一般口演18
自然言語処理

2020年11月21日(土) 11:15 〜 12:38 F会場 (イベントホール・特設会場1)

[4-F-1-03] 形態素解析が日本語医学 BERTモデルに与える影響

*和田 聖哉[1]、武田 理宏[1]、真鍋 史朗[1]、小西 正三[1]、松村 泰志[1] （1. 大阪大学大学院医学系研究科医学専攻 情報統合医学講座医療情報学）

*Shoya Wada[1], Toshihiro Takeda[1], Shiro Manabe[1], Shozo Konishi[1], Yasushi Matsumura[1] （1. 大阪大学大学院医学系研究科医学専攻 情報統合医学講座医療情報学）

キーワード：Natural Language Processing, Deep Learning, Data Mining

文脈に沿った単語分散表現の獲得を可能にする Bidirectional Encoder Representations from Transformers (BERT)の発表後，医学領域に適合した英語 BERTモデルが複数公開され、様々な医療自然言語処理タスクで精度向上に寄与したことが報告されている。我々は小規模医学コーパスと日本語 Wikipediaコーパスを組み合わせて事前学習を行うことで、実用に足る日本語医学 BERTモデルが構築可能であることをこれまでに報告した。
日本語の自然言語処理では、英語のように単語間にスペースがある言語と違い、形態素解析で適切な言語単位に単語分割した後に行う必要がある。日本語形態素解析には、意味処理には構文解析との連携に強い JUMAN、音声のことを考えた場合には最小単位で単語分割を行う unidic、情報抽出には固有表現が積極的に登録されている neologdを選択することが好ましいとされる。一般ドメインでは異なる形態素解析を適用した BERTモデルが京都大学、東北大学、情報通信研究機構などから公開され、タスクに応じた使い分けが報告されているが、日本語医学 BERTモデルでは、そのような報告はまだない。
そこで我々は、医療系の自然言語処理タスクで BERTを用いる際に適した形態素解析について調査する。具体的には、まず医学教科書の記述から、① 記載されている文章から疾患領域の分類を行うマルチクラス文書分類タスク、②教科書に記述されている疾患名、症状、検査所見等を抽出する固有表現抽出タスクの2つを整備する。次に形態素解析に用いる辞書を切り替えて BERTの事前学習モデルをそれぞれ構築し、各々の精度を評価する。

形態素解析が日本語医学 BERT モデルに与える影響

和田　聖哉[*1]、武田　理宏[*1]、真鍋　史朗[*1]、小西　正三[*1]、松村　泰志[*1]
[*1] 大阪大学大学院医学系研究科　医療情報学

The Effect of Morphological Analysis on The Japanese Medical BERT Model

Shoya Wada[*1], Toshihiro Takeda[*1], Shiro Manabe[*1], Shozo Konishi[*1], Yasushi Matsumura[*1]
*1 Department of Medical Informatics, Osaka University Graduate School of Medicine

With the development of contextual embeddings such as Bidirectional Encoder Representations from Transformers (BERT), the performance of information extraction from free texts by natural language processing has remarkably improved in both the general and medical domains.

When we perform natural language processing in Japanese, we need to apply tokenization by morphological analysis or other methods in pre-processing. In the general domain, several Japanese pre-trained BERT models with different morphological analyses have been published, however, in the medical domain, only one model has been published yet.

We have reported that it is possible to pre-train a practical Japanese medical BERT model from a small corpus. Therefore, we evaluated some tokenization methods (morphological analysis with MeCab or subword segmentation with SentencePiece) suitable for Japanese medical BERT models in terms of the medical document disease domain classification task. We confirmed that the accuracy varies by 1.11 points depending on the different tokenization methods. We also confirmed that the models we pre-trained had about 4 points higher accuracy than the published models.

