読影レポート間の類似度データセットの構築と予備実験

多田太郎 [1]　森川みどり [2]　那須照広 [2]　山本和英 [1]
[1] 長岡技術科学大学　[2] 株式会社ワイズ・リーディング
{tada,yamamoto}@jnlp.org,　{m-morikawa,t-nasu}@ysreading.co.jp

1　はじめに

医療文書のひとつに読影レポートがある．読影レポートは，病院の放射線科での画像検査に伴い作成される報告書である．読影レポートの作成は，医師に取って大きな負担となっている現状がある．

また近年，BERT（Bidirectional Encoder Representations from Transformers）が自然言語処理分野の様々なタスクに用いられ，高い精度を示している．英語の医療文書を対象とする報告では，ドメインに特化させた BERT だけでなく，一般ドメインの大規模コーパスで学習した BERT モデルも高い精度を示している．

本稿では，読影レポート作成の際に参考とする過去の読影レポートを得たいとの要望に対する取り組みとして，読影レポート所見テキスト間の類似度予測タスクのデータセットを構築し，BERT 学習済みモデルを含めた複数モデルで読影レポート間の類似度予測を行った．

2　関連研究

電子カルテを始めとする医療ドメインの文書を対象とした研究は、特に英語で書かれた文書を対象としたもので活発であり，多くの利用可能な資源が存在し，学術的な多くの取り組みが報告がされている．日本語の医療ドメイン文書の検索システムの研究では，小野ら [1] や岡本ら [2]，土井ら [3]，荒牧ら [4] などがある．

岡本らの論文では，診療文書内の文章に Observation，Diagnosis，Treatment の分類情報を付与し検索に適用している．土井らは，TF-IDF の情報を基に単語を基底とした文書ベクトル空間法を用いた類似症例検索を行っている．小野らは国際疾病分類である ICD コードと TF-IDF による単語の重み付けから最大全域木のデンドログラムを作成し，医療文書である退院サマリーのテキストの分類を行った．荒牧らは，連想検索エンジンである GETA をベースとした全文検索を行う症例検索システムを実際に 7 年間運用し報告をしている．

3　読影レポートについて

読影レポートは，病院での CT(コンピュータ断層撮影： computed tomography) 等を用いた画像検査後に検査画像を医師が読影し，所見や診断を記載した報告書である．

病院内の画像検査の流れを図 1 に示す．各診療科の医師は患者を診察し，必要に応じて画像検査を行なう放射線科へ検査依頼を行なう．放射線科では検査依頼を受けた後，検査オーダー文書に基づき患者の画像検査を行う．次に，撮影された検査画像から所見と診断を記載した一次読影レポートが作成される．その後，症例や病院の体制により作成しない場合もあるが，放射線科内の別の医師が患者の検査画像と作成された一次読影レポートを踏まえ，新たに読影を行い，一次読影レポート同様に所見と診断を記載した二次読影レポートを作成する．

