2C4-OS-21a-1

電子カルテに蓄積された臨床データと背景情報を活用し，フェノタイピングを精緻化する基盤技術開発の取り組み

Approach that uses the clinical data and its background information accumulated in EHR systems, and develops architecture that makes phenotyping more accurate

石井　雅通[*1]
Masamichi Ishii

[*1] 富士通株式会社
Fujitsu Ltd.

It is necessary to develop phenotyping algorithms that extract patients' phenotype data from clinical records written in natural language in order to achieve clinical & research support systems with Artificial Intelligence technology that use electronic medical record data as a case data base. This article addresses three points to make it: co-occurrence statistics calculated from corpus of clinical text records with their background information, clinical text editor annotating clinical metadata simultaneously, and supervised learning with continuous feedback.

1. はじめに

次世代シーケンサーの登場により遺伝子検査のコストが劇的に低減する傾向にあり、次世代医療として個別化医療や，精密な医療（precision medicine）の早期実現が求められている．そうしたゲノム医療の実現のためには遺伝子型（genotype）の情報に加えて，その表現型(phenotype)の獲得が必須であり，電子カルテに蓄積されてきた臨床情報の活用に対する期待が高まってきている．しかしながら，電子カルテから取得できる臨床情報には構造化データと非構造化データが混在しており，臨床情報の一部の活用が始まったばかりである．そのため、臨床研究にあたっては必要とされる表現型データを疾患別の患者レジストリとして再入力することが常態となっている．

臨床研究を促進するため，電子カルテに蓄積された非構造化データから臨床的に意義のある表現型データを半自動的に抽出するフェノタイピング(phenotyping)技術が求められている．精度よいフェノタイピングが実現すれば，電子カルテに日々蓄積される臨床データを巨大な症例データベースとして臨床研究に更に活用できる．

そこで我々は医学・医療にける人工知能の実現に向けて，自由記載されたカルテ記事のテキストからのフェノタイピングの精緻化をめざす．その第一歩としてカルテ記事に対するアノテーション技法を考案したので報告する．

2. フェノタイピングを精緻化するアプローチ

医療従事者によるカルテ記事の作成と同時にその内容を解析して必要なアノテーションを付加していくために，過去に電子カルテシステムに蓄積されたカルテ記事のテキスト本文に加えて，当該記事作成に関連する付帯情報（背景情報）を取得する．背景情報としては、カルテ記事の対象となった患者の属性情報（年齢，性別，生活習慣，病歴等），記載した医療従事者の属性情報（職種，所属診療科，専門領域等），カルテ記事作成時の環境情報（外来・病棟等の入力場所，記載した文書種別等）を想定している．取得したテキストに対して形態素解析を行い，用語の頻度，n-gram 単位での共起頻度等の統計情報を作成する．統計情報を利用して，カルテ記事で使用される用語について標準医学用語（シソーラスの解決含む）及び標準マスタ類（病名，医薬品，検査項目等）との対応づけ，特に医薬品に関してはジェネリック医薬品を含む、薬効や成分表記などの表現の粒度の違い及び表記ゆれを，必要に応じてカルテ記事入力者からのフィードバックを効率よく適切に獲得する．フィードバックにもとづいて非構造化データに対するアノテーションを充実させていくことで，フェノタイピング時の解析をより正確に，精緻化していくことが期待できる．更に，フィードバックにもとづく注釈付きのコーパスを教師あり学習データとして活用することで，アノテーション候補のランキング精度を持続的に向上させる仕組みとすることができる．

フェノタイピングにあたっては自然言語処理で形態素解析し，係り受け解析，意味解析等を実施することになるが，その際に医学・医療専門用語としてのターミノロジー，医学・医療オントロジー，カルテ記事の文脈に関する知識等が必要となってくる．しかしながら，ゲノム医療や最先端医療では医学や技術の進展に伴って新語が次々と発生するため，3文字略語などの類似度の高い頭語が頻出し，こうしたデータベースを保守するコストが大きい．この課題については電子カルテ入力時の日本語入力環境からリアルタイムに操作ログを取得し，解析することで電子カルテを運用しながらの未知語抽出を可能とし，抽出した未知語を新語として辞書に追加する機能を備えた．前述の仕組みによって新語登録時から背景知識と併せて統計処理されるため，従前の用語辞書登録に比して，迅速なアノテーションへの反映も期待できる．

以下に実装の手順を例示する．

2.1 準備処理

① 電子カルテ等から背景情報を抽出する．背景情報としては，患者の属性に関する情報，記載した医療従事者に関する情報，医療施設に関する情報（診療科，病棟・外来診察室の場所等）である．

