
一般口演B

[OB-1] 一般口演 B

新しい自然言語処理エンジンを用いた電子カルテデータ構造化と疾患別特徴語抽出

2019年6月7日(金) 09:45 ～ 10:15 第2会場 (熊本市民会館 2F 大会議室)

[OB-1] 新しい自然言語処理エンジンを用いた電子カルテデータ構造化と疾患別特徴語抽出

野口 怜 (群馬大学 医学部附属病院 システム統合センター)

新しい自然言語処理エンジンを用いた 電子カルテデータ構造化と疾患別特徴語抽出

野口 怜^{*1}, 鳥飼 幸太^{*1}, 齋藤 勇一郎^{*1}

^{*1} 群馬大学医学部附属病院 システム統合センター

Structuration of Electronic Medical Record and Extraction of Disease-Specific Feature Terms Using a Novel Natural Language Processing Engine

Rei Noguchi ^{*1}, Kota Torikai^{*1}, Yuichiro Saito^{*1}

^{*1} System Integration Center, Gunma University Hospital

抄録: 電子カルテのテキストデータは、患者に関する重要かつ一貫したデータを含み、利活用が大きく期待されているが、非構造化データのために扱いが難しくまだ十分に活用されていない。本研究では、電子カルテデータの診療活用に向け、退院サマリを対象として、辞書なしで文構造から用語切り出しを可能とする新しい自然言語処理エンジンを用いた分析を行った。1 症例 1 レコードの形式にデータを構造化した上で、各疾患を特徴付ける用語（特徴語）の抽出を行った。その結果、主訴・既往歴といった自由テキスト記述に対して、疾患分類に繋がる疾患別の特徴語が抽出できた。今後、その他テキスト記述項目への拡張や検査データなどと組み合わせることで、疾患分類モデルの構築へ繋げることが期待される。

キーワード 電子カルテ、退院サマリ、自然言語処理、データ構造化、テキストマイニング

1. はじめに

診療録などの電子カルテのテキストデータには、患者の診察結果や検査結果、治療経過などが一貫して詳述されていることから、患者の診断・治療への活用が大きく期待されている。しかしながら、テキストデータはそのままでは分析が困難なため、自然言語処理によるデータ前処理が不可欠である。従来研究もここに多くのリソースが割かれ、工数のかかる辞書整備が必要な「形態素解析」による前処理が主流であるが、前処理に工数がかかるあまり、肝心の分析がアドホックになりがちである。電子カルテデータの診療活用の実現に向けては、活用の姿を十分に想定した上で、テキストデータを最適なデータ形式に構造化し、目的に沿った形で自然言語処理を行う必要がある。本研究は、将来的に患者の間診や診察記事内の単語出現パターンから候補疾患提示などを行う疾患分類モデルの構築を想定し、その第一ステップとして、退院サマリの主訴などから各疾患を特徴付ける用語（特徴語）の出現パターン抽出の方法論構築を行う。退院サマリを 1 症例 1 レコードの形式の構造化データに変換した上で、辞書なしで文構造からテキストを「エンティティ」と呼ばれる意味単位に分割する新技術を用いて効果的に特徴語の抽出を試みる。

2. 方法

1) 対象データ

本研究では、電子カルテのテキストデータとして、比較的定型化され、可読性が高い退院サマリ（退院時に作成される患者の経過記録の要約文書）を用いた。群馬大学医学部附属病院の電子カルテに保管されている、主に循環器内科、呼吸器・アレルギー内科、消化器・肝臓内科で用いられている退院サマリ 1 年分 1,678 件（期間：2017.4.1~2018.3.31）を抽出した。

2) 前処理 1：退院サマリデータの構造化

上記の退院サマリ 1,678 件を匿名化した上で python (version 3.6.0) の分析環境 (Jupyter Notebook) に読み込み、正規表現（文字列のパターン抽出を行うための表記法）により、診断病名や主訴、現病歴、既往歴といったセクションごとにテキストを分割して 1 症例 1 レコードの形式に構造化した (Fig.1)。さらに、複数の診断病名がある場合は、1 病名ごとに分割して構造化した。

3) 前処理 2：診断病名の標準化

退院サマリに記載された診断病名は、医師による自由記載のため、表記揺れが多く、そのままでは疾患別の分析が困難である。そこで、電子カルテなどから抽出された実際の臨床現場で用いられる病名（出現形）と、国際標準の疾病分類に基づ

く病名（標準病名）とを対応付けた病名辞書「万病辞書」[1]を用いて、診断病名を標準病名に置き換えることで名寄せした。また、記載病名には、「～憎悪」「～疑い」などの接尾辞や、補足事項の記述がある場合があったため、部分一致（最長一致）を活用することで置換精度を上げた。

