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

Corresponding author
Natalia Viani
Via Ferrata 5, 27100, Pavia (PV), Italy
Telephone: +39 0382-985981
E-mail: natalia.viani01@universitadipavia.it

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.

Keywords: Information Extraction; Natural Language Processing;

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 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. It relies on semantic lexicons to identify the relevant concepts [6]. 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.

Table 1. Related work.

Paper/tool	Language	Method	Knowledge representation	Limitations
MedLEE [6]	English	Rule-based (dictionary + context rules)	Lexicon	• Not directly applicable to Italian • No possibility to extract event-attribute relations
MetaMAP [7], [14]	English, Italian	Rule-based (dictionary)	UMLS Metathesaurus	• No variant generation when applied to Italian • No possibility to extract event-attribute relations
CTAKES [8]	English	Rule-based (dictionary), Supervised machine learning	Customizable dictionaries, UMLS Metathesaurus	• Not directly applicable to Italian • Event-attribute relation extraction not addressed
Esuli et al. [12]	Italian	Supervised machine learning	-	• Requires annotated data
Attardi et al. [13]	Italian	Supervised machine learning	Dictionary features	• Requires annotated data
Alicante et al. [15]	Italian	Rule-based (dictionary), Unsupervised clustering	UMLS Metathesaurus, Italian pharmaceutical dictionary	• Event-attribute relation extraction not addressed • No software available
Spasić et al. [18]	English	Rule-based (ontology-driven)	Domain ontology	• Not easily extendible to Italian • Domain specific • Focused on two concept types (finding and anatomical region)
Mykowiecka et al. [19]	Polish	Rule-based (ontology-driven)	Domain ontology	• Not easily extendible to Italian • Domain specific • Manual curation of complex rules
Toepfer et al. [20]	German	Rule-based (ontology-driven)	Domain ontology	• Not easily extendible to Italian • Domain specific • General object definition and simple attribute structure

Articles and tools revised to motivate our research. The column “Limitations” illustrates the reason why the specific approach is not applicable to our task.

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).

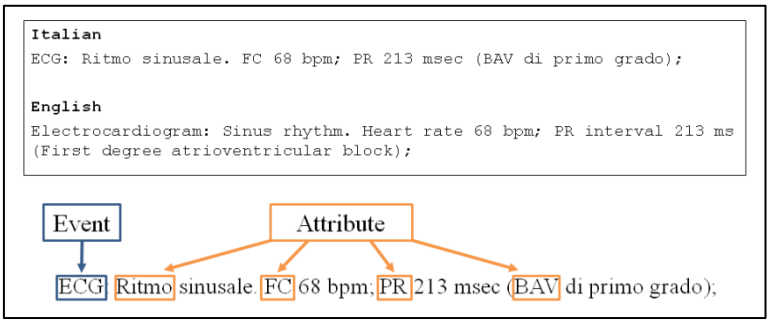

Fig 1. One example sentence. In the example sentence one event and four attributes are highlighted. For clarity, both the Italian text and its English translation are shown.

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.

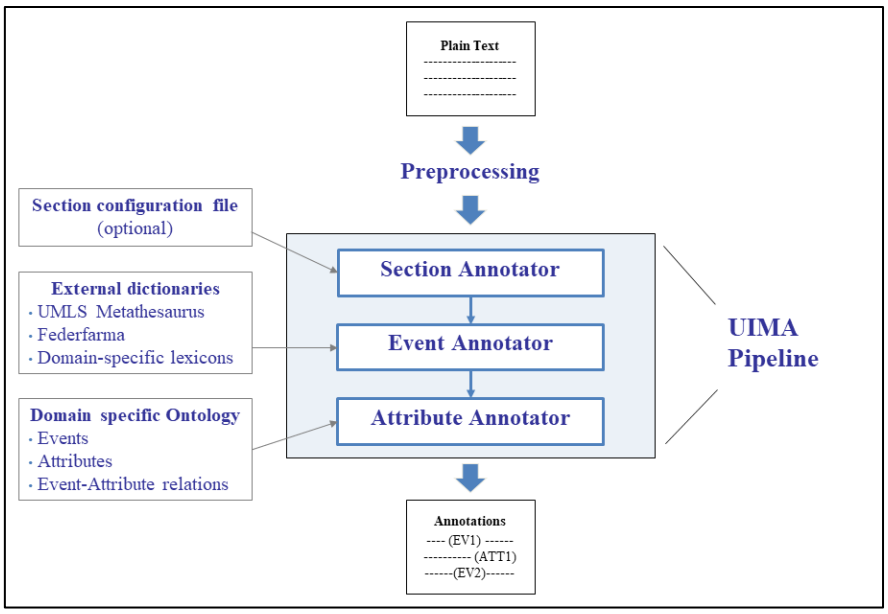

Fig 2. Information extraction pipeline. Preprocessed texts are given as inputs to the pipeline. The pipeline processes texts by annotating sections (Section Annotator), events (Event Annotator), and attributes with related values (Attribute Annotator).

