The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

医療画像レポート所見の構造化
Medical Image Report Findings Structurization

謝 剣維 Xie Jianwei
鳥海 不二夫 Toriumi Fujio

東京大学工学系研究科システム創成学専攻
Graduate School of System Innovation, the University of Tokyo

Structuring of a medical report has various applications. But most of them are only for supporting professionals to write well-structured reports, which means it cannot structurize reports without manual checks. In this study, we propose an automatic method to structurize medical image report findings. It extracts important information like disease names or symptoms in sentences and marks them with proper tags. With these tags, professionals can easily get the information they need from the report. Furthermore, we apply the method to a real dataset and get remarkably good results.

1. はじめに

一般テキストに対しての構造化については様々な研究が行われているが，医療テキストについては充分研究が進んでいない．一方で，カルテのデジタル化により医療テキストのデータがどんどん増えていると共に，機械学習や深層学習や自然言語処理技術も急速に発展している．このようなバックグラウンドから，大量な医療テキストデータを利用することで，自動診断や医療支援システムなどのシステムを作り出すことが可能ではないかと期待される．しかし，自然言語テキストを直接そのまま機械学習や深層学習に適用しても良い結果が得られない場合が多く，何らかの事前の処理を行わなければならない．そこで，本研究では画像レポート所見の構造化によって，機械学習に適したフォーマットに書き換えることを目的として研究を行った．

2. 研究目的

医療レポート所見とは，CT 画像等で気づいた点について医師自分の判断に基づいてレポートに書いたものである．例えば：「所見：副腎の腫大や腹部大動脈周囲リンパ節腫大は認められない．膵に腫大を認めない．」等である．

医療レポート所見は診断の根拠になるものなので，診断そのものではない．このような医療レポートは各医師が自由に記述するために，フォーマットに規定がない．そのため，機械学習などで用いるには構造化を行った方が効率が良い．

構造化とは電子レポートに重要な情報を抽出，フォームのある文書に書き換えることである．そして構造化した文書は原文より機械学習，深層学習にも使いやすいと期待できる．例えば：

原文：肺野には肺内転移を疑う．

構造化後：<部位>肺野<部位>には<所見要素>肺内転移<所見要素>を<主張>疑う<主張>．

ここには，部位，所見要素，主張と三つのタグ名がある．部位とは人体の部位，所見要素とは医療画像に示した重要な画像情報，主張とは医者の判断であるように，各部分にその部分の医療専門属性を持つタグを付ける．こういう構造化を行うことで，医療画像を教師データとして利用することが容易となる．例えば数十万枚の CT 画像から，どのような画像であれば「肺内転移」があったと判断したか自動的にタグがつけられるようになる．さらにこの教師データを使って医療画像の認識や分類が可能となると期待される．

最終的には各部分にその部分の医療専門属性を持つタグを付けることが目的であるが，現段階ではまず身体部位や医療処置・手術と病気名の検出をおこなう．

連絡先：謝 剣維，東京大学 工学系研究科システム創成学専攻，070-1583-0405，xiejianwei0803@gmail.com

3. データ

本研究では，埼玉医科大学から提供された医療画像レポート 512,322 件を用いた．各レポートには，一件ごとに「性別，検査種，検査部位，所見，診断」という5つの項目がある．検査種については 40 種類がある（CT，MRI，PET 検診など）．部位については 79 個が存在する（頭部系，骨，胸部系など）．

医療文書の構造化には医療専門用語や辞書が必要となるが，必ずしも医療専門用語辞典に必要な用語がすべて掲載されているとは限らない．そこで，本研究では，特定の部位と検査種について構造化手法を見つけ，それあらゆる部位や検査に拡張するという手順を用いる．

4. 辞書作成

形態素解析と構文解析において良い効果を得られるために，医療専門用語の辞書が不可欠である．

辞書のソースとして以下の四つが考えられる．

① 医学辞書の利用
② 辞書から用語抽出
③ 構文解析による専門用語の抽出
④ 人手による辞書の作成

本研究では主に①と②を使用する．これから用いた辞書と辞書からの用語抽出処理を述べる．

4.1 利用辞書

医療用語構造化のために以下の医療辞書を利用する．

(1) 人体部位の辞書

日本解剖学会の用語集．これは 9000 個の用語が入っている用語集で，一般用語，人体についての用語，骨学，関節学；靭帯学，筋学，内臓学，脈管学，神経系，感覚器の 9 つのカテゴリーに分けられる．実際検出を行った結果がほぼ全部の人体部位が検出できる．

(2) 病気や医療処置の辞書:

① ICD 標準病名マスター：

病態毎に選んだ代表病名「病名表記」が収載された病名辞書である.

