Information Extraction in the Medical Domain

Article *in* Journal of Information Technology Research · April 2015

DOI: 10.4018/jitr.2015040101

Aicha Ghoulam
Hassiba Benbouali University of Chlef
4 PUBLICATIONS **9** CITATIONS

SEE PROFILE

Fatiha Barigou
Université Oran 1, Ahmed Benbella, Oran, Algeria
36 PUBLICATIONS **83** CITATIONS

SEE PROFILE

Ghalem Belalem
University of Oran1, Ahmed Ben Bella
114 PUBLICATIONS **371** CITATIONS

SEE PROFILE

Information Extraction in the Medical Domain

Aicha Ghoulam, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria

Fatiha Barigou, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria

Ghalem Belalem, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria

ABSTRACT

Information Extraction (IE) is a natural language processing (NLP) task whose aim is to analyse texts written in natural language to extract structured and useful information such as named entities and semantic relations between them. Information extraction is an important task in a diverse set of applications like bio-medical literature mining, customer care, community websites, personal information management and so on. In this paper, the authors focus only on information extraction from clinical reports. The two most fundamental tasks in information extraction are discussed; namely, named entity recognition task and relation extraction task. The authors give details about the most used rule/pattern-based and machine learning techniques for each task. They also make comparisons between these techniques and summarize the advantages and disadvantages of each one.

Keywords: *Electronic Medical Report, Extraction of Semantic Relations, Information Extraction, Medical Named Entities Recognition, Medical Relation Extraction*

1. INTRODUCTION

The amount of information written in natural language and available in electronic format is increasing. Due to their unstructured nature, however, manual analysis of this huge information is challenging and labor intensive. To address these concerns we need new techniques of structured extraction to access useful information. Information Extraction (IE) can relieve some of these problems by offering access to relevant information without requiring the end user of the information to read the text.

As it is mentioned in (Jiang, 2012), extraction of structured information from text dates back to the '70s, it started gaining much attention when DARPA (Defense Advanced Research Projects Agency) initiated and funded the Message Understanding Conferences (MUC) in the '90s,. MUCs defined information extraction as filling a predefined template that contains a set of predefined slots like a terrorism template

DOI: 10.4018/jitr.2015040101

used in MUC-4. Template filling is a complex task and systems developed to fill one template cannot directly work for a different template. In MUC-6, a number of template-independent subtasks of information extraction were defined; these include named entity recognition, and relation extraction.

Early information extraction systems like the ones that participated in the MUCs were rule-based with manually coded rules. They use linguistic extraction patterns developed by humans to match text and locate information units. They can achieve good performance on a specific target domain, but it is labor intensive to design good extraction rules, and the developed rules are highly domain dependent. Realizing the limitations of these manual developed systems, researchers turned to statistical machine learning approaches. With the decomposition of information extraction systems into components such as named entity recognition, many information extraction subtasks can be transformed into classification problems or sequence labeling, the first one can be solved by standard supervised learning algorithms such as support vector machines and maximum entropy models, and the second one because information extraction involves identifying segments of text that play different roles, it can be solved by hidden Markov models and conditional random fields.

The IE is a research subject that covers many areas like customer care, personal information management, bio-informatics, community web sites. As it is mentioned in (Berrazega, 2012); these applications require IE for searching and responding queries.

To facilitate these search capabilities, information extraction is often needed as a preprocessing step to enrich document representation or to populate a database.

As the volume of medical knowledge double every five years according to some studies as it mentioned in (Ben Abacha & Zweigenbaum, 2011a), and recorded in unstructured formats, development of medical information extraction techniques have gained immense popularity. They include identification of biomedical and/or medical named entities, relations between the entities, or events associated like the one developed in (Zweigenbaum & Tannier, 2013).

Noticeable efforts have been invested in the medical domain. Examples include the work of (Harkema et al., 2005) who applied AMBIT in clinical and biomedical texts to extract key information. Aronson (2001) used MetaMap tool to recognize and categories medical terms.

In this paper, we focus on the two most fundamental tasks in information extraction, namely, named entity recognition and relation extraction in the medical field. We will compare some works using rule/pattern-based and machine learning approaches in term of used corpus, coverage and precision. The remainder of this paper is divided as follows: Section 2 presents the information extraction concept and approaches of extraction. Section 3 introduces information extraction in the medical domain, look at the related work on medical information extraction, and then initiate a comparative study. Finally, section 4 presents our conclusions and perspectives.

2. INFORMATION EXTRACTION

2.1. Definition

Information Extraction has been defined in the literature review by many researchers (Sarawagi, 2007) and (Jiang, 2012). The most common definition is that IE is an automatic process for extracting structured information which can be relevant for a particular domain from unstructured documents like free text that are written in natural language (e.g. news article, clinical reports) or semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. The obtained data are then arranged to be incorporated into machine readable databases and ontologies which, in turn, are used to improve applications such as Question Answering engines or Information Retrieval systems.