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.

2.3 Dictionary Lookup

The Event annotator identifies events by searching for entries in external dictionaries. As the main sources, we used the Italian version of UMLS, and the FederFarma Italian dictionary of drugs [23]. The Italian UMLS Metathesaurus includes 5 knowledge sources and about 141700 distinct concepts. The FederFarma dictionary contains about 6500 drug names and 4100 active principles. As additional resources, we manually created two vocabularies containing domain-specific procedures and a few events of interest. To expand the list of concepts to be searched for, we also considered a dictionary of common acronyms.

To identify UMLS concepts, we used the CTAKES UMLS Lookup Annotator [8]. We restricted the search to the semantic types representing problems, diagnostic procedures, and treatments. Findings were not considered at this point, but were searched for as event attributes.

Our dictionary lookup approach relies on string matching. To account for plural forms, the search is performed on TextPro normalized tokens. To identify the context in which events are mentioned, we use the ConText algorithm [24]. Specifically, we translated the ConText lexicon to Italian, and we used the algorithm as is to identify negations, hypothetical conditions, and the event experiencer.

2.4 Ontology-driven information extraction

The Attribute annotator relies on a domain ontology to extract attributes and associated values. The ontology is structured into Event and Attribute classes. Events are linked to their attributes through ontology relations, and the same attribute can be connected to multiple events. For example, the information regarding an ECG test is formalized as an Event ("ECGTest") with many Attributes, representing its results and findings (e.g., "AverageHeartRate", "Rhythm"). Some of these Attributes are shared with the Holter test as well.

All the concepts in the ontology are related to a regular expression. Each attribute is characterized also by a set of properties, such as the value. Values can be numeric or categorical. In the first case, the attribute properties include also the unit of measurement. Fig 3 shows an example of the properties for the ECG event and two of its attributes. From this figure, it is possible to notice that the only language-dependent components of the ontology are the properties "hasRegularExpression" and "hasStringValue", which are in this case specified in Italian.

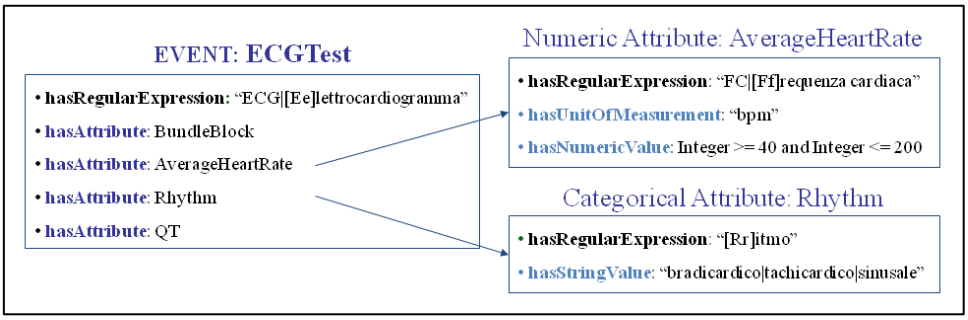

Fig 3. Ontology event and attribute properties. Events and Attributes are characterized by a "hasRegularExpression" property that allows searching for related occurrences in the text. Events are related to their Attributes through "hasAttribute" properties. Attributes are characterized by properties representing constraints on their possible values.

2.4.1 Ontology-driven annotation

The ontology was developed in Protégé [25]. To facilitate its use in the annotation process, the Protégé OWL content is automatically converted to an XML file, which includes events, attributes, and the relationships among them, without additional metadata. This file is given as input to the Attribute annotator.

This annotator matches each event found by the dictionary lookup to the corresponding concept in the ontology, and uses the relations to identify the attributes to be searched for. Then, it exploits the information previously extracted by the pipeline (e.g., presence of sections, semantic types of events) to define event-specific lookup windows where the identified attributes should be searched for. These lookup windows consist of either paragraphs (for tests) or sentences (for drug prescriptions).

As a final step, the annotator uses the regular expressions included in the ontology to extract all the attributes and their values. For similar attributes (e.g., basal, stress, and recovery QT interval), disambiguation is achieved by including in the ontology appropriate attribute modifiers, to be identified in the context surrounding the concept.

2.4.2 Ontology development and refinement

To design and refine the ontology we used a development set including 4429 reports. The steps we followed are shown in Fig 4.

