Using Local Grammar for Entity Extraction from Clinical Reports

Aicha Ghoulam[1], Fatiha Barigou[2], Ghalem Belalem [3], Farid Meziane[4]

[1, 2,3]*Departmnt of Computer Science, Faculty of Exact and Applied Sciences, University of Oran, Ahmed BenBella, Algeria*
[4]*School of Computing, Science and Engineering, University of Salford, Manchester, UK*

***Abstract* — Information Extraction (IE) is a natural language processing (NLP) task whose aim is to analyze texts written in natural language to extract structured and useful information such as named entities and semantic relations linking these entities. Information extraction is an important task for many applications such as bio-medical literature mining, customer care, community websites, and personal information management. The increasing information available in patient clinical reports is difficult to access. As it is often in an unstructured text form, doctors need tools to enable them access to this information and the ability to search it. Hence, a system for extracting this information in a structured form can benefits healthcare professionals. The work presented in this paper uses a local grammar approach to extract medical named entities from French patient clinical reports. Experimental results show that the proposed approach achieved an F-Measure of 90. 06%.**

***Keywords* — Information Extraction, Electronic clinical reports, medical entities recognition, natural language processing.**

I. INTRODUCTION

RECOGNITION and classification of named entities in texts is recently considered as an important task in automatic natural language processing (NLP) as they play a significant role in various types of NLP applications, especially in Information Extraction, Information Retrieval, Machine Translation, Question-Answering and , Parsing/Chunking.

The amount of information written in natural language and available in electronic forms is increasing, making the development of intelligent tools to process this information an urgent need for practitioners such as health care professionals. Information Extraction is gaining an increased attention by researchers, who seek to acquire knowledge from this huge amount of natural language content. Many approaches have been proposed to extract valuable information from texts in different fields, with the medical domain being one of them.

We note that the volume of medical information is constantly increasing. According to [1] it doubles every five years and this wealth of information is difficult to access because it is stored in unstructured formats. This is particularly true for the case of clinical reports (CRs) where information such as pathologies, medical history and diagnoses are recorded in a textual format, is ever increasing and becoming difficult to search and access. CRs can have a positive impact on the quality of care, patient safety and efficiencies in medical procedures. However, without accurate and appropriate content in a usable and accessible form, these benefits may not be achieved. Developing a system for extracting unstructured information can benefits healthcare professionals.

These kinds of systems have become very necessary tools; they will enable researchers to access accurate data and the required information, and reducing the time spent by doctors on making decisions about patients' diseases. Hence, the main motivation of this work is to develop an automated system for extracting named entities from clinical reports.

Firstly, most of the elements in CRs are name entities (e.g.., the names of patients, diseases, symptoms, and drugs) that can be used in various applications, such as seeking information to diagnose new patients, conducting epidemiological studies, statistical analysis, and data mining. However, these CR are difficult to analyze due to their unstructured nature and the large volume of records available. Secondly, most existing medical entities extraction systems are devoted to English. Research in the French language is still at its initial stages [2].

In this research, we propose to use an original approach based on local grammars to extract medical entities from French clinical reports. The local grammar based approach has recently been applied to extract proper nouns in English, Chinese, French, Korean, Portuguese, and Turkish news texts [3]. This approach was first used to discuss recursive phrases that are commonly found in specialist literature like biochemistry and then extended to extract date, time and address expressions from letters.

In this work, we study the application of local grammars for extracting medical entities from French clinical reports. We focus on the extraction of the following named entities; disease, symptom, treatment, drugs and clinical reviews. The rest of the paper is organized as follows Section 2 summarizes the task of named entity extraction and work related to the medical field. In section 3 we describe information extraction and methods. Section 4 describes the proposed system and our contribution to extract medical entities. Section 5, presents evaluation results concerning proposed system performance.

DOI: 10.9781/ijimai.2015.332

Finally, this paper ends with conclusions and some ideas for future works.

II. RELATED WORKS

The Named Entity Extraction (NEE) task aims to recognize named entities and classify them into categories like organization names, person names, location names, date and time expressions, monetary amounts and documents' references [4]. Named Entity Extraction systems are based on two main approaches: the rule based and the machine learning approaches [5]. Hybrid systems combine these two approaches [6].

The Rule-Based approach is a manual technique based on a specific domain extraction rules written by an expert using morphological and syntactic information like trigger words, capitalization, and gazetteer. This approach gives very good results however requires great human efforts and a considerable time for data analysis and rule writing. It is time consuming during development and it lacks portability which limits its extension to other domains.

