Journal of Biomedical Informatics 72 (2017) 8–22

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

DrugSemantics: A corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics

Isabel Moreno [a], Ester Boldrini [a], Paloma Moreda [a,*], M. Teresa Romá-Ferri [b]

[a] Department of Software and Computing Systems, University of Alicante, Alicante, Spain
[b] Department of Nursing, University of Alicante, Alicante, Spain

ARTICLE INFO

Article history:
Received 25 January 2017
Revised 24 April 2017
Accepted 12 June 2017
Available online 15 June 2017

Keywords:
Corpus
Reliability
Precision
Named Entity Recognition
Spanish
Summary of Product Characteristics

ABSTRACT

For the healthcare sector, it is critical to exploit the vast amount of textual health-related information. Nevertheless, healthcare providers have difficulties to benefit from such quantity of data during pharmacotherapeutic care. The problem is that such information is stored in different sources and their consultation time is limited. In this context, Natural Language Processing techniques can be applied to efficiently transform textual data into structured information so that it could be used in critical healthcare applications, being of help for physicians in their daily workload, such as: decision support systems, cohort identification, patient management, etc. Any development of these techniques requires annotated corpora. However, there is a lack of such resources in this domain and, in most cases, the few ones available concern English.

This paper presents the definition and creation of DrugSemantics corpus, a collection of Summaries of Product Characteristics in Spanish. It was manually annotated with pharmacotherapeutic named entities, detailed in DrugSemantics annotation scheme. Annotators were a Registered Nurse (RN) and two students from the Degree in Nursing. The quality of DrugSemantics corpus has been assessed by measuring its annotation reliability (overall $F = 79.33\%$ [95%CI: 78.35–80.31]), as well as its annotation precision (overall $P = 94.65\%$ [95%CI: 94.11–95.19]). Besides, the gold-standard construction process is described in detail. In total, our corpus contains more than 2000 named entities, 780 sentences and 226,729 tokens. Last, a Named Entity Classification module trained on DrugSemantics is presented aiming at showing the quality of our corpus, as well as an example of how to use it.

1. Introduction

Nowadays, there is a large amount of information on health and healthcare [1]. Examples of this huge quantity of information available are PubMed [2], a repository that comprises more than 25 million documents on biomedical literature, or the information stored for each patient on its own Electronic Health Record (EHR) during day-to-day care. Due to the high value of such data, exploiting this textual information is critical to: (i) improve healthcare quality; (ii) drive medical innovation research; and (iii) reduce healthcare costs [1]. Nevertheless, healthcare providers have difficulties to use such quantity of information during their professional practice mainly due to two reasons. On the one hand, they have a limited consultation time (i.e. often less than 10 min). On the other hand, the required information by them is stored in many and different sources [3].

Envision yourself in a primary health care consultation. A general physician attends to a patient that shows several health problems, for instance: overweight, diabetes and hypercholesterolemia. This patient is being monitored with diet, exercise and various medications but, during check-ups monitoring, he/she is not improving. The physician needs to know whether the negative evolution of his/her weight and his/her cholesterol are related or not to the medications employed for his/her treatment (an oral hypoglycemic and a lipid-regulating agents).

Before reaching a conclusion, the physician should analyze a wide range of specialized documents of different sizes and sources. The most relevant ones are: (i) patient EHR, accessible through many and different applications; (ii) Summaries of medicinal Product Characteristics (SPC) or package leaflets for the patients medications, available on medicines agencies web sites at

* Corresponding author at: University of Alicante, Apdo. de correos 99, E-03080 Alicante, Spain.
E-mail addresses: imoreno@dlsi.ua.es (I. Moreno), eboldrini@dlsi.ua.es (E. Boldrini), moreda@dlsi.ua.es (P. Moreda), mtr.ferri@ua.es (M.T. Romá-Ferri).