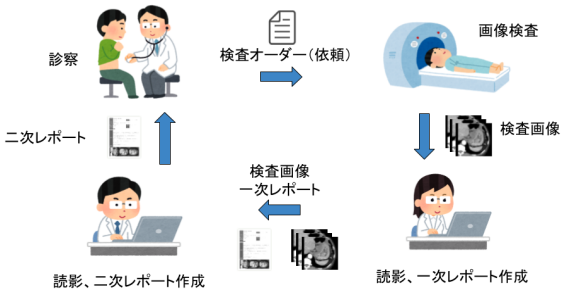


図 1　病院での画像検査の流れ

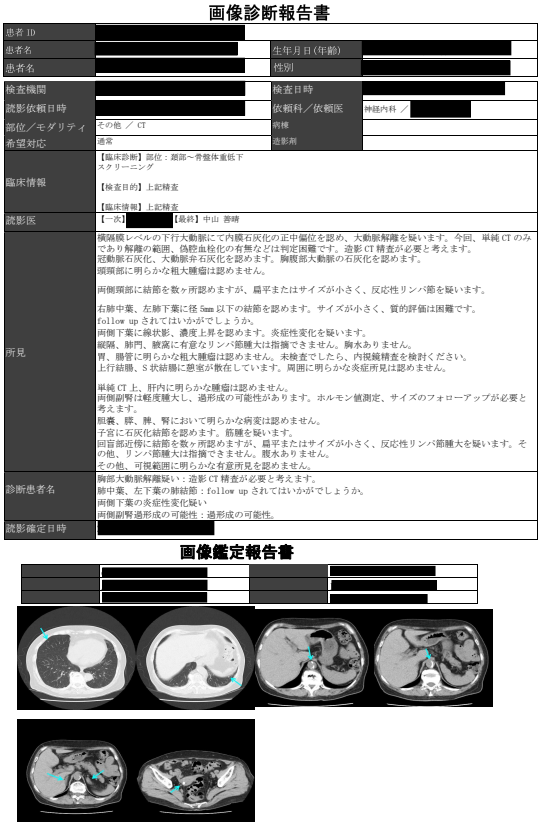
この一次読影レポートと二次読影レポートは，上記の様に別々の医師によって二段階で作成され，ダブルチェックの役割を担っている．放射線科で作成された二次読影レポートは，患者の検査依頼を行った診療科の医師に検査報告として提出される．図 2 に読影レポートの例を示す．読影レポートには検査画像に対する所見や診断の他に，検査方法や身体部位，検査依頼を行った診療科などの情報が含まれている．

画像検査においては，検査装置の進歩に伴い患者の体内をより細かく撮影することが可能となった．しかし，撮影枚数が増加し，1 つの検査で数百枚の画像を読影する必要がある場合もある．さらに，放射線科での 1 日の検査数も基本的に多い．読影レポートの作成作業は，担当する医師の大きな負担となっている [5]．

4　データセットの構築

本稿では，実際に読影レポートの作成に携わる医療分野の専門家の協力を得て，文書間類似度予測タスクのデータセットを構築した．データセットの構築に用いるデータは，181,645 検査分，約 310MB のテキストデータである．データセットを構築する上で考慮すべき読影レポートの特徴として，以下の様な点が挙げられる．

- 検査画像に写る全ての身体部位についての所見がレポートに記載される．



画像診断報告書

患者 ID	[redacted]		
患者名	[redacted]	生年月日(年齢)	[redacted]
患者名	[redacted]	性別	[redacted]
検査機関	[redacted]	検査日時	[redacted]
読影依頼日時	[redacted]	依頼科／依頼医	神経内科 ／ [redacted]
部位／モダリティ	その他 ／ CT	病棟	
希望対応	通常	造影剤	
臨床情報	【臨床診断】部位：頭部～骨盤体重低下スクリーニング 【検査目的】上記精査 【臨床情報】上記精査		
読影医	【一次】[redacted]【最終】中山 善晴		
所見	横隔膜レベルの下行大動脈にて内膜石灰化の正中偏位を認め、大動脈解離を疑います。今回、単純 CT のみであり解離の範囲、偽腔血栓化の有無などは判定困難です。造影 CT 精査が必要と考えます。 冠動脈石灰化、大動脈弁石灰化を認めます。胸腹部大動脈の石灰化を認めます。 頸頭部に明らかな粗大腫瘤は認めません。 両側頸部に結節を数ヶ所認めますが、扁平またはサイズが小さく、反応性リンパ節を疑います。 右肺中葉、左肺下葉に径 5mm 以下の結節を認めます。サイズが小さく、質的評価は困難です。 follow up されてはいかがでしょうか。 両側下葉に線状影、濃度上昇を認めます。炎症性変化を疑います。 縦隔、肺門、腋窩に有意なリンパ節腫大は指摘できません。胸水ありません。 胃、腸管に明らかな粗大腫瘤は認めません。未検査でしたら、内視鏡精査を検討ください。 上行結腸、S 状結腸に憩室が散在しています。周囲に明らかな炎症所見は認めません。 単純 CT 上、肝内に明らかな腫瘤は認めません。 両側副腎は軽度腫大し、過形成の可能性があります。ホルモン値測定、サイズのフォローアップが必要と考えます。 胆嚢、膵、脾、腎において明らかな病変は認めません。 子宮に石灰化結節を認めます。筋腫を疑います。 回盲部近傍に結節を数ヶ所認めますが、扁平またはサイズが小さく、反応性リンパ節腫大を疑います。その他、リンパ節腫大は指摘できません。腹水ありません。 その他、可視範囲に明らかな有意所見を認めません。		
診断患者名	胸部大動脈解離疑い：造影 CT 精査が必要と考えます。 肺中葉、左下葉の肺結節：follow up されてはいかがでしょうか。 両側下葉の炎症性変化疑い 両側副腎過形成の可能性：過形成の可能性。		
読影確定日時	[redacted]		

