Table Des Matières

FSEA …../ANNEE 1
RESUME 2
ABSTRACT 4
REMERCIEMENTS 5
DEDICACES 6
TABLE DES MATIERES 7
LISTE DES TABLEAUX 10
LISTE DES FIGURES 11
INTRODUCTION ET PROBLEMATIQUE 12
1. INTRODUCTION 13
1.1. Contexte de la thèse 13
1.2. Problématique 14
1.3. Contributions 15
1.4. Liste De Publications 16
1.5. Organisation de la thèse 17
PREMIERE PARTIE : ÉTAT DE L'ART 18
2. EXTRACTION D'INFORMATION MEDICALE 19
2.1. Introduction 19
2.2. L'extraction d'information 20
2.2.1. Définition 20
2.2.2. Les tâches d'extraction d'information 20
2.3. La Reconnaissance des Entités Nommées 21
2.3.1. Les classes des entités nommées 22
2.3.2. Les difficultés de la catégorisation des entités nommées 22
2.4. Extraction des Relations Sémantiques 23
2.4.1. Relations paradigmatiques 24
2.4.2. Relations syntagmatiques 24
2.5. Les Méthodes d'extraction d'information 24
2.5.1. Les méthodes à base de règles 24
2.5.2. Les méthodes d'apprentissage automatique 25
2.5.3. Les méthodes hybrides 27
2.5.4. Les méthodes à base d'ontologies 27
2.6. Mesures d'évaluation des Systèmes d'extraction d'information 28
2.7. Travaux sur l'extraction d'information médicale 29
2.7.1. Reconnaissance des entités médicales 29
2.7.1.1. Systèmes à base de règles 29
2.7.1.2. Systèmes à base d'apprentissage automatique 31
2.7.1.3. Systèmes utilisant l'approche hybride 33
2.7.2. Extraction des relations médicales 34
2.7.2.1. Systèmes à base de patrons 34
2.7.2.2. Systèmes par apprentissage automatique 36
2.7.2.3. Systèmes utilisant l'approche hybride 37
2.8. Conclusion 37
3. LES ONTOLOGIES ET LA RECHERCHE D'INFORMATION 39
3.1. Introduction 39
3.2. Les ontologies 39
3.2.1. Définition 40
3.2.2. Les composants de base d'une ontologie 40
3.2.2.1. Les concepts 40
3.2.2.2. Les propriétés 41
3.2.2.3. Les relations 41
3.2.2.4. Les instances 41
3.2.2.5. Les axiomes 41
3.2.3. Classification des ontologies 41
3.2.3.1. Classification de Van Heijs 41
3.2.3.2. Classification de Guarino 42

3.2.3.3. Classification de Lassila et McGuinness 43
3.2.4. Méthodologie de modélisation ontologique 44
3.2.4.1. La méthodologie de Uschold et King 44
3.2.4.2. La méthode METHONTOLOGY 45
3.2.4.3. La méthode ON-TO-KNOWLEDGE 46
3.2.4.4. La méthode ARCHONTE 47
3.2.5. Les formalismes de représentation des ontologies 48
3.2.5.1. Les graphes conceptuels 48
3.2.5.2. Les logiques de description 49
3.2.6. Les langages de représentation des ontologies 49
3.2.7. Les ressources sémantiques dans le domaine médical 51
3.2.7.1. Les dictionnaires 51
3.2.7.2. Les taxonomies 51
3.2.7.3. Les thésaurus 52
3.2.7.4. Les ontologies 53
3.3. La recherche d'information 54
3.3.1. Notions de base de la RI 54
3.3.1.1. Document 54
3.3.1.2. Requête 54
3.3.1.3. Modèle de représentation 55
3.3.1.4. Le processus de recherche 55
3.3.1.5. Fonction de pondération 55
3.3.2. Système de recherche d'information (SRI) 56
3.3.2.1. L'indexation 56
3.3.2.2. L'appariement requête-document 57
3.3.3. Les modèles de recherche d'information 57
3.3.3.1. Le modèle booléen 58
3.3.3.2. Le modèle vectoriel 59
3.3.3.3. Le modèle probabiliste 59
3.3.4. Les problèmes de la recherche d'information classique 60
3.3.5. Expansion de la requête 60
3.3.5.1. Utilisation d'un corpus de documents 61
3.3.5.2. Utilisation de Ressources sémantiques 62
3.3.5.3. Enrichissement basée sur les logs 62
3.3.5.4. Enrichissement basée sur des données du web 63
3.3.6. Évaluation des systèmes de recherche d'information 63
3.3.7. Recherche d'information médicale à base d'ontologie 66
3.4. Conclusion 69
DEUXIEME PARTIE : APPROCHES PROPOSEES 70
4. RECONNAISSANCE DES EM ET RM A PARTIR DES RC ECRITS EN FRANÇAIS 71
4.1. Introduction 71
4.2. Les avantages du système proposé 71
4.3. Notre objectif 73
4.4. Principe de la solution proposée 73
4.4.1. Les grammaires locales 74
4.4.2. Reconnaissance des entités médicales 75
4.4.2.1. Processus de reconnaissance des entités médicales 76
4.4.2.2. Étude expérimentale et discussion des résultats 79
4.4.3. Extraction des relations médicales 82
4.4.3.1. Étude expérimentale et discussion des résultats 85
4.4.4. Bilan 87
4.4.5. Construction de l'ontologie orthopédique 88
4.4.5.1. Les composants de l'ontologie 88
4.4.5.2. Construction de l'ontologie orthopédique 89
4.5. Conclusion 93
5. RECHERCHE D'INFORMATION MEDICALE 94
5.1. Introduction 94
5.2. Les objectifs de la solution proposée 95
5.3. Architecture adoptée 95
5.3.1.1. La phase d'indexation 96
5.3.1.2. La phase d'analyse de requête 96
5.3.1.3. La phase d'expansion de requête 96

5.3.1.4. La phase de recherche 98
5.4. Les méthodes d'expansion de requête proposées *98*
5.4.1. Expansion des entités médicales 98
5.4.2. Expansion par extraction de relations sémantiques dans le contexte de la requête 100
5.4.3. Expansion par reformulation booléenne de la requête 101
5.4.4. Étude expérimentale et résultats obtenus 101
5.4.5. Comparaison avec la méthode classique 103
5.4.6. Discussion des résultats 103
5.5. Architecture distribuée à large échelle proposée *105*
5.5.1. Exemples de systèmes de recherche d'information à large échelle 107
5.5.2. Proposition d'un système de recherche sémantique à large échelle 109
5.5.2.1. Vue d'ensemble 109
5.5.2.2. Les outils utilisés dans ce système 110
5.5.2.3. Description du système de recherche à large échelle proposé 112
5.5.2.4. Processus d'indexation et d'extraction 113
5.5.2.5. Processus de recherche 114
5.6. Conclusion *115*
CONCLUSION ET PERSPECTIVES 116
BIBLIOGRAPHIE 120

Liste Des Tableaux

TABLEAU 2.1 : TRAVAUX REPRESENTATIFS DE LA TACHE DE RE MEDICALES A BASE DE REGLES 30
TABLEAU 2.2 : TRAVAUX REPRESENTATIFS DE LA TACHE DE RE MEDICALES PAR APPRENTISSAGE AUTOMATIQUE 31
TABLEAU 2.3 : TRAVAUX REPRESENTATIFS DE LA TACHE D'ERS PAR UTILISATION D'UNE METHODE A BASE PATRONS 35
TABLEAU 2.4 : TRAVAUX REPRESENTATIFS DE LA TACHE D'ER PAR APPRENTISSAGE AUTOMATIQUE 36
TABLEAU 3.1 CLASSIFICATION DES ONTOLOGIES SELON VAN HEIJS, GUARINO, LASSILA ET MCGUINNESS 44
TABLEAU 3.2 RECHERCHE D'INFORMATION MEDICALE A BASE D'ONTOLOGIE 67
TABLEAU 4.1 : EXEMPLE D'ENTITES MEDICALES ET LEURS TYPES 75
TABLEAU 4.2 : EXEMPLE DE PATRONS POUR CHAQUE CATEGORIE 78
TABLEAU 4.3 : TABLE DE CONFUSION POUR CHAQUE CATEGORIE D'ENTITE MEDICALE 80
TABLEAU 4.4 : ÉVALUATION DU SYSTÈME DE RECONNAISSANCE DES EM 81
TABLEAU 4.5 : RESULTATS MACRO-MOYENNE DE PRECISION, RAPPEL ET F-MESURE DU SYSTEME DE RECONNAISSANCES DES EM 81
TABLEAU 4.6 : EXEMPLES DE PATRONS LINGUISTIQUES 84
TABLEAU 4.7 : ÉVALUATION DU SYSTEME D'EXTRACTION DES RELATIONS SEMANTIQUES POUR CHAQUE TYPE DE RELATION 86
TABLEAU 4.8 : RESULTAT MACRO-MOYENNE DE PRECISION, DE RAPPEL ET F-MESURE DU SYSTEME D'EXTRACTION DES RM 86
TABLEAU 5.1 : CONTENU DE L'ONTOLOGIE "ONTO_ORTHOPÉDIQUE" EN NOMBRE 97
TABLEAU 5.2 : LA MOYENNE DU RAPPEL, PRÉCISION, F-MESURE, MAP ET R-PRÉCISION OBTENUE EN UTILISANT LES DIFFÉRENTES MÉTHODES D'EXPANSION 102
TABLEAU 5.3 : LE TAUX D'AMÉLIORATION PAR RAPPORT À LA MÉTHODE CLASSIQUE 103
TABLEAU 5.4 : L'INTERPOLATION PRÉCISION/RAPPEL MOYENNE OBTENUE POUR DIFFÉRENTES MÉTHODES 104
TABLEAU 5.5 : EXEMPLES DE SYSTÈMES DE RECHERCHE D'INFORMATION À LARGE ÉCHELLE 107
TABLEAU 5.6 : COMPARAISON ENTRE LES SYSTÈMES DE RI OPEN SOURCE POUR UNE COLLECTION DE 2.7GB 111

Liste Des Figures

FIGURE 1.1 : LES CONTRIBUTIONS DE LA THESE 16
FIGURE 3.1 : TRIANGLE D'ARISTOTE 40
FIGURE 3.2 : LES PRINCIPALES ETAPES DE LA METHODOLOGIE DE USHOLD ET KING 44
FIGURE 3.3 : PROCESSUS DE DEVELOPPEMENT ET CYCLE DE VIE DE METHONTOLOGY (GOMEZ-PEREZ ET AL., 2003) 45
FIGURE 3.4 : LES PRINCIPALES ETAPES DE LA METHODOLOGIE OTK (SURE & STUDER, 2003) 46
FIGURE 3.5 : CONSTRUCTION D'ONTOLOGIE SELON LA METHODE ARCHONTE. 48
FIGURE 3.6 : LES LANGAGES DE REPRESENTATION DES ONTOLOGIES 49
FIGURE 3.7 : LES RESSOURCES SEMANTIQUES 51
FIGURE 3.8 : PROCESSUS EN U DE RECHERCHE D'INFORMATIONS 55
FIGURE 3.9 : CLASSIFICATION DES MODELES DE RECHERCHE D'INFORMATION 58
FIGURE 3.10 : EXEMPLE REPRESENTANT LE MODELE VECTORIEL (BOUGHANEM, 2006) 59
FIGURE 3.11 : JEU DE RESULTATS DES DOCUMENTS 64
FIGURE 4.1 : LES AVANTAGES DU SYSTEME D'INFORMATION PROPOSE 72
FIGURE 4.2 : SCHEMA GENERAL DU SYSTEME D'EXTRACTION D'INFORMATION MEDICALE. 74
FIGURE 4.3 : EXEMPLE D'UNE GRAMMAIRE LOCALE SOUS FORME DE GRAPHE 74
FIGURE 4.4 : ARCHITECTURE DU SYSTEME DE RECONNAISSANCE DES EM 76
FIGURE 4.5 : EXEMPLE DES ENTRÉES DANS UN DICTIONNAIRE DES NOMS DE MALADIE DANS UNITEX. 77
FIGURE 4.6 : EXEMPLE D'UNE REGLE D'IDENTIFICATION D'UNE ENTITE MEDICALE ET SA GRAMMAIRE LOCALE DANS UNITEX 77
FIGURE 4.7 : PERFORMANCE DU SYSTEME DE RECONNAISSANCE DES ENTITES MEDICALES 81
FIGURE 4.8 : SCHEMA GENERAL DU SYSTEME D'EXTRACTION DE RELATIONS 82
FIGURE 4.9 : LES QUATRE RELATIONS CIBLES EXPRIMEES AVEC LE DIAGRAMME DE CLASSE UML 83
FIGURE 4.10 : EXEMPLE D'UNE GRAMMAIRE LOCALE DE LA RELATION 'TRAITE' 84
FIGURE 4.11 : EXEMPLE D'UNE GRAMMAIRE LOCALE DE LA RELATION 'DETECTE' 85
FIGURE 4.12: CONSTRUCTION DE L'ONTOLOGIE "ONTO_ORTHOPEDIQUE". 90
FIGURE 4.13 : REPRESENTATION DE L'ONTOLOGIE SOUS PROTEGE 4.3 92
FIGURE 5.1 : ARCHITECTURE PROPOSÉ DU SYSTÈME DE RECHERCHE SÉMANTIQUE 96
FIGURE 5.2 : ALGORITHME D'EXPANSION DE REQUÊTE 97
FIGURE 5.3 : EXPANSION DES ENTITÉS MÉDICALES PAR SYNONYMES ET HYPONYMES 99
FIGURE 5.4 : EXPANSION PAR EXTRACTION DE RELATIONS SÉMANTIQUES DANS LE CONTEXTE DE LA REQUÊTE 100
FIGURE 5.5 : PERFORMANCE OBTENUE POUR CHAQUE MÉTHODE EN UTILISANT LA MOYENNE DE RAPPEL, DE PRÉCISION, DE FM, DE MAP, ET R- PRÉCISION 103
FIGURE 5.6 : COURBE DE RAPPEL / PRÉCISION 105
FIGURE 5.7 : ARCHITECTURE DU SYSTÈME (LIN ET AL., 2015) 107
FIGURE 5.8 : ARCHITECTURE DU SYSTÈME SOBHY ET AL., (2012) 108
FIGURE 5.9 : VUE D'ENSEMBLE DU SYSTÈME DE RECHERCHE À LARGE ÉCHELLE 109
FIGURE 5.10 : ARCHITECTURE GLOBALE DE NOTRE SYSTÈME DE RECHERCHE D'INFORMATION À LARGE ÉCHELLE 112
FIGURE 5.11 : PROCESSUS D'INDEXATION ET D'EXTRACTION 113
FIGURE 5.12 : PROCESSUS DE RECHERCHE DISTRIBUÉ 114

Introduction et Problématique

1. Introduction

1.1. Contexte de la thèse

Aujourd'hui, le domaine médical dispose d'une grande quantité d'information notamment celle qui existe dans les rapports cliniques; ce sont des documents médicaux rédigés par les médecins dans les hôpitaux, dans les soins ambulatoires et dans les services médicaux publics. Ces rapports médicaux exprimés en format textuel comportent des informations telles que les pathologies, les antécédents médicaux et les diagnostics. Notons que le médecin a besoin de consulter ces rapports porteurs d'informations pour qu'il puisse prendre une décision dans les brefs délais pour la prise en charge des patients.

Les moteurs de recherche médicaux comme CISMeF[1], PubMed[2], et BioPortal[3] sont basés sur des articles scientifiques et mêmes sur des livres médicaux; mais les termes médicaux qu'ils contiennent sont beaucoup plus généraux. En effet, dans ce cas deux points sont intéressants 1)-un système de recherche spécialisé pour les médecins est nécessaire, 2)-ce système est basé sur les anciens rapports médicaux; un tel système permet aux médecins de rédiger une requête exprimé en langage naturel et de retourner un ensemble de documents jugés pertinents. Dans cette thèse, ce système de recherche est basé sur l'extraction des informations contenues dans les rapports cliniques, nous nous intéressons donc par : i)-la recherche d'information (RI), et ii)-l'extraction d'information (EI).

Recherche d'information :

La recherche d'information a pour objectif l'étude des modèles et systèmes d'interaction entre des utilisateurs humains et des corpus de documents, en vue de la satisfaction de leurs besoins d'information.

Extraction d'information :

L'extraction d'information consiste à analyser des textes pour en obtenir des informations en vue d'une application précise. D'une autre façon, l'extraction d'information est le processus qui permet d'obtenir automatiquement des informations structurés à partir des documents sous format libre, ces informations sont destinées à créer ou alimenter un entrepôt de données.

[1] http://www.chu-rouen.fr/cismef/
[2] https://www.ncbi.nlm.nih.gov/pubmed/
[3] https://bioportal.bioontology.org/

1.2. Problématique

La recherche d'information classique est affectée par plusieurs problèmes dans différent domaine tel que le domaine médical où les termes importants (comme les entités médicales) sont rencontrés plusieurs fois. Toutefois, un terme peut apparaître plusieurs fois dans une collection de documents, mais à chaque fois avec une signification différente. Les systèmes de recherche d'information classique traitent les variations d'un terme comme étant des termes différents. Par exemple c'est le cas des synonymes et des abréviations. Cela affecte la recherche et nécessite soit l'intégration des thésaurus ou soit des médecins spécialistes pour spécifier toutes les variations possibles dans leur requête s'ils souhaitent récupérer tous les documents pertinents.

Dans les rapports cliniques, les médecins utilisent des entités médicales ou une propre terminologie pour décrire l'état d'un malade, (exemple : abréviations, termes traduits, termes largement utilisées). De plus, les requêtes des médecins contiennent aussi des entités médicales ce qui donne l'opportunité aux moteurs de recherche d'améliorer leur compréhension. Ces concepts ou entités médicales peuvent avoir plusieurs variations et c'est le cas par exemple des synonymes (exemple: "gonarthrose" et "arthrose du genou"), ou d'exprimer une même relation entre deux entités médicales ayant une même catégorie sémantique (exemple: "orthèse genou" et "gonarthrose", "prothèse genou" et "gonarthrose"). Pour faire face à ce problème, ceci nécessite l'intégration des thésaurus ou des ontologies médicales pour spécifier toutes les variations des entités médicales contenues dans la requête pour pouvoir récupérer tous les documents pertinents. Bien qu'il existe des ressources sémantiques dans le domaine médicale (exemple: MeSH[4], SNOMED int[5], NCIt[6]) la plupart des relations qui existent dans ces ressources sont des relations hiérarchiques (hyponymie, hyperonymie et de synonymie), ces ressources manquent de relations syntagmatique (Embarek, 2008). Pour remédier à ce problème et pour assurer une recherche de qualité, nous avons utilisé les techniques d'extraction d'information pour pouvoir alimenter une ontologie médicale contenant les entités médicales et les relations syntagmatiques qui les relient. Cette ontologie est intégrée dans un système de recherche d'information médicale pour étendre la requête de l'utilisateur.

[4] Medical Subject Heading
[5] Systematic Nomenclature Of MEDecine
[6] National Cancer Institute Thesaurus

Cinq tâches distinctes caractérisent les principales capacités fonctionnelles des systèmes EI actuels: la reconnaissance des entités nommées, la résolution de coréférence, le remplissage de patrons d'entités, l'extraction de relations, et la description d'évènement. Parmi ces tâches, nous nous intéressons à la reconnaissance des entités nommées et l'extraction de relations. Dans ce travail de thèse, nous étudions l'impact de l'extraction d'information dans un système de recherche d'information. Ainsi, nous avons divisé cette étude en deux grande parties; dans la première partie, nous avons développé une ontologie médicale à partir des informations extraites, dans ce cas nous avons entamé deux tâches ; (i) la reconnaissance des entités médicales et, (ii) l'extraction de relations sémantiques dans les rapports médicaux. Dans la deuxième partie de la thèse, nous avons proposé une approche sémantique de recherche d'information; cette approche est basée sur l'ontologie construite. Cette dernière est utilisée pour étendre la requête de l'utilisateur dans un système de recherche d'information.

De nos jours, le médecin a besoin d'un système de recherche pour lui faciliter l'accès à ces rapports médicaux dans les brefs délais. Les systèmes de recherche centralisés deviennent insuffisants pour manipuler des informations à large échelle, ils sont inadéquat pour traiter un nombre important de requêtes sur l'index. Pour assurer une recherche rapide et scalable nous proposons une architecture à large échelle basée sur le Cloud Computing pour la représentation de l'index et l'ontologie dans un environnement distribué.

1.3. Contributions

Nos contributions se tournent autour de trois principaux axes comme illustré dans la Figure 1.1 : a) l'extraction d'information à partir des rapports médicaux, b) la recherche d'information basée sur les informations extraites et c) la proposition d'une architecture du système de recherche d'information médicale à large échelle.

Dans le premier axe; après une étude sur les systèmes d'extraction d'information dans le domaine médical (Ghoulam et al., 2015a), nous avons conclu qu'il existe deux grandes méthodes d'extraction d'information, une méthode dite à base de règles (ou à base de patrons), et une méthode dite à base d'apprentissage automatique.

Nous avons noté que le nombre de travaux réalisés et utilisant la méthode à base de règles est plus élevé par rapport à ceux qui utilisent l'apprentissage automatique, en particulier dans la communauté française qui manquent de corpus annotés. Ceci nous a conduits à utiliser dans ce travail de thèse une méthode à base de règles. Nous proposons d'exploiter les grammaires locales; dans un premier temps pour la reconnaissance des entités médicales (Ghoulam et al.,

2015b). Et dans un deuxième temps pour l'extraction de relations médicales. Les informations extraites alimentent une ontologie médicale orthopédique.

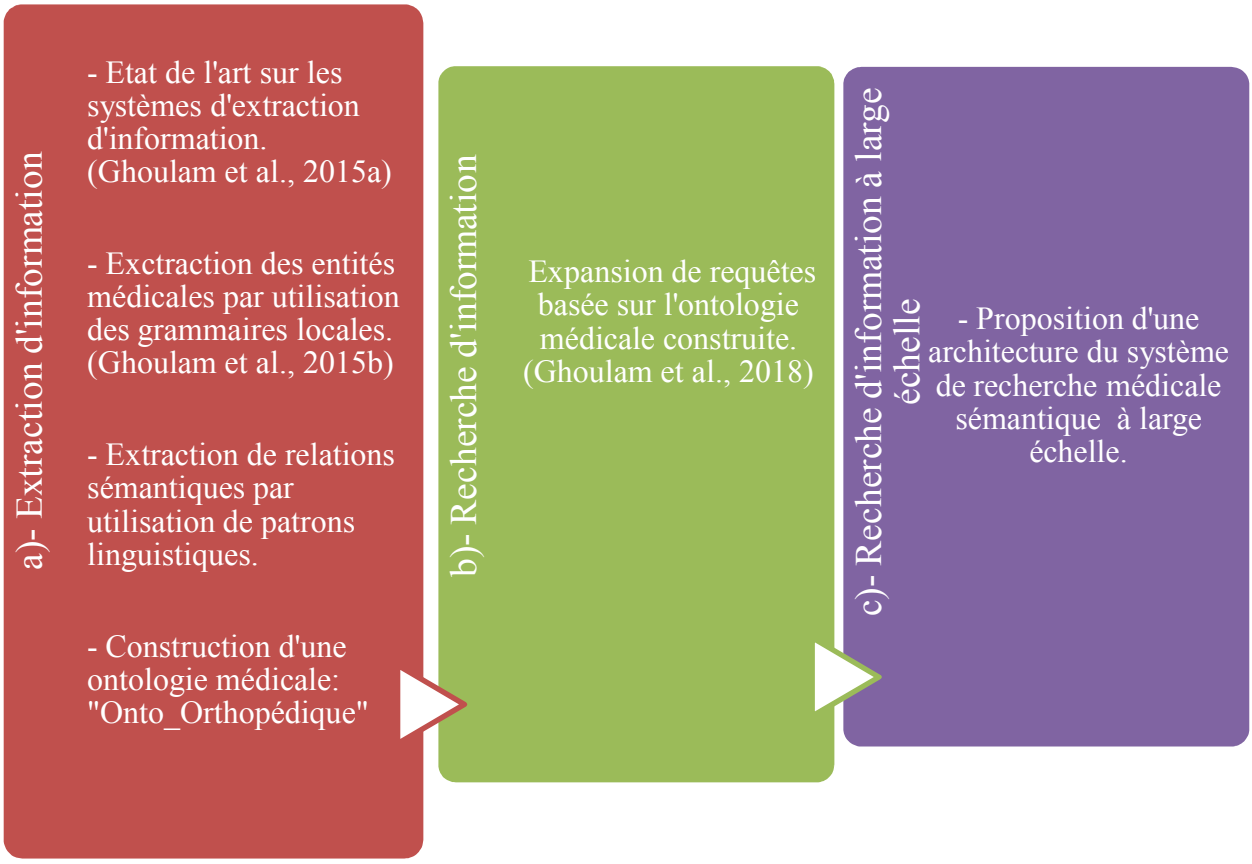

FIGURE 1.1 : LES CONTRIBUTIONS DE LA THESE

Dans le deuxième axe, nous nous intéressons à l'intégration de l'ontologie orthopédique dans un système de recherche d'information. Pour améliorer la qualité de recherche, nous faisons appel à cette ontologie pendant la recherche d'information pour étendre la requête de l'utilisateur.

Dans le troisième axe, une architecture du système de recherche sémantique à large échelle est proposée. Nous proposons d'utiliser le Cloud Computing pour la représentation de l'index et l'ontologie dans un environnement distribué.

1.4. Liste De Publications

- Ghoulam, A., Barigou, F., & Belalem, G. (2015a). **Information Extraction in the Medical Domain**. Journal of Information Technology Research, 8(2): 1-15.
- Ghoulam, A., Barigou, F., Belalem, G., & Meziane, F. (2015b). **Using Local Grammar for Entity Extraction from Clinical Reports**. International Journal

of Artificial Intelligence and Interactive Multimedia, 3(3): 16-24. DOI: 10.9781/ijimai.2015.332.

- Ghoulam, A., Barigou, F., Belalem, G., & Meziane, F. (2018). **Query expansion using external resource in medical domain**. International-journal-intelligent-information-technologies, 14(3): In press.

1.5. Organisation de la thèse

Cette thèse est composée d'une introduction générale et de deux grandes parties; la première partie est un état de l'art et la deuxième partie représente nos contributions.

La première partie contient deux chapitres d'état de l'art;

Le chapitre 2 représente l'extraction d'information médicale, y compris l'extraction des entités médicales ainsi que l'extraction des relations dans le domaine médical.

Le chapitre 3 est consacré aux ontologies et à la recherche d'information, nous commençons ce chapitre par un état de l'art sur les ontologies, nous entamons la classification et la construction des ontologies. Nous poursuivons par la définition des notions de base de la recherche d'information, et les problèmes liés à la recherche classique. Nous terminons ce chapitre par des travaux sur la construction d'ontologie dans le domaine médical et sur l'utilisation d'une telle ressource sémantique dans la recherche médicale.

La deuxième partie est consacrée à représenter nos contributions, elle contient deux chapitres;

Le chapitre 4 présente de façon détaillée notre approche d'extraction des entités et des relations médicales à partir des rapports cliniques pour alimenter une ontologie médicale orthopédique.

Le chapitre 5 est consacré à représenter nos contribution dans la recherche d'information sémantique par l'intégration de l'ontologie orthopédique dans un système de recherche d'information médical dans le but d'étendre la requête de l'utilisateur. Par la suite, nous proposons une architecture distribuée à large échelle de ce système de recherche sémantique.

Enfin, nous présentons les conclusions et les perspectives pour ce travail de recherche.

Première Partie : État de l’Art

2. Extraction d'Information Médicale

2.1. Introduction

La quantité d'information disponible sous forme électronique ne cesse de croître, en particulier dans le domaine médical où elle double tous les cinq ans, comme indiqué dans (Ben Abbacha, et al., 2011), ce qui rend le développement de systèmes intelligents pour traiter cette information un besoin urgent pour les praticiens; tels que les professionnels de la santé. Cette mine d'informations médicales se heurte à des difficultés d'accès, car celles-ci sont conservées dans des formats non structurés, et c'est le cas des Rapports Cliniques (notés plus loin par RCs). Ce sont des textes écrits par des médecins pour décrire des cas pathologiques et dont le contenu est particulièrement riche d'entités médicales et de relations entre ces entités. Les RCs sont ainsi considérés comme une importante source de connaissances qui peut servir dans de nombreuses applications médicales telles que la recherche d'information médicale, la prise de décisions médicales, les études épidémiologiques et l'exploration de données. Cependant, comme il s'agit souvent de textes non structurés, cette information est difficilement accessible. L'extraction d'informations médicales peut apporter une solution à ce type de problème.

L'Extraction d'Information (EI) est une tâche de Traitement Automatique du Langage Naturel (TALN) dont l'objectif est d'analyser des textes écrits en langage naturel pour extraire des informations structurées et utiles telles que les entités nommées et les relations sémantiques entre ces entités. L'EI est une tâche importante dans un ensemble diversifié d'applications telles que la bibliographie biomédicale, le service à la clientèle, les sites Web communautaires, la gestion de l'information personnelle (Berrazega, 2012).

Dans ce chapitre, nous présentons quelques notions de base relatives au domaine d'extraction d'information. Nous abordons deux tâches d'extraction : la reconnaissance des entités nommées et l'extraction de relations. Nous terminons ce chapitre par une étude comparative de quelques systèmes d'extraction d'information pour les deux tâches susmentionnées.

2.2. L'extraction d'information

2.2.1. Définition

L'extraction d'information est le processus automatique, qui permet d'extraire des informations pertinentes et précises à partir de documents non structurés en langage naturel et permet leur sauvegarde sous une forme structurée du type formulaire ou base de données (Sarawagi, 2007).

Le domaine de l'EI n'a acquis sa maturité qu'avec l'émergence des conférences d'évaluation MUC[7] (Message Understanding Conferences). Sept conférences ont eu lieu entre 1987 et 1998 afin d'évaluer les systèmes d'extraction d'information (Even, 2005). Ces conférences traduisent la volonté d'un ensemble de groupes de chercheurs en extraction d'information travaillant dans le cadre d'un projet de l'US Navy sur l'étude des messages de la marine, pour définir des méthodes communes et des corpus de référence afin de pouvoir comparer leurs systèmes.

2.2.2. Les tâches d'extraction d'information

Cinq tâches ont été spécifiées par l'évaluation de la MUC-7 (Piskorski & Yangarber, 2013). Ces tâches étaient axées sur l'extraction d'information dans des enregistrements. Les tâches données ci-dessous, sont adaptées à partir des définitions de MUC-7 (Chinchor & Marsh, 1998) :

- *Reconnaissance des entités nommées (REN) :* elle implique la reconnaissance des entités telles que les noms des organisations, les noms des personnes, les noms de lieux, les dates et les montants monétaires. La tâche a été considérablement élargie pour couvrir à la fois des choses concrètes et abstraites dans le texte. Dans le domaine médical, par exemple, cela peut inclure les noms de maladies et les noms de médicaments.
- *La tâche d'extraction de relations :* c'est la tâche qui consiste à détecter et à caractériser les relations sémantiques entre les entités dans le texte. Dans le domaine médical, il peut s'agir, par exemple, d'une relation entre une maladie et un médicament pour exprimer un traitement.

[7] https://en.wikipedia.org/wiki/Message_Understanding_Conference

- *La résolution de coréférence :* c'est une tâche qui détermine les expressions linguistiques qui se réfèrent à une même entité du monde réel. Elle n'a pas été largement appliquée aux documents médicaux (Ware et al, 2012). Formellement, la coréférence se compose de deux expressions linguistiques; antécédent et anaphore. L'anaphore est l'expression dont l'interprétation dépend de celle de l'autre expression (c'est-à-dire l'associant à une entité réelle ou abstraite du monde réel). L'antécédent est l'expression linguistique sur laquelle une anaphore dépend. Dans l'exemple : « *Laila se tenait à la fenêtre, elle regardait les passants..* », « *Laila* » et le pronom « *elle* » se réfèrent à la même entité. Donc, « *Laila* » est un antécédent, et « *elle* » est l'anaphore.

- *Remplissage de patrons :* les informations à extraire comme des entités, des relations et des événements sont pré-spécifiées dans des structures définies par l'utilisateur appelées modèles (ou patrons), chacun consiste en un certain nombre de slots (ou attributs), qui doivent être instanciés par un système d'extraction d'information pendant le traitement du texte.

- *La description d'évènements :* les différents résultats du remplissage de patrons sont reliés afin de produire les relations possibles entre ces patrons, mais également les évènements concernant certains de ces patrons et induits par des informations contenues dans les textes. On obtient ainsi une description des évènements contenus dans les documents analysés. Les patrons de scénario définissent les types de relations et d'entités formant les évènements à identifier (Even, 2005). Dans (Sun et al., 2013), un événement médical a été défini comme tout ce qui est cliniquement important et qui peut également être mis en correspondance avec le temps. Les auteurs ont créé dans le défi i2b2 2012; un corpus de relations temporelles cliniques qui comprend des événements cliniques, des expressions temporelles et des relations temporelles.

2.3. La Reconnaissance des Entités Nommées

L'extraction ou la reconnaissance des entités nommées est une étape importante dans le processus d'extraction d'informations à partir de textes. Elle a été inventée pour la première fois lors de la sixième conférence d'information MUC-6 (1995) (Nadeau & sekine, 2007).

Dans le domaine médical, une entité nommée est définie comme un mot unique ou une expression composée qui désigne un objet médical, par exemple une maladie, un symptôme

ou un médicament. Ces entités nommées spécifiques au domaine médical sont appelées entités médicales. Nous pouvons citer comme exemples :

- Les Maladies ou problèmes comme le Cancer, l'Alzheimer;
- Les Traitements comme la Radiothérapie;
- Les Tests ou examens médicaux comme les tests sanguins;
- Les Symptômes comme la Fièvre et les Vomissements;
- Les médicaments comme Panadol et Humex.

2.3.1. Les classes des entités nommées

Les sous tâches de reconnaissance des entités nommées proposent de distinguer trois (3) grandes classes d'entités (Jayan, 2013) :

- *ENAMEX « Named Entities » :* les expressions de noms propres incluant les noms de personnes, les noms de lieux et les noms des organisations, sous-catégorisées en «Organisation», «Personnes», et « Location ».
- *TIMEX « Temporal Expression » :* les expressions temporelles comme les dates et les heures, sous-catégorisées en «Date» et «Time».
- *NUMEX « Number Expressions » :* les expressions numériques telles que les expressions monétaires et les pourcentages, sous catégorisées en «Money» et «Pourcentage».

