一般口演B

[OB14] 一般口演 B

2020年6月6日(土) 15:25 ～ 15:55 第2会場 (Zoom)

[OB14-01] 医学分野に対応した日本語 BERT モデルの構築
Development of a Japanese BERT Model for Medical Domain

*和田 聖哉[1]、蒲原 純[2]、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、2. 大阪大学医学部）

*Shoya Wada[1], Jun Kamohara[2], Toshihiro Takeda[1], Shiro Manabe[1], Shozo Konishi[1], Yasushi Matsumura[1] （1. Department of Medical Informatics, Osaka University Graduate School of Medicine, 2. Faculty of Medicine, Osaka University）

©一般社団法人 日本医療情報学会

医学分野に対応した日本語 BERT モデルの構築

和田 聖哉[*1], 蒲原 純[*2], 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, [*2] 大阪大学医学部

Development of a Japanese BERT Model for Medical Domain

Shoya Wada[*1], Jun Kamohara[*2], Toshihiro Takeda[*1],
Shiro Manabe[*1], Shozo Konishi[*1], Yasushi Matsumura[*1]
*1 Department of Medical Informatics, Osaka University Graduate School of Medicine,
*2 Faculty of Medicine, Osaka University

Bidirectional Encoder Representations from Transformers (BERT)の発表後，様々な事前学習モデルが提案・公開されている．日本語環境では PubMed のように一般利用可能なデータベースが存在しないため，BERT モデルの事前学習に必要なコーパスを医学分野のみで実施することは困難である．本研究では医学分野に対応した日本語 BERT モデル構築へのアプローチとして，(1) Vocabulary の修正，(2) 追加学習の実施，(3) 提案手法による事前学習の比較を行い，医学文書分類タスクでその性能を評価した．その結果，Baseline と比較して(1) 0.91 ポイント，(2) 2.84 ポイント，(3) 5.60 ポイントの精度上昇を確認した．

キーワード 自然言語処理，機械学習，BERT，未定義語

1. はじめに

深層学習による End-to-End 学習の出現により，自然言語処理の精度は飛躍的に向上した．Transformer をベースとした言語モデル (Bidirectional Encoder Representations from Transformers: BERT) [1]は，大規模な生コーパスで事前学習を行い，各タスクで fine-tuning することにより，医学ドメインも含む様々なタスクで state-of-the-art を更新している[2].

日本語環境では，PubMed のように一般利用可能な医学データベースは存在しないため，医学分野に対応した公開 BERT モデルはまだない．日本語 BERT モデルでは Wikipedia を事前学習に用いたモデル [a]が公開されているが，このモデルを医療分野に適用する際には，専門的な表現（「疼痛」「痒疹」など）の扱いに注意する必要がある．

BERT の vocabulary は，out of vocabulary (OOV)を減らすために，Byte Pair Encoding (BPE)による単語以下の token（サブワード）で構成されている．サブワードの導入で低頻度語が表現可能になるが，それでも先に挙げた「疼」，「痒」は日本語 BERT モデルの vocabulary から脱落している．OOV が対象文字列に含まれる場合，BERT モデルはその単語を「[UNK]」（未定義語）として扱う．

我々は，日本語医療分野に対応した BERT モデル構築時の課題を解決するため，(1) Vocabulary の修正，(2) 追加学習の実施，(3) 提案手法による事前学習の比較を行い，その効果を医学文書分類タスクで評価した．.

2. 方法

1) BERT モデルの構築

我々の環境で日本語 BERT モデルを再現し（Baseline），公開モデル（Published）と同等の精度が得られることを確認後，提案モデルと比較した．Published と同様に，設定値は $\text{BERT}_{\text{BASE}}$ と同じ値を使用し，事前学習には日本語 Wikipedia のダンプデータ（20190901, 2.6 GB）を用いた．

2) Vmod の vocabulary

Baseline と Vmod の違いは vocabulary にある．Wikipedia から BPE で 30,000 語を取得し，その vocabulary で今日の診療プレミアム 2017 年版（医学書院）に含まれる 15 冊の医学書本文をトークン化した際に，未定義語となる単語を同定した．そこに BPE を適用して追加する 2,000 語を設定し，vocabulary_size を 32,000 に調整した．

3) 医学文書による事前学習の追加

日本語 Wikipedia による事前学習終了後に医学文書による事前学習を追加したモデルを準備した．医学文書として，今日の診療プレミアム 2017 年版（83 MB）を使用し，60epochs で追加学

[a] https://github.com/cl-tohoku/bert-japanese

習した.

