E-012

事前学習済み BERT モデルの語彙の追加を伴う伝統医学分野への適用と評価
Application of Pretrained BERT model with vocabulary expansion for the Field of Traditional Medicine

伊藤 元斗 [1]　関 隆志 [2]　高橋 晶子 [1]　力武克彰 [1]
Haruto Ito　Takashi Seki　Akiko Takahashi　Yoshiaki Rikitake

1. はじめに

近年、高齢化が進む中、健康寿命の延伸を図るために、伝統医学特有の疾病の発生を未然に防ぐ概念や、方剤などの治療に関する知見を活用することが期待されている。

伝統医学では、医師の診察から「証」と呼ばれる患者の状態を特定し、診断に用いる。その過程では、病態図の作成など、複数の症状と証の因果関係を適切に結びつける必要がある。

そこで、診断の支援を目的として証同士の因果関係を文献から抽出する手法や、医師の知識を補完するための証の検索システム、およびデータベースが構築されている[1]。

証の検索システムや因果関係の抽出は、機械学習に基づいた自然言語処理技術によって実現されている。普段我々が接している西洋医学に基づく医療では、自然言語処理技術が医療記録の自動解析や診断予測、ヘルスケアチャットボットなどに応用されており、伝統医学に基づく医療でも、同様の応用を実現することが期待される。

2. 研究目的

本研究は、文書の分類や固有表現抽出など、文の解析に特化した自然言語処理技術の BERT[2]に着目し、日本語の伝統医学分野に適用させ、伝統医学に基づく診療の補助やセルフケアに対して応用することを目的としている。

本稿では、BERT を伝統医学分野に適用させるための手法と、手法の比較を目的としたマルチクラス分類問題による性能の評価について述べる。

3. 関連研究

証の検索システムや因果関係の推論は、単語の分散表現(N 次元のベクトル)を得ることを主目的とした Word2Vec や SCDV などの、機械学習に基づいた自然言語処理技術により実現された。証の検索システムでは、入力文および証 DB 内の単語や文のベクトルのコサイン類似度を計算し、類似度の高い順で検索結果を提供しており、分散表現を用いた表記ゆれ吸収を実現している[3]。

近年、機械学習に基づいた自然言語処理技術として、Transformer のアーキテクチャを採用したモデルが注目されており、BERT もこの一種である。大規模なコーパスを用いた事前学習により、英語版や日本語版、多言語版の事前学習済みモデルが作成され、インターネット上で配布されている。また、事前学習済みのモデルに対して、比較的少量なデータを用いたファインチューニングを行うことで、様々なタスクで良い性能が得ることができる。

伝統医学分野では、BERT と Bi-LSTM を組み合わせたモデルによる診断結果の分類の予測などが行われている[4]。

1 仙台高等専門学校
National Institute of Technology, Sendai College
2 フジ虎ノ門整形外科病院
Fuji Toranomon Orthopedic Hospital

4. 研究手法

BERT による文書の分類や文の類似度の計算、因果関係の抽出など、診断支援手法の実装を行うために、伝統医学分野の学習を行った事前学習済みモデルが必要である。

日本語版の Wikipedia などの大規模なコーパスで事前学習を実施した BERT のモデルである”cl-tohoku/bert-base-japanese-v3”[5] (以後東北大版 BERT と記載)に対し、比較的少量の伝統医学文献を用いて追加事前学習を行い、伝統医学分野に適用したモデルを比較的低コストで作成する。

東北大版 BERT に対する追加事前学習を行うにあたり、学習の実施前に未知語の削減を目的としたモデルへの語彙の追加と対応する単語ベクトルの初期化を実施する。

4.1 語彙の追加と単語ベクトルの初期化

伝統医学文献中の専門用語や特殊な漢字は、東北大版 BERT では未知語として処理され、未知語の出現によりファインチューニング後の性能が低下する可能性がある。

そこで、伝統医学の用語辞典から約 3500 の用語を東北大版 BERT の語彙に追加し、未知語に関する問題を解決する。

この際、追加した語彙に対応する単語ベクトルを、何らかの方法で初期化する必要がある[6]。そこで、初期化手法の差異を確認するために、以下の 4 つの手法でそれぞれ初期化を行ったモデルを作成し、追加事前学習を行った。

