大規模事前学習言語モデルによる日本語所見文を用いたCOVID-19 肺炎の自動検出

鈴木脩右 [1]　明石敏昭 [2]　橋本正弘 [3]　大竹義人 [4]　村尾晃平 [5]　狩野芳伸 [1]
[1] 静岡大学大学院 総合科学技術研究科　[2] 順天堂大学 医学部 放射線診断学講座
[3] 慶應義塾大学 医学部 放射線科 (診断)　[4] 奈良先端科学技術大学院大学 先端科学技術研究科
[5] 国立情報学研究所 医療ビッグデータ研究センター
ssuzuki@kanolab.net　t.akashi.sg@juntendo.ac.jp　m.hashimoto@rad.med.keio.ac.jp
otake@is.naist.jp　k-murao@nii.ac.jp　kano@inf.shizuoka.ac.jp

概要

CT 画像に付随する放射線読影レポートの所見文から，COVID-19 肺炎であるか否かを自動検出する検出器を構築した．読影レポートの件数は多くないため，検出器には複数の大規模事前学習言語モデルを用い比較した．提案手法はベースラインより優れ，大規模事前学習言語モデルの有効性を示した．なかでも，我々が構築した RoBRTa ベースの検出器は Accuracy スコア 0.918 と最高精度を達成した．また，SHAP を用いて検出に影響を与える上位 9 単語を算出し，提案モデルが肺の症状を重視することが示唆された．

1　はじめに

COVID-19 肺炎が世界中で流行している．日本医学放射線学会は日本医用画像データベースを用いた COVID-19 肺炎の CT[1)] 画像収集と教師データ作成を行い，AI 画像診断を利用した COVID-19 肺炎の国内での発生検出に取り組んでおり，一定の成果を得ている [1]．現在，自然言語処理によるアプローチを取り入れた高精度な検出システムの実現が期待されている．

本研究では，CT 画像に付随する日本語放射線読影レポートを用いて，所見文から COVID-19 肺炎を自動検出する検出器の構築を試みる．我々は日本医学放射線学会が収集したデータからデータセットを構築する．データセットが小規模であることから，大規模事前学習言語モデルを検出器に用い高精度なモデルを実現する．

1）　Computed Tomography: コンピュータ断層撮影

2　関連研究

日本語放射線読影レポートを用いたタスクに大規模事前学習モデルを適応した研究では，多田ら [2]，Kuwabara ら [3]，Nakamura ら [4]，本田ら [5] 等がある．

Kuwabara らは小規模な疾患 2 値分類データセットを構築し，事前学習済みの BERT[6] を利用することで，小規模データの学習でありながら高精度な分類器を構築した．Honda らは約 32 万件の放射線科読影レポートを用いて BERT の事前学習を行い，12 クラス疾患分類タスクにおいて優れた分類器を構築した．

我々の知る限り，日本語放射線読影レポートを用いた COVID-19 肺炎の自動検出に関する研究は存在しない．

3　データセット

放射線読影レポートとは，放射線科の読影医が CT 等の検査画像から，検査画像に対する所見や診断を記載した報告書である．所見文の文例を図 3 に示す．本研究では所見文を入力し，COVID-19 肺炎の自動検出を行う．

【所見文】
yyyy/mm/dd CTと比較しました．
左肺底区に斑状陰影あり．前回CTで見受けられたすりガラス影によく似ている印象．
右肺底区胸膜下の網状影は増悪．一部結節状に見える箇所もあります．
全体として肺炎後変化または軽微な肺炎を疑います．
脂肪肝，右腎結石，両側腎嚢胞を認めません．

図 1　所見文文例 (作例)

3.1 使用データ

日本医学放射線学会が収集したデータを使用する．

PCR 検査結果データ は PCR 検査結果を元に，3 種類のラベル (陰性，陽性，不明) を付与したデータである．「不明」は検査結果が判定不能のものである．本研究では「陰性」と「陽性」を使用する．

COVID-19 流行前データ は COVID-19 流行前に収集されたデータに対して，COVID-19 らしさを示す 4 種類のラベル (典型的所見，疑わしい所見，非典型的所見，肺炎ではない所見) を付与したデータである．本研究では「典型的所見」を陽性データ，「非典型的所見」と「肺炎ではない所見」を陰性データとして扱う．

