電子カルテのテキストを対象とする UTH-BERT を用いた一日単位の症状の有無判定

國方 淳[*1], 間島行則[*1,2], 横井 英人[*1,2]
[*1] 香川大学医学部附属病院 臨床研究支援センター
[*2] 香川大学医学部附属病院 医療情報部

Daily symptom detection using UTH-BERT from texts in electronic medical records.

Jun Kunikata[*1], Yukinori Mashima[*1,2], Hideto Yokoi[*1,2]
[*1] Clinical Research Support Center, Kagawa University Hospital
[*2] Department of Medical Informatics, Kagawa University Hospital

電子カルテからの患者の症状の自動抽出は重要なタスクであるが，一般に症状は自由記載のテキストデータとして保存されているため抽出は容易ではない．一方で近年，自然言語処理のブレイクスルーとして Google 社により深層学習における BERT モデルが開発され，さらにそれを医療テキストで追加学習した UTH-BERT モデルが東京大学より公開された．そこで，今回我々は UTH-BERT を用いて当院の電子カルテの1ヶ月分のテキストを対象とし，悪心嘔吐と下痢について患者ごとの一日単位の症状の有無判定をタスクとして学習と評価を行った．

延べ 10,805 日分の学習データと 2,951 日分のテストデータをダウンサンプリングして学習と評価を行った結果，悪心嘔吐について ROC-AUC 0.994，Precision-Recall-AUC 0.924，下痢について ROC-AUC 0.998，Precision-Recall-AUC 0.939 と比較的高い精度が得られるとともに，従来の方法が不得手とする離れた単語間の関係性を正しく認識できる傾向がみられた．また，UTH-BERT は医療テキストで事前学習されていることから，単純なタスクであれば比較的少ないデータ量でも性能を発揮しやすいことが示唆された．

キーワード 自然言語処理，機械学習, BERT, UTH-BERT

1. はじめに

医療施設における症状発生のモニタリングやリアルワールドデータを用いた大規模研究において，患者に生じた症状を自動的に収集することは重要なタスクである．しかし，電子カルテには病名や処方などの情報がフォーマット化された構造化データとして保存されているのに対し，症状の情報は多くの場合，自由記載のテキストとして記録されているためプログラムを用いて自動抽出するのは容易ではない．

我々は第 39 回医療情報学連合大会にて自然言語処理と機械学習の手法を用いた症状の有無判定について報告した[1]が，近年の機械学習分野の発展は目覚ましく，特に自然言語処理の分野では Google 社が 2018 年に発表した BERT[2]が大きな進展をもたらした．本稿では東京大学の医療 AI 開発学講座が BERT を用いて医療テキストで学習を行った UTH-BERT[3]を用いた悪心嘔吐および下痢症状の有無判定の手法とその性能評価について報告する．

2. 方法

1) データ

香川大学医学部附属病院の 2015 年 10 月の入院患者の電子カルテデータより医師記載，看護師記載，リハビリ等の実施記録，救急診療記録をテキストデータとして抽出した．患者ごとの 1 日分のテキストを単位として，10 月 1 日から 24 日の延べ 10,805 日分のテキストデータを学習データ，25 日から 31 日の延べ 2,951 日分のテキストデータをテストデータとした．

2) ラベリング

評価者がテキストを読み，記載日に症状があると判断した場合を「症状あり」，ないと判断した場合を「症状なし」とラベリングした．症状の有無は 1 日ごとの判定であるため，記載日と異なる日に症状があった場合は「症状なし」とした．「症状あり」とラベリングされたのは悪心嘔吐の学習データで 651 日(6.0%)，テストデータで 179 日(6.1%)，下痢の学習データで 547 日(5.1%)，テストデータで 128 日(4.3%)であった．

3) 前処理

悪心嘔吐および下痢のそれぞれについて，スクリーニング文字列を定義し，スクリーニング文字列を含まない文を削除した．定義したスクリーニング文字列を Table.1 に示す．その後は UTH-BERT の前処理プロセスに従った．

Table.1 スクリーニング文字列

悪心・嘔吐	悪心，おしん，嘔，吐，おうと，おうき，はいた，はいて，はき，はく，vom，naus，戻す，戻し，もどす，もどし，気分，きぶん，気持ち，きもち，ムカ，むか，胃部不快，えずく えずき，えずい，えづく，えづき，えづい
下痢	下痢，げり，水様，泥状，軟便，ゆるい，緩い，diar，便，うんち，うんこ

4) 学習と評価

プログラミング言語として Python 3.6.9，深層学習ライブラリとして Tensorflow 2.3.1 を使用し，pre-trained モデルとして UTH-BERT-BASE-128 を用いた．悪心嘔吐および下痢のそれぞれについて学習データを対象として「症状あり」「症状なし」の二値分類タスクの fine tuning を行い，テストデータを用いて予測・評価を行った．

3. 結果

Fine tuning 後のモデルを用いたテストデータに対する予測結果の混同行列を Table.2 と Table.3 に，予測能の各評価指標を Table.4 と Table.5 に示す．

Table.2 悪心嘔吐判定の混同行列

	Predict Positive	Predict Negative
Actual Positive	166	13
Actual Negative	56	2716

Table.3 下痢判定の混同行列

	Predict Positive	Predict Negative
Actual Positive	125	3
Actual Negative	33	2790

Table.4 予測能の評価指標 1

	Accuracy	Precision	Recall	F1-score
悪心嘔吐	0.977	0.748	0.927	0.828
下痢	0.962	0.791	0.977	0.874

Table.5 予測能の評価指標 2

	ROC-AUC	Precision-Recall-AUC
悪心嘔吐	0.994	0.924
下痢	0.992	0.939

Table.2, 3, 4 は症状ありの確率が 0.5 以上と予測された場合を Predict Positive としたときの値を示している．

4. 考察

悪心嘔吐，下痢のいずれにおいても Precision-Recall-AUC が 0.92 以上を示しており，比較的高い判定精度を達成できた．前研究の悪心嘔吐を対象として Bag of Words と非深層学習的機械学習を用いた方法と比較して Precision-Recall-AUC で 0.89 からの上乗せがあり，その要因として「悪心は咳嗽時にみられる」のように症状と動詞の間に別の語句が入った文や，「悪心の出現はみられない」のような否定文をより正しく判定できる傾向がみられたことが挙げられる．しかし，データ量の不足により学習がすぐに飽和しており，本来の性能を発揮するためにはより多くの学習データの収集やデータオーグメンテーションが必要になると考えられる．

5. 結語

UTH-BERT は症状の有無判定タスクにおいて良好な性能を示した．また，UTH-BERT は医療テキストで事前学習されていることから，二項分類のような単純なタスクであれば比較的少ないデータ量でも性能を発揮しやすいことが示唆された．

参 考 文 献

[1] 國方淳，十河智昭，間島行則他：電子カルテに記載されたテキストを対象とした機械学習による日単位での嘔気嘔吐症状の有無判定：第 39 回医療情報学連合大会論文集, 498-501, 2019.

[2] J Devlin, MW Chang, K Lee, et al: BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv Prepr arXiv181004805, 2018.

[3] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, et al: A clinical specific BERT developed with huge size of Japanese clinical narrative, medRxiv, 2020.