一般口演 | 知識工学
一般口演9
自然言語処理・テキストマイニング
2021年11月20日(土) 09:10 〜 11:10 E会場 (2号館2階222+223)

[3-E-1-01] 詳細なアノテーション基準に基づく症例報告コーパスからの固有表現及び関係の抽出精度

*柴田 大作[1]、河添 悦昌[1]、篠原 恵美子[1]、嶋本 公徳[1] （1. 東京大学大学院 医学系研究科 医療AI開発学講座）
*Daisaku Shibata[1], Yoshimasa Kawazoe[1], Emiko Shinohara[1], Kiminori Shimamoto[1] （1. 東京大学大学院 医学系研究科 医療AI開発学講座）

キーワード：Information Extraction, Natural Language Processing, Machine Learning

【背景】診療において重要な情報である患者の症状や所見などはフリーテキストとしてのみ記録されるため、これら情報の利用を容易にするため構造化することが期待される。著者らはこれまで、診療テキストに対する網羅的な所見アノテーション基準を開発し、50種の固有表現と35種の関係を人手によりアノテートした症例報告コーパスを公開してきた。このアノテーションを精度良く再現することができれば、診療テキストをソースとする詳細な構造化データが得られる可能性がある。一方、このコーパスに出現する固有表現と関係の種類数は、先行研究で使用される他コーパスと比較して多いため、どの程度の精度でアノテーションを再現できるかが不明である。【目的】症例報告コーパスからの情報抽出を、固有表現抽出と関係抽出の2つのタスクとして定義し、機械学習による情報抽出の精度を評価する。【方法】前述の症例報告コーパスを利用した。1症例の文字数の平均は1,917、固有表現の平均は361個、関係の平均は347個であった。1症例は平均して12行で構成され、1行1文とした。1文あたりの文字数の平均は330であった。機械学習モデルは診療録テキストで事前学習済みのTransformer（ BERT-base）をベースとし、固有表現抽出と関係抽出を同時に行う Jointモデルを採用した。全2,194文のうち、 BERTの最大入力長である512単語を超えた20文は除外した。【結果】5分割交差検証によるマイクロ F1値の平均を評価指標とした。固有表現抽出の精度は87.2、関係抽出の精度は60.0であった。【考察】固有表現抽出は比較的高い精度を示したが、関係抽出は固有表現抽出と同程度の精度に至らなかった。固有表現抽出は多くの先行研究が報告されているが、関係抽出についてはその数が少ない。今後、精度の向上に向けた技術開発が必要と考える。

詳細なアノテーション基準に基づく症例報告コーパスからの固有表現及び関係の抽出精度

柴田大作[*1]、河添悦昌[*1]、篠原恵美子[*1]、嶋本公徳[*1]
*1 東京大学大学院 医学系研究科 医療 AI 開発学講座

The Accuracy of Entity and Relation Extraction from Case Reports

Daisaku Shibata[*1], Yoshimasa Kawazoe[*1], Emiko Shinohara[*1], Kiminori Shimamoto[*1]
*1 Graduate School of Medicine and Faculty of Medicine

[Background] Significant information related to medical data of the patients is often written in a free-text form in clinical records. It is expected that these records will be automatically structured and utilized for research. The already released case report corpus has manually annotated 70 type of entities and 35 type of relations. If a system that extracts information (entity and relation) is built from the corpus, structured data could be obtained from these clinical texts. [Objective] The information extracted from the corpus was divided into entity recognition (NER) and relation extraction (RE) groups, and their performances were evaluated. [Method] This study utilized the aforementioned corpus containing 183 cases. Furthermore, the case report consisted of an average of 1,915 characters, 394 entities, 387 relations, and 12 sentences. The report was split by a newline character, and a line was taken as a sentence. Finally, 2,194 sentences were obtained. A joint entity and relation extraction model based on Bidirectional Encoder Representations from Transformers was used. [Result] The evaluation results revealed that the micro-averaged F1 scores of NER and RE were 0.932 and 0.764, respectively. [Discussion] NER showed a relatively high performance; however, RE did not show it on the same level. Several studies reported the results of NER from clinical texts; however, very few studies reported those of RE. Therefore, study related to the RE performance is necessary.