On the other hand, the machine learning approach, is a trainable technique that is capable to improve its ability to extract information from input automatically or under supervision, but it requires large annotated corpora for training, which are both expensive and time-consuming to train the models [7]. Many different models have been proposed over the years. The most prominent of these are Hidden Markov Models (HMM) [8], Support Vector Machine (SVM) and Conditional Random Fields (CRF) [9].

Several studies have used the NEE task, [5,7,10]. Most systems were mostly interested with named entity like organization names, place names, date expressions and numeric expressions [11] with different languages [12] and gave promising result. Recently, NEE has been applied to the medical field to extract entities such as protein names, gene names, disease names and treatments from medical documents [7]. Various systems have been developed, using rule-based approaches, including MedLEE [13], SymTex [14], MetaMap [15] and MedIX [16].

The MedIX system [16] was applied to patient CRs using natural language processing techniques. It performs some processes such as preprocessing the text, tokenizing, and tagging, recognizing special formatting and then it identifies entities and classifies them into categories that included patient name, disease name and symptom names. Others classify entities into problem, treatment, test classes [9] and drug properties [17].

Authors in [18] proposed an approach relying on linguistic pattern and canonical entities to extract five categories of medical entities from CRs namely, disease name, treatment name, drug name, and test and symptom names. Other systems extract useful entities from radiology and mammography reports to identify patients with lung cancer [19] or with tuberculosis [20].

Recent studies are mostly based on the machine learning approach; [1] and [21] employ support vector machines to attribute semantic categories to each word in discharge summaries. Markov models-based methods are also frequently used [8]. Others used unsupervised methods were based on seed term collection [22].

In the past couple of years, researchers have been exploring the use of machine learning algorithms in the clinical concept detection. To promote the research in this field many organizations such as ShARe/CLEF, SemEval have organized a few clinical NLP challenges. Both rule based [23,24,25] and machine learning based methods as well as hybrid methods [26,27,28,29] were developed. In this shared-task sequential labeling algorithms (i.e., CRF) [30,31,32,33,34,35], and machine learning methods (i.e., SVM) [36] have been demonstrated to achieve promising performance when provided with a large annotated corpus for training.

The system that was top-ranked in the SemEval 2014 Task 7 among all participating teams is given in [32]; authors developed three ensemble learning approaches for recognizing disorder entities consisting of an ensemble learning-based approach and a Vector Space Model based method for disorder entity encoding. Extracted features from clinical notes were used to train two machine learning algorithm-based entity recognition models, CRF and Structural Support Vector Machines (SSVMs). These two models were ensembled with MetaMap. Their approaches achieved top rank in both subtasks (disorder entities recognition and encoding), with the best F-measure of 81.3% for entity recognition and the best accuracy of 74.1% for encoding, indicating that their proposed approaches are promising.

Another work [37] presented a comparison of two approaches to automatically de-identify medical records written in French; rule based system and CRF based system. They achieved an F-measure of 84.3% and 88.3% for each system respectively in cardiology reports. They achieve an F-measure of 68.1% and 63.8% for each system respectively in fœtopathologie reports. They concluded that a rule based system allowed them to achieve good results on nominative and numerical data, and the machine learning approach performed best on more complex categories.

III. INFORMATION EXTRACTION AND METHODS

Information Extraction (IE) has been defined in the literature by many researchers [38, 39]. 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 [5]:

- Named Entity Recognition (NER), involves the recognition of named entities such as organizations, persons, locations, dates and monetary amounts. 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 includes for example relation between disease and drug.
- Co-reference 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 [40].
- 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.
- Event Description, [41] 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.

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

- The Rule-Based IE methods: rule-based methods extract the information by rules, and these rules can be generated by human hand-coded, or by learning from examples. The most representative examples of this kind of systems are FASTUS [42], GE NLTOOLSET [43], PLUM [44] and PROTEUS [45]; these systems are well described in [46]. 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, it involves a great human effort and a considerable time for data analysis and rule writing. It is time consuming during development [55].
- 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 [5]. 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 (HMM), Maximum Entropy Markov Models (MEMM), SVM or even a vector classification model for which the features are not terms, but graph metrics [47] and CRF. Other studies used unsupervised machine learning methods; a class of problems in which one seeks to determine how the data are organized such as clustering; a common technique for statistical data analysis used in many fields as used in [48].
- 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 [49]; it usually performs a pattern matching procedure like a form of finite-state machines which relies on a set of extraction rules.
- IE using Ontology: Ontology is a formal and explicit specification of a shared conceptualization; it plays a crucial role in the process of IE. The relation between ontologies and IE is involved in two tasks: on the one hand, Ontology is used for information extraction; IE needs ontologies as part of the understanding process for extracting relevant information [50]; on the other hand, information extraction is used for populating and enhancing a domain ontology from the web as shown in [51]; they developed an ontology of a scene from the essential semantic components for the semantic structuring of the Web3D. The construction of ontology for the definition of tridimensional spaces will allow the Web3d to standardize the development of scenarios and the creation of manufacture agents that will make easier the modeling and texturing processes.

