

# 読影レポートの知識を埋め込んだ言語モデルを作成するための事前学習手法

中野騰久<sup>1</sup> 田川裕輝<sup>1</sup> 尾崎良太<sup>1</sup> 谷口友紀<sup>1</sup> 大熊智子<sup>1</sup>  
鈴木裕紀<sup>2</sup> 木戸尚治<sup>2</sup> 富山憲幸<sup>2</sup>

<sup>1</sup> 富士フイルム株式会社 <sup>2</sup> 大阪大学大学院医学系研究科  
{norihisa.nakano, yuki.tagawa, ryota.ozaki,  
tomoki.taniguchi, tomoko.ohkuma}@fujifilm.com  
{y-suzuki, kido, tomiyama}@radiol.med.osaka-u.ac.jp

## 概要

様々な自然言語処理タスクにおいてニューラル言語モデルが優れた性能を実現している。本論文では読影レポートというドメインに特化した言語モデルを作成するための手法を提案する。一般的なドメインのテキストではなく、読影レポートテキストを用いて事前学習をゼロから行う。また事前学習時に行う Masked Language Model において、入力内の固有表現と固有表現間の関係性に着目してマスク対象と入力範囲幅を決定することで、モデルにドメイン知識を埋め込む。上記手法によって作成した言語モデルを臓器分類、固有表現抽出、関係性判断の三つのタスクにおいて評価した結果を報告する。後段タスクと同一のドメインコーパスで言語モデルを作成する有用性を示し、提案手法によって臓器分類と関係性判断の二つのタスクで性能向上を確認した。

## 1 はじめに

読影レポートとは放射線科医が CT や MRI などの医療画像から読み取った異常所見の有無や疾患の症状について記述したものである。読影レポートには医師の知見が含まれており、これらを抽出することで様々な用途に利活用することが期待できる。そのためには読影レポートを構造化し、再利用可能な形式に変換することが必要である。我々はより高精度な情報抽出による構造化を図るため、読影レポートに最適な言語モデルの構築を目指している。

Bidirectional Encoder Representations from Transformers(以後,BERT)[1] が発表されて以降,transformer[2] をベースとしたニューラル言語モデルが優れた性

能を実現している。特定ドメインのタスクにおいては、そのドメイン文書を用いて事前学習したモデルを使用することが望ましく、医療分野においては Bio-BERT[3] や clinical-BERT[4] などが挙げられ、日本語においては UTH-BERT[5] が発表されている。本論文で取り扱う読影レポートにおいても、臓器や疾患などドメイン固有の語句が多く存在する。読影レポートを用いて事前学習をゼロから行うことで、言語モデルはドメイン固有の語句を上手く表現できるようになり、読影レポートを取り扱うタスクにおいて性能向上に期待できる。

BERT の事前学習では入力の一部をマスクし、前後の文脈からマスクされた語句を予測する Masked Language Model(以後,MLM) と呼ばれるタスクを行う。しかし従来手法では入力からランダムにマスク対象を選択するため、特に大規模なコーパスの入手が難しいドメイン領域の場合、そのドメイン固有の重要な語句の学習機会の損失が指摘される。また医師は部位や病変、病変の性状から疾患を導出し、読影レポートを記述する。MLM を行う際に、このような病変と部位や病変と疾患などの関係性に基づき、読影レポート固有の重要語句を中心にマスク対象とすることで言語モデルにドメイン知識を埋め込むことができる。加えて読影レポートの所見文は臓器毎に記述されるため,MLM における入力範囲幅を臓器単位にまとめることで、マスク対象とする固有表現を予測するための文脈情報を確保する。

本論文では読影レポートというドメイン領域において、ドメインコーパスを用いてゼロから事前学習すること、また言語モデルにドメイン知識を埋め込むように事前学習を行うことで、より高性能な言語

モデルの作成を目標とし、後段タスクである臓器分類、固有表現抽出、関係性判断の3つのタスクにおいて評価を行う。

## 2 関連研究

Lin ら [6] は MLM 時に固有表現を中心にマスク処理を行う Entity-BERT を提案した。Entity-BERT は、PubMedBERT[7] をベースにさらに臨床テキストである MIMIC[8] を用いて継続的に学習させ、この時入力に含まれる固有表現とそれ以外の語句から、それぞれ設定した割合でランダムにマスク対象を選択する。彼らはその割合について複数の組み合わせを比較した結果、固有表現のみではなく、それ以外の語句も同時にマスク対象とした方がモデルの性能が良くなると報告しており、固有表現以外の語句は意味情報に加え、構文情報を持つと説明した。またマスク対象とする割合を大きくしすぎると、学習が困難になる可能性についても報告している。適切な割合で固有表現を中心にマスク処理を行うことで、ドメイン固有の重要な表現を学習する機会が増え、三つの臨床テキストタスクにおいて従来手法によって作成された言語モデルよりも優れた性能を発揮することを示した。