http://dx.doi.org/10.1016/j.jbi.2017.06.013
1532-0464/

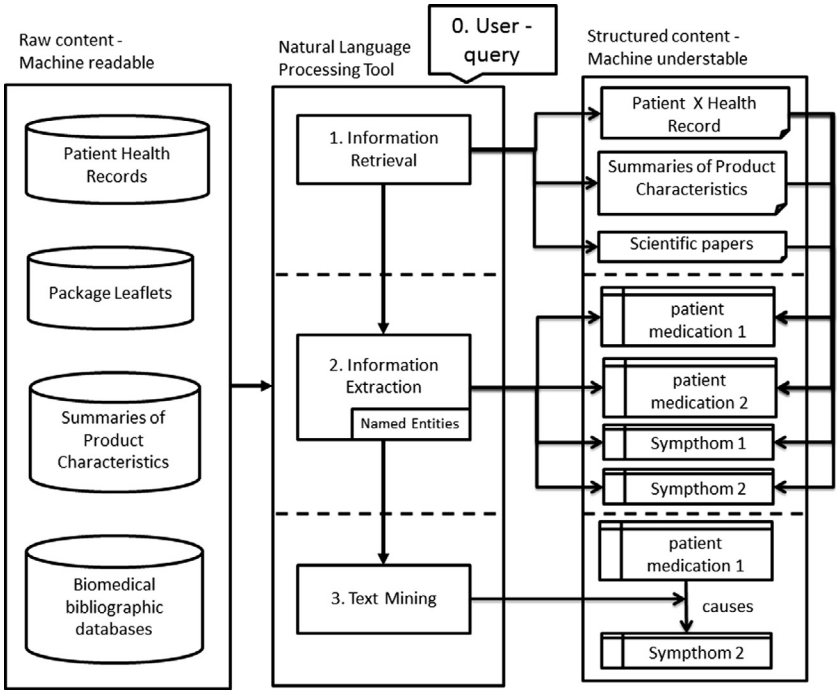

Fig. 1. Example of an hypothetical tool (center box) to allow pharmacoterapeutic monitoring using Natural Language Processing techniques, as in the example described in Section 1 Introduction. Its input is raw content (left box) and its output is structured content (right box).

international[1] or national[2] levels; and (iii) scientific papers indexed in biomedical bibliographic databases, such us MEDLINE [2],[3] or Scopus [4].[4] For healthcare providers, the analysis of all the information contained in every information source is unmanageable [5–7]. Thus they would need a tool that displays, at a glance, every document relevant to the patient condition with a single query based on their information needs [8].

Natural Language Processing (NLP) is a field of research that addresses the obstacles mentioned above. Its aim is to provide mechanisms to transform unstructured textual information, easy to understand for humans, into structured data that can be exploited by computer processes for different purposes [1,9]. So, NLP techniques can be employed to achieve the aforementioned tool for healthcare providers. Fig. 1 illustrates how to solve this problem using NLP: first, Information Retrieval (IR) techniques could be applied. In this way, relevant documents, that satisfy an information need, could be found from a large collection of documents [10, p. 1]. In our pharmacotherapeutic monitoring example, relevant documents would be the patient EHR together with SPCs and scientific papers related to the patient's condition. Afterwards, Information Extraction (IE) techniques could be employed. In this manner, textual and explicit relevant information could be extracted from the retrieved documents in the previous step [11, pp. 94–95][12, pp. 814–815]. In our example, the relevant information could be all the medication that a patient is currently taking and his symptoms (from the EHR) as well as signs commonly associated to these medications (from SPCs and scientific papers), among other relevant information. Then, Text Mining (TM) techniques could be used. This area is in charge of finding information that is not specified explicitly in the document, therefore, further inference is needed [13]. In the case we are dealing with, it would be to discover a reason for the negative evolution of weight and cholesterol level. That is, whether the patients current medication, all of which have been extracted previously from explicit information, is interacting with each other or not. Finally, all the obtained information (both explicit and implicit) would be displayed in an organized and summarized manner to the physician, in order to facilitate reaching a conclusion.

Building such tool for healthcare providers is not a trivial task. This is because IR is a mature area [5,14] where several IR systems have been developed to retrieve documents that satisfies an user's information need. Some examples are PubMed [2] (professional level) or Google [15] (wide variety of users). However, there is still plenty of work to do, in both IE and TM techniques, to reach suitable results for many and different user profiles [16].

Progress in any of these techniques relies heavily on annotated corpora. This is due to the fact that these resources have mainly two purposes: (i) development – to assist during the creation of rules and statistical models that will control the behavior of a system; and (ii) evaluation – to provide reference data against which to assess the performance of a system. Nevertheless, annotated corpora for the health domain present two main barriers.

On the one hand, there is a limited number of annotated corpora [17] and existing ones do not consider all relevant information for pharmacotherapeutic care, as Section 2 will show. Therefore, the goal of our research is the construction of DrugSemantics, a pharmacotherapeutic corpus to tackle a part of the IE problem. This resource contains annotations of Named Entities (NE) relevant to the pharmacotherapeutic care. A NE represents a mention of a semantic category in a text [11,18]. In this field, these NEs categories refer to important information for the prescription and monitoring processes of pharmaceutical products [3,19] and relates to concepts such as medicines[5] or clinical conditions.[6]

[1] European Medicines Agency has more than 937 authorized medications – January 2017 (http://www.ema.europa.eu/).

[2] Spanish Agency of Medicines and Medical Devices contains more than 13,500 medications on the market – January 2017 (http://www.aemps.gob.es/).

[3] MEDLINE comprises more than 26 million citations (January 2017).

[4] Scopus includes more than 60 million records from journals and books (January 2017).

[5] For example: trade names of medicines ("*Conacetol*®") or active substances ("*Paracetamol*" – acetaminophen in English).

[6] For instance: therapeutic indications, contraindications or inter-current illness.