②標準手術・処置マスター辞書：
この中の手術・処置名称テーブルを使用した. このテーブルはレセプト電算処理マスター名称, 公開されている学会用語集からの名称および一般のマスターからの手術・処置名称などの情報をもたせて格納した辞書である.

4.2 辞書から用語抽出

テキスト中には病名や処置その用語丸ごと出現するとは限らないので, 細かい用語を得る処理を行い, 辞書に追加する必要となる. 例えば:

「歯の交換期障害」という病名があるが,「歯の交換期｜に起きる｜障害」という文も存在するので,「障害」という症状名を検出するには用語抽出が必要となる. ここでは, 病状名が用語の最後に付与されることが多いことに注目し, 以下の手順で抽出を行う.

抽出手順 :

① 辞書に載っている用語内部を逆順にし, それらをソートすれば, 違う部位の同じ症状が集まる. 例えば:
a) 瘍腫位部 A
b) 瘍腫位部 B
c) 害障断横髄頚
d) 害障期年更性男
e) 害障期換交の歯

② 前後の用語と比較, 最大共通部分を取る. ここの例の場合では, b)が a)と c)比較する時, a)との共通部分が多いので「瘍腫」の抽出ができた. 同じく,「障害」も抽出できた.

③ 抽出した用語をもう一度逆順にすることで, 病床名を抽出できる.「瘍腫→腫瘍」「害障→障害」.

5. 医療用抽出の前処理

まず, 前処理として正規表現で日付の検出, 数字と単位の検出, 文頭特殊表現の削除と三つの部分に分ける. それは, 形態素解析や構文解析する際にノイズとなるため, 検出できた部分を一つ全体として扱う必要がある.

5.1 日付

日付では具体的な日付と連続の時間を表す部分を検出する. 例えば:

① 2007. 1. 1　10 時 10 分

② 一ヶ月前

などが抽出対象となる. 学習データ 1000 文において日付に使われる文字を抽出し, それらが長さ 3 文字以上に連続出現した文字列を日付として抽出した.

5.2 数字と単位

数字と単位では病症サイズや範囲を表す部分を検出する. 例えば:

① 10mm x 10mm

② 2 ~ 3cm

などが抽出対象となる. 学習データ 1000 文に使われる符号と単位を抽出し, それらが長さ 3 文字以上に連続出現した文字列を数字と単位として抽出した.

5.3 文頭特殊表現

文頭特殊表現では文の頭に意味のない符号や数字を削除する. 表 1 に削除例を示す. 文書列は処理前の文書. 削除部分列は処理後削除した部分である.

抽出方法:特殊表現はテストデータ以外の 1000 文に第一文字が漢字, ひらがな, カタカナではないものを人手で確認して抽出した.

5.4 抽出精度の評価

以上の前処理をテストデータ 1000 文で検証した結果, 表 4 のように高い適合率再現率で正しく前処理を行うことができることを確認した.

表 1: 文頭特殊表現削除例

文書	削除部分
・両側上顎洞に液体の貯留を認めます.	「・」
1．両側上顎洞に液体の貯留を認めます.	「1．」

6. 医療専門用語 IDIOM 化

医療専門用語の IDIOM とは医療に関する意味を持つ文法的に一つの成分として存在する用語である. 図 1 に IDIOM 化の例を示す. ここに「右側頭後頭葉内側」は一つ医療専門用語の IDIOM だが, 形態素解析では「右」,「側」,「頭」,「後頭葉」,「内側」と五つの文節に分けられる. IDIOM 化とはこれらの文節を組み合わせて医療専門用語の IDIOM を検出し, 最後に意味相応のタグを付ける(図 1 はタグ付け部分を示していない).

原文：
右側頭後頭葉内側に梗塞巣を認めます

形態素解析結果：
右 | 側 | 頭 | 後頭葉 | 内側 | に | 梗塞 | 巣 | を認めます。

IDIOM化後：
右側頭後頭葉内側 | に | 梗塞巣 | を | 認めます。

図 1: IDIOM 化例

このステップは京都大学黒橋・河原研究室開発した構文解析ツール KNP(構文解析は形態素解析の結果に基づく必要となるため, 形態素解析ツール Juman++を使用した)を使用して構文解析を行った.