Five separate component tasks, which illustrate the main functional capabilities of current IE systems, were specified by recent MUC-7 evaluation (Nadeau & sekine, 2007), The tasks were centred around extracting information into relational records, known as templates. The tasks are given below, adapted from the MUC-7 task definitions (Chinchor & Marsh, 1998):

- Named Entity Recognition (NER), involves the recognition of named entities such as organizations, persons, locations, dates and monetary amounts. The task has been greatly expanded to cover both concrete and abstract things in text. In the clinical domain, this might include entities such as disease and drug;
- Relation Extraction (RE) task, is the task of detecting and characterizing the semantic relations between entities in text. In the clinical field, it include for example relation between disease and drug;
- Coreference Analysis task, is a task which determine linguistic expressions that refer to the same real-world entity in natural language, has not yet been widely applied to clinical documents (Ware et al, 2012). Formally, coreference consists of two linguistic expressions antecedent and anaphor. The anaphor is the expression whose interpretation (i.e., associating it with an either concrete or abstract real-world entity) depends on that of the other expression. The antecedent is the linguistic expression on which an anaphor depends. in the sentences given in (Zheng et al., 2011): "Have reviewed the electrocardiogram. It shows a wide QRS with a normal rhythm but no delta waves.", "the electrocardiogram" and "It" refer to the same entity which is the electrocardiogram. So, "the electrocardiogram" is the antecedent, and "It" is the anaphor;
- Template Filling, The information to be extracted like entities, relationships and events in natural language texts is pre-specified in user defined structures called templates (or objects), each consisting of a number of slots (or attributes), which are to be instantiated by an IE system as it processes the text; and
- Event Description, (Sun et al., 2013) defined a medical event as anything that is clinically important and that can also be mapped to a timeline. They created the i2b2 2012 challenge; a clinical temporal relation corpus that includes clinical events, temporal expressions, and temporal relations. The clinical event were defined in the i2b2 2012 challenge to include: 1- clinical concepts (such as problems, tests, and treatments), 2- clinical departments (such as 'surgery' or 'the main floor'), 3- evidentials (ie, events that indicate the source of the information), and 4- occurrences (ie, events that happen to the patient, such as 'admission', 'transfer', and 'follow-up').

The temporal expressions, capturing dates, times, durations, and frequencies. Temporal relations, or temporal links, indicate whether and how two events, two temporal expressions, or an event and a temporal expression are related to each other in the clinical timeline.

Current IE systems do not generally extract MUC-style templates. In the Automatic Content Extraction programme (ACE), a successor to MUC, tasks are merged into one task for each of entities, relations and events (Doddington et al., 2004):

In this paper we are concerned with only two tasks; named entity recognition and relation extraction in the medical field. Section three gives a study of recent medical systems in information extraction task that can be broken down into Named Entity Recognition and Relation Extraction using two approaches rule based and machine learning. And then compare them using the standard evaluation metrics.

2.2. Information Extraction Methods

Usually an information extraction system supports one of the two basic methods of extraction, namely,.rule-based information extraction method, and statistical information extraction method.

2.2.1. Rule-Based IE Methods

Rule-based methods extract the informaton by rules, and these rules can be generated by human hand-coded, or by learning from examples.

Early information extraction systems in MUC were human hand-coded rules based systems, they use rules written by knowledge engineers and developed by designers who must know the formalism for writing those rules for the particular system used, to match text and locate information units; The most representative examples of this kind of systems are FASTUS (Hobbs et al., 1992), GE NLTOOLSET (Krupka et al, 1992), PLUM (Ayuso et al, 1992) and PROTEUS (Yangarber & Grishman, 1998); these systems are well described in (Kaiser & Miksh, 2005). They can achieve good performance on the specific target domain.

Human hand-coded rule-based system, in some sources also called knowledge engineering method, gives very good results, however, involves a great human effort and a considerable time for data analysis and rule writing. It is time consuming during development.

Later systems try to automatically learn such patterns from labeled data. These supervised approaches usually need a training process which requires users to provide training examples, for example, provide some tagging data. So by the help of the training dataset, the supervised approaches are able to learn a pattern. The most representative examples of this kind of systems is AutoSlog (Riloff, 1993).

Rule-based IE methods for named entity recognition generally work as follows: A set of rules is either manually defined or automatically learned. A rule consists of a pattern and an action. A pattern is usually a regular expression defined over features of tokens. For example, to label any sequence of tokens of the form "*Mr. X*" where *X* is a capitalized word as a person entity, the rule can be defined as shown in Box 1.

The left hand side is a regular expression that matches any sequence of two tokens where the first token is "*Mr.*" and the second token has the orthography type *FirstCap*. The right hand side indicates that the matched token sequence should be labeled as a person name.

Also rule-based IE methods for relation extraction generally work similarily.

For example the pattern "*X is treated by Y*" where X and Y are named entities, can extract the following relation: *is treated by (X, Y).*

Hand rule-based systems are more useful in closed domains where human involvement is both essential and available. In open-ended domains like opinion extraction from Blogs, the flexibility of statistical methods is more appropriate.

2.2.2. Statistical Learning IE Methods

Statistical learning methods or Machine Learning (ML) methods; are trainable techniques able to improve their ability to extract information from input automatically or under supervision see the survey of (Nadeau & sekine, 2007). Most recent studies use supervised machine learning starting from a collection of training examples; the idea of supervised learning is to study the features of positive and negative examples of information to be extracted (e.g. entities, relations, attributes) over a large collection of annotated documents and design rules that capture instances of a given type.

Many different models have been proposed over the years. The most prominent of these are Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), Support Vector Machine (SVM) and Conditional Random Fields (CRF) as developed in (Abachaet

Box 1.

$$\left(Token = \text{“}Mr.\text{”}orthography\ type = FirstCapitalized\right) \rightarrow person\ name$$

al., 2011c). CRFs and SVM are now established as the state-of-the-art methods and have shown clear advantages over MEMMs and HMMs both theoretically and empirically. The very important advantage of a machine learning based system is that it can be transferred to a different domain easily as long as specific texts and a person who can annotate them are available. But sometimes those texts are problematic or expensive to obtain or there is a lack of useful documents on which an algorithm can learn.