1. We started developing the ontology and the Attribute annotator by looking at 20 reports (*set for ontology design*) randomly selected in the development set. To identify events and attributes of interest, we analyzed both the information written in the text and the data stored in the TRIAD system. For example, we noticed that most reports include sections that describe specific diagnostic tests, with related results. As these results are also reported in TRIAD, we considered them relevant for the extraction, and discussed with clinicians the most important items to capture. After this manual analysis, we created an event for each of the identified tests (e.g., “ECG test”), and defined attributes suitable to store all the relevant results (e.g., “heart rate”). In addition, we exploited both domain knowledge and the TRIAD structure to define which attribute types we should consider (numeric or categorical). As a result of these analyses, we obtained a first version of our system (*system version 1*).

2. To evaluate the performance, we ran *system version 1* on the whole development set. We then iteratively improved the ontology and the annotator according to the results of an error analysis (Fig 4, step 2). In particular, we enhanced regular expressions and added modifiers for few attributes in the ontology. As regards the identification of attribute names, we improved the definition of event-specific lookup windows. Thanks to the performed changes, we obtained a *system version 2*.

3. For the final evaluation, we ran *system version 2* on an independent test set (1003 reports).

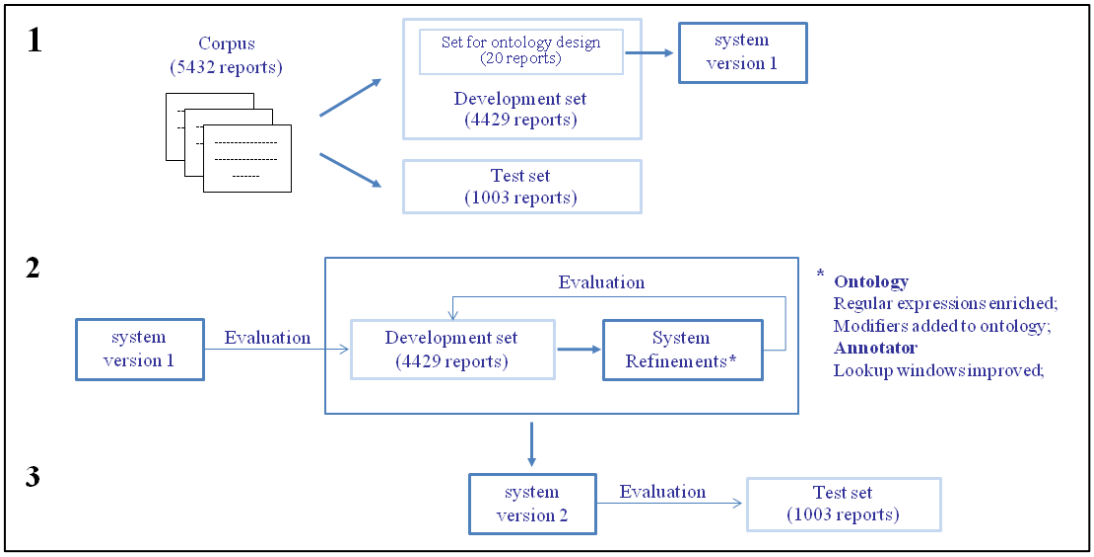

Fig 4. Ontology design and refinement. (1) The ontology was first designed by looking at 20 reports randomly selected in the development set (*system version 1*). (2) The system was then iteratively refined on the whole development set (*system version 2*). (3) The performance of the final version was evaluated on an independent test set.

2.5 Evaluation design

2.5.1 System main validation

Since we did not have annotated reports, we decided to evaluate the proposed approach against TRIAD, the hospital system that stores data on diagnoses, tests, prescriptions, and other relevant events. It is important to point out that there is not an exact alignment between the information written in reports and the data available in TRIAD: some information could have been written in the documents but not transferred to TRIAD, or the electronic data can come from sources different from the reports.

The steps of the evaluation are shown in Fig 5. For each event extracted by our pipeline (*System Event*), we first looked for the matching entry in TRIAD. To match events, we compared the date stored in TRIAD to

the date of the report where the event was found. For those events that we could retrieve, we matched each of the items extracted by our pipeline (*System Items*) to the corresponding data item in TRIAD. To evaluate the performance, we considered those items extracted by the pipeline for which a TRIAD entry was found (*TRIAD Items*). For each of these items, a *correct annotation* corresponds to an exact match between the system and the TRIAD value.

The performance of the system was computed on single events as the ratio between correct annotations and TRIAD items (*% correct annotations*).

