一般口演 | データベース・データウェアハウス
一般口演16
データベース・サーバ構築
2021年11月20日(土) 16:30 ～ 17:30 D会場 (2号館1階212)

[3-D-2-02] 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

*杉本 賢人[1]、和田 聖哉[1]、小西 正三[1]、岡田 佳築[1]、真鍋 史朗[1]、武田 理宏[1]、松村 泰志[1] （1. 大阪大学大学院　医学系研究科）
*Kento Sugimoto[1], Shoya Wada[1], Shozo Konisi[1], Katsuki Okada[1], Shiro Manabe[1], Toshihiro Takeda[1], Yasushi Matsumura[1] （1. 大阪大学大学院　医学系研究科）
キーワード：Radiology Reports, Data Ware House, Natural Language Processing, Secondary Use

【背景】 放射線レポートには読影医が記述した診断における重要な情報が記述されており，この情報は，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．このような背景から，我々は放射線レポートを構造化された形式に変換し，それらを利活用するための基盤となるシステムの構築に取り組んできた．我々の先行研究では，機械学習を用いて，事前に定義した要素に関する用語をレポートから抽出し，それらを高い精度で構造化することに成功している．しかし，抽出した情報をどのような形で蓄積し，利活用するべきかについての設計／及び構築に関して依然として課題として残っていた．
【方法】 まず，病院に蓄積されている放射線レポートのデータベースから抽出プログラムを用いて， CSV形式でデータをエクスポートする． CSVファイルから対象のモダリティに条件を絞って， DWH内にインポートする．インポートされたレポートについて，先行研究で開発した構造化システムを用いて，情報抽出を行う．尚，現時点では格納先のテーブルとして，患者情報や画像のメタ情報にレポートの「所見・確信度」の情報を付与したテーブルと更に自由度が高い解析を可能とするため JSON形式でデータを格納したテーブルの2つを用意した．これらの一連の処理は，定期的にバッチ処理により実行され，差分追加の形でデータを蓄積していく．放射線レポートを構造化形式に変換する機能は Pythonを用いて作成し，データベースについては PostgreSQLを採用した．
【結果】 現在，大阪大学医学部附属病院の他施設 DWHサーバ上で本環境を構築し，テストフェーズまでが完了した．現在，運用に向けた準備を進めている．
【結語】 放射線レポートの幅広い二次利用に向けて，レポートの内容を構造化する変換モジュールを開発し，その運用に向けたシステムを設計・開発した．

©一般社団法人 日本医療情報学会

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

杉本 賢人[*1], 和田 聖哉[*1,2], 小西 正三[*1], 岡田 佳築[*1,2],
真鍋 史朗[*1], 武田 理宏[*1], 山本 征司[*3], 松村 泰志[*1,4]
*1 大阪大学大学院医学系研究科 医療情報学
*2 大阪大学大学院医学系研究科 変革的医療情報システム開発学寄附講座
*3 大阪大学医学部附属病院 医療情報部
*4 国立病院機構大阪医療センター

Development of a Structuring System for the Secondary Use of Radiology Reports

Kento Sugimoto[*1], Shoya Wada[*1, 2], Shozo Konishi[*1], Katsuki Okada[*1, 2],
Shiro Manabe[*1], Toshihiro Takeda[*1], Masashi Yamamoto[*3], Yasushi Matsumura[*1, 4]
*1 Department of Medical Informatics, Osaka University Graduate School of Medicine
*2 Department of Transformative System for Medical Information, Osaka University Graduate School of Medicine
*3 Division of medical informatics, Osaka University Hospital
*4 National Hospital Organization Osaka National Hospital

Abstract in English comes here.

Radiology reports include clinical information about observed structures, diagnostic possibilities, and recommendations for treatment plans. Such information is also valuable for observational study and clinical decision support. However, since radiology reports usually have a free-text format, it is challenging to extract clinical information from radiology reports. To overcome this, we have been tried to build a system that convert free-text radiology reports into a structured format. First, we export the records from the radiology report table in the local hospital database and save them in CSV format. Then, the CSV files are imported into radiology report table in the common DWH database. The records in the table are organized by a structured algorithm, and output into JSON and tabular format. To verify the practicality of the system, we estimated the processing time. The results showed that the structured algorithm took a certain amount of time to process but was not operationally problematic.

