Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 55 – 61

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Named entity recognition over electronic health records through a combined dictionary-based approach

Alexandra Pomares Quimbaya[a*], Alejandro Sierra Múnera[b], Rafael Andrés González Rivera[a], Julián Camilo Daza Rodríguez[b], Oscar Mauricio Muñoz Velandia[a,b], Angel Alberto Garcia Peña[a,b] , Cyril Labbé[c]

[a]*Pontificia Universidad Javeriana, Cra. 7 #40-62, Bogotá 110231, Colombia*
[b]*Hospital Universitario San Ignacio, Cra. 7 #40-62, Bogotá 110231, Colombia*
[c]*Laboratoire d'Informatique de Grenoble équipe SIGMA, 700 avenue centrale,Saint-Martin-d'Hères 38400, France*

Abstract

In health care information systems, electronic health records are an important part of the knowledge concerning individual health histories. Extracting valuable knowledge from these records represents a challenging task because they are composed of data of different kind: images, test results, narrative texts that include both highly codified and a variety of notes which are diverse in language and detail, as well as ad hoc terminology, including acronyms and jargon, far from being highly codified. This paper proposes a combined approach for the recognition of named entities in such narrative texts. This approach is a composition of three different methods. The possible combinations are evaluated and the resulting composition shows an improvement of the recall and a limited impact on precision for the named entity recognition process.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Named Entity Recognition; Electronic Health Records; Text Mining

* Corresponding author. Tel.: + 57-1-3208320 ext 5291; fax: + 57-1-3208320 ext 5338
E-mail address: pomares@javeriana.edu.co

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.123

1. Introduction

Electronic health records (EHR) constitute an important resource not just for tracing single patient histories, but for population studies with clinical or administrative purposes. The nature of EHR, however, presents multiple challenges for doing so. Essentially, what we have is a typical knowledge extraction task where a combination of structured and unstructured data must be processed, including medically codified classifications, images, test results and narrative text. This paper will focus on the challenge of extracting information from narrative text contained within EHR through combined named entity recognition.

A large body of work has presented an array of methods to deal with biomedical text[1,2]. However, as pointed out in Leaman et.al[3], biomedical text is a highly codified result edited for clarity and intended at a large audience, while clinical narrative text contained in EHR is written by healthcare professionals about a single patient and is aimed at colleagues or themselves. This implies a variety of notes which are diverse in language and detail, as well as ad hoc terminology, including acronyms and jargon, far from being highly codified and standard. In addition, EHR are often filled under time pressure and with low motivation due to the fact that it takes time away from actual patient care. As a result, EHR narrative text usually suffers from low quality reflected in: variable semantics, structure without formal sentences, missing punctuation, missing expected words, misspelling or heterogeneous styles and jargon[3]. Moreover, independently of the motivation or resulting quality, the clinical language implies additional challenges, including term variability, ambiguity and complexity, lack of fine-grained classifications and data availability[4]. As such, many existing natural language processing approaches become ineffective or insufficient for it.

In this paper, we place attention particularly in named entity recognition (NER), for which specific challenges have also been identified, including which inference algorithm to use and how to use external knowledge resources (e.g. gazetteers)[5] or dealing with diverse medical fields, costly text annotations and different languages[6]. This paper, proposes an approach for dictionary-based, combined NER aimed at improving entity recall dealing with the aforementioned challenges associated to clinical narratives. To do so, we present related works in Section 2. We then go on to present the proposed strategy in Section 3. Section 4 then presents the evaluation results of applying the strategy to a standard data set. Finally, Section 5 presents some conclusions and suggests avenues for future work.

2. Related work

There have been proposed a wide variety of tools and methods to improve natural language processing of medical text. Table 1 summarizes the characteristics of a group of relevant works in NER in medical and biomedical fields.

	NER Technique	Entities	Features	Limitations
Chang 2002[11]	Pre-defined Dictionary	Acronyms and initials for health information resources, Human genome acronyms	Entity Recognition Search Abbreviations	Limited to abbreviations and acronyms
Jiang 2011[12]	Conditional Random Fields CRF	Medical Problems, Test, Treatments, Status	Entity Recognition	It needs a model training process
Jimeno 2008[13]	Exact Matching Flexile Matching	Pathologic Functions, Sign or Symptom, Cell or Molecular Dysfunctions, Findings	Entity Recognition	NER is limited to scientific texts
Aramaki 2009[15]	Conditional Random Fields CRF	Remedy, Medical Operations, Test, Examinations, Diseases, Symptoms, Medications	Entity Recognition Events Recognitions Date Times Negative Events	It needs a model training process
Zhang 2013[16]	Seed Term Collection Boundary Detection Entity Classification	Disorders, Therapeutic or Preventive Procedures, Laboratory Procedures, Clinical Drug, Test	Entity Recognition	It does not recognize typing and orthographic errors
Skeppstedt 2014[18]	Conditional Random Fields CRF Lemmatization Part-of-speech tagging	Disorders, Findings, Pharmaceutical, Body Structure	Entity Recognition NER in English and Swedish medical texts	It needs a model training process Previous annotation process developed by Physician or