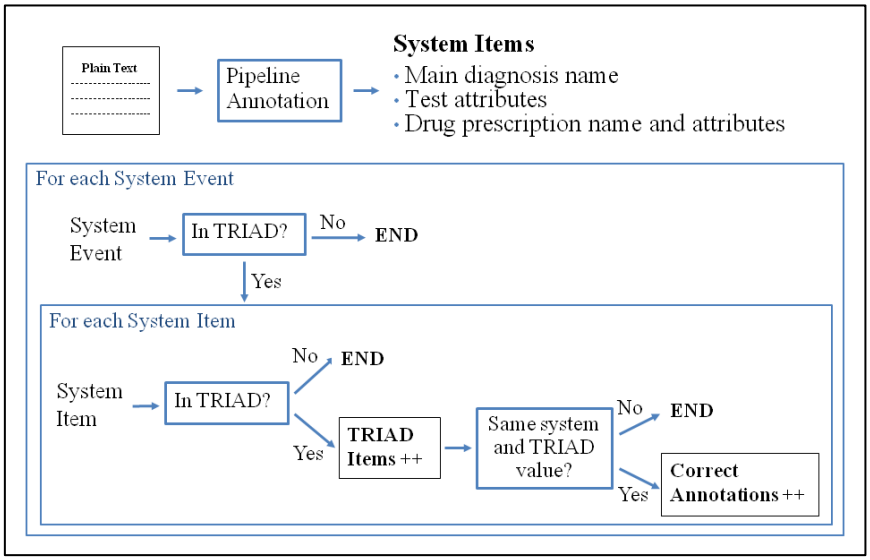

Fig 5. System main validation for one report. To evaluate the performance, we considered those items extracted by the pipeline (*System Items*) for which a corresponding TRIAD entry was found (*TRIAD Items*). For each of these items, a *correct annotation* is defined as an exact match between the system and the TRIAD value.

In Fig 6, we report an example of the performed evaluation for one short report, translated to English for convenience. The figure shows the report (Input Text), the information extracted by our system (Extracted Information), and the matching entry in the TRIAD database (TRIAD Holter ECG matching rows). In this case, the report includes a date ("05/08/2012") and one event (an Holter ECG test) with four attributes (Rhythm, Average Heart Rate, ST Elevation, ST Elevation Type). The extracted date is used to retrieve the matching entry in TRIAD, through the Visit Date field. For each attribute related to the extracted event, the system values and the TRIAD values are compared. Out of the four System Items, the number of correct annotations is three (percentage of correct annotations = 75%). In particular, due to the presence of the negated sentence "*no significant elevation in the other right precordial leads*", the system extracts an "ST

Elevation Type" attribute with an "absent" value. However, for this attribute, the correct value reported in TRIAD is "saddleback".

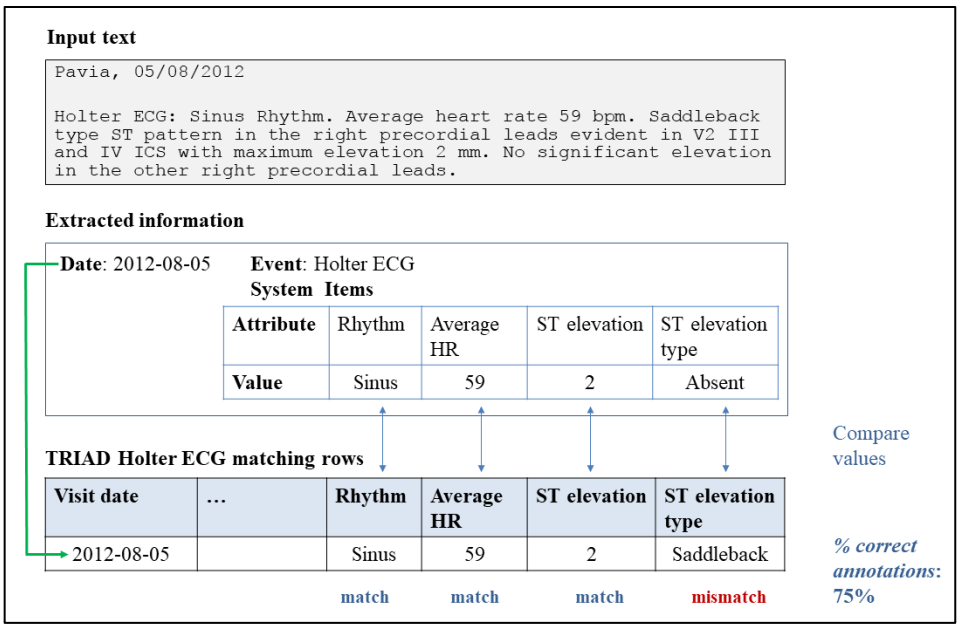

Fig 6. Validation example. The sample text includes the report date, one Holter ECG test and four of its attributes. The date is used to retrieve the matching entry in the TRIAD Holter ECG table. The comparison between the system values and the TRIAD values leads to a percentage of correct annotations of 75%.

