Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Information extraction from Italian medical reports: An ontology-driven approach

Natalia Viani[a,*], Cristiana Larizza[a], Valentina Tibollo[b], Carlo Napolitano[b], Silvia G. Priori[b,c], Riccardo Bellazzi[a,b], Lucia Sacchi[a]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100, Pavia, PV, Italy
[b] IRCCS Istituti Clinici Scientifici Maugeri, Via Salvatore Maugeri 10, 27100, Pavia, PV, Italy
[c] Department of Molecular Medicine, University of Pavia, Via Forlanini, 27100, Pavia, PV, Italy

ARTICLE INFO

Keywords:
Information extraction
Natural language processing

ABSTRACT

Objective: In this work, we propose an ontology-driven approach to identify events and their attributes from episodes of care included in medical reports written in Italian. For this language, shared resources for clinical information extraction are not easily accessible.

Materials and methods: The corpus considered in this work includes 5432 non-annotated medical reports belonging to patients with rare arrhythmias. To guide the information extraction process, we built a domain-specific ontology that includes the events and the attributes to be extracted, with related regular expressions. The ontology and the annotation system were constructed on a development set, while the performance was evaluated on an independent test set. As a gold standard, we considered a manually curated hospital database named TRIAD, which stores most of the information written in reports.

Results: The proposed approach performs well on the considered Italian medical corpus, with a percentage of correct annotations above 90% for most considered clinical events. We also assessed the possibility to adapt the system to the analysis of another language (i.e., English), with promising results.

Discussion and conclusion: Our annotation system relies on a domain ontology to extract and link information in clinical text. We developed an ontology that can be easily enriched and translated, and the system performs well on the considered task. In the future, it could be successfully used to automatically populate the TRIAD database.

1. Introduction

Textual reports written during clinical practice represent a great source of clinical knowledge. To help physicians access this knowledge and raise their awareness about the importance of this information to improve clinical decisions, the development of systems that automatically extract relevant information from clinical narratives is essential [1,2].

Natural language processing (NLP) methods have been successfully applied to the analysis of English clinical texts [3]. However, advances in other languages have been limited by the lack or poor coverage of resources [4]. In this work, we address the problem of developing NLP techniques that could be applied to the analysis of medical reports written in Italian.

In the Italian healthcare setting, outpatient encounters and hospital stays are frequently described in textual reports, often without following any standard template or format. Despite the availability of this rich textual content in the health information systems (HIS), automatically performing queries to draw meaningful conclusions is still not possible, due to the unstructured nature of the information.

Starting from this observation, this paper is focused on clinical information extraction (IE) from Italian medical reports, with the main goal of obtaining structured data that can be automatically queried and examined. This would fill the gap between data availability and actionable knowledge. In particular, we are interested in extracting the clinical events that occur in the episodes of care, such as diagnoses, diagnostic procedures, and treatments. In medical reports, these events are often mentioned together with a set of attributes (e.g., clinical variables), with specific values. Extracting these attributes and their values is important to fully identify all event-related information. In this paper, we will address two main research questions:

- Can an automated ontology-driven approach convert Italian textual reports into structured information that can be queried and

[*] Corresponding author.
E-mail address: natalia.viani01@universitadipavia.it (N. Viani).

https://doi.org/10.1016/j.ijmedinf.2017.12.013
Received 11 April 2017; Received in revised form 30 November 2017; Accepted 16 December 2017

examined?

- Is it possible to guide the IE process to preserve some semantic relations between the mentioned entities (e.g., an ECG test, and the heart rate measured during it)?

To answer these questions, we have defined a novel approach that relies on a domain ontology to guide the IE process in an NLP pipeline. This ontology formally specifies the concepts to be extracted and the relations among them. Concepts in the ontology include clinical events and their attributes. Examples of relevant events could be diagnostic procedures or drug prescriptions. As regards possible attributes, diagnostic procedures could be related to their results, while drug prescriptions are linked to dosages and frequencies. The main idea is to obtain a knowledge model that not only can be easily extended, but that is also almost language-independent.

As a clinical case to support the design of the ontology and the development of the IE system, we considered a set of Italian medical reports in the Molecular Cardiology domain.

1.1. Related work

Many IE systems have been developed to deal with English clinical narratives. Such systems rely on a variety of approaches, based either on rules and lexicons (e.g., UMLS [5]) or on machine learning. MedLEE is a rule-based system aimed at extracting and encoding clinical information in textual patient reports [6]. It relies on semantic lexicons to identify the relevant concepts. MetaMap searches for UMLS Metathesaurus entries in text, and includes different processing steps, such as variant generation and word sense disambiguation [7]. CTAKES allows performing several tasks (e.g., named entity recognition, co-reference resolution, relation extraction) through different NLP modules, which can be customized using both dictionaries and machine learning [8]. In general, the interest for clinical IE has grown over the past few years; specific competitions have been organized as well, leading to the development of both supervised and unsupervised approaches (e.g., 2010 i2b2 Challenge [9], 2013 ShARe/CLEF eHealth task [10], SemEval-2015 task [11]).