Keywords: Deep Learning, Natural Language Processing, Data Mining

1. 緒論

文脈に沿った単語分散表現の獲得を可能にする Bidirectional Encoder Representations from Transformers (BERT)[1]の発表後、医学領域に適合した英語 BERT モデルが複数公開され、様々な医療自然言語処理タスクで精度向上に寄与したことが報告されている。[2) 3)]

日本語に関しては、日本語 Wikipedia で事前学習を行った BERT モデルが京都大学、東北大学、情報通信研究機構などから公開されている。また、医療分野では、東京大学から日本語で記載された大規模なカルテ記録で事前学習を行ったモデルが公開されている (UTH-BERT)。[4)]

日本語の自然言語処理では、英語のように単語間にスペースがある言語と違い、前処理の段階で文章もしくは文を適切な言語単位（例：文章→文→句→単語→文字）へトークン化した後に、それを機械学習モデルに入力してタスクの学習と推論を行う。日本語のトークン化手法では、フリーテキストを言語上で意味を持つ最小単位（形態素）に分け、それぞれの品詞や変化などを判別して分割する形態素解析が代表的である。日本語形態素解析には、構文解析との連携に強い JUMAN、最小単位で単語分割を行う Unidic、固有表現が積極的に登録されている NEologd などがある。一般ドメインでは異なる形態素解析を適用した BERT モデルが公開され、タスクに応じた使い分けが報告されているが、[5) 6)] 日本語医学 BERT モデルは現在 UTH-BERT のみが利用可能であるため、そのような比較をした報告はまだない。

我々は小規模医学コーパスと日本語 Wikipedia コーパスを組み合わせて事前学習を行うことで、実用に足る日本語医学 BERT モデルが構築可能であることを以前報告した。[7)] その技術を活用し、医療系の自然言語処理タスクで BERT を用いる際に適した形態素解析について調査した結果を報告する。

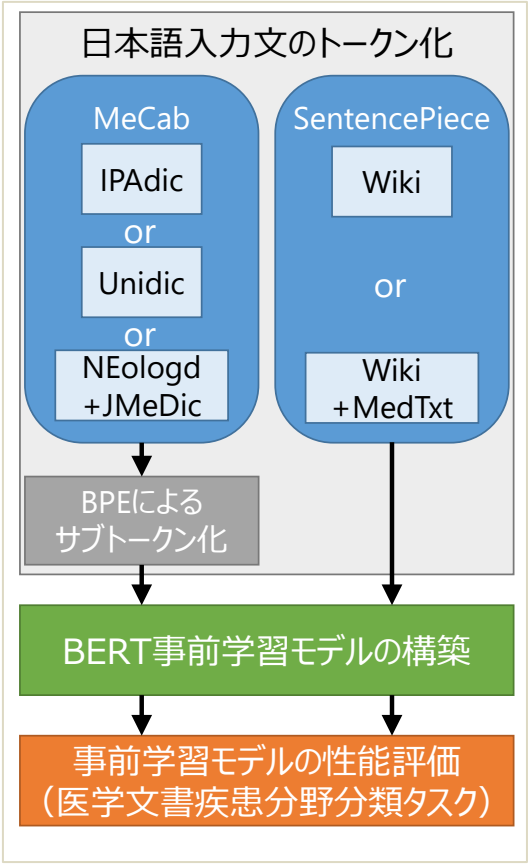

図 1　研究概要

2. 目的

本研究の概要を図 1 に示す。まず、複数の日本語文書のトークン化手法を適用して BERT の事前学習モデルをそれぞれ構築する。次に、Web で公開されている医学知識の記述から、記載されている文章から疾患領域の分類を行うマルチクラス文書分類タスクを整備してデータセットを作成し、各々の精度評価を行う。

3. 方法

3.1 日本語フリーテキストのトークン化

日本語文書のトークン化には、MeCab[1]による形態素解析と、SentencePiece[8]によるサブワード分割をそれぞれ事前学習の前処理に実装した。分割結果の事例を巻末表 1 に示す。

3.1.1 MeCab から利用可能な形態素解析辞書

IPA 辞書は、IPA 辞書、IPA コーパス に基づき Conditional Random Field (CRF) でパラメータ推定した辞書である。[1]IPA 辞書を用いて MeCab による形態素解析を行った場合、その分割単位は後述する UniDic 辞書、NEologd 辞書を使用した場合と比較すると、その中間に位置すると本研究では考えた。

UniDic 辞書は UniDic 辞書, BCCWJ コーパスに基づき CRF でパラメータ推定した辞書であり、国立国語研究所の規定した斉一な言語単位(短単位:日本語文章を形態論的側面に着目し、最小単位(文字)の結合により構築)に基づく。[9]