- ランダムな初期化 **(手法 1)**
- ランダムに初期化したベクトルを正規化 **(手法 2)**
- 「伝統医学」の文ベクトルにランダムなノイズを加えた初期化 **(手法 3)**
- 文献の記述と[MASK]トークンを用いた初期化 **(手法 4)**

4.1.1 手法 3 の手順

① 追加する語彙の各単語ベクトルを手法 1 と同様にランダムに初期化
② 追加した語彙の各単語ベクトルを、ランダムに初期化したベクトルの平均を取ったベクトルで減算
③ 東北大版 BERT に”[CLS]伝統医学[SEP]”の入力を与え出力の各トークンのベクトルの平均を計算し、追加した語彙の各単語ベクトルに加算
④ 各単語ベクトルのノルムを東北大版 BERT の既存の単語ベクトルのノルムの平均値に揃える

4.1.2 手法 4 の手順

BERTRAM[7]から着想を得た、より簡易的な手法である。

① 追加する語彙に関する説明文、もしくは語を含む文について、語に対応する部分を[MASK]トークンに置き換える (例：「[MASK]とは、肺の…」)
② 東北大版 BERT に[MASK]トークンに置き換えた文を入力し、出力のうちの[MASK]トークンに該当する部分の語のベクトルを得る
③ ②で得たベクトルから[MASK]トークンの単語ベクトルを減算して、ベクトルのノルムを東北大版 BERT の既存の単語ベクトルのノルムの平均値に揃える

4.2 追加事前学習による伝統医学文献の学習

追加事前学習[8]は、大規模なコーパスで事前学習済みのモデルに対してさらに事前学習を行う手法で、比較的低コストで伝統医学文献の知識を学習させることが期待できる。

今回は、追加事前学習用のコーパスとして、伝統医学の用語辞書 1 冊の文字起こしテキストデータから作成したコーパス(約 2.4MB)を用いた。

作成したコーパスは、ある用語に関する説明文を 1 文ごとに改行して記載し、異なる用語の説明文同士は空行で区切られているテキストファイルである。用語の文字列が説明文に含まれていない場合を考慮し、説明文の最初に「用語 A に関する説明。用語 A は「ようごえー」と読む。」のような 2 文を追加した。

東北大版 BERT に語彙の追加を行ったモデルに対して、学習率を 5e-5、エポック数を 200、ウォームアップステップを全体のステップ数の 10%に設定し、学習を行った。

5. マルチクラス分類問題による評価

追加事前学習による性能の変化と、語彙の追加の有無、単語ベクトルの初期化手法による性能の変化を確認するために、各モデルに対して分類問題のデータセットでファインチューニングを実施し、その性能を比較する。

使用するモデルは、語彙の追加を行い、単語ベクトルを各手法で初期化して追加事前学習を実施し作成した 4 つのモデルに、ベースとなった**東北大版 BERT** と語彙の追加を行わずに追加事前学習のみを実施した**語彙の追加なし**のモデルを加えた計 6 つのモデルを用いる。

分類問題のデータセットは、livedoor ニュースコーパス[9]から作成した分類問題のデータセットと、学習に用いた伝統医学文献の分類索引を元に作成した、伝統医学(基本)と伝統医学(臨床)の分類問題のデータセットを用いる。

通常の日本語文に対する性能を測る目的で、livedoor ニュースコーパスから作成したデータセットを用い、伝統医学文献中の日本語文に対する性能を測る目的で、伝統医学文献から作成した 2 つの分類問題のデータセットを用いる。

データセットを学習用:評価用の 8:2 に分割し、学習率を 5e-5、3 エポックの学習を行った後、残り 2 割の評価用データで式 1 のマクロ平均 F1 スコアを計算する。

各モデルと各データセットの組み合わせごとに 10 回ずつファインチューニングに実施し、それぞれの試行で算出したマクロ平均 F1 スコアを平均した結果を表 1 に示す。

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{n} \times \sum_{i=1}^{n} \mathrm{F1}_i \qquad (1)$$

$$\mathrm{F1}_i = 2 \times (\mathrm{R}_i \times \mathrm{P}_i) \div (\mathrm{R}_i + \mathrm{P}_i)$$

$$\mathrm{R}_i = \mathrm{TP}_i \div (\mathrm{TP}_i + \mathrm{FN}_i), \qquad \mathrm{P}_i = \mathrm{TP}_i \div (\mathrm{TP}_i + \mathrm{FP}_i)$$

