BERTを用いた「要注意」タグ付き読影レポートの検出

[1] 中村 優太　[2] 花岡 昇平　[3] 野村 行弘　[1] 中尾 貴祐　[3] 三木 聡一郎
[1,2] 渡谷 岳行　[2] 吉川 健啓　[3] 林 直人　[1,2] 阿部 修

[1] 東京大学医学系研究科生体物理医学専攻
[2] 東京大学医学部附属病院 放射線科
[3] 東京大学医学部放射線医学教室 コンピュータ画像診断学／予防医学講座

{yutanakamura-tky, hanaoka-tky, nomuray-tky, tanakao-tky, smiki-tky, watadat-tky, takeharu-yoshikawa, naoto-tky, abediag-tky}@umin.ac.jp

1 はじめに

読影レポートとは, 単純 X 線, CT[1], MRI[2], 核医学などの画像検査の結果を記載した文書であり, 画像診断に従事する医師 (以下, 読影医) によって作成される. 近年, 重篤な画像所見が偶然発見されたことが読影レポート内で指摘されたにも関わらず, それが外来や病棟などで患者の診療にあたっている医師 (以下, 依頼医) に認識されないまま見過ごされ, 結果的に必要な追加検査や治療などが遅れるという, いわゆる読影レポート未読問題に対する関心が高まっている.

東京大学医学部附属病院 (以下，当院) では, 読影レポート未読問題対策の一環として, 医療安全部門の主導のもと, レポーティングシステムに「要注意」タグ付与機能を実装し, 2019 年 9 月 9 日より運用を開始した. これは, 読影レポート作成時に, 依頼医が想定していないと思われ, かつ放置すると患者の予後に重大な影響を及ぼす可能性がある所見を認めた際に, レポーティングシステム上で当該読影レポートに対して「要注意」タグを付与するというものである.

本研究の目的は, 上記対策の一助とするため, BERT (Bidirectional Encoder Representations from Transformers)[1] を用いて「要注意」タグ付きの読影レポート (以下, 要注意レポート) の自動検出を試みることである. 注意が必要な読影レポートの自動検出を試みた先行研究はこれまでに複数報告されている. しかし, 多くは検出対象を (1) 特定の画像所見 [2, 3, 4], (2) 直接的な注意喚起表現 [5, 6], (3) 依頼医への推奨内容の一部 [5, 6] などに限定しており, 読影医からの多様な注意喚起を包括的に扱えていない可能性がある. また, その他の先行研究 [7] を含めても, 画像所見が依頼医にとって想定外であるかどうかを考慮したものや, 実際の画像診断業務の中で付与されたタグを教師データとするものは, 我々の知り得た範囲では存在しない.

[1]コンピュータ断層撮影 (Computed Tomography).
[2]核磁気共鳴画像法 (Magnetic Resonance Imaging).

2 読影レポートおよび当院における未読問題対策について

読影レポートは主に以下の内容から構成される.

依頼状　臨床病名, 検査目的
本文　所見, 診断

このうち, 読影医が実際に記載するのは本文にあたる所見と診断のみである. 臨床病名と検査目的は画像診断の参考のために提示されるものであり, 依頼医が画像検査依頼時に入力したものがそのまま反映される.

所見欄では原則として, 撮像範囲内のすべての部位について異常所見の有無を判定し, 存在する異常所見を描写し, 想定される病態や疾患の候補を提示する. 追加すべき検査, 手術適応, 治療効果判定などについて情報提供を行う場合もある. 診断欄には依頼医の要求や医学的状況などを総合的に踏まえた要点を記載する. ただし, 実際には読影レポートの体裁は読影医の自由な裁量に任されており, 所見欄と診断欄の使い分けも流動的となりうる.

当院のレポーティングシステムにて要注意レポートが作成されると, 依頼医に診療端末上でアラートが送信されるほか, 読影レポートの閲覧時に要注意レポートである旨が強調表示される. また同時に, 放射線部門から病院へと情報提供がなされ, 病院から依頼科に対応を促す際などに利用される. 運用開始から 2020 年 1 月までの約 3 ヶ月間に作成された読影レポートのうち, 約 0.5%が要注意レポートとなっている.

表 1: 要注意レポートの例 (一部を抜粋のうえ改変. 偶発的所見にあたる内容を下線で示す).

例 1: 直接的な注意喚起があるもの (**太字部**).