IV. PROPOSED APPROACH

In this study; we use and evaluate a rule based approach relying on local grammar the motivation and the description of this approach is presented in this section.

A. Benefits of the proposed system

Fig.1 show some benefits of such system for clinical staff. An UML use case diagram is used to describe the expected functionalities of the proposed system. Medical named entities recognition, as shown in Fig.1, is essential to built new systems to help doctor and clinical staff in their work. Doctors need quick and easy access to quality information resources to be able to make informed decisions regarding patient care; they also need systems to help them answer clinical questions.

1) Question-Answering systems:

i. Clinical staff asks to obtain medical response.
ii. A research in medical ontology must be done.
iii. The construction of medical ontology based on medical entity recognition and relation extraction between medical entities.

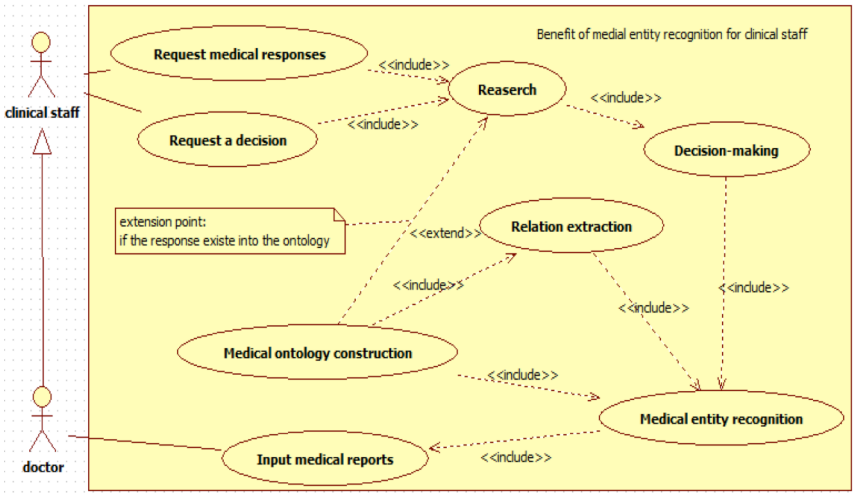

Fig.1. Medical entity recognition's benefits

iv. The extraction of relation between medical entities task necessities that the medical entities must be chucked.

2) Decision support system:

i. Clinical staff requests a decision.
ii. A research in past problems is done to make decision.
iii. Past problems input by doctors must be checked using medical entity recognition to facilitate research for similar cases.

B. Local grammar based approach

The Local Grammar (LG) approach was initiated by Harris [52] to discuss recursive phrases that are commonly found in specialist literature like biochemistry (immunology) [53]. Harris defines a local grammar as a way of describing syntactic restrictions of certain subsets of sentences which are closed under some or all of the operations in the language.

More specifically, LG is a way of recognizing the behavior of words that are used in a specific domain, finding how these words are used in sentences and inferring their usage patterns.

For example, Traboulsi [53] considered frozen expression as a subset of sentences that have some syntactic restrictions.

Certain expressions such as 'compound words' (e.g. stock market) are strictly frozen and others are partially frozen and are included in expressions such as the director of a small company, the director of a doctoral thesis as illustrated in the following patterns:

(financial + stock + E) market
Director of (company + thesis)
The 20 March (next + 2006)

Local grammar were extended by Gross [54] to extract date, time and address from letters. Gross defined LG as a finite state grammar and used it for finding words related by prefixation, suffixation, and sentences having similar syntax.

