

一般口演

一般口演1

病院情報システム1（電子カルテ・文書管理）

2018年11月23日(金) 10:15 ～ 11:45 E会場 (5F 501)

[2-E-1-2] 電子カルテにおける症状記載データ自動判別システムの構築

○富田 和恵¹, 竹村 匡正¹, 櫻井 理沙², 上村 幸司², 平松 治彦² (1.兵庫県立大学大学院応用情報科学研究科, 2.国立循環器病研究センター情報統括部)

近年、臨床研究領域において IT化が進み、データが電子的に集積されるようになっている（EDC）。医療機関においても ICTの導入が進み、診療記録を電子的に扱う電子カルテの導入が進んでいる。臨床研究で利用されるデータは、診療録記載（SOAP記載）としてフリーテキストで記載されていることが多く、SOAP記載を EDCで利用可能な形でデータ化することができれば、電子カルテおよび EDC双方の要求を満たしながら、連携が可能になると考えた。本研究では、電子カルテと EDCの連携の第1ステップとして、臨床研究において、疾患領域に関わらず収集することが要求される有害事象データを含む「症状記載」を対象として、SOAP記載から自動判別するシステム（分類器）を構築することを目的とした。対象は、国立循環器病研究センターの電子カルテの SOAP記載のうち、症状が記載されていると考えられる S、O項目の記載データ10000件とした。S、O項目の記載データに対して、形態素解析器 MeCabならびに医療単語辞書を用いて単語に分割し、形態素解析を実施した結果、医療単語を含む全ての形態素数は13856個、医療単語数は3480個であった。また、臨床試験業務経験者が S、O項目の記載データを確認し、症状記載あり（正例）に“1”、症状記載なし（負例）に“0”を付し、教師データを作成した。S、O項目に出現する単語をパラメータとして、教師あり機械学習を実施し、全形態素を対象にしたものと、医療単語を対象にしたものそれぞれに関して、正負2値分類を行った。分類器は線形サポートベクターマシンを利用し、分類器の評価方法としては、10分割交差検定法を用いた。分類器の精度は、全形態素で82.7%の正答率となり、本システムの有効性が示唆された。今後は本システムの精度を上げるための対象選定を考慮するとともに、症状記載ありと自動判別された SOAP記載データから、症状のみを抽出し、EDCに取り込むシステムの開発に繋げていくことが望まれる。

電子カルテにおける症状記載データ自動判別システムの構築

富田和恵^{*1}、竹村匡正^{*1}、櫻井理紗^{*2}、上村幸司^{*2}、平松治彦^{*2}

*1 兵庫県立大学大学院応用情報科学研究科、*2 国立循環器病研究センター情報統括部

Auto-classification system development of symptom data in electronic medical record

Kazue Tomita^{*1}, Tadamasa Takemura^{*1}, Risa Sakurai^{*2}, Koji Uemura^{*2}, Haruhiko Hiramatsu^{*2}

*1 Graduate School of Applied Informatics University of Hyogo,

*2 National Cerebral and Cardiovascular Center

In this study, as a first step of the connection between electronic medical records and EDC, I focused on "symptom description" including adverse event data, which is required to be collected irrespective of the disease area, and aimed to construct a system (classifier) that automatically classify whether the description include symptom(s) from the descriptions in SOAP. The subjects were 10000 descriptions of S and O items that are considered to have symptoms described in the SOAP description entered on the electronic medical record system by physician and nurse at National Cerebral and Cardiovascular Center.

The accuracy of the classifier was 82.7% for all words, suggesting the effectiveness of this system. And the classifier containing all words was more accurate than those which restricted subjects to medical words. That shows that the words other than medical ones were also effective for the classification. In the future, it is desirable to change target to improve the accuracy of this system. The final goal is to develop a system that extracts only symptoms from the electronic medical record data automatically and export the symptom description to EDC.