NEologd 辞書は多数の Web 上の言語資源から得た新語を追加することでカスタマイズされた MeCab 用のシステム辞書であり、多くの固有表現を収録している。[10] 本辞書を基にした形態素解析による分割結果を、本研究では長単位として扱った。

万病辞書 (J-MeDic)は医療従事者が記載した電子カルテや退院サマリから、症状や病名に関連する語を広く抽出したデータを基に構築された辞書であり、MeCab から利用可能な辞書形式が公開されている。[11]

本研究では、IPA 辞書、UniDic 辞書を用いた MeCab による形態素解析をそれぞれ **with MeCab-IPAdic**、**with MeCab-Unidic** とする。また、UTH-BERT では、NEologd 辞書と J-MeDic の両方を MeCab による形態素解析時に利用している。[4] 簡単のため、これを **with MeCab-MedTerm** と表記して他の外部辞書使用例と区別する。

3.1.2 SentencePiece

SentencePiece はサブワード(事前にテキストを単語分割して各単語の頻度を求めた後、高頻度の単語は 1 単語として扱い、低頻度語はより短い単位(文字や部分文字列)に分割する手法)をベースにしたアルゴリズムである。[8] 事前の形態素解析による単語分割は不要で生文から直接分割を学習し、学習した言語モデルに基づいて複数のサブワード分割候補を確信度に従い提供する。

本研究では、MeCab による形態素解析を適用した事前学習モデルと区別するために、**with SPM** と表記する。

3.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT を用いた自然言語処理は2段階で構成される。まず、大規模な生コーパスを用いて教師なし事前学習を行う。次に、そこで獲得した重みで初期化し、最後に分類などのタスクに合わせた全結合層を1層追加したモデルを作成し、教師あり学習でパラメータの微調整(fine-tuning)を行い、推論モデルを最適化する。

医学 BERT モデルを構築する際の課題は、大量の医学文書を数 GB 単位で収集する作業である。英語では PubMed 等のリソースを利用可能であるが、日本語では大規模医学知識コーパスを自由に利用に出来る場所はまだない。我々は小規模医学コーパスからでも日本語 Wikipedia コーパスを組み合わせて同時に事前学習を行うことで、実用に足る日本語医学 BERT モデルが構築可能であることを以前報告した。[7] その技術を活用し、異なるトークン化手法を適用した BERT の事前学習モデルをそれぞれ構築し、本研究で比較した。

Byte Pair Encoding (BPE: 後述)[12] による vocabulary 構築と日本語医学文書分類タスクの fine-tuning には、それぞれ huggingface の Tokenizers, Transformers を使用した。[13] SentencePiece を用いた事前学習モデルでは、vocabulary の学習とフリーテキストのトークン化に標準モジュールを利用した。[2]

3.2.1 日本語医学 BERT モデルの事前学習

以前我々の提案した手法に則り、日本語医学 BERT モデルを構築した。[7] すなわち、一般ドメインの文書には日本語 Wikipedia のダンプデータ(約 2.6GB)を、医学文書には今日の診療プレミアム(医学書院)に収録されている教科書 15 冊(計 83MB)を使用した。医学文書から生成される事前学習インスタンスを一般ドメイン文書の 32 倍(ファイルサイズ比率)で学習するように事前学習データを構成した。

BERT の vocabulary は、out of vocabulary を減らすために、Byte Pair Encoding (BPE)[12] によるサブワードで構成されている(例:「脳梗塞」、「心筋梗塞」が低頻度語で、「脳」や「心筋」が単語として閾値以上の頻度で存在する場合、「梗塞」をサブワードとして、それぞれ「脳 ##梗塞」、「心筋 ##梗塞」("##" はサブワードを区別するための記号)と表現出来るので、個別に vocabulary へ登録するよりも効率化出来る)。前出の一般文書と医学文書をそのまま BPE で学習した場合、医学文書量が一般文書比べて圧倒的に少ないため、医学文書に出現する表現の多くは低頻度語として扱われる。その結果、医学用語の表現に向かない vocabulary が構築される可能性がある。これを回避するため、BPE による vocabulary 構築時に、一般文書はそのまま扱う一方で、医学文書を 32 倍モジュールに読み込ませてボキャブラリを構築した。これにより擬似的に一般文書サイズと医学文書サイズを均一に扱うようにした。vocabulary のサイズは全て 32,000 語で統一した。また、SentencePiece を用いた vocabulary 構築時も、そのサイズを全て 32,000 語に設定した。また、SentencePiece によるサブワードの分割候補には、公開モデル [14] に倣い、ベスト解を採用した。