Keywords: Information Extraction, Natural Language Processing, Machine Learning

1. 緒論

診療において重要な情報である患者の所見や症状などは診療テキストに自由記載されることが多い。そのため、診療テキストから興味のある情報を構造化して抽出する技術の開発が期待される。本稿では、テキストに出現する興味ある表現(所見や人体部位など)を抽出することを固有表現抽出、固有表現間の関係を分類することを関係抽出とし、これらをまとめて情報抽出とする。

診療テキストから情報抽出を行う研究は、英語、日本語を問わずこれまでにも多く報告されている。Patel [1]らは経過記録や退院サマリなどを含む診療テキストに出現する 11 種類の固有表現 (例: 所見や人体部位)に対して人手によりアノテーションを実施し、Conditional Random Fields (CRF)を用いた固有表現抽出を行ったところ、91.6 の F 値で抽出できることを報告した。荒牧ら [2]は、症例報告に出現する病名の事実性を考慮した固有表現抽出 (実際に患者が罹患している病名であれば陽性病名、そうでなければ陰性病名)を CRF により実施した。その結果、陽性病名は F 値が 83.3 の精度で抽出できる一方、陰性病名はその頻度の低さ、出現する文脈の違いなどから陽性病名と比較して F 値が 55.4 と低い精度となることを明らかにした。Fraser ら [3]は退院サマリに出現する病名や症状、医薬品や治療表現、検査やバイタルサインに対してアノテーションが付与された i2b2 2010 データセット [4]と生物医学論文の概要に出現する Unified Medical Language System の concept や semantic type に対してアノテーションが付与された MedMentions データセット [5]を用い、Bidirectional Encoder Representations from Transformers (BERT) [6]による固有表現抽出を行った。その結果、BERT を用いることで従来手法よりも高い精度を得られることを明らかにし (F 値で最大 1.00 ポイントの差)、診療テキストからの固有表現抽出に対する BERT の有効性を示した。Yada ら [7]は、日本語の診療記録と読影レポートに出現する 5 種類の医療表現 (病名・症状、人体部位、特徴・測定値、変化、時間)に対してアノテーションを実施し、BERT による固有表現抽出を行ったところ、 日本語の診療テキストにおいても BERT を用いることで従来手法よりも高い精度を得られることを明らかにした (F 値で 1.04 ポイントの差)。

このように診療テキストからの固有表現抽出に関する研究は数多く報告されているが、関係抽出まで踏み込んだ研究は少ない。情報抽出の結果を真に有益なものとするためには、固有表現の抽出だけでは不十分であり、固有表現間の関係まで分類する必要があると考えられる。そのため本研究では、固有表現抽出だけではなく、関係抽出まで考慮した診療テキストからの情報抽出を行う。

2. 目的

本研究では、固有表現抽出と関係抽出を同時に行う BERT をベースとした Joint モデルを構築し、それを用いた症例報告からの固有表現及び関係の抽出精度を評価する。また、誤り分析を通して、異なるドメインのテキストで事前学習された 2 つの BERT (診療テキストで事前学習した UTH-BERT [8]と日本語 Wikipedia で事前学習した NICT-BERT[1])の抽出精度の違いを考察する。

3. 方法

3.1 実験材料

実験材料として、篠原らによって開発された症例報告コーパス [9]を使用する。このコーパスは、厚生労働省の指定難病名と「例」という文字列をタイトルに含み、2000 年以降に出版された症例 183 件のテキストから構成されている (1 つの症例報告で複数の症例について報告されている場合は、症例ごとに分割されている)。また、70 種類の固有表現と 35 種類の関

[1] https://alaginrc.nict.go.jp/nict-bert/index.html

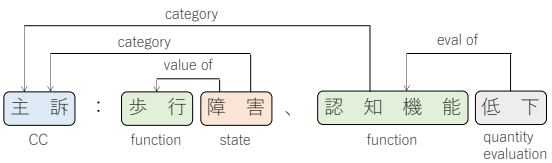
係についてアノテーションが文字単位で付与されており、各文書は改行で文単位に分割されている。コーパスの統計情報を表 1 に示す。以下、固有表現タグと関係ラベルはそれぞれ *ent:タグ名称*、*rel:ラベル名称*のように接頭辞+イタリックで表記する。

3.2 コーパスの前処理

本コーパスは文字単位でアノテーションが付与されている(図 1 の(a))が、実験にあたり単語単位に変換する必要がある。例えばある文を UTH-BERT へ入力する場合、形態素解析器 MeCab [10]を用い、辞書に万病辞書[2]と Neologd [11]を指定して形態素解析を行う必要がある (図 1 の(b))。しかしこの場合、図 1 の(a)と(b)からわかるようにアノテーションでタグが付与された範囲と形態素解析で得られた形態素の範囲にギャップが生じる。具体的には、「歩行」という文字列には固有表現タグとして *ent:function* が、「障害」という文字列には固有表現タグとして *ent: finding* がアノテーションされている。しかし、形態素解析を実施すると辞書に登録される「歩行障害」が一つの形態素となり、アノテーション結果と紐づけることが困難になる。これは日本語などの単語境界が明確でない言語において起こる問題であり、直ちに解決することは難しい。そのため本研究では、事前にアノテーション情報を利用して文を固有表現単位に分割し、形態素解析を実施する。例えば、「主訴:歩行障害、認知機能低下」という文であれば、「主訴/:/歩行/障害/、/認知機能/低下」のように分割された文字列に対して形態素解析を実施する (図 1 (c))。この前処理はアノテーションが付与されない未知のテキストに対する処理とは乖離していることに留意が必要である。