The machine learning techniques have demonstrated remarkable results in the general domain and hold promise for medical information extraction, however, as Chapman et al. (2011) say IE, especially in the medical domain, is expensive. It requires large volumes of high quality,manually annotated example text which are both expensive and time-consuming to train the models.

So the main shortcoming of supervised machine learning techniques is the requirement of a large annotated corpus; the unavailability of such resources and the prohibitive cost of creating them; lead researchers to two alternative learning methods; the semi-supervised learning and unsupervised learning techniques like that developed in (Zhang & Elhadad, 2013). The first one involves a small degree of supervision, like a set of seeds, for starting the learning process. The second one has the idea of clustering; for example gathering named entities into clustered groups based on the similarity of context and relying on lexical resources like WordNet. The survey of these techniques is well presented by (Nadeau & Sekine, 2007).

2.2.3. Wrapper Induction

Many approaches for data extraction from web pages have been developed to transform the web pages into program-friendly structures such as a relational database. Wrapper induction system considers web pages as a source data. It is a program that wraps an information source like a database server, or a web server (Chang et al., 2006); it usually performs a pattern matching procedure like a form of finite-state machines which relies on a set of extraction rules.

In the Web environment, The aim of a wrapper is to locate relevant information in semi-structured data ((e.g. HTML, XML) and to convert it into a self-described representation for further processing. Here, wrappers (e.g. WIEN presented by (Kushmerick, 2000)) are not constrained by natural language processing, they can take advantage of predefined HTML (XML) templates which implicitly classify the data found in a document:

There are many wrapper systems that are well described in (Chang et al., 2006); they classify them into four classes: manually-constructed IE systems (e.g. TSIMMIS, Minerva, XWrap), supervised IE systems (e.g. RAPIER, WIEN), semi-supervised IE system (e.g. IEPAD, OLERA) and unsupervised IE systems (e.g. DeLa, RoadRunner) and compare them in three dimension: task difficulties, technique used and automation degree.

2.3. IE using Ontology

Ontology-based information extraction task (OBIE) has recently emerged as a subfield of IE. Ontology is a formal and explicit specification of a shared conceptualization, it plays a crucial role in the process of IE.(Ritesh & Suresh, 2014).

Ontologies represent an ideal knowledge background in which to base text understanding and enable the extraction of relevant information. This may enable the development of more flexible and adaptive IE systems than those relying on manually composed extraction rules.

In (Wimalasuriya & Dejing, 2010) an OBIE system is defined as a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using an ontology definition language such as the Web Ontology Language (OWL). Those authors describe a close relation between OBIE and the Semantic

Web. OBIE systems generate semantic content which is known as Semantic Annotation for the Web pages.

The relation between ontologies and IE is involved in two tasks (Nedellec & Nazarenko, 2005): on the one hand, Ontology is used for information extraction; IE needs ontologies as part of the understanding process for extracting the relevant information (Gurulingapa et al., 2012b); on the other hand, information extraction is used for populating and enhancing a domain ontology.

According to (vicient, 2011), two different kinds of methods involving IE and ontologies are used: (i) the ontology-based IE method and (ii) the ontology-driven method. The first one uses a domain-specific ontology in its extraction process, it is document driven; it tries to identify entities starting from a particular document (or set of documents) and trying to annotate them according to the input ontology. On the contrary, in the second method, the idea is to consider each of the ontological elements and to use them to search for resources (e.g. Web pages) that can provide interesting information related to each component of the ontology.

In the medical domain, multiple standardised ontologies are available (e.g. UMLS[1], SNOMED CT[2], MeSH[3]). This external knowledge is exploited by extraction tools to identify meanings in sentences and to identify relevant text snippets given an extraction task.

2.4. Evaluation Criteria

According to (Piskorski & Yangarber, 2012), the overall quality of extraction depends on multiple factors, including, i.e.: (a) the level of logical structures to be extracted (entities, co-reference, relationships, events), (b) the nature of the text source (e.g., online news, web pages, news articles, blogs …), (c) the domain in focus and the language of the input.

Most information extraction systems use precision (P), recall (R) and their combination into F-Score to measure performance. The first shows the system's accuracy, the second the coverage, and the third is the harmonic mean between the first two.

The precision and recall metrics were adopted from the information retrieval research community. They measure the system's effectiveness from the user's perspective, i.e., the extent to which the system produces all the appropriate output (R) and only the appropriate output (P). This can be performed by annotating a corpus provided by expert curators then compare with the automatic annotations, see the survey given by (Camps et al., 2012).

To define P, R and F-score formally, let Nc denote the number of correctly extracted output by the system; Nt denote the total number of output extracted by the system (includes output that have been extracted correctly, incorrectly and overgenerated4) and T The number of manually annotated elements which includes all the items that the user wants the system to extract.

Precision is the ratio between items that have been correctly extracted and the total number of extracted items: $P = \frac{Nc}{Nt}$. Recall is the ratio between textual elements correctly extracted by the system, and textual elements that are manually annotated: $R = \frac{Nc}{T}$. F-Score combines these two into a single score and is defined by the following equation: $F - score = \frac{\left(B^2 + 1\right)PR}{B^2 P + R}$. By means of the parameter β it can be determined whether the recall (R) or the precision (P) score is weighted more heavily. when β equals 1, i.e., recall and precision are of equal importance, the metric is called the harmonic mean (F1-score).

3. INFORMATION EXTRACTION IN THE MEDICAL FIELD

The amount of information has increased exponentially in all areas including the medical field where clinical data concerning the medical histories, pathologies, and personal information of a patient are in form of medical reports written by doctors and are difficult to access, as they are often in unstructured form. To make access to patient data easily, structured data is required. This is where Natural Language Processing and more precisely Information Extraction is needed. It has a long history of research and of use with medical records, as reviewed most recently by (Meystre et al., 2008).