画像鑑定報告書

図 2　読影レポートの例

- 所見で記載される全ての身体部位について陽性・陰性の記載がある．
- 診断が同じであっても検査方法が異なる場合，参考となる文書にはならない．
- 撮影する検査部位（胸部などの大まかな部位）が重なっていても，異なる範囲（胸部と胸部～上腹部など）であると，検査画像が異なるものとなるため参考とならない場合がある．

データセットは，英語の医療ドメインの文類似度予測データセットである ClinicalSTS[6] を参考に構築した．ClinicalSTS では，文のペア群の各文ペアに対して，医療分野の専門家により 0 から 5 の 6 段階で類似度が付与されている．本稿で構築したデータセットは，上記のデータセットを参考に類似度を 0 から 3 の 4 段階に設定した．表 1 は，ClinicalSTS を参考に定めた読影レポート間の類似度付与基準である．本稿の文書類似度予測は，二次読影レポート作成時に参考となる過去の読影レポートを得ることが目的であることから，構築したデータセットには一次読影レポートの所見欄のテキストを用いた．また，同理由およびデータセットに含む文書ペアの選定は専門家が行わないことから，高い類似度のスコアリング基準に幅をもたせた．

文書ペア収集の際は，スコア付与後のデータセットでのスコアの偏りを少なくするため，表 1 内のトピックには身体部位を，同等の目安には疾病を，詳細には読影レポート内の所見欄テキストの詳細を置き換えて収集作業を行った．また，放射線科のサポートを目的とする観点から，上記の読影レポートの特徴を考慮し，検査方法についても文書ペアで異ならないようにした．

スコアリングを行った専門家は，データセットを構築する目的やスコアリング方法の説明を受け，表 1 の一般化された文書ペアの関係性をふまえ，専門知識と読影レポート作成に関わる経験を基にスコアリングを行った．表 2 は，専門家らが実際にスコアリングを行う際に，専門家らが定め使用したスコアリング基準である．最終的なデータセットは，一次読影レポート所見テキストの 260 ペアで，各ペアに 1 つずつ類似度が付与されている．また，データセットに含まれる所見テキストの平均の長さは 155.5 文字である．表 5 は構築したデータセットの一例である．

5　実験

構築したデータセットをランダムで学習用 200 ペア，開発用 30 ペア，テスト用 30 ペアに分割し，実験に使用した．類似度予測の手法には，嶋中ら [7] の BERT[8] を用いた翻訳文の自動評価モデルを使用した．本手法は，翻訳元の文と翻訳先の文のペアを BERT モデルに入力し，出力された文ペアの分散表現から翻訳の質を予測するモデルであり，翻訳文の自動評価タスクで高い精度を示している．

使用する BERT モデルには，Wikipedia[*1]で学習されたモデル[*2]を用いた．英語の医療ドメインにおいても，Clinical BERT [9] など，PubMed [*3]などの医療ドメインの大規模コーパスと一般ドメインの大規模コーパスにより学習されたモデルが公開され高い性能を示している．また，同論文内で比較されている一般ドメインの大規模コーパスのみで学習された BERT のモデルも，医療ドメインに特化したモデルと比較すると，同ドメインのタスクに対して精度は劣るものの，決して低い性能を示しているわけではない．そのため，タスクにもよるが，医療ドメインのタスクであっても，利用可能な言語資源や計算コストの都合を踏まえて一般ドメインの大規模コーパスで学習された BERT モデルを使用する選択は有効であると考えた．実験を行う際のファインチューニングには，構築したデータセットの学習用ペアのみを用いた．

構築したデータセットへの前処理としては，neologdn[*4]で正規化を行った後、否定文および挨拶文の削除を正規表現を用いて行った．評価にはピアソン相関を使用した．比較手法とした TF-IDF，Doc2Vec[*5] および USE（Universal Sentence Encoder）[*6]は，文書間

[*1] https://www.ncbi.nlm.nih.gov/pubmed/
[*2] https://github.com/cl-tohoku/bert-japanese
[*3] https://www.ncbi.nlm.nih.gov/pubmed/
[*4] https://pypi.org/project/neologdn/
[*5] https://radimrehurek.com/gensim/models/doc2vec.html#gensim.models.doc2vec.Doc2Vec
[*6] https://tfhub.dev/google/

表 1　文書類似度予測タスク構築のために定めた類似度基準

スコア	内容
0	2 つの文は異なるトピックにある．
1	2 つの文は同等ではないが，同じトピックに関するものである．
2	2 つの文は同等ではないが，詳細を共有している．
3	2 つの文は，完全にまたはほぼ同等で，同じことを意味している．
	あるいは，ほぼい同等であるが，いくつかの情報が異なる/欠落している．