表 1　症例報告コーパスの統計情報

文章数		183
固有表現タグの種類		70
関係の種類		35
1文書毎	文字数 (S.D)	1915 (696)
	単語数 (S.D)	972 (330)
	固有表現数 (S.D)	394 (129)
	関係数 (S.D)	387 (127)
	文数 (S.D)	12.0 (4.5)



(a) アノテーション例

主訴/：/歩行障害/、/認知/機能低下

(b) 形態素解析結果

主訴/：/歩行/障害/、/認知機能/低下

(c) 恣意的な形態素解析結果

図 1 コーパスの前処理の例: スラッシュは単語境界を示す。

[2] http://sociocom.jp/~data/2018-manbyo/index.html

3.3 実験

3.3.1 実験データのフォーマット

本研究では、固有表現抽出を文の各単語に最適なラベルを付与するタスクとし、Inside-Outside-Beginning (IOB2)形式を採用した。IOB2 形式は、固有表現の開始位置である単語に B(Beginning)、開始位置ではない単語に I(Inside)、固有表現ではない単語に O(Outside)タグを付与する。これに加え、B と I タグの接尾辞として固有表現タグを付与する。

関係抽出では、ある単語 (head)に対して、関係の矢印が流入する先の単語 (tail)であるか否かと関係の種類を同時に予測するタスクとした。例えば図 1 の(c)では、「認知機能」という単語が head、「主訴」という単語が tail で、関係は *rel: category* となる。フォーマットの例を表 2 に示す。

表 2 ラベリングの例: tail の数値は番号に対応する。

番号	単語	IOB2	Tail	関係の種類
0	主訴	B-CC	-	
1	:	O	-	
2	歩行	B-function	-	
3	障害	B-state	0, 2	catefory, value_of
4	、	O	-	
5	認知機能	B-function	0	catefory
6	低下	B-quantity_evaluration	5	eval_of

3.3.2 機械学習モデル

機械学習モデルとして、固有表現抽出と関係抽出を同時に行う Joint モデルを採用した。モデルは単語埋め込み層、線形層、CRF 層、タグ埋め込み層と関係分類層から構成される。詳細を以下に、モデルの概要を図 2 に示す。

1. 単語埋め込み層
   入力文の単語系列を$X = (x_1, x_2, \dots, x_i, \dots, x_n)$とし、$x_i$ は i 番目の単語を示す。また、対応するタグ系列を$y = (y_1, y_2, \dots, y_i, \dots, y_n)$とし、$y_i$ は i 番目のタグを示す。単語系列Xを BERT へ入力することで、単語の埋め込み表現$e = (e_1, e_2, \dots, e_i, \dots, e_n)$を得る。なお、$e_i$は 768 次元である。
   $$e = BERT(X)$$
2. 線形層
   単語埋め込み層で取得した埋め込み表現e_iの線形変換を行う。$W \in \mathbb{R}^{C \times 768}$、$b \in \mathbb{R}^C$であり、C は固有表現タグの総数である。
   $$P(y|x_i) = We_i + b$$
3. CRF 層
   CRF はタグ間の依存関係をモデル化し、タグ系列全体の遷移確率を考慮することが可能となり、入力系列に対して尤もらしいタグ系列を得ることができる。学習時は以下の$\log P(y|X)$を最小化する。
   $$\log P(y|X) = -s(X, y) + \log add(s(X, \tilde{y}))$$
   $$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$$
   $\tilde{y}$は全ての考えられるタグ系列、Aは遷移スコアの行列であり、$A_{y_i, y_{i+1}}$は、タグy_iからy_{i+1}への遷移スコアを示す。そして、Pは線形変換した単語の埋め込み表現であ

り、$P_{i,j}$は i 番目の単語の j 番目のタグの確率を示す。また、予測において単語系列 X に対する最適なタグ系列は以下で算出する[12]。

$$y^*_{ner} = argmax(s(X, \tilde{y}))$$

4. タグ埋め込み層
CRF 層で得られたタグ系列の埋め込み表現$le \in \mathbb{R}^{n\times 30}$を取得する。なお、学習時は正解のタグ系列を、予測時は CRF 層の予測結果をタグ埋め込み層へ入力する。得られたタグの埋め込み表現は、単語の埋め込み表現eと結合する。