In this paper, we refer to medical information extraction as information extraction carried out on medical text. By medical text we mean just the unstructured, textual portion of the patient medical record.

In the medical domain several authors experimented with more simple systems focused on specific IE tasks and on a limited number of different types of information to extract. A more recent review focused specifically on clinical IE from electronic health record; between (1995 and 2008) is well described in (Meystre et al., 2008).

In this paper, we will focus only on medical named entity extraction (MNER) and medical relation extraction (MRE) from clinical reports. Many recent researches are also interested by co-reference resolution (Dai et al., 2012), (Chen et al., 2012), (Yan et al., 2012), (Zheng et al., 2011), template filling and event extraction (Ananiadou et al., 2010), (Jindala & Dan, 2013), (Zweigenbaum & Tannier, 2013) in the medical filed. But we are concentrating for the two fundamental tasks: MNER and MRE for two reasons; first, named entity recognition is the most fundamental task in the information extraction system. Extraction of more complex structures such as relations and events depends on accurate named entity recognition as a preprocessing step. For example, in question answering system, candidate answer strings are often named entities that need to be extracted and classified first. Second, relation extraction is another important task in information extraction to detect and characterize the semantic relation between two entities into one of the predefined relation type. It is used in many application such as question answering to determine a precise answer, and then provide users with better search experience.

3.1. Medical Entities

For medical domain, a named entity is defined as a single word term or multi-words phrase that denotes a medical object, for instance a disease, symptom or drug.

Named entities specific to medical domain are called medical entities, as examples we can cite:

- Diseases or Problems; there are many diseases in the real world like Cancer, Alzheimer;
- Treatments like radiotherapy;
- Tests or medical exams like blood testing;
- Symptoms like fever and vomiting;
- Drugs or medicaments like Panadol and Humex.

The task of medical named entity recognition from clinical reports and medical records is an important task required not only in the Question Answering systems but even in Information Retrieval systems and other domains. Wang, (2009); extract clinical named entities in 11 entity types like (finding, procedure, body, substance… etc..) from clinical notes.

3.2. Relations between Medical Entities

Many applications in information extraction, natural language understanding, or information retrieval require an understanding of the semantic relations between entities. There are several types of semantic relations, grouped into two main families, paradigmatic relations and syntagmatic relations see the survey of (Berrazega, 2012).

3.2.1. Paradigmatic Relations

Paradigmatic relations are relations operating mainly on concepts of the same class. Usually, a hierarchical relation named vertical links represents these types of relations; which are used to organize concepts as a tree, like in thesaurus of Medical Subject Headings (MeSH) or metathesaurus of Unified Medical Language System (UMLS). Among this type of relation, we can mention relation of antonymy, synonymy and hypernymy.

3.2.2. Syntagmatic Relations

The task of medical relation extraction aim to extract relations between two (or more) medical entities it's under name of syntagmatic relations.

Syntagmatic relations are a semantic links occurring between two (or more) linguistic units present in an expression. They are identified by the study of syntactic forms in texts, and by a predicate; for example: we can cite specific relations in medical domain such as "*X should be treated by Y*" or "*X for the treatment of Y*".

There are many examples of relations such as developed in (Sun et al., 2013) to extract temporal relations, between the clinical events and temporal expressions.

3.3. Related Works

In this section we provide an overview of some previous efforts in MNER and relation extraction between those entities from medical reports.

The medical field has been the subject of several works; (Harkema et al., 2005) introduced AMBIT a text analysis system to facilitate access to patient clinical records. It involves mining radiology reports to extract signs of lung cancer, locations in the lung and relationships between these signs and locations expressed in the reports.

Ben Abacha & Zweigenbaum, (2011c) extract medical entities like problem, treatment, and test with a semantic method relying on MetaMapPlus based on UMLS using two English corpora i2b2 and Berkeley.

Embarek & Ferret. (2012) recognized medical entities like disease, symptom, medicament, exams, and treatment. They used rule-based method combined with morpho-syntactic patterns; and used the EQUER corpus (French scientific articles) downloaded from CISMeF website for evaluation. Barigou et al., (2011) developed MedIX as a tool for extracting entities and their properties from French clinical reports. In other works, those authors used a cellular automaton to extract medical information from clinical reports where rules used by MedIX are transformed into Boolean ones (Barigou et al., 2012).

Some works are interested by extraction of drugs properties like drug's name, drug's dosage, duration, frequency and reason like developed in (Spasicet al., 2010). Other works; focused on extraction of relations between diseases and drugs entities (Gurulingappa et al., 2012b).

Conditional Random Fields (CRF) has recently been shown to be well suited for MNER. This technique is used by (Huang & Hu, 2013) with semantic type of the UMLS meta-thesaurus to extract disease entities.

To extract semantic relationships between medical entities many authors have interested like relation between problem and treatments (e.g. cures, prevents, and side effect) as shown in (Ben Abacha & Zweigenbaum, 2011b), and between disease and treatment as in (Embarek & Ferret, 2012); authors used semi-automatic pattern-based approach to extract phrases from corpus and then select manually the phrases indicating the relation to extract.

Gurulingapa et al. (2012b) applied SVM to extract adverse drug effects from MEDLINE corpus, Minard et al., (2012) trained a SVM classifier to identify relations between problem and treatment and between problem and test, using i2b2 corpus.

3.4. Comparative Study

In this section, we analyze the related works according to the two tasks MNER and MRE based on the two approaches for IE described above; rule-based and machine-learning approaches.