2.5.2 Multilingual extendibility

One of the goals of the evaluation was to assess the extendibility of the proposed approach to other languages. To this end, we adapted our pipeline to the analysis of English text by using the English versions of TextPro and ConText, and the English translation of external dictionaries. Also, we translated the section configuration file. For adapting the ontology, we only translated the regular expressions, without performing any additional changes. As regards the Attribute annotator, we did not change the definition of lookup windows, as it is language-independent.

To test the multilingual extension of our pipeline, we had to identify a suitable English corpus. Although the Molecular Cardiology Unit outpatient service is mostly delivered to Italian subjects, we were able to find 37 reports written in English (prepared for foreign patients), which we used for the evaluation. In particular, we used 10 documents as a guide to translate regular expressions, and the remaining 27 reports as the test set. To be consistent, also the evaluation of the English pipeline was conducted against TRIAD.

3 Results

The developed ontology contains 11 events and 61 attributes: 44 attributes are numeric, the others are categorical. The main diagnosis of the patient represents one event identified by its name, and currently has no attributes. Drug prescriptions represent another event, identified by a name, with three attributes (dose, frequency, and format). The other events are diagnostic procedures, each with several attributes.

We evaluated the system on the five events most frequently stored in TRIAD: main diagnosis, prescribed drugs, and three diagnostic tests (ECG, Holter ECG, and Effort stress test). For prescribed drugs, we report both the evaluation on drug names only, and on drug names with associated dosages. Given that drug format and prescription frequency are not included in TRIAD, we did not evaluate these two attributes.

3.1 System main validation

We ran *system version 1* on the development set (Table 2). *System version 1* was then improved following the steps shown in Fig 4. The resulting *system version 2* was run both on the development and on the test sets. Results on the test set are shown in Table 2; we obtained a similar performance on the development set (data not shown).

Table 2. Evaluation results.

System version	Set	Event Name	System Items (a)	TRIAD Items (b)	Correct Annotations (c)	% Correct Annotations (d)
1	**Dev (4429 docs)**	Main Diagnosis	4202	4077	3607	88.5%
		ECG	26669	22546	21352	94.7%
		Holter ECG	26767	21538	19058	88.5%
		Effort Stress Test	9683	3978	2367	59.5%
		Prescribed Drug	8720 (8270*)	2436 (4584*)	2186 (2860*)	89.7% (62.4%*)
2	**Test (1003 docs)**	Main Diagnosis	927	913	845	92.6%
		ECG	7452	5070	4885	96.4%
		Holter ECG	7173	5127	4757	92.8%
		Effort Stress Test	2543	1118	1064	95.2%
		Prescribed Drug	1999 (1999*)	538 (930*)	435 (672*)	80.9% (72.3%*)

Evaluation of *system version 1* on the development set (4429 documents) and of *system version 2* on the test set (1003 documents). **a**: number of items extracted by the system **b**: number of extracted items for which an entry was detected in TRIAD; **c**: correct annotations; **d**: *% correct annotations*, computed as c/b. Cells marked with * are related to the results for drug names and dosages.

In the evaluation of *system version 1*, we obtained a percentage of correct annotations of 88.5% for diagnoses. The error analysis showed that most errors were due to problems in matching diagnosis names in reports with the corresponding entries names in TRIAD. As regards tests, the system achieved the best percentage of correct annotations for ECGs (94.7%), followed by Holter tests (88.5%), which are described with more complex sentences. On the other hand, we obtained a poor performance for Effort Stress tests (59.5%). In this case, the main issue was the misclassification of attributes that are written in the same way (e.g., “QT” for QT interval), but are related to different test phases (e.g., baseline, stress, and recovery QT length). Regarding drug prescriptions, drug names identification led to good results (89.7%). However, extracting drug dosages was not trivial (62.4%) because, while the TRIAD database contains only daily doses, reports often include sentences with both unit dosages and frequencies of administration.

In the evaluation of *system version 2* on the test set, a better performance was achieved for almost all events. For diagnoses, we exploited physicians' knowledge to refine the mapping of the terms used in TRIAD to those used in the reports. The most significant improvement concerned effort stress tests, with an increase in

the percentage of correct annotations from 59.5% to 95.2%. The only decrease in performance was given by drug names (from 89.7% to 80.9%). In this case, many of the non-matching drug names were due to the insertion of erroneous data in TRIAD: two very similar drugs ("Metoprolol" and "Metoprolol Retard") were frequently stored with the same name.

3.2 Multilingual extendibility

After adapting our pipeline to the analysis of English text, we ran the resulting system on the English test set (27 reports), obtaining promising results (Table 3). However, for events that are described with long sentences (Holter ECG, Effort Stress tests) we obtained a slightly lower performance with respect to the Italian counterpart. This was probably caused by a few translations that were not straightforward, mainly due to syntactic differences among languages (e.g., word order).

Table 3. Multilingual extendibility results.