$$le = label\ embedding(tag)$$
$$h = Concat(e, le)$$

5. 関係分類層
単語間の関係を分類する関係分類層は Zhang ら[13]の head-selection モデルを参考にした。単語x_iから単語x_jの関係$z(h_i, h_j)$を以下で算出し、学習時は交差エントロピー損失を最小化する。

$$z(h_i, h_j) = V^T \cdot \tanh(U \cdot h_i + W \cdot h_j)$$

ここで、$V \in \mathbb{R}^{798\times K}$、$U \in \mathbb{R}^{798\times 798}$、$V \in \mathbb{R}^{798\times 798}$である。また予測時は、確率がもっとも高いクラスを単語 i から単語 j への関係とする。

$$y^*_{RE} = argmax(r(h_i, h_j))$$

3.3.3 実験設定

BERT の入力は固定長に制限されるため、入力文の最大単語数を510単語とし、これを超える単語数から構成される文は実験データから除外した。UTH-BERT と NICT-BERT はそれぞれ前処理の方法 (形態素解析に使用する辞書)が異なるため、前者では 510 単語以内に収まっていたデータでも、後者において 510 単語を超える場合がある (この逆も存在する)。そのため、UTH-BERT と NICT-BERT で学習データの数が異なる。最終的には UTH-BERT のデータが 182 症例、2,172 文となり、NICT-BERT のデータが 182 症例、2,170 文となった。

実験では最適化関数に Adam を使用し、CRF 層の出力結果に基づいて算出される損失と関係分類層の出力結果に基づいて算出される損失の合計値を最小化するようにモデルを学習した。また、epoch は 120、バッチサイズ は 16、学習率の初期値は 1e-3 とし BERT 部分のみ 3e-5 とした。

3.3.4 評価方法

5 分割交差検証による Macro-F1、Micro-F1 の平均値を評価指標とした。データ分割の単位は症例単位とし、1 症例中に含まれる全ての文は訓練、検証もしくはテストデータのいずれかのみに存在するようにした。訓練データの 20%を検証データとし、検証データにおける Micro- F1 (固有表現抽出の Micro-F1 と関係抽出の Micro-F1 の平均値)がもっとも大きかったエポックのモデルを採用しテストデータを評価した。

固有表現抽出では、モデルが固有表現とした表現のうち実際に固有表現である表現の割合を Precision、実際に固有表現である表現のうち、モデルが固有表現と予測した表現の割合を Recall とし、Precision と Recall の調和平均によりそれぞれのタグの F 値を算出した。

関係抽出では、$(head, tail, rel)$を単語ごとに正解値と予測値から作成し (*head* と *tail* に関係がない場合は「関係なし」を付与)、正解データと予測結果の三つ組が完全に一致した場合は正解、それ以外は不正解としてそれぞれのラベルの F 値を算出した。例えば、(*歩行*, *障害*, *category*)が正解値である場合に、(*歩行*, *障害*, *eval_of*)と予測された場合は不正解となる。また評価において、正解ラベルが「関係なし」であるものは除外した。 これは、ほとんどの固有表現の間は「関係なし」で正解付けられ、結果が限りなく 1 に近い値となるためである。

3.3.5 誤り分析

コーパスにおいて出現頻度の高い固有表現タグと関係ラベル、また患者の状態を表現するために有用な固有表現タグと関係ラベルを選択し、これらについて誤り分析を行う。対象とした固有表現タグと関係ラベルの詳細を表 3 に示す。

4. 結果

4.1. 各 BERT の固有表現抽出と関係抽出の精度

以下、固有表現抽出(Named Entity Recognition)を NER、関係抽出(Relation Extraction)を RE と表記する。

NER の実験結果を表 4、誤り分析の対象とした固有表現タグの F 値を表 5、RE の実験結果を表 6、誤り分析の対象とした関係ラベルの F 値を表 7 にそれぞれ示す。NER の Micro-F1 は UTH-BERT が有意に高い値を示したが、Macro-F1 は有意差を認めなかった(表4)。また、RE の Macro-F1 と Micro-F1 共に 2 種の BERT 間で有意差を認めなかった(表 6)。

表 4 NER の実験結果

BERT	平均	Precision (S.D)	Recall (S.D)	F1 (S.D)	95%CI (F1)
UTH	Micro	0.928 (0.007)	0.936 (0.004)	0.932 (0.006)	0.923-0.940
	Macro	0.792 (0.027)	0.787 (0.027)	0.782 (0.029)	0.736-0.817
NICT	Micro	0.911 (0.004)	0.915 (0.003)	0.913 (0.003)	0.909-0.919
	Macro	0.789 (0.021)	0.768 (0.014)	0.772 (0.015)	0.749-0.794

表 5 誤り分析の対象とした固有表現タグの F 値

固有表現タグ	UTH		NICT		差
	F値 (S.D)	データ数	F値 (S.D)	データ数	
state	**0.922** (0.012)	2615	0.889 (0.002)	2598	0.033
body	**0.938** (0.004)	1373	0.904 (0.017)	1366	0.034
item	**0.922** (0.018)	1005	0.907 (0.004)	1003	0.015
clinical_test	**0.923** (0.015)	589	0.910 (0.013)	587	0.013
PN_Positive	**0.960** (0.004)	1015	0.953 (0.005)	1010	0.007
PN_Negative	**0.967** (0.007)	351	0.956 (0.005)	348	0.011
time	**0.973** (0.010)	692	0.959 (0.009)	686	0.014
value	**0.961** (0.007)	883	0.959 (0.005)	883	0.002
unit	**0.968** (0.010)	737	0.960 (0.006)	736	0.008

表 6 RE の実験結果

BERT	平均	Precision (S.D)	Recall (S.D)	F1 (S.D)	95%CI (F1)
UTH	Micro	0.763 (0.019)	0.766 (0.011)	0.764 (0.013)	0.744-0.780
	Macro	0.617 (0.046)	0.549 (0.030)	0.567 (0.035)	0.511-0.616
NICT	Micro	0.738 (0.010)	0.747 (0.006)	0.743 (0.005)	0.735-0.750
	Macro	0.578 (0.026)	0.524 (0.019)	0.537 (0.020)	0.507-0.557

表 7 誤り分析の対象とした関係ラベルの F 値

関係ラベル	UTH		NICT		差
	F値 (S.D)	データ数	F値 (S.D)	データ数	
value_of	**0.820** (0.012)	3924	0.802 (0.009)	3901	0.018
site	**0.727** (0.018)	1292	0.689 (0.016)	1285	0.038
unit	0.944 (0.012)	792	**0.947** (0.007)	792	-0.003
method	**0.724** (0.028)	887	0.706 (0.027)	885	0.018

4.2 予測誤りの集計

以下に、NER で UTH-BERT と NICT-BERT の F 値の差が大きかった *ent: state*、*ent: body*、*ent: value*、*ent: unit* における予測誤りの集計結果と RE で F 値の差が大きかった *rel: value_of* と *rel: site* における予測誤りの集計結果を示す。紙面の都合により他のタグの分析結果は割愛する。

4.2.1 ent:state の誤り例

ent: state において、UTH-BERT でのみ予測を誤った表現は 344 個、NICT-BERT でのみ予測を誤った表現は 679 個、両方の BERT で予測を誤った表現は 571 個あった。

UTH-BERT では、「嚢胞」、「周堤」や「萌出」などの表現が未知語となり、予測するタグを誤る事例がもっとも多く 9 件あった。NICT-BERT でも同様に、「黄疸」、「腫瘤」や「狭窄」などの表現が未知語となり、予測するタグを誤る事例がもっとも多く 73 件あった。また両方の BERT では「月経」に対して付与するタグを誤る事例 (年齢を意味する *ent: age* や時間を意味する *ent: time* と予測)がもっとも多く、それぞれ 19 件あった。

4.2.2 ent:body タグの誤り例

ent: body において、UTH-BERT でのみ予測を誤った表現は 132 個、NICT-BERT でのみ予測を誤った表現は 322 個、両方の BERT で予測を誤った表現は 193 個あった。

UTH-BERT では、「胃/体」の「胃」と「体」をそれぞれ *ent: body* と予測する事例がもっとも多く 6 件あった。また NICT-BERT では、「脾」、「僧帽弁」や「脾臓」などの表現が未知語となり、予測するタグを誤る事例がもっとも多く 50 件あった。両 BERT では「軟骨」に対して付与するタグを誤る事例がもっとも多く 4 件あり、例えば「喉頭/軟骨/炎」は「喉頭」と「軟骨」がそれぞれ *ent: body* であるが、これをまとめて *ent: body* とする事例や、人体組織を意味する *ent: tissue* と予測する事例などがあった。

4.2.3 ent:value タグの誤り例

ent: value において、UTH-BERT でのみ予測を誤った表現は 48 個、NICT-BERT でのみ予測を誤った表現は 70 個、両方の BERT で予測を誤った表現は 89 個あった。

