様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ２９ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ０７１９６

研究課題名（和文）機械学習を用いた画像診断レポートからの情報抽出と利活用に関する研究

研究課題名（英文）A Study on Extracting Information from Diagnostic Imaging Reports Using Machine Learning and Its Utilization

研究代表者
武田　理宏（Takeda, Toshihiro）

大阪大学・大学院医学系研究科・教授

研究者番号：７０５０６４９３

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：画像診断レポートから機械学習を用いた自然言語処理による情報抽出を行った。最初に固有表現抽出を用いて、「観察物」、「臨床所見」、「部位」、「変化」、「特徴」、「サイズ」の表現を抽出した。次に、抽出した固有表現の関係抽出を行った。最後に、文章の確信度を「確定」から「否定」まで5段階の尺度基準を分類した。
構築した情報抽出モデルを、大阪府下5施設の医療機関のデータウェアハウスに蓄積された画像診断レポートに適応し、JSON形式で出力を行った。
画像診断レポートの見落とし対策として、構築した情報抽出モデルを用いて、がんが含まれる画像診断レポートを抽出するプログラムを構築した。

研究成果の学術的意義や社会的意義
近年、リアルワールドデータ（RWD）の利活用が議論されるが、電子カルテデータで利活用の対象となるのは、レセプト・DPC情報、病名、処方、注射、検体検査結果などに限定されることが多い。
本研究は、フリーテキストで記載されるRWDを利活用するための取り組みの一つである。画像診断レポートから情報抽出することで、画像で診断される疾患を有する患者を正しく抽出することができる。また、画像で治療判定を行う疾患の、治療効果を検証することが可能となる。
我々は、がん所見が含まれるレポートを抽出することに成功した。これは、近年、社会問題となっている画像診断レポートの見落とし対策として、活用することが可能である。

研究成果の概要（英文）：We extracted information from diagnostic imaging reports using natural language processing with machine learning. Firstly, First, named entity recognition was performed to extract "observation" "clinical finding," "anatomical location," "change," "characteristics," and "size" expressions. Next, the relation extraction between "observation" and the other extracted expressions was performed. Finally, we classified the confidence level of the sentences into five scale criteria, ranging from "definite" to "denial". Our information extraction model was adapted to diagnostic imaging reports stored in the data warehouses of five medical facilities in Osaka Prefecture and output in JSON format.
To prevent diagnostic imaging reports from being overlooked, we built a program to extract diagnostic imaging reports that contain cancer using our information extraction model.

研究分野：医療情報学

キーワード：リアルワールドデータ　自然言語処理　機械学習　画像診断レポート　二次利用

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、リアルワールドデータとして多施設の電子カルテデータベースに蓄積された臨床データの利活用が取り組まれている。データの利活用は構造化、標準化されたデータが対象となるため、その範囲はレセプト・DPC、処方・注射オーダ、検体検査結果などに限定される。利活用できる電子カルテ蓄積データを増やしていくことは、医療情報学分野では喫緊の課題である。経過記録や検査レポート、退院サマリには、データ利活用が望まれる情報が多く記述される。しかし、前向き研究等でデータ収集を目的にテンプレート入力等を行っていない限り、これらの記録はフリーテキストで入力されるため、データの利活用は進んでいない。

多くの分野で、フリーテキストデータに対する自然言語処理（NLP: Natural Language Processing）に機械学習が用いられるようになり、その精度が上がっている。一方、電子カルテデータに対する自然言語処理は工学的知識に加え、形態素に対するタグ付けや取得された知識に対するシソーラスの構築に医学知識が必要なこと、解析対象が日本語であること、医療機関外にデータの持ち出しが容易でないことから、研究はまだ十分に進んでいない。

２．研究の目的

本研究では、フリーテキストで記述された画像診断レポートの自然言語処理を研究の対象とする。画像診断レポートは教育を受けた放射線読影専門医が主治医に所見を伝えるために記述するため、ある程度の質が担保されていること、臓器ごとに患者病態を反映する情報が記述されるためデータの利活用が期待されることが研究対象とする理由である。機械学習により収集された用語は、医学知識を持った研究者によりシソーラスを整備する必要がある。さらに、収集された情報を実際に利活用することで、その有用性を評価する。