【臨床病名】 人工肛門造設後
【検査目的】 人工肛門閉鎖術前につき精査目的です。thin slice も撮影お願いします
【所見】 直腸癌術後。局所再発を認めません。有意なリンパ節腫大を認めません。肝転移、肺転移は指摘できません。右乳腺 C 領域に造影効果を伴う 18mm 大の腫瘤性病変が疑われます。**未知の病変でしたら超音波検査等での精査が望まれます。**
【診断】 直腸癌術後。右乳腺腫瘤疑い。**未知の病変でしたら超音波検査等での精査が望まれます。**

例 2: 直接的な注意喚起がないもの.

【臨床病名】 肺癌疑い
【検査目的】 頭部精査目的
【所見】 前交通動脈と連続し前上方に突出する 8mm 大の動脈瘤が疑われる。両内頚動脈サイフォン部に石灰化が見られる。脳室拡大なし。両側水晶体術後。副鼻腔や乳突蜂巣の含気は保たれている。
【診断】 前交通動脈に嚢状瘤、8mm 大。

表 2: 要注意レポートの注意喚起表現.

	全体[3]	test set[3]
注意喚起表現あり	57	13
+精査を推奨	45	10
+経過観察を推奨	12	3
+治療を推奨	2	0
注意喚起表現なし	39	4
計	96	17

表 3: 要注意レポートの偶発的所見.

	全体[3]	test set[3]
結節, 腫瘍	69	15
出血	5	0
血栓	5	0
消化管穿孔	4	2
動脈瘤	4	0
ほか	12	0

表 4: Baseline の検索語候補.

注意, 慎重, 懸念, 危惧, 早急, 再検, 経過, follow, フォロー, 評価, 確認, 精査, 検索, 検討, 否定, 除外, コンサルト, いかが, お願い, 望ま[4]

なお, 注意喚起には高度な医学的判断を要するため, 要注意レポートとするか否かの基準は個々の読影医に委ねられており, 現時点では統一的な基準が設定される予定はない.

3 課題設定および手法

3.1 課題設定

Out-of-vocabulary に関する検討

汎用的な事前学習済みモデルを語彙を変更せずに読影レポートへと適用した際に, 未知語として扱われる語の割合とその種類を調査した.

要注意レポートの検出

「要注意」タグの運用開始以降に作成された読影レポートに対し, 要注意レポートであるか否かをその記載内容のみから 2 クラス分類によって推測した.

3.2 データセット

2019 年 9 月 9 日から 12 月 16 日までの間に当院で施行された CT 検査に対する読影レポート 15,750 件を対象とした. このうち要注意レポートは 96 件 (0.61%) であった. 要注意レポートは, 表 1-3 に示すように, 直接的な注意喚起の有無や, 依頼医への推奨内容, および対象とする偶発的所見の種類においてさまざまであった.

データセットはクラス比を保持したまま 8:2 に分割し, それぞれ train set,test set とした. 不均衡性を緩和するため, train set, test set の両者において要注意レポートのみを 10 倍に oversampling した.

3.3 事前学習済み BERT モデルの fine-tuning

BERT の英語版モデルには BioBERT[8] や ClinicalBERT[9] などの生物医学領域に特化したものが存在するが, 日本語版については本記事執筆時点ではそのようなモデルは存在しない. このため本研究では, 日本語版 Wikipedia 記事および SentencePiece[10] によって事前学習済みの BERT モデル [11] を使用した.

この BERT モデルが [CLS] トークンに対して出力する特徴量ベクトルを, softmax 関数を活性化関数とする単層パーセプトロンを用いて 2 次元ベクトルに変換することにより分類器を構成した.

以下のように, 学習時および推論時の入力内容を変えた 2 通りの fine-tuning を行なった.

[3]Oversampling 前のもの.
[4]"望まれる", "望ましい"などとの部分一致を目的としている.

- **BERT**$_{\mathrm{R}}$: 所見と診断をつなげて一続きにした本文を入力に用いる.
- **BERT**$_{\mathrm{O+R}}$: 依頼状を 1 文目, 本文を 2 文目とする Sequence pair を入力に用いる.

上記内容を事前学習と同一の SentencePiece モデルによって分かち書きし, 先頭 512 トークンを使用した.

Fine-tuning は BERT モデルの全 Encoder 層と後続の Pooler およびパーセプトロンの重み行列に対して行い, それ以外の重み行列は固定した.

Batch size は 4, optimizer は学習率を 5×10^{-7} に固定した Adam[12], epoch 数は 3 とし, loss function には binary cross-entropy を用いた. ハイパーパラメータは train set を用いた cross-validation による探索を通じて決定した.

