# BERTを用いた医療文書からの固有表現抽出

○五井野 琢也　濱上 知樹 (横浜国立大学)

## Named Entity Recognition from Medical Documents by Fine-Tuning BERT

∗Takuya Goino and Tomoki Hamagami (Yokohama National University)

**Abstract**− In recent years, electronic health records(EHRs) have rapidly become widespread.Large-scale medical data contains useful information and is one of the data that is expected to be utilized. The description content of EHRs contains many non-grammatical and fragmented expressions. The technology that automatically recognizes named entities in a document is called Named Entity Recognition, and is essential for maximizing the use of document data and applying it to various tasks of natural language processing. In this research, in addition to extracting disease names and symptoms from EHRs using BERT, we also identified modality (5 types) that represent the subjectivity of speakers and writers. The results were compared using multiple BERT models pre-trained with the Japanese corpus.

**Key Words:** electronic health records(EHRs), Named Entity Recognition, BERT

## 1　はじめに

医療文書の電子化が急速に普及している．特に電子カルテは大学病院規模の病院で月に２０万以上もやり取りされるという統計がある [1]．大規模な医療データには有用な情報が含まれ，最も活用が期待されているデータの 1 つである．電子カルテの記述内容については, 医療従事者に任されてきた面から非文法的, かつ断片化した表現が多く含まれた自然言語文で構成される．医療文書中の固有表現（病名，症状など）を自動で認識する技術は固有表現抽出 (NER) と呼ばれ，文書データを最大限に活用し類似症例検索，診断支援などの自然言語処理の様々なタスクに応用していく上で必要不可欠な基礎技術となっている．

日本語は漢字，ひらがな，カタカナの混合表記を用い，かつ外来語を多く輸入するため表記ゆれ (同一概念を指す複数の表記群) の問題が他の言語に比べて大きい．さらに，電子カルテは多忙な業務の合間に作成されるため，略語表記が頻出し，かつ書き間違え，打ち間違え，外国語表記 (英語，ドイツ語など) や記号表記が頻用され，一層の複雑さを持っている [1]．そのため，辞書やルールを基に固有表現を抽出するルールベースの手法には限界があり，機械学習を用いた手法が用いられている．

近年，汎用言語モデル BERT[2] が自然言語処理タスクで精度向上に寄与しているが，大規模コーパスで事前学習を行うモデルのため，コーパスや処理単位の影響が大きい．

本研究では，小規模ではあるが病名，症状の固有表現ラベルと 5 種類のモダリティラベルをアノテーションしたデータセットを作成し，日本語のコーパスで異なる条件の事前学習を行った 4 種類の汎用言語モデル BERT を用いて，NER とモダリティ推定を行い精度を比較した．モダリティ推定とは病名，症状が実際に生じているか否かの事実性を分類するタスクである．事実性の多くは用言部に付随するモダリティと呼ばれる言語表現によって判別されるためモダリティ推定と呼ばれる．医療文書には否定表現 (Negative) が多く現れ、医師がどのように考え診察したかが分かる場合がある．データセットは BERT モデルと同じ tokenizer を利用し，トークン化したトークンに対してアノテーションを施した．

## 2　関連研究

日本語の医療文書を対象とした NER の研究がある．Yano[3] は双方向 LSTM と 出力層に CRF を用いた深層ニューラルネットワークによる文字ベース系列ラベリングを用いて，疾患名の識別と疾患のモダリティ推定に取り組んでいる．田川ら [4] は読影所見を対象に 8 種類の固有表現と 3 種類のモダリティラベルを設計し，基本的な系列ラベリング手法の精度比較を行い BERT+CRF が最も高い精度を獲得したことを報告している．また，訓練したモデルを利用し生成した読影所見の評価を行っている．

