Extraction of Clinical Information from Clinical Reports: an Application to the Study of Medication Overuse Headaches in Italy.

Cristiana Larizza, Matteo Gabetta, Lina Maria Rojas Barahona, Giusseppe Milani, Elena Guaschino, Grazia Sances, Cristina Cereda, Riccardo Bellazzi

Extraction of Clinical Information from Clinical Reports: an Application to the Study of Medication Overuse Headaches in Italy.

Cristiana Larizza, PhD[1], Matteo Gabetta, MD[1], Lina Maria Rojas Barahona, PhD[1], Giuseppe Milani, MD[1], Elena Guaschino, MD[2], Grazia Sances, MD[2], Cristina Cereda, PhD[2], Riccardo Bellazzi, PhD[1]

[1]Dipartimento di Informatica e Sistemistica, Università di Pavia, Pavia, Italy
[2]Institute of Neurology, IRCCS C. Mondino Foundation, Pavia, Italy

Abstract

A i2b2-Pavia pilot project has been recently activated at the Headache Centre of the C. Mondino Institute of Neurology, in Pavia, with the aim of investigating Medical Overuse Headaches. The software infrastructure so far implemented automatically extracts and integrates data coming from different sources into a repository purposely designed for multidimensional inspection. A great effort has been devoted to train a Natural Language Processing system able to extract medical concepts from Italian clinical reports.

Introduction

One of the most common neurologic disorders is headache, that frequently evolves into Medical Overuse Headache (MOH), the major avoidable cause of headache disability in the developed world (Pascual *et al.*, 2001[1]). Within the i2b2 project (Informatics for Integrating Biology and the Bedside, https://www.i2b2.org/[2]) the Laboratory of Bio-Medical Informatics of the University of Pavia has recently undertaken the implementation of a software infrastructure able to 1) retrieve and anonymise clinical reports stored into legacy databases; 2) automatically parse and analyse them with the purpose of extracting and encoding the detailed diagnoses into the International Headache Classification (ICHD-2); 3) find which primitive headaches evolved into MOH; 4) extract the medications prescribed to the patients and, finally 5) store the extracted data into a multidimensional database where also structured data coming from spreadsheet files have been saved. A great effort has been devoted to train a Natural Language Processing (NLP) system able to analyse Italian texts containing medical terms and extract detailed information about the evolution of the headache.

Methods

The overall procedure for extracting relevant data on headache patients from unstructured reports is centered on a NLP based tool able to analyze Italian clinical reports. It consists of a pipeline of modules that parse unstructured text and extract specific medical concepts (diagnoses and medicaments). In particular, it includes several text mining modules rely on some form of lexical and syntactic analysis accurately configured to work on Italian medical texts (sectionizer, tokenizer, POS tagger, chunker) and some specific modules purposely developed to automatically extract and encode the diagnoses into the International Headache Classification (ICHD-2); extract the medications overused and the medications suggested to the patients, and, finally, detect the primary headaches associated with MOH, that is which kind of headache evolved into MOH. This information is in fact crucial to better characterize and study such disorders. Moreover, the system stores the data extracted from the reports into a database purposely designed for further inspection and analysis through appropriate exploration tools.

Results

We have currently evaluated the NLP modules submitting 149 annotated reports related to MOH to the physicians. The results so far obtained are satisfactory and show the potential utility of such system to recover automatically information relevant for studying headaches. As further step we plan to retrieve a greater number of reports, create a reference gold standard and start a deeper and more extensive evaluation of the NLP tool.

References

1. Pascual J., Colas R., Castillo J. Epidemiology of chronic daily headache. Curr. Pain Headache Rep., 2001, **5** (6) : 529-536
2. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, Chueh HC. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc. 2007 Oct 11:548-52.