Journal of Biomedical Informatics 63 (2016) 22–32

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes

Emma Chiaramello[a], Francesco Pinciroli[a,b], Alberico Bonalumi[c], Angelo Caroli[c], Gabriella Tognola[a,*]

[a] Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (IEIIT), Consiglio Nazionale delle Ricerche (CNR), Piazza L. da Vinci, 32, 20133 Milano, Italy
[b] e-HealthLAB, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Piazza L. da Vinci, 32, 20133 Milan, Italy
[c] UOC Sistemi Informativi e Informatici, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico Milano, Via Francesco Sforza, 35, 20122 Milano, Italy

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 12 July 2016
Accepted 17 July 2016
Available online 18 July 2016

Keywords:
Information extraction
Unstructured clinical notes
Italian language
MetaMap
Failure analysis
EHR data reuse

ABSTRACT

Information extraction from narrative clinical notes is useful for patient care, as well as for secondary use of medical data, for research or clinical purposes. Many studies focused on information extraction from English clinical texts, but less dealt with clinical notes in languages other than English.

This study tested the feasibility of using "off the shelf" information extraction algorithms to identify medical concepts from Italian clinical notes. Among all the available and well-established information extraction algorithms, we used MetaMap to map medical concepts to the Unified Medical Language System (UMLS). The study addressed two questions: (Q1) to understand if it would be possible to properly map medical terms found in clinical notes and related to the semantic group of "Disorders" to the Italian UMLS resources; (Q2) to investigate if it would be feasible to use MetaMap as it is to extract these medical concepts from Italian clinical notes.

We performed three experiments: in EXP1, we investigated how many medical concepts of the "Disorders" semantic group found in a set of clinical notes written in Italian could be mapped to the UMLS Italian medical sources; in EXP2 we assessed how the different processing steps used by MetaMap, which are English dependent, could be used in Italian texts to map the original clinical notes on the Italian UMLS sources; in EXP3 we automatically translated the clinical notes from Italian to English using Google Translator, and then we used MetaMap to map the translated texts.

Results in EXP1 showed that the Italian UMLS Metathesaurus sources covered 91% of the medical terms of the "Disorders" semantic group, as found in the studied dataset. We observed that even if MetaMap was built to analyze texts written in English, most of its processing steps worked properly also with texts written in Italian. MetaMap identified correctly about half of the concepts in the Italian clinical notes. Using MetaMap's annotation on Italian clinical notes instead of a simple text search improved our results of about 15 percentage points. MetaMap's annotation of Italian clinical notes showed recall, precision and F-measure equal to 0.53, 0.98 and 0.69, respectively. Most of the failures were due to the impossibility for MetaMap to generate meaningful variants for the Italian language, suggesting that modifying MetaMap to allow generating Italian variants could improve the performance. MetaMap's performance in annotating automatically translated English clinical notes was in line with findings in the literature, with similar recall (0.75), F-measure (0.83) and even higher precision (0.95). Most of the failures were due to a bad Italian to English translation of medical terms, suggesting that using an automatic translation tool specialized in translating medical concepts might be useful to obtain better performances.

In conclusion, performances obtained using MetaMap on the fully automatic translation of the Italian text are good enough to allow to use MetaMap "as it is" in clinical practice.

1. Introduction

In the last few years, the development of tools to extract information from clinical digital documents has become one of the most interesting research areas in the field of medical information tech-

* Corresponding author.
E-mail address: gabriella.tognola@ieiit.cnr.it (G. Tognola).

http://dx.doi.org/10.1016/j.jbi.2016.07.017
1532-0464/© 2016 Elsevier Inc. All rights reserved.

nology, as patient information has the potential to make a significant impact in many aspects of healthcare and biomedical research [1]. Recent works have demonstrated this potential for individual patient care (such as with clinical decision support systems [2,3]), to improve health literacy [4], for public health (e.g. in bio-surveillance [5]), and in biomedical research (e.g. in cohort identification [6,7], in the discovery of unknown correlations between diseases [8,9] or in pharmacovigilance [10,11]).

Although portion of patient information exists as structured data, a huge part of clinical data is still in the narrative form. Unstructured texts allow clinicians to easily report on various types of information about patients, i.e. symptoms, pathologies, medical histories, findings, medical treatments and drugs, and to make hypotheses, diagnoses, prescriptions, and suggestions using a free-form structure. Clinical notes can be very heterogeneous, with flexible formatting, short and telegraphic phrases and typically contain abbreviations, acronyms, misspellings and terms specific to the particular medical context [1,12]. This narrative text is adopted in many different clinical settings, e.g. in ambulatory clinical practice, and presents several difficulties to be analyzed for automated information extraction and for secondary use for research or clinical purposes.