2.3.2. Les difficultés de la catégorisation des entités nommées

Pour réaliser un système de reconnaissance des entités nommées, il est intéressant de déterminer un ensemble de catégories. Plusieurs éléments peuvent être pris en compte ce qui provoque un ensemble de problèmes et de difficultés liés à la catégorisation automatique (Poibeau, 2005). Voici quelques difficultés :

- *La polysémie des entités nommées :* c'est-à-dire une entité qui a plusieurs sens ou significations différentes. On distingue :
  - *Homonymie :* Orange désigne une ville, ou une société.
  - *Métonymie :* Leclerc peut être un homme d'affaire, un supermarché, ou un groupe financier.

- *Facettes :* Jacques Chirac le président de la république ou le maire de Paris.

- *Référentialité des entités nommées :* dans l'exemple de (Poibeau, 2005) issu d'un corpus de biologie « Abd-a interagit avec trx », a-t-on affaire à des entités nommées ou à des termes? « Abd-a » constitue-t-il un nom de gènes ou une classe? Les réponses ne sont pas toujours claires.

- *Catégorisation à partir d'une ontologie (plusieurs variantes) :*
  - *Exemple :* « Jacque Chirac », « J. Chirac », « le président Jacque Chirac », « l'ancien président français ».

- *La coordination :* Il s'agit d'une ou plusieurs entités.
  - *Exemple :* "Bill and Hillary Clinton flew to Chicago last month". Il existe trois possibilités (Tannier, 2005) de catégoriser les entités nommées :
    - <PERS>Bill</PERS> and <PERS>Hillary Clinton</PERS> flew to Chicago last month
    - <PERS>Bill and Hillary Clinton</PERS> flew to Chicago last month
    - <PERS> Bill Clinton</PERS> and <PERS> Hillary Clinton </PERS> flew to Chicago last month

- *L'imbrication :* Il s'agit d'une entité incluant d'autres entités (Tannier, 2005).
  - *Exemple :*
    - <ORG>L'Université de Corte</ORG> ou
    - <ORG>L'Université de <LOC>Corte</LOC><ORG>

- *Les frontières des entités nommées* (application des règles).

2.4. Extraction des Relations Sémantiques

De nombreuses applications dans l'extraction d'information, la compréhension du langage naturel et la recherche d'information nécessitent une extraction de relations sémantiques entre les entités. L'extraction de relations est généralement précédée par la tâche de reconnaissance des entités. Les relations sémantiques sont regroupées en deux familles principales (Nirenburg & Raskin, 2004). Il s'agit des relations paradigmatiques et des relations syntagmatiques.

2.4.1. Relations paradigmatiques

Les relations paradigmatiques sont des relations fonctionnant principalement sur des concepts de la même classe. Habituellement, ce type de relations représente des relations hiérarchiques nommées liens verticaux. Elles sont utilisées pour organiser les concepts sous forme d'un arbre, comme dans le thésaurus Medical Subject Headings (MeSH[8]) ou dans les Meta-thesaurus comme l'Unified Medical Language System (UMLS[9]). Parmi ce type de relations, on peut citer les relations de l'antonymie, de la synonymie et de l'hyperonymie.

2.4.2. Relations syntagmatiques

La tâche d'extraction de relations médicales vise à extraire les relations entre deux ou plusieurs entités médicales, connu sous le nom de relations syntagmatiques.

Les relations syntagmatiques sont des liens sémantiques qui se produisent entre deux (ou plusieurs) unités linguistiques présentes dans une expression. Ils sont identifiés par l'étude des formes syntaxiques dans les textes, et par la présence d'un prédicat. Par exemple, on peut citer des relations spécifiques dans le domaine médical comme : « X doit être traité par Y » ou « Y pour le traitement de X ». Les deux exemples précédents montrent qu'il existe une relation de traitement entre la maladie X et le traitement Y. Il existe de nombreux autres exemples d'expressions telles que développées dans (Sun et al., 2013) pour extraire des relations temporelles entre les événements cliniques et les expressions temporelles.

2.5. Les Méthodes d'extraction d'information

Les systèmes d'extraction d'information reposent généralement sur deux (2) approches principales : l'approche dite à base de règles et l'approche à base d'apprentissage automatique (Nadeau & Sekine, 2007). Il existe des systèmes hybrides qui combinent ces deux approches. Il y'a aussi des méthodes à base d'ontologie (Serrano et al, 2011).

2.5.1. Les méthodes à base de règles

Les méthodes à base de règles s'appuient sur l'utilisation des règles qui sont construites manuellement à partir des corpus d'un domaine donné pour identifier les entités.

Les premiers systèmes d'extraction d'information dans les MUCs étaient des systèmes à base de règles; ils utilisaient des règles écrites manuellement et développées par des concepteurs (experts) qui devaient connaître le formalisme d'écriture de ces règles pour un

[8] https://www.ncbi.nlm.nih.gov/mesh
[9] https://www.nlm.nih.gov/research/umls/

système particulier. Les exemples les plus représentatifs de ce type de systèmes sont FASTUS (Hobbs et al., 1992), GENLTOOLSET (Krupka et al, 1992), PLUM (Ayuso et al, 1992) et PROTEUS (Yangarber & Grishman, 1998). Ces systèmes sont bien détaillés dans (Kaiser & Miksh, 2005).

Les méthodes à base de règles, également appelées des méthodes d'ingénierie de connaissances dans certaines sources (Ghoulam et al., 2015a) donnent de bons résultats. Elles nécessitent cependant un grand effort et un temps considérable pour l'analyse des données et l'écriture des règles.

Les méthodes de reconnaissance des entités nommées à base de règles fonctionnent généralement comme suit : un ensemble de règles est défini manuellement ou acquis automatiquement. Une règle est constituée d'un patron et d'une action. Un patron est généralement une expression régulière définie à partir des caractéristiques des mots. Par exemple, pour étiqueter toute séquence de mots de la forme "M. X" où X est un mot en majuscule qui se réfère à une entité de type personne, la règle peut être définie ainsi (Jiang, 2012) :

(Token= 'M.' type d'orthographe = premier en majuscule) → *nom d'une personne.*

En outre, les méthodes d'EI basées sur des règles pour l'extraction de relations fonctionnent généralement de manière similaire. Par exemple dans le patron "X est traité par Y" où X et Y sont des entités nommées, on peut extraire la relation suivante : "est traitée par (X, Y)".

Les systèmes basés sur la construction manuelle des règles sont plus intéressants dans des domaines fermés où l'intervention de l'être humain est à la fois essentielle et disponible. Dans des domaines ouverts comme l'extraction d'opinions à partir de blogs, la souplesse des méthodes statistiques est plus appropriée (Ghoulam et al., 2015a).

2.5.2. Les méthodes d'apprentissage automatique

Les méthodes d'apprentissage automatique ou encore les méthodes d'apprentissage statistique, sont des techniques d'entraînement capables d'extraire automatiquement, par exemple, des entités étiquetés dans un jeu de données. Ces méthodes nécessitent un large corpus de textes étiquetés pour apprendre à identifier des entités.

Il existe différents types d'apprentissage; l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. La plupart des études récentes utilisent

l'apprentissage supervisé à partir d'une collection d'exemples étiquetés (Nadeau & Sekine, 2007). L'idée de l'apprentissage supervisé consiste à étudier les caractéristiques des entités à extraire sur une vaste collection de documents annotés, où un expert doit être disponible pour étiqueter correctement ces exemples.

De nombreux modèles d'apprentissage ont été proposés et appliqués à la REN; les plus importants sont les Modèles de Markov cachés (HMM) (Vinciarelli & Favre, 2005), les modèles de Markov d'entropie maximum (MEMM) (Borthwick et al., 1998), les machines à vecteur de support (SVM) (Ben abacha & Zweigenbaum, 2011) et les champs aléatoires conditionnels (CRF) (He & Kayaalp, 2008). L'avantage important d'un système basé sur l'apprentissage automatique est qu'il peut être transféré facilement dans un autre domaine différent.

Les techniques d'apprentissage automatique ont démontré des résultats remarquables dans tous les domaines, en général, et dans le domaine médical en particulier. Cependant, Chapman et ses collègues (Chapman et al., 2011) ont montré que ce genre de techniques est coûteux et plus particulièrement dans le domaine médical. Il nécessite de grands volumes de textes, des exemples de haute qualité, manuellement annotés, qui sont à la fois coûteux et longs pour entraîner les modèles. Ainsi, la principale barrière des techniques d'apprentissage supervisé est l'exigence d'un grand corpus annoté. L'indisponibilité de ces ressources et le coût prohibitif de leur création; ont conduit les chercheurs à deux méthodes d'apprentissage alternatives; l'apprentissage semi-supervisé et les techniques d'apprentissage non supervisé comme celles développés dans (Zhang & Elhadad, 2013).

La première méthode implique un petit degré de supervision, tel qu'un ensemble d'exemples de départ (seeds en Anglais) pour lancer le processus d'apprentissage (Thenmalar et al., 2015).

La deuxième méthode est fondée sur le regroupement (appelée Clustering); aucun expert n'est fourni. Le Clustering est un algorithme d'apprentissage non supervisé. À partir d'échantillons d'apprentissage non étiquetés on cherche des régularités sous-jacentes, par exemple rassembler des entités nommées en un groupe de cluster basé sur la similarité du contexte et s'appuyer sur des ressources lexicales comme WordNet[10]. L'étude de ces techniques est bien présentée dans (Jiang, 2012).

[10] https://wordnet.princeton.edu/

2.5.3. Les méthodes hybrides

Les méthodes hybrides utilisent des règles écrites à la main mais construisent aussi une partie de leurs règles à l'aide d'informations syntaxiques et d'informations sur le discours tirées de données d'entraînement grâce à des algorithmes d'apprentissage automatique. Ces systèmes combinent les deux méthodes précédentes c'est-à-dire celles à base de règles et à base d'apprentissage automatique. La combinaison des deux méthodes peut se faire en commençant par exemple (i) par la méthode manuelle à base de règles comme utilisée dans (Ben abacha et Zweigenbaum, 2011) où l'hybridation est faite par transformation des résultats à base de règles de MetaMapPlus au format BIO puis considérer ces résultats comme des attributs pour le classifieur CRF. (ii) ou par la méthode d'apprentissage comme utilisé dans (Zribi et al., 2010) où l'hybridation est faite par utilisation de l'algorithme d'apprentissage des règles RIPPER pour la détection des entités nommées puis les résultats obtenus sont validés par un ensemble de règles construites manuellement en calculant un score de validation pour chaque règle.

2.5.4. Les méthodes à base d'ontologies

La tâche d'extraction d'information basée sur l'ontologie (OBIE; Ontology-based Information Extraction) a récemment émergé comme un sous-domaine de l'EI (Ritesh & Suresh, 2014).

Dans (Wimalasuriya & Dejing, 2010), un système OBIE est défini comme un système qui traite un texte en langage naturel non structuré ou semi-structuré à travers un mécanisme guidé par des ontologies pour extraire certains types d'informations. Il présente la sortie par utilisation d'un langage de définition d'ontologie tel que le langage OWL (Ontology Web Language). Ces auteurs décrivent une relation étroite entre OBIE et le Web sémantique.

La relation entre ontologie et EI est impliquée dans deux (2) tâches (Nedellec & Nazarenko, 2005): d'une part, l'ontologie est utilisée pour l'extraction d'information; l'EI a besoin d'ontologies dans le cadre du processus de compréhension pour extraire l'information pertinente (Gurulingappa et al., 2012); d'autre part, l'extraction d'information est utilisée pour peupler et améliorer des ontologies de domaine (Denis & Wasito, 2017).

Deux types différents de méthodes impliquant l'EI et les ontologies sont utilisés (Vicient, 2011) : (i) la méthode EI basée sur l'ontologie et (ii) la construction d'ontologie basée sur l'EI. La première méthode utilise une ontologie spécifique au domaine dans son processus d'extraction, elle essaie d'identifier des entités à partir d'un document particulier (ou d'un

ensemble de documents) et de les annoter en fonction d'une ontologie d'entrée. Dans le domaine médical, de multiples ontologies standardisées sont disponibles (par exemple UMLS, SNOMED CT[11], MeSH). Au contraire, la deuxième méthode; se base sur des techniques d'extraction afin de peupler ou d'enrichir une ontologie. Cette dernière peut être utilisée dans d'autres applications.

2.6. Mesures d'évaluation des Systèmes d'extraction d'information

Les métriques d'évaluation utilisées pour évaluer les systèmes d'extraction d'information ont été inspirées de celles utilisées en recherche d'information (Boufaden, 2005). Les mesures de Rappel et de Précision sont adaptées au domaine de l'extraction d'information pour pouvoir évaluer le silence et le bruit lors d'une tâche d'extraction.

Le *rappel* détermine le taux des informations correctes extraites par rapport au nombre total d'informations pertinentes disponibles dans le texte analysé (Even, 2005).

$$\mathbf{Rappel} = \frac{\mathbf{N_{correctes}}}{\mathbf{N_{total}}} \qquad \mathbf{2.1}$$

Avec $N_{correcte}$ le nombre d'informations correctes extraites par le système, et N_{total} le nombre total d'informations attendues (nombre d'information pertinentes dans le texte).

$$\mathbf{Silence} = \mathbf{1} - \mathbf{Rappel} \qquad \mathbf{2.2}$$

La *précision* calcule le rapport entre le nombre d'informations extraites qui sont correctes et le nombre total d'informations extraites (correctes et incorrectes).

$$\mathbf{Précision} = \frac{\mathbf{N_{correctes}}}{\mathbf{N_{correctes}} + \mathbf{N_{incorrectes}}} \qquad \mathbf{2.3}$$

Avec $N_{incorrectes}$ le nombre d'informations incorrectes extraites par le système.

$$\mathbf{Bruit} = \mathbf{1} - \mathbf{Précision} \qquad \mathbf{2.4}$$

F-mesure est une combinaison des mesures de Rappel et de Précision.

$$\mathbf{F-mesure} = \frac{(\mathbf{1}+\boldsymbol{\beta}\mathbf{2}) * \mathbf{Précision} * \mathbf{Rappel}}{\boldsymbol{\beta}\mathbf{2} * (\mathbf{Précision} + \mathbf{Rappel})} \qquad \mathbf{2.5}$$

Pour une valeur de $\beta = 1$, le Rappel et la Précision ont la même importance, et donc :

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

$$\mathbf{F-mesure} = \frac{\mathbf{2*Précision*Rappel}}{\mathbf{Précision+Rappel}} \quad \mathbf{2.6}$$

2.7. Travaux sur l'extraction d'information médicale

Plusieurs travaux ont été proposés pour extraire des informations dans le domaine médical (Ghoulam et al, 2015a), nous présentons dans cette section quelques uns sur la reconnaissance des entités médicales ainsi que l'extraction de relations reliant les entités.

2.7.1. Reconnaissance des entités médicales

Plusieurs travaux et systèmes de reconnaissance des entités médicales ont été mis en œuvre; ils permettent d'identifier et de catégoriser les entités médicales. Ces systèmes utilisent différentes approches (à base de règles, par apprentissage ou hybride) pour un même corpus ou différents corpus (des articles scientifiques, rapports cliniques…etc.) de différentes langues (Français, Anglais,…).

2.7.1.1. Systèmes à base de règles

Le Tableau 2.1 résume quelques travaux de reconnaissance d'entités médicales à base de règles. L'évaluation de chaque système est donnée dans le même tableau; elle est basée sur les métriques de précision (P), rappel (R) et F-mesure (F).

	Référence	Contribution : reconnaissance de	Corpus	Techniques	P (%)	R (%)	F (%)
Approches à base de règles	Harkema et al. (2005)	Signes de cancer des poumons et localisation du cancer	Rapport de radiologie (Anglais)	AMBIT	69.00	83.00	75.00
	Spasic et al. (2010)	Propriétés de médication : nom, mode, fréquence, durée, raison	I2b2 2009 (Anglais)	Patrons linguistiques et règles sémantiques	86.00	77.00	81.00
	Ben Abacha & Zweigenbaum (2011)	Problèmes, traitements et tests	I2b2 2010	Tree Tagger & Meta Map Plus	48.68	56.46	52.28
			Berkeley		23.43	42.47	30.20
	Barigou et al. (2011)	Noms des patients, maladies, signes, médication	Rapports cliniques (Français)	TreeTagger & dictionnaires & Règles syntaxico-sémantiques	98.90	68.80	81.15
	Barigou et al. (2012)	Noms des patients, maladies, signes, médication	Rapports cliniques (Français)	Cellular automaton Boolean inference	92.00	89.00	86.89

	Embarek & Ferret, (2012)	Maladie, signe, médication, examen et traitement	Articles scientifiques (Français)	Patrons morpho-syntaxiques	90.00	84.00	86.00
	Van et al. (2016)	Anatomie, Produits chimiques et médicaments, procédure, troubles, Régions géographiques, êtres vivants, objets, phénomènes, physiologie, Appareils.	MEDLINE (Français)	Utilisation de l'UMLS	50.10	37.60	42.90
			EMAEA (Français)		47.60	23.20	37.60

Tableau 2.1 : Travaux représentatifs de la tâche de RE médicales à base de règles

Harkema et al. (2005) introduisent le système AMBIT. Il s'agit d'un environnement pour l'acquisition d'informations médicales et biologiques à partir de textes. Le processus d'extraction d'informations comprend trois (3) étapes majeures : (i) étape lexicale et terminologique, (ii) étape syntaxique et sémantique, (iii) étape de discours. Ce système est utilisé pour extraire différentes entités telles que les signes du cancer et les localisations du cancer du poumon. Ils ont obtenu un score de F-mesure égale à 75%.

Le système donné par (Spasic et al., 2010) a obtenu 81% de F-mesure ; trois phases ont été planifiées dans ce système: la phase de prétraitement linguistique, la phase d'extraction des patrons, et la phase de remplissage des templates. Cette approche est principalement basée sur un dictionnaire de noms de médicaments pour extraire les propriétés des médicaments comme le nom, le mode d'administration…etc. Les auteurs ont utilisé aussi le méta-thésaurus UMLS.

Ben Abacha & zweigenbaum (2011) extraient à partir de deux corpus I2b2 2010 et Berkeley les entités médicales de types problème, traitement et test. Les résultats obtenus pour les deux corpus n'ont pas les mêmes performances, pour le corpus I2b2 2010 la F-mesure a été 58.28% et de 30.20% pour le corpus Berkeley. Ceci est dû aux caractéristiques de chaque corpus. Le corpus I2b2 utilise un vocabulaire assez spécifique tel que les abréviations conventionnelles des termes médicaux et les abréviations des mots indépendants du domaine. Le corpus I2b2 a été annoté selon des critères bien définis pour être pertinent pour le défi, tandis que le corpus de Berkeley a été annoté avec des règles différentes et avec moins de mesures de contrôle.

Le même principe a été adopté par (Barigou et al., 2011, 2012). Ils ont réussi à extraire des entités telles que : le nom du patient, les maladies, les symptômes, les médicaments à partir des rapports cliniques écrits en français. L'évaluation est réalisée sur un petit corpus et les

résultats montrent que le second système basé sur l'automate cellulaire (Barigou et al., 2012) est capable de couvrir plus d'entités. L'automate cellulaire repose sur une bibliothèque de règles et un lexique de noms propres pour identifier les entités, le système donne des résultats intéressants mais nécessite une évaluation sur un grand nombre de rapports. Les auteurs ont souligné que le rappel était faible en raison de l'absence de règles couvrant toutes les expressions des symptômes et de dosages, etc,

Embarek & Ferret (2012) ont développé Esculape un système de Question/Réponse ; les auteurs ont conçu des patrons morphosyntaxiques pour extraire cinq types d'entités médicales à partir des articles scientifiques écrits en Français. Par l'utilisation des différents dictionnaires, ils ont réussi à extraire des entités de type maladie, symptôme, médicament, examen et traitement.

Afzal et al. (2015) développent un système de reconnaissance des entités médicales, ils traduisent des termes de l'UMLS en Anglais vers le Français par le biais de Google traducteur et Microsoft Bing. Van et al. (2016) utilisent le corpus Français QUAERO, leur méthode se base sur un dictionnaire composé de terminologies françaises de l'UMLS.

2.7.1.2. Systèmes à base d'apprentissage automatique

	Référence	Contribution: extraction de	Corpus	Techniques	P (%)	R (%)	F (%)
Approches à base d'apprentissage automatique	Ben Abacha & Zweigenbaum, (2011)	Problème, traitement et test	I2b2 2010 (Anglais)	SVM	43.65	47.16	45.33
				BIO & CRF	70.10	83.31	76.17
	Huang et al. (2013)	Noms des Maladies	Bio text corpus	CRF	65.98	49.67	56.67
	Wu et al. (2015)	Problèmes, procédures, tests, médications	Corpus clinique (Chinois)	Réseaux de Neurone Profond	92.37	93.21	92.8
	Afzal et al. (2015)	Dix types d'entités (anatomie, procédure, trouble…)	MEDLINE (Français)	CRF	71.1	62.5	66.5
			EMEA (Français)		75.1	76.1	75.6
	Ho-Dac et al. (2016)	Dix types d'entités (anatomie, procédure, trouble…)	QUAERO corpus médical (Français)	CRF	-	-	-
	Sunil & Ashish, (2016)	Noms des maladies	NCBI dataset	Réseaux de Neurone Récurrent	84.87	74.11	79.13

Tableau 2.2 : Travaux représentatifs de la tâche de RE médicales par apprentissage automatique

Le Tableau 2.2 résume quelques travaux de reconnaissance des entités médicales par l'apprentissage automatique. L'évaluation de chaque système est donnée dans le même tableau; elle est basée sur les métriques; précision (P), rappel (R) et F-mesure (F).

Les approches d'apprentissage automatique sont également utilisées pour extraire des entités médicales; un tel système est développé par (Ben abacha & Zweigenbaum, 2011) pour extraire des entités de type problème, traitement et test du corpus i2b2 2010. Ils ont utilisé deux modèles différents, à savoir SVM et CRF. Les meilleurs résultats sont obtenus par le classificateur CRF avec l'utilisation du format BIO "Beginning Inside Outside" et avec des caractéristiques lexicales et morphosyntaxiques combinées avec des attributs sémantiques.

Huang & Hu (2013) ont développé un système d'extraction d'entités de type maladie, ils ont utilisé le classificateur CRF entrainé sur des caractéristiques orthographiques et morphologiques ainsi que des caractéristiques des entités. Ils ont proposé une nouvelle méthode utilisant l'information contextuelle sémantique au niveau de la phrase comme l'une des caractéristiques discriminantes pour la reconnaissance des entités de type maladie. La méthode exploite les types sémantiques des maladies dans le Meta-thesaurus UMLS. Dans cette étude, seuls les concepts de type sémantique « maladie » ou « syndrome » sont conservés. Les résultats montrent une amélioration avec l'ajout du type sémantique de « maladie » ou de « syndrome » comme caractéristique pour entrainer le modèle CRF. Une augmentation de 0.72% pour la F-mesure, de 1.05 pour la précision et de 0.52 pour le rappel.

Wu et al. (2015) ont étudié une méthode d'apprentissage profonde pour la tâche de reconnaissance des entités médicales dans les documents cliniques Chinois. Ils ont développé un réseau de neurones profond pour générer les mots à partir d'un grand corpus non étiqueté grâce à un apprentissage non supervisé. Le système réalise une F-mesure de 92.8%.

Afzal et al. (2015) ont entraîné des modèles de CRF pour chaque type d'entité pour chacun des deux corpus Français (MEDLIN et EMEA). Ho-Dac et al. (2016) développent un système basé sur une méthode d'apprentissage supervisé, pour cela ils ont utilisé le CRF. De plus, ils ont utilisé des ressources externes pour étiqueter les mots (liste de termes de SNOMED et VIDAL).

Sunil & Ashish (2016) ont proposé différents modèles de réseaux de neurones récurrents pour la tâche de reconnaissance des noms de maladies et leur classification en quatre catégories prédéfinies. Ces modèles ont obtenu des performances améliorées lorsqu'ils sont

appliqués au corpus NCBI pour la tâche de reconnaissance des noms de maladies avec une F-mesure de 79.13%.

2.7.1.3. Systèmes utilisant l'approche hybride

Pour la méthode hybride qui combine les approches à base de règles et les approches à base d'apprentissage automatique, il y en a un dans (Ben Abacha & Zweigenbaum, 2011); le modèle CRF encodé par le format BIO est combiné avec la méthode sémantique MetaMapPlus pour extraire des entités médicales du corpus i2b2 2010, ils ont obtenu une F-mesure égale à 77,55%.

Un système de reconnaissance basé sur une méthode hybride est défini dans (Jingchi et al., 2015), ils ont utilisé l'UMLS pour identifier les entités et ils ont entraîner des modèles CRF pour catégoriser les entités du corpus MEDLINE. Le système a obtenu une F-mesure égale à 45.3%.

Un autre système développé par (Eva et al., 2015); ils combinent trois classifieurs: CRF pour les entités non intégrés, le CRF pour les entités intégrés et le SVM pour détecter leur classe sémantique dans le corpus MEDLINE.

Dans le challenge i2b2 2010, un corpus anglais de référence annoté a été construit pour trois tâches : la tâche d'extraction des concepts médicaux, la tâche de classification d'assertion et la tâche de classification des relations.

Uzuner et al. (2011) ont présenté un état de l'art des travaux qui ont participé dans ce challenge et ayant utilisé ce corpus avec l'évaluation de chaque tâche. Ils ont conclu que les systèmes basés sur l'apprentissage automatique peuvent être améliorés par l'utilisation combinée des systèmes à base de règles pour déterminer les concepts médicaux. Selon la tâche, les systèmes à base de règles peuvent soit fournir des données pour l'apprentissage automatique, soit post-traiter la sortie de l'apprentissage automatique.

Névéol et al. (2014) ont développé un corpus médical en Français qui s'appelle QUAERO French Medical Corpus. Il a été développé en tant que ressource pour la tâche de reconnaissance des entités nommées et la normalisation. Il est composé d'un ensemble de titres d'articles scientifiques indexé par MEDLINE et des textes de monographies sur les médicaments publié par EMEA[12].

[12] European Medicines Agency

Dans le CLEF eHealth[13] de 2015; la tâche d'extraction d'information à partir de textes cliniques; particulièrement la sous tâche de reconnaissance des entités cliniques utilise le corpus de (Névéol et al., 2014), un ensemble d'équipes a participé dans cette tâche (reconnaissance et normalisation des entités sont définie selon les groupe sémantiques de l'UMLS). Ces équipes utilisent différentes approches. Le meilleur système est celui de l'équipe ERASMUS (Afzal et al., 2015), il a obtenu une meilleure performance en terme de F-mesure pour les corpus EMEA et MEDLINE (Névéol et al., 2015).

Dans le CLEF eHealth[14] de 2016 (Névéol et al., 2016), le corpus de (Névéol et al., 2014) a été encore utilisé, les données publiées en CLEF eHealth 2015 sont utilisées comme un ensemble d'entraînement et un nouveau test a été diffusé à CLEF eHealth 2016. Un deuxième corpus désigné sous le nom « CepiDC Causes of Death Corpus » a été utilisé; il comprend des descriptions des causes de décès en texte libre en Français, rapportées par les médecins dans les causes normalisées des formes de décès. Ce dernier corpus est utilisé pour le codage à partir de la CIM-10[15].

2.7.2. Extraction des relations médicales

Dans cette section, nous analysons des travaux connexes selon la tâche d'extraction de relation (ER) sémantiques basée sur les deux approches : (i) à base de patrons et (ii) à base d'apprentissage automatique. Les tableaux 2.3 et 2.4 montrent quelques travaux représentatifs de la tâche d'extraction de relations sémantiques dans le domaine médical. Le Tableau 2.3 résume des travaux utilisant des approches à base de patrons et le Tableau 2.4 résume des travaux utilisant des approches à base d'apprentissage automatique. L'évaluation de chaque système est donnée dans le même tableau.

2.7.2.1. Systèmes à base de patrons

Des patrons linguistiques construits semi-automatiquement tenant en compte le type sémantique des entités médicales dans le réseau sémantique de l'UMLS sont utilisés dans (Ben Abacha et Zweigenbaum, 2011) pour extraire les relations entre maladie et traitement. Trois relations sémantiques (cure, prevent et side effect) sont extraites à partir du corpus de MEDLINE. Leur système a obtenu une F-mesure de 67.23%.

[13] https://sites.google.com/site/clefehealth2015/task-1/task-1b
[14] https://sites.google.com/site/clefehealth2016/task-2
[15] Classification International des Maladies, 10ème révision

Dans (Embarek & Ferret, 2012) un processus semi-automatique été utilisé pour extraire des modèles linguistiques de relations. Les auteurs sélectionnent des phrases qui se réfèrent à une relation cible et valident la présence de la relation entre les entités. Le système a obtenu 66.0% de F-mesure.

	Référence	Contribution : extraction de	Corpus	Techniques	P (%)	R (%)	F (%)
Approches à base de patrons	Ben Abacha et al. (2011)	Relations problème-traitement	MEDLINE 2001 (Anglais)	Construction semi-automatique de patrons	75.72	60.46	67.23
	Embarek & Ferret (2012)	Maladie-traitement Maladie-symptôme Maladie-médicament Maladie-examen	Articles scientifique (Français)	Construction semi-automatique de patrons	83.00	55.00	66.00
	Meng & Morika (2015)	Relations entre les emplacements du nodule et les types du nodule, Relations entre les tailles du nodule, Relations entre la taille du nodule et l'emplacement du nodule, Relation entre les dates et la taille du nodule.	Rapports de radiologies (Anglais)	Patrons lexicaux	-	-	94.00 98.00 94.00 92.00
	Lafourcade & Ramadier (2016)	Cause, localisation	Rapports de radiologies (Français)	Patrons linguistiques & contraintes (Patrons sémantiques)	-	-	72.00 54.00

Tableau 2.3 : Travaux représentatifs de la tâche d'ERs par utilisation d'une méthode à base patrons

Meng & Morioka (2015) ont proposé un Framework pour générer automatiquement des modèles lexicaux. Quatre tâches d'extraction de relations ont été évaluées sur des rapports de radiologies. (i) relations entre les emplacements du nodule et les types du nodule. Le système a obtenu une F-mesure de 94%. (ii) relations entre les tailles du nodule, le système obtient une F-mesure de 98%. (iii) relations entre la taille du nodule et l'emplacement du nodule, le système a obtenu 94%, (iv) relation entre les dates et la taille du nodule, l'évaluation de la tâche donne une F-mesure de 92%.

À partir d'un sous-ensemble de rapports de radiologie écrits en Français, les auteurs dans (Lafourcade & Ramadier, 2016) ont construit des modèles linguistiques avec des contraintes sémantiques pour extraire des relations sémantiques telles que (causes, symptômes, localisations, etc.). Le système obtient une F-mesure de 72.0% pour les relations de cause et 54.0% pour les relations de localisation.

2.7.2.2. Systèmes par apprentissage automatique

Ben Abacha & Zweigenbaum (2011) ont extrait des relations médicales sémantiques à partir du corpus MEDLINE. Le même type de relations choisies dans l'approche à base de patrons a été extrait. Un classifieur SVM a été utilisé pour déterminer la relation entre deux entités médicales, pour cela trois types de caractéristiques sont utilisés; les caractéristiques lexicales, morphosyntaxiques et sémantiques. Deux méthodes ont été expérimentées, la première s'appelle « Multi-class machine learning »; dans cette méthode un seul modèle est utilisé pour tous les types de relations et une classification multiple de phrases dans trois classes (chaque classe représente un type de relation), le système a obtenu 90.52% de F-mesure. La deuxième méthode s'appelle « Mono-class machine learning » utilise trois modèles différents chacun associé à une seule relation cible. Le système a atteint une F-mesure de 91.49%.

	Référence	Contribution : extraction de	Corpus	Techniques	P (%)	R (%)	F (%)
Approches à base d'apprentissage automatique	Ben Abacha et al. (2011)	Relations problème-traitement	MEDLINE 2001 (Anglais)	SVM multi-classe	90.52	90.52	90.52
				SVM mono-classe	91.96	91.03	91.49
	Gurulingappa et al. (2012)	Relations Maladie-médicament	MEDLINE (Anglais)	SVM	86.00	89.00	87.00
	Minard et al. (2012)	Relations : Maladie-traitement, Maladie-test	I2b2 2010 (Anglais)	SVM	80.00	63.00	70.00
	Xinbo et al. (2016)	Problèmes-traitements, Problèmes-tests, Problèmes-problème	I2b2 2010 (Anglais)	CRF	-	-	80.00
	JiYoung et al. (2017)	Relation de synonymie & d'hyponymie des entités de types processus, tâche, et matériel.	Articles scientifiques (Anglais)	Neural Network	65.80	63.30	64.50

Tableau 2.4 : Travaux représentatifs de la tâche d'ER par apprentissage automatique

Le travail de (Gurulingappa et al., 2012) se concentre sur l'adaptation d'une machine SVM dans un système d'extraction de relations pour l'identification et l'extraction des effets indésirables liés aux médicaments à partir des rapports de cas de MEDLINE. L'ensemble de données utilisé pour l'entrainement et la validation du système d'extraction des relations se base sur le corpus de l'ADE. Ce corpus contient 2972 rapports de MEDLINE qui sont annotés manuellement par trois annotateurs. Le corpus contient des annotations de 5063 médicaments, 5776 conditions (maladies, symptômes, etc.) et 6821 relations entre les médicaments et les

conditions représentant des effets indésirables. Pour l'entraînement de leur classificateur SVM, Gurulingappa et al. (2012) ont utilisé des dictionnaires pour l'identification des médicaments et des conditions pour générer des relations fausses qui ne relèvent pas dans des relations d'effet négatif. Le système a obtenu une F-mesure totale de 87.0%.

Le classificateur SVM a également été utilisé par (Minard et al., 2012) avec des caractéristiques lexicales, syntaxiques et sémantiques pour extraire les relations de type traitement-maladie et test-maladie. Le système a obtenu 70.0% de F-mesure.

Une méthode basée sur le modèle CRF a été proposée par (Xinbo et al., 2016) comme modèle de classification des relations. Les auteurs ont validé le modèle sur le corpus d'i2b2 2010. Ils ont extrait des caractéristiques des concepts, ces caractéristiques sont optimisées par un modèle d'apprentissage. Le système a obtenu 80.0% de F-mesure.

