International Journal of Reliable and Quality E-Healthcare
Volume 7 • Issue 1 • January-March 2018

Concept Attribute Labeling and Context-Aware Named Entity Recognition in Electronic Health Records

Alexandra Pomares-Quimbaya, Pontificia Universidad Javeriana, Bogotá, Colombia

Rafael A. Gonzalez, Pontificia Universidad Javeriana, Bogotá, Colombia

Oscar Mauricio Muñoz Velandia, Hospital Universitario San Ignacio, Pontificia Universidad Javeriana, Bogotá, Colombia

Angel Alberto Garcia Peña, Hospital Universitario San Ignacio, Pontificia Universidad Javeriana, Bogotá, Colombia

Julián Camilo Daza Rodríguez, Hospital Universitario San Ignacio, Pontificia Universidad Javeriana, Bogotá, Colombia

Alejandro Sierra Múnera, Hospital Universitario San Ignacio, Pontificia Universidad Javeriana, Bogotá, Colombia

Cyril Labbé, Laboratoire d'Informatique de Grenoble, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France

ABSTRACT

Extracting valuable knowledge from Electronic Health Records (EHR) represents a challenging task due to the presence of both structured and unstructured data, including codified fields, images and test results. Narrative text in particular contains a variety of notes which are diverse in language and detail, as well as being full of ad hoc terminology, including acronyms and jargon, which is especially challenging in non-English EHR, where there is a dearth of annotated corpora or trained case sets. This paper proposes an approach for NER and concept attribute labeling for EHR that takes into consideration the contextual words around the entity of interest to determine its sense. The approach proposes a composition method of three different NER methods, together with the analysis of the context (neighboring words) using an ensemble classification model. This contributes to disambiguate NER, as well as labeling the concept as confirmed, negated, speculative, pending or antecedent. Results show an improvement of the recall and a limited impact on precision for the NER process.

KEYWORDS

Concept Attribute Labeling, Electronic Health Records, Named Entity Recognition, Text Mining

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. In fact, physicians often complain that EHR systems are oriented towards information storage, but lack the ability to provide information extraction as well, mainly a consequence of the unstructured nature of the information (Menasalvas, Rodriguez-Gonzalez, Costumero, Ambit, & Gonzalo, 2016). Actually, the extraction task involves a combination of structured and unstructured data, including codified clinical classifications, images, test results and narrative text, among others. This paper will focus on the recognition of pre-established entities of interest in EHR extracted from clinical systems. This falls within the task of named entity recognition (NER), responsible for extracting entities and relationships between them, within a specific domain,

DOI: 10.4018/IJRQEH.2018010101

typically relying on dictionaries, ontologies and thesaurus to do so (Menasalvas et al., 2016). As such, this paper proposes an approach to NER, which is aimed at improving precision and recall by combining different NER methods.

There is a large body of work on methods and tools to process biomedical text in general (Menasalvas et al., 2016). However, as pointed out in Leaman et.al (Leaman, Khare, & Lu 2015), biomedical texts are a highly codified result edited for clarity and intended at a large audience, while clinical narrative texts contained in EHR are written by healthcare professionals about a single patient and are 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 practice, these results in EHR systems are country, hospital and even service dependent (Menasalvas et al., 2016). 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 or education. 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 (Menasalvas et al., 2016). Moreover, independently of the motivation or resulting quality, the clinical language per se implies an additional series of challenges, including term variability, ambiguity and complexity, lack of fine-grained classifications, results followed by units or dosages, incomplete syntactic components in sentences, as well as data availability (Dong, Qian, Guan, Huang, Yu, & Yang, 2016). In NER terms, ambiguity is one of the biggest challenges, because concepts of interest are frequently hypothetical, negated or include temporal relationships (Menasalvas et al., 2016). As such, many existing natural language processing (NLP) approaches become ineffective or insufficient for clinical narrative text.

Moreover, despite numerous NER proposals for EHR, the vast majority are limited to medical text written in English (Menasalvas et al., 2016). Given that NER relies on dictionaries, several are already available, including Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) or International Classification of Diseases (ICD); these are also in English with either no translations or limited versions in other languages. In the context of this paper, EHRs belong to a Spanish speaking hospital and it has already been recognized that for event extraction from EHRs in Spanish, the lack of annotated corpora is perhaps the main difficulty (Casillas, Pérez, Oronoz, Gojenola, & Santiso, 2016). Despite there being ways to avoid a language-specific annotated corpus through supervised machine learning, training the models is costly and using it in complex language sets runs into major performance issues (Dong et al., 2016). In addition, the choice of inference algorithms and managing heterogeneous medical fields further complicates medical NER (Casillas et al., 2016; Dong et al., 2016).

This paper proposes an approach for dictionary-based, combined NER aimed at improving entity recall dealing with the aforementioned challenges associated to clinical narratives extracted from EHR systems. To do so, it presents a proposal that combines three different NER techniques and takes into consideration the contextual words around the entity of interest to determine whether the concept is confirmed, negated, speculative, pending or an antecedent. To our best knowledge, our approach is the first including this level of detail in the expression of sense for the healthcare domain in Spanish. The rest of this paper continues with related works. We then go on to present the proposed strategy and present the evaluation results of applying the strategy to a standard data set. Finally, we present some conclusions and suggests avenues for future work.

BACKGROUND

Natural language processing of medical text has a wide variety of tools and methods. For our research work, we made an analysis of some relevant work in NER in the medical and biomedical fields in Table 1, where we proceed to highlight the following characteristics: