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

第41回医療情報学連合大会
(第22回医療情報学会学術大会)
COI開示

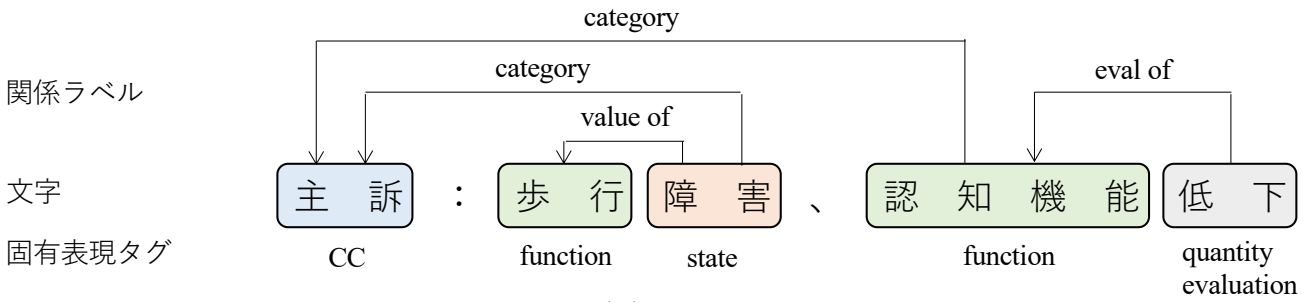
- 演題名:詳細なアノテーション基準に基づく症例報告コーパスからの固有表現及び関係の抽出精度
- 筆頭演者名: 柴田大作

私が発表する今回の演題について開示すべきCOIは以下のとおりです。

寄付講座: I&H株式会社（阪神調剤グループ），
株式会社EMシステムズ

背景：診療テキストからの情報抽出

- 診療テキストに自由記載される患者の症状や所見を自然言語処理技術で抽出することが期待
- 固有表現抽出に関する研究はこれまでも報告されているが，関係抽出まで踏み込んだ研究は少ない
  - 情報抽出の結果を有効に活用するためには **固有表現間の関係の情報**も重要
  - いつ，誰に，どのような疾患が，どこに生じたのかなどの情報も必要



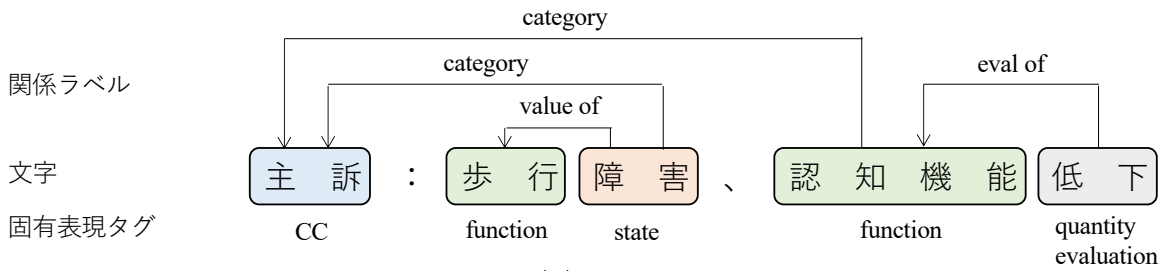
研究の目的

1. 固有表現抽出と関係抽出を同時に行うBERTをベースとするJointモデルの精度を評価
2. 異なるドメインのテキストで事前学習された2つのBERTにおける精度の違いを考察

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)