② 電子カルテの構造化された臨床情報に加えて，カルテ記事等の非構造化テキストを抽出する．抽出した臨床情報の格納先は SS-MIX2 ストレージなどの標準化形式の利用が望ましい．

③ 前記①②で準備した背景情報および臨床情報を入力として形態素解析を行い品詞分解する

④ 形態素解析にあたっては医学用語辞書，MEDIS 標準マスタ等に存在する用語であれば当該用語に対する注釈を

連絡先：石井雅通，富士通株式会社，東京都大田区新蒲田 1-17-25，03-6424-6215，m.ishii@jp.fujitsu.com

付加する. 注釈には当該用語の属性情報も付加する(例えば医薬品に対する薬効コード等).

⑤ 用語ごとに TF-IDF(Term Frequency-Inverse Document Frequency)法により tf-idf 値を算出する
算出するテキストは背景情報の種別ごと, すなわち患者別, 医療従事者別, 医療施設別に集計して算出する.

2.2 注釈つき診療記録の作成

⑥ 日本語入力装置から入力中のテキストデータを取得する
⑦ 入力中データを形態素解析器等により品詞分解する.
⑧ 医学用語辞書, 標準コードマスタ類に存在しない用語の場合は, 医学用語(新語追加分)データベースに追記する.
⑨ 医学用語(新語追加分)は既存の医学用語辞書に新語として識別できる区分を付して追記する.
⑩ 現在入力中のカルテ記事の対象患者の臨床情報を取得する
⑪ 現在入力中のカルテ記事の背景情報を電子カルテより取得する
⑫ 前記⑩⑪⑦の処理結果を用語ごとに対応づけを行い, ⑤で作成した背景情報つき診療用語統計と比較可能とする
⑬入力用語に対する解釈として2つ以上の選択肢が存在する場合に, 前記⑤で作成した背景情報つき診療用語統計を利用した重み付けによりランキングする.
⑭ 特定条件を満たす重み付けスコアが複数ある場合に入力者に対して正解の選択(フィードバック)を要求する.
⑮ 前記⑭で取得した入力者からのフィードバックにもとづいて注釈つき単語情報を取得し, 注釈つき診療記録テキストを出力する.
⑯注釈つき診療記録テキストを電子カルテシステムに保存する. これにより電子カルテのテキスト記事が生成時からシソーラスや標準マスタの標準コードと対応づけされた注釈を保持した記事となる.
⑰前記⑮へ引き渡す注釈つき診療記録テキストの情報を, 背景情報と併せて, 教師あり学習データとして出力する
⑱ 背景情報つき教師あり学習データは蓄積されて
⑲背景情報つき教師あり学習データに基づき, 背景情報つき診療用語統計は改定される. これにより運用を通して重み付けが最適化される.

3. 想定するユースケースの例

同一英字の頭語を識別するケースを例示する.

〔入力データ1〕
頭語 フルスペル; 日本語表記; 関連する診療科
ASD autosensitized dermatitis;自家感作性皮膚炎; 皮膚科
ASD applicator skin distance; 装着器皮膚間隔; 皮膚科
ASD aortic septal defect; 大動脈中隔欠損; 小児心臓外科
ASD atrial septal defect; 心房中隔欠損; 循環器科

1. 入力者が入力装置で(Aortic Septal Defect の意で)ターム「ASD(Enter)」と入力する.または入力中の文の形態素解析の結果として「ASD」を抽出する.
2. 「⑥入力中データ取得部」が入力を受け取る.
3. 「⑦用語解析処理」で医学用語辞書、標準コードマスタ類から「ASD」の文字列の選択候補を抽出
4. 「⑫背景情報マッチング処理」が、「⑩入力中患者診療情報取得部」が取得した「患者臨床情報」を取得する
5. 「⑫背景情報マッチング処理」が、「⑩入力中データ背景情報取得部」が取得した「背景情報」を取得する

〔入力データ2〕
【患者臨床情報】
病名:大動脈肺動脈中隔欠損[icd10:Q21.4]、
心不全、気管支炎
症状:呼吸困難、息切れ
検査項目:心エコー(超音波)検査
投薬:利尿薬(摘要病名:心不全)

〔入力データ3〕
【背景情報:利用者】職種:医師、専門医:小児外科
【背景情報:患 者】年齢:10才、外来受診:小児心臓外科
【背景情報:環 境】入力場所:外来ブース11
[予約枠:小児心臓外科]、記載様式:プログレスノート