３．研究の方法

３－１．自然言語処理

大阪大学医学部附属病院の胸部 CT レポート、腹部 CT レポートを検証の対象とし、3 名の医学専門家（臨床医 2 名、放射線技師 1 名）、医学生がアノテーション作業を行った。

３－２．エンティティ抽出（Named entity recognition (NER)）

前処理として、正規表現を使用してレポートを文に分割し、各文を MeCab でトークン化した。トークン化したシークエンスをモデルに使用した。NER タスクで広く使われているタグ付けフォーマットである IOB2 フォーマット（B タグと I タグはそれぞれエンティティの始まりと内側を表し、O タグはエンティティの外側を表している）を用いて、Observation（観察物）、Clinical finding（臨床所見）、Anatomical location（部位）、Change（変化）、Characteristics（特徴）、Size（サイズ）、Certainty（確信度）のタグ付けを行った。NER のための最先端の深層学習モデル、BiLSTM-CRF、BERT、BERT-CRF を比較した。

３－３．リレーション分類（Relation Extraction）

NER で抽出したエンティティを Observation と Clinical finding に関するエンティティ（Object）とそれらの属性エンティティ（Attribute）に分類し Object と Attribute の関係有無を学習するためのモデルを構築した。このモデルは, Object と Attribute の位置情報が分かるように工夫した入力テキストを与えることで、その Object と Attribute の関係の有無の二値分類を出力するよう設計した。複数の文にまたがる関係を抽出するために、分割された分が一つの連続したシークエンスに連結した。深層学習モデルとして、BiLSTM attention model と BERT を用いて、その性能を比較した。

３－４．確信度判定

我々は確信度について，高い順から Definite（高い確信度を持って診断）、Likely（他の鑑別すべき疾患よりも可能性が高い）、May represent（鑑別疾患のうち、可能性が高い疾患を明言していない）、Unlikely（他の鑑別すべき疾患よりも可能性が低い）、Denial（疾患の存在を否定している）の 5 段階に分類した。NER で抽出した Observation（観察物）、Clinical finding（臨床所見）に対しタグ付けを行い、BERT を用いて、このタグを予測するモデルを構築した。

確信度判定は、特に Likely、May represent、Unlikely の判定は難しい場合ある。そこで、構築したモデルは、「Strict」（予測と Ground Truth が一致した場合のみを正解として判定する）と「Relaxed」（予測と Ground Truth の scale が±1 の場合は誤りを許容する。ただし、Denial は Strict と同等の評価）の 2 つの基準を用いて精度評価した。

３－５．構造化データベースの構築

各医療機関の DWH に蓄積されている放射線レポートのテーブルからデータ抽出プログラムを用いて、多施設共通 DWH にデータをインポートする。この際、放射線領域における標準コードである JJ1017 とマッピングを行う。次に、インポートされた多施設共通 DWH の放射線レポートテーブルのレコードを対象に構造化処理を実施する。

構造化アルゴリズムは大きく本研究で構築した「エンティティ抽出」、「リレーション分類」、「確信度分類」の 3 つのサブモジュールから構成される。

構造化の結果は、Object と Modifier や Object 同士のリレーションの配列を持つ型として表現される。ただし、配列の要素数やリレーションのカーディナリティはレポートによって様々であり、これをリレーショナルデータベースのテーブル形式で表現する場合、事前に定義した特定のスキーマに依存することになるため、データの柔軟性が損なわれてしまう。そこで、レポートのメタ情報は、従来のテーブル形式のフィールドで保持しながら、構造化の結果をユニバーサルな形式で表現するため JSON 型で表現する形でテーブルを設計した。構造化テーブルでは患者 ID、オーダ番号、グループ番号、版、検査日、検査時刻をプライマリキーに持ち、構造化結果を JSON 形式で記述する。

３－６．重要所見が書かれる画像診断レポートの抽出

JSON 形式で構造化した画像診断レポートに対して、重要所見判定モジュールを適応する形とした。本研究において重要所見はがん所見を対象とした。

最初に医学生 3 名が、腫瘍に関するフラグのついた所見に対して、良性（Benign）、悪性（Malignant）、判断できない（Indeterminate）のコードを付与した。重要所見の判定所見は、抽出した Observation、Clinical Finding の用語から腫瘍コードが付与されており、Malignant または Indeterminate のコードが付与された所見のうち、確信度判定で unlikely 以上の判定が行われたものを抽出した。これだけの条件では偽陽性となる症例が多く出るため、除外条件として、抽出された要素が observation の場合、①その要素に紐づく clinical finding を参照し、その用語に腫瘍コードが付与されていない、もしくは Benign が付与されているもの、②その要素に紐づく change 表現を参照し、その用語が「改善・変化なし」の場合（変化表現についても、事前に頻出用語に「改善・変化なし」のコーディングを実施）とした。