正常肺データ は正常肺の症例を集めたデータである．本研究では陰性データとして扱う．

アノテーションデータ は本研究で独自にアノテーションしたデータである．データベースの中には，PCR 検査結果データのようなラベルは付与されていないが COVID-19 の症状について記載されているデータが存在する．これらのデータを抽出し，記載内容を元に手作業で 3 種類のラベル (陰性，陽性，不明) のいずれかを付与した．この内「陰性」と「陽性」を使用する．

3.2 データセットの構築

3.1 節で述べたデータを元に 2 種類のデータセットを構築した．

性能評価用データセット は正常肺データ以外のデータで構築した．データサイズは 3,317 件 (陰性:1,664 件，陽性:1,653 件) である．本データセットで交差検証を行い，各モデルでの性能を評価した．

実タスクを想定したデータセット は PCR 検査結果データ及び正常肺データから評価データを，評価データに使用していないデータから訓練データをそれぞれ構築したデータセットである．データサイズは，評価データ 300 件 (陰性:150 件，陽性:150 件)，訓練データ 3,084 件 (陰性:1,581 件，陽性:1,503 件) である．

4 大規模事前学習言語モデル

本研究で使用した大規模事前学習言語モデルについて述べる．各モデルは huggingface/transformers [7] で実装した．

4.1 事前学習済み BERT

本研究で用いた事前学習済み BERT[2] について説明する．モデルサイズは全て base サイズ (層数 12，次元数 768，ヘッド数 12，最大入力長 512) である．

TOHOKU-BERT は東北大学が公開しているモデル[3] である．日本語 Wikipedia で事前学習されており，単語分割は MeCab+WordPiece で行う．MeCab の辞書は Unidic を使用する．

複数のバージョンの内，‘bert-base-japanese-v2’ を使用した．

TOHOKU-BERT-char は東北大学が公開しているモデルである．日本語 Wikipedia で事前学習されており，単語分割は文字単位で行う．

複数のバージョンの内，‘bert-base-japanese-char-v2’ を使用した．

UTH-BERT は東京大学が公開しているモデル [8] である．日本語で記載された大規模医療テキストで事前学習されており，単語分割は MeCab+WordPiece で行う．MeCab の辞書は ipadic-NEologd と万病辞書 [9] を併用する．

SP-BERT は Kikuta が公開しているモデル [10] である．日本語 Wikipedia で事前学習されており，単語分割は SentencePiece で行う．

Radiology-BERT は本田らが構築したモデル [5] である．放射線読影レポートで事前学習されており，単語分割は Byte Pair Encoding (BPE) で行う．なお本モデルは一般公開されていない．

4.2 RoBERTa の事前学習

我々は，約 85 万件からなる放射線読影レポートの所見文[4] を使用し base サイズの RoBERTa [11] を構築した．

本モデルの特徴として，1 つ以上の文書から連続した最大入力長分の単語を入力とした学習を行わない．代わりに，所見文が最大入力長に満たない場合は，診断文を加え事前学習データの単語数を増加させた．また，計算コストの都合から，Mixed Precision [12] によるメモリの節約，Gradient Accumulation [13] による batch サイズの増加をしている．

辞書に ipadic-NEologd と万病辞書を併用した MeCab+WordPiece で単語分割し，Whole Word Masking

2） モデルの呼称は本論文独自のものが含まれる点に留意されたい．
3） https://github.com/cl-tohoku/bert-japanese
4） 3 節で述べたデータは除いている．

での Masked Language Modeling の事前学習を，語彙数 32,000，batch サイズ 128 で 30 epoch 行った．
本モデルを **Radiology-RoBERTa** とする．

5 実験

COVID-19 肺炎の自動検出を陰性/陽性の 2 値分類タスクとみなし，2 種類の実験を行った．

5.1 ベースライン

ベースラインには 2 種類のモデルを用意した．

ルールベース 特定単語 (GGO，GGN，すりガラス，浸潤影，consolidation，grand glass，pneumonia，網状影，肺炎，間質性) を含む所見文を陽性，それ以外を陰性と判断する．

BiLSTM Radiology-RoBERTa と同様の所見文データで事前学習した Word2Vec を初期重みとする BiLSTM モデルである．辞書に ipadic-NEologd と万病辞書を併用した MeCab+WordPiece で単語分割した．パラメータは語彙数 32,000，埋め込み次元 128，隠れ状態 512，層数 2，最大入力長 512 とした．

5.2 評価指標

分類タスクで一般的に用いられる Accuracy，Precision，Recall，F_1 に加え，Specificity を評価指標とした．Specificity は 1 式で求める．

$$\text{Specificity} = \frac{TrueNegative}{FalsePositive + TrueNegative} \quad (1)$$