6. 「⑫背景情報マッチング処理」にて、〔入力データ1〕と〔入力データ2〕〔入力データ3〕のそれぞれの項目との一致度を測定する

ASD autosensitized dermatitis; 自家感作性皮膚炎
→ 一致なし

ASD applicator skin distance; 装着器皮膚間隔
→ 一致なし。

ASD aortic septal defect; 大動脈中隔欠損
→ 〔入力データ2〕病名:大動脈肺動脈中隔欠損の「大動脈」「中隔欠損」が部分マッチ。
→ 〔入力データ3〕専門医:小児外科の「小児」「外科」が部分マッチ
→〔入力データ3〕患者:外来受診:
「小児心臓外科」が完全マッチ
→〔入力データ3〕入力場所:外来ブース11、
「小児心臓外科」が完全マッチ

ASD atrial septal defect; 心房中隔欠損
→ 〔入力データ2〕病名:心不全 の「心」が部分マッチ。

7. 「⑬重み付けによるソート処理」で背景情報つき診療用語統計からマッチングした単語ごとの共起統計を取得する

アノテーションなしコーパスの共起統計
0.009「ASD」と「小児外科」
0.016「ASD」と「小児心臓外科」

アノテーションありコーパスの共起統計
0.012「ASD(aortic septal defect)」と「小児外科」
0.019「ASD(aortic septal defect)」と「小児心臓外科」
0.011「ASD(atrial septal defect)」と「心不全」

例として、完全マッチ 1.0 ポイントに対して、部分マッチ 0.3 ポイントの重み付けで共起頻度と掛け合わせたスコアを算出する. アノテーションあり共起統計とアノテーションなし共起統計のスコアは1:1で合計した.
(以下のスコアは説明のための例示)

ASD autosensitized dermatitis; 0.000
ASD applicator skin distance; 0.000
ASD aortic septal defect; 0.0413
ASD atrial septal defect; 0.0033

8. スコアを降順にソートして、注釈つき単語選択候補リストとして引き渡す

ASD aortic septal defect; 0.0413
ASD atrial septal defect; 0.0033
ASD autosensitized dermatitis; 0.000
ASD applicator skin distance; 0.000

9.「⑬フィードバック提示部」では注釈つき単語選択候補リストに基づき候補が1つであれば，自動決定モードであれば「⑮注釈つき単語選択候補リスト生成部」に引き渡す．確認モードであれば，「⑭フィードバック入力部」に引き渡す．候補が2つ以上の場合は「⑭フィードバック入力部」に引き渡す．

10.「⑭フィードバック入力部」では，ユーザから注釈つき単語選択候補リストの情報を提示してユーザから正解の選択情報を得る．
「⑬フィードバック提示部」は選択結果にもとづき「⑮注釈つき単語選択候補リスト生成部」に引き渡す．

11.「⑮注釈つき単語選択候補リスト生成部」は注釈つき診療記録テキストを生成する

注釈つき診療記録テキストの実装例として，XML形式で注釈を属性情報としてアノテーションする場合を示す．
<medicalTerm dictionary=”標準医学用語” standardTerm=” aortic septal defect”>ASD</medicalTerm>