学習環境は以下の通りである: Intel Core 3.6GHz, 64 GB memory, Ubuntu 16.04 LTS, NVIDIA GeForce RTX 2080 Ti 11GB x1, PyTorch 1.3.1, Torchtext 0.4.0, Scikit-learn[13] 0.20.2, Transformers 2.2.1, SentencePiece 0.1.83.

3.4 Baseline 手法

一部の先行研究 [5] では, 追加検査などを促す表現を部分一致によって検索している. また, 一般的なレポーティングシステムも, 当院採用のものも含め, 部分一致による文書検索機能を提供している場合がある.

これを踏まえ, 読影医が注意を促す際によく使用されると思われる 20 語を検索語の候補とし (表 4), このうち n 語 ($n = 1, 2, 3, 4, 5$) を用いて読影レポート本文に対して部分一致による OR 検索を行った際に検出性能が最大となる検索語の組合せを探索した.

3.5 評価指標

評価指標には Area Under the ROC curve (AUROC) を用いることとし,Recall,Precision,F1 score も算出した. なお, baseline 手法に対する AUROC は, 検索語のうち 1 つでも部分一致すればスコア 1, そうでなければスコア 0 として算出した.

4 結果と考察

4.1 Out-of-vocabulary に関する検討

Oversampling 前の全読影レポートの本文を SentencePiece によって分かち書きすると, 未知語として扱われたトークンは全体の 1.87%であった. 汎用的な事前学習済みモデルの語彙を変更せずとも, 一部の語を除いては, 概ね問題なく認識されたと考えられる. 未知語は 495 語あり, このうち英字 1 字のものを除くほとんどは医学用語に関係していた. 表 5 に英字 2 字以上の未知語 (415 語), 英字以外のみからなる未知語 (56 語) の一部をそれぞれ示す.

表 5: 読影レポート中の未知語の例.

医学用語などの略称
MRI, RFA, ESD, IPMN, DVT, PET, PE, TA, LAD, VP, RR, GIST, SMA, EUS, FDG, RT, US, DLB, HR, EVAR, SMV, STA, ABG, GGN, LMT, HGS, ...
医学用語またはその部分文字列
盂, 彎, 鬆, 癪, 扼, 辷, 瀰, 彎, 癆, 嚥, 腟, 胤, 簇, 腓, 篩, 縦, 蝸, 褥, 菲, 痔癆, 痰, 蠕, 疸, 羸, 穹窿, 噌, 睾, 鈎, 潅, 喀痰, 癆, 癇, 橈, 渣, 痔, 喀, 穹, 疣, 悸, 皺, 撓, 夾, 嗽, 嗄, 殻, 蔟, 浣, 欝, 頤, 嗟, 踝, 齲, 紮, 惻, ...

4.2 要注意レポートの検出

表 6, 図 1 に示すように,baseline 手法では検索語 5 語を用いたとき AUROC が最大で 0.8582 に達した. このとき F1 score は 0.5963 であった. BERT を用いた場合, 読影レポート本文を入力とした場合に baseline よりも高い AUROC が得られ, Recall も上昇した. 一方, False positive も増加し, これにより F1 score は baseline より低下した. また, 依頼状と本文の Sequence pair を BERT に入力した場合は AUROC, F1 score ともに baseline より低値となった. 依頼状と本文の関係が適切に学習されていなかった可能性がある.

表 7 に示すように,BERT では直接的な注意喚起を行なっていない要注意レポートも検出可能であった. また, ごく少数のデータからの推測ではあるが, 偶発的所見の種類を問わず検出可能であったと思われる. これらは, 対象となる表現や所見の種類を限定した先行研究に対する優位性と考えられる.

本研究の限界は主に 3 点ある. 1 点目はデータの少なさである. 特に, 不均衡性も相まって test set 内の要注意レポートは一層少なく, その多様性は現時点では十分に網羅されていない可能性がある. 2 点目は, 画像所見が依頼医にとって予想外であるかどうかの判別が困難だったことである. 重篤な所見を指摘している読影レポートのうち, 依頼医の想定外のものとみなされなかったために「要注意」タグが付与されなかったものが,BERT によって要注意レポートと誤分類されていることがあり, 偽陽性率の上昇の一因となっていた.

表 6: 要注意レポートの検出結果.

