画像診断レポートからの構造化データの抽出

杉本 賢人[*1], 和田 聖哉[*1], 島井 良重[*1], 山畑 飛鳥[*1],
武田 理宏[*1], 真鍋 史朗[*1], 松村 泰志[*1]
[*1] 大阪大学大学院医学系研究科 医療情報学

Extraction of Structured Information from Radiological Reports

Sugimoto Kento[*1], Wada Syoya[*1], Shimai Yoshie[*1], Yamahata Asuka[*1],
Takeda Toshihiro[*1], Manabe Shiro[*1], Matsumura Yasushi[*1]
[*1]Department of Medical Informatics,
Osaka University Graduate School of Medicine

フリーテキストで記述された画像診断レポートを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，機械学習を用いて，画像診断レポートから「部位」と「所見」に関する用語を網羅的に抽出し，それらを組み合わせて構造化データを作成した．また，二次利用を考えて，書き手の表記ゆれや同義語などの表現を 1 つの代表用語に変換する処理を行った．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単純 X 線画像のレポート 319,130 件を利用した．機械学習による用語抽出の精度は，平均の F1-score が 0.94 であり，全レポートから 1,788 件の部位表現，8,807 件の所見表現を獲得した．また，構造化処理を行うことで，824,539 レコードの構造化データを抽出した．

キーワード 自然言語処理，画像診断レポート，機械学習，情報抽出，構造化

1. はじめに

画像診断レポートには，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，臨床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．そこで，我々は二次利用に向けて，画像診断レポートから，「部位」と「所見」の組の構造化データの抽出を行った．構造化のための「部位」と「所見」に関する用語は，事前に機械学習を利用して抽出した．

2. 方法

1) 用語の抽出

レポートから「部位」と「所見」に関する用語の抽出は，Bidirectional LSTM-CNN-CRF[1]モデルを用い，Python によりシステムを構築した．このモデルは，再帰ニューラルネットをベースとしたモデルであり，文章から目的の用語を高精度で認識できることが知られている．文を形態素に分割したものをモデルに与えることで，各形態素に部位や所見などのラベル情報が出力される．ラベル情報としては，「部位」に関する用語を「部位」とその位置情報を示す「部位修飾」に分けてラベルを定義した．また，所見に関する用語は，その結語によって意味が異なるため，「所見あり，所見疑い，所見なし」を区別できるようラベルを定義した．複数の形態素を 1 つのかたまりとして認識するため，IOB2 フォーマット[2]を用いてラベルを拡張した．系列の各形態素は，Word Embedding によって，ベクトル表現にマッピングし，形態素の各文字は，Char Embedding によりベクトル表現にマッピングし，2 つのベクトルを結合して LSTM に入力した．

(1) 対象データ

2000 年から 2017 年の間に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単純 X 線画像の所見レポート 319,130 件を対象とした．

(2) 学習データの作成

全レポートから，無作為に 5,000 件を抽出し，文単位に分割後，各文を形態素単位に分割した．その後，各形態素に，部位や所見などのラベル情報を人手で付与した．

2) 構造化処理

機械学習により，抽出した「部位」と「所見」の用語を 1 つの組として構造化データを作成した．部位列は，「部位修飾」と「部位」のラベルの付い

た用語を連結したものを抽出した. 各レコードは, レポート中の「部位」と「所見」の用語の位置情報に基づいて作成した.

3) 代表用語への変換

レポート中の用語は, 1 つの概念が書き手の表記ゆれや同義語などにより, 複数の用語で表現されている. そこで, 抽出した用語とそれに対応する 1 つの概念に相当する代表用語とを紐づけた対応テーブルを事前に部位と所見別に作成し, 構造化後の用語を代表用語に変換した. 部位の対応テーブルについては, 事前に領域を定義して, 各用語を対応付けた. 所見に関しては, 事前に必要な表現を定義することが困難であったため, 用語の出現頻度に基づき集約を行った. これらの作業は, 複数の医療従事者(医師, 放射線検査技師)らと議論して行った.

3. 結果

1) 用語の抽出

500 件の評価データを用意して, 抽出精度を評価した結果を表 1 に示す.

表 1　カテゴリ別の抽出精度

カテゴリ	Precision	Recall	F1-Score
部位修飾	0.97	0.98	0.97
部位	0.95	0.95	0.95
所見(肯定)	0.94	0.93	0.93
所見(疑い)	0.88	0.90	0.89
所見(否定)	0.93	0.97	0.95
合計(平均)	(0.93)	(0.94)	(0.94)

結果から, モデルがレポート中の部位・所見表現を高精度で抽出できており, その予測精度は高いことがわかる.

2) 構造化処理

全レポート 319,130 件から,「異常所見なし」など所見が書かれていないものを除いた 273,740 件から構造化データを抽出した. 処理後, 824,539 レコードの構造化データが作成された.

3) 代表用語への変換

レポート中に出現した用語数と定義された代表用語数を表 2 に示す.

表 2　部位･所見別の代表用語数

	用語数	代表用語数
部位	1,788	21
所見	8,807	121

「部位」は事前に定義した 21 領域に各用語を変換した. いずれにも当てはまらない用語は「その他」として変換した.「所見」は出現頻度に基づき, 121 の代表用語に変換した.

4. 考察

機械学習による用語の抽出により, 高い精度で「部位」と「所見」の用語を抽出できた. しかし, 部位との区別が難しい所見表現などにおいて, いくつかの抽出誤りが見られた.

代表用語の定義に関して,「部位」表現は事前に領域を定義して整理した. 本研究では, 二次利用の用途として, 医用画像認識の学習データを想定しており, その目的に合わせて領域を定義しているが, 幅広い二次利用を考える場合, 汎用的な代表用語の定義方法の検討が必要である.

5. 結語

フリーテキストで記述された画像診断レポートから,「部位」と「所見」の組の構造化データを抽出できた.

参 考 文 献

[1] Ma X, Hovy E: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.arXiv, 2016.

[2] Erik F. Tjong Kim Sang, J. Veenstra: Representing Text Chunks, In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, 173-179, 1999