## 3 データ

大阪大学医学部附属病院から提供された約 11 年分の匿名化済み読影レポートを用いて言語モデルの作成と評価を行った。言語モデルの作成には約 670 万文、性能評価には言語モデル作成用コーパスに含まれない約 9 千文を使用した。全ての読影レポートに対して予め、文毎に臓器ラベルの付与、固有表現の抽出と固有表現ラベルの付与、抽出された固有表現間の関係性ラベルの付与を行った (図 1)。

臓器ラベルは肺や心臓などの主要な臓器からなる全 31 種のラベルを、固有表現ラベルは解剖区域、病名、病変、性状、変化、形状変化、計測項目、計測結果、撮影条件からなる全 9 種のラベルを定義した。膨

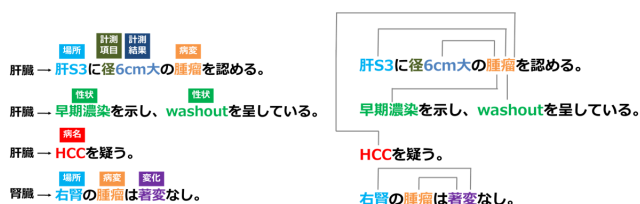


図 1 ラベル付与例

左: 臓器ラベルと固有表現ラベル 右: 関係性ラベル

表 1 コーパスにラベルを付与する際のモデル性能  
臓器分類 固有表現抽出 関係性判断

| dev / test | 0.913 / 0.912 | 0.870 / 0.852 | 0.960 / 0.957 |
|------------|---------------|---------------|---------------|
|------------|---------------|---------------|---------------|

大なデータに人手でラベルを付与するのは困難なため、読影レポートの一行を一入力として作成した BERT モデルを用いて各タスクの推論を行い、言語モデル作成用コーパスにラベルを付与した。推論に用いたモデルの性能を表 1 に示す。

## 4 提案手法

先行研究である Lin ら [6] はドメイン固有の語句を中心にマスクすることの有用性を示したが、マスクする固有表現の選別や、モデルに入力するテキスト量については言及していない。MLM において従来は一文もしくは二文を入力としており、同一入力内の文脈情報のみではマスクされた語句を予測するには不十分である可能性が指摘される。そのような場合にドメイン領域にとって重要な固有表現の単語ベクトル構築に悪影響を及ぼすと考えられる。我々はマスク対象とする固有表現を予測する手がかりとして、マスク対象と関係のある固有表現が効果的であると考え、マスク対象を同一入力内に関係を持つ固有表現を有するものに限定する。また読影レポートの所見文は臓器毎にまとめて記述されるため、MLM における入力を臓器単位にまとめることでマスク対象とする固有表現を予測するための文脈情報を確保する。提案手法におけるマスク対象の決定フローを図 2 に示す。

マスクされた固有表現と関係を持つ固有表現を含む文脈情報を用いて、言語モデルはマスク対象を予測することができる。これによって読影レポートにおける部位や病変、病変の性状から疾患を導出する医師の診断プロセスに似た学習を行い、モデルにドメイン知識を埋め込む。

## 5 実験

### 5.1 設定

言語モデルの事前学習には約 670 万文の読影レポートテキストを使用した。語彙は、NEologd 辞書<sup>1)</sup>を組み込んだ MeCab で単語分割した後、文字単位のサブワードに分割し 3852 の語彙を構築した。バッチサイズは 32、入力の最大長は 512 で 100 万ステップ