表 2　専門家らによるテキスト間の類似度スコアリング基準

スコア	内容
0	共通項なし．
1	共通の部位についての記載があるが， 主病変が異なっている．
2	主病変が同じで表現が違っている．
3	主病変が同じで表現もほぼ同じである．

のコサイン類似度を人手スコアとのピアソン相関に用いた．TF-IDF の計算にはデータセットおよびデータセットに含めなかった読影レポートも用い，Doc2Vec のモデルの学習もデータセットに含めなかった読影レポートで行った．学習済みで利用可能な他のモデルとして USE も実験に用いた．

6　結果・考察

表 3 に実験結果を示す．読影レポートを用いて学習した Doc2Vec の出力が，最も人手との相関が高い結果となった．Wikipedia で学習した BERT モデルを使用した分類器および USE は，5 節の通り，より多くの読影レポートを用いた TF-IDF に近い相関を得ることが出来た．表 4 に手法同士のピアソン相関を示す．学習済みモデルを用いた手法とそうでない手法の間で比較的低い相関となった．このことからも，Clinical BERT など一般ドメインと医療ドメインの両方を用いて学習したモデルの精度が高い現状が理解できる．

また本稿では，否定文を削除するというシンプルなクリーニングを行った．本稿で実験に用いた手法全てでそうであるが，クリーニング後にテキストが短くなった際に単語の重なりから大きく影響を受ける例がみられた．これらは，単語の出現頻度を確保した上で，固有名詞をより大きな粒度の単語として学習することにより，ある程度避けることができると考えられる．読影レポートの詳細な情報をより正確に得るため，特徴に沿ったより細かなクリーニングを行う場合，表記ゆれなどの観点から単語の重なりが必ずしもテキスト内に現れるとは限らない．その場合，分散表現手法が有利になってくる可能性が考えられる．

本稿で構築したデータセットにより，教師ありの手法を用いることが可能となった．しかし，結果を踏まえると，一般ドメインの大規模コーパスで学習された BERT モデルのファインチューニングを行なうには，データセットの規模が十分とは言えない．しかし，構築したデータセットを基に擬似データセットを生成するなど，現状から精度を向上させるための新たな取り組みが可能な状態となった．

4 節で述べた読影レポートの特徴は，読影レポートを扱う上での前処理にも通ずる特徴である．そのため，所見テキスト内の記載内容の位置関係やメインの検査部位とそうではない検査部位を考慮するなど，テキストのクリーニングにも改善可能な点が多い．これらより，モデルの精度向上の可能性は多くあり，タスクで求められる精度にもよるが，医療ドメインのタスクであっても，一般ドメインのコーパスで学習されたモデルを用いることが選択肢になりうると考えられる．

表 3　実験結果

手法	人手とのピアソン相関
ランダム	0.066
TF-IDF	0.556
Doc2Vec	0.627
USE	0.552
BERT	0.552

表 4　手法同士のピアソン相関

	TF-IDF	Doc2Vec	USE	BERT
TF-IDF	-	0.772	0.497	0.376
Doc2Vec	-	-	0.418	0.247
USE	-	-	-	0.665
BERT	-	-	-	-

7　おわりに

本稿では，読影レポートにおける類似文書を得るための取り組みとして，文書間類似度予測タスクのデータセットの構築とデータセットを用いての実験を行った．データセットは小規模ではあるが，実際に医療分野の専門家によるアノテーションがなされており，今後の研究手法の選択肢を増やすものとなった．また，学習済み BERT モデルを用いた実験では，小規模のデータセットによるファインチューニングのみで，学習に読影レポートを用いた他の手法に近い精度を得ることが出来た．