## 3　データセットの作成

NTCIR-11[5] MedNLP-2 タスクで仮想の患者を想定して医師により記述された，模擬患者の病歴報告を使用したテストコレクションと NTCIR-12[6] MedNLPDoc で診療情報管理士のためのテキストである「ICD コーディングトレーニング第 2 版」をデータとして使用したテストコレクションをデータセットとして用いる．NTCIR-11 のデータセットは病名・症状とそのモダリティのアノテーションが付与されている．NTCIR-12 に対しては非医療従事者が NTCIR-11 のデータセットを基に手動でアノテーションを付与した．アノテーションに用いた 5 種類のモダリティラベルの詳細を Table 1 に示す．

Table 1: Modality types

| ラベル | 説明 |
|---|---|
| Positive | 実際に患者が患っている |
| Negative | 患っていない |
| Suspicion | 患っている疑い、可能性 |
| Family | 患者の家族の病気 |
| Family+Negative | 患者の家族に無い病気 |

NER として系列ラベリングを行う．系列ラベリングは形態素解析によってトークン化されたデータ列（文字列，単語列）を受け取り，出力として個々のトークンに

対して固有表現ラベルを付与するタスクである．IOB2（Inside-outside-beggining）方式を採用し固有表現のラベルを付与する．その例を Table 2 に示す．##はサブワードに分割された単語である．B は固有表現のトークン開始位置，I は固有表現の継続，O はトークンが固有表現でないことを意味する．

Table 2: Examples of IOB2 labeling

| トークン (文) | IOB2 タグ |
|---|---|
| 頻 | B-Positive |
| 拍 | I-Positive |
| 傾向 | I-Positive |
| の | O |
| 心 | B-Positive |
| ##房 | I-Positive |
| 細 | I-Positive |
| 動 | I-Positive |
| を | O |
| 有する | O |
| 慢性 | B-Positive |
| 心不全 | I-Positive |
| ( | O |
| 虚 | B-Suspicion |
| 血 | I-Suspicion |
| 性 | I-Suspicion |
| 心 | I-Suspicion |
| 疾患 | I-Suspicion |
| の | O |
| 可能 | O |
| 性 | O |
| あり | O |
| ) | O |

Table 3: Statistics of the created dataset

| モダリティ | 件数 (+追加数) |
|---|---|
| Positive | 1359 |
| Negative | 517 |
| Suspicion | 13(+50) |
| Family | 166 |
| Family+Negative | 3(+50) |
| 文数 | 1916(+95) |

作成したデータセットの統計量を Table 3 に示す．モダリティの Suspicion と Family+Negative はデータを作成するために用いたデータセットに十分なラベル数が無かったため，使用したデータセットの文を模倣しダミーデータとして 50 個ずつ固有表現を含む文を作成し追加した。

## 4 実験手法

本研究では，事前学習済みの BERT モデルの最終層に CRF を追加し正解ラベルを出力させるように Fine-tuning を行った．

### 4.1 CRF

CRF(Conditional Random Field) は，系列ラベリングを解くために用いられる手法である．入力系列 $X = x_0, x_1, ..., x_n$ が与えられたとき，出力ラベル系列が $Y = y_0, y_1, ..., y_n$ となる条件付確率 $P(Y|X)$ を以下の式で定義する．

$$P(Y|X) = \frac{1}{Z_X} \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, X)\right) \quad (1)$$

$Z_X$ は全てのラベル系列を考慮したときに確率の和が 1 になるようにするための正規化項で

$$Z_X = \sum_{Y} \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, X)\right) \quad (2)$$

である．$f_k$ は素性関数，$\lambda_k$ は素性関数 $f_k$ に対する重みで，$P(Y|X)$ の対数尤度を最大化ように学習する．学習した識別機を用いて，入力系列 $X$ に対する最適な出力ラベル系列 $y*$ は以下の式で決定する．

$$y* = \arg\max_{Y} P(Y|X) \quad (3)$$

### 4.2 BERT

BERT（Bidirectional Encoder Representation from Transformer）は，Attention 機構のみで構成される Encoder-Decoder モデルである Transformer[7] の Multi-Head Attention(Encoder 部分) を用いて双方向に学習を行うモデルである．Fig. 1 に BERT の構成を示す．E は入力トークンの埋め込み，Trm は Transformer の Encoder 部分，T は最終的に出力される文脈を考慮したトークンの埋め込みである．

