モデルにおいては，事前学習のマスク穴埋めタスクの際に，1 つの単語に対応する複数のサブワードは全てマスクする Whole Word Masking を採用する．また，学習の高速化のため，単精度浮動小数点数（FP32）と半精度浮動小数点数（FP16）を組み合わせた自動混合精度（Automatic Mixed Precision：AMP）演算を導入する．

D　副標目付与タスクのデータ作成詳細

はじめに 2.2 節で示した論文データのうち事前学習に用いていないものから，（抄録，索引語，副標目集合）の組を抽出する（副標目集合は空でもよい）．次に，このままでは副標目集合が空の事例数が多いため，全ての索引語について副標目集合が空の抄録を取り除く．最後に，索引語のカテゴリでフィルタリングを行い医学に特に関係する索引語を含む組だけを残す．このようにして得られる 359,434 件の組を，抄録が分割をまたがないという条件のもと，6:2:2 で訓練用・バリデーション用・テスト用に分割する．

E　病名コード付与タスクの訓練詳細

まず，訓練データのラベルである病名コードを標準病名マスター[11]を用いて病名に変換し，（診療記録，病名）の組を作成する．そして，同一の組に含まれる診療記録と病名の埋め込みの距離が近くなるよう対照学習を行う．なお，埋め込みは各モデルの `[CLS]` トークンに対応する最終出力層から得る．また，対照学習における負例は，訓練バッチ内の異なる 2 つの組から 1 つずつ取得した診療記録と病名のペア (in-batch negatives) とする．

推論時は，入力の診療記録に対して最も埋め込みの近い病名を順に出力し，標準病名マスターを用いて病名コードに変換する（最終的に異なる 10 個の病名コードを予測とする）．

11）https://www2.medis.or.jp/stdcd/byomei/index.html