3.2.2 Fine-tuning(日本語医学文書分類)

日本語事前学習モデルを、医学文書の疾患分野分類タスクで評価するために、MSD マニュアルプロフェッショナル日本語版 [3]から疾患分野毎にページを取得してスクレイピングを行った。学習データの形式は、各文書の本文を入力文、その疾患分野を正解ラベルとした(マルチクラス文書分類:予測は与えられたラベル候補のうち、いずれか一つが選択される)。

[1] https://taku910.github.io/mecab/
[2] https://github.com/google/sentencepiece
[3] https://www.msdmanuals.com/ja-jp/professional

表.1 医学文書疾患分野分類タスク

Class	Number
小児科	330
感染性疾患	245
神経疾患	154
婦人科および産科	152
その他のトピック	151
外傷と中毒	149
皮膚疾患	127
血液学および腫瘍学	120
消化器疾患	118
心血管疾患	106
泌尿器疾患	103
肺疾患	103
筋骨格疾患と結合組織疾患	90
眼疾患	79
内分泌疾患と代謝性疾患	78
精神障害	74
耳鼻咽喉疾患	70
肝胆道疾患	63
老年医学	48
免疫学；アレルギー疾患	44
栄養障害	40
歯科疾患	31
計	2,475

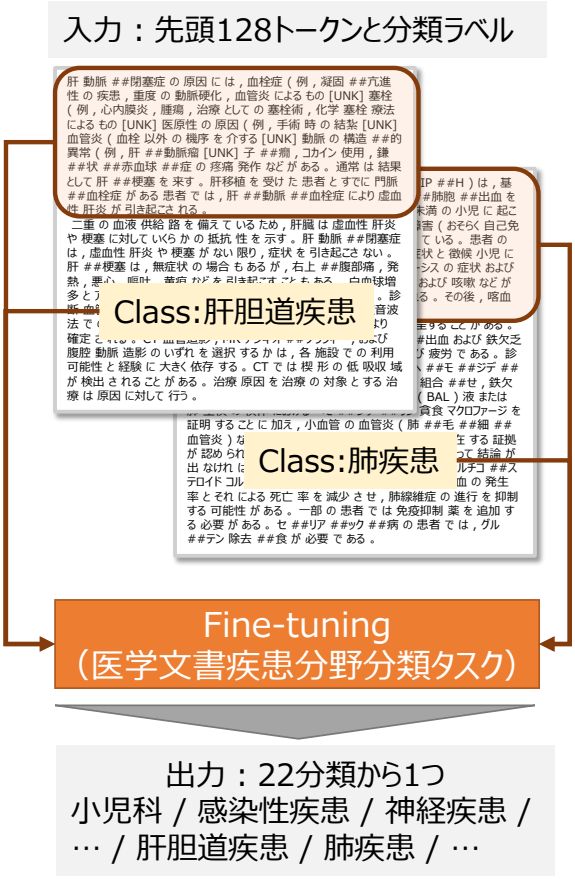

図 2 Fine-tuning の概要

BERT の事前学習モデルでは入力トークン数に上限値がある。我々はそれを 128 に設定して事前学習モデルを構築した。Fine-tuning に用いる入力長を各事前学習モデルで揃えるために、入力には各文書の先頭 128 トークンまでを利用した（図 2）。層化 5 分割交差検証を、異なる初期シード値 5 種類で実施して分類精度の平均値を比較した。各クラスに属するデータ数は表 1 に示した通り、不均衡データであった。評価指標には個別の事例の分類精度を反映する micro-F1 を採用した。

3. 2. 3 比較対象となる事前学習モデル

本研究で構築した事前学習モデルと今回比較する公開モデルをここで整理する。

MedBERT は 3.2.1 項で示した手法で我々が構築した事前学習モデルである。文書のトークン化手法の違いで、with {MeCab-MedTerm, MeCab-IPAdic, MeCab-Unidic, SPM-Wiki+MedTxt}が存在する。