手法	検索語 (AUROC を最大化する組合せ)	AUROC	F1 score	Recall	Precision
Baseline	1 語 (精査)	0.6437	0.4184	0.2941	**0.7246**
	2 語 (精査, 経過)	0.7471	0.5014	0.5294	0.4762
	3 語 (精査, 経過, 検討)	0.8037	0.5612	0.6471	0.4955
	4 語 (精査, 経過, 検討, いかが)	0.8325	0.5926	0.7059	0.5106
	5 語 (精査, 経過, 検討, いかが, 確認)	0.8582	**0.5963**	0.7647	0.4887
$BERT_R$		**0.9107**	0.3731	**0.8824**	0.2366
$BERT_{O+R}$		0.8493	0.4396	0.7059	0.3191

図 1: 要注意レポート検出の ROC 曲線.

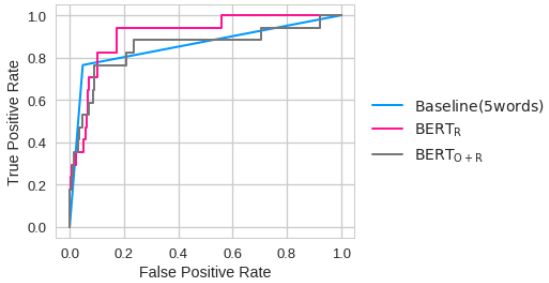

表 7: 要注意レポートの特徴ごとの Recall.

	Baseline(5 語)	$BERT_R$	$BERT_{O+R}$
注意喚起表現あり	**1.00**	0.92	0.85
＋精査を推奨	1.00	1.00	0.90
＋経過観察を推奨	1.00	0.67	0.67
注意喚起表現なし	0.00	**1.00**	0.50
結節, 腫瘍	0.87	**0.93**	0.87
消化管穿孔	0.00	**1.00**	0.00
全体	0.76	**0.94**	0.76

依頼状と本文を Sequence pair として入力する試みも奏功したとはいえず, さらなる手法の検討が必要である. 3 点目は評価指標や loss function の適切さが定かでないことである. 現時点では「要注意」タグの運用開始から日が浅いため,Recall,Precision のどちらを重視すべきかについての知見は十分でなく, 今後の運用を通じて検討していく必要がある.

5 おわりに

事前学習済み BERT を用いることにより,「要注意」タグの付与された読影レポートの検出を行った. 読影レポート未読問題へのよりよい対処のため, さらなるデータの蓄積が待たれるとともに, 読影レポート本文にとどまらない様々な情報の活用が期待される.

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[2] R. Lou, D. Lalevic, C. Chambers, H. M. Zafar, and T. S. Cook. Automated Detection of Radiology Reports that Require Follow-up Imaging Using Natural Language Processing Feature Engineering and Machine Learning Classification. *J Digit Imaging*, Sep 2019.

[3] Matthew C. Chen, Robyn L. Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E. Chapman, David B. Larson, Curtis P. Langlotz, Timothy J. Amrhein, and Matthew P. Lungren. Deep learning to classify radiology free-text reports. *Radiology*, Vol. 286, No. 3, pp. 845–852, 2018. PMID: 29135365.

[4] 今井健, 荒牧英治, 梶野正幸, 美代賢吾, 大江和彦. 医学用語属性と構文情報を用いた診断報告書からの重要所見情報の抽出. 言語処理学会 第 12 回年次大会, pp. 89–92, 2006.

[5] T. Mabotuwana, C. S. Hall, S. Dalal, J. Tieder, and M. L. Gunn. Extracting Follow-Up Recommendations and Associated Anatomy from Radiology Reports. *Stud Health Technol Inform*, Vol. 245, pp. 1090–1094, 2017.

[6] E. Carrodeguas, R. Lacson, W. Swanson, and R. Khorasani. Use of Machine Learning to Identify Follow-Up Recommendations in Radiology Reports. *J Am Coll Radiol*, Vol. 16, No. 3, pp. 336–343, Mar 2019.

[7] X. Meng, M. V. Heinz, C. H. Ganoe, R. T. Sieberg, Y. Y. Cheung, and S. Hassanpour. Understanding Urgency in Radiology Reporting: Identifying Associations Between Clinical Findings in Radiology Reports and Their Prompt Communication to Referring Physicians. *Stud Health Technol Inform*, Vol. 264, pp. 1546–1547, Aug 2019.

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. 2019.

[9] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019.

[10] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018.

[11] Yohei Kikuta. Bert pretrained model trained on japanese wikipedia articles. https://github.com/yoheikikuta/bert-japanese, 2019.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.