JiYoung et al. (2017) ont proposé un système d'extraction de relations entre les entités de types: processus, tâche, et matériel; à partir d'un ensemble d'articles scientifiques basé sur une méthode supervisée. Ils ont étudié l'utilisation des réseaux de neurones artificiels. Leur modèle est classé au premier rang dans la tâche 10 de SemEval-2017. Leur système a obtenu une F-mesure de 64.5%.

2.7.2.3. Systèmes utilisant l'approche hybride

Ben Abacha & Zweigenbaum (2011) combinent une méthode à base de patrons linguistiques avec un apprentissage supervisé par un SVM. L'hybridation avec un système de classification multiple a obtenu une F-mesure de 93.73% et l'hybridation avec un système mono-classification a obtenu 94.07% de F-mesure.

Dans les Tableaux 2.3 et 2.4, la plupart des corpus utilisés par différents auteurs pour l'extraction de relations sont en Anglais. Jusqu'à 2013, il n'existait pas de corpus annoté pour le Français, ce qui empêche la communauté française d'utiliser des techniques d'apprentissage automatique dans la tâche d'extraction de relations et même pour la tâche de reconnaissance des entités médicales. En 2014, un corpus Français médical QUAERO développé par (Névéol et al, 2014) a été utilisée pour la tâche de reconnaissance et normalisation des entités cliniques en 2015 et en 2016.

2.8. Conclusion

Dans ce premier chapitre, nous nous sommes essentiellement intéressés à l'étude des systèmes d'extraction d'information d'une manière générale et du domaine médical en

particulier. Lors de la sixième et la septième conférence MUC, la tâche d'extraction d'information est découpée en cinq sous-tâches. Parmi ces sous tâches; on trouve la reconnaissance des entités nommées et l'extraction de relations.

Les entités nommées par exemple permettent un accès particulièrement pertinent au contenu des documents, de plus le repérage et la catégorisation de ces entités représentent un enjeu crucial pour l'analyse et la compréhension des textes. Ils aident en effet d'autres systèmes ou applications dans leurs processus d'analyse.

Les relations qui existent entre les entités doivent être extraites en raison de leur importance, à titre d'exemple; de créer de nouvelles bases de connaissances structurées utiles pour toute application, ou encore augmenter les bases de connaissances actuelles par l'ajout des entités et des faits au thésaurus ou aux bases de données et même de supporter les questions par de bonnes réponses.

Plusieurs systèmes d'extraction ont été réalisés dont certains ont été développés par utilisation des méthodes à base de règles et d'autres par des techniques d'apprentissage automatique.

Dans le domaine médical, l'extraction d'information n'est qu'à ses débuts. Après une étude faite sur les différents systèmes d'extraction d'information dans le domaine médical (Ghoulam et al, 2015a), nous avons constaté que les systèmes d'extraction d'information à base de règles sont plus nombreux comparés avec ceux à base d'apprentissage automatique en particulier dans la communauté Française et ceci est dû au manque de corpus annotés, ce qui explique notre choix d'utiliser une méthode à base de règles pour développer notre système d'extraction d'information médicale. Mais, depuis 2015 et 2016 la tendance converge vers l'utilisation des méthodes à base d'apprentissage automatique.

3. Les Ontologies et la Recherche d'Information

3.1. Introduction

La Recherche d'Information (RI) est une branche qui étudie la construction des Systèmes de Recherche d'Information (SRI). Ce dernier, a pour objectif de retrouver à partir d'une base de documents stockés dans des ordinateurs, les documents pertinents en réponse à une requête d'un utilisateur, et qui correspond au besoin d'information de ce dernier.

Les SRI traditionnels traitent par exemple les synonymes et les abréviations d'un terme comme étant des termes différents. Cela affecte la recherche et nécessite soit l'intégration d'une ressource sémantique telle que les ontologies et les thésaurus soit des utilisateurs pour spécifier toutes les formes possibles d'un terme.

Les ontologies sont des méthodes très intéressantes et efficaces pour représenter et structurer les connaissances d'un domaine donné. Les ontologies sont construites pour partager et réutiliser les informations stockées (Vandecasteele, 2012). Actuellement, leur champ d'application s'élargie considérablement. Les ontologies sont utilisées dans l'ingénierie des systèmes, la recherche d'information, la traduction et d'autres domaines nécessitant une formalisation de la connaissance.

Ce chapitre est composé de deux parties : la première partie est consacrée aux ontologies, nous donnons en premier lieu la définition d'une ontologie et ses éléments de base, ensuite nous présentons les classifications des ontologies. Nous discutons par la suite les méthodologies proposées pour construire une ontologie ainsi que les formalismes de représentation des ontologies. Dans la deuxième partie, nous commençons par les notions de bases de la RI, par la suite nous présentons les modèles de la RI. Nous nous intéressons aussi aux problèmes de la recherche classique, et les solutions proposées pour faire face à ses problèmes, en particulier l'Expansion de Requête (ER). Nous terminons ce chapitre par des travaux sur la construction des ontologies et sur l'utilisation d'une telle ressource dans des SRI pour étendre les requêtes des utilisateurs dans le domaine médical.

3.2. Les ontologies

Les ontologies sont employées dans l'Intelligence Artificielle (IA), le Web Sémantique (WS), le Génie Logiciel (GL) et l'Informatique Biomédicale (IB) comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde.

3.2.1. Définition

Au cours des dernières décennies, le terme « ontologie » a été adopté par des informaticiens, d'abord dans le domaine de l'IA et plus récemment aussi dans d'autres domaines. Au sein de ces domaines, le terme est utilisé dans un sens plus étroit que dans le contexte de la philosophie. L'une des premières définitions a été donnée par (Neches, 1991) qui a défini une ontologie comme suit :

"*Une ontologie définit les termes de base et les relations comprenant le vocabulaire d'un sujet ainsi que les règles de combinaison des termes et des relations pour définir les extensions du vocabulaire*" (Neches, 1991).

Quelques années plus tard, (Gruber, 1993) définit une ontologie comme suit :

"*Une ontologie est une spécification explicite d'une conceptualisation*".

Nous pouvons définir une ontologie comme étant l'ensemble de termes représentant un domaine donné et une spécification de leur sens plus toutes les relations reliant ces termes.

3.2.2. Les composants de base d'une ontologie

3.2.2.1. Les concepts

Un concept, représente l'idée que l'on se fait d'un terme. Également; appelé classe dans certains travaux ou outils. Dans la Figure 3.1 selon le triangle d'Aristote, un concept peut être divisé en trois parties : un terme (signe ou label), une notion (idée) et un ensemble d'objets (Baneyx, 2007).

Exemple: la chose ou l'objet du monde (L'arbre Cactus), le signe ou le terme c'est-à-dire la chaine de caractère ou la photo qui désigne cet objet particulier dans le monde (Cactus, le terme qui désigne cet arbre parmi tous les arbres), et le concept, ou l'idée de l'arbre Cactus.

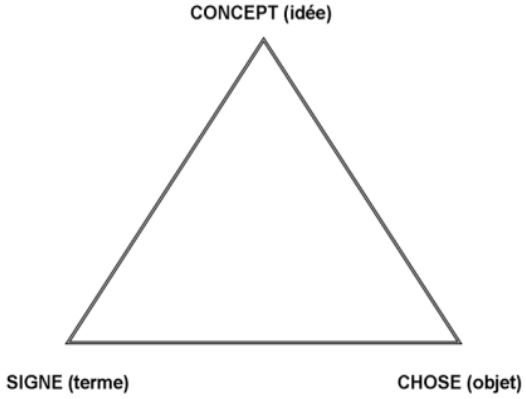

FIGURE 3.1 : TRIANGLE D'ARISTOTE

3.2.2.2. Les propriétés

La propriété ou attribut est une caractéristique associé à un concept et qui peut généralement être dotée d'une valeur. Les attributs peuvent avoir plusieurs facettes décrivant le type de l'attribut, la cardinalité de l'attribut.

3.2.2.3. Les relations

Les relations sont un lien entre deux concepts du domaine. Généralement les ontologies contiennent des relations binaires. La relation la plus utilisée est la relation de subsomption (is-a), elle établit l'hiérarchie de la structure ontologique.

3.2.2.4. Les instances

Les instances sont utilisées pour représenter des éléments ou des individus dans une ontologie. L'instanciation est de créer une instance individuelle d'un concept, et la renseigner avec les valeurs des attributs.

3.2.2.5. Les axiomes

Les axiomes servent à modéliser des phrases qui sont toujours vraies (Gruber, 1993). Ils sont normalement utilisés pour représenter des connaissances qui ne peuvent pas être formellement définies par les autres composants.

3.2.3. Classification des ontologies

Selon l'objectif principal pour lequel une ontologie est conçue, plusieurs classifications d'ontologies ont été proposées. On distingue trois (3) classifications principales :

3.2.3.1. Classification de Van Heijs

Van Heijst, (1995) a classifié les ontologies selon deux (2) dimensions : (i) la quantité et le type de structure de la conceptualisation et (ii) le sujet de la conceptualisation. Pour chaque dimension Van Heijst a regroupé un ensemble de type d'ontologies.

- *La quantité et le type de structure :*

  a) *Les ontologies terminologiques :* spécifier les termes utilisés pour représenter la connaissance dans un domaine de discours. Par exemple; les lexiques.

  b) *Les ontologies de l'information :* spécifier la structure d'enregistrement des bases de données. Par exemple; des schémas de base de données.

c) *Les ontologies de modélisation de connaissances :* spécifier les conceptualisations de la connaissance. Ce type d'ontologies a une structure interne plus riche que les ontologies de l'information et elles sont adaptées à leur utilisation particulière.

- *Sujet de la conceptualisation :*

a) *Les ontologies génériques :* elles sont un peu similaires aux ontologies de domaine, mais les concepts qu'elles définissent sont considérés comme génériques dans de nombreux domaines, en conséquence, elles peuvent être réutilisés dans plusieurs domaines.

b) Les ontologies de domaine : elles expriment des conceptualisations spécifiques à des domaines particuliers, dans notre cas c'est le domaine médical. Elles peuvent être réutilisées.

c) *Les ontologies d'application :* ne sont pas nécessairement réutilisables et elles contiennent toutes les définitions nécessaires. Elles modélisent les connaissances requises pour une application particulière. Elles peuvent inclure des concepts provenant des ontologies génériques et des ontologies de domaines et peuvent contenir des extensions spécifiques de méthodes et de tâches.

d) *Les ontologies de représentation :* clarifier la conceptualisation du formalisme de représentation des connaissances.

3.2.3.2. Classification de Guarino

Guarino, (1998) a définit deux (2) nivaux pour classifier les ontologies : (i) niveau de précision et (ii) niveau de dépendance.

- *Niveau de précision :*

a) *Les ontologies de référence ou hors ligne :* peuvent être utilisées pour établir un consensus sur le partage d'un vocabulaire.

b) *Les ontologies implémentées (partageables) ou en ligne :* sont constituées d'un ensemble minimal d'axiomes, dans un langage d'expressivité minimale, pour prendre en charge des services limités, afin d'être partagés entre des utilisateurs déjà d'accord sur une conceptualisation.

- Niveau de dépendance :

a) *Les ontologies de haut niveau :* ces ontologies décrivent des concepts très généraux comme l'espace, le temps, la matière, l'objet, l'événement, l'action, etc. Ces ontologies sont indépendantes d'un problème ou d'un domaine particulier (elles peuvent être utilisées par de grandes communautés d'utilisateurs).

b) *Les ontologies de domaine :* sont des ontologies qui décrivent le vocabulaire lié à un domaine générique.

c) *Les ontologies de tâches :* sont des ontologies qui décrivent une tâche ou une activité générique.

d) *Les ontologies d'application :* décrivent des concepts dépendant à la fois d'un domaine particulier et d'une tâche (habituellement des ontologies de domaines et de tâches spécialisées).

3.2.3.3. Classification de Lassila et McGuinness

Lassila & McGuinness, (2001) ont classifié les ontologies selon la richesse de leur structure interne. Les ontologies qui contiennent au moins des hiérarchies formelles de type "est-une" sont considérés comme des ontologies simples et les ontologies qui incluent d'autres formalismes, frames, restrictions et contraintes sont appelées ontologies structurées (McGuinness, 2003). Une représentation possible de leur classification pourrait être comme suivant :

- *Les ontologies simples :* ce sont les ontologies qui incluent les hiérarchies formelles de type "est -un (is-a)"; de plus elles peuvent avoir :
  - des relations d'instance formelles,
  - des informations sur la propriété; et
  - les valeurs de restriction.
- Les ontologies structurées : ce sont les ontologies qui incluent de plus par rapport aux ontologies simples ;
  - les spécifications des classes disjointes,
  - les relations inverses et
  - des contraintes logiques générales

La classification des ontologies est résumée dans le Tableau 3.1 :

Van Heijs	**Guarino**	**Lassila et McGuinness**
Quantité et type de structure: -Les ontologies terminologiques. -Les ontologies de l'information. -Les ontologies de modélisation de connaissances.	*Niveau de précision:* -Les ontologies de référence ou hors ligne. -Les ontologies implémentées (partageables) ou en ligne	*Les ontologies simples:* -les hiérarchies formelles de type "est-un (is-a)".
Sujet de la conceptualisation: -Les ontologies génériques -Les ontologies de domaine -Les ontologies d'application -Les ontologies de représentation	*Niveau de dépendance:* -Les ontologies de haut niveau. -Les ontologies de domaine. -Les ontologies de tâches. -Les ontologies d'application	*Les ontologies structurées:* -les hiérarchies formelles avec spécification des classes disjointes, et les relations inverses.

Tableau 3.1 Classification des ontologies selon Van Heijs, Guarino, Lassila et McGuinness

3.2.4. Méthodologie de modélisation ontologique

Plusieurs méthodologies ont été proposées pour la construction des ontologies. Nous abordons dans cette section les méthodologies les plus importantes qui ont réussi à acquérir une certaine crédibilité dans le monde des concepteurs d'ontologies.

3.2.4.1. La méthodologie de Uschold et King

Uschold & King (1995) ont défini une méthodologie qui décrit les étapes du processus de modélisation de l'ontologie. Les auteurs divisent ce processus en plusieurs étapes, qui sont décrites dans la Figure 3.2 :

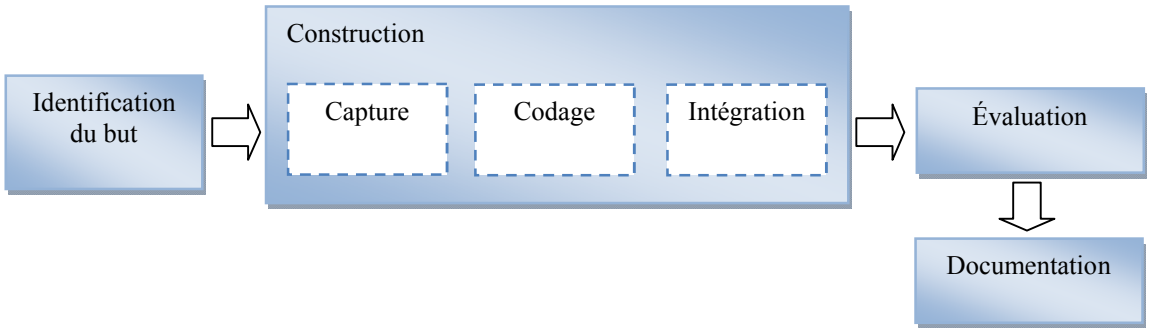

FIGURE 3.2 : LES PRINCIPALES ETAPES DE LA METHODOLOGIE DE USHOLD ET KING

On peut résumer les étapes de cette méthodologie dans les points suivantes :

- Identifier le but.
- Construire l'ontologie en plusieurs étapes :
  - Capture d'ontologie.
  - Codage de l'ontologie.
  - Intégration des ontologies existantes.
- Évaluation.
- Documentation.

3.2.4.2. La méthode METHONTOLOGY

METHONTOLOGY est une méthodologie étendue décrite dans (Gomez-Perez et al., 2003). Cette méthodologie prend en compte trois genres d'activité; les activités de développement, les activités de gestion et les activités de soutien qui sont décrites dans la Figure 3.3 :

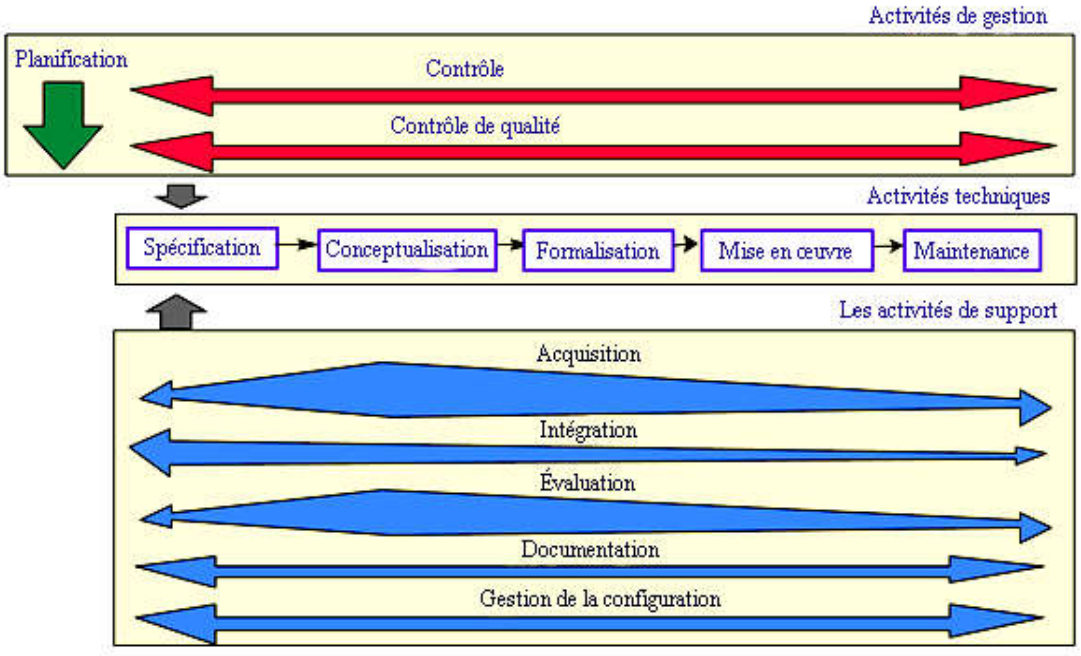

FIGURE 3.3 : PROCESSUS DE DEVELOPPEMENT ET CYCLE DE VIE DE METHONTOLOGY (GOMEZ-PEREZ ET AL., 2003)

Concernant le développement, les activités sont :

- La spécification : établit informellement ou formellement (des questions de compétence) le but et la portée de l'ontologie.
- La conceptualisation :
    - Construire un glossaire des termes (avec des définitions, des synonymes et des acronymes) qui seront inclus dans l'ontologie;
    - Classer les termes en une ou plusieurs taxonomies de concepts;
    - Définir les relations binaires entre les concepts;
    - Construire le dictionnaire de concepts (attributs de classe);
    - Détailler le dictionnaire de concepts (cardinalité, relations inverses, propriétés, etc.);
    - Définir les axiomes et les règles.
- Formalisation.
- La mise en œuvre.
- Maintenance.

3.2.4.3. La méthode ON-TO-KNOWLEDGE

La méthodologie ON-TO-KNOWLEDGE (OTK) est largement décrite dans (Sure & Studer, 2003), elle a été influencée par les propositions méthodologiques de Uschold et ses collègues et de METHONTOLOGY. Le processus méthodologique est divisé en cinq étapes décrites dans la Figure 3.4

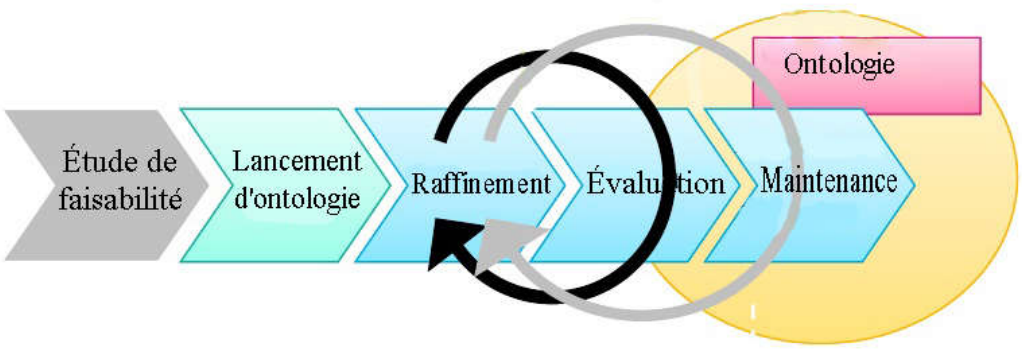

FIGURE 3.4 : LES PRINCIPALES ETAPES DE LA METHODOLOGIE OTK (SURE & STUDER, 2003)

On peut résumer les étapes de ON-TO-Knowledge dans les points suivants :

- Étude de faisabilité;

- Initiation du développement de l'ontologie;
    - Décrire le but, le domaine et la portée de l'ontologie.
    - Décrire les directives de conception;
    - Décrire les sources de connaissances;
    - Poser des questions de compétence;
    - Analyser les sources de connaissances (construire le lexique initial);
    - Créer une description semi-formelle de l'ontologie (brouillon).
- Raffinement: La connaissance est acquise et formalisée dans une approche cyclique;
    - Le processus d'élicitation des connaissances avec des experts du domaine (basé sur l'entrée de la phase de lancement), modification ou extension de l'ontologie préliminaire;
    - Formalisation de l'ontologie.
- Évaluation;
- Application.

3.2.4.4. La méthode ARCHONTE

ARCHONTE (ARCHitecture for ONTological Elaborating), proposée par (Bachimont, 2000), est une méthode de construction d'ontologie à partir de textes. Plusieurs études se sont basées sur cette méthode pour construire leurs ontologies telles que dans (Baneyx, 2007), (Charlet et al., 2012).

On peut résumer les étapes de cette méthode comme suivant :

- Choix des termes pertinents du domaine et normalisation de leur sens puis justification de la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père.
- Formalisation des connaissances, par exemple ajouter des propriétés à des concepts, des axiomes.
- Opérationnalisation dans un langage de représentation des connaissances.

Cette méthodologie comporte trois étapes décrites dans la Figure 3.5

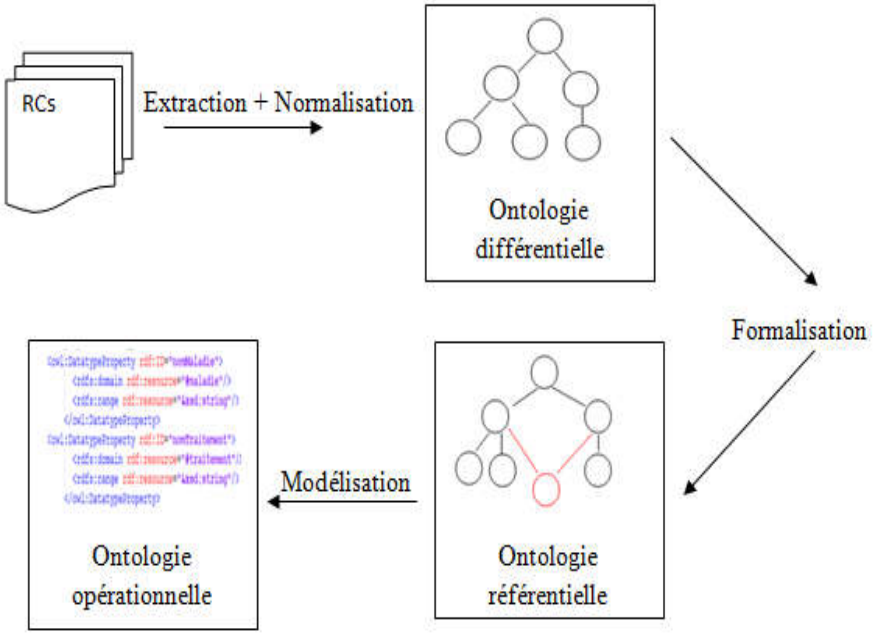

FIGURE 3.5 : CONSTRUCTION D'ONTOLOGIE SELON LA METHODE ARCHONTE.

3.2.5. Les formalismes de représentation des ontologies

Il existe plusieurs formalismes de représentation des ontologies telles que les graphes conceptuels et les logiques de description.

3.2.5.1. Les graphes conceptuels

Les graphes conceptuels sont des graphes qui appartiennent à la famille des réseaux sémantiques, ils sont introduits par (Sowa, 1984). Ce formalisme est utilisé par des travaux comme ceux de (Baneyx, 2007) pour représenter les connaissances du domaine de la pneumologie et ceux de (Embarek, 2008) pour représenter les connaissances médicales.

Le modèle des graphes conceptuels se décompose en deux parties :

- Une partie terminologique pour représenter les types de concepts, les types de relations et les instances des types de concepts. Donc il existe à ce niveau trois ensembles disjoints : l'ensemble des types de concepts (Tc), l'ensemble des types de relations (Tr) et l'ensemble des marqueurs individuels qui sont les instances (M).
- Une partie assertionnelle pour la représentation des assertions du domaine étudié.

3.2.5.2. Les logiques de description

Les logiques de description sont des langages formels, elles se basent sur la logique du premier ordre. La modélisation des connaissances d'un domaine à l'aide des logiques de description comporte deux niveaux, la T-box et la A-box:

- La T-box (T pour la terminologie) décrit les connaissances générales d'un domaine et contient les déclarations des primitives conceptuelles organisées en concepts et relations. Ces déclarations décrivent les propriétés des concepts et des relations.
- La A-box (A pour les assertions) décrit les connaissances factuelles d'un domaine et représente une configuration précise. Elle contient les déclarations d'individus, instances des concepts qui ont été définis dans la T-box.

Parmi les travaux utilisant les logiques de description comme formalisme pour représenter les connaissances, on trouve le travail de (Zaidi–Ayad, 2012) pour construire une ontologie en arabe appliqué au saint Coran.

3.2.6. Les langages de représentation des ontologies

L'une des décisions clés à prendre dans le processus de développement d'une ontologie est de sélectionner le langage (ou l'ensemble des langages) dans lequel l'ontologie sera implémentée et utilisée. En 2002, le W3C a mis en place un groupe de travail pour le développement de langages standards pour modéliser les ontologies utilisables et échangeables sur le Web. Ce sont les langages de balisage pour la représentation des ontologies (ontology mark-up languages) comme montre la Figure 3.6.

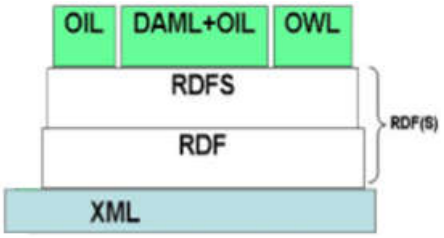

FIGURE 3.6 : LES LANGAGES DE REPRESENTATION DES ONTOLOGIES

Dans cette section nous allons citer les plus utilisés dans la représentation des connaissances ontologiques.

RDF correspond à (Ressource Description Framework), RDFs; le "s" veut dire schémas, c'est une extension de RDF. Le RDF est un modèle de représentation sémantique des

informations du Web qui utilise la syntaxe de XML. Généralement, il est sous forme d'un ensemble de triplets (sujet, prédicat, objet); le « sujet » représente la ressource à décrire; le « prédicat » représente un type de propriété applicable à cette ressource; l'« objet » représente une donnée ou une autre ressource.

Un autre langage qui a été construit comme une extension de RDF; OIL (Ontology Inference Layer). Ce langage utilise des logiques de description pour donner une sémantique claire à ses primitives de modélisation. OIL a été défini dans l'objectif de permettre la spécification et l'échange d'ontologies (Gómez-Pérez, 2004).

DAML (DARPA Agent Markup Language) comme OIL a été développé comme une extension de RDF (S). La dernière version de DAML se combine avec OIL (DAML+OIL). Ce dernier combine des opérations d'intersection, de restriction de valeur, négation, union (Gómez-Pérez, 2004).

Un autre langage s'appelle OWL (Ontology Web Language), créé par le W3C, dérivé du langage (DAML + OIL), et s'appuie sur RDF. Ce langage permet de représenter des ontologies sur le web. Une ontologie OWL est composée d'un en-tête (métadonnées), d'axiomes et de faits. Les axiomes concernent la définition complète ou partielle de concepts et de relations. Les faits concernent des individus pour lesquels on donne des valeurs aux propriétés des classes dont ils sont les instances (Xavier, 2005).

OWL est divisé en trois couches ou (sous langages) : OWL Lite, OWL DL, et OWL Full :

- OWL Lite étend le RDFs et rassemble les fonctionnalités les plus courantes de OWL, il est généralement destiné aux utilisateurs qui ont seulement besoin de créer des taxonomies de classe et des contraintes simples.
- OWL DL est un peu plus complexe, il inclut le vocabulaire OWL complet.
- Enfin, OWL Full est la version la plus complexe d'OWL, elle offre plus de flexibilité pour représenter les ontologies.

3.2.7. Les ressources sémantiques dans le domaine médical

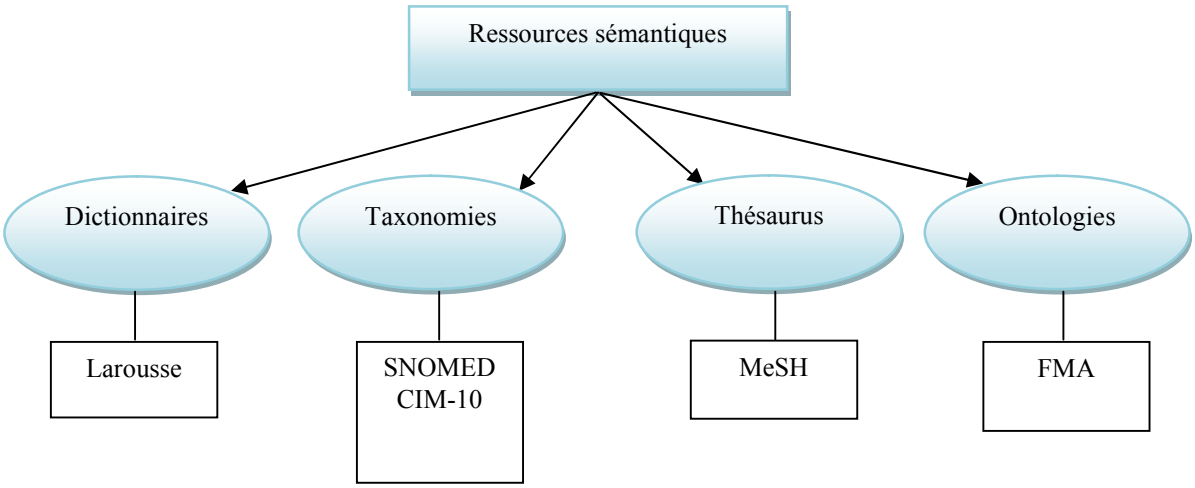

FIGURE 3.7 : LES RESSOURCES SEMANTIQUES

3.2.7.1. Les dictionnaires

Ce sont des ressources qui contiennent les définitions des termes et aussi leurs synonymes (éventuellement leurs antonymes). Mais la structuration des ces ressources y reste peu développée (Audeh, 2014).

3.2.7.2. Les taxonomies

Ce sont des ressources qui peuvent être considérées comme des catégorisations (classifications) pour un domaine précis. La structure d'une taxonomie est exprimée par un seul type de relation, ce qui donne la possibilité de généraliser et de préciser un seul sens. Exemple: SNOMED, CIM-10.

- SNOMED[16] (Systematized Nomenclature of Medicine), la nomenclature systématisée de MEDicine est une nomenclature de type classification multiaxiale et multi-domaines comportant à l'origine sept axes : topographie, morphologie, étiologie, altération fonctionnelle, nosologie, actes médicaux (Baneyx, 2007). La SNOMED est une des classifications médicales la plus complète mais un même concept peut y être décrit de différentes façons et rien n'empêche de le créer par combinaison de concepts inconsistants (Baneyx, 2007). Ce modèle pose des problèmes, par exemple : des termes de différents axes ne sont pas complètement indépendants entre eux, l'axe Maladie fait

[16] https://www.snomed.org

souvent double emploi, certains concepts peuvent apparaître dans plusieurs axes. La SNOMED a évolué en SNOMED-RT (RT: Root Procedure) puis en SNOMED-CT (CT: Clinical terms), fusion de la SNOMED-RT et SNOMED-CT une terminologie britannique de la NHS (services de santé britanniques).

- CIM-10[17] Classification internationale des maladies et des problèmes de santé connexes, 10^{e} révision (connue sous la "CIM-10") est une liste de classifications médicales codant notamment les maladies, signes, symptômes, circonstances sociales et causes externes de maladies ou de blessures, publiée par l'organisation mondiale de la santé (OMS). La CIM-10 permet le codage des maladies, des traumatismes et de l'ensemble des motifs de recours aux services de santé. Les affections (symptômes, maladies, lésions traumatiques, empoisonnements) et les autres motifs de recours aux services de santé sont répertoriés dans la CIM avec une précision qui dépend de leur importance, c'est-à-dire de leur fréquence et de l'intensité du problème de santé public qu'ils posent (par exemple, le chapitre des maladies infectieuses est le plus gros et le plus détaillé parce que ces maladies sont la première cause mondiale de morbidité et de mortalité). Le code des entrées dans la CIM-10 est composé d'un ensemble de caractère ; le premier caractère est une lettre correspondant au chapitre suivie de chiffres pour indiquer les maladies définies à un niveau général.

3.2.7.3. Les thésaurus

Les éléments de base dans un thésaurus sont les concepts, au contraire à un dictionnaire ou à une taxonomie. Chaque concept a un identifiant, une définition, et contient les termes qui peuvent être employés pour le dénoter. Les relations sémantiques les plus fréquentes dans un thésaurus sont les relations d'équivalence (synonymie/antonymie) et les relations hiérarchiques (hyperonyme/hyponyme) (Audeh, 2014). Exemple d'un thésaurus: MeSH.

