一般口演A

[OA3] 一般口演 A

2020年6月6日(土) 10:50 〜 11:30 第1会場 (Zoom)

[OA3-01] 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
Efforts to information extraction from radiology report and transformation into a structured format

*杉本 賢人[1]、和田 聖哉[1]、山畑 飛鳥[1]、小西 正三[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）

*Kento Sugimoto[1], Shoya Wada[1], Asuka Yamahata[1], Shozo Konishi[1], Toshihiro Takeda[1], Shiro Manabe[1], Yasushi Matsumura[1] （1. Department of Medical Informatics, Osaka University Graduate School of Medicine）

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

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

Efforts to information extraction from radiology report and transformation into a structured format

Kento Sugimoto, Shoya Wada, Asuka Yamahata, Shozo Konishi,
Toshihiro Takeda, Shiro Manabe, 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，放射線レポートに記載されている重要な情報の対象について検討し，機械学習を用いてそれらの情報を抽出すること試みた．また，二次利用を考えて，抽出した情報を構造化形式に変換するためのシステムを設計した．また，構造化形式に変換するため ，抽出した情報同士の関係性を分類するモデルを構築した．機械学習を利用した情報抽出の精度は，平均の F1-score が 0.95 であり，高い精度のモデルが構築できた．同様に構造化のための関係分類モデルの F1-score も 0.96 と高い分類精度を達成した．構造化後の情報は幅広い二次利用先で柔軟に加工できることを考慮し，JSON 形式に変換して保存した．

キーワード 自然言語処理，放射線レポート，機械学習，構造化

1. はじめに

放射線レポートには，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，放射線レポートを幅広い二次利用のデータソースとするため，レポートに含まれる情報について整理し，それらを構造化形式に変換することを試みた．

2. 方法

1) 情報モデルの定義

我々は，先行研究にて提案されている放射線レポートの情報モデル[1]に加えて，RadLex[2]の分類を参考にして，情報モデルの定義を検討した．我々は，「臓器や部位に関する表現（Anatomical entity）・観察物を示す表現（Imaging observation）・臓器の異常所見を示す表現（Clinical finding）・肯定／否定などの表現（Certainty descriptor）・観察物の特徴を示す表現（Characteristics descriptor）・観察物などのサイズを示す表現（Size descriptor）・観察物などの変化状態を示す表現（Change descriptor）・観察物や臓器の異常状態に関する読影医の解釈を示す表現（Interpretation）」の8つの概念を定義した．（以降，これらの概念を「エンティティ」と呼ぶ）．

2) 対象データ

本研究では，2010 年から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 画像のレポート（118,078 件）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．

3) データセットの作成

蓄積されたレポートから，無作為に 540 件を抽出して，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 名の医療従事者により実施された．アノテーターは，フリーテキスト形式のレポートを文単位に分割したものを渡し，適切な範囲に情報モデルで定義したエンティティを付与した．

4) エンティティ抽出

構造化に向けて，レポートからエンティティに関する範囲を抽出する必要がある．我々は，レポート内に含まれるエンティティ情報を深層学習モデル（BiLSTM-CRF [3]）を用いて抽出した．

5) 関係分類

抽出したエンティティを「Imaging observation, 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．モデルは，関係抽出タスクで高精度を記録している Soares ら[4]の手法を改良して構築した．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された．

6) 構造化

関係ありと分類された組について，その組から「Object-Attribute-Relation」のトリプレットを作成した．その情報を汎用的な構造データの表現記法として知られる JSON（JavaScript Object Notation）形式に変換した．

3. 結果

1) エンティティ抽出

アノテーション済の 540 件のレポートを 378 件（訓練），54 件（検証），108 件（評価）に分割した．各エンティティの F1-score を表 1 に示す．表 1 から全てのエンティティを高い精度で抽出できていることが分かる．

表 1　エンティティ別の F1-score

エンティティ	F1-score
Anatomical entity	0.941
Imaging observation	0.962
Clinical finding	0.950
Certainty descriptor	0.976
Characteristics descriptor	0.895
Size descriptor	0.986
Change descriptor	0.930
Interpretation	0.942
Total (average)	0.951

2) 構造化

データセットとして，無作為に抽出したレポートにエンティティ情報を付与した結果から，Object エンティティと Attribute エンティティの組み合わせ候補 7,504 組を作成し，関係有無のラベルを付与した．データセットは，7:1:2 の割合で訓練用，検証用，評価用に分割した．関係分類モデルの F1-score は 0.961 と高い分類精度を記録した．

4. 考察

1) エンティティ抽出

エンティティ別の結果では「Characteristics descriptor」が他より若干低いことが分かる．これは，特徴を表す表現は非常に多様であり，また，書き手により多くの表現が存在したためである．

2) 関係分類

関係分類の F1-score は 0.961 であり，これは入力テキストの Object と Attribute 候補の関係有無を正確に分類できたことを意味している．ただし，簡単な関係分類で間違えた例もあり，モデルの見直しやルールベースとの hybrid システムにより更なる正確な構造化を目指したい．

5. 結語

放射線レポートから，機械学習により二次利用に用いるための必要な情報を抽出した．また，構造化に向けて，それら情報の関係の有無を分類し，分類結果を JSON 形式に変換した．

参 考 文 献

[1] R.M.J. Ricky K. Taira, Stephen G. Soderland. Automatic Structuring of Radiology Free-Text Reports. Radiographics. 21 (2001),237–45.

[2] Langlotz CP. RadLex: A New Method for Indexing Online Educational Materials. RadioGraphics. 2006;26(6).

[3] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. In: Proceedings of NAACL-HLT 2016. 260–70

[4] L. B. Soares, N. Fitzgerald, J. Ling, and T. Kwiatkowski, Matching the Blanks: Distributional Similarity for Relation Learning, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2019. 2895–2905