双方向 Transformer 言語モデルを用いた CT 画像報告書のテキスト情報からの入院予測

橋本 正弘[*1,2]，泉 啓介[*2,3]，洪 繁[*2,4]，陣崎 雅弘[*1,2]
*1.慶應義塾大学医学部放射線科学（診断），*2.慶應義塾大学メディカル AI センター，
*3.慶應義塾大学医学部内科学（リウマチ・膠原病），
*4.慶應義塾大学医学部坂口光洋記念講座（システム医学）

Predicting hospitalization from textual information of CT image report using bidirectional transformers for language understanding model

Masahiro Hashimoto[*1,2], Keisuke Izumi[*2,3], Shigeru Ko[*2,4], Masahiro Jinzaki[*1,2]
*1 Department of Radiology, Keio University School of Medicine
*2 Keio Medical AI Center, Keio University
*3 Division of Rheumatology, Department of Internal Medicine, Keio University School of Medicine
*4 Department of Systems Medicine, Keio University School of Medicine

抄録：CT.画像報告書の未読報告書に対処する優先順位を決定するため、画像報告書の所見文章から検査後 90 日以内の入院を予測する機械学習を行った。日本語 Wikipedia の文章で事前学習した Bidirectional Transformers for Language Understanding（BERT）モデルを転移学習させる事で、従来の形態素解析、TF-IDF によるベクトル化、gradient boosting を用いる方法より精度を向上させることができた。非医療文章で医療用専門用語の頻度が低い日本語 wikipeida の文章を用いて事前学習した BERT モデルを転移学習することで医療記録である CT 画像報告書にも応用できることを示した。

キーワード 深層学習、転移学習、自然言語処理、画像報告書

1. はじめに

当院では画像報告書の未読・既読管理を行っている。しかしながら、多くの報告書は次回外来まで未読のままであり、管理対象となる未読報告書が膨大で、対処を要する画像報告書の全てに迅速に対処できているとは言えない。そこで、画像報告書に対する優先順位を決定するための指標を作成するため、機械学習を用いて画像報告書の文章から、検査後の入院の有無を予測するモデルの構築を検討した。

Bidirectional Transformers for Language Understanding（BERT）は双方向の Transformer で言語モデルを学習することで汎用性を獲得し、様々な自然言語タスクにおいて精度が向上したと報告されている[1]。そこで、本研究では BERT と従来手法の比較を行った。

2. 方法

1) 対象データ

倫理委員会の承認を得て、2012 年 1 月 1 日から 2018 年 3 月 31 日までに慶應義塾大学病院で撮影された健康診断以外の外来の CT 画像報告書の所見欄文章、2012 年 1 月 1 日から 2018 年 6 月 30 日の入院オーダーを対象とした。なお、医療記録の研究利用に不同意を表明している患者、並びに医療記録へのアクセス制限が設定されている患者のデータは除外した。

2) 前処理

CT 画像報告書 195,836 件をランダムに学習用 146,874 件、バリデーション用 24,478 件、検証用 24,484 件に分割した。CT 撮影後に入院オーダーが出され、90 日以内に入院した「入院あり」とそれ以外「入院なし」を教師データとした。学習用、バリデーション用、検証用データの入院あり/入院なしはそれぞれ 16,690 / 130,184 件、2,797 / 21,681 件、2,690 / 21,794 件であった。不均衡データであったため、学習用データは入院ありを over sampling し、検証用データでは入院なしを under sampling を行い、概ね同数となるように調整した。

3) 環境

Intel Xeon 2.2GHz × 2, 128 GB memory,

NVIDIA Tesla P100 16G×4, Ubuntu 16.04 LTS, Docker17.06.1 , nvidia driver384.66 , nvidia-docker1.0.1-1, Python 3.7.1 TensorFlow 1.13, Scikit-learn 0.20, MeCab 0.996

4) 機械学習手法

(1) BERT

日本語 wikipedia で事前学習して公開されている model[2]をもとに、学習用データを用いて転移学習を実施した。日本語テキストの分かち書きは事前学習と同様に unigram language model[3]の実装である SentencePiece[4]を利用し、vocabulary size は 32,000 を用いた。

(2) 従来手法

画像報告書の所見欄文章を MeCab で形態素解析して分かち書きを行い、TF-IDF によるベクトル化を行った。ベクトル長は 2,000 を用いた。機械学習として gradient boosting を用い、学習用およびバリデーション用データを用いて学習を行い、検証用データを用いて検証した。

3. 結果

Table.1 BERT モデルの結果

BERT	入院ありと予測	入院なしと予測
入院あり	1904	786
入院なし	631	2059

Precision 0.75、recall 0.71、f-value 0.73

Table.2 従来モデルの結果

従来手法	入院ありと予測	入院なしと予測
入院あり	1710	980
入院なし	720	1970

Precision 0.70、recall 0.64、f-value 0.67

4. 考察

日本語 Wikipedia で事前学習した BERT モデルを転移学習させることで、CT 画像報告書の所見文章から 90 日以内の入院を予測する精度が従来モデルより向上した。

BERT は様々な自然言語タスクにおいて高い精度が得られると報告されている[1]。しかしながら、BERT の学習を行うために必要な文章数が膨大で、計算コストも高いことから、これまでの報告は公開コーパスを対象としたものに限られており、日本語の文章を対象とした報告や、医療記録の文章を対象とした報告もなかった。

本研究では非医療文章である日本語 wikipeida を用いて事前学習した BERT モデルを画像報告書にも応用できることを示した。今後、BERT を比較的文章量が限られる他の医療記録の文章にも応用できる可能性を示すことができた。

5. 結語

日本語 Wikipedia で事前学習した BERT モデルを fine tuning する事で、画像報告書からの入院の予測精度を従来手法より向上させることができる。

6. 謝辞

本研究は、AMED の課題番号 JP18lk1010025s0701 の支援を受けて実施された。

参考文献

[1]Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

[2]Yohei Kikuta, BERT Pretrained model Trained On Japanese Wikipedia Articles, GitHub, GitHub repository: https://github.com/yoheikikuta/bert-japanese (accessed on 6 Feb.2019)

[3]Taku Kudo: Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, https://arxiv.org/abs/1804.10959

[4]https://github.com/google/sentencepiece (accessed on 6.Feb.2019)