For certain expressions such as dates, times, and other types of proper names, it appears impossible to individually identify the set of all possible constructions and much more effective a

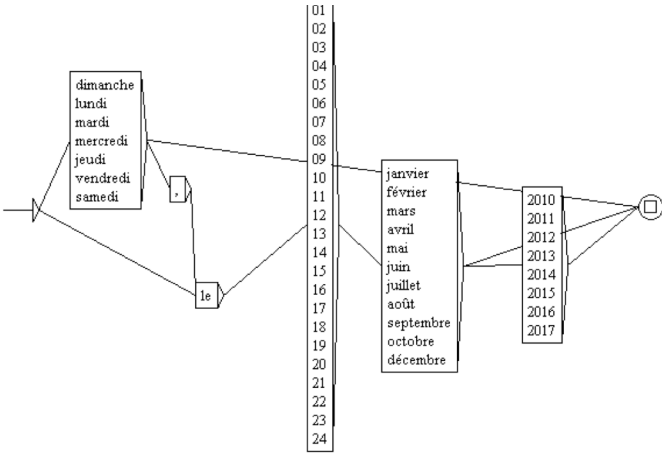

Fig.2. Example of a local grammar for French date expressions

representation in the form of automata. This representation is easy to be read of course if the graphs are well arranged. We give in Fig.2 an example of a local grammar for French date expressions.

It can recognize expressions like: "*dimanche, le 02 septembre 2014*".

Local grammars as finite state local automata have been used by [3] to recognize English person names in textual documents and then extended it to extract Arabic person names in [24].

C. Local Grammar based Approach for Extracting Named Medical Entities

In this work we study French CR to extract medical named entities using local grammar. In table 1, we gave the classes of entities and examples for each one.

We noticed that medical entities occur frequently at constructions having consistent structures in the proximity of Reporting Words (RWs) like "*consulte pour" (*consulting for), "*présentant*" (having) in the case of disease entities, "*signe de*" (sign of) in the case of symptom entities which are sufficiently frozen to be described in the form of local grammars. An example of these local grammars is shown in Fig. 3.

TABLE I
MEDICAL NAMED ENTITIES EXAMPLE

ENTITY	MEDICAL ENTITY EXAMPLE
Disease	Masse du pancreas (Mass of the pancreas)
Symptom	Anorexie (anorexia) Amaigrissement (weight loss) Déshydratation (dehydration)
Clinical Review	Scanner AP échographie Abdomino-Pelvienne (abdomino- pelvic ultrasound)
Treatment	Alimentation orale légère (Lightweight oral feeding) réhydratation 1 fl (rehydration 1 bottle)
Medication	Cefacidal 1g , Gentamicine 80 mg, Flagyl 1 fl

This graph is able to recognize constructions like:

- [Un malade nommée X présente une **fistule de fémur droite**]
  (A *patient named X has a* ***right femoral fistula)***
- [*Un malade Y consulte pour un* ***traumatisme lombaire***]
  (*A patient named Y consults for* ***lumbar trauma***)

The boxes labeled <disease>, <organ>, <location>, <adjective> are the names of sub-graphs that recognize candidates of disease names, organ names (anatomy), location, and adjectives respectively. Local grammar graphs containing sub-graphs shows similarity to recursive transition networks.

To extract medical entities from French clinical reports written in a free and natural language, our contribution adopts the following approach:

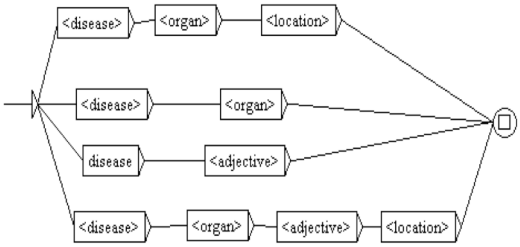

Fig.3. Example of the local grammars of disease entity

- Construction of different Gazetteers;
- Construction of medical entities classification rules;
- Describing the rules in the form of local grammars.

D. System Architecture

Figure 4 shows the architecture of the system. Our system has two major components: the gazetteers and the Grammars.

Pre-processing task:

It is necessary to properly delimit the clinical report into meaningful units. Most natural language processing solutions expect their input to be segmented into sentences, and each

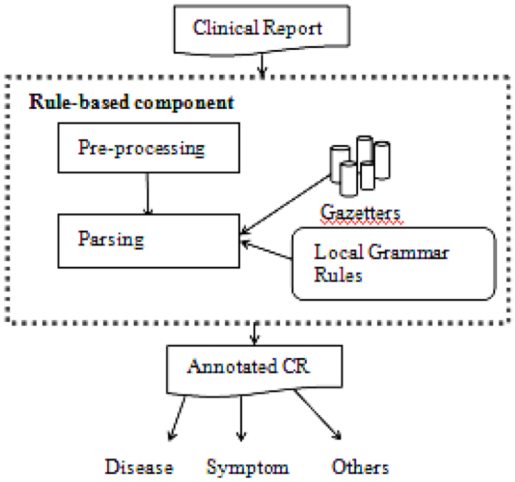

Fig.4. Architecture of the system

sentence into tokens; so for that we used the Unitex[1] open source for splitting CR into sentences and splitting sentences into tokens.