Keywords: Secondary use, Radiology Report, Natural Language Processing, Data Ware House

1. 背景

放射線レポートには読影医が記述した診断における重要な情報が記述されており、この情報は、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。しかし、放射線レポートはフリーテキストで書かれていることが多く、二次利用が困難であることが知られている。この課題の解決するためには、フリーテキストのレポートを二次利用可能な構造化されたフォーマットとして蓄積できるような基盤システムが求められるが、日本語を対象としたそのようなシステムは、我々が知る限り存在しない。

このような背景から、我々は放射線レポートを構造化された形式に変換し、それらを利活用するための基盤となるシステムの構築に取り組んできた。我々の先行研究では、機械学習を用いて、レポートに書かれている重要な情報を抽出し、それらを高い精度で構造化することに成功している[1-3]。しかし、抽出した情報をどのような形で蓄積し、利活用するべきかについてのシステム設計に関しては依然として課題として残っていた。

2. 目的

現在、我々は大阪大学医学部附属病院を中心として、各施設のデータウェアハウス（以降、ローカル DWH と略す）に蓄積されている放射線レポートを横断的に集約し、それらの二次利用に向けて、各施設に共通フォーマットのデータウェアハウス（以降、多施設共通 DWH と略す）を構築し、共通のデータ基盤を構築することを進めている。

本稿では、我々が大阪大学医学部附属病院に構築した多施設共通 DWH について、そのシステムの詳細や実用性について報告する。

3. 方法

3.1 データの取り込み

まず初めに、ローカル DWH として、大阪大学医学部附属病院の 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、CSV 形式でデータをエクスポートする。プログラムは日次バッチ処理にて実行され、抽出条件として、「レポート確定日＝バッチ処理日の前日」のレコードのみを対象とする。処理日の前日となるのは、バッチ処理が深夜 1 時頃に起動することを予定しているためである。次に、CSV ファイルから多施設共通 DWH の放射線レポートテーブルにデータをインポートする。ローカル DWH の放射線レポートテーブルは病院によってテーブルのスキーマが異なる可能性があるため、インポート時に共通のスキーマにマッピングするように処理しておく。また、ローカル DWH に蓄積されている放射線レポートには、施設独自のモダリティコードが付与されているため、インポート時、JJ1017（放射線領域における標準コード）とマッピングを行い、コードを標準化しておく。JJ1017 マスタは施設独自のモダリティコードとそれに対応する標準コードのレコードを各施設が用意して運用することを想定している。また、レコードは構造化の対象となるモダリティコ

ードのみを登録してくことで、未登録のモダリティコードの放射線レポートを構造化の対象外とする。

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。構造化処理も日次バッチ処理によって行い、エクスポートと同様に「レポート確定日＝バッチ処理日の前日」のレコードを対象とする。システムの全体図については図 1 の通り。構造化テーブルについては次項以降で説明する。

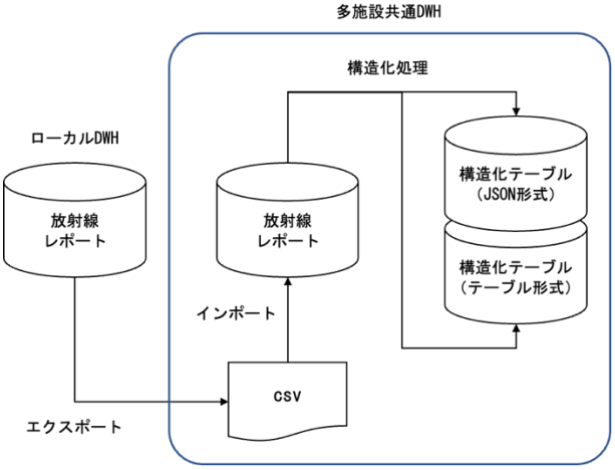

図 1 構造化システムのプロセスフロー

3.2 構造化アルゴリズム

構造化アルゴリズムは大きく「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ている。まず、エンティティ抽出処理では、放射線レポートの所見／診断フィールドに格納されているフリーテキストの文字列を入力として、事前に定義した表 1 のエンティティの用語を抽出する。

表 1　エンティティー覧

カテゴリー	名称
Object	Observation
Object	Clinical finding
Modifier	Anatomical location modifier
Modifier	Characteristics modifier
Modifier	Change modifier
Modifier	Size modifier