The first task in automated information extraction from clinical notes is the semantic annotation of relevant concepts in the text, i.e. the detection of noun phrases and their classification according to the semantic categories of interest [13,14]. Semantic annotation requires two components: a linguistic processing tool to detect noun phrases and a knowledge source to classify the identified concepts in the proper semantic category. The goodness of the annotation depends both on the efficiency of the linguistic tool and the quality and coverage of the knowledge sources [15,16].

Natural Language Processing (NLP) methods are those linguistic processing tools that are needed to detect noun phrases from unstructured text. Many NLP tools were originally developed to extract information from biomedical literature. In the last few years these tools have been largely adapted and used in the analysis of clinical notes [1,12]. Among the earliest attempts to develop NLP applications for the medical domain, the Linguistic String Project-Medical Language Processor (LSP-MLP) developed in 1986 [17] and the Medical Language Extraction and Encoding System (MedLEE) [18,19] are the most known examples. A more recent NLP tool by the National Library of Medicine (NLM) – MetaMap [20] – was developed to extract medical concepts from biomedical literature. In the last decade MetaMap has been also used to analyze narrative texts different from scientific papers, such as e-mails [21], patients-reports [22] and clinical notes (e.g. to identify specific respiratory conditions [5], epilepsy [6], to extract clinical problems for inclusion in the patient's electronic problem list [23,24], and to discover adverse drug reactions [25]). MetaMap uses computational linguistic techniques to identify words and phrases in English free text and map them to the Unified Medical Language System (UMLS), a knowledge source developed by the US National Library of Medicine as a starting point for the development of biomedical language processing algorithms [26].

Recently, efforts were made to develop new and *ad hoc* information extraction algorithms based on linguistic processing tools for the identification of medical concepts in biomedical texts written in languages different from English, such as for the Spanish [27], French [28,29], Portuguese [30] and Swedish [14] languages. Only a few studies focused on using already existing, or "off the shelf", linguistic algorithms to process non-English texts [15,31,32].

The goal of the present study was to understand if using a linguistic tool developed to process English text, such as MetaMap, could be suitable to extract medical concepts from clinical texts written in other languages, in particular in Italian. In the view to assess the ease of use of an annotation tool in clinical practice, our idea was not to develop a new "Italian MetaMap", but to slightly adapt the existing tool. To the best of our knowledge, no previous works focused on the identification of medical concepts from Italian clinical notes using MetaMap.

Dealing with Italian language is a challenging task, as Italian language is more inflective than English and is characterized by a higher number of compounds of words, particularly in the medical lexicon. In addition, the coverage of medical terminology resources in Italian is lower than in English.

The present study addressed two specific research questions:

(Q1) are the medical terms found in Italian clinical notes and related to the "Disorders" semantic group properly mapped in the Italian UMLS terminological resources?
(Q2) is it feasible to use MetaMap "as it is" to identify medical concepts related to the "Disorders" semantic group from Italian clinical notes?

To answer to Q1, we carried out experiment "EXP1" in which we investigated how many medical concepts in a set of unstructured clinical notes written in Italian could be mapped to the UMLS Italian medical sources.

To answer to Q2, we followed two strategies. The first one was to test the feasibility of using MetaMap to directly analyze Italian text. The first strategy was tested in experiment "EXP2" in which (i) we first assessed how the different processing steps used by MetaMap, which are English dependent, could be used for other languages, such as Italian and then (ii) we run MetaMap to map the original Italian texts using a modified version of the UMLS knowledge source consisting in the Italian UMLS Metathesaurus sources only.

In the second strategy we followed to address question Q2, we assessed the feasibility of using MetaMap to analyze an English translated version of the original Italian text. We carried the experiment "EXP3" out, in which we automatically translated the clinical notes from Italian to English using Google Translator, and then we used MetaMap with its default knowledge source to annotate the translated texts. The approach of combining Google Translator and MetaMap seemed feasible to be applied in the context of extracting medical concepts in a clinical setting for several reasons: it is based on open source tools, it has low computational impact so that it can be used in clinical practice, and it does not require high expertise or training skills. Last but not least, the combination of Google Translator and MetaMap to identify medical concepts in texts written in a language different from English was described earlier in [31,32] to be a promising method. It is to note that our study differed from [31,32] because we tested the approach on a different language (Italian in our study, Spanish in [31,32]) and on a different type of text (we used clinical texts while in [31,32] they analyzed articles from biomedical literature), so it would be interesting to compare if our findings would be different from those reported by the Spanish researchers as a result of the different language and the different type of documents.

2. Materials and methods

2.1. MetaMap processing steps

MetaMap is based on six processing steps: (1) tokenization, (2) parsing, (3) variant generation, (4) candidate retrieval, (5) candidate evaluation, and (6) mapping construction.

As shown in Table 1, some processing steps are based on rules and lexical resources built on English grammar and are therefore strictly language specific, while others use algorithms that are