BERT は教師なしの大規模な文章データを用いて，ランダムに選んだトークンをマスクし，マスクされたトークンを周りの文脈から予測する Masked Language Model と，2 文を入力し連続する文か否かを判定する Next Sentence Prediction という 2 つのタスクで事前学習を行う．事前学習された BERT をタスクごとの教師ありデータで Fine-tuning することで自然言語処理の複数のタスクで高い精度を獲得している．

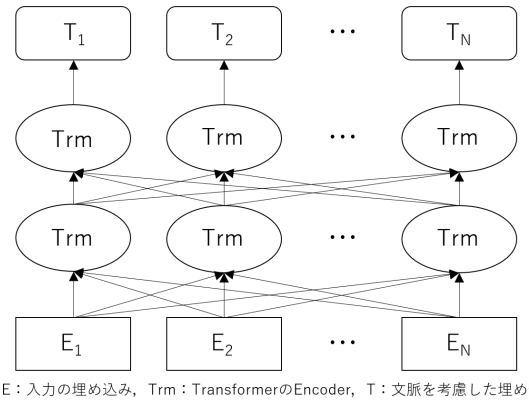


Fig. 1: BERT model architecture.

## 5 実験

### 5.1 実験概要

BERT+CRF を用いて病名、症状とそのモダリティ(5 種類) の NER とモダリティ推定を行った。また、事前学習コーパスとトークナイズが異なる 4 種類の BERT を用いて結果を比較した．

Table 4: Pre-training of BERT model

| Model | Pre-training corpus | Tokenization | Vocaburaries |
|---|---|---|---|
| tohoku-char-BERT | Wikipedia | character | 4,000 |
| tohoku-BERT | Wikipedia | Mecab(IPA 辞書) + Wordpiece | 32,000 |
| tohoku-wwm-BERT | Wikipedia | Mecab(IPA 辞書) + Wordpiece | 32,000 |
| UTH-BERT | Clinical text | Mecab(NEologd，万病辞書) + Wordpiece | 25,000 |

### 5.2 実験設定

本研究では、Table 4 に示す BERT を用いて比較を行った．東北大学が公開している日本語の Wikipedia で事前学習をを行った 3 種類の BERT(tohoku-char-BERT，tohoku-BERT,tohoku-wwm-BERT とする)[8] と，東京大学が公開している日本語の医療文書で事前学習を行った UTH-BERT[9] である．tohoku-wwm-BERT はトークンに対応するサブワードを含めてマスクする Whole Word Masking を行い Masked Language Model で事前学習を行う．また，UTH-BERT は MeCab[10] の辞書に医療分野の専門用語の切り出しを目的とした万病辞書 [11] を用いている．BERT+CRF の実装には Transformers[12] を利用した．上記のモデルを Batch size=32，学習率=$5 \times 10^{-5}$，epoch 数=15 で Fine-tuning した。

精度評価には 5 分割交差検証を用いた。データセットを 5 分割し，5 分の 4 を学習データ，5 分の 1 の半分を検証データ，もう半分をテストデータとする．各交差で学習データに対する Micro-F1 が最も高い epoch 数のモデルを採用した．

### 5.3 評価指標

適合率 (Precison)，再現率 (Recall)，F1(F-measure) を用いた。指標の算出にはトークナイズされたトークンに対して正しいラベルが付与されたか否かで識別する CoNLL-2003[13] と同様の方法を利用した．Micro-F1 は全てのラベルで真陽性，偽陽性，偽陰性，真陰性を集計しそれらを用いて算出する．

### 5.4 実験結果

Tble 5 に抽出した固有表現 (病名，症状) の評価結果を示す．各交差ごとに Micro-F1 を算出しマクロ平均をとった．UTH-BERT の Micro-F1 値が最も高かった．また、東北大学の BERT モデルの比較から，NER においてトークナイズは単語単位よりも文字単位の方が精度が高いことが示された．