System version	Set	Event Name	System Items (a)	TRIAD Items (b)	Correct Annotations (c)	% Correct Annotations (d)
2 (adapted to English)	**EN test (27 docs)**	Main Diagnosis	27	19	18	94.7%
		ECG	183	78	74	94.9%
		Holter ECG	115	65	56	86.2%
		Effort Stress Test	110	39	34	87.2%
		Prescribed Drug	91 (91*)	20 (31*)	17 (23*)	85% (74.2%*)

Multilingual extendibility evaluation on the English test set (27 documents). **a**: number of items extracted by the system; **b**: number of extracted items for which an entry was detected in TRIAD; **c**: correct annotations; **d**: % *correct annotations*, computed as c/b. Cells marked with * are related to the results for drug names and dosages.

4 Discussion

4.1 Main contributions in relation to existing research

Identifying information in clinical free-text is a challenging task. To perform clinical IE, several systems have been proposed in the literature, especially for English text. Many systems include machine learning modules, developed through supervised techniques. To both develop such modules and enable their evaluation, the availability of annotated corpora is essential. For the Italian language, however, shared

clinical corpora are not easily available. The main contribution of our paper is the development of an ontology-driven approach for clinical IE that, embedded in a well-performing NLP methodology, allows the analysis of the Italian language without the need for annotated data. Besides extracting clinical events, this approach allows capturing attributes of interest and linking them to the events they are related to.

To define events and related attributes, we developed an ontology for the cardiology case study by analyzing both the information written in the reports and the data stored in the TRIAD database. Also, we organized a few encounters with physicians to verify the information to be extracted. The ontology, iteratively refined on a development set, is used to guide the information extraction process. We evaluated the performance of the final system on an independent test set, achieving high percentages of correct annotations. The information extracted by the system was validated against the data stored in TRIAD. To assess multilingual extendibility, we translated the ontology to the English language, obtaining promising results.

As mentioned in the Related Work section, a few works have used ontologies for clinical NLP. In the following, we will discuss the main differences between our approach and previous efforts in this area. Spasić et al. developed an ontology-driven system, KneeTex, that extracts information from English knee MRI reports [18]. KneeTex is focused on the extraction of findings and anatomical regions, with possible classifiers which are defined in the ontology. Although the resulting system performs very well for the considered task, the developed ontology is strongly domain-specific. In our ontology, instead, the definition of events and attributes is general, thus facilitating an extension to other clinical domains. Mykowiecka et al. proposed an ontology-driven system to analyze mammography reports in Polish [19]. The proposed system works well on the analyzed clinical domain. However, a considerable manual effort was put into complex rules engineering. As a main difference, our ontology is automatically translated into an XML file, which is then given as an input to the NLP pipeline. Moreover, the approach we use to extract and relate information has a smaller dependency on syntax. Toepfer et al. used an ontology to extract objects (with attribute and values) from German transthoracic echocardiography reports [20]. The developed IE system performs well. The proposed ontology structure is similar to ours, especially as regards the definition of attributes and values. As one main difference, we do not consider objects, but events characterized by a semantic type. Moreover, we relate events to attributes that can be either numeric or categorical, rather than considering only textual variants. As a final consideration, while Toepfer et al. defined the ontology in a semi-automatic

way, our ontology was developed by manually analyzing reports. In the future, it would be interesting to explore the possibility of automatically developing the ontology from free-text [26,27]. To this end, concepts automatically extracted from UMLS could be exploited.

Formalizing information through ontologies brings several advantages. First, although ontologies are built for a specific domain, they allow easy updates and extensions to account for new information. Our approach, for example, could be applied to the analysis of reports coming from other clinical domains. To adapt the system, only the ontology (and possibly the external dictionaries) should be updated. Second, thanks to the flexible Event-Attribute structure, it is possible to relate the same attribute to different events, reusing shared concepts multiple times. In addition, the inclusion of regular expressions in the ontology makes the proposed approach easily language extensible. To analyze reports in another language, it would be in principle sufficient to translate the regular expressions. As another interesting aspect, ontologies can be particularly suitable to assess and manage data quality [28].

Overall, the system proposed in this paper achieved a good performance on items extraction and linking. An alternative approach to process non-English texts is to automatically translate documents, and use one of the systems available for English [14]. However, there are two reasons why we decided not to go for this solution. First, we did not want to introduce errors due to automatic translation. Second, we were interested in extracting and linking events and attributes, a feature that is not the main focus of available clinical NLP systems.