比較した結果で示されるように, IDIOM 化により, 文にある情報が明確となる. これによって, 文中のどこにどんな症状や所見要素が記述されているかを抽出することが可能となった.

6.1 IDIOM 化処理:

「左視床, 両側基底核に陳旧性小梗塞を認める.」を例文とし, 処理の手順を紹介する.

① 構文解析を行い，句を単位として文を分割する．図 2 に示すように「左視床」，「両側基底核」，「陳旧性小梗塞」と三つの句が分けられる．（一つ枠内は一つ句と呼ぶ．助詞は除く．）

② 4 章で作成した「人体部位の辞書」と「病気や医療処置辞書」及びそれらから抽出した専門用語で例文に最大マッチングする（最大マッチングというのは二つの用語がマッチングできた場合，長い方を選ぶ）．

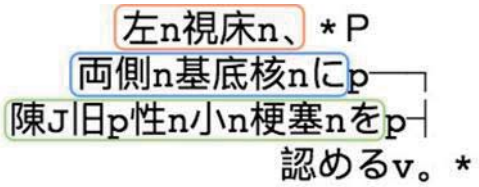

図 2: 構文解析結果

③ 図 3 の例の場合，辞書に「左視床」，「基底核」，「梗塞」，「小梗塞」載っている．「梗塞」と「小梗塞」に長いほうの「小梗塞」を選び，マッチングを行う）．

図 3: 辞書の最大マッチング

④ 方向や方位を表す単語が辞書に載っていないため，「右」，「右側」，「両側」の 15 個方向や方位を表す単語を用意する．②のマッチング結果の隣にこれらの単語が存在すれば，その結果を拡張する．例えば，「基底核」の前に「両側」が存在するため，結果を「両側基底核」に拡張する．

⑤ ①と③の結果を組み合わせる．句の中に辞書に載っている用語が存在すれば，その句全体を医療専門用語 IDIOM として扱う．この例の場合では，②に得た「左視床」，「両側基底核」が①の句と一致するため，検出はできたと判断する．しかし，「小梗塞」は「陳旧性小梗塞」に含まれる時，その句の全体（陳旧性小梗塞）を医療専門用語 IDIOM として検出した．

7. 結果

医療専門用語 IDIOM 正確検出の定義を以下のように考え，構造化の効果を検証した．

① 完全一致

② 修飾性の単語を含めている（例：「軽度な梗塞」は正確）

③「の」を含む IDIOM は別々であれ一つであれ，両方とも正確（例：「脳室の拡大」の場合，「脳室」と「拡大」，あるいは「脳室の拡大」，両方とも正確）

以上，①～③を満たしたものを構造化に正解したものとし，2000 個レポートから 7393 文に対し構造化した．

その中からランダムに 100 文を抽出し，構造化の精度を人手で評価した．結果は表 2 に示す．100 文中，医療用語 IDIOM が 210 箇所存在し、検出箇所が 188 になり、その中 173 個の医療用語 IDIOM を正確に検出した。検出失敗した部分については、15 個誤検出と 37 個検出できないという結果になる。

表 2: 医療用語 IDIOM 検出結果

	Actual Positive	Actual Negative	Recall
Predicted Positive	173	37	82.38%
Predicted Negative	15	---	
Precision	92.02%		

これによって，提案した構造化手法によって precision: 92.02% と recall: 82.38% という十分な精度で構造化が実現できていることが明らかとなった．

8. 今後の展望

これまでに，医療専門用語の検出とその結果を利用した構造化を行った．この結果を利用して，これからさらに正確な医療専門用語や知識などの抽出をできるように分析を行う．また他に，自動診断や医療デジタル支援の構築，さらに医者がレポートを書く時の自動アドバイス，構造化のある文書を候補として提供することなどが期待できる．

謝辞

本研究で利用した医療データをご提供いただいた埼玉医科大学山根准教授に感謝します．

参考文献

[今井健 05] 今井健: 画像診断報告書からの所見抽出，博士論文，2005.

[荒牧 英治 17] 荒牧 英治: 医療言語処理 (自然言語処理シリーズ)，コロナ社，2017.

[谷川原綾子 16] 谷川原綾子，放射線技術学に関する用語集の日本語表記と意味記述に関する比較，日本放射線技術学会雑誌 72.3 (2016): 203-208.