Extracting Diagnoses and Drug-Abuse Patterns from Italian Clinical Reports of Patients with Headache Disorders

Matteo Gabetta[a], Cristiana Larizza[a], Lina Rojas Barahona[a], Elena Guaschino[b], Grazia Sances[b], Cristina Cereda[b], Riccardo Bellazzi[a]

[a] *Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy*
[b] *Headache Unit, IRCCS C. Mondino Foundation, Institute of Neurology, Pavia, Italy*

Abstract

The ever increasing interest in extracting valuable information from heterogeneous unstructured biomedical texts has yielded important contributions. However, most of the advances concern processing English reports and literature; this growth has not been equal for other languages, including Italian. This work is focused on processing discharge summaries, written in narrative Italian, related to patients admitted into a Headache Unit with the aim of extracting the type of discharge headache and, in case of a diagnosis of Medication Overuse Headache, also the original headache type leading to it. The evaluation of the system has given satisfactory results encouraging the application of these techniques for extracting automatically other relevant clinical information not available in structured form.

Keywords:

Text mining, Headache disorders, Natural language processing

Introduction

One of the most common neurologic disorders is headache that can lead to *Medication Overuse Headache* (MOH) for patients taking frequently headache relief drugs. Within the i2b2 project, founded by the NIH, the Laboratory of Bio-Medical Informatics (BMI) of the University of Pavia has undertaken the implementation of a software infrastructure with the aim of supporting the clinical research at the Headache Unit of the Neurological Institute, IRCCS C. Mondino at Pavia, Italy, in order to obtain an automatic and accurate mapping of free-text reports onto structured diagnosis information. The first pilot study is focused on the specific task of automatically extracting relevant clinical data related to MOH. Within this Information Extraction process, a relevant effort has been made in order to setup a software infrastructure for processing Italian clinical reports and to develop custom modules able to extract complex diagnoses.

Methods

For processing the reports we use the GATE framework, on which we implemented an Information Extraction pipeline inspired by the one available within the i2b2 project, that underwent substantial modifications and extensions to cope with our specific task.

The components included in the pipeline are: the *section splitter* (aimed at identifying thematic sections within the document), the *text tokenizer* (separating its atomic elements), the *POS-tagger* (assigning a specific POS-Tag to each element) and the *NP-chunker* (pointing out specific syntactical structures); through these stages the document is set up to be processed by the last module, the *diagnosis extractor*, which, relying on ICHD-2 classification, actually extracts diagnosis-concepts from the text.

Results

The test of the system has been performed on a set of discharge reports that were submitted to a physician who manually annotated both principal diagnosis and, when available, also the initial headache disorders evolved into it, in order to provide a *"gold standard"*. To assess the system performance we performed a standard "*precision-recall"* evaluation on all the documents for two different tasks, one considering the single annotation correctness and the other considering the whole documents. The results for the first task are: precision 0.9675 and recall 0.9425. The second task achieved: precision 0.8767 and recall 0.8707.

Conclusion

The purpose of our work is to exploit the vast amount of unstructured clinical information usually available in a Hospital Information System. The relevance of the project is justified by the emerging translational research in which it is required to plan genetic studies over large populations of individuals sharing specific phenotypical characteristics that can be collected also with systems like the one here presented. In this work we solved important issues involving not only the application of NLP techniques in the MOH domain, but also the processing of Italian language, efforts not completed by other groups to our knowledge. Future directions involve refining of the diagnosis extraction system as well as implementing a module for the extraction of the therapies prescribed to the patient.