Table 5: Comparison of Named Entity F1

| Model | Micro-F1 |
|---|---|
| tohoku-char-BERT | 0.919 |
| tohoku-BERT | 0.894 |
| tohoku-wwm-BERT | 0.907 |
| UTH-BERT | **0.931** |

Table 6 にモダリティラベルごとの評価結果を示す．各交差ごとにそれぞれのモダリティラベルに対して適合率，再現率，F1 を算出しマクロ平均をとった．Positive と Suspicion では tohoku-char-BERT の精度が高く、それ以外のラベルでは UTH-BERT の精度が最も高かった．特に単語単位の分割を行った東北大学の BERT が Family ラベルで精度が低くなった．

また，ダミーデータとして固有表現を 50 個ずつ追加した Suspicion と Family+Negative は，件数が少ないにも関わらず高い評価結果となっているが，作成したダミーデータが短文，単調であったことが原因だと考えられる．

## 6 考察

### 6.1 未知語に関する考察

作成したデータセットを BERT モデルの tokenizer を利用し分割した際に，未知語として扱われたトークンの割合は tohoku-char-BERT で 0.16%，tohoku-BERT，tohoku-wwm-BERT で 1.37%，UTH-BERT で 0.31% であった．特に tohoku-BERT，tohoku-wwm-BERT の未知語として扱われたトークンのうち 52.7%が固有表現 (病名、症状) に関係するトークンであった．

### 6.2 サブワードに関する考察

tohoku-BERT，tohoku-wwm-BERT，UTH-BERT は語彙リストを基にサブワードに分割される．テストデータの固有表現のうちサブワードに分割されたトークンの割合は touhoku-BERT，tohoku-wwm-BERT で 25.2%，UTH-BERT で 16..5%であった．また，予測に失敗した固有表現のうちサブワードであった割合は，tohoku-BERT で 22.8%，tohoku-wwm-BERT で 22.3%，UTH-BERT で 12.8%であった．サブワードに分割されたトークンに対しても抽出できていることが分かる．

### 6.3 epoch 数に関する考察

実験では 15epoch まで学習し，最も Micro-F1 が高いものを採用している．Fig.2 に各 epoch に対する Micro-F1 の推移を示す．tohoku-char-BERT は最大 epoch 付近で最も高い値を獲得しており，学習を進めることでより高い値を獲得する可能性がある．一方で UTH-BERT は，交差毎に最高値を出す epoch 数に違いがあり、値の推移から最適な epoch 数を推測することが難しい．そのため，最適な epoch 数の決め方を検討する必要がある．

### 6.4 データセットにする考察

作成したデータセットは小規模であるため，固有表現とモダリティ表現が十分に網羅されていない．そのため、大規模なデータセットを用いた実験が必要である．また，東北大学の BERT と比べて UTH-BERT は，少ない epoch 数で高い精度を獲得していることから，医療分野のより複雑な自然言語処理のタスクで力を発揮すると考えられる．

## 7 おわりに

本研究では，固有表現 (病名，症状) と 5 種類のモダリティラベルをアノテーションしたデータセットを作成し BERT+CRF を用いて NER とモダリティ推定を

Table 6: Comparison of modality labels by BERT models

| Model | Positive | | | Negative | | | Suspicion | | |
|---|---|---|---|---|---|---|---|---|---|
| | 適合率 | 再現率 | F1 | 適合率 | 再現率 | F1 | 適合率 | 再現率 | F1 |
| tohoku-char-BERT | 0.927 | 0.953 | **0.939** | 0.930 | 0.946 | 0.938 | **1.000** | 0.897 | **0.931** |
| tohoku-BERT | 0.909 | 0.957 | 0.932 | 0.937 | 0.934 | 0.934 | 0.986 | **0.900** | 0.926 |
| tohoku-wwm-BERT | 0.907 | **0.958** | 0.932 | 0.941 | 0.943 | 0.940 | 0.986 | 0.883 | 0.910 |
| UTH-BER | **0.932** | 0.944 | 0.938 | **0.952** | **0.957** | **0.954** | **1.000** | 0.893 | 0.927 |