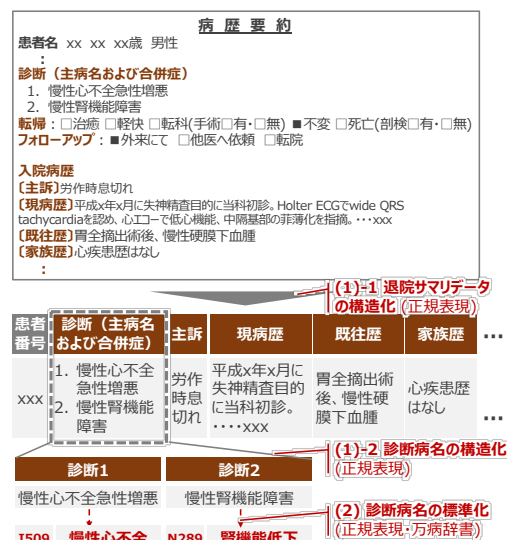


Fig.1 退院サマリ例とデータ前処理の流れ

4) 前処理 3 : エンティティの分割

疾患別の特徴語を抽出するため、構造化したデータの「主訴」と「既往歴」の記述テキストを対象に、エンティティ分割を行なった。エンティティ分割では、InterSystems 社の自然言語処理エンジン IRIS NLP [2] を用いた (Fig.2)。

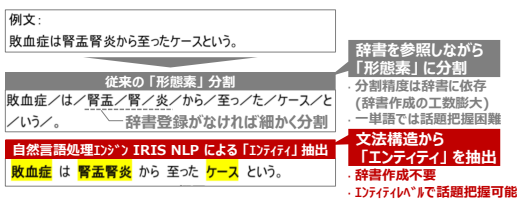


Fig.2 エンティティ分割のイメージ

3. 結果

全 1,678 件のうち、フォーマット崩れなどで標準病名にマッチしなかった 62 件を除いた計 1,616 件について、主病名で分類を行った。上位から、肺腺癌 67 件、肝細胞癌 65 件、慢性心不全 64 件、発作性心房細動 47 件、うつ血性心不全 42 件、持続性心房細動 36 件と続いた。ここでは、上位占めた心疾患の分析結果例を述べる。主訴・既往歴それぞれでエンティティの頻度を集計し、各病名における最頻出のエンティティをハイライトし

た (Table.1-2)。主訴においては、心房細動系の疾患 (ICD 分類 I48) では「動悸」が最上位だったのに対し、心不全系の疾患 (同 I50) では「呼吸困難」が最上位となった。一方、既往歴では、いずれの疾患においても高血圧に関する特徴語 (高血圧/高血圧症/ht) や糖尿病が上位に上がった。

Table.1 「主訴」の出現エンティティ集計

ICD 10	主病名 (カッコ内は退院サマリ件数)	「主訴」出現エンティティ										総計
		呼吸困難 動悸	呼吸器 苦勞	息切れ 苦勞	なし	下腿浮腫	体動困難	体重増加	呼吸時の 苦勞困難	…		
	発作性心房細動 (47)	33	1		1					…	70	
I48	持続性心房細動 (36)	10	6	9	2	4	1			…	45	
	心房細動 (21)	7	1	2	1	4				…	30	
I500	うつ血性心不全 (42)	2	19	6	4	5		2	4	1	3	60
I509	慢性心不全 (64)	2	14	18	14	6	4	3	2	4	1	100
総計		54	34	33	28	13	13	6	6	5	4	305

Table.2 「既往歴」の出現エンティティ集計

ICD 10	主病名 (カッコ内は退院サマリ件数)	「既往歴」出現エンティティ										総計
		高血圧	高血圧症	脂質異常症	糖尿病	慢性心不全	高尿酸血症	緑内障	心臓手術	心臓病	...	
	発作性心房細動 (47)	8	11	9	6	4		2	2		...	259
I48	持続性心房細動 (36)	5	5	4	5	5	2	2	1	1	...	162
	心房細動 (21)	6	4	3	1	2	1	1	2		...	100
I500	うつ血性心不全 (42)	7	2	6	5	5	4	5	3		...	303
I509	慢性心不全 (64)	14	9	8	11	13	7	4	3	3	6	529
総計		40	31	30	28	25	18	12	11	8	7	1,353

4. 考察

主訴では、「動悸」「呼吸困難」という特徴語により、ICD10 の上 3 桁レベルでの疾患分類が可能であった。既往歴における疾患名では、分類には繋がらないものの、「高血圧」や「糖尿病」は一般的な心疾患リスク要因と合致しており、本研究手法の妥当性が示唆される。一方で課題としては、既往歴に見られた高血圧/高血圧症/ht のような表記揺れが存在しており、これらも万病辞書やエンティティのベクトル化などによりグルーピングできれば、さらなる分類精度の向上が期待される。

5. 結語

新しい自然言語処理エンジンを用いて、主訴・既往歴というテキスト記述から、疾患を特徴付ける用語を得ることができた。本手法をその他のテキスト記述にも拡張し、疾患の「特徴量」として構造化データに格納し説明変数として用いることで、診断支援に向けた疾患分類モデル構築へ繋げたい。

参考文献

- [1] Ito K, et al., LREC, 2365-2369, 2018
- [2] Bronselaer G, et al., journal of intelligent systems, vol. 27, 970-993, 2012