UTH-BERT は日本語で記載された大規模なカルテ記録をデータセットとして事前学習を行ったモデルで、東京大学が公開している。その vocabulary は、NEologd 辞書と J-MeDic を併用して MeCab で形態素解析を行った医療テキストに BPE を適用したサブワードから構成されている。トークン化手法を明示する場合には、**UTH-BERT with MeCab-MedTerm** と記載する。

日本語 Wikipedia で事前学習を行ったモデルを東北大学が公開している。その vocabulary は日本語 Wikipedia に対して、MeCab-IPAdic で形態素解析を行った後に BPE を適用したものから構築されている。本研究ではこれを **TOHOKU-BERT with MeCab-IPAdic** と呼び、一般ドメインのコーパスで事前学習したモデルと考えた。また、日本語 Wikipedia のトークン化に SentencePiece を用いて日本語 Wikipedia で事前学習を行ったモデルも利用可能である。[14]これを **BERT with SPM-Wiki** とする。

4. 結果

各トークン化手法の出力結果の違いを、「慢性膵炎では，糖尿病性ケトアシドーシス(DKA)の合併は比較的稀である．」という文章を入力にして、巻末表 1 に示した。TOHOKU-BERT with MeCab-IPAdic には、「膵」が登録されていないため、「膵炎」が未知語として「[UNK]」に置換された。MedBERT with MeCab-IPAdic ではその問題が解消されていた。また、「ケトアシドーシス」が一語で表記出来ていた。MeCab-MedTerm で形態素解析を行った場合、「慢性膵炎」は一語で扱われるようになった。SentencePiece を用いた場合、文書分割位置は品詞に基づかない。「である」に関して、MeCab では「で / ある」と分割されたが、SentencePiece の場合は頻出表現と扱われたために分割されなかった。また、日本語 Wikipedia のみで学習させた言語モデル (SPM-Wiki) に比べ、医学文書も同時に学習させた言語モデル (SPM-Wiki+MedTxt) では、医学文書で頻出する表現（「の合併」「は比較的」「稀である」など）が一つのトークンとして扱われた。その結果、文章の分割数は 12 となり、各トークン化手法の中で最も小さくなった。

医学文書の疾患分野分類タスク精度を各事前学習モデルで評価した（図 3）。MedBERT は UTH-BERT を含めた既存

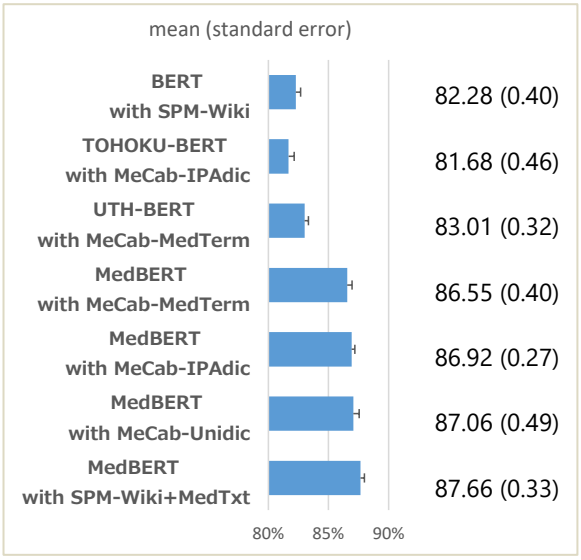

図 3　各事前学習モデルの医学文書疾患分野分類精度

公開モデルに比べ、全体として約 4 ポイント精度が高かった。最も精度の高いモデルは SentencePiece を用いたモデルであり、87.66 %の精度で文書分類が出来た。MedBERT で MeCab による形態素解析を利用したもののうち、MedTerm が他の 2 つに比べ約 0.4 ポイント精度が低かった。

トークン化手法の違いによる影響を入力文字列の観点から評価した。まず、最大入力長 128 トークンの制限下で入力される文字数を比較した（表 2）。MedBERT のうち、with SPM-Wiki+MedTxt は今回比較したトークン化手法の中で最も入力文字数が多く、279.4 であった。これは他の MedBERT で用いたトークン化手法と比較したところ、1.2～1.4 倍に相当した。