Gazetteers:

The gazetteer contains diseases names, symptoms names, clinical reviews, treatment and medications, medical adjectives, organs and so on. These Dictionaries are in electronic format; we have assembled them from different web site:

- A dictionary of adjectives[2] containing 514 entries.
- A dictionary of organ (Atlas: human body)[3] containing 384 entries.
- A dictionary of diseases[4,5] containing 343 entries.
- A dictionary of treatments[6].
- A dictionary of clinical reviews[6] containing 28 entries
- A dictionnaire of symptoms[7] containing 67 entries
- A dictionary of drugs[8,9]
- A list of French medical reporting words or trigger words
- A dictionary of medical names.

Grammars:

The grammar performs recognition and extraction of medical entities from clinical reports based on combination of regular expression patterns in the form of local grammars. A deep contextual analysis of various French clinical reports was performed using the Unitex open source software to build local grammars based on keywords or trigger words forming a window around medical entities.

1 http://www -igm.univ-mlv.fr/~unitex
2 http://www.linternaute.com/dictionnaire/fr/definition/abdominal/
3 http://www.doctissimo.fr/html/sante/atlas/index.htm
4 http://www.passeportsante.net/problemes-et-maladies-p69
5 http://www.vulgaris-medical.com
6 http://www.e-sante.fr/
7 http://www.vulgaris-medical.com/symptomes
8 http://www.eurekasante.fr/medicaments/alphabetique
9 http://www.doctissimo.fr/html/medicaments/medicaments.htm

TABLE IV
DETAILED EVALUATION ON THE CLINICAL REPORTS.
PRECISION (P), RECALL (R) AND F-MEASURE (F)

CATEGORY	P	R	F
Disease	0,921	0,800	0,856
Symptom	0,971	0,917	0,943
Treatment	1,000	0,765	0,867
Clinical Review	1,000	0,941	0,969
Drug	1,000	0,765	0,867

Example rule:

The following rule recognizes a disease name composed of medical name followed by a medical adjective and human organ based on a proceeding disease indicator pattern which is the RW.

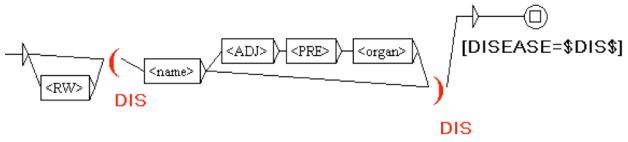

Fig.5. Example of Local Grammar in Unitex

(*name + ws+ adjectives +French prepositions + ws + organ(s)*)

Corresponding Local Grammar:

The following local grammar corresponds to the above rule using the Unitex editor:

Writing conventions:

- ws: whitespace.
- Name: dictionary of medical names.
- ADJ: dictionary of medical adjectives.
- PRE: dictionary of French prepositions.
- Organ: dictionary of human organ.

Example:

The following disease name would be recognized by the above local grammar shown in Fig.5:

"Masse tumoral du colon."; [Tumor mass of the colon]

We created a set of rules using Unitex to classify different medical named entities into disease, symptoms, clinical review, drugs and treatment from French clinical reports. Some examples of rules for each class are given in the table II below:

TABLE II
MEDICAL NAMED ENTITIES RULES EXAMPLE

CATEGORY	MEDICAL ENTITY EXAMPLE	RULE EXAMPLES
Symptom	Anorexie	(symptom name)
Clinical Review	Scanner AP	(test name)
Treatment	réhydratation 1 fl	(treatment name + ws + nbr + ws +unit)
Medication	Cefacidal 1g	(name drug+ws+nbr+unit)

V. EXPERIMENTAL STUDY

In this section we describe the data and metrics used to test our approach experimentally and discuss the different results.

A. Data set: clinical reports

We analyzed more than 50 French clinical reports to construct rules for medical named entities, and evaluated the system by using 30 new clinical reports from urology patients and general medicine at the hospital of CHLEF (Algeria). We have annotated the dataset with the help of a doctor. Five classes of medical entities were studied: Disease, Symptom, Treatment, Clinical review, Drug or medication. (so, 80 clinical reports have been collected in total: 50 for the development of rules and 30 for the evaluation of the system)

B. Metrics

Standard metrics for evaluating named-entity extraction are used to measure the accuracy of the proposed approach. We calculate precision, recall, and F-measure. They are defined as:

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F - Measure = 2 * (Precision * Recall)/(Precision + Recall)

Where:

- TP: True Positives; number of medical entities that were identified correctly.
- FP: False Positives; number of medical entities that were detected by the system and were not present in the report.
- FN: False Negatives; number of medical entities that were present in the report but system failed to detect them.