1) <https://github.com/neologd/mecab-ipadic-neologd>

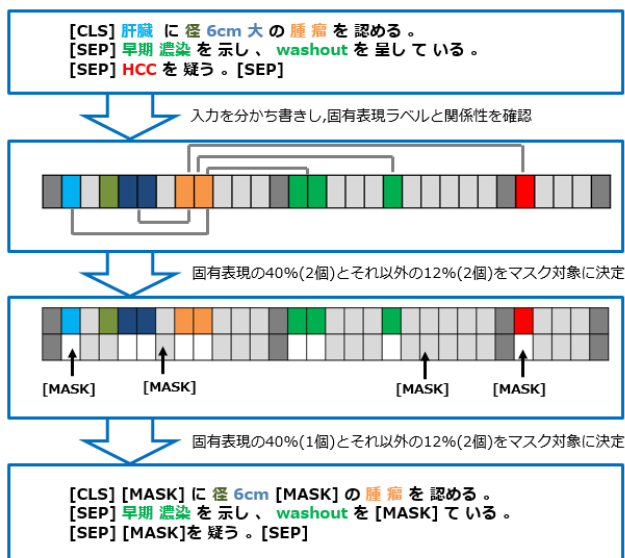


図2 提案手法におけるマスク対象決定フロー

学習を回した。本実験では以下に述べる四つのマスク対象の決定手法を比較する。

- **1, random)** 従来手法と同様に入力内の 15 % をランダムに選択する。
- **2, Entity)** Lin ら [6] と同様に入力内の固有表現の 40 % と、それ以外の語句の 12 % をランダムに選択する。
- **3, Relation)** 入力内の固有表現の内、同一入力内の固有表現間において関係を有する固有表現の 40 % と、それ以外の語句の 12 % をランダムに選択する (図 2)。

特殊トークンである [CLS],[SEP] はマスク対象から除外し、マスク対象となる候補を選択した後は従来手法と同様に,[MASK] トークンへの置換, 変更なし, ランダム他の語句への置換をそれぞれ 80 %,10 %,10 % の比率で行う。モデルへの入力範囲は文単位と臓器単位の二つを比較する。臓器単位の入力, 臓器分類によって推論された臓器ラベルを基に, 同じラベルが付与された連続する読影レポートテキストをまとめることで作成した。

また Liu ら [9] の報告に基づき, 本実験の事前学習では Next Sentence Prediction は行わなかった。

## 5.2 性能評価方法

作成した言語モデルの性能評価には, 臓器分類, 固有表現抽出, 関係性判断の三つの自然言語処理タスクにファインチューニングして行う。言語モデル作成に使用しない 9 千文の読影レポートテキストに

対して人手でアノテーションし, dev と test のデータセットを作成した。また日本語 BERT の比較対象として, 日本語 Wikipedia で学習された東北大学乾研究室の BERT モデル [10] と日本語臨床テキストで学習された UTH-BERT[5] を使用した。

**臓器分類:** 入力されたテキストがどの臓器もしくは部位に属するものかを判断する 31 クラスの多クラス分類タスクである。読影レポートは対象となる一文のみでは部位や臓器を特定できないような文が存在するため, 判断対象となるテキスト  $x_i$  とその前文である  $x_{i-1}$  を同時にモデルに入力する。BERT モデルに分類層を追加したネットワークを用いて, 最終層の先頭である [CLS] トークンを特徴量として判断する。

**固有表現抽出:** 入力されたテキストから固有表現を抽出するタスクである。固有表現に対して 9 種類のラベルを定義している。読影レポート内において同一の臓器について記述されたテキストを同時にモデルに入力する。手法は田川ら [11] の報告に基づき, BERT-CRF を採用した。

**関係性判断:** 固有表現抽出によって抽出された固有表現間に関係性があるか否かを判断するタスクである。BERT モデルに分類層を追加したネットワークを用いて, 対象となる二つの固有表現とその間の文脈を特徴量として判断する。

## 5.3 結果

表 2 に dev データセットと test データセットに対する実験結果を示す。値は Micro-F1 値である。

ドメインコーパスを用いた言語モデルは既存モデルよりも高い性能を発揮することが確認できた。事前学習で用いた読影レポート (670 万文, 0.3GB) は既存モデルのコーパスである日本語 wikipedia (3000 万文, 4GB) や臨床テキスト (1 億 2000 万行) と比較して少量であったが, 後段タスクと同一ドメインで学習することの有用性が示された。固有表現間の関係性に着目したマスクの決定手法によって, 臓器分類と関係性判断のタスクにおいて性能向上を確認したが, 固有表現抽出タスクにおいては効果が見られなかった。また入力を臓器単位にまとめることでほとんどの場合に性能向上が確認された。