次に、抽出したエンティティのリレーション分類を行う。抽出した各エンティティから、Object と Modifier のペアおよび Object 同士のペアを抽出し、これらが文脈的に「リレーションがあるか」どうかを機械学習で分類する。最後に、エンティティ抽出で取り出した Object の確信度を 5 段階の順序尺度（高い順から Definite, Likely, May represent, Unlikely, Denial）で付与する。これについても、機械学習を用いて文脈から自動で付与される。これらのモジュールを介すことで、フリーテキストで記述されたレポートから「そのレポートの核となるエンティティ（Object）は何か、その確信度はどのくらいか、また、それらの修飾表現（Modifier）はどのようなものか」といった情報が構造化されて表現することが可能となる。

構造化アルゴリズムの全体図については図 2 に示した。各機能の詳細については、我々の先行研究[1-3]を参照されたい。

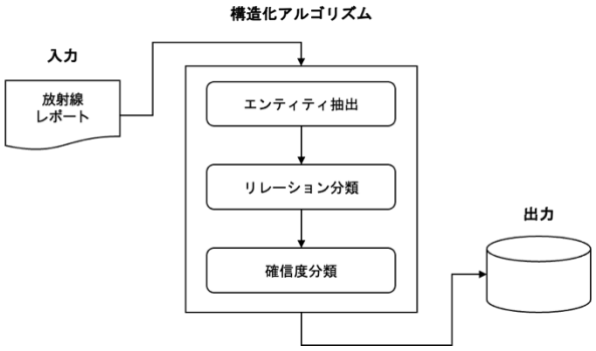

図 2 構造化アルゴリズムのシステムフロー

3.3 構造化データのフォーマット

我々は幅広い二次利用や様々なシステムとの連携を容易に行えることを目指し、構造化の処理結果を格納するテーブルを設計した。

3.3.1 JSON 形式

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。テーブルレイアウトは表 2 の通り。

表 2　構造化テーブル（JSON 形式）

No	PK	名称
1	●	患者 ID
2	●	オーダ番号
3	●	グループ番号
4	●	版
5	●	検査日
6	●	検査時刻
7		構造化結果（JSON）

テーブルのプライマリキー（PK）としては、No.1~6 までのフィールドで、ローカル DWH の放射線レポートテーブルに保持されたレコードの情報をそのまま格納する。構造化アルゴリズムの処理結果は、JSON 形式で表現し、テーブルの No.7 のフィールドに格納する。これは、PostgreSQL データベースの JSON 型を使用することで実現している。これにより、JSON でデータを保持しつつ、PostgreSQL が用意している JSON 関数や JSONPATH といったツールを使用し、JSON 形式のデータを高速に検索することが可能となる。

このテーブルは、JSON という汎用性の高い形式で構造化された結果を表現しているため、特定のアプリケーションで求められるデータのスキーマに依存しない点が特徴である。上述した JSON 関数や JSONPATH を利用して、ユーザーが直接結果を分析することも可能であるが、どちらかと言うと、目的のアプリケーションに応じて、開発者が JSON の形式から特定のスキーマに加工してデータを扱うという用途を想定している。JSON 型のデメリットとしては、通常の SQL での検索や Excel に直接エスクポートして分析するなどの従来の使い方が難し

く、JSON データの扱いに慣れていないユーザーが直接利用するのは難しい点が挙げられる。そこで、直接的な利用を容易にするため、我々は同時に従来の形式のテーブルを用意した。

3.3.2 テーブル形式

テーブル形式のスキーマを表 3 に示す。No.1~6 のフィールドとしては、JSON 形式と同様にローカル DWH の放射線レポートテーブルに保持されたレコードの情報をそのまま格納する。

表 3　構造化テーブル（表形式）

No	PK	名称
1	●	患者 ID
2	●	オーダ番号
3	●	グループ番号
4	●	版
5	●	検査日
6	●	検査時刻
7	●	オブジェクトシーケンス番号
8		オブジェクト名称
9		確信度

No.7 以降が構造化アルゴリズムの処理結果となる。No.7 は、エンティティ抽出で得た各オブジェクトをシーケンス番号で管理する。No.8 には、レポート内のオブジェクトの文字列を格納する。No.9 には確信度分類で付与したオブジェクトの確信度スケールを格納する。