The ontology proposed in this work was developed for clinical text in Italian. To evaluate the feasibility of a multilingual extension based on translation, we adapted our pipeline to process a set of English clinical reports. Although the considered corpus is small, this preliminary evaluation shows encouraging results. Despite this, we have to point out that the syntactic structure of sentences can be different across languages: it might be not straightforward to translate some concepts, especially if these concepts are expressed by several words. To mitigate this issue, one option could be using attribute names composed of only one word (e.g., “Regurgitation” instead of “Mitral regurgitation”), and including all the other identifying words (e.g., “Mitral”) in a suitable modifier, to be searched for in a lookup window. However, this solution would introduce a layer of complexity into the definition of the ontology. Finally, since our multilingual extendibility assessment was focused on the ontology, it has not taken into account the different coverage of

dictionaries across languages. Analyzing this aspect would be instead relevant in a more comprehensive assessment including the whole NLP pipeline.

4.2 Potential clinical impact

The proposed IE approach has several potential implications to clinical practice. First of all, relevant information is extracted by using specific relations that are defined through a domain knowledge formalization. This feature allows converting textual content into a structured format that verifies an event-attribute logic and can be easily queried. Ensuring a timely access to this data is of paramount importance to facilitate patient review and support clinical decision.

As pointed out by Botsis et al., NLP methods can play an important role in reducing the health data that is unavailable, inaccessible or incomputable [29]. Integrating structured data extracted from text into research repositories can effectively facilitate the reuse of collected data. To this end, the information extracted through our pipeline is currently being integrated into an i2b2 data warehouse together with data coming from other sources [30]. As a future application, the system could also be used to automatically populate the TRIAD system, saving a lot of manual work. In this case, using an automatic extraction system would be useful to check the manually entered items to improve data quality.

4.3 Limitations

The approach we proposed has some limitations. First, given the unavailability of annotated data, we could not evaluate the performance on a gold standard dataset. Given the large size of our corpus, manually annotating all the documents would be hard and time consuming. To overcome this issue, we validated our system annotations against the data included in TRIAD. However, we could only consider item values that are both extracted from reports and found in TRIAD, thus focusing on the percentage of correct annotations. To evaluate false negatives, we should consider data that are available in TRIAD, but not extracted from reports. Since additional sources could have been used to fill in the database, this computation cannot be automatically performed. To evaluate false positives, on the other hand, we should consider data extracted from reports, but not available in TRIAD. Given that data entered in TRIAD is not guaranteed to be complete, it is hard to evaluate false positives as well. As one final remark, even when the same items are present in both reports and TRIAD, there could be human errors in data entry.

Another limitation of our approach concerns the structure of the considered reports. In the Italian clinical setting, the structure of clinical free text reports is in general defined by the specific center that issues the report to the patient. For this reason, completely unstructured documents are as frequent as more structured texts organized in sections and paragraphs. In the majority of the documents considered in this paper, the content is organized in a clear way, and it is possible to identify specific sections, some of which actually correspond to clinical events (e.g., “ECG test”, “Holter ECG test”). This feature may have affected the results that we obtained, which may not extend equally well to other clinical corpora. To further assess this potential issue, we are currently extending our pipeline to process pathology reports for breast cancer patients. As a related topic, to ultimately assess the feasibility of an approach based on ontology translation for multilingual extendibility, it would be necessary to consider a larger corpus, with more variability. Unfortunately, we could not retrieve many documents that were originally written in English, and we decided not to automatically translate reports. Still, the evaluation we conducted can be considered as a preliminary assessment, with promising results.

As a final limitation, we used regular expression matching to look for attribute names and values in the text. However, we did not take into account possible variants of concepts (except those already included in the ontology). Also, we did not address the presence of typing errors, such as misspellings or wrong data values. In case of poor data quality, such errors could affect the extraction process, leading to an increase in the number of false negatives. To assess the quality of our corpus, we performed a qualitative review of the reports included in the set for ontology design, and we noticed that most reports follow a regular structure, without major typing errors. As future work, we will perform a systematic evaluation of the corpus relying on a set of annotated documents, which will enable a quantitative assessment of the data quality on the basis of the analysis of false negatives.

5 Conclusion

The approach illustrated in this work relies on a domain ontology to extract events and attributes from medical text in the Italian language. To identify and formalize relevant concepts, we leveraged domain knowledge and information written in reports. The developed ontology can be easily enriched and translated. Our preliminary multilingual assessment indicates that our approach could be extended to other languages.

The obtained results show that the developed system performs well, thus it could be successfully used to analyze languages such as Italian, where shared corpora and resources may not be easily accessible. As future work, we plan to explore the feasibility of methods based on supervised machine learning, too. To this end, we are currently manually annotating a subset of the available documents. As another possibility to explore supervised approaches, the annotations currently produced by our IE system, which are partially validated against a structured database, could be considered in the future as a pseudo-gold standard corpus.

Acknowledgments

We would like to acknowledge the Centre for Health Technologies (CHT) at the University of Pavia for making this work possible. We would also like to thank Prof. Guergana Savova, Dr. Timothy Miller, and the Natural Language Processing laboratory at the Computational Health Informatics Program (CHIP) at Boston Children's Hospital for their valuable help and the insights provided during this research.