方法：症例報告コーパス

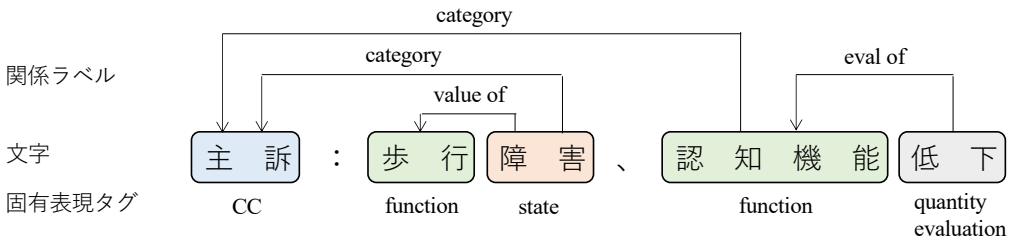
- 症例183件のテキストで構成 [1]
  - 厚生労働省の指定難病名と「例」という文字列をタイトルに含み，2000年以降に出版された症例報告の症例セクション
    - 1つの症例報告で複数の症例について報告されている場合は症例ごとに分割
  - 各テキストは改行で文単位に分割
  - 70種類の固有表現と35種類の関係が文字単位でアノテーション



[1] 篠原 et al. 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.

方法：コーパスの前処理

- 日本語BERTを使用する場合，事前にテキストの形態素解析が必要
  - 形態素解析器や辞書はBERTに依存
  - アノテーション情報と得られる形態素の間にギャップ
    - 「歩行」と「障害」は異なる固有表現としてアノテーションされているが，「歩行障害」で1つの形態素として分かち書き
  - 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てから形態素解析を実施
    - 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



アノテーション例

主訴/：/歩行障害/、/認知/機能低下

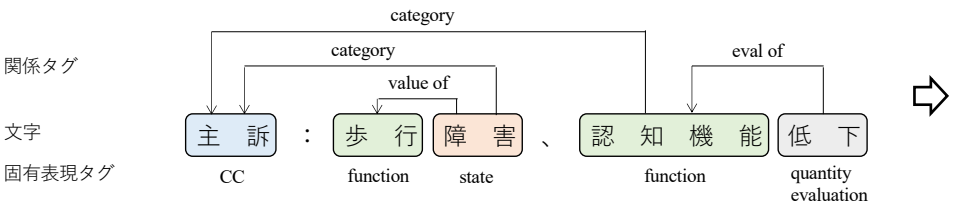
形態素解析結果

主訴/：/歩行/障害/、/認知機能/低下

恣意的な形態素解析結果

方法：タスク設定

- 症例報告からの情報抽出を2つのタスクで実施
  - 固有表現抽出: Named Entity Recognition（NER）
    - 文の各単語に最適なラベルを付与するタスク
    - Inside-Outside-Begining2 (IOB2)形式でラベリング
  - 関係抽出: Relation Extraction（RE）
    - head ある単語に対して、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予測するタスク
      - 「認知機能」がhead, 「主訴」がtail, 関係が*category*
      - 「低下」がhead, 「認知機能」がtail , 関係が*eval_of*



