言語モデルを利用した疾患症状表現の教師なし固有表現認識

和田 聖哉[*1], 武田 理宏[*1], 真鍋 史朗[*1], 小西 正三[*1], 岡田 佳築[*2], 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] 大阪大学大学院医学系研究科 ゲノム情報学共同研究講座

Unsupervised Named Entity Recognition for Disease Symptom Expression with Language Model

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

深層学習は飛躍的な精度向上を自然言語処理にもたらしたが，その達成にはまず高品質な教師データを十分量準備する必要がある．その作成コストは多大であるが，人手で作成する前にある程度の情報を付与すること(プレアノテーション)は，データ作成者の負担軽減に繋がる．我々は，言語モデルの事前学習で行われる Masked language modeling に着目し，MASK 化された token の予測候補はプレアノテーションに有用な情報を持つと考え，疾患名・症状表現の固有表現認識で検証した．その結果，予測候補 top10 で提案手法は precision 60.1 %，recall 89.0 %を示し，固有表現辞書に登録されていない表現へのプレアノテーションに有用であることを確認した．

キーワード 自然言語処理，機械学習，BERT，プレアノテーション

1. はじめに

深層学習の導入後，自然言語処理の精度は飛躍的に向上し，特に Transformer をベースとした言語モデル[1]の登場により，高精度な目的タスクの解決が実現した．

高性能な分類モデルを構築するには高品質な教師データを十分量準備することが重要であるが，その作成には高いコストを要する．例えば入力文から対象文字列を抽出する固有表現認識タスクにおいては，「1．文字列の選択，2．分類クラスの付与」という作業が要求される．この労力を軽減するために，予め所有している固有表現辞書とのテキスト照合を行い，人手作業前に情報を付与する手法(プレアノテーション)を適用することも考えられるが，辞書の拡充自体と拡充後の計算負荷増大と，それぞれに課題がある．

言語モデル BERT の事前学習では，入力文中の token をランダムにマスク化し，それを周囲の token から予測する課題 (Masked language modeling: MLM)の最適化が行われる．我々はこの token の予測候補を利用することで，プレアノテーションに有用な情報を，パラメータチューニングを行わずに取得出来るのではないかと考えた．そこで本研究では，疾患名や症状表現に関する固有表現を対象に，予め構築した日本語医学 BERT モデルによる教師なし固有表現認識の可能性について検証する．

2. 方法

1) 教師なし固有表現認識

固有表現認識は，入力文中から固有表現を抽出し，それを予め定義された固有表現分類に分類するタスクと定義される．一般的に言語モデルを用いた固有表現認識では，入力文に含まれる各 token に対してそれぞれの分類ラベルが付与された情報を教師データとして用い，パラメータチューニングを行うことで課題を解く(教師あり学習)．本研究ではパラメータチューニングを行わずに事前学習モデルの出力をそのまま利用することを，「教師なし固有表現認識」と定義した．

2) 提案手法

1. 入力文を GiNZA [2]で処理し，名詞(複合名詞含む)を特定する．
2. 特定した名詞を[MASK] token に置換し一文中に一つの[MASK]が存在するように入力インスタンスを生成する．すなわち，一文中に複数の名詞が存在する場合は，各々が個別に

MASK 化されたインスタンスを生成する.

3. 入力インスタンスを事前学習モデルに入力し, [MASK] token の予測候補を上位 N 個 (N=1, 2, …, 10) まで取得する.
4. 予測 token が疾患名・症状表現かどうかを判定する. この判定では, ICD-10 対応標準病名マスタ (Ver.5.05)から構築した固有表現辞書との照合を行う. topN のうち一つでもマッチすれば, MASK 化された token に該当する表現は疾患名・症状表現であると判定する.

3) 実験設定

(1) 事前学習済み日本語医学 BERT モデル

以前我々の提案した手法 [3]を用いて, 日本語 Wikipedia と今日の診療プレミアム(医学書院)から構築したモデルを準備した. 日本語形態素解析器には, GiNZA に組み込まれている SudachiPy を採用した.

(2) 評価データ

日本語 Wikipedia で ICD-10 コードが与えられている記事から 50 記事をサンプリングした. 各記事の文章を GiNZA で処理し, 品詞タグ情報から複合名詞を含めた名詞を全対象事例とした (6,479 例). その表現が疾患名もしくは症状表現か否かを目視確認し, 評価データを作成した(正例:負例=1,843: 4,636).

(3) 評価指標

提案手法の精度評価には, 適合率 precision, 再現率 recall とそれらの調和平均 F1-score を用いた. [MASK] token の予測候補採用数が精度に与える影響を確認するために, 候補数 N を 1 から 10 まで変化させて各指標を評価した. 提案手法の比較対象には, 固有表現辞書に収録されている表現との完全一致を用いた. また, これら 2 つの手法を併用した時の精度も確認した.

3. 結果

評価データを用いた提案手法の精度を Figure 1 に示す. 固有表現辞書の precision は 98.6 %, recall は 55.2 %であった. 一方提案手法において top10 を採用した場合, precision は 60.1 %まで低下するが, recall は 89.0 %まで上昇した. これら 2 つを組み合わせると, recall は 93.5 %まで上昇した.

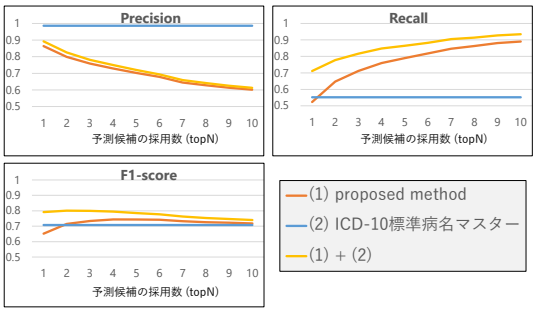

Fig. 1 提案手法の精度

4. 考察

教師データによるパラメータチューニングを行わずに, 事前学習モデルの出力を利用した固有表現認識の精度を確認した. 固有表現辞書による表現同定は正確だが, 見落としが多い. その一方で, 提案手法は高い recall を有しており, 未知の固有表現でも有効に機能することが期待される. さらに, 既知の表現を特定できる固有表現辞書を用いた手法と提案手法を組み合わせると, 精度の上乗せ効果も確認できた.

5. 結語

提案手法による疾患名・症状表現の固有表現認識は, 標準病名マスタに収録されていない表現も高精度に回収できた。回収した表現の正確度は不十分であるが, それに人手判定を組み合わせて固有表現辞書の拡充を並行して行うことで、全体の精度向上や教師あり固有表現認識への展開が期待出来る。

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] 松田寛, 大村舞, 浅原正幸: 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習, 言語処理学会第 25 回年次大会講演論文集, F2-3, 2019.

[3] Wada S, Takeda T, Manabe S, et al.: A pre-training technique to localize medical BERT and enhance biomedical BERT, arXiv preprint, arXiv:2005.07202, 2020.