3.4.1. Medical Entity Recognition

The Table 1 summarizes works using rule based and machine learning approaches; these works developed systems for MNER. The evaluation

Table 1. Representative works of medical entity recognition task

	Ref.	Contribution: Extraction of	Corpus	Techniques	Precision (%)	Recall (%)	F1-Score (%)
Rule based approaches	Harkema et al., (2005)	Signs of lung cancer and locations	English Radiology reports	AMBIT	69.00	83.00	75.00
	Spasic et al., (2010)	Drugs properties: name, mode/route, frequency, duration, reason	English i2b2 2009	Linguistic pattern and semantic rules	86.00	77.00	81.00
	Ben Abacha & Zweigenbaum (2011c)	Problem, treatment and test	English I2b2 2010	TreeTagger & MetaMapPlus	48.68	56.46	52.28
			Berkeley		23.43	42.47	30.20
	Barigou et al., (2011)	Patient's name, disease, symptom, medication,	French Clinical reports	TreeTagger & dictionaries	98.90	68.80	81,15
	Barigou et al., (2012)	Patient's name, disease, symptom, medication,		Cellular automaton Boolean inferring	92.00	89.00	86,89
	Embarek & Ferret, (2012)	Disease, symptom, medication, exams, treatment	French scientific articles	Morpho-syntactic pattern	90.00	84.00	86.00
Machine Learning approaches	Ben Abacha & Zweigenbaum (2011c)	Problem, treatment, test	English I2b2 2010	SVM	43,65	47,16	45,33
				BIO-CRF	70,10	83,31	76,17
	Huang et al., (2013)	Diseases	Biotext corpus	CRF	65,98	49,67	56,67

performance of each system is given in the same table; they made evaluation according to the precision, recall and F1-score.

Ben Abacha & Zweigenbaum (2011c) used I2b2 2010 and Berkeley corpora to extract three types of entities; *problem*, *treatment*, and *test*. The results obtained on the two corpuses are not on the same scale of performance; F1-score is 52.28% for I2b2 2010 and 30.20% for Berkeley. This is due to the characteristics of each corpus. The I2b2 corpus uses a quite specific vocabulary such as conventional abbreviations of medical terms and abbreviations of domain-independent words. The I2b2 corpus was annotated according to well-specified criteria to be relevant for the challenge, while Berkeley corpus was annotated with different rules and less control measures.

Embarek & Ferret (2012) developed a question answering system; the authors conceived morpho-syntactic patterns to extract five types of medical entities from French scientific articles. With the help of different dictionaries they succeeded in extracting *disease*, *symptom, medicament*, *exams* and *treatment* entities.

The same principle was adopted by (Barigou et al., 2011, 2012); they managed to extract entities like *patient's name*, *disease*, *symptom*, and *drug name* from French clinical reports. Evaluation is performed on a small corpus and the results show that the second system which is based on cellular automaton (Barigou et al., 2012) is able to cover more entities. The cellular automaton relies on a library of rules and a lexicon of proper nouns to identify entities, it gives very interesting results but need to be evaluated in a collection of more reports. These authors highlighted that the recall was low due to the insufficient rules to cover the diversity of expression of symptom, finding, and dosage.

Harkema et al.(2005); Spasic et al.(2010) used English corpus which is different in size and content, to extract different entities and categories; the system given by (Spasic et al., 2010) achieved 81% F1-score; they used three steps; linguistic pre-processing, pattern matching and template filing. This approach is primarily dictionary-based where authors used the meta-thesaurus of UMLS and assembled semi- automatically a dictionary of medications names to extract drugs properties like name, mode/route, frequency, duration, and reason. Harkema et al., (2005) introduced AMBIT; a processing framework for acquisition of medical and biological information from text; the information extraction process comprise three major stages lexical and terminological processing, syntactic and semantic processing, and discourse processing. The AMBIT system is used to extract different entities: signs cancers and locations, from radiology reports of lung cancer. They achieved 75% F1-score.

Machine learning approaches are also used to extract medical entities; such a system is developed by (Ben abacha and Zweigenbaum, 2011c) for extracting *problem*, *treatment* and test entities from the i2b2 2010 corpus. They used two different models namely SVM and CRF. The best results are obtained by CRF classifier using "Beginning Inside Outside" format (BIO) with lexical and morpho-syntactic features combined with semantic features.

Huang & Hu, (2013) developed a system to the disease named entity extraction using CRF classifier trained on orthographical, morphological and concept features of entities. They proposed a new method which uses the sentence level semantic contextual information as one of discriminative features for disease named entity recognition. The method takes advantage of semantic types related to disease in UMLS metathesaurus by fuzzy dictionary lookup. In this study only those concepts with semantic type of "DISEASE" or "SYNDROME" are kept. The results show that by adding "DISEASE" or "SYNDROME" semantic type as feature to train the CRF model, it achieves overall 0.72 increase of F1-score, with 1.05 and 0.52 increase in precision and recall.

For the hybrid method which combine rule based approaches with machine learning approaches there is one in (Ben Abacha & Zweigenbaum, 2011b); Conditional Random Fields (CRF) encoding with "Beginning Inside Outside" format (BIO) is combined with the semantic method MetaMapPlus to extract medical entities from I2b2 2010 corpus and it obtained 77.55% F1-score.

3.4.2. Discussion

Regarding the rule-based systems, we can observe two results from the Table 1; first, the work of Spasic et al., (2010) is the most effective MNER system which achieved a F1-score of 81,00% comparing with other woks using english corpora. Second, the work of Barigou et al. (2012) which abtained 86,89% is among the best systems using French corpora.