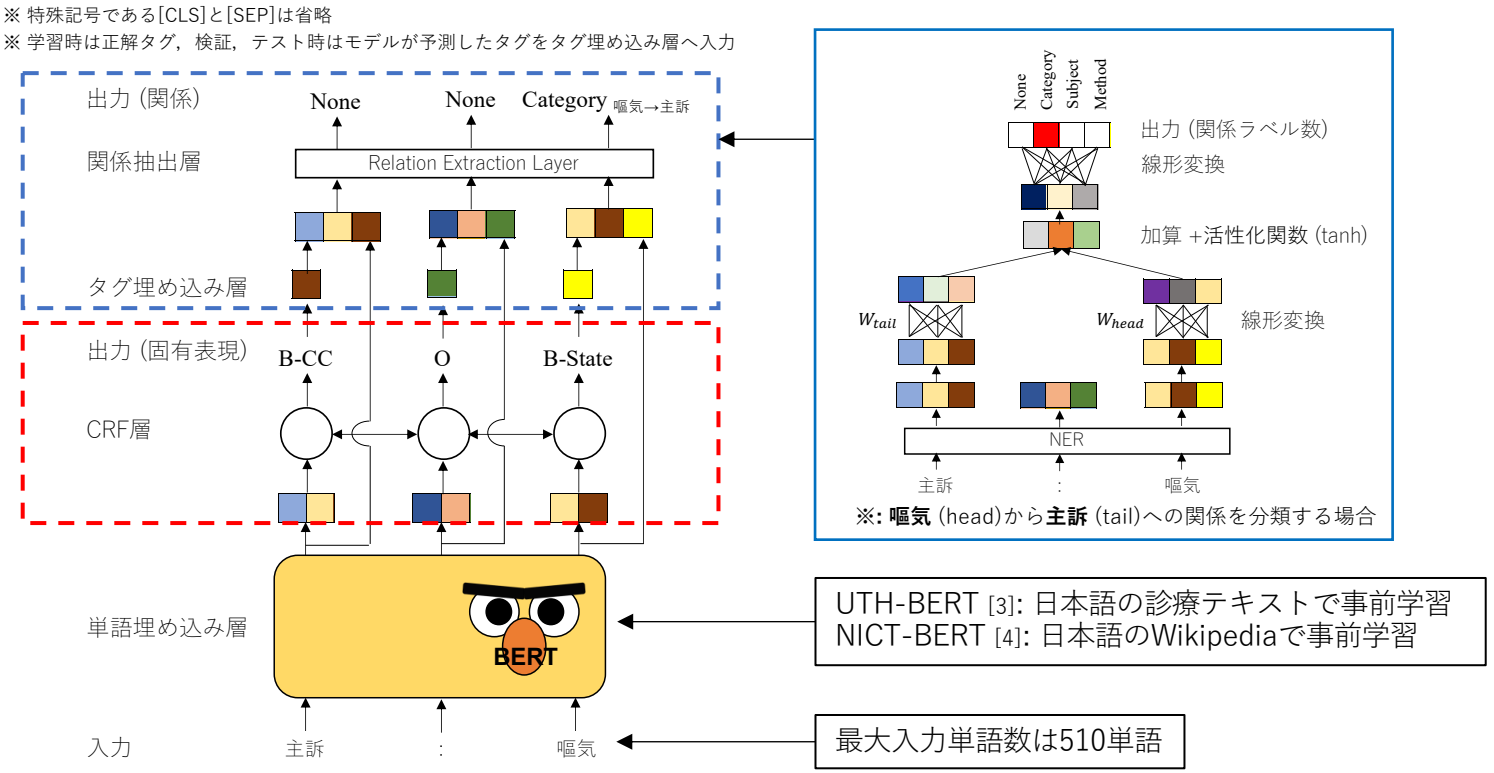
アノテーション

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

実験データの形式

方法：機械学習モデル

- NERにはBERT-CRFを使用し，REには Head Selection [2]を応用したネットワークを使用



[2] Zhang, Xingxing, Jianpeng Cheng, and Mirella Lapata. "Dependency parsing as head selection." arXiv preprint arXiv:1606.01280 (2016).
[3] Kawazoe Y, et al. A clinical specific BERT developed using a huge Japanese clinical text corpus. PLoS One. 2021 Nov 9;16(11):e0259763.
[4] https://alaginrc.nict.go.jp/nict-bert/index.html

方法：実験設定

- 学習データ（症例報告コーパス）の詳細
  - 学習データはUTH-BERTが182症例の2,172文,
    NICT-BERTが182症例の2,170文
    - BERTへの入力は最大で510単語であるため,
      これを超える単語数から構成される文は削除
    - BERTごとに前処理の方法が異なるため,
      どちらかのBERTでのみ510単語を超える文も存在
- 機械学習モデルのパラメーターの詳細
  - 最適化にはAdam，バッチサイズ は16，エポックは120,
    学習率の初期値は1e-5とし，BERTのみ3e-5に設定
  - NERの出力結果に基づいて算出される損失とRE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学習