- Le thésaurus MeSH[18] (Medical Subject Heading) est une ressource médicale, conçu par le NLM (National Library of Medecine) depuis 1960. Le MeSH comprend des descripteurs, des qualificatifs, des paires (descripteur/qualifiant) et les enregistrements de concepts supplémentaires. Les descripteurs MeSH sont disposés dans une structure à la fois alphabétique et hiérarchique. Il a été traduit

[17] http://taurus.unine.ch/icd10
[18] https://www.nlm.nih.gov/mesh/

en français par l'INSERM (Institut national de la santé et de la recherche médicale). Il est utilisé par des systèmes de recherche pour l'indexation, le catalogage et la recherche d'information dans les documents biomédicaux relatifs à la santé.

3.2.7.4. Les ontologies

Les ontologies se situent dans le niveau de connaissance, car dans ces ressources les concepts sont liés par des relations sémantiques, ces relations permettent d'extraire et de calculer des faits, et donc amène à produire de la connaissance (Audeh, 2014).

Les ontologies sont construites pour décrire un monde réel; les thesaurus facilitent l'accès à des contenus; les taxonomies permettent de classer des ressources dans des dossiers, et des catégories.

Exemple d'une ontologie dans le domaine médical : l'ontologie FMA.

- FMA[19] (Foundational Model of Anatomy) est une ontologie de référence pour l'anatomie humaine. Elle concerne la représentation de toutes les classes anatomiques et les relations nécessaires pour la modélisation symbolique de la structure phénotypique du corps humain dans une forme qui soit compréhensible par l'homme et qui soit également traitable par une machine. Le FMA est une open source, mis à la disposition d'utilisateurs qui peuvent récupérer des parties de la modélisation pour l'intégrer dans leur propre ontologie.

Vandecasteele, (2012) a regroupé en trois points l'intérêt de construire une ontologie; (i) les ontologies améliorent la communication (entre les humains, entre les systèmes informatiques, entre les humains et systèmes informatiques). (ii) Potentialités d'inférence informatique; les ontologies permettent la représentation et la manipulation des informations. (iii) Potentialités de la réutilisation des connaissances; les ontologies permettent la structuration et l'organisation d'un domaine.

Dans la deuxième partie de ce chapitre, nous entamons la RI et l'utilisation des ontologies dans un SRI.

[19] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

3.3. La recherche d'information

La signification du terme recherche d'information peut être très large. Cependant, en tant que domaine d'étude académique, la recherche d'information peut être définie comme suit:

La recherche d'information (RI) consiste à trouver généralement des documents de nature non structurée (généralement du texte) qui satisfont un besoin informationnel à partir de grandes collections (généralement stockées sur des ordinateurs) (Christopher et al., 2008).

Selon (Greengrass, 2000) le terme RI fait référence à la recherche des enregistrements non structurés, c'est-à-dire d'enregistrements constitués principalement de textes en langage naturel du format libre. Bien sur, d'autres types de données peuvent également être non structurés, par exemple des images photographiques, des vidéos, etc. Cependant, le domaine RI s'est concentrée sur la recherche du texte en langage naturel, un accent raisonnable compte tenu de l'importance et du volume immense des données textuelles sur internet et dans les archives privées.

Selon (Fishkin, 2005), la recherche d'information est une partie de l'informatique qui étudie la recherche de l'information (pas de données) à partir d'une collection de documents écrits. Les documents récupérés visent à satisfaire un besoin d'information de l'utilisateur exprimé en langage naturel.

3.3.1. Notions de base de la RI

Le processus de recherche d'information illustré dans la Figure 3.8, comprend les principaux concepts suivants :

3.3.1.1. Document

Une unité de recherche. Il peut s'agir d'un paragraphe, d'une section, d'un chapitre, d'une page Web, d'un article ou d'un livre entier (Baeza-Yates & Ribiero-Neto, 1999).

3.3.1.2. Requête

La RI cherche généralement à trouver des documents dans une collection donnée qui répondent à un besoin d'information donné. Le sujet ou le besoin d'information est exprimé par une requête, formulée par l'utilisateur (Greengrass, 2000). Le besoin en information est une expression mentale d'un utilisateur, alors qu'une requête est une représentation possible de ce besoin (Boughanem, 2006).

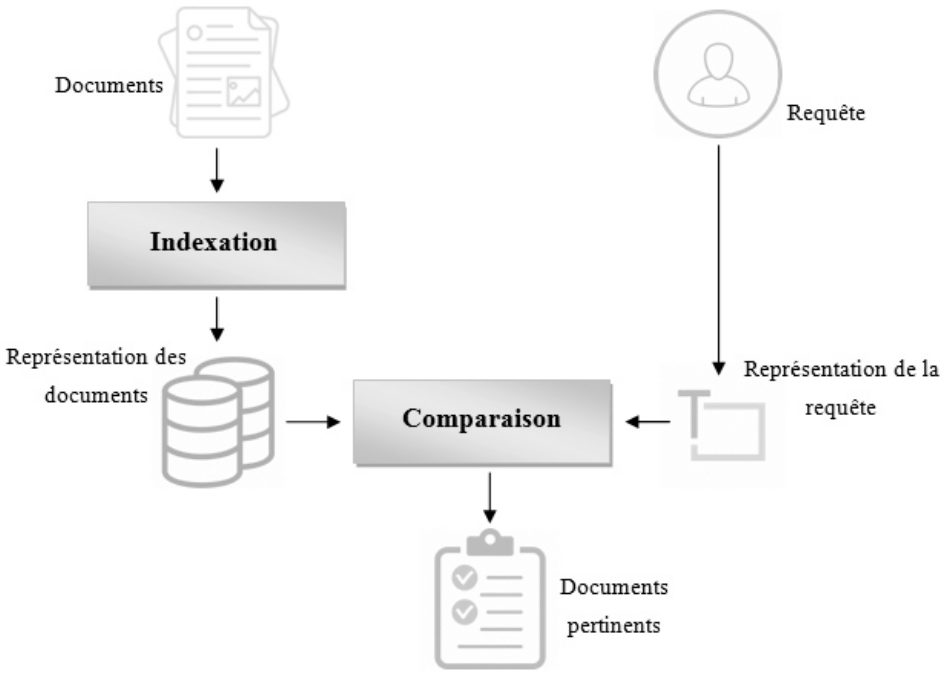

FIGURE 3.8 : PROCESSUS EN U DE RECHERCHE D'INFORMATIONS

3.3.1.3. Modèle de représentation

Le modèle de représentation est un processus qui définit un ensemble de règles et notations permettant la traduction d'une requête ou document à partir d'une description brute vers une description structurée. Ce processus permet d'extraire à partir d'un document ou d'une requête une représentation « sémantique » de son contenu. Ce processus est appelé « indexation ». L'index, le résultat de l'indexation, est une liste de termes ou groupes de termes, appelés descripteurs dans la communauté de RI.

3.3.1.4. Le processus de recherche

Le processus de recherche permet d'associer à une requête, l'ensemble des documents pertinents qui doivent être retournés. Le modèle de recherche d'information est lié au modèle de représentation des documents et des requêtes.

3.3.1.5. Fonction de pondération

La pondération est de caractériser l'importance des termes dans un document en associant un poids à un terme de sorte que les termes importants aient un poids fort.

L'approche de pondération la plus répandue et qui représente le mieux son contenu sémantique est « TF-IDF ». Elle est utilisée dans différentes versions par la majorité des systèmes de recherche d'information.

- TF (term frequency) : La "fréquence du terme" est la fréquence d'occurrence du terme donné dans un document donné (Salton & McGill, 1983).
- IDF (Inverse of Document Frequency) : En revanche, la fréquence de document inverse est une statistique "globale" qui mesure l'importance d'un terme donné dans une collection entière de documents.

3.3.2. Système de recherche d'information (SRI)

Un système de recherche d'information intègre un ensemble de modèles pour la représentation des unités d'information (documents et requêtes) ainsi qu'un processus de recherche qui permet de sélectionner l'information pertinente en réponse au besoin exprimé par l'utilisateur à l'aide d'une requête (Tamine, 2000).

L'objectif d'un système de recherche d'information est de sélectionner les documents pertinents pour une requête d'utilisateur. Deux principales phases caractérisant ce système :

3.3.2.1. L'indexation

L'indexation est un processus de conversion du contenu des documents; elle consiste à extraire les termes des documents. Le résultat de l'indexation constitue ce qu'on appelle l'index.

L'indexation peut être caractérisée par son mode et sa fonction de pondération, Elle représente le contenu d'un document ou d'une requête par des informations qui couvrent mieux son contenu sémantique. Selon (Boughanem, 2006) l'indexation est le processus qui permet de construire un ensemble d'éléments « clés » permettant de caractériser le contenu d'un document afin de retrouver ce document en réponse à une requête.

De manière générale, l'indexation se décompose en trois phases :

- L'extraction des termes du document,
- La normalisation des termes,
- La pondération des termes.

L'indexation peut se faire de différentes manières; elle peut être manuelle, automatique ou semi-automatique :

- *Manuelle :* ce mode fait appel à l'intervention active d'un spécialiste du domaine ou d'un documentaliste pour qu'il analyse chaque document du corpus.

- *Automatique :* sans intervention humaine, ce mode fait appel à un processus entièrement automatisé pour analyser les documents.
- *Semi-automatique :* ce mode est une combinaison des deux autres modes (manuel et automatique), il suggère un expert en indexation pour établir des relations sémantiques entre les termes, de plus l'utilisation d'un ordinateur pour identifier les termes significatifs.

3.3.2.2. L'appariement requête-document

C'est la correspondance entre les termes des documents et ceux de la requête. Le processus d'appariement requête-document permet d'associer à chaque document une valeur de pertinence vis à vis une requête. Les documents ayant une pertinence positive à la requête sont sélectionnés. Généralement le degré de pertinence est calculé à partir d'une fonction notée SC (q, d), ou q est une requête et d est un document. Il existe deux (2) principes d'appariement (matching) : (i) appariement exact et (ii) appariement approché.

- *Appariement exact :* La requête spécifie de manière précise les critères recherchés. Le résultat de la recherche est un ensemble de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas ordonnés.
- *Appariement approché :* La requête décrit les critères recherchés dans un document. Le résultat est un ensemble de documents sensés être pertinents pour la requête. Les documents sont sélectionnés selon un degré de pertinence et sont ordonnés.

3.3.3. Les modèles de recherche d'information

Le rôle d'un modèle de RI est de fournir un cadre théorique pour la modélisation de la mesure de pertinence. Il existe trois principaux modèles, ils sont montrés dans la Figure 3.9 (Mataoui, 2007).

- Les modèles ensemblistes.
- Les modèles algébriques.
- Les modèles probabilistes.

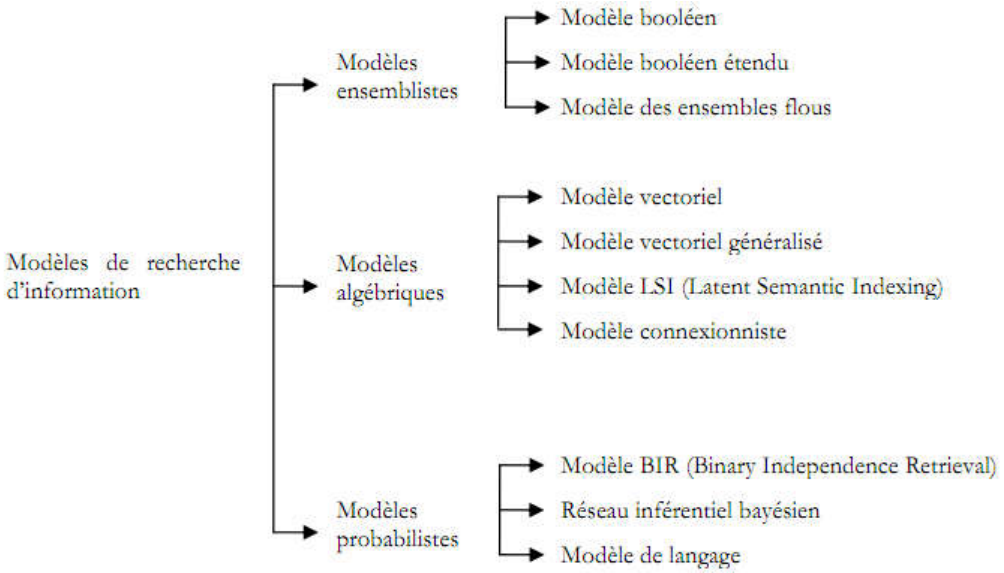

FIGURE 3.9 : CLASSIFICATION DES MODELES DE RECHERCHE D'INFORMATION

Une stratégie ou un modèle de recherche est un algorithme qui prend une requête q et un ensemble de documents d1, d2, ..., dn et identifie le coefficient de similarité SC (q, di) pour chacun des documents, 1 =< i <= n. Selon (Van Rijsbergen, 1979), toutes les stratégies de recherche sont basées sur la comparaison entre la requête et les documents. Les distinctions faites entre les différents types de stratégies de recherche peuvent parfois être comprises en examinant le langage de requête, c'est-à-dire le langage dans lequel l'information nécessaire est exprimée. La nature du langage de requête dicte souvent la nature de la stratégie de recherche.

3.3.3.1. Le modèle booléen

Les modèles ensemblistes ou les stratégies de recherche booléenne reposent sur la théorie des ensembles, ils récupèrent les documents qui sont "vrai" pour la requête. La formulation booléenne n'a de sens que si les requêtes sont exprimées en fonction des termes de l'index (ou mots-clés) et combinés par les opérateurs logiques habituels, conjonction (ET), disjonction (OU) et négation (NON). Ces connecteurs logiques permettent d'effectuer des opérations d'union « OU », d'intersection « ET » et de différence « NON » entre les ensembles de résultats associés à chaque terme.

Le modèle booléen est le premier modèle de la RI, généralement un document est représenté par un ensemble de termes. La requête est un ensemble de mots connectés par des opérateurs booléens; ce modèle est caractérisé par un appariement exact basé sur la présence ou l'absence des termes de la requête dans les documents.

3.3.3.2. Le modèle vectoriel

Les modèles algébriques ou encore vectoriels se basent sur la théorie algébrique. Le modèle vectoriel est proposé par Salton dans le système SMART (System for the Mechanical Analysis and Retrieval of Text) (Salton, 1981). Dans ce modèle, la requête et chaque document sont représentés comme des vecteurs dans l'espace des termes. Une mesure de similarité entre les deux vecteurs est calculée (Grossman et Frieder, 2004).

Exemple:

Documents : D1 = 2 T1 + 3 T2 + 5 T3, D2 = 3 T1 + 7 T2 + T3

Requête : Q = 0 T1 + 0 T2 + 3 T3

Le document le plus pertinent à la requête est D1 car plus le vecteur associée est similaire à celui de la requête plus le document est pertinent comme montre la Figure 3.10.

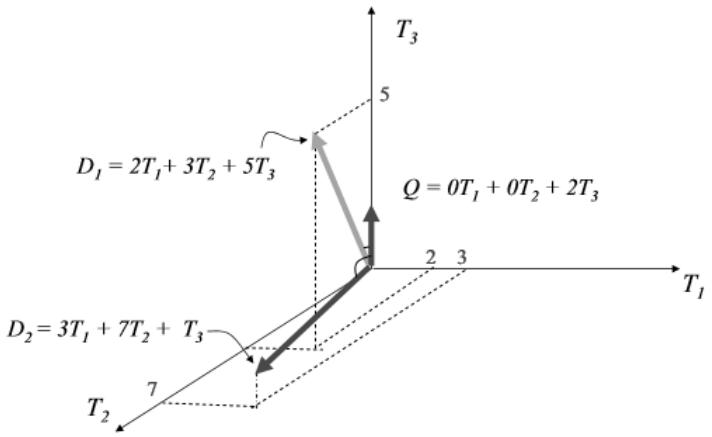

FIGURE 3.10 : EXEMPLE REPRESENTANT LE MODELE VECTORIEL (BOUGHANEM, 2006)

3.3.3.3. Le modèle probabiliste

Les modèles probabilistes se basent sur la théorie des probabilités. Ces modèles estiment la probabilité qu'un document donné soit pertinent pour une requête donnée.

La probabilité qu'un terme apparaisse dans un document pertinent, est calculée pour chaque terme de la collection. Pour les termes qui apparaissent dans une requête et dans un document, la mesure de similarité est calculée comme la combinaison des probabilités de chacun des termes correspondants (Grossman et Frieder, 2004).

3.3.4. Les problèmes de la recherche d'information classique

Habituellement dans les systèmes de recherche d'information (SRI) classiques, l'utilisateur formule son besoin en information par une requête; en revanche, un système de recherche fournit les documents les plus pertinents qui satisfont sa requête. Cependant, il y a beaucoup de difficultés pour développer un SRI efficace.

Hammache (2013) a défini deux problèmes de la recherche d'information classique : l'ambiguïté des mots qui se réfère à des mots lexicalement identiques et portant des sens différents « problème de polysémie ». Et la disparité des mots; qui se réfère à des mots lexicalement différents mais portant un même sens « problème de synonymie ». Selon Bhatnagar & Pareek (2014), les concepts peuvent être décrits par des mots différents dans les requêtes et/ou dans les documents de l'utilisateur.

De nombreuses techniques ont été proposées pour résoudre ces problèmes; l'une de ces techniques est l'expansion (appelé aussi enrichissement ou élargissement) de requête.

3.3.5. Expansion de la requête

Beaucoup de travaux ont été réalisés dans le domaine de l'expansion des requêtes. L'ambiguïté du langage naturel et la difficulté d'utiliser un seul terme pour représenter un concept d'information exigent une expansion de requêtes (Bhogal et al., 2007). Le but principal de l'expansion de requêtes est d'ajouter de nouveaux termes significatifs à la requête initiale pour trouver de nouveaux documents pertinents non proposés avec la requête initiale (i.e. améliorer le rappel et donc réduire le silence), et de placer les documents pertinents déjà trouvés par la requête initiale en début de la liste des résultats (i.e. améliorer la précision et réduire le bruit).

Dans la littérature, nous trouvons plusieurs appellations du mot « expansion », par exemple; enrichissement, reformulation, raffinement, retour de pertinence (Audeh, 2014). Boubekeur (2008) a utilisé le mot reformulation de requêtes, qui consiste, à partir d'une requête initiale exprimée par l'utilisateur, d'enrichir cette requête qui répond mieux à son besoin informationnel.

Dans (Tamine, 2000), la reformulation de requêtes est considérée comme étant un mécanisme adaptatif de modification de requête qui a des conséquences très avantageuses sur les résultats de recherche. Cette modification de requêtes en poids et/ou structure peut être basée sur diverses techniques : utilisation de thesaurus, utilisation des résultats de recherche

locale, injection de pertinence de l'utilisateur etc. Son principe est de modifier la requête de l'utilisateur par ajout de termes significatifs et/ou la ré-estimation de leur poids.

Selon Hammache (2013), la reformulation ou l'expansion de la requête est un processus permettant la construction d'une nouvelle requête. Elle est souvent opérée par ajout et/ou réévaluation des poids des termes de la requête initiale.

Dans cette thèse, nous utilisons le terme d'expansion pour désigner l'enrichissement de requêtes initiales en utilisant une ontologie spécifique au domaine de l'orthopédie.

L'expansion de requêtes peut être manuelle, interactive ou automatique :

- *Expansion manuelle :* repose sur l'expertise des utilisateurs pour décider des termes à inclure dans la nouvelle requête.
- *Expansion interactive :* l'utilisateur choisit les termes d'expansion à partir des suggestions fournies par le système.
- *Expansion automatique :* tout le processus d'expansion est invisible pour l'utilisateur. L'expansion se fait par exemple grâce à l'utilisation de thésaurus linguistiques.

La possibilité d'enrichir la requête initiale de l'utilisateur s'avère importante dans le processus de recherche d'information, car son but essentiel est de permettre aux utilisateurs d'avoir un résultat satisfaisant; c'est-à-dire avoir plus de documents pertinents. Les chercheurs ont développé des techniques efficaces pour l'expansion de requêtes. Une étude de ces techniques est donnée dans (Carpineto & Romano, 2012). Les sources de sélection des termes d'expansion peuvent être regroupées en quatre groupes: (corpus de documents, ressources sémantiques, les logs, données du web).

3.3.5.1. Utilisation d'un corpus de documents

Dans cette catégorie l'enrichissement peut être; global ou local.

Approche globale : Le corpus entier est considéré pour sélectionner les termes d'expansion. Toute la collection de documents est utilisée pour l'enrichissement de requête. Par exemple dans (Jing & Croft, 1994), les auteurs ont proposé une technique globale basée sur la cooccurrence des termes dans le corpus, ils sélectionnent les termes d'expansion les plus similaires à la requête.

Approche locale : Deux techniques sont considérées; la première s'appelle le retour (ou la rétroaction) de pertinence où les termes sont sélectionnés à partir d'un ensemble initial de documents retrouvés en réponse à la requête d'origine (Elayeb et al., 2011; Picariello et al., 2007). La deuxième technique s'appelle le pseudo retour de pertinence; c'est une technique aveugle de retour de pertinence, elle est basée sur l'hypothèse que les documents les premiers classés sont pertinents. Cette technique a été revue par Pragati & Pareek. (2014). Ils ont proposé une hybridation basée sur le corpus, un algorithme génétique floue et la notion de similarité sémantique. Colace et al., (2015) ont proposé l'approche des paires de mots pondérées pour élargir la recherche. Cette structure est extraite de l'ensemble des documents obtenus grâce au retour de pertinence, puis ajoutée à la requête initiale.

3.3.5.2. Utilisation de Ressources sémantiques

Dans cette catégorie, les chercheurs intègrent la notion de sémantique par l'utilisation de connaissances linguistiques externes comme l'ontologie WordNet[20]. Les approches proposées font l'objet de plusieurs études. Abbache et al. (2014) ont utilisé Arabic WordNet pour étendre des requêtes en Arabe, ils ont ajouté à la requête les synonymes de ses termes. Cette méthode n'a pas donné de bons résultats par rapport à la méthode interactive. Pour cela, ces mêmes auteurs ont proposé une méthode à base de règles d'association pour sélectionner automatiquement de WordNet Arabe les synonymes qu'ils faut ajouter à la requête initiale (Abbache et al., 2016). Une amélioration des résultats de la recherche en termes de MAP a été constatée.

Une ontologie générale a été utilisée dans (Audeh et al., 2014) pour l'expansion de requêtes c'est l'ontologie « Yago ». Elle sélectionne de l'ontologie des noms alternatifs pour chaque entité nommée, consistant principalement en tous ses labels possibles.

3.3.5.3. Enrichissement basée sur les logs

Cette technique d'enrichissement de requête est basée sur l'analyse des journaux de recherche appelés aussi historique des sessions de recherche (logs). L'idée est d'extraire implicitement les associations de requêtes suggérées par les utilisateurs Web (Carpineto & Romano, 2012), c'est-à-dire utilisation des informations collectées de façon indirecte suite à des interactions des utilisateurs pendant la recherche d'information. Les logs contiennent généralement des requêtes des utilisateurs, suivies des URL des pages Web sur lesquelles ces derniers ont fait des cliques dans les pages de résultats de recherche. Des modèles sont

[20] http://wordnetweb.princeton.edu

extraits à partir des historiques (logs) dans (Wang & Zhai, 2008), ces modèles sont utilisés pour décider si un terme doit être ou non ajouté à la requête.

3.3.5.4. Enrichissement basée sur des données du web

Généralement ces approches se basent sur les données du web telles que l'utilisation des textes d'ancrage (anchor texts) et aussi par l'utilisation du Wikipédia. Elles sont considérées comme des ressources indirectes (Carpineto & Romano, 2012).

Les textes d'ancrage est la zone réactive sur laquelle le visiteur clique pour déclencher le lien qui le mène de la page actuelle à la page de destination de ce lien. Les textes d'ancrage et les requêtes de recherche réelles des utilisateurs sont très similaires car la plupart des textes d'ancrage sont des descriptions de la page de destination. Dans (Kraft & Zien, 2004), les textes d'ancrage sont classés par l'utilisation de plusieurs critères qui correspondent le mieux à la nature spécifique des données, tels que le nombre d'occurrences d'un texte d'ancrage et le nombre de termes et de caractères qu'il contient.

La méthode basée sur les documents de Wikipédia et les hyperliens est proposée dans (Arguello et al., 2008). Aussi des articles spécifiques de Wikipédia ont été utilisés dans (Xu et al., 2009).

3.3.6. Évaluation des systèmes de recherche d'information

Dans la RI, il est très important de comparer les approches ou encore de mesurer la performance relative d'une approche par rapport à d'autres approches (Boughanem, 2006). Selon (Christopher et al., 2008), afin de mesurer l'efficacité de la recherche d'information, et de procéder à des évaluations automatiques, une collection de tests est nécessaire et doit être composée de trois éléments :

- Un ensemble de documents ;
- Un ensemble de requêtes de test sur l'ensemble de documents du même corpus;
- Un ensemble de jugements de pertinence, généralement une évaluation binaire de pertinence ou de non-pertinence pour chaque paire requête-document.

L'approche standard d'évaluation des systèmes de recherche s'articule autour de deux notions importantes; pertinence et non pertinence de documents.

Grossman & Frieder (2004) ont distingué les catégories de documents qui correspondent à toute requête émise, ils ont illustrés cette distinction par un schéma représentatif donné dans la Figure 3.11ci-dessous :

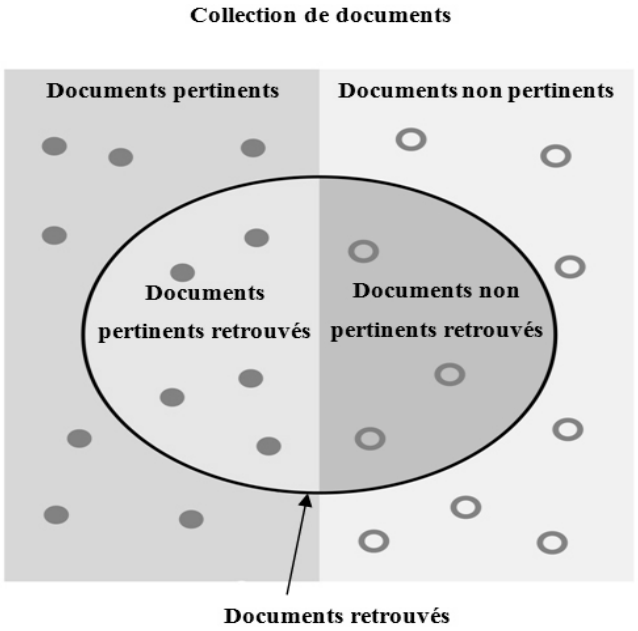

FIGURE 3.11 : JEU DE RESULTATS DES DOCUMENTS

Généralement l'évaluation se base sur la capacité d'un système de recherche d'information à sélectionner des documents pertinents, et cela repose sur des facteurs essentiels; le Rappel, la Précision, le MAP, et la R-précision.

- *La Précision :* est le rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents retrouvés. La précision fournit une indication de la qualité de l'ensemble de réponses (Grossman & Frieder, 2004).

$$\textbf{Précision} = \frac{\textbf{Nombre total de documents pertinents retrouvés par le système}}{\textbf{Nombre total de documents retrouvés par le système}} \qquad \textbf{3.1}$$

- *Le Rappel :* prend en compte le nombre total de documents pertinents; c'est le rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents de la collection qui sont jugés pertinents (Grossman & Frieder, 2004). En fait c'est la proportion de documents pertinents retrouvés par le système parmi tous ceux qui sont pertinents.

$$\textbf{Rappel} = \frac{\textbf{Nombre total de documents pertinents retrouvés par le système}}{\textbf{Nombre total de documents pertinents dans le corpus}} \qquad \textbf{3.2}$$

L'idéal pour un système de RI est d'avoir de bons taux de Précision et de Rappel en même temps. Les deux métriques ne sont pas indépendantes. Il existe des mesures complémentaires au rappel et précision, il s'agit du silence et bruit.

- Le silence est une notion complémentaire au rappel, elle est définie par :

$$\mathbf{Silence = 1 - Rappel} \qquad \mathbf{3.3}$$

- Le bruit est une notion complémentaire à la précision, elle est définie par :

$$\mathbf{Bruit = 1 - Précision} \qquad \mathbf{3.4}$$

- *La MAP (Mean Average Precision) :* est la moyenne des précisions obtenues à chaque fois qu'un document pertinent est retrouvé (Baccini, 2010). Le MAP a été introduit dans le TREC2 pour sa capacité à résumer les mesures de précision à 11 points de rappel. Le MAP est calculé en deux étapes (Hammache, 2013) :
  - Dans la première étape, on calcule la précision moyenne pour une requête donnée ($AveP(q)$), ainsi pour chaque document d_i pertinent retrouvé on calcule sa précision qui est égale au nombre de documents pertinents retrouvés sur le rang de ce document; pour les documents retrouvés non pertinents leur précision est égale à zéro.

$$\mathbf{P(d_i)} = \begin{cases} \frac{\mathbf{r_{ni}}}{\mathbf{n_i}} & \mathbf{si\ d_i\ est\ retrouvé\ pertinent} \\ \mathbf{0} & \mathbf{sinon} \end{cases} \qquad \mathbf{3.5}$$

Ou :

- n_i le rang du document di qui a été retrouvé et qui est pertinent pour la requête.
- r_{ni} est le nombre de documents pertinent retrouvé au rang ni.

  - La précision moyenne pour une requête donnée est alors obtenue en calculant la moyenne des précisions des documents pertinents, exprimée ainsi :

$$\mathbf{AveP\ (q)} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{P(d_i)} \qquad \mathbf{3.6}$$

Où : N est le nombre total de documents pertinents pour la requête q.

- Dans la deuxième étape, on calcule la précision moyenne pour un ensemble de requêtes, en effectuant la moyenne des précisions moyennes de chaque requête, exprimée ainsi :

$$\mathbf{MAP} = \frac{\sum_{k=1}^{NQ} \mathbf{AveP}(q_k)}{NQ} \qquad \mathbf{3.7}$$

Ou : NQ est le nombre total de requêtes.

- La R-précision : Pour un ensemble de requêtes donné, la R-précision est la précision à R, Où R est le nombre de documents pertinents pour Q.

$$\mathbf{R-precision} = \frac{r}{R} \qquad \mathbf{3.8}$$

r : nombre de documents pertinents retournés par le système à la position R où R est le nombre de documents pertinents dans le corpus.

3.3.7. Recherche d'information médicale à base d'ontologie

Le domaine médical, comme tout autre domaine de spécialité plus important et plus traité, il est caractérisé par la spécificité de sa terminologie, la complexité et la richesse de son vocabulaire. Ce domaine intègre un grand nombre de terminologies et d'ontologies médicales.

L'utilisation et la construction des ontologies médicales ont été le centre d'intérêt de plusieurs travaux dans la recherche médicale y compris l'expansion de requêtes et la codification des pathologies et aussi la construction des systèmes de question-réponse médicale qui sont un cas particulier des systèmes de recherche d'information. Le Tableau 3.2 résume quelques systèmes de recherche à base d'ontologie ainsi que quelques ontologies construites.

Dans (Jovic et al., 2007) les auteurs présentent un processus détaillé pour représenter les connaissances du trouble de l'insuffisance cardiaque à l'aide d'ontologies. L'ontologie a été ensuite utilisée dans un système d'aide à la décision médicale.

Dans (Díaz-Galiano et al., 2009) les auteurs ont utilisé le thésaurus MeSH pour l'expansion de requêtes. Les termes associés aux descripteurs MeSH sont considérés comme des synonymes, utilisés pour enrichir les requêtes. Les expériences ont montré des améliorations dans la performance de la recherche d'information dans le domaine médical.

Référence	Contribution	Corpus	Techniques	Nbr. de requêtes	Résultats
Jovic et al. (2007)	Construction d'ontologie du trouble de l'insuffisance cardiaque	Directives pour le diagnostic et le traitement de l'insuffisance cardiaque chronique (Anglais)	Termes médicaux & synonymes (UMLS)	-	-
Díaz-Galiano et al. (2009)	Expansion de requête	Collection multimodale (images et texte) de cas cliniques (Anglais)	MeSH	-	-
Dhombres et al. (2010)	Construction d'ontologie du domaine prénatale	-Comptes rendus d'échographie. -Comptes rendus de radiopédiatrie. -Documents de référence du domaine. (Français)	ON-TO-MENELAS FMA, CIM-10, ORPHANET, CCAM	-	-
Embarek & Ferret, (2012)	Construction d'ontologie médicale	Articles scientifiques (Français)	Extraction d'information	-	-
Charlet et al. (2012)	Construction d'ontologie médicale	Rapports cliniques (Français)	Extraction d'information + CIM-10, CCAM, SNOMED	-	-
Mohameth et al. (2012)	Expansion de requête	MuCHMORE (Anglais)	OBIRS -feedback + MeSH	23	MAP = 0.3708
Khadim et al. (2014)	Expansion de requête	Pages Web (Anglais)	SVM + MeSH + UMLS	50	p@5= 0.5720 p@10 = 0.5460 NDCG@5 = 0.5702 NDCG@10 = 0.5574
Chen et al. (2016)	Expansion de requête	MEDLINE & CNKI (Anglais)	Ontologie de l'hépatite	5	
Zulkarnaine et al. (2016)	Construction d'ontologie médicale	Rapports cliniques (Anglais)	FMA, SNOMED, Rad Lex	-	
Yangyang et al. (2017)	Expansion de requête	PubMed (Anglais)	MeSH	30	p@10= 05129 p@20= 0.4763 MAP= 0.4885

Tableau 3.2 Recherche d'information médicale à base d'ontologie

Pour des perspectives d'automatisation de procédures de codification médico-économique et d'intégration dans un système d'aide à la décision, (Dhombres et al., 2010) ont construit une ontologie médicale du domaine prénatale basée sur l'analyse des documents du domaine et la

réutilisation des ressources terminologiques telles que ON-TO-MENELAS, FMA, CIM-10, ORPHANET, CCAM.

Embarek & Ferret, (2012) ont construit une ontologie pour la médecine générale par extraction des entités médicales et de relations entre ces entités. L'ontologie a été intégrée dans un système de questions/réponses pour aider à trouver des réponses à des questions de l'utilisateur.

Charlet et al. (2012) ont développé une ontologie médicale basée sur les rapports cliniques et la réutilisation des ressources spécialisés tels que CIM-10, CCAM et SNOMED V3.5 afin d'être intégrée dans un système de recherche médicale.