Table III describes in more details those metrics.

TABLE III
EVALUATION METRICS

		EXPERT (DOCTOR)	
		YES	NO
SYSTEM	YES	TP	FP
	NO	FN	TN

C. Experimental Results

In this study, we experiment the approach we have described in section 3 to recognize medical entities from clinical reports. Five categories were studied and the results are discussed in this section.

Fig. 6 shows the precision, recall and F-measure for each class. Analysis of the experiments allowed us to observe that the overall performance of our system over the five categories is good. The results are shown in table IV below.

The insufficient coverage of the diversity of all medical entities in our small set of rules explains the low results in recall. The system failed to recognize entities due to the insufficient numbers of entries in dictionnaries and insufficient

rules for identifying different entities especially, treatment and drugs entities.

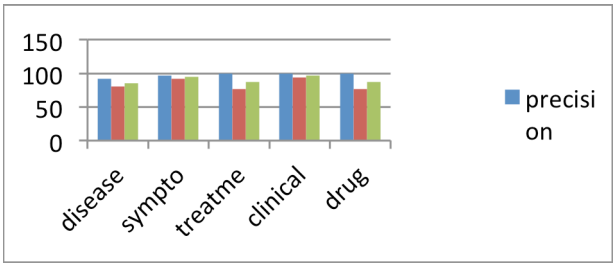

Fig.6. Performance system

Generally, the system performs well achieving and it gives a macro precision of 97,84% and a macro-recall of 83,78% which are the averge as it's shown in table V.

TABLE V
MACRO-AVERAGE RESULTS FOR PRECISION (P), RECALL (R) AND F-MEASURE (F) OF OUR SYSTEM.

EVALUATION	P %	R %	F %
AVERAGE	97,84	83,78	90,06

These results are very interesting and need to be evaluated in a larger collection of clinical reports, and this is very important.

VI. CONCLUSION

The work done in this paper is an attempt to broaden the coverage for medical entity extraction by incorporating the French clinical reports.

We used a rule based approach relying to the local grammar to extract medical entities from French clinical reports. The experimentations show that the rule based approach allows obtaining a good precision, but having a disadvantage to require a great human efforts and a considerable time compared to the high variability and the complex structure of the clinical reports.

One of the most important obstacles in identifying medical entities is the high terminological variation in the medical domain. In other hand the evolution of entity naming such as new abbreviations, names for new diseases or drugs constitute obstacles which can limit the scalability of the local grammar approach. Also the main limitation of the approach is their lack portability which limits their extension to other medical domains.

We plan to extract medical entities by machine learning, starting from a collection of training examples; the idea is to study the features of positive and negative examples of medical entities to be extracted over a collection of annotated documents with the need of doctor and design rules that capture instances of a given type. Therefore the hybridization will be a performance evaluation for future work.

REFERENCES

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

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

[3] H. N. Traboulsi, "Named Entity Recognition: A Local Grammar-based Approach", Ph.D. dissertation, Dept of Computing, Surrey Univ. Guild ford, U.K, 2006.

[4] T. Poibeau, "Boosting the robustness of a named entity recognizer", International Journal of Semantic Computing, 2009, 32(1), pp 77-98.

[5] D. Nadeau, S. Sekine, "A survey of named entity recognition and classification", journal of linguistic investigations, 2007, 30(1), p .3-26.

[6] M. Mohammed Oudah, K. Shaalan, "A pipeline Arabic Named Entity Recognition Using a Hybrid Approach", in proceedings of COLING 2012, Mumbai: Technical Papers, pp 2159–2176.

[7] S. Meystre, G. Savova, K. Kipper-Schuler, J. Hurdle, "Extracting Information from Textual Documents in the Electronic Health Record: A Review of recent Research", year book of Medical Informatics. 2008, pp. 44-128.

[8] Y. He, M. Kayaalp. "Biological entity recognition with Conditional Random Fields.", In AMIA Annu Symp Proc, pp 293-297, 2008.

[9] F. Barigou, B. Beldjilali, B. Atmani, "MedIX : A Named Entity Extraction Tool from patient clinical reports", International Conference on Communication, Computing and Control Application, Hammamet, Tunisia, March 3-5, 2011, pp.488-494 .