Generally speaking, recognition of medical entities give again good result when using hand coded rulesd than machine learning approaches; but in the work of (Ben Abacha and Zweigenbaum, 2011c), we can see that IE system performs better with machine learning method than rule based method. Using the i2b2 2010 corpus, the CRF model gives 76,17 F1-score instead of 52,28% in the case of rule-based system.

In the 2010 i2b2 /VA challenge an annotated reference standard English corpus was given for three tasks: medical concept extraction task, an assertion classification task and a relation classification task. Uzuner et al. (2011) presented state of the art of works participant using this reference standard with the evaluation for each task. They concluded that the machine learning-based systems participating could be improved with rule-based systems to determine medical concepts. Depending on the task, the rule-based systems can either provide input for machine learning or post-process the output of machine learning. For example, Ben Abacha & Zweigenbaum (2011b) combined CRF) classifier with the semantic method MetaMapPlus to extract medical entities from I2b2 2010 corpus and they obtained 77.55% F1-score, an improvement of 1,38% .

3.4.3. Medical Relation Extraction

Embarek & Ferret (2012) used a semi-automatic process to extract linguistic patterns of relations; they select phrases referring to a target relation and validate the presence of relation between entities. The system achieved 66% of F-score.

Ben Abacha & Zweigenbaum (2011b) used patterns constructed manually for extracting relation between disease and treatment entities like *cure*, *prevent* and *side effect* from Medeline corpus. Their system obtained 67.23% F-score.

The work of (Gurulingapa et al., 2012b) focuses on the adaptation of a SVM machine learning-based relation extraction system for the identification and extraction of drug-related adverse effects from MEDLINE case reports. The data set used for training and validation of the relation extraction system is the ADE corpus. The ADE corpus contains 2972 MEDLINE case reports that are manually annotated by three annotators. The corpus contains annotations of 5063 drugs, 5776 conditions (e.g. diseases, signs, symptoms), and 6821 relations between drugs and conditions representing clear adverse effect implications. For training their SVM classifier, Gurulingapa et al. (2012b) used dictionaries for the identification of drugs and conditions to generate false relations that do not fall into adverse effect relations. The system achieved an overall F-score of 0.87. The authors conclude that optimization of feature representation to include additional features for instance from syntactic sentence parse trees may further improve the results.

The SVM classifier was also used by (Minard et al., 2012) with lexical, syntactic and semantic features to extract disease-related treatment and disease-related test. The system achieved 70% of F-score.

Ben Abacha & Zweigenbaum, (2011b) extract semantic relations from medeline corpus, by combining a pattern-based method with a SVM machine learning-based relation. A multi-classification system achieved 93.73% of F-score and a mono-classification system obtained 94.07% of F-score.

3.4.4. Discussion

Uzuner et al. (2011) observed the lack of context in some of the relations found in the reference standard, indicating the possible use of domain knowledge in the annotation of these examples. In some other cases, the complexity of the language got in the way of relation extraction via machine-learning systems. The difficulty of classifying these relations comes from lack of explicit contextual information that describes the relations and/or the complexity of the language used in presenting the relations. While deeper syntactic analysis may help with the complex language, in the absence of context, domain knowledge may provide a good starting point.

The same comment as in MNER; most corpuses used by different authors in MRE system are in English (see Table 2). To date, there is no annotated corpus for French, thereby preventing French community to use machine learning techniques.

4. CONCLUSION

The area of medical research is attracting researchers, which explain its exploitation in several real applications.

Information extraction from electronic medical report is recent and is gaining more interest among doctors and researchers. Little work has been conducted on information extraction in the medical domain compared to IE done in the biomedical field.

In this paper we conducted a study on some relevant works concerned with IE in the medical domain. We have selected only works using medical reports.

Different approaches have been applied to tackle the problem of MNER and relations extraction, these are: rule based, dictionary matching based, and machine learning-based techniques.

We noted that the number of work performed with the rule-based approach is higher

Table 2. Representative works of medical relation extraction task

	Ref.	Contribution Extraction of	Corpus	Techniques	Precision (%)	Recall (%)	F-Score (%)
Rule based approaches	Ben Abacha et al., (2011b)	Disease-treatment relations	English MEDLINE 2001	Semi-automatic patterns	75.72	60.46	67.23
	Embarek & Ferret, (2012)	Semantic relations disease-treatment, disease-symptom, disease-drug, disease-exams	French scientific articles	Semi- automatic patterns	83.00	55.00	66.00
Machine Learning approaches	Ben abacha et al., (2011b)	Semantic relations disease-treatment	MEDLINE 2001	SVM multi-class Machine Learning	90,52	90,52	90,52
				SVM mono-class Machine Learning	91,96	91,03	91,49
	Gurulingapa et al., (2012b)	extraction of drug related adverse effects disease-drug	MEDLINE	SVM	86,00	89,00	87,00
	Minard et al., (2012)	Relation extraction disease-treatment, disease-test	I2b2 2010	SVM	80,00	63,00	70,00

than that of the machine learning approach, particularly in the community using the French language. We realized that this choice is due to the lack of annotated corpus particularly in the French community.

For the past years, systems have been developed using rule-based approaches. Here the use of different dictionaries is important to obtain good results. Updating these dictionaries is essential because for new wording in medical concepts. The rule based and dictionary based approaches lacks prediction power.

Actually, machine learning based approaches have demonstrated as been the most robust method for medical IE due to its capability of prediction of new wording based on learned patterns. SVM and CRF classifiers are the most used models and give good results compared to others machine learning techniques.

Different system are evaluated using different corpuses, however, to be able to interpret results and to compare those systems, they must use the same corpuses.

Each system offers a number of specifications, but we cannot say that one system is better than another since each system employs a different environment of evaluation. But globally we can see that the machine learning methods are currently the best.