Mohameth et al. (2012) ont proposé une méthode basée sur le retour de pertinence et le thésaurus MeSH pour l'expansion de requêtes. La nouvelle requête construite est basée sur un indicateur de performance maximale. Les résultats obtenus montrent une amélioration dans le rappel et la précision.

Khadim et al. (2014) ont proposé d'utiliser des ressources externes pour améliorer la recherche d'information dans le domaine biomédical. Ils élargissent la requête par l'utilisation de vocabulaires contrôlés tels que le MeSH et l'UMLS. Ils ont utilisé des sites Web médicaux pour l'évaluation de leur système. L'expérimentation montre que la combinaison des synonymes des termes et des relations sémantiques pour l'expansion de requêtes donne de mauvais résultats que l'utilisation des synonymes seuls.

Chen et al. (2016) ont proposé une approche d'expansion sémantique basée sur une ontologie médicale de l'hépatite. Ils se basent sur trois types d'expansion; expansion par synonymes, expansion par hyperonyme-hyponyme et expansion par mots similaires. L'expérimentation sur un grand corpus a montré que l'utilisation simultanée de ces trois types d'expansion améliore la précision de recherche.

Pour éviter la construction d'une nouvelle ontologie de zéro, Zulkarnaine et al. (2016) ont proposé une méthodologie pour développer une nouvelle ontologie médicale par la réutilisation des ontologies biomédicales existantes telles que FMA, SNOMED-CT et RadLex.

Yangyang et al. (2017) ont proposé une approche multi-analyse sémantique pour la recherche d'informations médicales. Cette approche d'expansion de requêtes est basée sur l'utilisation d'une ontologie médicale MeSH. Des expériences sur des collections d'articles médicaux PubMed montrent que cette approche est faisable et efficace par rapport à d'autres

approches traditionnelles en recherche médicale. Ils ont utilisé des articles sur l'hypertension à partir de PubMed. Ils n'ont pas développé et exploité les relations sémantiques dans leur approche.

3.4. Conclusion

Nous avons étudié dans ce chapitre deux domaines importants et fortement liés à savoir la recherche d'information et les ontologies, nous avons abordé les technologies utilisées dans la construction d'une telle ressource, ainsi que les formalismes de représentation des connaissances et les langages pour exploiter ces ontologies. Dans la deuxième partie, nous nous sommes intéressés aux SRI, le but de chaque système de recherche d'information est de satisfaire les besoins des utilisateurs. Ces derniers sont préoccupés par un seul problème : celui de pouvoir récupérer tous les documents dont ils ont besoin d'une façon rapide et efficace. Nous avons vu que l'utilisation des ressources sémantiques telles que les ontologies améliore énormément la qualité et la couverture de la recherche d'information.

Parmi les problèmes de la recherche classique, nous trouvons l'ambiguïté et la disparité des termes de la requête; ces deux problèmes sont cruciaux dans la recherche d'information, et c'est ce qui empêche d'avoir toutes les réponses pertinentes.

Une solution de ces problèmes est proposée dans ce chapitre, c'est l'expansion ou l'enrichissement de la requête, plusieurs approches d'expansion ont été proposées dans la littérature médicale, les plus répandues et utilisées sont les approches basées sur l'utilisation de ressources sémantiques. Dans le domaine médical, plusieurs ressources ont été développées telles que les ontologies médicales, les thésaurus médicaux dans diverses spécialités médicales.

Dans ce travail de thèse, dans le but d'améliorer la qualité de recherche, une ontologie orthopédique est construite pour étendre la requête de l'utilisateur dans un système de recherche médical et c'est ce que nous expliquons dans les chapitres suivants.

Deuxième Partie : Approches Proposées

Dans cette partie, nous développons les approches proposées pour l'extraction d'information et la recherche d'information. Cette partie comprend deux chapitres:

Le chapitre 4 décrit l'architecture du système d'extraction d'information médicale. Nous nous sommes intéressés à la reconnaissance des entités médicales et des relations médicales. Cette extraction a pour objectif le peuplement d'une ontologie médicale dans le domaine orthopédique que nous appelons "Onto_Orthopédique". Cette dernière va servir à étendre la requête de l'utilisateur pour la recherche d'informations dans les rapports médicaux.

Dans le chapitre 5 nous avons étudié l'impact de cette extraction d'information pour la recherche d'information, pour cela nous avons proposé des méthodes d'expansion de requêtes pour améliorer les performances du système de recherche d'information médicale. Ces méthodes ont été évaluées avec l'utilisation de l'ontologie "Onto_Orthopédique" construite.

Par la suite, nous avons proposé une architecture d'un système de recherche à large échelle où plusieurs médecins peuvent y accéder simultanément sur un Cloud privé. Ils peuvent avoir un résultat de recherche pour chaque requête saisie dans les brefs délais par utilisation des outils de stockage distribué et par le traitement parallèle.

4. Reconnaissance des EM et RM à partir des RC écrits en Français

4.1. Introduction

La reconnaissance des entités nommées et l'extraction de relations sémantiques ont été appliquées dans le domaine médical pour extraire des entités médicales (EM) telles que les noms de protéines, noms de gènes, noms de maladies et de traitements à partir de documents médicaux (Meystre et al., 2008), ainsi que les relations médicales (RM) telles que la relation qui relie une maladie à un traitement (Ben Abacha & Zweigenbaum, 2011). L'extraction d'information médicale permet d'analyser des documents médicaux non structurés qui, dans notre cas, ce sont des rapports cliniques sous forme de texte, pour extraire des informations médicales destinées à alimenter une ontologie, cette dernière est envisagée dans la recherche d'information médicale pour l'expansion des requêtes.

Ce chapitre présente notre contribution dans le cadre de l'extraction d'information médicale. Nous présentons un système d'extraction d'information à base de règles à partir des rapports cliniques écrit en Français. Dans ce système, nous commençons par la reconnaissance des EM ensuite, nous passons à l'extraction des RM reliant ces entités. Nous construisons par la suite une ontologie orthopédique que nous appelons "Onto_Orthopédique" basée sur la méthodologie "ARCHONTE".

4.2. Les avantages du système proposé

Nous proposons un système d'extraction d'information pour les deux premières tâches d'extraction; la reconnaissance des entités médicales, et l'extraction de relations sémantiques reliant ces entités.

Les médecins ont besoin d'un accès facile et rapide à des ressources d'information de qualité pour être en mesure de prendre des décisions éclairées concernant les soins appropriées pour les patients. Ils ont aussi besoin de systèmes pour les aider à répondre aux questions cliniques.

La Figure 4.1 montre un diagramme de cas d'utilisation UML qui exprime quelques avantages du système d'extraction proposé. Ce diagramme décrit les fonctionnalités attendues du système. Comme le montre ce diagramme, la tâche de reconnaissance des entités médicales et la tâche d'extraction de relations sont deux tâches essentielles. Ces

dernières aident à la construction du nouveau système médical. Nous pouvons résumer quelques avantages de ces deux tâches d'extraction dans les points suivants :

- *Aide à la construction d'ontologies médicales :* la reconnaissance des entités médicales ainsi que l'extraction de relations médicales aident à la construction d'une ontologie médicale. Cette dernière, répertorie toutes les entités médicales qui ont été identifiés pendant le processus de reconnaissance ainsi que le lien entre ces entités. L'ontologie construite peut être utilisée dans d'autres applications telles que la recherche d'information médicale.
- *Aide à la construction de systèmes médicaux de questions-réponses:* le processus d'analyse de question a besoin d'un système de reconnaissance des entités médicales et d'extraction de relations pour pouvoir répondre précisément au questions des utilisateurs par exemple le système MEANS (Ben Abbacha, 2012) utilise un système d'extraction des EM et RM pour analyser la question de l'utilisateur,.
- *Aide à la construction de systèmes de recherche d'information médicale:* pour être en mesure de satisfaire le besoin informationnel d'un utilisateur en lui retournant des documents pertinents, le processus d'analyse de requêtes a besoin d'un système de reconnaissance d'entités médicales et d'extraction de relations reliant ces entités pour étendre la requête de l'utilisateur en ajoutant des synonymes ou d'autres entités en relation.

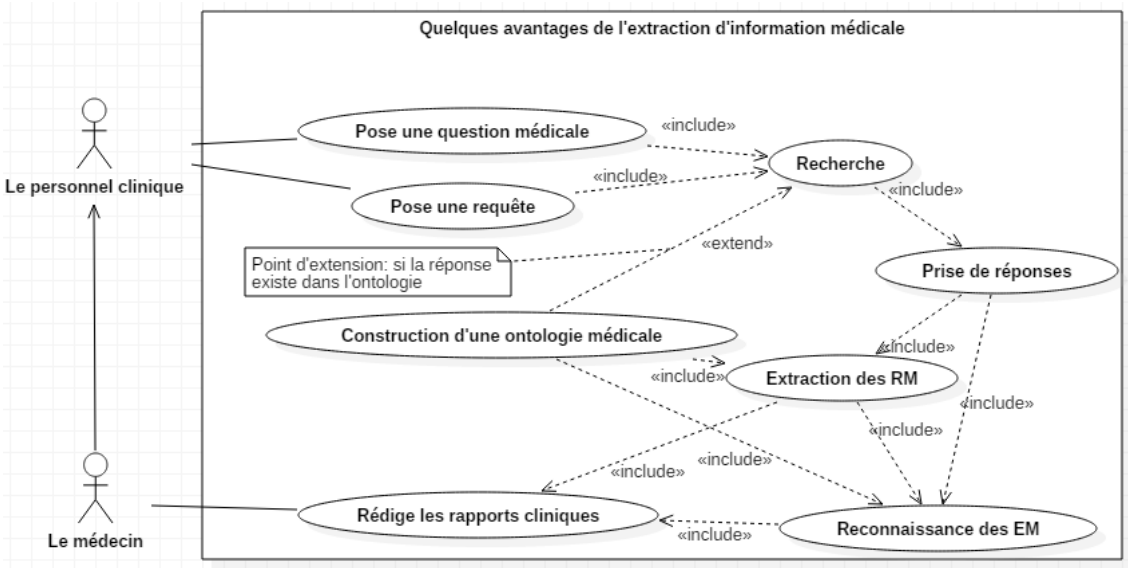

FIGURE 4.1 : LES AVANTAGES DU SYSTEME D'INFORMATION PROPOSE

4.3. Notre objectif

Les rapports cliniques (RC) sont des rapports écrits par les médecins. Ils contiennent des informations médicales telles que les pathologies, les antécédents médicaux et les diagnostics; ils sont enregistrés dans un format textuel non structuré. Ainsi, l'exploitation de l'information contenue dans ces rapports pour répondre aux besoins des médecins est une tâche non aisée.

Généralement, les RC peuvent avoir un impact positif sur la qualité des soins. Toutefois, sans un contenu approprié sous une forme utilisable et accessible, ces avantages peuvent ne pas être atteints. Notre objectif dans ce travail de thèse est le développement d'un système d'extraction d'information médical à partir des RC pour le peuplement d'une ontologie médicale qui va servir à la recherche d'information. Ce système d'extraction est nécessaire et peut faire bénéficier les professionnels de soins de santé. Il est devenu un outil très nécessaire pour accéder à des données précises et aux informations requises, et de plus, il va réduire le temps consacré par les médecins dans la recherche en limitent ainsi les délais de prise de décisions concernant le diagnostic et les recommandations.

4.4. Principe de la solution proposée

Pour extraire de l'information à partir des RC, notre solution suit deux (2) étapes :

- La reconnaissance des EM et
- L'extraction de RM.

Dans l'étape de la reconnaissance des EM nous utilisons une approche fondée sur des règles transformés en grammaires locales. Dans l'étape d'extraction de relations sémantiques, nous proposons d'utiliser une approche qui prend en compte des patrons linguistiques, ces derniers sont transformés aussi en grammaires locales. La motivation et la description de cette approche sont présentées dans cette section.

La Figure 4.2 montre un schéma illustratif global du système d'extraction d'information médicale proposé.

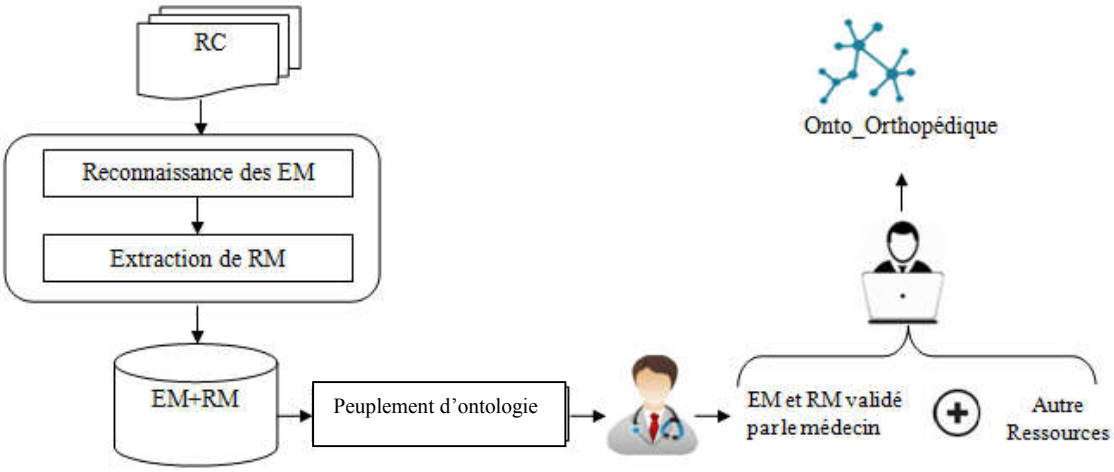

FIGURE 4.2 : SCHEMA GENERAL DU SYSTEME D'EXTRACTION D'INFORMATION MEDICALE.

4.4.1. Les grammaires locales

Une grammaire locale (Gross, 1997) est une représentation par automate de structures linguistiques. Elle est représentée sous formes de graphes. Cette grammaire se présente sous la forme de graphe lexicalisé et permet de représenter et de localiser de manière très précise dans les textes des constructions locales comme les dates, les déterminants numéraux et d'autres expressions linguistiques.

Le logiciel Unitex[21] intègre un éditeur de graphes permettant la construction de grammaires locales. Dans ce logiciel, les états sont laissés implicites et les symboles de l'alphabet d'entrée sont représentés dans des boites connectées entre elles par des arcs non étiquetés. La Figure 4.3 présente un graphe qui décrit la formation d'un groupe nominal (GN) en français exprimée avec la syntaxe Unitex.

Le symbole en forme de flèche est l'état initial du graphe. Le symbole composé d'un rond contenant un carré est l'état final du graphe. La grammaire reconnaît les séquences décrites par les chemins allant de l'état initial à l'état final.

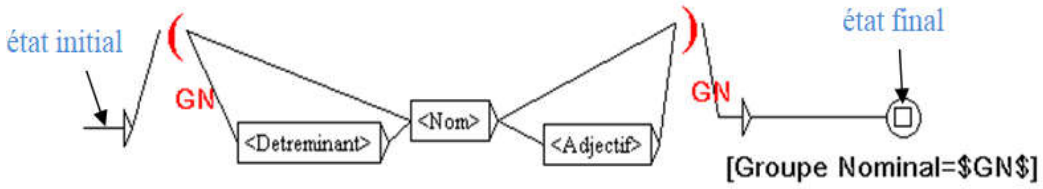

FIGURE 4.3 : EXEMPLE D'UNE GRAMMAIRE LOCALE SOUS FORME DE GRAPHE

[21] http://www -igm.univ-mlv.fr/~unitex

Les grammaires Unitex sont des variantes des grammaires algébriques, également appelées grammaires hors-contexte. Les grammaires Unitex intègrent la notion de transducteur. Cette notion, empruntée aux automates à états finis, signifie qu'une grammaire peut produire des sorties (Paumier, 2016).

Une grammaire algébrique est constituée de règles de réécriture, l'exemple de la Figure 4.3 représente une grammaire qui reconnaît un groupe nominal en Français; l'ensemble de règles de production (R) est représenté comme suivant:

R= { $\boldsymbol{GN} \rightarrow < \boldsymbol{Determinant} >< Nom >$, $\boldsymbol{GN} \rightarrow < \boldsymbol{Determinant} >< Nom >< Adjectif >$

$\boldsymbol{GN} \rightarrow < Nom >$ }

4.4.2. Reconnaissance des entités médicales

Dans l'étape de reconnaissance, nous étudions des rapports cliniques écrits en français pour extraire des entités médicales. L'ensemble des entités retenues pour la construction des règles sont : nom des maladies, des symptômes, des examens, des traitements et des médicaments ((Ghoulam et al, 2015b).

Dans le Tableau 4.1, nous avons illustré par des exemples chaque type (classe) d'entité médicale.

Type entité	**Exemple d'entité médicale**
Maladie	Traumatisme lombaire
Symptôme	Douleur lombaire
Examen	Radiologie
Traitement	Corset
Médicament	Paracétamol, Solupred

Tableau 4.1 : Exemple d'entités médicales et leurs types

Pour extraire ces entités à partir des rapports cliniques, nous adoptons la démarche suivante :

- Construction des dictionnaires (Gazetteers);
- Construction des règles de classification des entités médicales;
- Description des règles dans le format des grammaires locales.

4.4.2.1. Processus de reconnaissance des entités médicales

La Figure 4.4 montre l'architecture du système de reconnaissance des EM. Ce système s'appuie sur deux composantes essentielles : les règles transformées en grammaires locales et les dictionnaires.

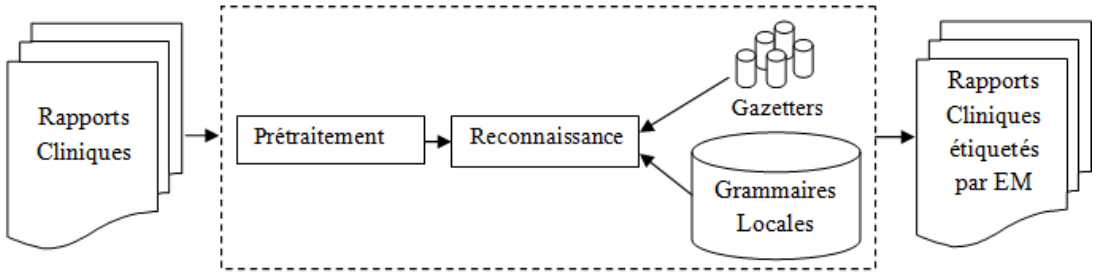

FIGURE 4.4 : ARCHITECTURE DU SYSTEME DE RECONNAISSANCE DES EM

Pour réaliser la reconnaissance des EM. Ce processus est effectué en deux (2) phases : la phase de prétraitement et la phase d'application des ressources linguistiques.

- *La phase de prétraitement :* Nous avons utilisé l'open source Unitex pour segmenter les rapports cliniques (RC). Unitex utilise une grammaire pour découper un texte en phrases ensuite en un ensemble de mots (appelés aussi tokens).

- *La phase application des ressources linguistiques :* Dans cette phase, un ensemble de ressources à savoir des grammaires locales et des dictionnaires sont appliqués pour identifier et catégoriser les entités médicales.

  a) *Application des dictionnaires :* nous avons construit manuellement un ensemble de dictionnaires (appelés aussi Gazetteers) pour les noms des maladies, noms des symptômes, examens cliniques, traitements et médicaments en plus, des dictionnaires pour les adjectifs médicaux et les noms des organes. Ces dictionnaires sont sauvegardés dans le format électronique (DELA); un exemple de quelques entrées du dictionnaire des noms des maladies est montré dans la Figure 4.5.

  Nous avons rassemblé les dictionnaires suivants à partir de différents site web médicaux :

- Dictionnaire des adjectifs[22] qui comprend 514 entrées.
- Dictionnaire des organes (Atlas : human body)[23] qui comprend 384 entrées.
- Dictionnaire des noms de maladies[24,25] qui comprend 343 entrées.
- Dictionnaire des noms de traitement[26] qui comprend 10 entrées.
- Dictionnaire des examens cliniques[27] qui comprend 28 entrées.
- Dictionnaire des symptômes[28] qui comprend 67 entrées.
- Dictionnaire des médicaments[29,30] comprend 8 entrées.
- Liste des mots déclencheurs (les marqueurs) médicaux.

```
alzheimers,alzheimer.maladie
asthmes,asthme.maladie
arthroses,arthrose.maladie
```

FIGURE 4.5 : EXEMPLE DES ENTRÉES DANS UN DICTIONNAIRE DES NOMS DE MALADIE DANS UNITEX.

b) *Application des grammaires locales :* une analyse manuelle des rapports cliniques a été réalisée pour extraire des règles pour chaque type d'entité. Ces règles sont traduites en grammaires locales à l'aide du logiciel open source Unitex.

La Figure 4.6 montre un exemple d'une règle et sa grammaire locale correspondant à une entité médicale de type maladie.

La règle : <RW> + ws + déterminant + ws + (nom_ médical + ws + adjectif)→ catégorie: Maladie

Grammaire locale correspondante :

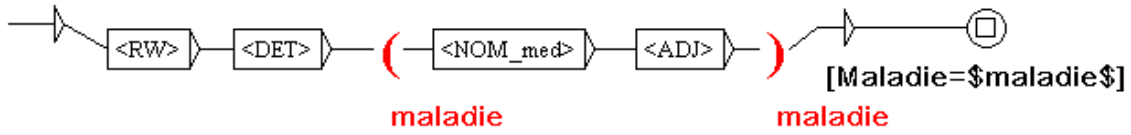

FIGURE 4.6 : EXEMPLE D'UNE REGLE D'IDENTIFICATION D'UNE ENTITE MEDICALE ET SA GRAMMAIRE LOCALE DANS UNITEX

[22] http://www.linternaute.com/dictionnaire/fr/definition/abdominal
[23] http://www.doctissimo.fr/html/sante/atlas/index.htm
[24] http://www.passeportsante.net/problemes-et-maladies-p69
[25] http://www.vulgaris-medical.com
[26] http://www.e-sante.fr/
[27] http://www.vulgaris-medical.com/
[28] http://www.vulgaris-medical.com/symptomes
[29] http://www.eurekasante.fr/medicaments/alphabetique
[30] http://www.doctissimo.fr/html/medicaments/medicaments.html

Conventions d'écriture :

- ws : un blanc.
- <RW> : un déclencheur (Reporting Word).
- <DET> : déterminant.
- < NOM_med> : nom médical.
- <ADJ> : adjectif.

La grammaire locale montrée dans la Figure 4.6 identifie et catégorise une entité médicale, exemple :

"*un malade qui (consulte pour)*^{RW} *(un)*^{DET} *(traumatisme)*$^{NOM\ médical}$ *(lombaire)*^{ADJ}".

L'entité: "*traumatisme lombaire*" est composée d'un nom médical "*traumatisme*" suivi d'un adjectif "*lombaire*".

Pour catégoriser certaines entités, nous nous basons sur les mots déclencheurs. Nous avons remarqué que les entités médicales dans le corpus étudié se trouvent fréquemment à proximité des mots déclencheurs. Par exemple le déclencheur "*consulte pour*", ou le déclencheur "*présente*" en cas des entités de type maladie.

Dans l'exemple précédent : "*un malade qui consulte pour un traumatisme lombaire*", l'entité: « *traumatisme lombaire* » sera catégorisée comme un nom d'une maladie.

Nous avons créé un ensemble de règles sous forme de grammaires locales par l'utilisation d'Unitex. Ces grammaires ont été appliquées pour catégoriser les différentes entités médicales.

Quelques exemples de patrons pour chaque catégorie sont donnés dans le Tableau 4.2 ci-dessous:

Catégories	**Exemple d'entité médicale**	**Exemples de patrons**
Maladie	Traumatisme lombaire	(nom médical + ws + adjectif)
Symptôme	Douleur lombaire	(douleur + ws + adjectif)
Examen_clinique	Radiologie	(Dictionary Lookup = nom d'un examen clinique)
Traitement	Corset	(Dictionary Lookup = nom d'un traitement)
Médicament	Paracétamol	(Dictionary Lookup = nom d'un médicament)

Tableau 4.2 : Exemple de patrons pour chaque catégorie

4.4.2.2. Étude expérimentale et discussion des résultats

Dans cette section, nous décrivons les données et les métriques utilisées pour tester expérimentalement notre approche et nous discutons des différents résultats obtenus.

A. Données expérimentales et nature des RC

Pour le développement du système d'extraction, nous avons préparé un corpus contenant un ensemble de rapports cliniques et ceci pour servir (i) premièrement à construire les règles de reconnaissance des entités médicales et les patrons linguistiques pour l'extraction de relations c'est ce qu'on appelle le corpus d'analyse et (ii) deuxièmement à la reconnaissance des entités et l'extraction de relations, c'est ce qu'on appelle le corpus de test.

Pendant notre visite au service de traumatologie de l'hôpital de Chlef[31]; nous avons remarqué qu'il n'existe pas de version électronique des rapports cliniques. Pour chaque patient admis au service, le médecin rédige manuellement sur papier des notes décrivant l'état du malade et les différents traitements recommandés. Aucun logiciel n'est utilisé pour la rédaction d'un tel rapport.

Les rapports cliniques sont consultés par les médecins du service pour avoir une idée de l'évolution de l'état du malade. Ces rapports diffèrent d'un document normal dans les points suivants; les informations dans les RC sont organisées en sections, environ cinq (5) sections sont considérées :

- L'identification de l'hôpital et du service de plus des informations sur le malade (nom, prénom, âge).
- L'historique du malade (des maladies chroniques, maladies antérieures).
- Le Diagnostic de l'état actuel.
- Le Bilan; y compris les examens cliniques faits sur le malade.
- L'ensemble des traitements prescrits et évolution.

Avec l'aide du médecin et le secrétaire médicale, nous avons constitué un ensemble de 80 rapports médicaux saisi et corrigé manuellement pour former le corpus médical.

- 50 rapports sont utilisés pour la construction des règles d'extraction, et
- 30 rapports sont utilisés pour l'évaluation du système d'extraction.

[31] Hôpital d'Ouled Mohamed CHLEF

B. Protocole expérimental

Les mesures standard d'extraction d'information telles que: la précision, le rappel et la F-mesure sont utilisées pour évaluer le système de reconnaissance des entités médicales. Nous calculons les valeurs de ces mesures selon les formules suivantes :

$$\text{Precision}=\frac{VP}{VP+FP}$$

$$\text{Rappel}=\frac{VP}{VP+FN}$$

$$\text{F-Mesure}=\frac{2*(\text{Precision}*\text{Rappel})}{\text{Precision}+\text{Rappel}}$$

Nous avons donc construit la table de confusion illustrée dans le Tableau 4.3 pour calculer la précision et le rappel de chaque catégorie d'entités.

		Expert (médecin)	
	Pour une catégorie d'entité	Oui	Non
Système	Oui	VP	FP
	Non	FN	VN

Tableau 4.3 : Table de confusion pour chaque catégorie d'entité médicale.

- VP : Vrai Positifs; nombre d'entités médicales correctement identifiées comme appartenant à une classe d'entité.
- FP : Faux Positifs; nombre d'entités médicales incorrectement identifiées comme appartenant à une classe d'entité.
- FN : Faux Négatifs; nombre d'entités médicales incorrectement rejetés d'une classe d'entité.
- VN : Vrai Négatifs : nombre d'entités médicales qui sont correctement rejetés d'une classe d'entité.

Nous avons ensuite calculé le rappel moyen, la précision moyenne et la F-mesure moyenne en considérant la macro-moyenne pour l'ensemble des catégories. Ces formules sont calculées ainsi:

$$Macro - Précision = \frac{\sum_{i=1}^{nc} Precision(c_i)}{nc}$$

$$Macro - Rappel = \frac{\sum_{i=1}^{nc} Rappel(c_i)}{nc}$$

Avec nc : représente le nombre de catégories des entités considérées dans cette étude.

C. Discussion des résultats

Dans cette section, nous présentons le résultat des expériences effectuées pour la tâche de reconnaissance des entités médicales à partir de rapports cliniques.

La Figure 4.7 montre la précision, le rappel et la F-mesure pour chaque catégorie. L'analyse des expérimentations nous a permis d'observer que la performance globale du système pour les cinq catégories est bonne. Les résultats de chaque catégorie en termes de précision, rappel et F-mesure sont présentés dans le Tableau 4.4.

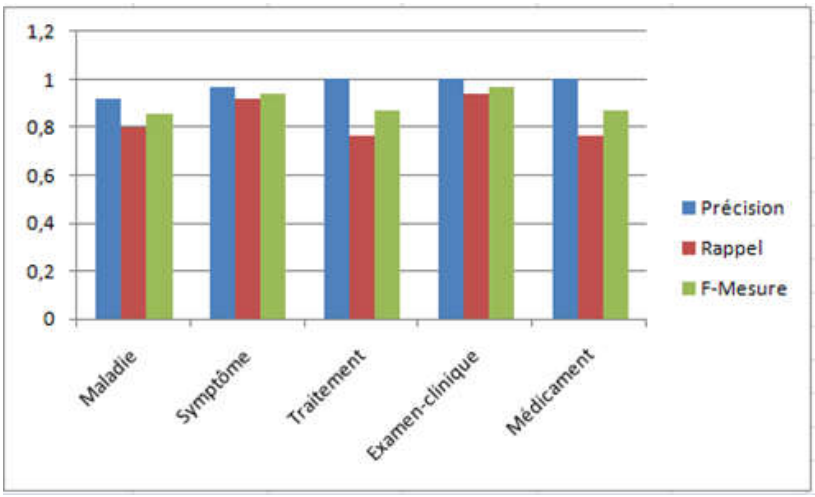

FIGURE 4.7 : PERFORMANCE DU SYSTEME DE RECONNAISSANCE DES ENTITES MEDICALES

Catégorie	Précision	Rappel	F-Mesure
Maladie	0,921	0,800	0,856
Symptôme	0,971	0,917	0,943
Traitement	1,000	0,765	0,867
Examen-clinique	1,000	0,941	0,969
Médicament	1,000	0,765	0,867

Tableau 4.4 : Évaluation du système de reconnaissance des EM

Évaluation	Macro-précision	Macro-Rappel	Macro-F-mesure
Moyenne	97,84%	83,78%	90,06%

Tableau 4.5 : Résultats macro-moyenne de précision, rappel et F-mesure du système de reconnaissances des EM

En général, le système à base de règles obtient de bons résultats en termes de macro-précision et macro-rappel qui représentent respectivement la moyenne des précisions et des rappels. Comme indiqué dans le Tableau 4.5, nous avons une précision globale de 97,84% et un rappel global de 83,78%. Ces résultats sont très intéressants mais doivent être évalués avec une plus grande collection de rapports cliniques.

L'analyse de ces résultats nous a permis de conclure que le système produit des résultats importants en termes de précision. Le petit nombre de rapports cliniques explique les bons résultats obtenus et donc la haute performance pour certains types d'entités. Le système n'a pas réussi à reconnaitre toutes les entités en raison du nombre insuffisants d'entrés des dictionnaires et aussi l'insuffisance des règles pour identifier les entités de types traitement et de types médicament ce qui explique la diminution dans le rappel de ces deux types d'entités par rapport à d'autres types d'entités.

4.4.3. Extraction des relations médicales

Nous définissons une relation sémantique médicale comme étant la relation qui existe entre deux entités ou deux concepts présents dans les rapports cliniques.

Dans cette étape, nous extrayons les relations sémantiques entre les entités identifiées dans l'étape décrite dans la section 4.4.2 en utilisant des modèles linguistiques sous forme de grammaires locales.

Ce système d'extraction de relations s'appuie sur les résultats du système de reconnaissance des entités médicales. Il utilise comme entrée les RCs étiquetés par les entités médicales identifiées par le système d'extraction des entités nommées ((Ghoulam et al, 2015b). On peut schématiser ce système dans la Figure 4.8.

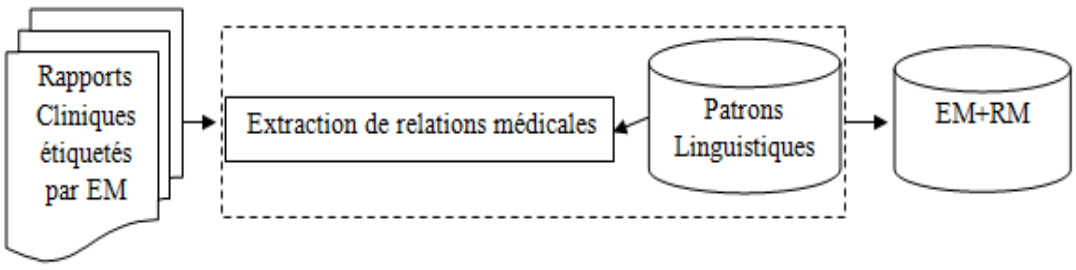

FIGURE 4.8 : SCHEMA GENERAL DU SYSTEME D'EXTRACTION DE RELATIONS

On se concentre sur quatre types de relations (voir Figue 4.9) :

- La relation Traite (maladie, traitement);
- La relation Detecte (maladie, examen-clinique);
- La relation Signe (maladie, symptôme);
- La relation Soigne (maladie, médicament).

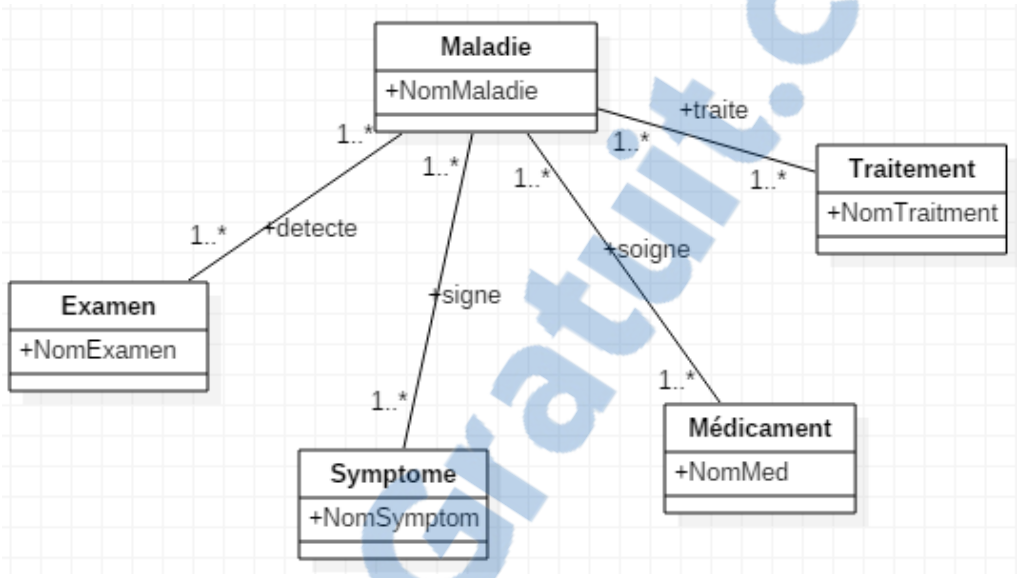

FIGURE 4.9 : LES QUATRE RELATIONS CIBLES EXPRIMEES AVEC LE DIAGRAMME DE CLASSE UML

L'approche que nous proposons pour extraire ces relations inclut les deux points suivants :

- La construction de patrons linguistiques pour chaque type de relation et
- La transformation des patrons en grammaires locales pour identifier la relation dans le corpus.