UTH-BERT では、「2」に付与するタグを誤った事例がもっとも多く 7 件あった。具体的には「2/本/の/鉗子」は「2」が *ent: value*、「本」が *ent: unit*、「鉗子」が *ent: device* であるが、これらをまとめて *ent: device* と予測した事例、「2 回」に対して *ent: detail* を付与した事例などがあった。NICT-BERT でも同様に「2」に付与するタグを誤った事例がもっとも多く 5 件あった。具体的には、「2/コース/後」の「2」は *ent: value*、「コース」は *ent: unit*、「後」は *ent: time* であるが、これらをまとめて *ent: time* と予測した事例、「ラクトミン/2/Cap」の「ラクトミン」は *ent: drug*、「2」は *ent: value*、「Cap」は *ent: unit* であるが、これらをまとめて *ent: drug* と予測した事例などがあった。両 BERT においては「5」に付与するタグを誤った事例がもっとも多くそれぞれ 17 件あった。具体的には、「長指屈筋/5/／/5」は「5」が *ent: value* であるが、最初の「5」を *ent: value*、後ろの「／5」を *ent: unit* と予測した事例、「週/5/日」は「週」が *ent: time_span*、「5」が *ent: value*、「日」が *ent: unit* であるがこれらをまとめて *ent: time_span* と予測した事例があった。

4.2.4 ent:unit タグの誤り例

ent: unit において、UTH-BERT でのみ予測を誤った表現は 41 個、NICT-BERT でのみ予測を誤った表現は 68 個、両方の BERT で予測を誤った表現は 37 個あった。

UTH-BERT では、「1U/L」に付与するタグを誤った事例がもっとも多く 4 回あった。例えば、「GPT/9/1/2/1/U/／/L」の「912」は *ent: value*、「1U／L」は *ent: unit* であるが、「9121」を *ent: value*、「U/L」を *ent: unit* と予測した事例があった (本来は「IU/L」であるが、コーパスのテキストの不備である。)。

NICT-BERT では、「μg／day」の予測を誤る事例がもっとも多く 7 件あった。これは「μ/g/／/day」の「μ」が未知語となることで、「μ」を *ent: outside*、「g/／/day」を *ent: unit* と予測してしまうものであった。

両 BERT では、「回/目」に付与するタグを誤った事例がもっとも多くそれぞれ 7 件あった。例えば、「2 回目」の「2」は *ent: value*、「回目」は *ent: unit* であるが、これをまとめて *ent: detail* や *ent: time* と予測する事例があった。

4.2.5 rel:value_of ラベルの誤り例

UTH-BERT において、本来は *rel: value_of* が付与される箇所に *rel: None* (関係なし)を付与した数は 2,868 件あり、その内訳は 1) 固有表現タグの予測誤りに起因する可能性のある誤りが 1,419 件、2) 固有表現タグは正しく予測できているが、関係ラベルの *rel: None* と予測したものが 1,449 件あった。また、*rel: value_of* と *rel: None* 以外の関係ラベルを付与した数は 781 件あり、その内訳は 3) 固有表現タグは正しく予測できたが、関係ラベルの分類を誤ったものが 599 件、4) 固有表現タグの予測を誤りかつ関係ラベルの分類を誤ったものが 182 件あった。NICT-BERT では、1)が 1,857 件、2)が 1433 件、3)が 560 件、4)が 188 件あった。

4.2.6 rel:site ラベルの誤り例

UTH-BERT において、本来は *rel: site* が付与される箇所に *rel: None* を付与した数は 1,094 件あり、その内訳は 1) 固有表現タグの予測誤りに起因する可能性のある誤りが 383 件、2) 固有表現タグは正しく予測できているが、関係ラベルの *rel: None* と予測したものが 711 件あった。

また、*rel: site* と *rel: None* 以外の関係ラベルを付与した数は 532 件あり、その内訳は 3) 固有表現タグは正しく予測できたが、関係ラベルの分類を誤ったものが 468 件、4) 固有表現タグの予測を誤りかつ関係ラベルの分類を誤ったものが 64 件あった。NICT-BERT では、1)が 638 件、2)が 693 件、3)が 464 件、4)が 70 件あった。

5. 考察

5.1 NER に関する考察

表 5 において、誤り分析の対象としたすべての固有表現タグについて、UTH-BERT が NICT-BERT よりも高い F 値を示した。その中でも、患者の状態を表す、*ent: state* と人体部位を表す *ent: body* は F 値の差が大きかった。患者の状態に関する表現は診療テキストでは多く出現するが、Wikipedia では出現頻度が低いと思われる。そのため、UTH-BERT ではより適当な埋め込み表現を事前に学習することができ、これが NICT-BERT と UTH-BERT との抽出精度の違いにつながったと考えられる。

一方、*ent: value*、*ent: unit*、*ent: PN_Positive* は数値、単位、肯定を意味する表現に付与される。これらはどのような種類のテキストでも使用される一般的な表現であり、診療テキストと Wikipedia の両方で頻出する表現であることから、それぞれの BERT で同程度の質の埋め込み表現が学習される。そのため、UTH-BERT と NICT-BERT 間の F 値の差が小さかったと考えられる。