REFERENCES

Alphones, E., Aubin,S., Bessieres, P., Bisson, G., Hamon, T., Lagarrigue, S., Nazarenko, A., Manine, A., Nedellec, C., Abdel vetah, M., Poibeau, T., & Weissenbacher, D. (2004). Extraction d'information appliquée au domaine biomédical: apprentissage et traitement automatique de la langue. *Presented at Actes de la Conférence Internationale sur la Fouille de textes (CIFT'04), La Rochelle, FRANCE.*

Ananiadou, S., Sampo, P., Tsujii, J., & Douglas, B. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, *28*(7), 381–390. doi:10.1016/j.tibtech.2010.04.005 PMID:20570001

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLSmetathesaurus: the MetaMap program. In AMIA AnnuSympProc, 17-21.

Ayuso, D., Boisen, S., Fox, H., Gish, H., Ingria, R., & Weischedel, R. (1992). BBN: Description of the PLUM system as used for MUC-4. *In Proceedings of the Fourth Message Understanding Conference (MUC-4)*, 169–176.

Barigou, F., Beldjilali, B., & Atmani, B. (2011). MedIX: ANamed Entity Extraction Tool from patient clinical reports. *International Conference on Communication, Computing and Control Application*, Hammamet, Tunisia, March 3-5, 488-494.

Barigou, F., Beldjilali, B., & Atmani, B. (2012). Using a cellular automaton to extract medical information from clinical Reports. *Journal of information processing system*, 8(1), 67–84.

Ben abacha, A., & Zweigenbaum, P. (2011a). Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, S5, Vol. 2, p1, doi: .10.1186/2041-1480-2-S5-S4

Ben abacha, A., & Zweigenbaum, P. (2011b). A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts. *Computational Linguistics and Intelligent Text Processing, 12th International Conference*, volume 6608 of Lecture Notes in Computer Science, February 20-26, Tokyo, Japan, 139-150.

Ben abacha, A., & Zweigenbaum, P. (2011c). Medical entity recognition: A comparison of Semantic and Statistical Methods. *Proceedings of the Workshop on Biomedical Natural Language Processing*, ACL-HLT, pages 56–64, Portland, Oregon, USA, June 23-24.

Berrazega, I. (2012). Temporal Information Processing: A Survey. [IJNLC]. *International Journal on Naturel Language Computing*, *1*(2).

Burcu, Y. (2007). Ontology-Driven Information Extraction. *Ph.D. Thesis. Vienna University of Technology.*

Campos, D., Matos, S., & Oliveira, JL. (2012). Biomedical Named Entity Recognition: A survey of Machine-Learning Tools. *lisence INTECH.*

Chang, C., Kayed, M., Girgis, MR., Shalan, K. (2006). A survey of web Information Extraction Systems. *IEEE transactions on knowledge and data engineering*, TKDE-0475-1104.R3

Chapman, W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, *18*(5), 540–543. doi:10.1136/amiajnl-2011-000465 PMID:21846785

Chen P, David H, Guoqing C.(2012). A rule based solution to co-reference resolution in clinical text. *J Am Med Inform Assoc,* 20:891–897. doi:10.1136.

Chinchor, N. A., & Marsh, E. (1998). Muc-7 information extraction task definition. *In Proceeding of the Seventh Message Understanding Conference (MUC-7)*, Appendices.

Dai, H.-J., Chen, C.-Y., Wu, C.-Y., Lai, P.-T., Tsai, R. T.-H., & Hsu, W.-L. (2012). Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. *Journal of the American Medical Informatics Association*, *19*(5), 888–896. doi:10.1136/amiajnl-2012-000808 PMID:22556185

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ace) program–tasks, data, and evaluation. *In Proceedings of LREC*, volume 4, 837–840. Citeseer.

Embarek, M., & Ferret, O. (2012). Esculape: Un système de question-réponse dans le domaine médical fondé sur l'extraction de relations. *TAL*, *53*(1), 69–99.

Fukuda, K., Tsunoda, T., Tamura, A., & Takagi, T. (1998). Toward Information Extraction: Identifying protein names from biological papers. *Proceedings of the Pacific Symposium on Biocomputing*. 1998:707-18.

Gurulingappa, H., Matteen-rajput, A., Robert, A., Flucky, J., Hofmann-Apitius, M., & Toldo, L. (2012a). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, *45*(5), 885–892. doi:10.1016/j.jbi.2012.04.008 PMID:22554702

Gurulingappa, H., Matteen-rajput, A., & Toldo, L. (2012b). Extraction of Adverse Drug Effects from Medical case Reports. In: Courtot M, editor. *International Conference Biomedical Ontologies*, 22-25. Graz, Austria.

Harkema, H., Ian, R., Gaizauskas, R., & Hepple, M. (2005). Information Extraction from Clinical Records. *In Proceedings of the 4th UK e-Science all Hands Meeting*, Nottingham, UK. CoxS.J. (ed.). EPSRC.

Hobbs, J. R., Appelt, D., Tyson, M., Bear, J., & Islael, D. (1992). SRI International: Description of the FASTUS system used for MUC-4. *In Proceedings fo the 4th Message Understanding Conference (MUC-4)*, 268–275.

Huang, Z., & Hu, X. (2013). Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. *International Journal of Machine Learning and Computing*. Vol.3, No. 6.