このテーブルは、従来のリレーショナルデータベースのテーブル形式でデータを格納するため、JSON 形式をサポートしていない従来のシステムとの連携が容易な点、テーブル形式のデータの分析に慣れているユーザーにとって馴染みやすい点がメリットに挙げられる。例えば、検索用途として、ユーザーは「ある特定の所見を疑っている患者」といったクエリを実行することで、目的のレコードを容易に取得することができる。その後、使い慣れた Excel などの表計算ソフトで自由に分析ができる。ただし、このテーブルは上述のような検索用途に限定したスキーマであるため、検索の柔軟性に制限がある点がデメリットとして挙げられる。また、テーブルをシンプルに表現するため、Modifier 情報を格納していないため、それらを含めた情報検索はできない。

3.4 開発環境

多施設共通 DWH サーバの OS は Windows Server 2016 を使用した。費用面の観点から、サーバに GPU は搭載しておらず、処理は全て CPU でのみ行われる。放射線レポートを構造化形式に変換するプログラムは Python を用いて作成した。データベースについては JSON 型をサポートしている PostgreSQL を採用した。

4. 結果

実用性を検討するため、バッチ処理の各処理に要する時間を計測した。

データベースからのエクスポートおよびインポートについてはいずれも 10 秒以内で処理が完了した。主処理となる構造化アルゴリズムについては 1 件あたり平均 20 秒程度の処理時間を要した。構造化処理については 3 つの深層学習アルゴリズムが直列的に動いているが、いずれも CPU で処理するという制約から処理に時間を要する結果となった。

現在、構造化システムでは、「胸部 CT／腹部 CT レポート」を処理対象としているが、大阪大学医学部附属病院では、2018 年以降、平均 110 件あたりのレコードが新しく作成されている。このことから、おおよその処理時間は 20 秒×110=2200 秒（≒36 分）と見積もれる。

5. 考察

放射線レポートの二次利用に向けて、レポートの構造化システムを構築した。結果を格納するテーブルのスキーマとしては、特定のアプリケーションに依存しないユニバーサルな形式でデータを表現するため、JSON 形式のフィールドを持つテーブルを用意した。さらに、ユーザーの利用性や他システムとの連携を考えて従来のテーブル形式のものも用意した。

実用性を検討するため、バッチ処理に要する処理時間について検討した。構造化に要する処理時間は 1 日あたり 36 分程度要するが、処理自体は深夜の時間帯にバッチ処理で動くことを想定しているため、この程度の処理時間であれば概ね運用上の実用性に問題はないものと考えられる。また、構造化アルゴリズムは最新の深層学習を採用しているが、全ての処理は CPU 上で動作させることが可能であり、専用の GPU マシンを用意する必要がないことは汎用性の点からもメリットの 1 つであると考える。

ただし、実用性と運用面において検討するべき課題もまだある。実用性の点で重要となるのが、検索の実用性と柔軟性を拡張するためのシソーラスの構築である。現在の構築している構造化システムは用語の表記揺れなどは吸収できておらず、整形された品質の高いデータセットとは言い難い。この点については今後のシステム改善において優先的に取り組むべき課題であると考えている。また、運用時の各機械学習モデルの改善の仕組みについても検討する必要がある。モデルを改善していくため Human-in-the-Loop（人間参加型）のシステムを構築していく必要があると考える。

本システムは現在、大阪大学医学部附属病院でのみ構築されているが、システム自体は汎用的であり、他院にも展開可能である。今後は他院でも同様のシステムを構築し、他施設で本システムを運用していく予定である。

5. 結語

放射線レポートの二次利用に向けて、レポートの構造化システムを構築した。

謝辞

本技術開発には、2018 年 10 月 11 日から参画した戦略的イノベーション創造プログラム（SIP）「AI（人工知能）ホスピタルによる高度診断・治療システム」の研究開発資金を一部充当し、開発を推進しています。

参考文献

[1] Sugimoto K, Takeda T, Oh J-H, et al. Extracting Clinical Terms from Radiology Reports with Deep Learning, Journal of Biomedical Informatics, 2021 ; 116 : 2103729.

[2] 杉本賢人，和田聖哉，山畑飛鳥ら，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，第 24 回日本医療情報学会春季学術大会，2020.

[3] 杉本賢人，和田聖哉，小西正三ら，杉本．放射線レポートの確信度スケールの分類手法の開発，第 25 回日本医療情報学会春季学術大会，2021.