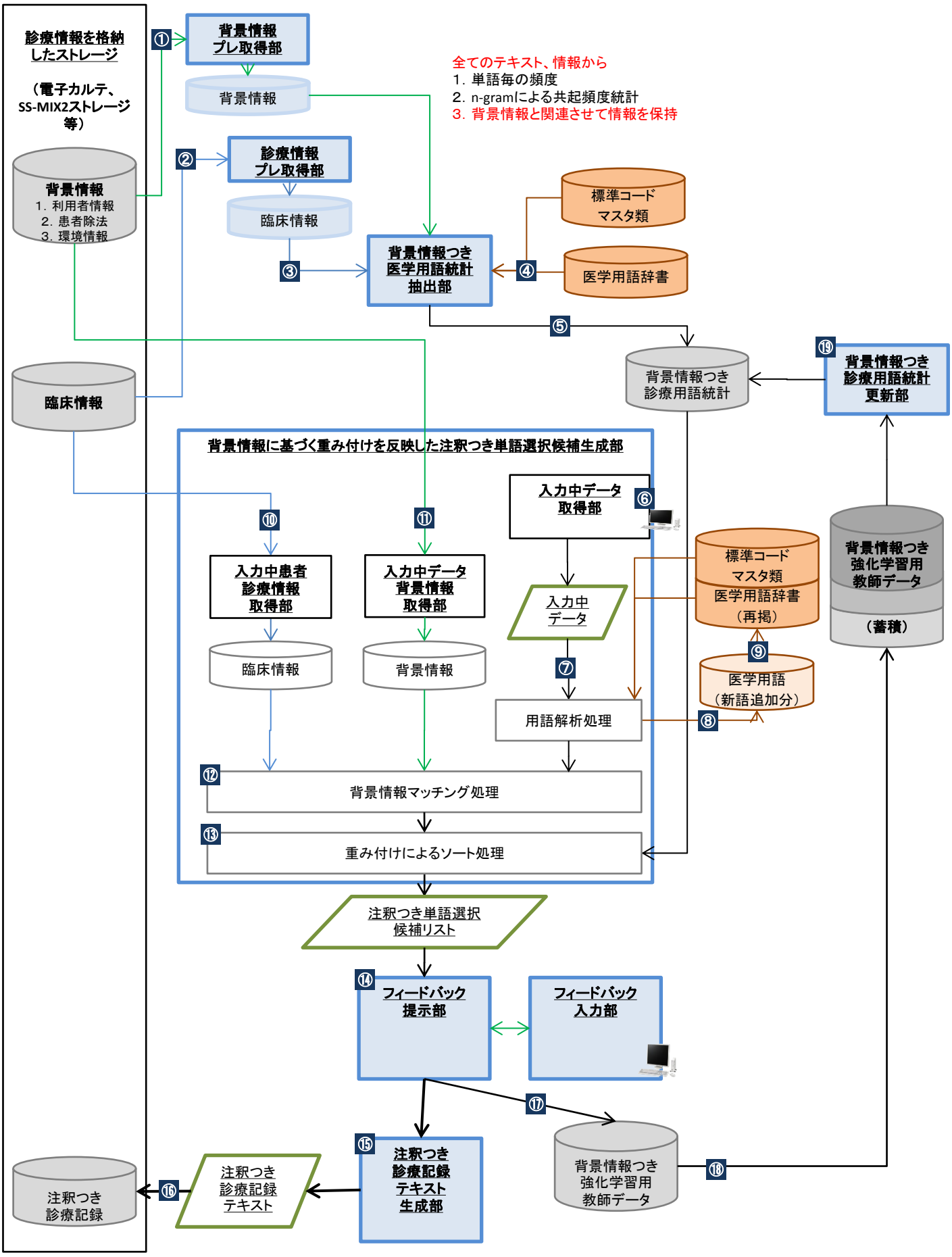
12. 注釈つき単語選択候補リストと入力者からのフィードバック（正解）は背景情報つき強化学習用教師データとして出力される

4. 今後の取り組み

本稿では，電子カルテに蓄積された臨床データを背景情報と共に活用し，フェノタイピングを精緻化する基盤技術開発の取り組みについて報告した．電子カルテへのカルテ記事記載時からアノテーションを付加することでコーパスの持続的なクレンジングを実現することを目的としている．また入力者からのフィードバックを活用することで，教師あり学習データを持続的に収集し増加させていくことができる特徴をもつ．教師あり学習データを機械学習等の技法により活用することで，あいまい語に対する正解候補提示の精度が向上することが期待できる．
今後はプロタイプシステムの構築し実証を重ねることで，フェノタイピングの実現に向けて最適な統計処理方法の開発や学習アルゴリズムの選択を進めていく予定である．

参考文献

[篠原 2015] 身体部位表現と解剖オントロジーのマッピングに関する基礎的検討，第 19 回日本医療情報学会春季学術大会抄録集，88-89p，2015


診療情報を格納したストレージ
（電子カルテ、SS-MIX2ストレージ等）
①
背景情報プレ取得部
背景情報
全てのテキスト、情報から
1. 単語毎の頻度
2. n-gramによる共起頻度統計
3. 背景情報と関連させて情報を保持
②
診療情報プレ取得部
臨床情報
背景情報
1. 利用者情報
2. 患者除法
3. 環境情報
③
背景情報つき医学用語統計抽出部
④
標準コードマスタ類
医学用語辞書
⑤
背景情報つき診療用語統計
⑲
背景情報つき診療用語統計更新部
臨床情報
背景情報に基づく重み付けを反映した注釈つき単語選択候補生成部
⑥
入力中データ取得部
⑩
入力中患者診療情報取得部
⑪
入力中データ背景情報取得部
入力中データ
⑦
用語解析処理
⑧
標準コードマスタ類
医学用語辞書（再掲）
⑨
医学用語（新語追加分）
背景情報つき強化学習用教師データ（蓄積）
臨床情報
背景情報
⑫
背景情報マッチング処理
⑬
重み付けによるソート処理
注釈つき単語選択候補リスト
⑭
フィードバック提示部
フィードバック入力部
⑰
⑮
注釈つき診療記録テキスト生成部
背景情報つき強化学習用教師データ
⑱
⑯
注釈つき診療記録テキスト
注釈つき診療記録