[10] M. Chau, J., Xu, H. Chen, "Extracting Meaningful Entity from Polices Narrative Reports", Proceeding of the National Conference for Digital Government Research, 2002, pp.271-275

[11] L. Kosseim, G. Lapalme, "EXIBUM: un système expérimental d'extraction d'information bilingue", Rencontre International sur l'extraction, le filtrage et le résumé automatique (RIFRA'98), 1998.

[12] K. Shaalan, "Person Name Entity Recognition for Arabic", Proceedings of the 5th workshop on important Unresolved Matters, p 24-17, 2007.

[13] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson, "A general natural language text processor for clinical radiology", Journal of the American Medical Informatics Association, 1994, 1(2), pp.161-174.

[14] P. Haug, L. Christensen, M. Gundersen, B. Clemons, S. Koehler, K. Bauer, "A natural language parsing system for encoding admitting diagnose ", American Medical Informatics Association Annual Symposium, AMIA 97, 1997, pp.814-818.

[15] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Meta thesaurus: the MetaMap program", American Medical Informatics Association Annual Symposium, AMIA'01, Washington, DC, USA, 2001, pp.17-21.

[16] A. Ben Abacha, P. Zweigenbaum, "Medical entity recognition: A comparison of Semantic and Statistical Methods", In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT, pages 56–64, Portland, Oregon, USA, June 23-24.

[17] I. Spasic, F. Sarafraz, J. Akeane, G. Nenadic, "Medication information extraction with linguistic pattern matching and semantic rules", Published by group.bmj.com, 2010.

[18] M. Embarek, O. Ferret, "Learning patterns for building resources about semantic relations in the medical domain", Proceedings of the International Conference on Language Resources and Evaluation, LREC'08, Marrakech, Morocco, 26 May - 1 June, 2008.

[19] H. Harkema, R. Ian, R. Gaizauskas, M. Hepple (2005). Information Extraction from Clinical Records. In Proceedings of the 4th UK e-Science All Hands Meeting http://www.allhands.org.uk/2005/proceedings/,2005.

[20] C. A. Knirsch, N. Jain, A. Pablos-Mendez, C. Friedman, G. Hripcsak, "Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Clinical Decision Support System". Journal Infection Control and Hospital Epidemiology, 1999, 19(2), pp.94-100.

[21] T. Sibanda, T. He, P. Szolovits, O. Uzuner, "Syntactically-informed semantic category recognition in discharge summaries", Proceeding of the Fall Symposium of the American Medical Informatics Association; Washington, DC, November, 2006.

[22] S. Zhang, N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts", Journal of Biomedical Informatics 46, 2013, p 1088-1098.

[23] J. Fan, N. Sood, Y. Huang. "Disorder Concept Identification from Clinical Notes An Experience with the ShARe/CLEF 2013 Challenge", Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, 2003, Valencia - Spain.

[24] S. Matos, T. Nunes, J. L. Oliveira. "BioinformaticsUA: Concept Recognition in Clinical Narratives Using a Modular and Highly Efficient Text Processing Framework", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, August 23-24, 2014, pages 135–139.

[25] S. Ramanan, S. Nathan. "RelAgent: Entity Detection and Normalization for Diseases in Clinical Records: a Linguistically Driven Approach", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 477–481, Dublin, Ireland, August 23-24, 2014.

[26] Y. Xia, X. Zhong, P. Liu, C. Tan, S. Na, Q. Hu and Y.Huang. " Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1", OnlineWorking Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, 2013, Valencia -Spain.

[27] J. D. Osborne, B. Gyawali, T. Solorio. "Evaluation of YTEX and MetaMap for clinical concept recognition", Online Working Notes of the CLEF 2013 Evaluation Labs andWorkshop, 23 - 26 September, 2013, Valencia - Spain.

[28] P. Pathak, P.Patel, V.Panchal, N. Choudhary, A. Patel, G. Joshi. "ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 278–283, Dublin, Ireland, August 23-24, 2014.

[29] K. Gojenola, M.Oronoz, A. Pérez, A. Casillas. " IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 361–365, Dublin, Ireland, August 23-24, 2014.

[30] A. Bodnari, L. Deleger, T. Lavergne, A. Neveol, P. Zweigenbaum. "A Supervised Named-Entity Extraction System for Medical Text", Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 – 26 September, Valencia - Spain.