## 6 分析

**未知語の影響:** 本実験の事前学習で使用したコーパスに含まれる固有表現は全部で約 16 万種類, 1800



表 2 実験結果 (dev/test)

|               | マスク方法  | 臓器分類          | 固有表現抽出        | 関係性判断         |
|---------------|--------|---------------|---------------|---------------|
| 日本語 wiki-BERT | random | 0.876 / 0.859 | 0.851 / 0.828 | 0.950 / 0.946 |
| UTH-BERT      | random | 0.829 / 0.831 | 0.860 / 0.840 | 0.932 / 0.892 |

| 入力単位                  | マスク方法    | 臓器分類                 | 固有表現抽出               | 関係性判断                |
|-----------------------|----------|----------------------|----------------------|----------------------|
| sentence <sup>1</sup> | random   | 0.902 / 0.895        | 0.868 / 0.855        | 0.959 / 0.960        |
| sentence <sup>1</sup> | Entity   | 0.904 / 0.898        | 0.867 / 0.843        | 0.958 / 0.942        |
| sentence <sup>1</sup> | Relation | 0.906 / <b>0.909</b> | 0.860 / 0.837        | 0.962 / 0.956        |
| organ <sup>2</sup>    | random   | 0.902 / 0.905        | <b>0.877 / 0.863</b> | 0.961 / <b>0.964</b> |
| organ <sup>2</sup>    | Entity   | 0.914 / 0.905        | 0.862 / 0.845        | 0.962 / 0.962        |
| organ <sup>2</sup>    | Relation | <b>0.915</b> / 0.908 | 0.865 / 0.842        | <b>0.963</b> / 0.963 |

1: 読影レポートの一文を一入力としたもの

2: 同一臓器について書かれた連続する文をまとめて一入力としたもの

万個であった。読影レポートを用いて作成した言語モデルには、本コーパスを用いて語彙を作成しているため未知語は現れない。しかし既存モデルで使用されるトークナイザーでは固有表現を入力した際に、半数以上である約 8.5 万種類の固有表現に未知語である [UNK] トークンが出現する。同じく評価用データには約 2400 種、2 万個の固有表現が含まれており、既存モデルにおいては約 670 個の固有表現に未知語が出現する。UTH-BERT の場合はトークナイズ時に万病辞書<sup>2)</sup>を用いているが、未知語の出現率に大きな差異は認められなかった。同じ医療分野であっても臨床テキストと読影レポートでは現れる固有表現が大きく異なると考えられる。

**入力範囲幅の影響:** 入力範囲幅を臓器単位にまとめることで性能の向上が確認できた。臓器単位の入力は全体の約 36 %が一入力内に二文以上のテキストを保有する(二文:17 %, 三文:7 %, 四文以上:12 %)。関連するテキスト群を受け取ることで、言語モデルが読影レポートの複数文に跨る構文情報を獲得できたことが示唆される。また本論文における評価タスクの入力が複数文に跨るものが多いため、事前学習時と後段タスクの入力が近いものとなっていることも原因として考えられる。

**コーパス内のノイズの影響:** 固有表現を中心にマスク対象とすることで、二つの自然言語処理タスクについて性能向上に期待できるが、固有表現抽出タスクにおいてはその効果が見られない。これはコーパスへのラベルの付与を行ったモデルの固有表現抽

出タスクの性能が他の二つのタスクの性能と比較して低く(表 1)、事前学習用コーパスに付与された固有表現に関する情報にノイズが多いことが一因として考えられる。よって精度の高い固有表現ラベルに絞ってマスク対象の候補とするなど、より高品質な情報を持つコーパスを用いることで改善が見込める。本実験では全てのタスクにおいて、コーパスへのラベルの付与を行ったモデルよりも高い精度のモデルが作成できている。よって作成したモデルを用いてコーパスへラベルを付与し、再度言語モデルを作成する、というサイクルを繰り返すことでより性能の高い言語モデルを作成することが可能と考える。

## 7 おわりに

本論文では、読影レポートに適した言語モデルの作成手法を提案し、臓器分類、固有表現抽出、関係性判断の三つのタスクにおいて性能を評価した。結果から事前学習に用いるコーパスが後段タスクと同一のドメインであることの有用性を示した。また言語モデルの事前学習である MLM において、臓器単位にまとめられたテキストを一入力とすることでモデルの性能が向上することを示し、さらにマスク対象を同一入力内に関係を持つ固有表現を有する固有表現に限定することで、二つのタスクで性能向上が確認された。今後の課題として、精度の高い固有表現ラベルに絞ってマスク対象を決定するなど、コーパスにおけるノイズの影響を調査することや最適な入力範囲幅やマスク選択割合の探索などが挙げられる。