| Family | | | Family+Negative | | |
|---|---|---|---|---|---|
| 適合率 | 再現率 | F1 | 適合率 | 再現率 | F1 |
| 0.990 | 0.848 | 0.904 | 0.991 | **1.000** | 0.995 |
| 0.974 | 0.697 | 0.753 | 0.985 | 0.947 | 0.961 |
| 0.972 | 0.768 | 0.827 | **1.000** | **1.000** | **1.000** |
| **0.993** | **0.943** | **0.965** | 1.000 | 1.000 | 1.000 |

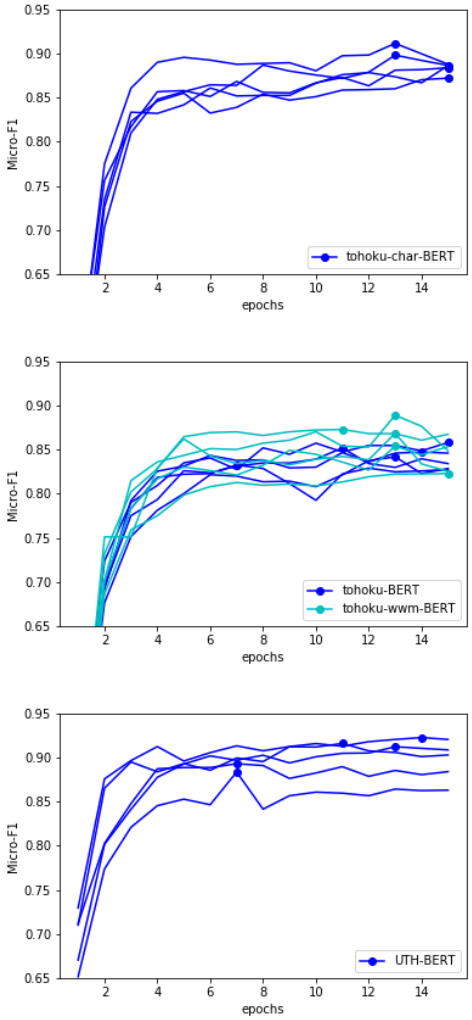


Fig. 2: Micro-F1 change by number of epochs

行った．また，事前学習の異なる 4 種類の BERT モデルを用いて精度の比較を行った．実験結果から医療文書を用いて事前学習した BERT が最も精度が高く，NER においては単語単位ではなく，文字単位の分割が有効であることを示した．

今後は，より大規模なデータセットを用いた Fine-tuning による抽出性能の分析と，モダリティラベルの設計を検討する必要がある．

## 参考文献


1) 荒巻英二：医療言語処理（自然言語処理シリーズ），第 12 巻，コロナ社，(2017)
2) Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, North American Chapter Of The Association For Computational Linguistics, (2019)
3) Ken Yano : Neural disease named entity extraction with characterbased bilstm+ crf in japanese medical text, CoRR, (2018)
4) 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，言語処理学会第 26 回年次大会，(2020)
5) Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma : Overview of the NTCIR-11 MedNLP-2 Task, Citeseer, (2014)
6) Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma : Overview of the NTCIR-12 MedNLPDoc Task, In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, (2016)
7) Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N. Gomez, LukaszKaiser, Illia Polosukhin : Attention Is All You Need,Neural Information Processing Systems, (2017)
8) https://github.com/cl-tohoku/bert-japanese
9) https://ai-health.m.u-tokyo.ac.jp/uth-bert
10) http://taku910.github.io/mecab/
11) http://sociocom.jp/ data/2018-manbyo/index.html
12) https://github.com/huggingface/transformers
13) Erik F. Tjong, Kim Sang, and Fien De Meulder : Introduction to the CoNLL-2003 shared task: Language-independent name entity recognition, In Proceedings of the Seventh Conference on Natural Language Learning, pp. 142-147, (2003)