方法：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て算出
  - 訓練/検証/テストの分割は症例単位で実施
    - 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
  - 訓練データの20%を検証データに使用
  - 検証データにおいてNERとREのMicro-F1の平均値が最も大きかったエポック時のモデルを用いてテストデータを評価

方法：固有表現抽出の評価方法

- CoNLL-2000の評価方法を使用 [5]
  - タグ毎のF1: PrecisionとRecallの調和平均で算出
    - Precision: モデルが固有表現と予測した表現の中で，実際に固有表現だった表現の割合
    - Recall: 固有表現である表現の中で，固有表現と予測された表現の割合
  - Macro-F1: タグのサンプル数が考慮されないF値
  - Micro-F1: タグのサンプル数が考慮されたF値

単語	主訴	:	腹痛	、	頭痛	既往	歴	:	糖尿病	、	気管支	喘息
正解	B-CC	O	B-FD	O	B-FD	B-PH	I-PH	O	B-FD	O	B-FD	I-FD
予測	B-CC	O	B-CC	O	O	B-PH	I-PH	O	B-PH	B-FD	B-FD	I-FD

⇨

正解

予測	CC	FD	PH	O
CC	1	1	0	0
FD	0	1	0	1
PF	0	1	1	0
O	0	1	0	3

⇨

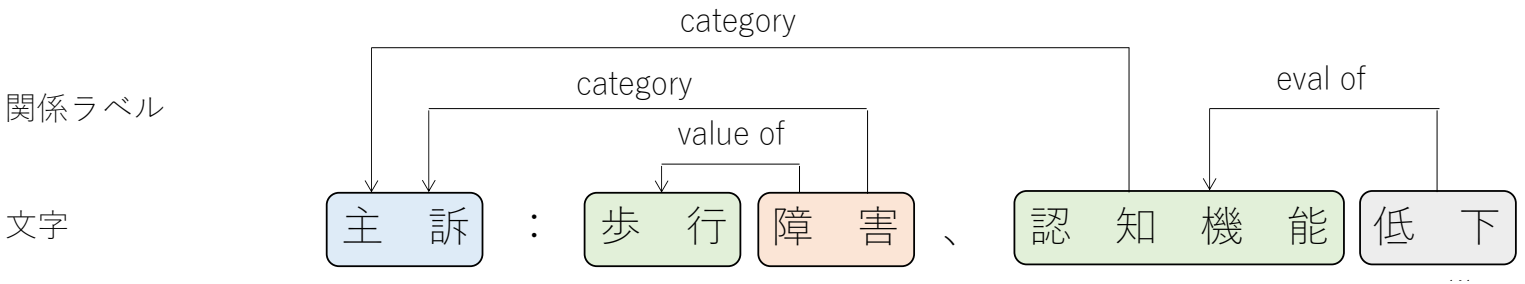
Macro-F1 = 0.56
Micro-F1 = 0.50
(TP=3, FP=3, FN=3)

算出例

[5] Sang, Erik F., and Sabine Buchholz. "Introduction to the CoNLL-2000 shared task: Chunking." arXiv preprint cs/0009008 (2000).

方法：関係抽出の評価方法

- 通常の分類問題と同様の評価方法でF1を計算



「障害」から他の単語への関係の正解ラベル: [category, None, value_of, None , None , None]

「障害」から他の単語への関係の予測結果: [category, value_of, value_of, category, None, None]

正解

予測	category	value_of	None
category	1	0	1
value_of	0	1	1
None	0	0	2

⇨ Macro-F1 = 0.67
Micro-F1 = 0.67
(TP=2, FP=2, FN=0)

※ Macro, Micro F1の算出時，正解ラベルがNoneである事例は無視する

結果：実験結果

BERT	指標	NER		RE	
		F1 (S.D)	95% CI	F1 (S.D)	95% CI
UTH	Micro	0.932 (0.006)	0.923-0.940	0.764 (0.013)	0.744-0.780
	Macro	0.782 (0.029)	0.736-0.817	0.567 (0.035)	0.511-0.616
NICT	Micro	0.913 (0.003)	0.909-0.919	0.743 (0.005)	0.735-0.750
	Macro	0.772 (0.015)	0.749-0.794	0.537 (0.020)	0.507-0.557