４．研究成果

４－１．エンティティ抽出（Named entity recognition (NER)）

構築したエンティティ抽出モデルの精度は、平均 F-1 値が BiLSTM-CRF が 96.1、BERT が 95.2、BERT-CRF が 95.4 と BiLSTM-CRF が最も良い性能を示した。胸部 CT レポート、腹部 CT レポートの BiLSTM-CRF を用いた各エンティティの抽出精度を表 1 に示す。

表 1．胸腹部 CT から各エンティティの抽出精度

Entity type	Chest			Abdomen		
	Precision	Recall	F-1	Precision	Recall	F-1
Observation	95.3	97.0	96.1	93.8	96.4	95.1
Clinical finding	95.3	93.2	94.2	97.9	96.4	97.1
Anatomical location	96.0	96.6	96.3	96.0	96.6	96.3
Certainty	98.7	98.5	98.6	99.4	98.9	99.1
Chang	89.5	91.5	90.5	92.4	93.6	93.0
Characteristis	89.1	90.0	89.5	88.5	85.3	86.9
Size modifier	99.3	98.2	98.7	98.6	98.9	98.7
Micro-averaging	95.6	96.0	95.8	96.2	96.5	96.3

４－２．リレーション分類（Relation Extraction）

構築した関係抽出モデルでは、BiLSTM-Attention モデルの平均 F1 値が 95.6、BERT の平均 F1 値が 97.6 エンティティ抽出モデルの実験結果とは対照的に、BERT モデルは BiLSTM attention モデルを F-1 値で 2.0%上回った。胸部 CT レポート、腹部 CT レポートの BERT を用いた各エンティティの関係抽出精度を表 2 に示す。

表 2．胸腹部 CT から各エンティティの関係抽出精度

Relation type	Chest			Abdomen		
	Precision	Recall	F-1	Precision	Recall	F-1
Modifier relation						
Anatomical location	97.9	97.8	97.9	96.2	99.0	97.6
Certainty	99.4	99.3	99.4	99.4	99.6	99.5
Change	96.0	94.7	95.4	94.7	95.4	95.0
Characteristics	95.3	94.9	95.1	95.5	97.5	96.5
Size	98.1	100.0	99.1	98.3	97.7	98.0
Evidence relation						
Clinical finding	95.7	97.8	96.7	89.8	91.1	90.4
Micro-averaging	97.7	97.8	97.7	96.6	98.3	97.4

４－３．確信度判定

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、合計 4,597 件の observation や clinical finding 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。スケール別のサンプル数は、Definite が 2,738 件、Likely が 66 件、May represent が 356 件、Unlikely が 117 件、Denial が 1,320 件であった。構築した機械学習モデルでの scale 別の F-1 値を表３に示す。

表 3. 各スケールで確信度判定精度

Grade	Strict	Relaxed
Definite	98.41	99.05
Likely	58.82	97.51
May represent	91.75	96.43
Unlikely	85.58	93.54
Denial	98.67	98.67
Avg.	97.10	98.59

４－４．重要所見が書かれる画像診断レポートの抽出

がん登録患者を参照して用意した評価セットにおける結果は除外ルールがない場合、感度（Sensitivity）は 98.20%、陽性適中率（PPV）は 88.19%を記録した。除外ルールがない場合、より多くのレポートを重要所見として判定することになるので、感度（Sensitivity）は高いが、陽性適中率（PPV）が低く、多くの偽陽性を伴う結果となっていることが分かる。除外ルールを適用することで、感度（Sensitivity）は 95.10%と 3.1%の悪化が見られるが、陽性適中率（PPV）は 96.09%と 7.9%の改善が確認された。