Despite the increasing research activity in clinical NLP, advances in non-English languages are still limited, mainly due to the lack of shared tools and resources. This is true also for the Italian language, which is the focus of this work. To extract information from Italian clinical text, one main challenge is represented by the unavailability of annotated resources. Currently, we could find only two corpora of Italian medical records that have been annotated and used to develop supervised algorithms. The first corpus includes 500 mammography reports annotated with 9 topics [12]. The second corpus consists of 10000 sentences annotated with medical entities and temporal expressions in a semi-automatic way [13]. To the best of our knowledge, though, these two corpora are not publicly available and cannot be reused for further exploration of supervised techniques.

As an alternative to supervised learning, approaches that do not require annotated data have been developed, too. Chiaramello et al. explored the usability of the MetaMap system to process Italian clinical notes [14]. They obtained two main results. First, they found that the Italian UMLS Metathesaurus has a smaller coverage with respect to the English version. Second, the unavailability of the “variant generation step” for Italian was identified as the main source of annotation failures. In another work, Alicante et al. proposed a system that extracts medical entities using dictionaries and standard NLP tools, and discovers relations among entities through clustering methods [15]. As a main result, they identified clusters corresponding to possible relations, and labeled them in an automatic way.

To guide the development of IE systems, it is possible to rely on domain ontologies, containing information on the concepts to be extracted [16–18]. For the English language, Spasić et al. proposed an ontology-driven system to extract information on findings and anatomical regions from magnetic resonance imaging (MRI) reports [18]. The developed ontology was used to guide and constrain the text analysis, and language processing was modeled through a set of sophisticated lexico–semantic rules.

Few works have dealt with ontology-driven IE on other non-English languages [19,20]. Mykowiecka et al. developed a rule-based system that extracts information from Polish clinical texts to fill in template forms [19]. To specify the information to be extracted, a domain ontology was designed, and manually translated into typed feature structures (TFSs). To extract information, TFSs were combined by manually written grammar rules. In another work, Toepfer et al. created a system that extracts objects (mostly body parts), attributes, and values from German clinical texts [20]. To formalize relevant concepts, a domain ontology was developed and refined by domain experts, in a semi-automatic and iterative way. In each iteration, the expert accepted or rejected annotations automatically extracted according to the ontology, including possible attribute values and variants (in form of either a string or a regular expression).

To the best of our knowledge, this is the first work that uses an ontology-driven approach to mine clinical data from medical reports written in Italian.

Table 1 presents a synthetic view of the literature revised in this section. For each work, we report the target language, the IE methodology (rule-based or machine learning), the information representation strategy (if available), and the limitations we have found for applying each methodology to our problem. In the Discussion section, we will provide a more detailed comparison between our approach and other ontology-driven methodologies.

2. Materials and methods

2.1. Dataset

The corpus considered in this work includes 5432 reports belonging to patients with inherited arrhythmias, such as Long QT Syndrome, and Brugada Syndrome. Documents were provided by the Molecular Cardiology Laboratories of the ICS Maugeri hospital in Pavia, Italy. This set of documents was obtained after cleaning the original corpus to remove a few duplicate instances and those reports that did not include a specific date.

All the considered reports contain the visit date, and most of them are organized in sections, including an anamnestic fitting, the family history, information on performed tests, and a conclusion with possible drug prescriptions. Currently, part of the data written in reports is manually entered in a hospital research system, named TRIAD (http://triad.fsm.it/cardmoc/).

In Fig. 1, we report an example of text containing five relevant concepts: an ECG event and four of its attributes (rhythm, heart rate, PR interval, and atrio-ventricular block).

2.2. NLP pipeline

To perform IE, we designed a pipeline made of different annotators, each with a specific role. The pipeline was implemented using the UIMA framework [21]. Fig. 2 shows the steps needed for the extraction of clinical events and their attributes. First, we use the TextPro tool to perform standard preprocessing (sentence splitting, tokenization, lemmatization, and part of speech tagging) [22]. Then, preprocessed texts are given as inputs to the pipeline.

The first UIMA annotator identifies sections in the text. This is done by using an optional configuration file that contains possible names for sections (e.g., “anamnestic fitting”). The second annotator identifies events. After events are extracted, the third annotator uses the ontology to identify their attributes of interest. In the next sections, we describe in detail the approaches used in the Event and Attribute annotators.