次に、サブワード出現頻度を一文書単位で評価した（表 3。SentencePiece によるトークン化手法では、入力文は品詞に従った単語の単位とは無関係に分割される。そのために「サブワード」の同定が困難であり、本検証では省略した）。MedBERT with MeCab-MedTerm におけるサブワード出現頻度は 0.100 であり、with MeCab-Unidic と比較して 2.9 倍大きかった。

5. 考察

専門書レベルの医学文書疾患分野分類タスクにおいて、本研究で構築した事前学習モデル (MedBERT) は既存モデルよりも高い精度で分類できることを確認した。また、トークン化手法の比較では、疾患名などの医学用語が辞書ベースで登録されている with MeCab-MedTerm が最も精度が低く、医学教科書と一般文書から分割位置を学習した言語モデルを採用した with SPM-Wiki+MedTxt が最も精度が高かった。

既存モデルと我々が構築したモデルの最も大きな違いは、事前学習用コーパスに医学教科書を使ったか否かにある。BERT with SPM-Wiki、TOHOKU-BERT with MeCab-IPAdic は、日本語 Wikipedia を事前学習に用いた、一般ドメイン向けの事前学習モデルである。UTH-BERT がこれらよりも文書分類精度で 0.73～1.33 ポイント高かったのは、事前学習に大規模カルテ記録を用いていることで医学知識をある程度学習出来ていたためと考えられる。しかしながら、医学教科書と大規模カルテ記録とではその記載内容は異なる。そのために分類に深い医学知識を要する本タスクにおいて、我々の構築した

表.2　一文書当たりの入力文字数平均値（標準偏差）

BERT with SPM-Wiki	222.3 (27.7)
TOHOKU-BERT with MeCab-IPAdic	198.5 (20.0)
UTH-BERT with MeCab-MedTerm	204.9 (21.6)
MedBERT with MeCab-MedTerm	223.8 (25.4)
MedBERT with MeCab-IPAdic	211.1 (23.1)
MedBERT with MeCab-Unidic	202.0 (20.9)
MedBERT with SPM-Wiki+MedTxt	279.4 (36.9)

表.3　一文書当たりのサブワード出現頻度（標準偏差）

TOHOKU-BERT with MeCab-IPAdic	0.107 (0.069)
UTH-BERT with MeCab-MedTerm	0.153 (0.085)
MedBERT with MeCab-MedTerm	0.100 (0.067)
MedBERT with MeCab-IPAdic	0.043 (0.056)
MedBERT with MeCab-Unidic	0.035 (0.049)

MedBERT の方が全体的に高精度になったものと思われる。このように医学分野の自然言語処理タスクにおいて事前学習に用いたコーパスで精度が異なるという報告に関しては、PubMed の文章から事前学習を行った BioBERT と、MIMIC-III の clinical note で事前学習を行った ClinicalBERT との比較で既に報告されており、我々の仮説を裏付けるものと言える。[3)]

MedBERT で使用したトークン化手法の中では、with SPM-Wiki+MedTxt が最も精度が高く、with MeCab-MedTerm が最も低かった。これは、トークン化手法の違いで、対象文書の入力文字数とサブワード出現頻度が異なることが影響していると考えられる。

表 2 では、各トークン化手法による入力文字数の違いを示した。最も精度の高かった with SPM-Wiki+MedTxt は今回比較したトークン化手法の中で最も入力文字数が多かった。これは本トークン化手法がタスク入力文を最も効率的に表現できることを意味する。この結果は、本タスクでは入力文字数の増加が精度向上に寄与する可能性を示唆している。

一方で、頻出表現が vocabulary に登録されているという点では with MeCab-MedTerm も同様であり、MeCab による形態素解析を行ったものの中では、入力文字数の平均値は 223.8 と最も多かった。しかしながら、その分類精度は MedBERT の中で最も低かった（with MeCab-Unidic と比較すると、0.51 ポイント低下）。サブワード分割がその一因と考え、その出現頻度を一文書単位で評価した（表 3）。MedBERT with MeCab-MedTerm におけるサブワード出現頻度は 0.100 であり、with MeCab-Unidic と比較して 2.9 倍大きかった。UniDic 辞書は日本語文章を形態論的側面に着目し、最小単位（文字）の結合による短単位で構成されている（巻末表 1 の事例を参照）。その結果、医学文書の分割において、UniDic 辞書を利用し