5.3 性能評価実験

3.2 節の性能評価用データセットで 10 分割交差検証を行った．分割は 訓練 : 検証 : 評価 = 8 : 1 : 1 である．評価指標は各モデル毎の平均値を算出した．

5.4 実タスクを想定した実験

3.2 節の実タスクを想定したデータセットを使用し，ベースライン，Radiology-BERT，Radiology-RoBERTa での実験を行った．初期値の影響を考え，各モデル毎に 5 回実験した．評価指標は各モデル毎の平均値を算出した．

また，大規模事前学習言語モデルベースの検出器の判断根拠を調査するため，SHapley Additive exPlanations (SHAP) [14] を用いて検出に影響を与える上位 9 単語を算出した．SHAP には各モデル毎のベストモデルを使用した．

6 結果・考察

6.1 性能評価結果

実験結果を表 1 に示す．ベースラインと比較して大規模事前学習言語モデルベースの検出器が優れた結果を出しており，本タスクにおける大規模事前学習言語モデルの有効性を示している．事前学習済み BERT の中では Radiology-BERT が最も優れており，事前学習データのドメインによる影響が精度向上に大きく寄与している可能性が高い．Radiology-RoBERTa が Accuracy スコア 0.918 と最高精度を達成しており，提案モデルの有効性を示している．

6.2 実タスクを想定した実験結果

実験結果を表 2 に示す．全体の傾向は 6.1 節と同様であり，提案モデルである Radiology-RoBERTa の有効性を示している．

SHAP による可視化結果を図 2 と図 3 に示す．

SHAP value に着目して述べる．陽性と比べて陰性の SHAP value が小さく，陰性を示す単語は少ないことを示唆している．Radiology-RoBERTa よりも Radiology-BERT の SHAP value が高く、一単語辺りに含まれる情報量の違いが起因すると考えられる．

具体的な単語に着目して述べる．Radiology-BERT は「肺」が含まれている単語が複数上位に含まれており、陽性の判断に診断部位を重視している．対して Radiology-RoBERTa の上位 9 単語には肺の症状が複数含まれており、陽性の判断に肺の症状を重視している．これより、Radiology-RoBERTa がより妥当な分類をしていることが示唆される．

7 おわりに

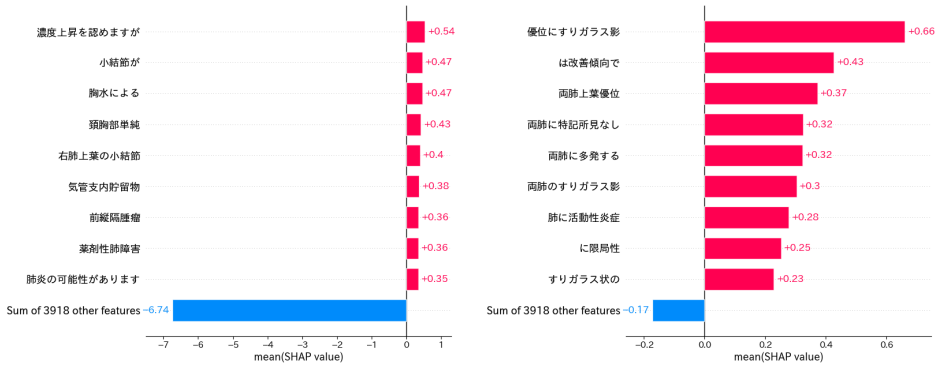
本論文では，所見文を用いる COVID-19 肺炎の自動検出器を構築した．データセットは多くないため，検出器には複数の大規模事前学習言語モデルを用い比較した．提案手法はベースラインより優れ，本タスクにおける大規模事前学習言語モデルの有効性を示した．なかでも，提案モデルである Radiology-RoBRTa は Accuracy スコア 0.918 と最高精度を達成した．また，SHAP を用いて検出に影響を与える上位 9 単語を算出した．この結果から，提案モデルが肺の症状を重視することが示唆された．

表 1　性能評価結果

	Accuracy	Precision	Recall	F_1	Specificity
ルールベース	0.637	0.587	**0.909**	0.713	0.369
BiLSTM	0.848	0.877	0.811	0.841	0.885
TOHOKU-BERT	0.904	0.916	0.889	0.902	0.918
TOHOKU-BERT-char	0.891	0.902	0.878	0.889	0.904
UTH-BERT	0.906	0.907	0.904	0.905	0.909
SP-BERT	0.893	0.902	0.881	0.891	0.905
Radiology-BERT	0.913	**0.932**	0.891	0.911	**0.935**
Radiology-RoBERTa (Ours)	**0.918**	0.928	0.906	**0.916**	0.931