がん登録を参照する場合、偏った集団での適用結果となっている可能性があるため、実臨床に近い通常の分布からサンプルリングしたレポートの性能についても評価した。この評価セットにおける陽性適中率（PPV）は 73.78%であり、がん登録を参照した場合の陽性適中率（PPV）と比較した場合に大幅に低い結果となった。

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　0件／うちオープンアクセス　10件）

１．著者名 Gon Yasufumi、Sasaki Tsutomu、Kawano Tomohiro、Okazaki Shuhei、Todo Kenichi、Takeda Toshihiro、Matsumura Yasushi、Mochizuki Hideki	４．巻 222
２．論文標題 Impact of stroke on survival in patients with cancer	５．発行年 2023年
３．雑誌名 Thrombosis Research	６．最初と最後の頁 109～112
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.thromres.2023.01.002	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kikuchi Masataka、Kobayashi Kaori、Itoh Sakiko、Kasuga Kensaku、Miyashita Akinori、Ikeuchi Takeshi、Yumoto Eiji、Kosaka Yuki、Fushimi Yasuto、Takeda Toshihiro、Manabe Shirou、Hattori Satoshi、Nakaya Akihiro、Kamijo Kenichi、Matsumura Yasushi	４．巻 20
２．論文標題 Identification of mild cognitive impairment subtypes predicting conversion to Alzheimer’s disease using multimodal data	５．発行年 2022年
３．雑誌名 Computational and Structural Biotechnology Journal	６．最初と最後の頁 5296～5308
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.csbj.2022.08.007	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Nakatani Daisaku、Dohi Tomoharu、Takeda Toshihiro、Okada Katsuki、Sunaga Akihiro、Oeun Bolrathanak、Kida Hirota、Sotomi Yohei、Sato Taiki、Kitamura Tetsuhisa、Suna Shinichiro、Mizuno Hiroya、Hikoso Shungo、Matsumura Yasushi、Sakata Yasushi	４．巻 4
２．論文標題 Relationships of Atrial Fibrillation at Diagnosis and Type of Atrial Fibrillation During Follow-up With Long-Term Outcomes for Heart Failure With Preserved Ejection Fraction	５．発行年 2022年
３．雑誌名 Circulation Reports	６．最初と最後の頁 255～263
掲載論文のDOI（デジタルオブジェクト識別子） 10.1253/circrep.CR-22-0006	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Sugimoto Kento、Takeda Toshihiro、Oh Jong-Hoon、Wada Shoya、Konishi Shozo、Yamahata Asuka、Manabe Shiro、Tomiyama Noriyuki、Matsunaga Takashi、Nakanishi Katsuyuki、Matsumura Yasushi	４．巻 116
２．論文標題 Extracting clinical terms from radiology reports with deep learning	５．発行年 2021年
３．雑誌名 Journal of Biomedical Informatics	６．最初と最後の頁 103729～103729
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jbi.2021.103729	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Wang Bowen、Takeda Toshihiro、Sugimoto Kento、Zhang Jiahao、Wada Shoya、Konishi Shozo、Manabe Shirou、Okada Katsuki、Matsumura Yasushi	4．巻 209
2．論文標題 Automatic creation of annotations for chest radiographs based on the positional information extracted from radiographic image reports	5．発行年 2021年
3．雑誌名 Computer Methods and Programs in Biomedicine	6．最初と最後の頁 106331～106331
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.cmpb.2021.106331	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Manabe Shirou、Takeda Toshihiro、Hattori Atsushi、Yamamoto Masashi、Shimai Yoshie、Namiuchi Yoshiki、Yamaguchi Junji、Yamada Tomomi、Konishi Shozo、Matsumura Yasushi	4．巻 210
2．論文標題 Practical use of a multicenter clinical research support system connected to electronic medical records	5．発行年 2021年
3．雑誌名 Computer Methods and Programs in Biomedicine	6．最初と最後の頁 106362～106362
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.cmpb.2021.106362	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Teramoto Kei、Takeda Toshihiro、Mihara Naoki、Shimai Yoshie、Manabe Shirou、Kuwata Shigeki、Kondoh Hiroshi、Matsumura Yasushi	4．巻 9
2．論文標題 Detecting Adverse Drug Events Through the Chronological Relationship Between the Medication Period and the Presence of Adverse Reactions From Electronic Medical Record Systems: Observational Study	5．発行年 2021年
3．雑誌名 JMIR Medical Informatics	6．最初と最後の頁 e28763～e28763
掲載論文のDOI（デジタルオブジェクト識別子） 10.2196/28763	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Sugimoto Kento、Takeda Toshihiro、Oh Jong-Hoon、Wada Shoya、Konishi Shozo、Yamahata Asuka、Manabe Shiro、Tomiyama Noriyuki、Matsunaga Takashi、Nakanishi Katsuyuki、Matsumura Yasushi	4．巻 116
2．論文標題 Extracting clinical terms from radiology reports with deep learning	5．発行年 2021年
3．雑誌名 Journal of Biomedical Informatics	6．最初と最後の頁 103729～103729
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jbi.2021.103729	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Toshihiro Takeda, Shirou Manabe, Atsushi Hattori, Junji Yamaguchi, Shozo Konishi, Yuichiro Yamamoto, Daiyo Takahashi, Yasushi Matsumura	４．巻 270
２．論文標題 An Automatic Image Collection System for Multicenter Clinical Studies.	５．発行年 2021年
３．雑誌名 Studies in Health Technology and Informatics	６．最初と最後の頁 23-27
掲載論文のDOI（デジタルオブジェクト識別子） 10.3233/SHTI200115	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kento Sugimoto, Toshihiro Takeda, Shoya Wada, Asuka Yamahata, Shozo Konishi, Shiro Manabe, Yasushi Matsumura	４．巻 270
２．論文標題 End-to-End Approach for Structuring Radiology Reports.	５．発行年 2021年
３．雑誌名 Studies in Health Technology and Informatics	６．最初と最後の頁 203-207
掲載論文のDOI（デジタルオブジェクト識別子） 10.3233/SHTI200151	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計7件（うち招待講演　1件／うち国際学会　1件）