A. La construction de patrons linguistiques de relations

L'utilisation de patrons linguistiques est l'une des méthodes les plus efficaces pour l'extraction automatique d'informations à partir de corpus textuels s'ils sont efficacement conçus. Pour construire des patrons pour une relation donnée, nous avons construits des patrons linguistiques à partir de 50 RC; le Tableau 4.6 résume quelques exemples de patrons linguistiques pour chaque catégorie.

Pendant cette étape de construction de patrons, nous avons observé que les relations sémantiques ne sont pas toujours exprimées avec des mots explicites comme ceux présentés dans les exemples du Tableau 4.6. Par conséquent, il est difficile de

construire des modèles ou des patrons qui peuvent couvrir toutes les expressions pertinentes.

Catégorie de la relation	Entités liées	Nombre de patrons	Exemples simplifiés
Detecte	Maladie-Examen	26	<Examen>confirme la présence de <Maladie>
Traite	Maladie -Traitement	22	<Maladie>, ayant subit le <Traitement>
Signe	Maladie -Symptôme	16	{tokens}*avec signe de <Symptôme>
Soigne	Maladie -Médicament	20	<Maladie> {tokens}*-<Médicament>

Tableau 4.6 : Exemples de patrons linguistiques

B. La transformation du patron linguistique en grammaire locale

Les patrons linguistiques exprimant les relations sémantiques sont représentés sous formes de grammaires locales. Quelques exemples de grammaires sont montrés dans les Figures 4.10 et 4.11.

Relation: Traite (maladie-traitement)

Exemple: Fracture de la rotule…traité par attèle en genouillère.

Patron linguistique: <maladie> traité par <traitement>.

Grammaire correspondante:

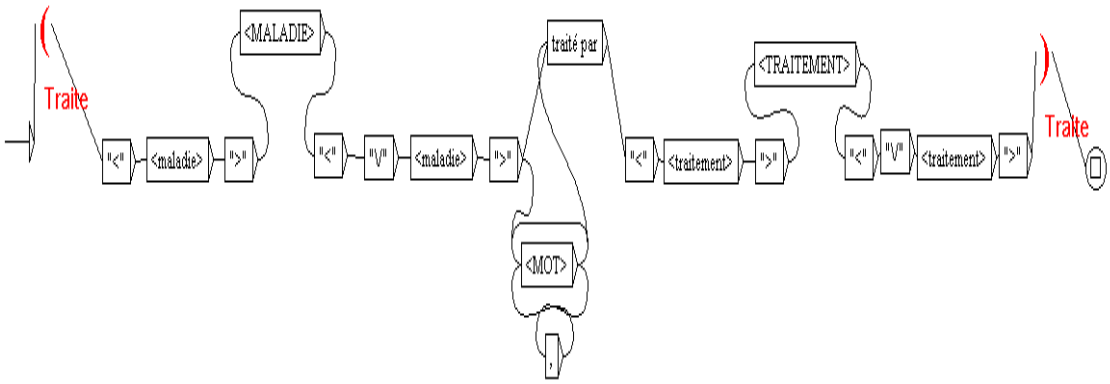

FIGURE 4.10 : EXEMPLE D'UNE GRAMMAIRE LOCALE DE LA RELATION 'TRAITE'

***Relation*:** Detecte (maladie-examen)

Exemple: radiologie : fracture de L01

Patron linguistique: <examen> : <maladie>

Grammaire correspondante:

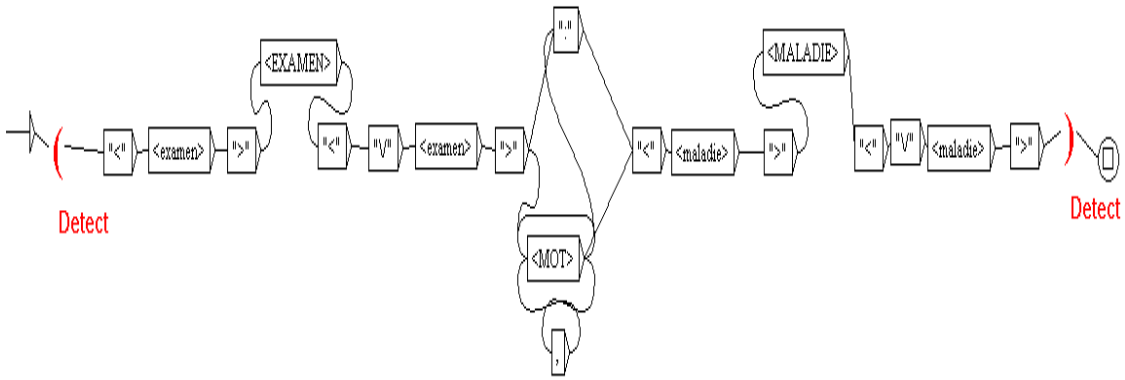

FIGURE 4.11 : EXEMPLE D'UNE GRAMMAIRE LOCALE DE LA RELATION 'DETECTE'

4.4.3.1. Étude expérimentale et discussion des résultats

Dans cette section, nous décrivons les données expérimentales et les métriques d'évaluation utilisées pour tester expérimentalement notre approche d'extraction de relations et de discuter les différents résultats obtenus.

A. Donnée expérimentale

Nous avons utilisé les mêmes rapports cliniques traités dans la tâche de reconnaissance.

B. Protocole expérimental

Nous avons utilisé les mêmes mesures d'évaluation standard décrites dans la sous-section (4.4.2.2 B) ; le rappel, la précision et la F-mesure.

- VP: Vrai Positifs; nombre de relations validées correctes.
- FP: Faux Positifs; nombre de relations détectées par le système et qui n'étaient pas confirmé par le médecin.
- FN: Faux Négatifs; nombre de relations qui figuraient dans le rapport, mais le système ne les a pas détectés.

C. Discussion des résultats

Dans cette section, nous présentons les résultats obtenus; Nous discutons certains problèmes et caractéristiques de l'approche proposée.

Nous évaluons notre système pour extraire quatre types de relations, en utilisant deux manières différentes de reconnaissance des EM : (i) manuelle: nous avons étiqueté manuellement les RCs par les EM et (ii) automatique: le système de reconnaissance décrit dans 4.2.2 fait l'étiquetage automatique des EM. Les résultats obtenus pour chaque cas sont présentés dans le Tableau 4.7.

Relations ↓	Etiquetage manuel des EM			Reconnaissance automatique des EM		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
Detecte	0.857	0.654	0.724	0.625	0.447	0.499
Traite	0.666	0.483	0.481	0.666	0.483	0.481
Signe	0.222	0.222	0.222	0.111	0.111	0.111
Soigne	0.5000	0.500	0.500	0.500	0.500	0.500

Tableau 4.7 : Évaluation du système d'extraction des relations sémantiques pour chaque type de relation

Le Tableau 4.8 présente les résultats obtenus en termes de macro-moyenne de précision, de rappel et F-mesure.

Évaluation	Reconnaissance manuel des EM			Reconnaissance automatique des EM		
	Macro-Précision	Macro-Rappel	Macro-F-mesure	Macro-Précision	Macro-Rappel	Macro-F-mesure
La moyenne	56.15%	46.50%	48.21%	47.57%	38.55%	39.80%

Tableau 4.8 : Résultat macro-moyenne de précision, de rappel et F-mesure du système d'extraction des RM

Pour l'extraction des relations de type '*Traite*', nous avons obtenu un rappel de 48,33%, 66,66% de précision et 48,18% de F-mesure. D'autres approches similaires à notre travail comme par celle de (Ben abacha & Zweigenbaum, 2011) ont obtenu un rappel de 60,46%, et 75,72% de précision et 67,23% de F-mesure. Dans (Lee et al., 2004), ils ont obtenu 84% de rappel, 48,14% de précision et 61,20% de F-mesure pour l'extraction des relations de traitement.

La limitation majeure de cette approche est la dépendance totale à des patrons reliant des entités médicales. Dans le Tableau 4.8, la comparaison entre les deux résultats obtenues, dans les deux cas de reconnaissance des EM montre que ces résultats sont approximativement convergents, ce qui signifie qu'il y'a un manque de patrons dans les RCs et ceci explique les faibles résultats dans le rappel.

Le système n'a pas réussi à extraire toutes les relations en raison du nombre insuffisant de patrons utilisés, en particulier entre les entités de type maladie et de type symptômes. En outre, en raison de la structure des RCs; dans ces rapports les médecins utilisent des entités médicales pour décrire l'état du malade mais sans utiliser d'expressions indiquant qu'il existe réellement des relations entre les entités existantes, cela exprime le manque de liens entre ces entités et ceci explique aussi les résultats faibles du rappel.

4.4.4. Bilan

Le travail réalisé est fondé sur l'extraction des informations à partir d'un corpus médical (rapports médicaux). Deux étapes le caractérisent: (i) la reconnaissance des entités médicales par utilisation d'une approche à base de règles avec modélisation de ces règles sous forme de grammaires locales et (ii) l'extraction des relations sémantiques syntagmatiques par utilisation d'une approche à base de patrons linguistiques.

Nous avons expliqué notre choix de l'approche à base de règles en raison d'absence de corpus médical annoté dans la communauté française. En général les expérimentations montrent que cette approche permet d'obtenir une bonne précision, mais présente l'inconvénient d'exiger un grand effort humain et un temps considérable pour construire les règles par rapport à la grande variabilité et la structure complexe des rapports cliniques.

L'un des obstacles les plus importants dans l'identification ou la reconnaissance des entités médicales est la variation élevée de terminologies dans le domaine médical. En revanche, l'évolution des entités médicales telles que les nouvelles abréviations, les noms de nouvelles maladies ou de nouveaux médicaments constituent des obstacles qui peuvent limiter l'évolutivité de l'approche à base de règles. Aussi la principale limitation de l'approche est leur manque de portabilité qui limite leur extension à d'autres domaines médicaux.

Pour l'extraction des relations sémantiques avec les patrons linguistiques; il a été difficile d'extraire les relations médicales existant entre les entités médicales à cause de l'absence de patrons linguistiques et ce, dû à la nature des RCs, car la relation entre deux entités médicales n'est pas toujours exprimée dans un RC par une expression claire. Par exemple: "*….consulte pour une <Maladie> fracture du coude </Maladie>….avec signe de <Symptôme>douleur du coude</Symptôme> et <Symptôme>gonflement</Symptôme>….*", le système détecte une relation entre la maladie "*fracture du coude*" et le symptôme "*douleur du coude*" à travers le patron "*avec signe de*", mais pas pour le symptôme "*gonflement*".

4.4.5. Construction de l'ontologie orthopédique

4.4.5.1. Les composants de l'ontologie

Comme nous l'avons vu, les ontologies et les représentations de connaissances sont propres à un domaine donné. Elles existent sous forme de concepts et de relations. Les composantes de notre ontologie orthopédique sont :

- *Les concepts :* dans cette ontologie orthopédique, le concept est porteur d'une connaissance. Il représente les différents types des entités étudiées dans ce travail, comme Maladie, Médicament, Examen, Symptôme et Traitement.
- *Les propriétés :* la propriété est la caractéristique d'un concept qui peut généralement être dotée d'une valeur. Si nous prenons l'exemple de Maladie (dans le domaine orthopédique c'est la fracture ou traumatisme) nous pouvons désigner quelques propriétés comme; nom de la maladie, la localisation de la maladie (gauche, droite, supérieur, inférieur, médiane).
- *Les entités médicales ou instances :* l'ensemble des objets auxquels le concept fait référence, sont appelés les instances. Dans notre ontologie orthopédique, ce sont les entités médicales.

  - Exemple :
    - Traumatisme, fracture sont des instances de Maladie.
    - Corset, ostéosynthèse sont des instances de Traitement.
    - Radiologie rachidienne est une instance du concept Radiographie.
    - Douleur lombaire est une instance du concept Symptôme.

- *Les relations médicales :* les relations syntagmatiques sont des relations dépendantes du corpus, spécifiques à un domaine, ici le domaine médical "orthopédie", elles ont un sens précis dans le domaine utilisé. Dans cette thèse Quatre types de relations syntagmatiques sont considérées; *Traite* (maladie, traitement), *Détecte* (maladie, examen), *Signe* (maladie, symptôme) et *Soigne* (maladie, médicament).

4.4.5.2. Construction de l'ontologie orthopédique

Pour la construction de l'ontologie orthopédique (Onto_Orthopédique), nous utilisons la méthode ARCHONTE définie par (Bachimont et al., 2002). Cette méthodologie est bien décrite dans le chapitre 3. La méthode ARCHONTE est une méthode ascendante de construction d'ontologie à partir des textes du domaine en trois étapes.

La Figure 4.12 montre les étapes de construction de l'ontologie "Onto_Orthopédique".

La procédure d'alimentation d'Onto_Orthopédique passe par deux (2) étapes essentielles :

- Extraction des informations y compris les entités médicales et les relations médicales à partir des rapports cliniques. Pour cela cinq classes sémantiques ont été considérées pour classer les entités médicales identifiées à partir des RCs (voir la sous-section 4.4.2). De même quatre classes de relations ont été définies pour typer les relations sémantiques extraites (voir sous-section 4.4.3).
- Les informations extraites sont validés par un médecin spécialiste du domaine orthopédique et même par des sites web médicaux. Les informations validées sont ajoutés à l'ontologie "Onto_Orthopédique", en précisant l'appartenance à une classe ou une sous-classe.

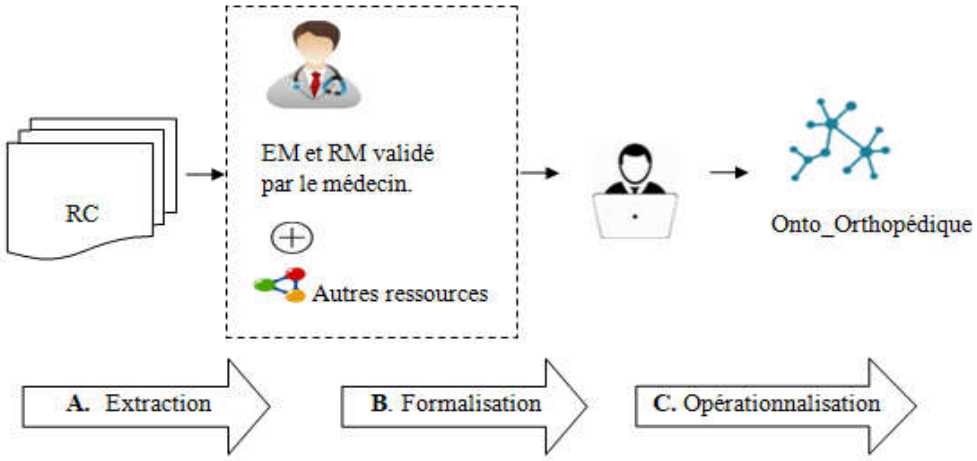

FIGURE 4.12: CONSTRUCTION DE L'ONTOLOGIE "ONTO_ORTHOPEDIQUE".

A. Étape 1 : Extraction des informations à partir des RC et Normalisation sémantique (Passage du corpus à l'ontologie différentielle)

Dans cette étape de l'extraction des informations à partir des RC, nous avons extrait des informations médicales à partir d'un corpus contenant un ensemble de rapports cliniques du domaine orthopédique. Les informations extraites sont des entités médicales et des relations sémantiques. Bachimont, (2002) propose d'introduire clairement le sens de chacun des concepts de l'ontologie. Dans notre cas, nous avons associé pour chaque entité médicale extraite un type sémantique (la classe point de vue ontologie). Pour cela cinq types sémantiques ont été considérés : Maladie, Symptôme, Traitement, Examen et Médicament. De même nous avons associé à chaque relation syntagmatique extraite un type sémantique. Cette étape d'extraction et d'affectation de type est bien décrite dans les sections 4.4.2 et 4.4.3.

La méthode ARCHONTE repose sur l'idée de préciser les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père. La structuration de ces concepts, en fonction des similarités et des différences partagées avec leurs concepts pères et leurs concepts frères dans un arbre, permet de passer à «l'ontologie différentielle». Avec l'aide du médecin et à partir d'autres ressources médicales (des sites web médicaux[32, 33, 34, 35]), nous avons pu structurer ces concepts.

[32] http://www.doctoralia.fr/maladies/discipline/traumatologie+-+chirurgie+orthopedique-1075
[33] http://www.chirurgie-orthopedie chanzy.com/traumatologie
[34] http://www-sante.ujf-grenoble.fr/SANTE/corpus/corpus.htm?menu=01+rub=02+disci=D21

B. Étape 2 : Formalisation de l'ontologie orthopédique (passage de l'ontologie différentielle à l'ontologie référentielle)

Dans l'ontologie différentielle chaque concept extrait est placé de manière précise dans la structure hiérarchique, en précisant les principes différentiels qui les définissent. Bachimont (2002) propose de définir quatre principes différentiels fondamentaux : le principe de communauté avec le père, le principe de différence avec le père, le principe de différence avec les frères, et le principe de communauté avec les frères.

Prenons comme exemple les deux entités médicales "*corset thoraco-lombaire*" et "*corset lombaire*" qui sont de type "Traitement". Ces deux entités sont des concepts frères dont le concept père est "*corset d'immobilisation*". Le principe de communauté avec le concept père est dans le traitement orthopédique car se sont des dispositifs médicaux orthopédiques sur mesure. Le principe de différence avec le père est que le concept père "*corset d'immobilisation*" est général, alors que le concept "*corset thoraco-lombaire*" ou le concept "*corset lombaire*" sont indiqué à une région bien défini du rachis. Le principe différentiel entre les concepts frères est dans la partie du rachis, car dans le cas de "*corset lombaire*" on s'intéresse à la partie lombaire, et dans le cas de "*corset thoraco-lombaire*" le traitement se fait dans la partie thoraco-lombaire du rachis. Le principe de communauté entre les concepts frères est l'immobilisation stricte des régions de la colonne vertébrale (le rachis). Avec l'aide du médecin nous avons construit l'ontologie différentielle, précisé les sous-classes et les instances des classes auparavant définies.

Exemple : traumatisme← traumatisme lombaire← fracture L1.

Aussi avec l'aide du médecin nous avons complété manuellement l'ontologie par l'ajout de synonymes.

Exemple : arthrose de genou = gonarthrose, gonflement du genou= épanchement de synovie.

A cette étape de la méthode ARCHONTE, on passe à l'ontologie référentielle ou l'ontologie formelle. La structure de la hiérarchie ne représente plus un arbre mais un treillis car les extensions des concepts peuvent avoir un sous-ensemble commun.

[35] https://www.hetop.eu/hetop

Exemple : pour les concepts "*Traumatisme Lombaire*" et "*Traumatisme Thoracique*", la formalisation permet de créer un nouveau concept "*Traumatisme Thoraco_Lombaire*".

C. Étape 3 : Opérationnalisation de l'ontologie (le langage utilisé pour représenter les informations de l'ontologie)

Dans cette étape, l'ontologie référentielle est traduite manuellement en une ontologie opérationnelle. L'ontologie référentielle est spécifiée dans un langage opérationnel. Nous avons choisi le langage OWL pour la représentation des connaissances car c'est le langage le plus expressif, il est reconnu par le W3C[36] comme le standard le plus utilisé dans la construction des ontologies pour le web sémantique. Nous utilisons le logiciel protégé 4.3 pour l'opérationnalisation d'Onto_Orthopédique car ce logiciel offre la possibilité de visualiser graphiquement l'ontologie. Le résultat des deux étapes est montré dans la Figure 4.13 à l'aide du logiciel Protégé 4.3.

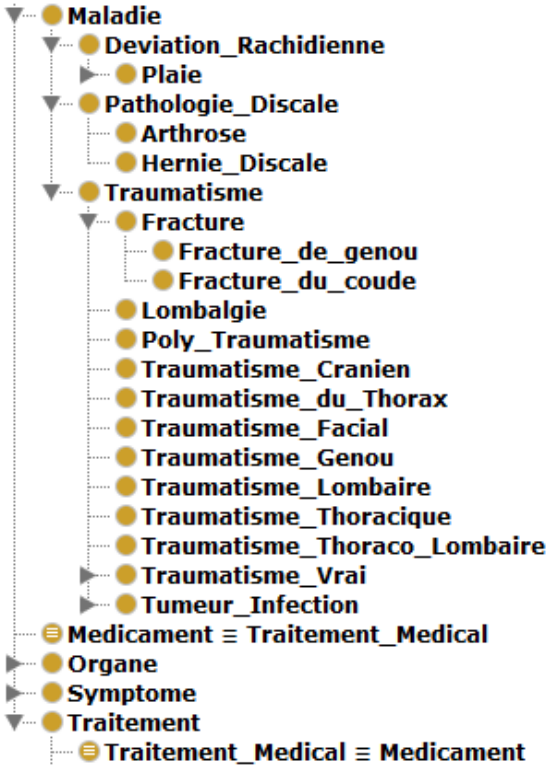

FIGURE 4.13 : REPRESENTATION DE L'ONTOLOGIE SOUS PROTEGE 4.3[37]

[36] http://www.w3.org/
[37] https://protege.stanford.edu/download/protege/4.3/installanywhere/Web_Installers/

4.5. Conclusion

Dans ce chapitre, nous proposons une méthode fondée sur l'extraction d'information médicale pour construire une ontologie orthopédique. Le processus d'extraction d'information passe par deux étapes; la reconnaissance des EM et l'extraction des RM.

Pour l'étape de reconnaissance des EM, nous utilisons une approche à base de règles pour extraire les entités médicales à partir des rapports cliniques écrit en français. Les résultats expérimentaux montrent que l'approche proposée a obtenu une F-Mesure moyenne de 90.06%.

Pour l'extraction des RM, nous utilisons une approche basée sur des patrons linguistiques qui identifient le lien sémantique entre deux entités médicales différentes. L'approche repose sur deux étapes principales: (i) la reconnaissance des entités médicales grâce à l'utilisation de l'approche à base de règles, et (ii) l'exploitation des patrons linguistiques extraits à partir des rapports cliniques.

Après avoir extrait les entités médicales et les relations sémantiques reliant ces entités, nous construisons l'ontologie orthopédique "Onto_Orthopédique". Nous nous sommes basés sur une méthodologie semi-automatique appelée "ARCHONTE" pour la construction et la structuration de l'ontologie. Onto_Orthopédique est créée dans le format OWL supporté par Protégé.

Dans le chapitre suivant, nous présentons l'utilisation de l'ontologie orthopédique "*Onto_Orthopédique*" dans un système d'expansion de requête pour améliorer la qualité de recherche d'information médicale.

5. Recherche d'information médicale

5.1. Introduction

Avec la quantité croissante des données disponibles dans le domaine médical, l'accès aux informations utiles et pertinentes en temps réel devient une tâche importante et primordiale pour les praticiens et les chercheurs. En effet, les systèmes de recherche d'information aident les utilisateurs dans leurs activités quotidiennes pour satisfaire leurs besoins. Habituellement, l'utilisateur formule son besoin d'information dans une requête; en retour, un système de recherche d'informations fournit les documents les plus pertinents censés satisfaire la requête de l'utilisateur. Cependant, il existe de nombreuses difficultés dans le développement de SRI efficace. L'une de ces difficultés est le problème de vocabulaire; car les utilisateurs peuvent exprimer leurs besoins en utilisant des mots différents ayant des significations similaires (c'est le cas des synonymes) et un même mot avec des significations différentes (c'est la polysémie). Selon (Bhatnagar & Pareek, 2014), les concepts peuvent être décrits par des termes différents dans les requêtes et / ou les documents des utilisateurs.

De nombreuses techniques ont été proposées pour résoudre ce problème; par exemple, on trouve les techniques d'expansion de requête. Depuis longtemps, l'expansion de requêtes (ER) a été une solution pour améliorer la qualité de la recherche d'un SRI. L'expansion de requête peut être réalisée de différentes façons, elle peut être manuelle (l'utilisateur choisit des termes d'expansion), interactive (l'utilisateur choisit les termes d'expansion à partir des suggestions du système) ou automatique (tout le processus est invisible pour l'utilisateur). Récemment, les systèmes sélectionnent des termes d'expansion à partir des ressources externes telles que les ontologies et les hiérarchies lexicales qui ont considérablement amélioré leurs résultats.

En médecine, la plupart des ontologies contiennent des concepts reliés par des relations hiérarchiques mais souffrent de l'absence de relations syntagmatiques (Embarek, 2008). Ainsi, une ontologie construite à partir des RCs est nécessaire pour modéliser les concepts du domaine médical et aussi ce type de relations syntagmatiques entre ces concepts.

Dans ce chapitre, nous étudions l'impact de l'extraction d'information pour la recherche d'information. Nous utilisons pour cela l'ontologie orthopédique "*Onto_Orthopédique*" décrite dans le chapitre 4 pour l'expansion de requêtes dans un SRI médicale. Trois

méthodes d'expansion sont proposées et comparées avec la méthode de recherche classique.

5.2. Les objectifs de la solution proposée

Pour résoudre les problèmes de vocabulaire liés à la recherche d'information médicale, nous avons choisi de développer un système de recherche sémantique, basé sur une ontologie du domaine médical construite à partir des rapports médicaux. Nous avons opté pour cette solution pour analyser et étudier l'impact de l'extraction d'information dans la recherche d'information médicale.

Le choix du développement d'un tel système de recherche sémantique est motivé par les faits suivants:

- Les rapports cliniques sont considérés comme une source d'information qui a un impact positif sur la qualité des soins.
- Les médecins ont besoin de consulter et de rechercher à travers ces rapports les informations qui leur permettent de prendre des décisions ou les orienter vers des traitements plus appropriés.
- Ce type de systèmes de recherche médicale devient de plus en plus sollicité. Il permet aux chercheurs d'accéder à des informations requises et de réduire le temps nécessaire pour prendre des décisions.

5.3. Architecture adoptée

Avant de détailler l'approche d'expansion de requête adoptée, la Figure 5.1 présente l'architecture proposée pour un système de recherche d'information médicale.

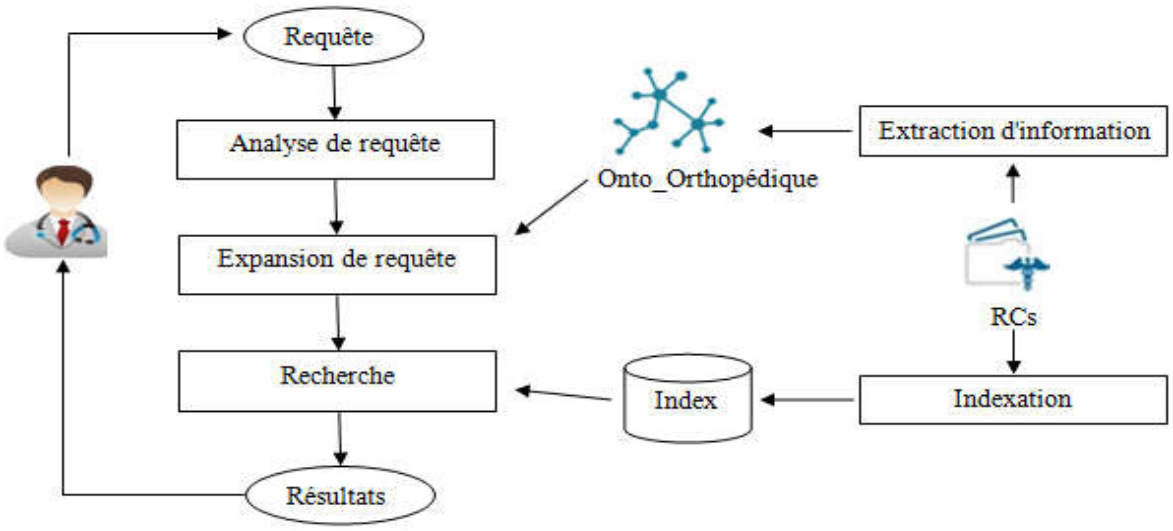

FIGURE 5.1 : ARCHITECTURE PROPOSÉ DU SYSTÈME DE RECHERCHE SÉMANTIQUE

Le système de RI que nous proposons se compose de quatre (04) phases: la phase d'indexation, la phase d'analyse de requête, la phase d'expansion de requête et la phase de recherche.

5.3.1.1. La phase d'indexation

Dans cette phase, l'indexation consiste à représenter le contenu des RCs dans un index. Nous utilisons le système open source Lucene[38], le résultat du traitement des RCs est enregistré dans un index. Nous avons employé l'analyseur "*French Analyser*" pour analyser les RCs qui sont écrits en français. Cet analyseur a pour fonctions d'extraire les différents termes de ces RCs et de les stocker dans un index.

5.3.1.2. La phase d'analyse de requête

Dans la phase d'analyse de requête, nous extrayons les entités médicales existantes dans la requête de l'utilisateur. Nous avons utilisé pour cela la même méthode développée dans (Ghoulam et al., 2015b). La sortie de cette phase est un ensemble d'entités médicales (E1, E2, ..., En) et un ensemble de termes (T1, T2, ..., Tk).

5.3.1.3. La phase d'expansion de requête

Dans la phase d'expansion de requête, nous utilisons l'ontologie "Onto_Orthopédique" pour étendre chaque entité médicale reconnue dans la requête.

[38] https://lucene.apache.org

L'ontologie orthopédique "Onto_Orthopédique" que nous avons conçue contient un ensemble d'entités médicales reliées par des relations sémantiques. Cette ontologie est conçue à partir des rapports cliniques réels. Nous avons enrichi cette ontologie par des synonymes et des descendants en utilisant un ensemble de sites web après validation par un médecin expert. Le Tableau 5.1 résume le contenu de cette ontologie.

Onto_Orthopédique	**Information**
Langage	OWL sous protégé
Nombre des EM	>140
Nombre de relations	>65

Tableau 5.1 : Contenu de l'ontologie "Onto_Orthopédique" en nombre

```
Algorithme d'expansion de requête
Entrée          Q : requête initiale
                W : ontologie "Onto_Orthopédique"

Intermédiaire   E, E1 et SD : listes vides des entités médicales
                R, RS listes vide de relations sémantiques
Sortie          Q1 requête étendue

Début
  E  ← Extraire les entités médicales à partir de Q;
  R  ← Extraire les relations médicales à partir de Q;
  E1 ← E ;
  Pour chaque élément e dans E1 faire
     SD ← Extraire synonymes et descendants de e à partir de W;
     E  ← E + SD;
  FinPour
  E1 ← E;

  Pour chaque couple (e1, e2) dans E faire
     RS ←Extraire les relations sémantiques entre(e1,e2) à partir de W;
     R  ← R + RS ;
  Fin Pour

  Pour chaque relation r dans R faire
  Pour chaque élément e dans E1 faire
     Extraire les entités qui sont en relation r avec l'entité e dans W;
     Ajouter ces entités à E;
  FinPour
  FinPour

   Q1 ← E;

   Retourner Q1;
Fin
```

FIGURE 5.2 : ALGORITHME D'EXPANSION DE REQUÊTE

Trois méthodes d'expansion sont proposées et évaluées. Nous décrivons par la suite chacune d'elle en détail. L'algorithme d'expansion de requête est donné dans la Figure 5.2.

5.3.1.4. La phase de recherche

Dans cette phase, la recherche est faite dans l'index. Le processus de recherche se résume comme suivant: notre système cherche dans l'index les entités médicales étendues dans la requête de l'utilisateur, ces entités sont des séquences de termes (termes suivants) dans l'index. Le système recherche les RCs qui contiennent ces entités.

5.4. Les méthodes d'expansion de requête proposées

Dans cette section, nous présentons les approches d'expansion de requête basées sur l'ontologie "Onto_Orthopédique". En fait, trois méthodes d'expansion de requêtes sont proposées afin de i)- reformuler la requête initiale, ii)- pourvoir les comparer par rapport à la méthode simple. Dans la première méthode, les entités médicales sont étendues par leurs synonymes et hyponymes. Dans la seconde, nous avons appliqué l'extraction de relations sémantiques dans le contexte de la requête. Et finalement dans la troisième méthode, la requête est étendue par les entités médicales (synonymes, hyponymes et les entités en relations sémantiques); la requête étendue est reformulée sous une forme booléenne.

5.4.1. Expansion des entités médicales

Dans la première méthode d'expansion, les entités médicales présentes dans la requête sont étendues par leurs synonymes et leurs hyponymes qui se trouvent dans l'ontologie "Onto_Orthopédique". Cette méthode se compose de quatre étapes; la reconnaissance des entités médicales, l'extraction des synonymes et hyponymes, l'expansion de requête et la recherche dans l'index. Ces étapes sont illustrées dans la Figure 5.3.

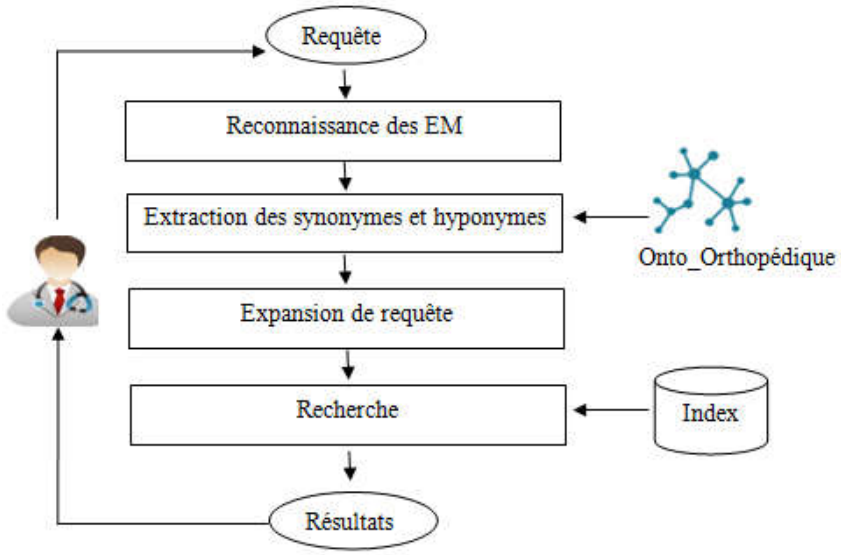

FIGURE 5.3 : EXPANSION DES ENTITÉS MÉDICALES PAR SYNONYMES ET HYPONYMES

L'étape1 : Reconnaissance des EM

Les médecins expriment leurs besoins d'information à l'aide d'une requête. Cette dernière peut contenir des entités médicales et pour cela c'est important de les reconnaître. Dans (Ghoulam et al., 2015b), une approche à base de règles est utilisée pour la reconnaissance des entités médicales. Dans cette étape, nous avons utilisé la même approche pour la reconnaissance des entités médicales.