- NERはMicro-F1，Macro-F1共にUTH-BERTを用いた方が精度が高い
  - UTH-BERTとNICT-BERTの差はMicroで0.02，Macroで0.01
- REもMicro-F1，Macro-F1共にUTH-BERTを用いた方が精度が高い
  - UTH-BERTとNICT-BERTの差はMicroで0.02，Macroで0.03

考察：固有表現抽出

タグ	意味	UTH		NICT		差	例		
		F値	タグ数	F値	タグ数				
state	患者の状態	0.922	2615	0.889	2598	0.033	異常,	所見,	腫瘤
body	人体部位	0.938	1373	0.904	1366	0.034	腹部,	肺,	下肢
value	数値	0.961	883	0.959	883	0.002	5,	2,	程度
unit	単位	0.968	737	0.960	736	0.008	mg/dl,	mmHg,	%
PN_Positive	肯定表現	0.960	1015	0.953	1010	0.007	認めた,	であった,	した
item	所見項目	0.922	1005	0.907	1003	0.015	体重,	身長,	血圧
clinical_test	臨床検査	0.923	589	0.910	587	0.013	検査,	MRI,	解析
PN_Negative	否定表現	0.967	351	0.956	348	0.011	なし,	なく,	認めず
time	時間	0.973	692	0.959	686	0.014	時,	後,	現在

- 患者の状態と人体部位はF値の差がBERT間で大きい
  - 患者の状態や人体部位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
    - UTH-BERTではより適当な埋め込み表現を学習できた可能性
- 数値，単位や肯定表現はF値の差がBERT間で小さい
  - 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においても頻出
    - それぞれのBERTで同程度の質の埋め込み表現が学習
- 所見項目，臨床検査，否定や時間表現は一定の差
  - 数値や単位と同様にどちらのテキストでも使用される表現であると考えられるが，UTH-BERTのF値が高い
  - 今度，データ数を増加させるなど更なる検討が必要

考察：関係抽出

ラベル	意味	UTH		NICT		例
		F値	ラベル数	F値	ラベル数	
value_of	headがtailの値である	0.820	3924	0.802	3901	疼痛 (ent: state)を認めない (ent: PN_Neg)
site	headがtailの部位である	0.727	1292	0.689	1285	四肢 (ent: body)の筋力 (ent: state)低下
method	headがtailにより得られる	0.724	887	0.706	885	聴診 (ent: clinical_test)上、異常は無い (ent: PN_Neg)
unit	headがtailの単位である	0.944	792	0.947	792	身長170 (ent: value)cm (ent: unit)

- *rel: value_of*, *site*, *method*はUTHの方が精度が高い
  - これらの関係ラベルは*ent: state*, *body*, *item*や*PN_Positive/Negative*である固有表現間に付与される傾向
  - 上記の固有表現タグの抽出精度（NER）はUTHの方が高い
- *rel: unit*はNICTの方が精度が高い
  - この関係ラベルは*ent: value*や*unit*である固有表現間に付与される傾向
  - 上記の固有表現タグの抽出精度はUTHの方が僅かに高いが，両者の間に大きな差はない

NERがREの精度に一定の影響を与えている
REの精度の改善にはNERの精度の更なる向上が必要

結語と今後の課題

- 結語
  1. 診療テキストで事前学習したモデルにおいて，固有表現抽出はMicro-F1で0.932，関係抽出は0.764
  2. 診療テキストに独特な患者の状態と人体部位のような表現は、診療テキストで事前学習したモデルが有効である可能性が示唆
- 今後の課題
  - 文字ベースのBERTを用いた情報抽出の精度評価
  - 本研究で学習したモデルを用いて，退院サマリなどを対象とした情報抽出の精度評価