た BPE ではサブワードの出現頻度が他に比べ低下する。しかし分割数は多くなる傾向があり、そのため入力トークン数の上限が存在する条件では、入力文字数の平均は 202.0 と他に比べ小さかった。それにも関わらず with MeCab-MedTerm より分類精度が 0.51 ポイント高かったという実験結果は、BPE によるサブワードの出現頻度増加が医学文書疾患分野分類タスクの精度低下に寄与する可能性を示唆している。言い換えると、本実験結果は、日本語 BERT モデルの構築時に疾患名などの固有表現を vocabulary へ積極的に反映させることが、タスク次第では不利になる可能性があることを示している。今後さらなる検証が必要である。

本研究にはいくつかの制限が存在する。まず、本研究で精度評価のために用いたデータセットは個人で構築したものであり、これが第三者による再現の障害になり得ると考えられる。該当 Web ページのスクレイピングを行えば本データセットの再構築は実現出来るが、完全な再現は不可能である。これは、日本語医学領域の自然言語処理において、利用可能な共用タスクの言語資源が少ないという課題が根底にある。

次に、本研究は文書分類タスクでの評価であり、他の自然言語処理タスクにおいては、最適なトークン化手法は本研究の報告とは異なる可能性がある。文法的な分割を考慮する形態素解析が不要な SentencePiece によるサブワード分割の場合、(複合)名詞単位で抽出することの多い固有表現抽出タスク(例:疾患名や症状に該当する文字列を抽出するタスク)では、トークン化の段階でエラーが生じる可能性がある。従って、一般的な自然言語処理と同様に、前処理段階におけるトークン化手法には、目的に沿った分割を採用するのが最適解になると思われる。しかし現時点で公開されている日本語医学 BERT モデルは UTH-BERT のみであり、一般日本語ドメインのように多様な選択肢はまだ存在しない。

最後に、本研究では以前我々が報告した、小規模の医学コーパスしか入手できない状況から医学 BERT モデルを構築する手法を適用している。[7] そのため大規模医学コーパスで事前学習したモデルが存在した場合には、最適なトークン化手法に関しては異なる結論に至る可能性がある。また、本研究で構築した事前学習モデルは性能面でそれに及ばない可能性がある。しかしながら、複数のトークン化手法を個別に適用した事前学習モデルのいずれもが既存モデルより高性能であった結果は、本手法の有用性と頑健性を示唆するものであり、トークン化手法の違いによる精度差は本報告と同様の傾向を示す可能性が高い。

6. Future works

研究制限で述べたように、医学文書の自然言語処理に適した事前学習モデルを構築する上では、まずそのベンチマークとなる指標の整備が必要となる。それと並行して、公開可能な事前学習モデルの構築について検討したい。

7. 結論

日本語医学 BERT モデルにおいて、医学文書疾患分野分類タスクの精度は前処理の文章トークン化手法の影響を受けることを確認した。医学分野に特化した共用タスクの整備や事前学習モデルの公開・検証が今後の課題である。

8. 謝辞

本技術開発には，2018 年 10 月 11 日から参画した戦略的イノベーション創造プログラム(SIP)「AI（人工知能） ホスピタルによる高度診断・治療システム」の研究開発資金を一部充当し，開発を推進しています．