L'étape 2 : Extraction des synonymes et des hyponymes

La sortie de l'étape précédente représente un ensemble de concepts et/ou des entités médicales. Dans cette étape, le système cherche des synonymes et des hyponymes pour chaque entité médicale reconnue à partir de l'ontologie "Onto_Orthopédique".

L'étape 3 : Expansion de requête

Nous avons ajouté les synonymes et les hyponymes des entités à la requête initiale. Par exemple, si un utilisateur écrit la requête suivante : "*traumatisme lombaire*", qui est un nom d'une maladie, cette requête peut être élargie pour inclure aussi: "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*".

L'étape 4 : La recherche dans l'index

Avec la reformulation de la requête initiale, la nouvelle requête est identifiée par le système qui renverra les rapports contenant ces entités médicales.

5.4.2. Expansion par extraction de relations sémantiques dans le contexte de la requête

La deuxième méthode comprend les mêmes étapes que la méthode précédente. La principale différence est l'ajout d'une étape supplémentaire qui est l'extraction de relations sémantiques dans le contexte de la requête, comme montre la Figure 5.4. Par conséquent, la reformulation de la requête est basée sur l'expansion de l'entité médicale et l'extraction des relations. Par exemple: si un utilisateur entre la requête "*traitement de traumatisme du rachi lombaire*"; la requête passe par la phase de reconnaissance des entités médicales pour reconnaître "*traumatisme du rachi lombaire*", comme une maladie.

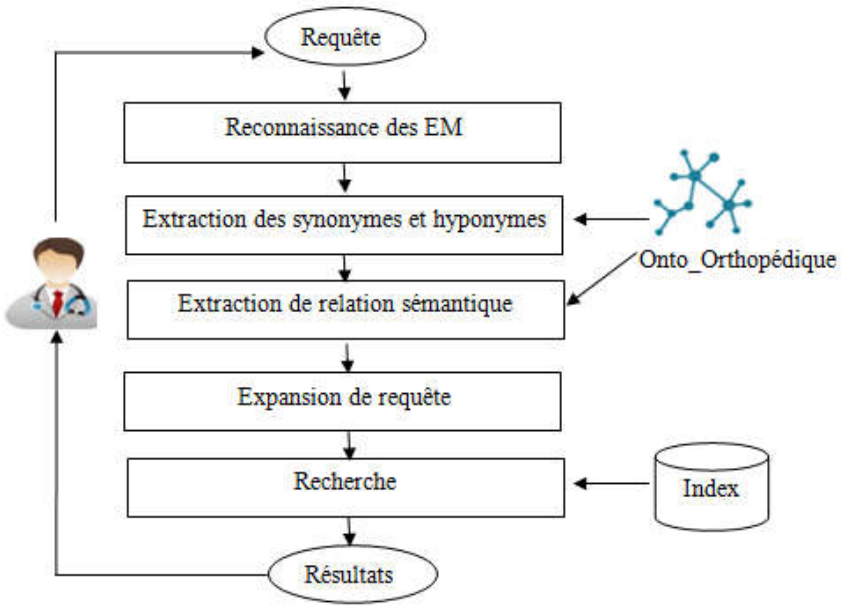

FIGURE 5.4 : EXPANSION PAR EXTRACTION DE RELATIONS SÉMANTIQUES DANS LE CONTEXTE DE LA REQUÊTE

Les synonymes et les descendants de la maladie "*traumatisme du rachi lombaire*" sont extraits de l'ontologie "Onto_Orthopédique", Ainsi la requête peut être étendue pour inclure: "*traumatisme du rachi lombaire*", "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*".

Dans la troisième phase, Le système extrait les relations entre le terme "traitement" et les entités médicales "*traumatisme du rachi lombaire*", "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*". Le terme "*traitement*" sera transformé en un nom de relation qui est «*traite*» comme suit: traite (maladie, X) où X sera trouvée à partir de la ressource externe. Dans notre exemple: traite ("*traumatisme du rachi*

lombaire", X); traite ("traumatisme lombaire", X), traite ("*rachi lombaire*", X), traite ("*fracture de L02*", X). A partir de la ressource externe; X = "*plaque vissée*", "*corset*", "*corset bivalve* ".

Enfin, la requête sera élargie pour inclure: "*plaque vissé*", "*corset*", "*corset bivalve*", "*traumatisme du rachi lombaire*", "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*".

5.4.3. Expansion par reformulation booléenne de la requête

Dans cette méthode, nous nous sommes intéressés à reformuler la requête de l'utilisateur sous forme d'une expression booléenne. Dans les deux méthodes précédentes, l'expansion de requête consistait juste à ajouter des entités médicales à partir de l'ontologie "Onto_Orthopédique" à la requête originale. La reformulation booléenne de la requête étendue suggère de passer à une expression booléenne à l'aide des opérateurs booléens.

La requête exemple que nous avons présentée dans la première méthode est: "*traumatisme lombaire*". Après expansion, la requête inclut les entités suivantes: "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*". Dans ce cas, la requête étendue sera reformulée sous forme booléenne sous forme de "*traumatisme lombaire*" OU "*rachi lombaire*" OU "*fracture de L01*" OU "*fracture de L02*".

La requête exemple présentée dans la deuxième méthode: "*traitement de traumatisme du rachi lombaire*"; a été étendue à : "*plaque vise*", "*corset*", "*corset bivalve*", "*traumatisme du rachi lombaire*", "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*". Cette requête étendue sera reformulée sous forme d'expression booléenne: [("*plaque vise*" OU "*corset*" OU "*corset bivalve*") ET ("*traumatisme du rachi lombaire*" OU "*traumatisme lombaire*" OU "*rachi lombaire*" OU "*fracture de L01*" OU "*fracture de L02*")].

En général, le connecteur 'ET' (AND en anglais) est utilisé pour relier des entités médicales qui de types différents. Contrairement, le connecteur 'OU' (OR en anglais) est utilisé pour lier les entités médicales de mêmes types.

5.4.4. Étude expérimentale et résultats obtenus

Dans cette section, nous décrivons l'ensemble de données et les métriques d'évaluation utilisés pour tester les approches et discuter les résultats obtenus. Pour l'évaluation des méthodes d'expansion et pour comparer les résultats obtenus, nous avons expérimenté

chaque méthode (T1, T2, T3, T4) par un ensemble de dix (10) requêtes. Les différents tests sont cités comme suit:

- Test1 (T1): recherche simple; recherche sans expansion ou recherche classique.
- Test2 (T2): recherche par expansion de requête; étendre la requête par utilisation des synonymes et des hyponymes.
- Test3 (T3): recherche par expansion de requête; étendre la requête par utilisation de l'extraction de relations dans le contexte de la requête.
- Test4 (T4): recherche par expansion de requête; étendre la requête par utilisation de l'extraction de relation et avec reformulation booléenne de la requête.

Corpus utilisé :

Nous avons recueillis 200 rapports cliniques orthopédiques français de l'hôpital Chlef (Algérie). Nous avons également utilisé un ensemble de dix (10) requêtes médicales contenant des entités médicales fournies par des médecins pour l'évaluation.

Nous avons utilisé des mesures standard de recherche d'information pour évaluer chaque méthode d'expansion. Ces mesures sont; le rappel, la précision, la F-mesure, la MAP et la R-précision, elles sont définies dans le chapitre 3 de l'état de l'art.

Le Tableau 5.2 ainsi que la Figure 5.5 montrent les moyennes du rappel, la précision, la F-mesure, la MAP et la R-précision obtenues par le système utilisant les différentes méthodes d'expansion.

Mesure↓ : Méthode→	T1	T2	T3	T4
Rappel	0.725	0.919	0.928	0.918
Précision	0.794	0.778	0.753	0.975
FM	0.698	0.791	0.774	0.944
MAP	0.882	0.907	0.895	0.976
R-précision	0.668	0.833	0.815	0.912

Tableau 5.2 : La moyenne du rappel, précision, F-mesure, MAP et R-précision obtenue en utilisant les différentes méthodes d'expansion

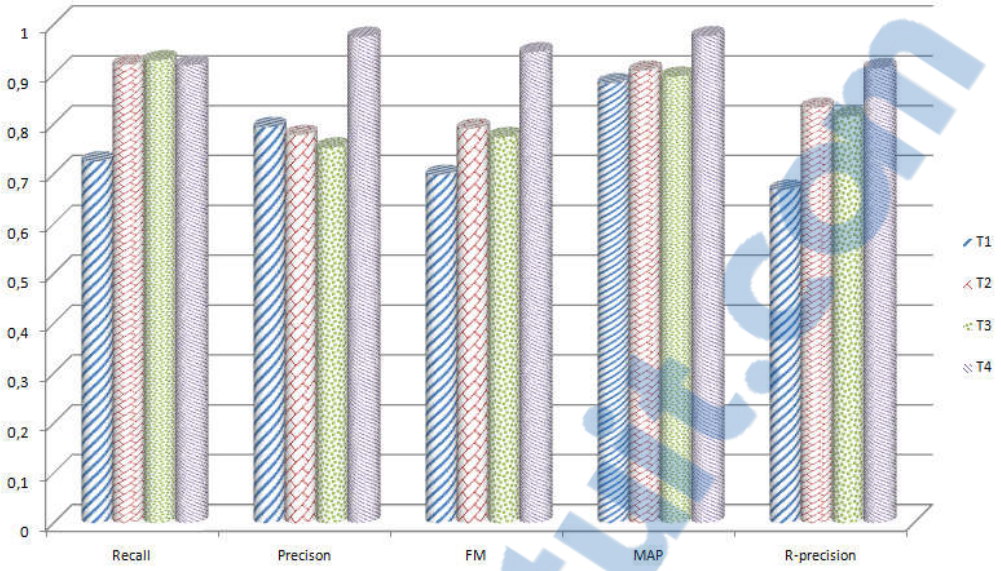

FIGURE 5.5 : PERFORMANCE OBTENUE POUR CHAQUE MÉTHODE EN UTILISANT LA MOYENNE DE RAPPEL, DE PRÉCISION, DE FM, DE MAP, ET R- PRÉCISION

5.4.5. Comparaison avec la méthode classique

La méthode classique ou méthode sans expansion (T1) est considérée ici comme une référence (Baseline) pour évaluer les performances des méthodes proposées. Pour pouvoir comparer les méthodes d'expansion de requête proposées, nous avons calculé le taux d'amélioration par rapport à la méthode classique (T1) comme illustré dans le Tableau 5.3.

Mesure↓ : Méthodes→	T1	T2	T3	T4
Rappel	0.725	0.919 (+26.76%)	0.928 (+28%)	0.918 (+26.62%)
Précision	0.794	0.778 (-2.01%)	0.753 (-5.16)	0.975 (+22.79%)
FM	0.698	0.791 (+13.32%)	0.774 (+10.88)	0.944 (+35.24%)
MAP	0.882	0.907 (+2.83%)	0.895 (+1.47)	0.976 (+10.65%)
R-précision	0.668	0.833 (+24.70)	0.815 (+22%)	0.912 (+35.52%)

Tableau 5.3 : Le taux d'amélioration par rapport à la méthode classique

5.4.6. Discussion des résultats

En ce qui concerne les résultats obtenus dans le Tableau 5.3; on constate une amélioration du rappel dans toutes les méthodes (T2, T3, T4) par rapport à la méthode classique (T1), ainsi qu'une amélioration notable de la MAP et de la R-précision.

L'augmentation du rappel signifie qu'il y a une augmentation du nombre de rapports pertinents recherchés par chaque méthode. En d'autres termes, nous concluons qu'à chaque fois la requête contient plus d'entités médicales plus le nombre de documents pertinents

augmente et donc le rappel s'améliore, et c'est le cas de la méthode d'expansion par extraction de relation (T3) où le rappel est amélioré de +28% par rapport à T1. Ainsi, l'utilisation d'une ontologie de domaine médicale dans l'expansion de requête améliore le rappel dans la recherche des informations médicales.

En revanche, la précision de la méthode classique (T1) était meilleure que celle de T2 et T3, ce qui signifie qu'il y'a beaucoup de documents non pertinents récupérés par le système; dans la méthode T2 la précision diminue de -2.01% et dans la méthode T3, elle diminue de -5.16% par rapport à la méthode classique (T1). Ceci nous a permis de conclure que les méthodes d'expansion de requête T2 et T3 que nous avons proposées n'améliorent pas la précision dans la recherche d'information médicale.

Contrairement à la méthode T4, elle montre une amélioration dans la précision avec un taux de +22.79% par rapport à la méthode classique (T1). Nous concluons, qu'avec une bonne reformulation de la requête étendue, la précision peut être améliorée.

En ce qui concerne la MAP et la R-précision, on peut observer qu'à partir de la Figure 5.5, la méthode T4 surpasse la méthode classique (T1) et aussi les deux autres méthodes T2 et T3.

Le Tableau 5.4 présente les résultats d'interpolation rappel/précision obtenus par les trois stratégies (T2, T3, T4) et même la méthode classique (T1). La précision/rappel sont présentés en montrant la précision moyenne interpolée à onze (11) niveaux de rappel standard.

Précision de→ Rappel↓	T1	T2	T3	T4
0.0	0.933	1.000	1.000	1.000
0.1	0.967	1.000	1.000	1.000
0.2	0.908	0.945	1.000	0.978
0.3	0.920	0.922	0.967	0.980
0.4	0.891	0.907	0.973	0.970
0.5	0.893	0.910	0.932	0.974
0.6	0.866	0.898	0.877	0.977
0.7	0.870	0.904	0.872	0.977
0.8	0.872	0.884	0.879	0.977
0.9	0.852	0.880	0.853	0.977
1.0	0.853	0.849	0.800	0.977

Tableau 5.4 : L'interpolation précision/rappel moyenne obtenue pour différentes méthodes

La Figure 5.6 montre la corrélation de précision/rappel à 11 points. Nous pouvons comparer les méthodes et voir l'importance de la méthode d'expansion de requête en se basant sur la reformulation booléenne de la requête (T4).

Les résultats expérimentaux de la Figure 5.6 montrent que toutes les méthodes d'expansion de requête réalisent une amélioration par rapport à la méthode classique T1. Pour la méthode T2 ; qui utilise l'expansion par des synonymes/hyponymes; elle améliore la recherche à huit (8) points de rappel (0,0 à 0,2 et 0,4 à 0,7 et 0,9), et la performance se dégrade à trois points (0,3, 0,8 et 1,0). De même, la méthode d'expansion T3 basée sur les synonymes /hyponymes et l'extraction de relation dans le contexte de la requête améliore la recherche à sept points (de 0,0 à 0,6) et la performance se dégrade à quatre points (0,7 à 1,0). Contrairement à la méthode d'expansion T4, comme elle est montrée dans la courbe de rappel / précision, elle améliore la recherche à tous les onze points de rappel.

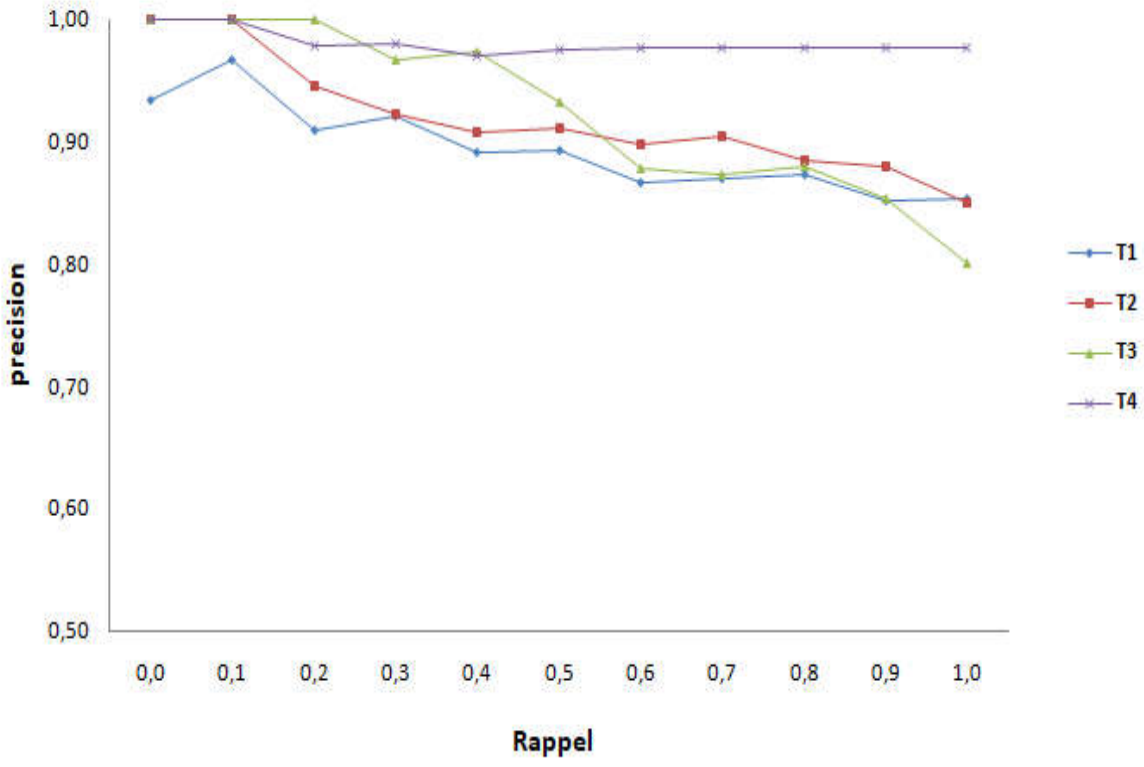

FIGURE 5.6 : COURBE DE RAPPEL / PRÉCISION

5.5. Architecture distribuée à large échelle proposée

Aujourd'hui, la production d'énormes quantités de données est sous le nom de "Big Data". Dans le domaine médical, chaque hôpital a ses propres RCs rédigés par des

médecins spécialistes. Ces hôpitaux ont du mal à trouver des moyens de rendre ces données utiles. Cependant, ce n'est pas une tâche facile. La quantité de données produites rend très difficile le stockage, la gestion, l'analyse et l'utilisation de ces données.

La base de tout système Big Data est un système de stockage de données distribué à grande échelle (Rashmi, 2016).

Généralement le "Big Data" est un nouveau domaine de recherche. Il est considéré comme un avantage pour plusieurs entreprises et organisations, car il les aide à prendre de meilleurs décisions dans les brefs délais (Sugam et al.,, 2014). Néanmoins, le Big Data est confrontée à des défis, nous listons dans ce travail de thèse quelques principaux défis du Big Data tel que: le volume de données, la vélocité, la variété et la sécurité et confidentialité.

1- Volume de données: l'un des principaux défis de Big Data à relever est de parvenir à stocker et à analyser toutes les informations. La quantité de données produites dans le monde entier est en croissance, elle augmente de téraoctets à même pétaoctets (Sugam et al.,, 2014). Ce volume de données rend le stockage un défi pour tout le monde et dans tous les domaines. Il existe de nombreuses solutions de stockage fiables telles que SAN (Storage Area Networks), stockage sur le Cloud (Cloud public, Cloud privé, Cloud hybride) (Li et al., 2016).

2- Vélocité (vitesse): les données doivent être collectées et traitées en temps réel. Il ne suffit pas simplement de stocker les Big Data, mais aussi de les générer et de les traiter rapidement pour atteindre les objectifs souhaités.

3- Variété: Les données peuvent prendre des formes très variées et très hétérogènes (par exemple: texte, image, vidéo).

5- Sécurité et confidentialité: la production de données fiables augmente le besoin de sécurité et de confidentialité. Il est donc essentiel que les spécialistes des données examinent ce besoin et traitent les données de manière à ne pas perturber la vie privée.

Le développement de divers outils d'analyse de Big Data a grandement facilité le traitement des données. Généralement les systèmes Big Data développés repose sur la solution de distribution, cette dernière veut dire prendre les données, les découper ou les partitionner sur une série de machine appelée "Cluster". La distribution va permettre d'augmenter le volume et d'avoir une bonne vitesse.

5.5.1. Exemples de systèmes de recherche d'information à large échelle

Système	Méthodes et Matériels utilisés	Avantages	Inconvénients
Lin et al., (2015)	Hadoop Cluster: -HDFS: pour la sauvegarde des rapports médicaux. - MapReduce: pour la création de l'index Lucene: -Cluster de recherche distribué. Maladie-symptôme treillis	- Framework basé sur les Cloud pour mettre en œuvre un service de diagnostic à domicile. - Indexation parallèle et cluster de recherche distribué.	-Modèle de sécurité suppose que tous les patients sont traités de la même manière. (problème de sécurité). -Indexation offline manque d'indexation en temps réel.
Sobhy et al., (2012)	Hadoop Cluster: Pour la scalabilité. HIPAA privacy + règles de sécurités.	Un système "MedCloud" scalable et sécurisé.	-Recherche non sémantique. -Recherche et indexation off line.

Tableau 5.5 : Exemples de systèmes de recherche d'information à large échelle

Ces dernières années, la quantité des rapports cliniques sauvegardés dans les hôpitaux est en augmentation dans tous les pays. Ce problème est connu sous le nom de "medical big data". Plusieurs systèmes de recherche à large échelle ont été proposés pour accéder rapidement et efficacement à des informations à partir des RCs des patients.

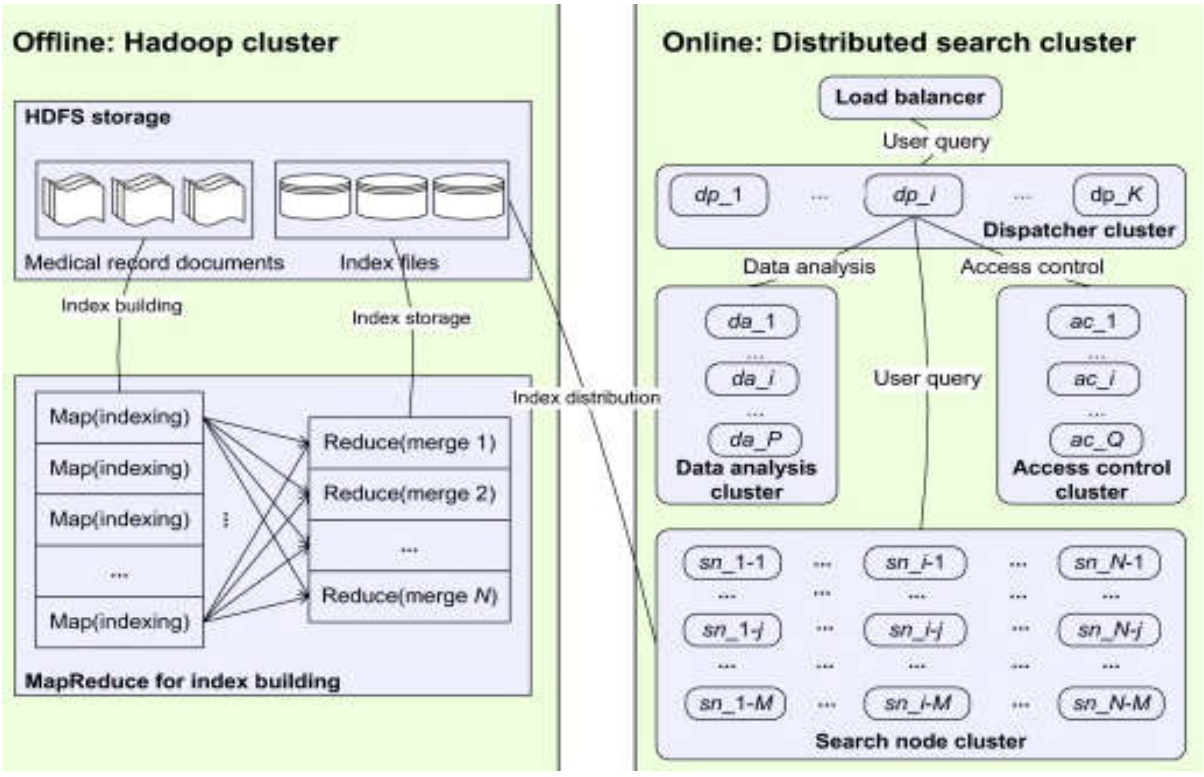

FIGURE 5.7 : ARCHITECTURE DU SYSTÈME (LIN ET AL., 2015)

Lin et al., (2015) ont proposé un Framework basé sur un Cloud informatique pour mettre en œuvre un service de diagnostic à domicile comme montre la Figure 5.7. Ce Cloud est utilisé pour le stockage et le repérage des RCs à large échelle. Ils ont utilisé Hadoop Cluster pour le stockage de données et pour la construction de l'index. Un Cluster de recherche distribué basé sur Lucene est conçu, de plus un réseau sémantique maladie-symptôme est conçu à partir des rapports cliniques pour aider les utilisateurs à déterminer quel type de maladie ils sont concernés. Les limitations de leur système réside dans : (i)- l'indexation ne se traite pas en temps réel, (ii)- absence de sécurité; tous les patients sont traités de la même manière. (iii)- dépendance totale au réseau maladie-symptôme pour le diagnostic à domicile.

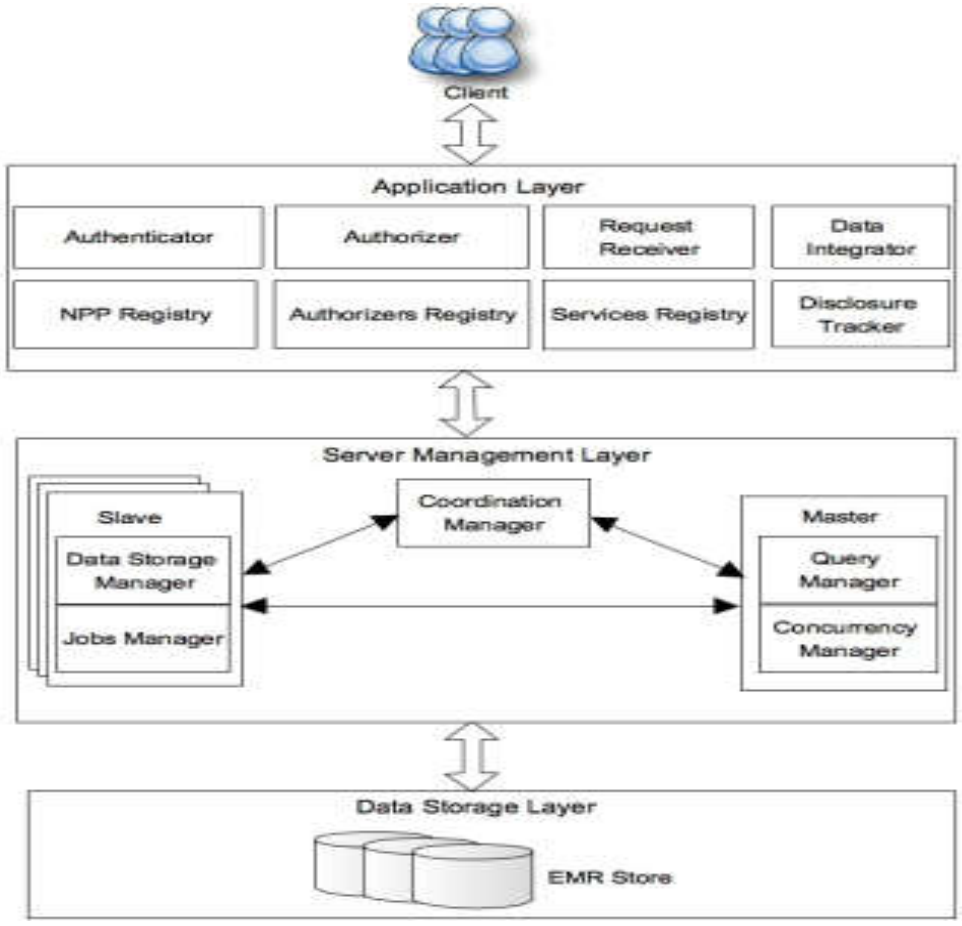

FIGURE 5.8 : ARCHITECTURE DU SYSTÈME SOBHY ET AL., (2012)

Sobhy et al., (2012) comme montre la Figure 5.8 ont proposé un système nommé "MedCloud" basé sur un Cloud informatique. Ce système utilise Hadoop Cluster pour le stockage de données et les préoccupations de confidentialité et de sécurité sont basées sur HIPAA (Health Insurance Portability and Accountability Act). Les limites de ce système

résident dans : (i)- ce système est basé sur une recherche traditionnelle simple, (ii)- ils n'ont pas traité l'indexation en temps réel.

5.5.2. Proposition d'un système de recherche sémantique à large échelle

5.5.2.1. Vue d'ensemble

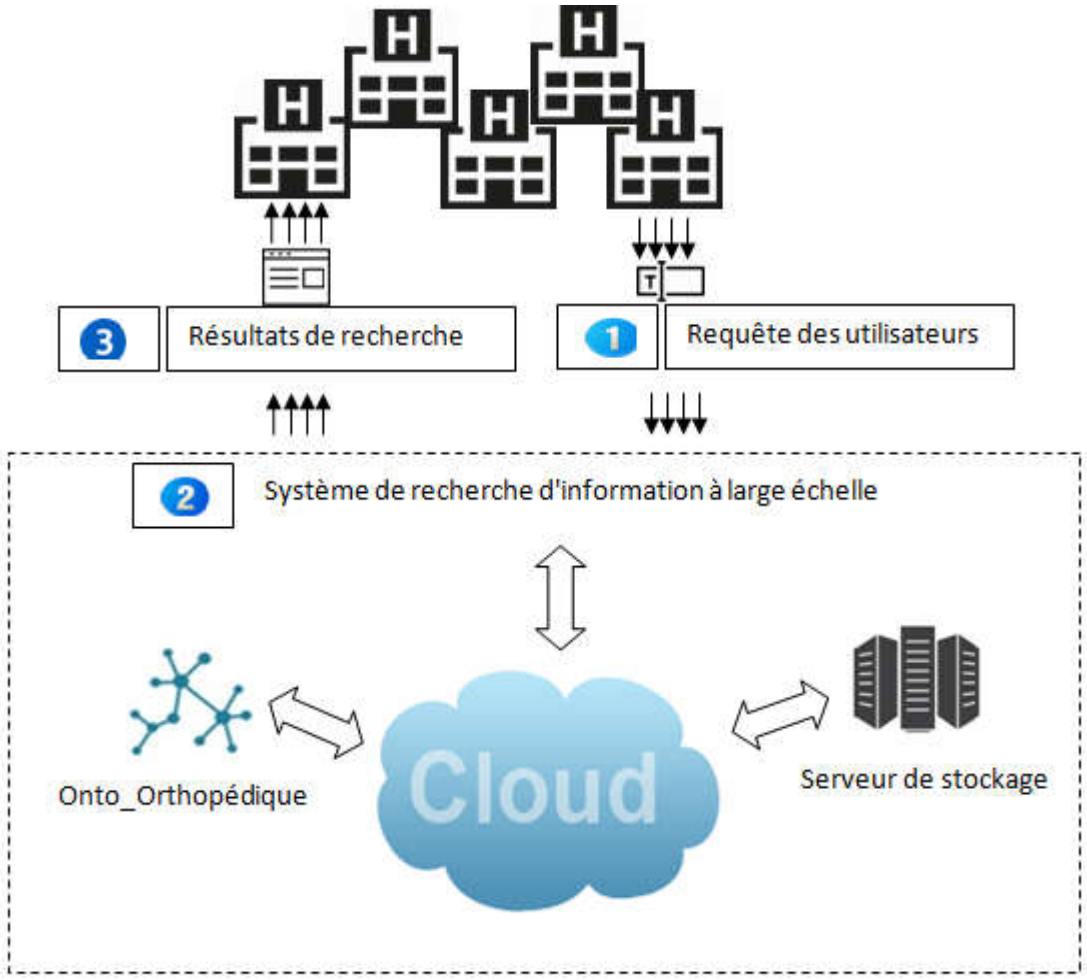

FIGURE 5.9 : VUE D'ENSEMBLE DU SYSTÈME DE RECHERCHE À LARGE ÉCHELLE

La Figure 5.9 montre une vue d'ensemble du système proposé en trois éléments essentiels:

(1) Les utilisateurs (le personnel médical) expriment leurs besoins informationnels sur n'importe quel navigateur.

(2) Le système de recherche d'information (qui sera hébergé dans les locaux du ministère de la santé ou dans un site à part) traite les requêtes du personnel médical et interroge l'ontologie "*Onto_Orthopédique*" pour l'expansion de la requête.

(3) Le système de recherche retourne (affichage des résultats) à partir d'un ensemble de documents, ceux dont le contenu est pertinent à la requête de l'utilisateur en utilisant Terrier[39].

Dans cette solution nous proposons un système de recherche médical à large échelle qui se diffère des autres systèmes de recherche de (Lin et al., 2015) et (Sobhy et al., 2012) dans les points suivants:

- Une recherche sémantique basée sur une ontologie médicale pour l'expansion de requête.

- Une recherche et indexation en ligne.

- Modèle de sécurité et confidentialité, nous proposons un Cloud privé qui permet un contrôle plus élevé des données où les utilisateurs ne sont pas traités de la même manière.

5.5.2.2. Les outils utilisés dans ce système

Nous décrivons dans cette section les outils utilisés dans ce travail pour construire le système. Nous proposons les outils suivants: *Crawlers, HDFS, MapReduce* et *Terrier*.

1- **Les crawlers (ou Spiders):** son rôle est de récolter les nouveaux RCs existant dans les disques durs des hôpitaux et les sauvegarder dans des clusters (Serveurs de stockage des RCs). Pour faire face au problème de l'indexation en temps réel nous proposons d'utiliser des Crawlers en ligne. Ces derniers ont pour mission de parcourir de façon autonome et automatique les différents Hôpitaux à la recherche des RCs ou d'éventuelles mises à jour.