１．発表者名 武田理宏
２．発表標題 画像レポート見落とし防止対策機能の標準実装に向けた取り組み
３．学会等名 第42回医療情報学連合大会（招待講演）
４．発表年 2021年

１．発表者名 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
２．発表標題 放射線レポートの確信度スケールの分類手法の開発
３．学会等名 第25回日本医療情報学会春季学術大会
４．発表年 2021年

１．発表者名
杉本 賢人、和田 聖哉、小西 正三、岡田 佳築、真鍋 史朗、武田 理宏、山本 征司、松村 泰志

２．発表標題
放射線レポートの二次利用に向けた構造化システムの開発に関する取り組み

３．学会等名
第42回医療情報学連合大会

４．発表年
2021年

１．発表者名
区 駿業、和田 聖哉、武田 理宏、岡田 佳築、真鍋 史朗、小西 正三、杉本 賢人、松村 泰志

２．発表標題
PDF 形式で保存される検査レポートからデータの抽出と構造化を実現するプログラムの開発

３．学会等名
第42回医療情報学連合大会

４．発表年
2021年

１．発表者名
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

２．発表標題
放射線レポートの確信度スケールの分類手法の開発

３．学会等名
第25回医療情報学春期学術集会

４．発表年
2021年

１．発表者名
杉本 賢人，和田 聖哉，小西 正三，武田 理宏，真鍋 史朗，松永 隆，松村 泰志

２．発表標題
放射線レポートのエンティティ抽出モデルの他部位・他病院への適用可能性の評価

３．学会等名
第40回医療情報学連合大会

４．発表年
2020年

１．発表者名 Bowen Wang, Toshihiro Takeda, Kento Sugimoto, Jiahao Zhang, Shoya Wada, Shozo Konishi, Shirou Manabe, Katsuki Okada, Yasushi Matsumura
２．発表標題 Semi Supervised Learning of Nodule Detection in Chest Radiographs
３．学会等名 11th Biennial Conference of the Asia-Pacific Association for Medical Informatics（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	松村　泰志 (Matsumura Yasushi) (90252642)	独立行政法人国立病院機構大阪医療センター（臨床研究センター）・その他部局等・機関長・部門長クラス (84414)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	杉本　賢人 (Sugimoto Kento)	大阪大学・医学系研究科・大学院生 (14401)	
研究協力者	張　家豪 (Zhang Jiahao)	大阪大学・医学系研究科・大学院生 (14401)	
研究協力者	王　博文 (Bowen Wang)	大阪大学・医学系研究科・大学院生 (14401)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関