2) <http://sociocom.jp/data/2018-manbyo/index.html>

---

## 参考文献

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.
- [3] Biobert: a pre-trained biomedical language representation model for biomedical text mining. 2019.
- [4] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020.
- [5] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. 2021.
- [6] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. 2021.
- [7] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pre-training for biomedical natural language processing, 2020.
- [8] Shen L Lehman LW Feng M Ghassemi M Moody B Szolovits P Celi LA Mark RG Johnson AE, Pollard TJ. MIMIC-III, a freely accessible critical care database.
- [9] 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, 2019.
- [10] Masatoshi Suzuki and Kentaro Inui. Bertbase-mecab-ipadic-bpe-32k-whole-word-mask. <https://github.com/cl-tohoku/bert-japanese>.
- [11] 田川裕輝, 西埜徹, 谷口元樹, 谷口友紀, 大熊智子, 若宮翔子, 荒牧英治. 生成された読影所見の自動評価に向けた固有表現認識とモダリティ推定, 2020.

表3 固有表現ラベルと詳細

| ラベル  | 定義  |
|------|---|
| 解剖区域 | 主には臓器などの部位を示す表現。「末端」、「底部」などの画像中の位置を表す名詞も含まれる。                                   |
| 病変   | 画像情報を客観的に観察して得られる病変情報である「腫瘤」、「結節」など。「認める」、「認めない」など断定的な表現で記述される。                 |
| 病名   | 観察された病変から診断される病名である「肺炎」、「肺癌」など。「疑われる」など断定的ではない表現で記述される。                         |
| 変化   | 「著変」、「増大」など。前回診断との比較から得られた観察結果を指す。  |
| 形状変化 | 臓器や部位の形状の変化に関する表現。臓器に対する「腫大」、「萎縮」など。  |
| 計測結果 | 計測された大きさや値に関する表現。「15mm」や「8mm大」など。   |
| 計測項目 | 病変のサイズや造影剤の吸収量などの計測した項目に関する表現。  |
| 撮影条件 | 撮像方法や撮像タイミングに関する表現。「単純CT」、「早期相」など。  |
| 性状   | 病変、形状変化の状態や性質といった特徴に関する表現。「(性状)を伴う(病変)」といった文脈や「～性」、「～状」という表現で病変と複合して記述されることが多い。 |

## Appendix

### データセットについて

本論文で利用した固有表現ラベルに関する説明を表3に示す。病変、変化、性状に関しては実際にそれが生じているのか否かの事実性についても固有表現抽出のタスク時に推定するために、固有表現ラベルに事実性を表す Positive と Negative を結合したもの(例えば病変 P や変化 N など)をラベルとして用意した。また関係性を考慮する組み合わせは(解剖区域, 病変), (性状, 病変), (病名, 病変), (変化, 病変), (計測結果, 病変), (解剖区域, 形状変化), (性状, 形状変化), (病名, 形状変化), (変化, 形状変化)の9パターンである。計測項目と撮影条件は関係性を持たないため、提案手法においてはマスク対象の候補から除外される。

またコーパス内の固有表現ラベルの分布を図3に示す。部位を表す解剖区域、読影のターゲットとなる病変、病変を説明する性状の順に多い。本論文の提案手法では、マスク対象の候補を抽出した後は固有表現ラベルの種類に関わらずランダムにマスク箇所を決定するため、重要であるが出現頻度の少ない病名に関する学習が不足している可能性がある。

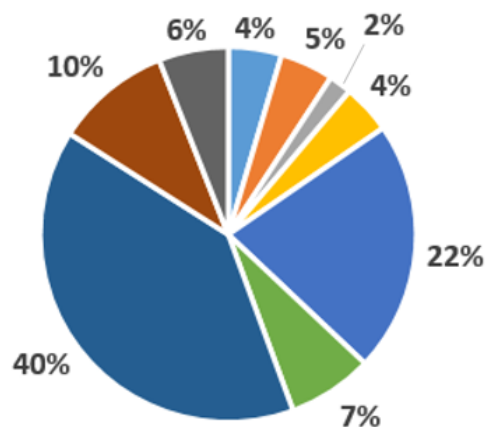


図3 コーパス内の固有表現ラベル分布