Keywords: Electronic Health Records, Adverse Drug Event, Clinical Research

1. 緒論

近年、臨床研究において IT 化が進み、紙の症例報告書・調査票 (Case Report Form: CRF) によって収集されていたデータが電子的データ取得 (Electronic Data Capture: EDC)¹⁾ によって集積されるようになってきている。EDC は、調査票回収のためにモニターが医療機関を訪問する必要がない、データセンターにおいて医療機関側で入力されたデータをリアルタイムに見ることができる (企業側が副作用等の安全性情報をリアルタイムに把握でき、迅速な対応を取ることができる)、入力時のチェック機能により初回回収時からある程度クリーニングされたデータが収集可能であること等により、症例データの固定までの時間を短縮できる等々の利点がある。また、昨今、世界同時開発を目指すグローバル試験においては、EDC の利用は必須である。一方、医療機関においても ICT の導入が進み、特に診療記録を電子的に扱う電子カルテシステム²⁾ の導入が進んでいる。

しかし、現状の EDC と電子カルテシステムの間でデータの共有は難しく、電子カルテ画面上に表示されているデータを、医療機関担当者が EDC に入力し直している。ここで、EDC で求められるデータの多くは所謂診療録記載 (SOAP 記載) としてフリーテキストで記載されている。臨床研究において疾患領域にかかわらず収集することが求められる有害事象データも、SOAP 記載における「症状記載」の内容に含まれており、これらのデータを人手で抽出することは手間がかかるのが問題である。

2. 目的

本研究では、電子カルテにおける SOAP 記載から自動的に内容を判別し、「症状記載」を自動的に抽出するシステムを構築することを目的とした。

3. 方法

3.1 対象

対象は、国立循環器病研究センターにて医師、看護師によって電子カルテシステム上に入力された SOAP 記載のうち、症状が記載されていると考えられる S、O 項目の記載データ 10000 件とした。本システムは、これら 10000 件のデータに対して、「症状記載」かどうかを人手で判定し、この判定結果を教師データとして機械学習を行い、症状記載が含まれているかどうかを自動的に判定するものである。なお、これらのデータの利用については、国立循環器病研究センター倫理委員会の審査を受けた。

3.2 実験方法

3.2.1 教師データの作成

電子カルテの S、O 項目の記載データに対して、臨床試験業務経験者によって症状記載かどうかを判定し、「症状記載あり (正例)」には「1」、「症状記載なし (負例)」には「0」を付し、教師データとした。

3.2.2 分類器の構築

次に、形態素解析³⁾を実施した。形態素解析では、医療単語辞書を搭載した形態素解析器 MeCab⁴⁾を用いて単語に分割し、学習データとした。

分類器の構築については、形態素解析器によって分かち書きされた用語に対して ID を付与し、各 SOAP 記載を ID を用いたベクトルで表現した。これらのベクトルデータおよび各 SOAP 記載の正例、負例データを用いて、サポートベクターマシン (Support Vector Machine: SVM)⁵⁾による機械学習⁶⁾

を行った。具体的には、scikit-learn⁷⁾ライブラリを用いた線形SVMを用いた分類器を構築し、分類を行った。

3.2.3 分類器の精度評価

分類器の精度評価については、10分割交差検定法⁸⁾を用いた。本研究では10分割交差検定を、各分割されたデータをテストデータ、残りのデータを学習データとして10回繰り返し、各回に算出された精度の平均をモデルの推定精度とした。また、分類実験として、全形態素を対象にしたものと、医療単語のみを対象にしたものについて評価を行った。

4 結果

4.1 臨床試験業務経験者による分類結果

臨床試験業務経験者が分類した結果は、S情報に関しては、正例（「症状記載」あり）数が967、負例（「症状記載」なし）数が3,672であり、O情報に関しては、正例数が1,285、負例数が4,076であった。全体の正例が2252、フレイが7748例であった。結果を表1に示す。

表1:分類結果

	S	O	全体
症状記載あり	967	1285	2252
症状記載なし	3672	4076	7748

4.2 形態素解析の結果

形態素解析の結果、医療単語を含む全ての形態素数は13856種類、医療単語数は3480種類であった。

4.3 分類器の性能判定

全形態素を用いた場合の10分割交差検定の結果を表2に、医療単語を用いた場合を表3に示す。「実際」は臨床試験業務経験者による症状記載の有無の判定で、「予測」は本研究で構築した分類器による症状記載の有無の判定である。また、正例数は症状記載あり、負例数は症状記載なしの数を示す。