また、*ent: item*、*clinical_test*、*PN_Negative*、*time* はそれぞれ所見項目、臨床検査、否定、時間を意味する表現に付与される。これらの表現は *ent: value* や *unit* などと同様に、どちらのテキストでも使用される表現であるが、UTH-BERT の方が高い F 値を示した。この原因については現状で考察できておらず、今後、データ数を増加させるなどを実施した上での更なる検討が必要である。

5.2 RE の誤り分析

表 7 において、*rel: value_of* は *ent: value* と *ent: item* や *ent: state* と *ent: PN_Positive*/*PN_Negative* などの固有表現間の関係として付与される。例えば、「身長 170cm」であれば、「170 (*ent: value*)」から「身長 (*ent: item*)」へ *rel: value_of* が付与され、「疼痛を認めない」であれば、「認めない (*ent: PN_Negative*)」から「疼痛 (*ent: state*)」へ付与される。また、*rel: site* は *ent: state* と *ent: body* や *ent: state* と *ent: item* などの固有表現間の関係として付与される。例えば、「手首に疼痛」であれば、「疼痛 (*ent: state*)」から「手首 (*ent: body*)」へ *rel: site* が付与され、「発作の頻度上昇」であれば「頻度 (*ent: item*)」から「発作 (*ent: state*)」へ付与される。そして、*rel: method* は *ent: PN_Positive* もしくは *PN_Negative* と *ent: state* などの固有表現間の関係として付与される。例えば、「異常を認めず」であれば「認めず (*ent: PN_Negative)*」から「異常 (*ent: state*)」へ付与される。さらに、*rel: unit* は *ent: value* と *ent: unit* である固有表現間の関係として付与される。例えば、「身長 170cm」であれば「cm (*ent: unit*)」から「170 (*ent: value*)」へ付与される。

上記 4 つの関係ラベルの F 値を比較すると、*rel: value_of*、*site*、*method* では UTH-BERT の精度が高く、*rel: unit* では NICT-BERT の精度が高いことが確認された。RE では、固有表現タグが重要な情報となる。これはあるタグに属する固有表現と別のあるタグに属する固有表現間に付与される関係ラベルは概ねいくつかに定まっているためである。そのため、NER の精度が RE の精度に大きな影響を与えると考えられる。この観点に着目すると *rel: value_of*、*site*、*method* において関係ラベルが付与されやすい固有表現タグである *ent: state*、*body*、*item* や *PN_Positive*/*PN_Negative* は UTH-BERT の方が NER の精度が高い。一方で *rel: unit* において関係ラベルが付与されやすい固有表現タグである *ent: value* や *unit* の精度は UTH-BERT の方が僅かに高いものの、両 BERT 間で大きな差はない。そのため、NER の精度が RE の精度に一定の影響を与えており、RE の精度を向上させるためには NER の精度の更なる向上が必要であると考えられる。

6. 結論

本研究では、2 種類の BERT をベースに NER と RE を同時に行う Joint モデルを採用し、症例報告コーパスを材料として情報抽出精度を評価した。いずれのモデルも、NER は高い精度を示したが、RE の精度は NER と比較して低く、実応用へ向けた更なる改善が必要であると考えられた。また、患者の状態を示す固有表現タグとそれに関連する関係ラベルの抽出精度は UTH-BERT の方が高いが、一般的なテキストにも出現する固有表現タグやそれに関連する関係ラベルでは UTH-BERT と NICT-BERT 間で精度に大きな差は確認されなかった。

また、今回得られた抽出精度を実診療で得られるテキストに外挿することはできない。そのため今後、症例報告コーパスで学習したモデルを用いて、退院サマリなどを対象とする情報抽出の精度評価を行う。

謝辞

本研究は MEXT 科研費 JP20H04279 の支援を受けた。

参考文献

1) Patel, P., Davey, D., Panchal, V., & Pathak, P. Annotation of a large clinical entity corpus. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2018 : 2033-2042.

2) 荒牧英治, 若宮翔子, 矢野憲, 永井宥之, 岡久太郎, & 伊藤薫. 病名アノテーションが付与された医療テキスト・コーパスの構築. 自然言語処理 2018 ; 25(1): 119-152.

3) Fraser, K. C., Nejadgholi, I., De Bruijn, B., Li, M., LaPlante, A., & Abidine, K. Z. E. (2019). Extracting umls concepts from medical text using general and domain-specific deep learning models. In Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019) 2019 : 157-167.

4) Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 2011 ; 18(5) : 552-556.

5) Murty, S., Verga, P., Vilnis, L., Radovanovic, I., & McCallum, A. Hierarchical losses and new resources for fine-grained entity typing and linking. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics ; Volume 1 : 97-109.

6) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 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 2018 : 15-18.

7) Yada, S., Joh, A., Tanaka, R., Cheng, F., Aramaki, E., & Kurohashi, S. Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases. In Proceedings of the 12th Language Resources and Evaluation Conference 2020 : 4565-4572.