4) 提案手法による事前学習の実施

日本語 Wikipedia のデータサイズに比べて医学文書のデータサイズは大きく乖離しているため，それらをそのまま事前学習に用いても，十分に医学知識を反映させたモデルとはなり得ない．そこで，医学文書を 30 倍に増幅させた上で，日本語 Wikipedia コーパスと混合し，事前学習データに用いた．使用する vocabulary は，この混合コーパスに BPE を適用し，Baseline，Vmod と同じ vocabulary_size で設計した．

5) 精度評価用データセット

MSD マニュアルプロフェッショナル版 [b]から疾患分野毎にページ(対象:2,475 ページ)を取得し，その疾患分野(「消化器疾患」,「小児科」など全 22 種)を分類ラベルとした．層化 5 分割交差検証で分類精度の平均値を比較した．評価用データセットの各クラスに属するデータ数は不均衡であるため，評価指標にはマシューズ相関係数を使用した．

6) 文書分類タスクの fine-tuning

fine-tuning は huggingface の transformers を改良して実装した．入力 token 数の上限値は 128 とした．

3. 結果

結果を Table 1 に示す．Vmod は Baseline と比較し，0.91 ポイント上昇した．医学文書の事前学習追加でその差は 0.38 ポイントまで縮まった．提案手法により構築したモデルはそれらを大きく上回る精度を示した(85.73 ポイント)．

Table.1 医学文書分類の精度比較

	Wikipedia	+Medical Texts
Published	79.63 ±3.54	-
Baseline	80.13 ±2.05	82.97 ±2.62
Vmod	81.04 ±2.60	83.35 ±1.77
Proposed	**85.73 ±1.81**	

(mean ± SD)

[b] https://www.msdmanuals.com/ja-jp/プロフェッショナル

Vmod と Baseline で判定の別れた事例を Fig. 1 に示す．Baseline では「眼瞼」が「[UNK]」と扱われ，モデルは「皮膚疾患」と予測した．一方，Vmod では未定義語は出現せず，予測モデルは正しく「眼疾患」と予測した．

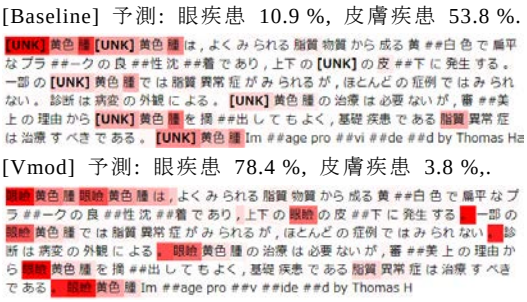

Fig.1 Baseline の誤判定事例(眼瞼黄色腫)
最終層の heads の Self Attention を合計し可視化した．赤色の濃さは attention の強さに相関する．

4. 考察

Baseline と Vmod の比較で，日本語医学分野に対応した BERT モデルには vocabulary の再設計が必要であることを示した．追加学習によりその性能差は縮まるが，Fig.1 で示したような事例の存在を考慮すると，未定義語の出現を減らす意義は少なからずあると言える．最後に，BERT の事前学習に単独で用いるには小さいコーパスでも，提案手法により有用な医学領域対応モデルが構築可能であることを示した．その効果は追加学習による精度上昇よりも大きかった．

5. 結語

日本語医学分野の BERT モデルを構築する上で大きな課題であった事前学習コーパス量について，本手法がその解決の一手となりうることを示した．

参 考 文 献

[1] J Devlin, MW Chang, K Lee, et al: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT 2019, 4171-4186, 2019.

[2] Y Peng, S Yan, Z Lu: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets, Proceedings of the BioNLP 2019 workshop, 58-65, 2019.