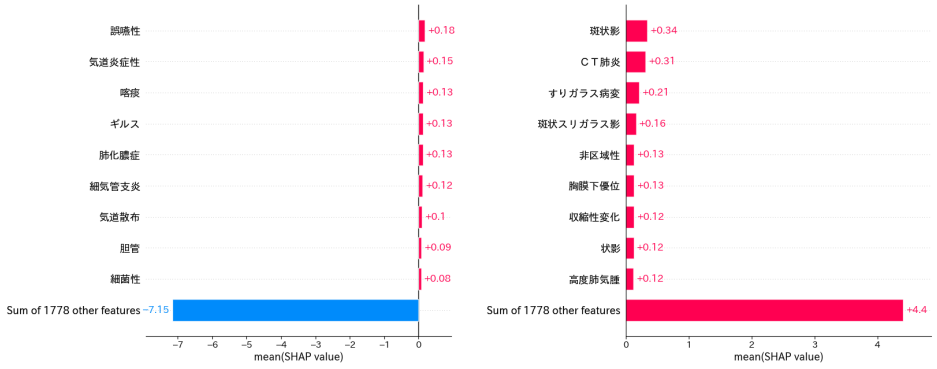
表 2　実タスクを想定した実験結果

	Accuracy	Precision	Recall	F_1	Specificity
ルールベース	0.670	0.631	0.820	0.713	0.520
BiLSTM	0.867	0.854	0.884	0.869	**0.849**
Radiology-BERT	0.887	0.852	0.939	0.893	0.836
Radiology-RoBERTa (Ours)	**0.893**	**0.855**	**0.949**	**0.899**	0.837



(a) 陰性検出に影響する上位 9 単語　　(b) 陽性検出に影響する上位 9 単語

図 2　Radiology-BERT の検出精度に影響する単語の可視化



(a) 陰性検出に影響する上位 9 単語　　(b) 陽性検出に影響する上位 9 単語

図 3　Radiology-RoBERTa の検出精度に影響する単語の可視化

謝辞

本研究は AMED の JP20lk1010036 課題番号の支援を受けたものです.

参考文献

[1] 明石敏昭, 待鳥詔洋, 青木茂樹. COVID-19 肺炎に対する日本医学放射線学会の対応と画像診断 AI への期待. **Medical Imaging Technology**, Vol. 39, No. 1, pp. 3–7, 2021.

[2] 多田太郎, 森川みどり, 那須照広, 山本和英. 読影レポート間の類似度データセットの構築と予備実験. 言語処理学会第 26 回年次大会発表論文集, pp. 1193–1196, 2020.

[3] Ryosuke Kuwabara, Changhee Han, Kohei Murao, and Shinichi Satoh. BERT-based few-shot learning for automatic anomaly classification from Japanese multi-institutional CT scan reports. **International Journal of Computer Assisted Radiology and Surgery**, Vol. 15, No. 1, pp. 148–149, 2020.

[4] Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Takahiro Nakao, Soichiro Miki, Takeyuki Watadani, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. **BMC Medical Informatics and Decision Making**, Vol. 21, No. 1, pp. 1–19, 2021.

[5] 本田修平, 大竹義人, 高尾正樹, 荒牧英治, 矢田竣太郎, 合田憲人, 佐藤真一, 橋本正弘, 明石敏昭, 菅野伸彦, 佐藤嘉伸. 大規模放射線読影レポートデータベースによる BERT モデルの事前学習とそれを用いた CT 画像の撮影目的の推定. 日本医用画像工学会大会予稿集, 39 回, pp. 56–56, 2020.

[6] 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45. Association for Computational Linguistics, 2020.

[8] 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.

[9] 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 **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. European Language Resources Association (ELRA), 2018.

[10] Yohei Kikuta. BERT Pretrained model Trained On Japanese Wikipedia Articles, 2019. https://github.com/yoheikikuta/bert-japanese.

[11] 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 preprint arXiv:1907.11692**, 2019.

[12] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed Precision Training. **arXiv preprint arXiv:1710.03740**, 2018.

[13] Joeri R. Hermans, Gerasimos Spanakis, and Rico Möckel. Accumulated Gradient Normalization. In **Proceedings of the Ninth Asian Conference on Machine Learning**, pp. 439–454. Proceedings of Machine Learning Research, 2017.

[14] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.