Jiang, J. (2012). Information Extraction from Text. Research Collection School of Information Systems. In Charu C. Aggarwal and ChengXiang Zhai (Eds.), Mining Text Data, Springer. 11-41. doi:10.1007/978-1-4614-3223-4_2

Jindala, P., & Dan, R. (2013). Extraction of Events and Temporal Expressions from Clinical Narratives. [US.]. *Journal of Biomedical Informatics*, *46*, S13–S19. doi:10.1016/j.jbi.2013.08.010 PMID:24022023

Kaiser, K., & Miksch, S. (2005). Information Extraction.A Survey.Vienna University of Technology.Asgaard-TR-2005-6.

Krupka, G., Jacobs, P., Rau, L., Childs, L., & Sider, I. (1992). GE NLTOOLSET: Description of the system as used for MUC-4. *In Proceedings of the 4th Message Understanding Conference (MUC-4)*, 177–185.

Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, *118*(1), 15–68. doi:10.1016/S0004-3702(99)00100-9

Meystre, S., Savova, G., Kipper-Schuler, K., & Hurdle, J. (2008). Extracting Information from Textual Documents in the Electronic Health Record: A Review of recent Research. *Yearbook of Medical Informatics*, 44–128. PMID:18660887

Minard, A., Ligozat, A., & Grau, B. (2012). Extraction de relations dans des comptes rendu hospitaliers. Dans Actes de IC2011, 22èmes Journées francophones d'Ingénierie des Connaissances, France.

Nadeau, D., &Sekine, S., (2007). A survey of named entity recognition and classification. *In journal of linguistic investigations*, 30(1), p .3-26.

Nedellec, C., & Nazarenko, A. (2005).Ontologies and Information Extraction.*LIPN Internal Report*. arXiv:cs/0609137.

Piskorski, J., & Yangarber, Y. (2013). Information extraction- past, present and future, *The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Multisource, Multilingual Information Extraction and Summarization, Publisher: Springer*, ISBN: 978-3-643-28568-4.

Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. *In Proc. of the 11th National conference on Artificial Intelligence*, 811–816.

Ritesh, S., & Suresh, J. (2014). Ontology-based information extraction: An overview and a study of different approaches. *International journal of computer Applications*, volume 87- N°4, 0975-8887.

Robert, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2007). The CLEF corpus: semantic annotation of clinical text. In *proceeding of the AMIA Symposium*, pp 625-629.

Sarawagi, S. (2007). Information extraction. *Foundations and Trends in Databases*, *1*(3), 261–377. doi:10.1561/1900000003

Spasic, I., Sarafraz, F., Akeane, J., & Nenadic, G. (2010). Medication information extraction with linguistic pattern matching and semantic rules. *Published by* group.bmj.com

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, *20*(5), 806–813. doi:10.1136/amiajnl-2013-001628 PMID:23564629

Uzuner, O., Brett, R. S., Shuying, S., & Scott, L. D. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556. doi:10.1136/amiajnl-2011-000203 PMID:21685143

Vicient, M. (2011). Ontology-based Information Extraction, *Master of Science thesis*.

Wang, Y. (2009). Annotating and recognising named entities in clinical notes. In: Proceedings of the ACL-IJCNLP 2009 student research workshop, ACLstudent '09. Stroudsburg (PA), USA: Association for Computational Linguistics; p. 18–26. doi:10.3115/1667884.1667888

Ware H, Charles J M, Vasudevan J, Oussama R.(2012). Machine learning-based coreference resolution of concepts in clinical documents. *J Am Med Inform Assoc*; 19:883e887. doi:.10.1136/amiajnl-2011-000774

Wimalasuriya, D., & Dejing, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, *36*(3), 306–323. doi:10.1177/0165551509360123

Yan X, Jiahua L, Jiajun W,Yue W, Zhuowen T,Jian-Tao S, Junichi T, Eric I-Chao C.(2012). A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. *J Am Med Inform Assoc,*19:897e905. doi:10.1136.

Yangarber, R., & Grishman, R. (1998). NYU: Description of the Proteus/PET system as used for MUC-7 ST. *In Proceedings of the 7th Message Understanding Conference: MUC-7, Washington, DC.*

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, *46*(6), 1088–1098. doi:10.1016/j.jbi.2013.08.004 PMID:23954592

Zheng, J., Wendy, W., Rebecca, S., & Guergana, K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, *44*(6), 1113–1122. doi:10.1016/j.jbi.2011.08.006 PMID:21856441

Zweigenbaum, P., &Tannier, X. (2013).Extraction des relations temporelles entre événements médicaux dans des comptes rendus hospitaliers. *les sables d'Olonne, TALN-Récital*, 17-21.

ENDNOTES

[1] http://www.nlm.nih.gov/research/umls/

[2] http://www.ihtsdo.org/snomed-ct/

[3] http://purl.bioontology.org/ontology/MESH

[4] If an element is recognised by the system but at the same time is not annotated manually it is overgenerated

Aicha Ghoulam graduated from Department of Computer Science, University of Chlef, Algeria. In 2010, she received his Magister degrees in Computer Science from Algiers University. She is currently a research member of Laboratory of Computer Science of Chlef. Her research interests include natural language processing, information extraction, information retrieval, knowledge-based system, pattern recognition.

Fatiha Barigou graduated from Department of Computer Science, University of Oran, Algeria. In 2012, she received his PhD degrees in Computer Science from the University of Oran. Dr. Barigou is currently a research member of Laboratory of Computer Science of Oran. Her research interests include natural language processing, information extraction, information retrieval, knowledge-based system, pattern recognition and data mining.

Ghalem Belalem graduated from Department of Computer Science, Faculty of Sciences, and University of Oran, Algeria, where he received PhD degree in computer science in 2007. He is now a research fellow of management of replicas in data replicas in data grid. His current research interests are distributed systems; grid computing, could computing and data grid placement of replicas and consistency in large scale systems and mobile environment.