放射線レポートの確信度 Scale の分類手法の開発

杉本 賢人, 和田 聖哉, 小西 正三, 岡田 佳築, 真鍋 史朗, 武田 理宏, 松村 泰志
大阪大学大学院医学系研究科 医療情報学

Classification of Certainty Scale of Radiology Reports

Kento Sugimoto, Shoya Wada, Shozo Konishi, Katsuki Okada,
Shiro Manabe, Toshihiro Takeda, Yasushi Matsumura
Department of Medical Informatics,
Osaka University Graduate School of Medicine

フリーテキストで記述された放射線レポートから必要な情報を抽出して，構造化データに変換することで，臨床研究や診断支援システムなどのデータソースとして活用できる．我々は，これまでに放射線レポートに記載されている重要な情報について，機械学習を用いて抽出する研究を進めてきた．本研究では，機械学習で抽出した情報のうち，観察物や臨床所見に関する情報を対象として，その確信度の scale を分類する手法について検討する．実験では，観察物や臨床所見の確信度を事前に定義した 5 段階に分類し，機械学習による分類モデルを構築した．評価指標として，厳密な一致のみを許容する「strict」な基準，正解の基準を緩めた「relaxed」な設定の 2 つで精度を評価した．実験では，F1-score が strict：97.33%，relaxed：98.49%を達成した．

キーワード 自然言語処理，放射線レポート，機械学習

1. はじめに

放射線レポートは，放射線医が記述した診断における重要な情報が記述されている．これらの情報は，観察研究や診断支援システムなど様々な分野での活用が期待されているが，フリーテキスト形式で記述されているため，利用が難しい．我々は，先行研究で放射線レポートから必要な情報を抽出するための情報モデルを提案し，その情報モデルに従った用語をレポート内から認識する機械学習手法における実験において高精度で対象用語を抽出できたこと示した[1]．また，抽出した用語同士の関係の有無を分類するための機械学習手法を提案し，高い分類性能を持つ機械学習モデルを構築きていることを示した[2]．

放射線レポートには，読影医が確信を持って観察物や臨床所見の有無を記述している場合や特定の疾患の可能性はあるが，断定はできないような書き方など表現は様々である．例えば，レポート上は，同じ「肺がん」という記述でも，実際の読影画像は heterogeneous であり，「肺がん」の有無を判別するだけでは，読影医がレポートに記述した情報量の多くを落としていることになる．また，依頼医にとって，読影医がどの程度記述した臨床所見に確信を持っているかをレポートから判断することはしばしば難しいことが知られている[3]．

本研究では，まず，先行研究の手法[1]を用いて，放射線レポートに記載された観察物や臨床所見の用語について認識する．その後，その用語の前後の文脈から，読影医がレポートに記述した情報の確信度（Certainty Scale）を評価する．

2. 方法

1) 対象データ

本研究では，大阪大学医学部附属病院の画像診断レポートシステムに蓄積されている胸部 CT 画像及び腹部 CT 画像のレポートを利用した．本研究は大阪大学医学部附属病院の観察研究倫理審査委員会の承認（承認番号 17166）を得て実施した．

2) アノテーション

まず，蓄積されたレポートから，無作為に 500 件（胸部レポート：300 件，腹部レポート：200 件）を抽出した．各レポート 500 件について，先行研究の手法[1]を用いて，合計 4,597 件の観察物や臨床所見の用語を認識し，各用語についてアノテーション作業を実施した．アノテーション作業は，3 名の医学生により実施された．アノテーターは，レポ

ートのタグ付けされた観察物や臨床所見の用語を含むテキストについて，それぞれ確信度 scale を後述する 5 段階の基準に沿って付与した．多数決により，決定できない事例については，医師が最終的な分類を行った．

3) 確信度の基準

確信度については，高い順に「Definite, Likely, May represent, Unlikely, Denial」の 5 段階の scale でガイドラインを設計した．

Table.1 にアノテーション結果を示す．

Table.1　グレード別のサンプル数

Grage	サンプル数
Definite	2,738
Likely	66
May represent	356
Unlikely	117
Denial	1,320
Total	**4,597**

4) 分類モデル

分類モデルには BERT[4]を採用した．分類には，原著論文と同様に BERT の先頭系列に付与される[CLS] token の出力を用いた．

3. 結果

アノテーション済のレポートを，4 分割交差検証により評価した．指標については，予測と正解のクラスが一致した場合のみ正解と判定する「strict」な基準，隣接する scale 間の誤りを許容する「relaxed」の基準で評価した（Table.2）．ただし，「Denial」に関しては「relaxed」の場合も，厳密な一致のみを正解とするよう判定した．

4. 考察

実験の結果，いずれの基準でも，Micro 平均では，高い性能を示した．これは，サンプル数が多かった Definite, Denial な結果が貢献している．一方 Likely や Unlikely はサンプル数が少なかったこともあり，十分な性能で予測できておらず，strict 設定における Macro 平均の精度は低かった．ただし，基準を緩和した「relaxed」で評価した場合は，これらの scale も F1-score が 90%を超えており，

Table.2　scale 別の F-1（%）

Grade	Strict	Relaxed
Definite	98.47	99.00
Likely	64.99	95.04
May represent	93.14	96.15
Unlikely	84.24	90.24
Denial	98.86	98.86
Macro avg	**87.94**	**95.86**
Micro avg	**97.33**	**98.49**

Macro 平均を押し上げていることが分かる．このことから，モデルはラベル数の少ない scale を正確に予測するのは困難であったが，近い scale を予測することは高い精度で可能であったということが示唆される．

今後は特定の臨床所見の確信度 scale を評価して本システムの実用性について検討したい．

1. 結語

放射線レポートに記載される観察物や臨床所見の確信度について 5 段階の scale 情報を付与した．BERT を用いた分類モデルを構築し，高い精度でその scale を分類できた．

参 考 文 献

[1] 杉本 賢人，和田 聖哉，山畑 飛鳥，他：機械学習を用いた画像診断レポートからの情報抽出，第 20 回日本医療情報学会学術大会 397-401, 2019.

[2] 杉本 賢人，和田 聖哉，山畑 飛鳥，他：放射線レポートからの情報抽出と構造化に関する取り組み，第 24 回日本医療情報学会春季学術大会, 2020.

[3] Khorasani R, Bates DW, Teeger S, et al. Is terminology used effectively to convey diagnostic certainty in radiology reports?, Academic Radiology 10(6)685-8, 2003.

[4] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding, In North American Association for Computational Linguistics (NAACL),2019