[31] A. Parikh ,Ah PVS, J. Mustafi, L. Agarwalla, A. Mungi. "ThinkMiners: Disorder Recognition using Conditional Random Fields and Distributional Semantics", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 652–656, Dublin, Ireland, August 23-24, 2014.

[32] Y. Zhang, J.Wang, B.Tang, Y.Wu, M. Jiang, Y. Chen, H. Xu. "UTH_CCB: A Report for SemEval 2014 – Task 7 Analysis of Clinical Text", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 802–806, Dublin, Ireland, August 23-24, 2014.

[33] G.Attardi, V. Cozza, D.Sartiano. "UniPi: Recognition of Mentions of Disorders in Clinical Text", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 754–760, Dublin, Ireland, August 23-24, 2014.

[34] J.Jonnagaddala, M. Kumar, H.J. Dai, E. Rachmani, C.Y. Hsu. "TMUNSW: Disorder Concept Recognition and Normalization in Clinical Notes for SemEval-2014 Task 7", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 663–667,Dublin, Ireland, August 23-24, 2014.

[35] G. Omid, R.J. Kate. "UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 828–832, Dublin, Ireland, August 23-24, 2014.

[36] J.Cogley, N. Stokes, J. Carthy. " Medical Disorder Recognition with Structural Support Vector Machines", Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 – 26 September, Valencia - Spain.

[37] C. Grouin, P. Zweigenbaum, "Automatic de-identification of French clinical record: comparison of rule based and machine learning approaches", In Proc MEDINFO 2013, Studies in Health Technology and Informatics, pages 476-480. Amsterdam, IOS Press, 2013.

[38] S. Sarawagi, "Information extraction. Foundations and Trends in Databases". (2007). Vol. 1, No. 3. 261–377.

[39] J. Jiang," Information Extraction from Text". Research Collection School of Information Systems. In Charu C. Aggarwal and ChengXiang Zhai (Eds.), (2012). Mining Text Data, Springer. 11-41.

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

[41] W.Sun, A. Rumshisky , & O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge". (2013). In the Journal of the American Medical Informatics Association. doi:10.1136/amiajnl-2013-001628.

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

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

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

[45] 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.

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

[47] H Cordobés,., A. Fernández Anta, L. F. Chiroque, F. Pérez, T. Redondo, and A. Santos, "Graph-based Techniques for Topic Classification of Tweets in Spanish", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 2, issue Special Issue on AI Techniques to Evaluate Economics and Happines, no. 5, pp. 31-37, 03/2014.

[48] K. Khan,and A. Sahai, "A fuzzy c-means bi-sonar-based Metaheuristic Optimization Algorithm", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 1, issue Regular Issue, no. 7, pp. 26-32, 12/2012.

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

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

[51] H.Bolivar-Baron,., R. Gonzalez-Crespo, and O. Sanjuan-Martinez, "Ontology of a scene based on Java 3D architecture.", International Journal of Interactive Multimedia and Artificial Inteligence, vol. 1, issue Special Issue on Business Intelligence and Semantic Web, no. 2, pp. 14-19, 12/2009.

[52] Z. Harris, "Theory of language and Information: A Mathematical Approach", Oxford & New York: Clarendon Press, 1991

[53] H. N. Traboulsi, "Arabic Named Entity Extraction: A Local Grammar–based Approach", Proceeding of the International Multiconference on Computer Science and Information Technology, 2009, pp. 139-143.

[54] M. Gross, "The construction of local grammars", in E.Roche & Y. Schabés (eds), Finite-State Language, Speech, and communication, MIT Press, 1997, pp.329-354.

[55] S. J. Bolaños-Castro, R. G. Crespo, V. H. Medina-García, "Patterns of software development process", International Journal of Interactive Multimedia and Artificial Intelligence, vol 1. Issue 4, pp. 33-40, 12/2011

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 exact and applied sciences, University of Oran, Algeria, where he received PhD degree in computer science in 2007. His current research interests are distributed system; grid computing, cloud computing and data grid placement of replicas, consistency, fault tolerance, economic models, energy consumption, Big data, and improved performance in large scale systems and mobile environment.

Farid Meziane Professor in Data and Knowledge Engineering. He holds a PhD in Computer Science from the University of Salford and is the head of the Informatics Research Centre. He is the Chair of the 20th International Conference on Application of Natural Language to Information Systems (NLDB2015) and has served on the programme committees of over 20 conferences. He is on the editorial board of 5 international journals. His research interests are in the broad area of data and knowledge engineering. This includes data mining, information extraction and retrieval, big data and the semantic web.