References

[1] A. Talaei-Khoei, T. Solvoll, P. Ray, N. Parameshwaran, Maintaining awareness using policies; Enabling agents to identify relevance of information, J. Comput. Syst. Sci. 78 (2012) 370–391.

[2] D. Demner-Fushman, W.W. Chapman, C.J. McDonald, What can Natural Language Processing do for Clinical Decision Support?, J. Biomed. Inform. 42 (2009) 760–772.

[3] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb Med Inf. (2008) 128–44.

[4] S. Velupillai, D. Mowery, B.R. South, M. Kvist, H. Dalianis, Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis, Yearb. Med. Inform. 10 (2015) 183–193.

[5] Unified Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls/ (accessed January 7, 2017).

[6] C. Friedman, A broad-coverage natural language processing system., Proc. AMIA Symp. (2000) 270–274.

[7] A.R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. JAMIA. 17 (2010) 229–236.

[8] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. JAMIA. 17 (2010) 507–513.

[9] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. JAMIA. 18 (2011) 552–556.

[10] S. Pradhan, N. Elhadad, B.R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W.W. Chapman, G. Savova, Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, J. Am. Med. Inform. Assoc. JAMIA. 22 (2015) 143–154.

[11] N. Elhadad, S. Pradhan, S. Lipsky Gorman, W.W. Chapman, S. Manandhar, G.K. Savova, SemEval-2015 task 14: Analysis of clinical text, in: Proc. 9th Int. Workshop Semantic Eval. SemEval 2015, 2015: pp. 303–10.

[12] A. Esuli, D. Marcheggiani, F. Sebastiani, An enhanced CRFs-based system for information extraction from radiology reports, J. Biomed. Inform. 46 (2013) 425–435.

[13] G. Attardi, V. Cozza, D. Sartiano, Annotation and extraction of relations from Italian medical records, in: Proc. 6th Ital. Inf. Retr. Workshop, 2015.

[14] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli, G. Tognola, Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes, J. Biomed. Inform. 63 (2016) 22–32.

[15] A. Alicante, A. Corazza, F. Isgrò, S. Silvestri, Unsupervised entity and relation extraction from clinical records in Italian, Comput. Biol. Med. 72 (2016) 263–275.

[16] I. Spasic, S. Ananiadou, J. McNaught, A. Kumar, Text mining and ontologies in biomedicine: making sense of raw text, Brief. Bioinform. 6 (2005) 239–251.

[17] D.C. Wimalasuriya, D. Dou, Ontology-based information extraction: An introduction and a survey of current approaches, J. Inf. Sci. 36 (2010) 306–323.

[18] I. Spasić, B. Zhao, C.B. Jones, K. Button, KneeTex: an ontology–driven system for information extraction from MRI reports, J. Biomed. Semant. 6 (2015) 34.

[19] A. Mykowiecka, M. Marciniak, A. Kupść, Rule-based information extraction from patients' clinical data, J. Biomed. Inform. 42 (2009) 923–936.

[20] M. Toepfer, H. Corovic, G. Fette, P. Klügl, S. Störk, F. Puppe, Fine-grained information extraction from German transthoracic echocardiography reports, BMC Med. Inform. Decis. Mak. 15 (2015).

[21] D. Ferrucci, A. Lally, UIMA: an architectural approach to unstructured information processing in the corporate research environment, Nat. Lang. Eng. 10 (2004) 327–348.

[22] E. Pianta, C. Girardi, R. Zanoli, The TextPro Tool Suite., in: Proc. 6th Ed. Lang. Resour. Eval. Conf., 2008.

[23] FederFarma, https://www.federfarma.it/ (accessed January 7, 2017).

[24] H. Harkema, J.N. Dowling, T. Thornblade, W.W. Chapman, Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports, J. Biomed. Inform. 42 (2009) 839–851.

[25] Protégé, http://protege.stanford.edu/ (accessed January 7, 2017).

[26] K. Liu, W.R. Hogan, R.S. Crowley, Natural Language Processing methods and systems for biomedical ontology learning, J. Biomed. Inform. 44 (2011) 163–179.

[27] J. Hoxha, G. Jiang, C. Weng, Automated learning of domain taxonomies from text using background knowledge, J. Biomed. Inform. 63 (2016) 295–306.

[28] S.T. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A.E.T. Yeo, A. Talaei-Khoei, Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature, Int. J. Med. Inf. 82 (2013) 10–24.

[29] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary Use of EHR: Data Quality Issues and Informatics Opportunities, Summit Transl. Bioinforma. 2010 (2010) 1–5.

[30] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. JAMIA. 17 (2010) 124–130.