TP_i = ラベル i のデータのうち正しくラベル i と予測した数
FP_i = ラベル i 以外のデータのうち間違えてラベル i と予測した数
FN_i = ラベル i のデータのうち間違えてラベル i 以外と予測した数
n = ラベルの種類の数(8 クラス分類問題であれば 8)

6. 考察

追加事前学習により、通常の日本語文の分類タスクに対する性能が低下し、伝統医学文献の日本語文の分類タスクに対する性能が向上した。また、語彙の追加の有無や単語ベクトルの初期化方法により、若干の性能差が生じた。

伝統医学(基本)は、伝統医学の基本的な事項に関する説明文が、どのトピック(概念、証、理論など)なのかを分類する問題である。語彙の追加なしのモデルの性能が一番高く、伝統医学分野の用語の正確な認識と理解が影響しない通常の日本語文の範疇にある説明文が若干多く含まれていることと、語彙の追加を行ったモデルに比べ、モデルの全体の重みを伝統医学文献に対して最適化できたことが理由として考えられる。

伝統医学(臨床)は、伝統医学の応用に関する説明文が、どの科に対応するのかを分類する問題である。用語の正確な認識と理解が必要で、語彙の追加を行ったモデルの方が性能向上する傾向にあると考えられる。

7. まとめ

約 2.4MB の伝統医学文献を用いた追加事前学習では伝統医学分野の性能と通常の日本語文に対する性能がトレードオフの関係になることが示された。語彙の追加の有無や単語ベクトルの初期化手法による若干の性能の変化については、他の伝統医学文献を元に作成したタスクや、単語ベクトルの初期化手法を用いた調査が必要であると考えられる。

謝辞

本研究は JSPS 科研費 JP23K11344 の助成を受けたものです。

参考文献

[1] Nakagawa R, et al., "Design of a Diagnostic Support Method Utilizing Interrogation Information in Traditional Chinese Medicine", Advanced Information Networking and Applications - Proceedings of the 35th International Conference on Advanced Information Networking and Applications AINA 2021, pp.177-187, (2021)

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanov, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp.4171-4186, (2019)

[3] 太田遥人, 関隆志, 高橋晶子, 力武克彰, “中医学のための単語埋め込みに基づく情報検索システムの研究”, 情報処理学会 第 84 回全国大会講演論文集 2022(1), pp.747-748, (2022)

[4] C. Xu, F. Yuan, S. Chen, "Research on Assistant Diagnostic Method of TCM Based on BERT", 2021 11th International Conference on Information Technology in Medicine and Education, pp.282-286, (2021)

[5] Hugging Face, "cl-tohoku/bert-base-japanese-v3 - Hugging Face", URL: https://huggingface.co/cl-tohoku/bert-base-japanese-v3, 最終更新日: 2023/5

[6] John Hewitt, “Initializing New Word Embeddings for Pretrained Language Models”, URL: https:/nlp.stanford.edu/~johnhew//vocab-expansion.html, (2021)

[7] Timo Schick, Hinrich Schütze, "BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3996–4007, (2020)

[8] 鈴木雅弘, 坂地泰紀, 和泉潔, 石川康, “金融文書を用いた追加事前学習言語モデルの構築と検証”, 言語処理学会第 28 回年次大会発表論文集, pp588-592, (2022).

[9] 株式会社ロンウイット, “livedoor ニュースコーパス”, URL: http://www.rondhuit.com/download.html#ldcc, 最終閲覧日:2023/5

表 1 . 各モデルのデータセットごとマクロ平均 F1 スコアの平均値の一覧

データセットの種類	東北大版 BERT	語彙の追加なし	手法 1	手法 2	手法 3	手法 4
livedoor ニュースコーパス	**0.882**	0.873	0.870	0.872	0.865	0.870
伝統医学(基本)	0.743	**0.793**	0.768	0.771	0.784	0.784
伝統医学(臨床)	0.843	0.862	0.870	0.861	**0.885**	0.879