表2:全形態素による分類結果

		予測	
		正例数	負例数
実際	正例数	816	1436
	負例数	296	7452

正解率 (Accuracy) = $(816+7452)/10000 : 82.7\%$

適合率 (Precision) = $816/(816+296) : 73.4\%$

特異度 (Specificity) = $7452/(7452+296) : 96.2\%$

偽陽性率 (False Positive Rate) = $296/(7452+296) : 3.8\%$

偽陰性率 (False Negative Rate) = $1436/(816+1436) : 63.8\%$

表 2: 医療単語による分類結果

		予測	
		正例数	負例数
実際	正例数	741	1511
	負例数	436	7312

正解率 (Accuracy) = $(741+7312)/10000 : 80.5\%$

適合率(Precision) = $741/(741+436) : 63.0\%$

特異度(Specificity) = $7312/(7312+436) : 94.4\%$

偽陽性率(False Positive Rate) = $436/(7312+436) : 5.6\%$

偽陰性率(False Negative Rate) = $1511/(741+1511) : 67.1\%$

5. 考察

全形態素を利用した場合の分類器の精度は82.7%、医療単語を用いた場合の精度は80.5%であった。ここで、対象を医療単語に限定した場合より全形態素を対象とした場合の方が精度は高く、判定には医療単語以外の単語も寄与していることが推測された。これは、例えば、SまたはO項目に症状が記載されていたとしても、それを打ち消す記載があった場合は、症状記載なし（負例）と判断される可能性が高いためであると推察される。

また、実際は「正」であるのに「負」と予測されたケースが全形態素を対象とした場合で63.8%、医療単語を対象とした場合で67.1%と高い値を示しており、基本的には症状でない、という判断をしている傾向がある。もっとも、今回の実験で用いたscikit-learnライブラリでは、デフォルトのパラメータ設定で利用しているため、より精度が上がる可能性もある。一方で、臨床研究で利用する場合には一定の精度および再現率が求められることが予測され、今後はどのような要件で利用するかについて検討する必要がある。

6. 結論

本研究では、電子カルテとEDCの連携の第1ステップとして、臨床研究において、疾患領域に関わらず収集することが要求される有害事象データを含む「症状記載」を対象として、電子カルテシステムにおけるSOAP記載から自動判別するシステムの構築を行った。教師データとしては、実際に臨床現場で記載されたSOAP記載に対して、専門者が症状記載と判定したものを教師データとし、SVMを用いて自動判別システムを作成した。結果、分類器の精度は、全形態素で82.7%、医療単語を対象としたもので80.5%の正答率であり、本システムの有効性が示唆されるとともに、カルテ記載データを臨床研究に利用できる可能性が示唆された。

参考文献

- 1) 治験効率化作業班. 治験情報のIT化に関する報告書ー治験実務の立場からの提言ー: 社団法人日本医師会治験促進センター, 2008.
[http://www.jmacct.med.or.jp/plan/files/it/ct_it_report_0811.pdf (cited 2018-Jan-08)].
- 2) 第7回標準的電子カルテ推進委員会. 電子カルテシステム構成概念(JAHISレベル3の対象範囲): 厚生労働省, 2005
[<http://www.mhlw.go.jp/shingi/2005/03/s0303-8a.html> (cited 2018-Jan-08)].
- 3) 奥村学. 形態素解析. 自然言語処理の基礎: コロナ社, 2015: 22-23.
[<http://taku910.github.io/mecab/> (cited 2018-Jan-08)]
- 4) [<http://www.sys.ci.ritsumeit.ac.jp/project/theory/svm/svm.html> (cited 2018-Jan-08)]
- 5) 株式会社システム計画研究所. Pythonによる機械学習入門: オーム社, 2016: 35-36.
[<http://scikit-learn.org/stable/> (cited 2018-Jan-08)]
- 6) [<http://musashi.osdn.jp/tutorial/mining/xtclassify/accuracy.html> (cited 2018-Jan-08)]