8) Kawazoe, Y., Shibata, D., Shinohara, E., Aramaki, E., & Ohe, K. A clinical specific BERT developed with huge size of Japanese clinical narrative. medRxiv 2020.

9) 篠原 恵美子, 河添 悦昌, 柴田 大作, 嶋本 公徳, 関 倫久. 医療テキストに対する網羅的な所見アノテーションのためのアノテーション基準の構築. 第 25 回日本医療情報学春季学術大会, 2021.

10) Kudo, T., Yamamoto, K., & Matsumoto, Y. Applying conditional random fields to Japanese morphological analysis. In Proceedings of the 2004 conference on empirical methods in natural language processing 2004 : 230-237.

11) Sato, T, Hashimoto, T & Okumura M. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval. In Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing 2017.

12) Zhang, X., Cheng, J., & Lapata, M. Dependency parsing as head selection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 2017 : 665-676.

13) Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2016 : 260-270.

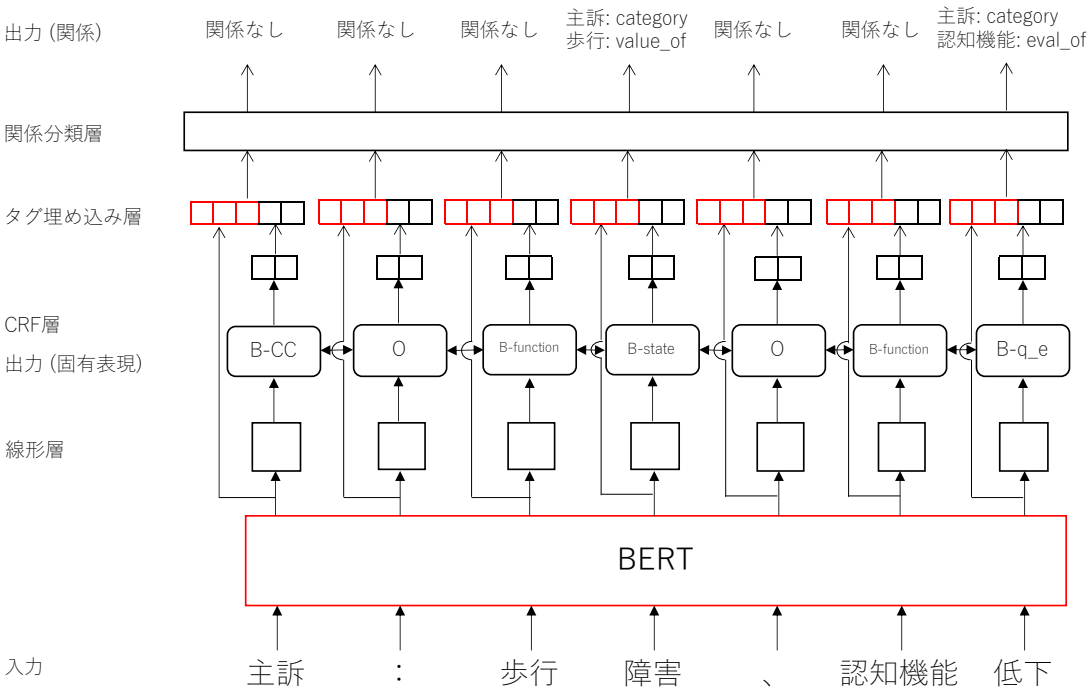


図 2 Joint モデルの概要

表 3 誤り分析の対象とした固有表現タグと関係ラベル

固有表現タグ	説明	例
state	患者の状態全般を示す表現。	吐き気、萎縮症、糖尿病
body	人体部位。特定の部位を示すもの。	肝、手足、眼瞼結膜
item	患者の状態を表すために参照される項目。	血糖値、HbA1c、食欲
clinica_test	臨床検査。itemとの違いは計測法を含むか否か。	神経学的検査、徒手筋力検査
PN_Positive	患者の状態があることを示す表現。	認め、認める、で、示す
PN_Negative	患者の状態がないことを示す表現。	認めず、ではなく、詳細不明
time	時間軸上における特定位置の時点や区間を示す表現	約10年前、その後、直後
value	検査値など身体や検体を測定し得られる数値。	7.5、5、165.0
unit	数値との組で表される単位。	mg/日、cm、行、kg/m^2
関係ラベル	関係ラベルの概略	例
value_of	sourceがtargetの値である。	**身長** (target)は**170** (source)cm
site	sourceがtargetの部位である。	**四肢** (target)の**筋力** (source)低下
unit	sourceはtargetの単位である。	身長は**170** (target)**cm** (source)であり
method	sourceがtargetの (方法に)より得られる。	**聴診** (target)上、**異常** (source)なし