謝辞

本研究は、平成 29-31 年学術研究助成基金助成金 挑戦的研究 (萌芽) 課題番号 17K18481 の助成を受けています。

universal-sentence-encoder-xling-many/1

表 5　読影レポート所見テキストのペア例

スコア	文書ペアの例
0	右股関節に、臼蓋低形成と亜脱臼が存在するようです。右股関節間隙は、荷重面にて狭小化しています。軟骨下骨の骨硬化を伴い、臼蓋、および大腿骨頭軟骨下に骨嚢胞を形成しています。右変形性股関節症の所見です。画像参照ください。
	右肘関節の変形、破壊が目立ちます。骨棘形成も見られます。肘関節内には多数の遊離体が形成されています。高度な変形性肘関節症の像です。骨折を疑う所見は指摘できません。骨腫瘍を疑うような病変はありません。その他、観察範囲内に明らかな異常所見を認めません。
1	肝には脂肪沈着を認めます。肝内に粗大な腫瘍性病変は指摘できません。胆、膵、脾、副腎、腎に有意な異常は指摘できません。腹水は指摘できません。腹部骨盤領域に病的リンパ節腫大は指摘できません。
	肝表面は不整で、脾腫を伴い、肝硬変の所見と考えます。腹水、腸管の浮腫、側副血行路発達による食道静脈瘤も肝硬変に伴う変化と考えます。肝内に HCC を疑う腫瘤は認めません。腹腔内に有意に腫大したリンパ節は認めません。CT では消化管に粗大な充実性腫瘍は指摘できません。描出範囲内の肺野に特記すべき異常所見は指摘できません。その他、明らかな異常は認めません。
2	子宮筋層内に多数の T2 強調像にて低信号の腫瘤を認め、子宮筋腫を疑います。最大のもので長径 60mm 程度です。前壁の筋腫は T2 強調像にてやや高信号で dynamic study にて漸増性の増強効果を認めます。比較的変性の乏しい筋腫を疑います。付属器に有意な器質病変を認めません。腹水を認めません。その他、骨盤腔内に有意な異常を認めません。
	子宮体部筋層内から漿膜下に、T2 低信号を示した径 9mm～42 × 31mm 程の結節が 3 ヶ所認められます。多発子宮筋腫が疑われます。子宮体部内腔に沿って、T2 低信号を示し、造影されない層状の病変が認められ、月経周期に伴った血腫の可能性があります。両側卵巣には明らかな病変は認められません。ごく少量の腹水貯留が認められます。明らかなリンパ節腫大は認められません。
3	膀胱左側壁に 35x23x35 径の腫瘤が見られます。内部は T2 強調画像で筋層と同等かやや高信号として描出されています。拡散強調画像では高信号として描出され、造影検査では、不整が強い造影剤増強効果を伴っています。膀胱癌所見と思われます。腫瘤は周囲脂肪織に広範に浸潤していますが周囲臓器や骨盤壁浸潤はなさそうです。（T3b）有意なリンパ節腫大は見られません。
	膀胱右腹側に内腔に突出する 30 径の乳頭状腫瘤が見られます。拡散強調画像で高信号として描出され、ADC 値に低下が見られます。膀胱癌が疑われる所見です。T2 強調画像でも腫瘍基部の低信号線には断裂が見られ、膀胱壁のへこみも見られます。周囲脂肪織への浸潤も疑われ、病期的には T3 と思われます。前立腺肥大が見られますが前立腺癌示唆する所見は見られません。腹水や有意なサイズのリンパ節腫大は見られません。

参考文献

[1] 小野大樹, 高林克日己, 鈴木隆弘, 横井英人, 井宮淳, 里村洋一. テキストマイニングによる退院サマリー自動分類の試み. 医療情報学, Vol. 24, No. 1, pp. 35–44, 2004.

[2] 岡本和也, 竹村匡正, 黒田知宏, 長瀬啓介, 吉原博幸. 文脈に基づく類似診療文書検索システム. 生体医工学, Vol. 44, No. 1, pp. 199–206, 2006.

[3] 土井俊祐, 木村隆, 関根正樹, 鈴木隆弘, 高林克日己, 田村俊世. 学会ホームページにおける類似症例検索システムの実運用と評価. 生体医工学, Vol. 49, No. 6, pp. 870–876, 2011.

[4] 荒牧英治, 岩尾友秀, 若宮翔子, 伊藤薫, 矢野憲, 大江和彦. 症例検索システムの試行運用に基づいた利用状況に関する基礎的検討. 医療情報学, Vol. 38, No. 4, pp. 245–256, 2018.

[5] 牧野恭子, 早川ルミ, 寺井公一, 深津博. 知識処理を活用した読影レポート作成支援システムの開発と評価. 人工知能学会論文誌, Vol. 23, No. 6, pp. 463–472, 2008.

[6] Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. Overview of biocreative/ohnlp challenge 2018 task 2: Clinical semantic textual similarity. 2018.

[7] 嶋中宏希, 梶原智之, 小町守. 事前学習された文の分散表現を用いた機械翻訳の自動評価. 自然言語処理, Vol. 26, No. 3, pp. 613–634, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[9] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Clinical NLP Workshop*, pp. 72–78, 2019.