9. 参考文献

1) Devlin J, Chang MW, Lee K, and 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: Human Language Technologies, Volume 1 (Long and Short Papers): 4171-4186.
2) Lee J, Yoon W, Kim S, Kim D, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (2020) 36(4): 1234-1240.
3) Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (2019): 72-78.
4) Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed with huge size of Japanese clinical narrative. medRxiv preprint (2020).
5) 柴田知秀，河原大輔，黒橋禎夫. BERT による日本語構文解析の精度向上．言語処理学会第 25 回年次大会 (2019): 205-208.
6) 浅原正幸，西内沙恵，加藤祥. NWJC-BERT: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析．言語処理学会第 26 回年次大会 (2020): 961-964.
7) Wada S, Takeda T, Manabe S, Konishi S, Kamohara J, and Matsumura Y. A pre-training technique to localize medical BERT and enhance BioBERT. arXiv preprint (2020): arXiv:2005.07202.
8) 工藤拓. サブワード正則化: 複数のサブワード分割候補を用いたニューラル機械翻訳．言語処理学会第 24 回年次大会 (2018): 37-40.
9) 岡照晃:「言語研究のための電子化辞書」，コーパスと辞書，講座 日本語コーパス 7，朝倉書店，2019: 1-28.
10) 佐藤敏紀，橋本泰一，奥村学．単語分かち書き用辞書生成システム NEologd の運用 ー 文書分類を例にして ー．自然言語処理研究会研究報告 (2016): NL-229-15.
11) MEDNLP. 医療言語処理グループ 万病辞書 Mecab 用辞書データ (2019) [http://sociocom.jp/~data/2018-manbyo/ (cited 2020-Mar-18)].
12) Sennrich R, Haddow B, and Birch A. Neural machine translation of rare words with subword units. In Proceedings of the ACL, 2016.
13) Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. arXiv preprint (2019): arXiv: 1910.03771.]
14) Kikuta Y. BERT Pretrained model Trained On Japanese Wikipedia Articles. GitHub repository (2019): [https://github.com/yoheikikuta/bert-japanese (cited 2020-Aug-25)].

巻末表 1 入力文「慢性膵炎では，糖尿病性ケトアシドーシス(DKA)の合併は比較的稀である．」のトークン化事例

事前学習モデル	入力文のトークン化	BPE によるサブワード分割	分割数
TOHOKU-BERT with MeCab-IPAdic	慢性 / 膵炎 / で / は / ， / 糖尿 / 病 / 性 / ケトアシドーシス / （ / DKA / ） / の / 合併 / は / 比較的 / 稀 / で / ある / ．	慢性 / [UNK] / で / は / ， / 糖尿 / 病 / 性 / ケ / ##ト / ##アシ / ##ドー / ##シス / （ / D / ##KA / ） / の / 合併 / は / 比較的 / 稀 / で / ある / ．	24
MedBERT with MeCab-IPAdic	同上	慢性 / 膵炎 / で / は / ， / 糖尿 / 病 / 性 / ケトアシドーシス / （ / D / ##KA / ） / の / 合併 / は / 比較的 / 稀 / で / ある / ．	20
UTH-BERT with MeCab-MedTerm	慢性膵炎 / で / は / ， / 糖尿病 / 性 / ケトアシドーシス / （ / DKA / ） / の / 合併 / は / 比較的 / 稀 / で / ある / ．	慢性膵炎 / で / は / 、 / 糖尿病 / 性 / ケ / ##ト / ##アシドーシス / （ / Ｄ / ##ＫＡ / ） / の / 合併 / は / 比較的 / 稀 / で / ある / ．	20
MedBERT with MeCab-MedTerm	同上	慢性膵炎 / では / ， / 糖尿病 / 性 / ケトアシドーシス / （ / D / ##KA / ） / の / 合併 / は / 比較的 / 稀 / で / ある / ．	17
MedBERT with MeCab-Unidic	慢性 / 膵炎 / で / は / ， / 糖尿 / 病性 / ケトアシドーシス / （ / D / K / A / ） / の / 合併 / は / 比較 / 的 / 稀 / で / ある / ．	慢性 / 膵炎 / で / は / ， / 糖尿 / 病性 / ケトアシドーシス / （ / D / K / A / ） / の / 合併 / は / 比較 / 的 / 稀 / で / ある / ．	21
BERT with SPM-Wiki	_ / 慢性 / 膵 / 炎 / では / ， / 糖尿病 / 性 / ケ / ト / アシ / ドー / シス / （ / DKA / ） / の / 合併 / は 比較的 / 稀 / である / ．	-	21
MedBERT with SPM-Wiki+MedTxt	_ / 慢性膵炎 / では / ， / 糖尿病性 ケトアシドーシス / （ / D / KA / ） / の合併 / は比較的 / 稀である / ．	-	12

入力文の分割位置を明示するため、“ / ”を挿入した。
BERT 事前学習モデルの BPE によるサブワード分割では、”##”が直前のトークンと連結して一単語となることを示す。
例)「ケ / ##ト / ##アシドーシス」:「ケトアシドーシス」が vocabulary にはないため、「ケ / ト / アシドーシス」で表現される。