2- Hadoop cluster: HDFS et MapReduce Hadoop est un Framework basé sur le Cloud, c'est une technologie de cloud computing distribué la plus populaire et la plus significative dans le domaine du Big Data. Haddop est utilisé pour l'exécution des applications sur un grand cluster. Hadoop est constitué de deux composantes principales: Hadoop Distributed File System (HDFS) et le Framework MapReduce. Le HDFS sauvegarde des fichiers volumineux dans un cluster. MapReduce est un programme qui permet le traitement parallèle d'un grand nombre de données et donc l'accélération d'indexation d'un nombre volumineux de données.

Dans notre proposition du système à large échelle, HDFS est adopté pour le stockage d'un grand nombre de RCs à partir des Hôpitaux et le MapReduce Framework est utilisé

[39] http://Terrier.org/

pour le traitement parallèle des RCs pour créer les fichiers d'index ainsi que pour l'extraction des EM et RM.

3- Cluster de recherche distribuée (Terrier) Dans ce système de recherche à large échelle nous proposons d'expérimenter avec Terrier. Dans ce système de recherche à large échelle nous proposons d'expérimenter avec Terrier. Dans (Middleto & Baeza-Yates, 2007) les auteurs ont réalisé une comparaison entre un ensemble de systèmes open source de RI. Le Tableau 5.6 montre les résultats de cette comparaison selon les critères suivants :

- Le temps d'indexation et la taille d'index pour une collection de 2.7 GB,
- La possibilité d'une indexation à large échelle pour une collection de 10 GB et
- Le langage de programmation utilisé.

Les trois systèmes de recherche étaient capable d'indexer la collection de 2.7 GB dans un temps qui diffère d'un système à un autre. Lucene a nécessité une heure pour terminer l'indexation or alors que Terrier n'a nécessité que 40 minutes et 12 secondes pour indexer les documents. Zettair était le meilleur des trois systèmes, l'indexation a durée seulement 4 minutes et 44 secondes. Un deuxième test a été fait ; il consistait à comparer leur capacité d'indexer une large collection de 10 GB de documents. Dans cette expérience, les auteurs ont constaté que Zettair a eu le meilleur résultat par observation du de la précision à différentes valeurs des premiers documents retournés. Dans le Tableau la précision pour les cinq premiers documents retournés est donnée (p@5) uniquement pour Terrier et Zettair qui étaient capables d'indexer toute la collection de 10GB sans dégradation considérable contrairement à Lucene dont le temps d'indexation a énormément augmenté ainsi il a été exclu de la comparaison pour cette collection (de 10GB).

Système de recherche	Langage	Temps d'indexation	Taille d'index	Large scale (10GB)	P@5
Lucene	Java	1:01:25	26%	Pas possible	-
Terrier	Java	0:40:12	52%	C'est Possible	0.2800
Zettair	C/C++	0:04:44	33%	C'est Possible	0.3240

Tableau 5.6 : Comparaison entre les systèmes de RI open source pour une collection de 2.7GB

Terrier est une plate forme académique de recherche d'information open source basé sur le langage de programmation Java. Le choix de cette plate forme est dû à deux points essentiels :

(i) Terrier offre un cadre pour l'évaluation des résultats de recherche des documents pour différentes application (Santos et al., 2011).

(ii) Terrier peut traiter les grandes collections de documents (Santos et al., 2011).

5.5.2.3. Description du système de recherche à large échelle proposé

Le système de recherche à large échelle que nous proposons est montré dans la Figure 5.10, cette dernière comprend les étapes suivantes:

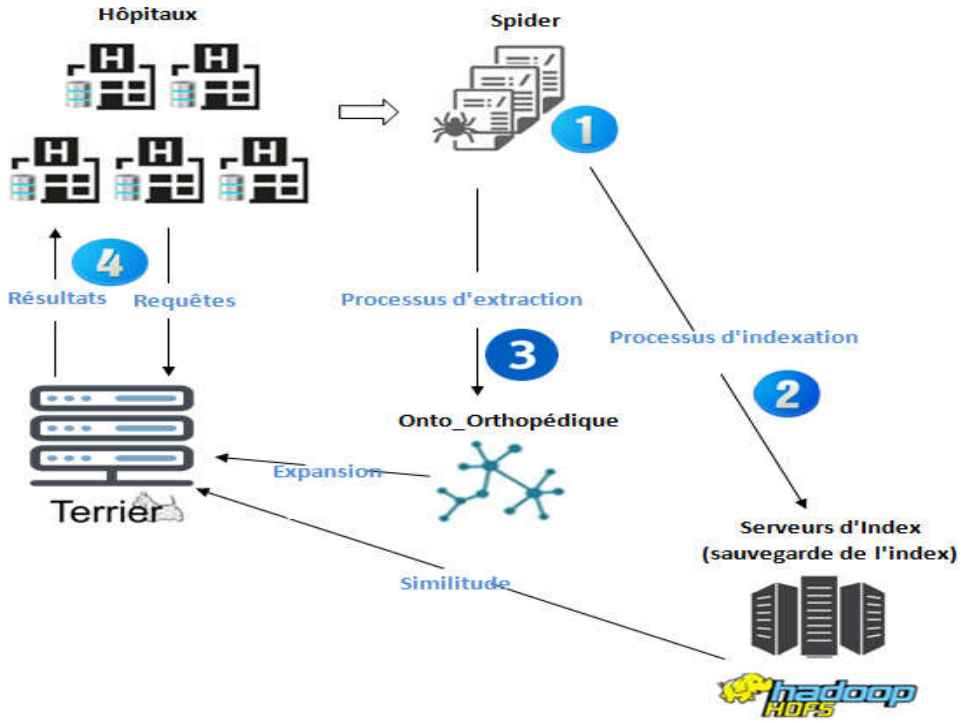

FIGURE 5.10 : ARCHITECTURE GLOBALE DE NOTRE SYSTÈME DE RECHERCHE D'INFORMATION À LARGE ÉCHELLE

1. Les nouveaux RC enregistrés au niveau des hôpitaux sont premièrement anonymisés par un système d'anonymisation des RC (nom de malade, nom de médecin, nom de l'organisation et les numéros de téléphone). Ensuite, ils sont explorés constamment par les spiders ou les crawler. Ces derniers fournissent ces RCs au module d'indexation pour les indexer et les stocker dans des serveurs de stockage en utilisant un système de fichier distribué HDFS.

2. Un processus d'indexation par utilisation de MapReduce (API qui existe dans le système Terrier[40]) est appliqué sur l'ensemble des RC enregistrés dans les serveurs, ce processus génère un index qui est à son tour enregistré dans des serveurs de stockage.
3. Un processus d'extraction des entités médicales et de relations est appliqué aussi sur les RC. Ces informations sont enregistrées dans l'ontologie "*Onto_Orthopédique*", après validation des médecins spécialistes.
4. Des requêtes de recherche sont lancée par les utilisateurs sur Terrier. Terrier retrouve les documents correspondants en interrogeant l'index de recherche. Les résultats de la recherche sont retournés aux utilisateurs.

5.5.2.4. Processus d'indexation et d'extraction

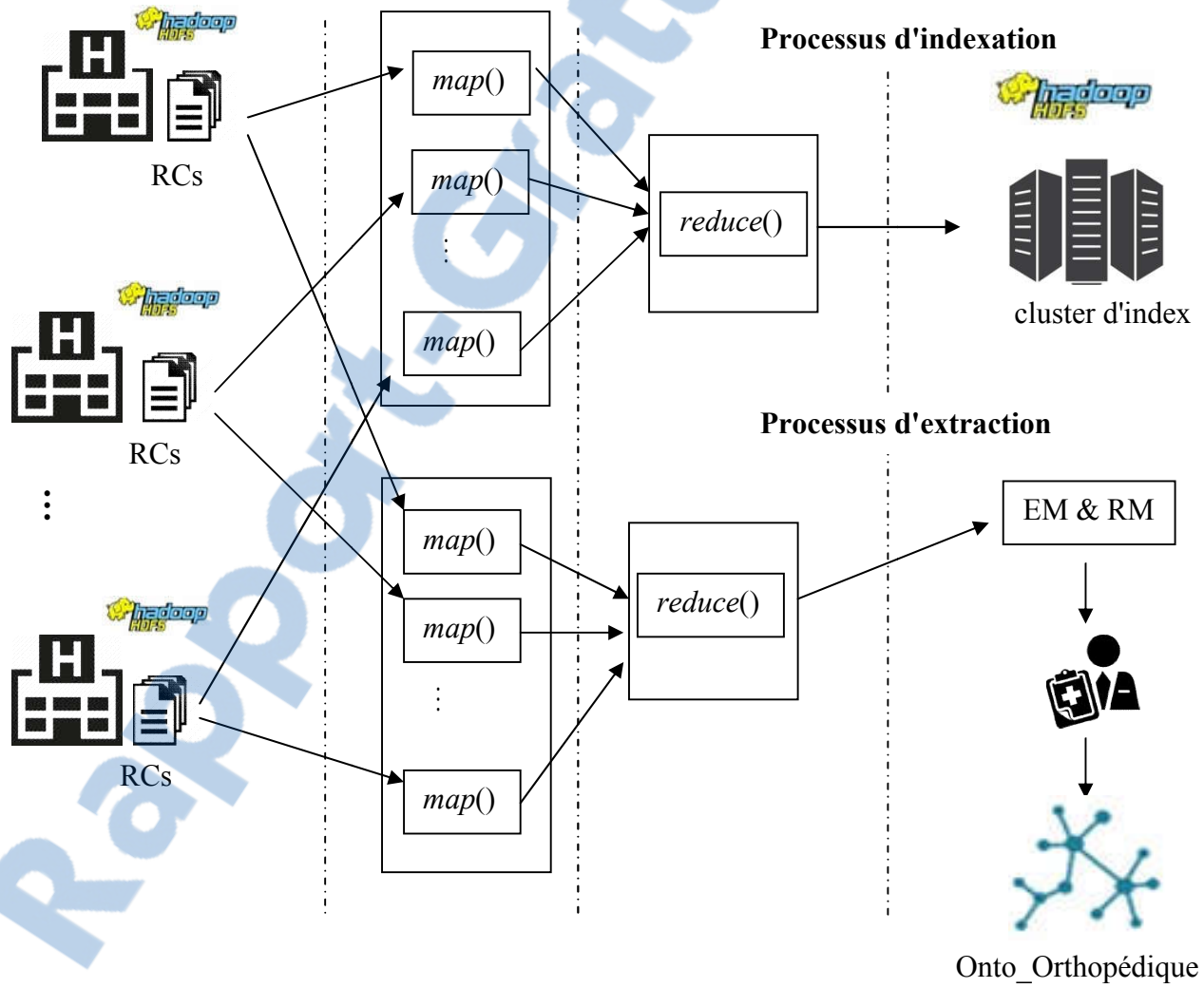

FIGURE 5.11 : PROCESSUS D'INDEXATION ET D'EXTRACTION

La Figure 5.11 montre les deux processus d'indexation et d'extraction, après avoir enregistré et collecté les RCs dans des serveurs de stockage (HDFS), le processus

[40] La version open source de Terrier contient un indexeur distribué basé sur Hadoop.

d'indexation est lancé en utilisant un traitement de données parallèle basé sur le MapReduce. Pendant la phase d'indexation, les fonctions de MapReduce sont appliquées pour extraire en parallèle tous les mots du RCs pour construire l'index et les enregistrer dans un cluster (HDFS). Nous proposons d'appliquer aux nouveaux RC récoltés par les Crawlers l'extraction des EM et RM. Les informations extraites sont validées par un expert et enregistrées dans "*Onto_Orthopédique*".

5.5.2.5. Processus de recherche

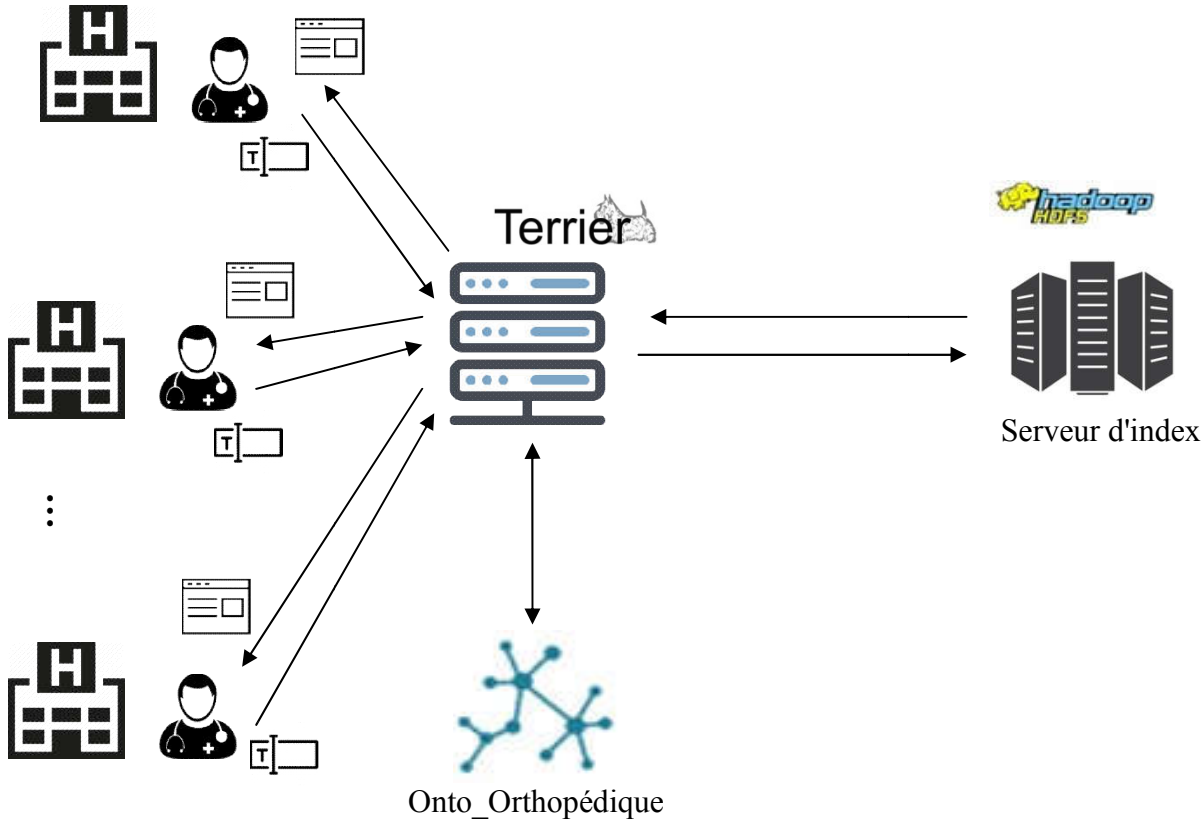

FIGURE 5.12 : PROCESSUS DE RECHERCHE DISTRIBUÉ

Le processus de recherche est montré dans la Figure 5.12. La recherche distribué consiste en un ensemble de phases. (i) les utilisateurs du système de recherche expriment leurs besoins informationnels à l'aide d'une requête via une interface. (ii) extraction des entités médicales contenues dans la requête de l'utilisateur. (iii) expansion sémantique de la requête. (vi) à la fin, une recherche dans l'index est effectué, et une liste des RCs jugés pertinents et leurs scores respectifs sont envoyés pour chaque utilisateur et ceci selon leur requête.

5.6. Conclusion

Dans ce chapitre nous avons proposé une méthode d'expansion de requête basée sur une ressource sémantique. Cette dernière est une ontologie médicale du domaine orthopédique. Elle est construite en se basant sur l'extraction des informations de type entités médicales et relations sémantiques qui les relient à partir d'un ensemble de rapports cliniques. Comparé à un système de recherche classique, les résultats obtenus des expérimentations de plusieurs techniques de recherche sémantique sont encourageants. Nous avons conclu que l'utilisation d'une ontologie de domaine construite à base d'entités et relations médicales extraites des RCs peut améliorer la recherche dans le domaine médical ce qui nous permet de conclure que l'extraction d'information a un impact positif sur la recherche médicale.

L'inconvénient majeur de l'extraction d'information pour la recherche est le traitement d'un grand nombre de RCs et de requêtes dans un système centralisé. Pour cela, nous avons proposé une architecture à large échelle du système de recherche en se basant sur l'extraction des informations. L'utilisation du Cloud Computing est considérée comme une solution qui permet la scalabilité et le parallélisme pour traiter un grand nombre de RCs et de requêtes.

CONCLUSION ET PERSPECTIVES

Le sujet de cette thèse traite la recherche d'information par utilisation de l'extraction d'information dans le domaine médical.

Nous avons développé dans cette thèse un système de recherche centralisé. Pour ce faire, nous avons commencé par deux tâches essentielles ; la première est la tâche de reconnaissance des entités médicales, la deuxième tâche est l'extraction de relations sémantiques à partir des rapports cliniques.

Dans la reconnaissance des entités médicales, nous avons développé un système de reconnaissance utilisant des rapports cliniques écrits en français. Comme les méthodes d'apprentissages restent encore un défi pour la langue française qui souffre du manque de corpus annoté surtout dans le domaine médical, nous avons proposé d'utiliser une méthode à base de règles ; elle se base sur des grammaires locales et des Gazetters pour la définition de règles. Nous avons considéré cinq classes pour catégoriser les entités médicales: maladie, traitement, examen, médicament et symptôme. Les résultats expérimentaux montrent que l'approche proposée a atteint une F-mesure moyenne de 90.06%.

Dans le développement du système d'extraction de relations, nous avons considéré les entités extraites dans la phase une pour pouvoir extraire les relations qui existent entre ces différentes entités. Quatre type de relations ont été étudiées : la relation « *traite »* qui relie les deux entités « *maladie » et «* traitement », la relation « *detecte* » qui relie l'entité « *maladie »* à l'entité *« examen »*, la relation « *soigne » qui relie* les deux entités « *maladie* » et « *médicament* », et enfin la relation « *signe* » entre « *maladie* » et « *symptôme* ». Ce processus d'extraction se base sur la définition d'un ensemble de patrons linguistiques représentés par des grammaires locales. Le système a atteint une précision moyenne de 47.57% et un rappel moyen de 38.55%. Le système n'a pas réussi à extraire toutes les relations en raison du nombre insuffisant de patrons linguistiques en particulier entre les entités de type maladie et symptôme. En outre, et en raison de la structure des RCs, les médecins utilisent plus d'entités médicales et moins de relations. Ce qui explique le faible rappel.

Après avoir extrait les entités médicales et les relations, ces dernières sont validées par le médecin orthopédique. Nous avons fait appel à la méthodologie "ARCHONTE" pour la construction de l'ontologie orthopédique. Avec l'aide du médecin et des sites web médicaux, nous avons pu construire une ontologie hiérarchique (ontologie différentielle). Nous avons utilisé l'open source Protégé 4.3 pour l'opérationnalisation de cette ontologie.

L'ontologie orthopédique "*Onto_Orthopédique*" construite est intégrée dans un système de recherche d'information pour être utilisée dans l'expansion de requêtes. Trois approches d'expansion de requête sont proposées:

1. Expansion des entités médicales par synonymes et descendants (hyponymes).
2. Expansion par extraction de relation sémantique dans le contexte de la requête.
3. Expansion par extraction de relation et reformulation booléenne de la requête.

Dans la première méthode d'expansion de requêtes, la requête est élargie avec les synonymes et descendants des entités médicales existantes dans la requête. Le système de recherche donne 79% en termes de F-mesure et 91% en termes de MAP.

Dans la deuxième méthode d'expansion, La requête de l'utilisateur s'enrichie avec utilisation de l'extraction de relations dans le contexte de la requête. Le système a donné une F-mesure de 77% et un MAP de 89%.

Dans la troisième méthode d'expansion, nous utilisons pour étendre la requête de l'utilisateur l'extraction de relations et la reformulation booléenne de la requête. Le système a donné une F-mesure de 94% et une MAP de 97%.

Nous avons comparé les résultats des méthodes d'expansion par rapport à la recherche simple (sans expansion). Nous concluons qu'à chaque fois la requête contient plus d'entités médicales plus le nombre de documents pertinents retrouvés augmente ce qui améliore le rappel. Ainsi, l'utilisation des ressources sémantiques dans l'expansion de requêtes améliore le rappel dans la recherche des informations médicales. Nous avons vu que la troisième méthode d'expansion fournit une amélioration dans la précision par rapport à la recherche simple et même par rapport aux premières et deuxièmes méthodes d'expansion. Ce résultat, nous a permis de conclure qu'avec une bonne reformulation de la requête étendue la précision de recherche peut être améliorée.

Par la suite, nous avons proposé l'architecture d'un système de recherche médicale à large échelle. Nous avons proposé d'utiliser le Cloud Computing qui promet un faible coût, une

évolutivité élevée, une disponibilité et une capacité de récupération en cas de sinistre. Il peut constituer une solution naturelle à certains problèmes rencontrés lors du stockage et de l'analyse des dossiers médicaux des patients.

Nous proposons comme perspectives dans ce travail de thèse les éléments suivants:

1. **Améliorer l'extraction d'information:**

   - *Pour la reconnaissance des entités médicales:* nous proposons l'utilisation d'une méthode d'apprentissage automatique ou d'une méthode hybride.
   - *Pour l'extraction de relations:* résoudre le problème d'absence de patrons dans les RCs, avec les deux solutions suivantes:

     a) développer une interface interactive qui permet aux médecins d'identifier la relation sémantique entre les entités lors d'absence de patrons.
     b) valider automatiquement l'existence de relations entre deux entités médicales en cas d'absence de médecins, ceci sera peut être possible soit à partir d'un ensemble d'articles scientifiques ou soit par une requête faite dans le moteur de recherche CISMeF[41].

2. **L'enrichissement de l'ontologie orthopédique:**

On peut enrichir l'ontologie orthopédique construite par les moyens suivants:

- développer une interface dédiée aux médecins pour enrichir cette ontologie; par ajout et validation de nouvelles entités médicales et de nouvelles relations comme les relations de synonymie et d'hyponymie.
- enrichir l'ontologie en utilisant des ressources médicales existantes telles que le MeSH et la SNOMED.

3. **Recherche d'information sémantique à large échelle**

Cette étude peut aussi être poursuivie par la considération à large échelle par implémentation de l'architecture proposée et l'évaluer sur un cas réel on prend en considération les fonctionnalités suivantes:

- Définir le privilège des médecins utilisateur du système par définition des fonctionnalités; (à titre d'exemple: pour les médecins professeur, les médecins

[41] http://www.chu-rouen.fr/cismef/

résidents, les CHU, les médecins des hôpitaux, les médecins de dispensaires; définir les fonctionnalités pour leurs permettre de visualiser les RCs ou non).

- Enregistrer les requêtes des médecins et leurs réponses (query logs), c'est le journal de requêtes pour l'utiliser dans d'autres applications.

Bibliographie

A

Abbache, A., Barigou, F., Belkredim, FZ., Belalem, G. (2014) .The use of Arabic WordNet in Arabic Information Retrieval. International Journal of Information Retrieval Research (IJIRR) 4 (3):54-65.

Abbache, A., Meziane, F., Belalem, G., Belkredim, FZ. (2016) .Arabic query expansion using WordNet and Association Rules. International Journal of Intelligent Information Technologies (IJIIT) 3 (12): 51-64.

Afzal, Z., Saber ,A. Akhondi, Herman van Haagen., Erik Van Mulligen., Jan A. Kors.(2015). Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. CLEF 2015 Online Working Notes. CEUR-WS.

Arguello, J., Elsas, J. L., Callan, J., & Carbonell, J. G. (2008). Document representation and query expansion models for blog recommendation. In Proceedings of the 2nd International Conference on Weblogs and Social Media. AAAI Press, p.10–18.

Audeh, B. (2014). Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web. Autre. Ecole Nationale Supérieure des Mines de Saint-Etienne, thèse de doctorat.

Audeh, B., Beaune, P., Beigbeder, M. (2014) .L'utilisation des entités nommées pour l'expansion sémantique des requêtes Web. Conference: EGC, Rennes, France.

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), p.169–176.

B

Baccini A, Sébastien Déjean, D. Kompaoré, Josiane Mothe. (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. Revue des Sciences et Technologies

de l'Information - Série TSI : Technique et Science Informatiques, Lavoisier, 2010, 29(3):289-308.

Bachimont, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In R. TEULIER, J. CHARLET & P. TCHOUNIKINE, Coordinateurs, Ingénierie des connaissances, chapitre 19. Paris : L'Harmattan. Article réédité en 2005 dans le cédérom associé au livre.

Bachimont, B., Isaac, A., Troncyr, R. (2002). Semantic commitment for designing ontologies : A proposal. In A. GOMEZ-PÉREZ & V. BENJAMINS, Coordinateurs, 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), volume LNAI 2473 of Lecture Notes in Artificial Intelligence, p.114–121.

Baeza-Yates. R & Ribiero-Neto, B. (1999). Modern Information Retrieval, Textbook Paperback, May 1999, ISBN: 020139829X.

Baneyx, A. (2007). Construire une ontologie de la Pneumologie Aspects théoriques, modèles et expérimentations. Autre [cs.OH]. Université Pierre et Marie Curie - Paris VI, thèse de doctorat.

Barigou, F., Beldjilali, B., Atmani, B. (2011). MedIX : A Named Entity Extraction Tool from patient clinical reports, International Conference on Communication, Computing and Control Application, Hammamet, Tunisia, March 3-5, p.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. (2011). 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, February 20-26, Tokyo, Japan, p.139-150.

Ben abacha, A., & Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of Biomedical Semantics, 2(5).

Ben Abacha, A., & Zweigenbaum, P. (2011). Medical entity recognition: A comparison of Semantic and Statistical Methods, In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACLHLT, June 23-24, Portland, Oregon, USA, p.56–64.

Ben Abacha, A. (2012). Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS. Université Paris Sud - Paris XI, thèse de doctorat.

Berrazega, I. (2012). Temporal Information Processing: A Survey. International Journal on Naturel Language Computing 1(2):1-14.

Bhatnagar, P., & Pareek, N. (2014). Novel Approach for Query Expansion Using Genetic Algorithm. International Journal of Information and Computation Technology 4(3):239-246.

Bhogal, J., Mcfarlane, A., Smith, P. (2007). A review of ontology based query expansion. Journal information processing and Management: an international journal 43(4): 866-886.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. Paper presented at the Sixth Workshop on Very Large Corpora. Association for Computational Linguistics, New Brunswick, New Jersey.

Boubekeur, F. (2008). Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets. Informatique [cs]. Université Paul Sabatier - Toulouse III, thèse de doctorat.

Boufaden, N. (2005). Extraction d'informations à partir de conversation téléphoniques spécialisées. Université de Montréal, thèse de doctorat.

Boughanem, M. (2006). Introduction à la Recherche d'Information. EARIA'06. Toulouse France.

C

Carpineto, C., Romano, G. (2012) .A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys 44(1):1-50.

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.

Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P., Vandenbussche, P. (2012). Construire une ontologie médicale pour la recherche d'information : problématiques

terminologiques et de modélisation. 23ièmes journées francophones d'Ingénierie des connaissances, Jun 2012, Paris, France, p.33-48.

Chen, Y., Lu, H., Shapiro, L., Ravensara S. T., & Li, L. (2016). An approach to semantic query expansion system based on Hepatitis ontology. J Biol Res (Thessalon) 2016 May 23(1):11-22.

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

Christopher, D Manning., Rghavan, P., Schütze, H. (2008). Introduction to Information Retrieval, Cambridge University Press. https://nlp.stanford.edu/IR-book.

Colace, F., De santo, M., Greco, L., Napoletano, P. (2015).Improving relevance feedback-based query expansion by the use of a weighted word pairs approach. Journal of the association for information science and technologie.66(11):2223–2234.

D

Denis., E. C, Wasito., I. (2017). Automatic ontology construction using text corpora and ontology design patterns (ODPS) in Alzheimer's disease. Journal of a science and information. (10)2:59-66.

Dhombres, F., Jouannic, J.M., Jaulent, M. C., Charlet, J. (2010). Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. Sylvie DESPRES. 21èmes Journées Francophones d'Ingénierie des Connaissances, Jun 2010, Nîmes, France. Ecole des Mines d'Alès, p.171-182.

Díaz-Galiano, M.C., Martín-Valdivia, M.T.,& Ureña-López, L.A. (2009): Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in Biology and Medicine; 39(4): 396-403.

E

Elayeb, B., Bounhas, I., Ben Khiroun, O., Bellamine Ben Saoud, N. (2011). Towards a Possibilistic Information Retrieval System Using Semantic Query Expansion. In the 4th international conference on Internet Technologies and Applications (ITA'11), Grout V., Oram D., et Picking R., (Eds), p. 308-316.

Embarek Mehdi. (2008). Un système de question-réponse dans le domaine médical : le système Esculape. Université Paris-Est, thèse de doctorat.

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.

Eva, D, Morlane-Hondere, F, Campillos, L., Dhouha, B., Swen, R., Lavergne, T. (2015). LIMSI @ CLEF eHealth 2015 - task1b. CLEF 2015 Online Working Notes. CEUR-WS.

Even, F. (2005). Extraction d'information et modélisation de connaissances à partir de Notes de communication orale. UFR sciences et techniques, Université de Nantes, France, thèse de Doctorat.

F

Fishkin, R. (2005). A Glossary of Information Retrieval Terminology. February 14th, 2005. https://moz.com/blog/a-glossary-of-information-retrieval-terminology.

G

Ghoulam, A., Barigou, F., Belalem, G., Meziane, F. (2015). Using Local Grammar for Entity Extraction from Clinical Reports. International journal of interactive multimedia artificial intelligence 3(3):16-24.

Ghoulam, A., Barigou, F., Belalem, G. (2015). Information Extraction in the Medical Domain. Journal of Information Technology Research 8(2): 1-15.

Ghoulam, A., Barigou, F., Belalem, G., & Meziane, F. (2018). Query expansion using external resource in medical domain. International-journal-intelligent-information-technologies. 14(3): In press.

Gomez-Pérez, A., Fernandez-Lopez, M., Corcho, O. (2003). Ontological engineering. With examples from the areas of knowledge management, e-commerce and the Semantic Web. Advanced information and knowledge processing. London; Springer.

Guarino, N. (1998). Formal ontology and information systems. In Formal ontology in information systems. Proceedings of the International Conference on Formal Ontologies in Information Systems of (FOIS'98), Trento, 6–8 June 1998, p.3–15.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University.

Greengrass, E. (2000). Information Retrieval: A Survey. 30 November. Publisher: University of Maryland 2000, p. 1-224.

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

Grossman, D. A., & Frieder, O. (2004). In Information Retrieval: Algorithms and Heuristics Dordrecht: Springer Netherlands.

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

H

Hammache, A. (2013). Recherche d'information: un modèle de langue combinant mots simples et mots composés. Université Mouloud Mammeri de Tizi-Ouzou Informatique thèse de doctorat.

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

He, Y., Kayaalp, M. (2008). Biological entity recognition with Conditional Random Fields, In AMIA Annu Symp Proc, p. 293-297.

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 for the 4th Message Understanding Conference (MUC-4), p. 268-275.

Ho-Dac, LM., Tanguy, L., Grauby, C., Hnub, N., Heu Mby, A., Malosse, J., Riviere, L., Veltz-Mauclair, A., Wauquier, M. (2016). LITL at CLEF eHealth2016: recognizing entities in French biomedical documents. CLEF 2016 Online Working Notes.CEUR-WS.

Huang, Z., & Hu, X. (2013). Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. International Journal of Machine Learning and computing. 3(6):494-498.

J

Jayan, J.P., Rajeev, R., Sherly, E. (2013). A Hybrid Statistical Approach for Named Entity Recognition for Malayalam Language. International Joint Conference on Natural Language Processing. Nagoya, Japan, p. 58–63.

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. p.11-41.

Jingchi, Jiang., Yi, Guan., and Chao, Zhao. (2015). WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition Based on CRF. CLEF 2015 Online Working Notes. CEUR-WS.

Jing, Y., & Croft, W. B. (1994). An association thesaurus for information retrieval, RIAO 94 Conference Proceedings, p. 146-160.

JiYoung , L., Dernoncourt , F., Szolovits , P. (2017). MIT at SemEval-2017 Task 10: Relation Extraction with Convolutional Neural Networks. arXiv preprint arXiv:1704.01523.

Jovic, A., Prcela, M., Gamberger, D. (2007). Ontologies in Medical Knowledge Representation. Proc. of Int. Conf. Information Technology Interfaces, p. 535 – 540.

K

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

Khadim, D., Mougin, F., Diallo. G. (2014).Query expansion using external resources for improving information retrieval in the biomedical domain. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014). Vol-1180:189-194.

KRAFT, R. & ZIEN, J. (2004). Mining anchor text for query refinement. In Proceedings of the 13th International Conference on World Wide Web. ACM Press, p. 666–674.

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), p. 177–185.

L

Lafourcade, M., and Ramadier, L., (2016). Semantic Relation Extraction with Semantic Patterns: Experiment on Radiology Report. Proceedings of the Tenth International Conference on Language Resources and Evaluation, portorož, Slovenia.

Lassila, O., & McGuinness, D. (2001). The role of frame-based representation on the Semantic Web. Technical Reports KSL-01-02, Knowledge Systems Laboratory, Stanford University.

Lee, C.H., Khoo, C., Na, J.C. (2004). "Automatic identification of treatment relations for medical ontology learning: An exploratory study". Proceedings of the Eighth International ISKO Conference 2004, p. 245-250.

Li J, Zhang Y, Tian Y. (2016). Medical big data analysis in hospital information system. In: Ventura Soto S, Luna JM, Cano A, editors. Big data on real-world applications. Rijeka (Croatia): InTech; 2016, p. 65–96.

Lin, W., Dou, W., Zhou, Z., Liu, C. (2015). A cloud-based framework for Home-diagnosis service over big medical data. Journal of Systems and Software .Volume 102, April 2015, P. 192-206.

M

Mataoui, M. (2007). Reformulation de requêtes dans les systèmes de recherches d'information dans des documents XML. Université M'hamed BOUGARA de Boumerdes, Algérie, thèse de Magister, Informatique.

McGuinness, D. L. (2003). Ontologies come of age. In Spinning the Semantic Web: Bringing the world wide web to its full potential, ed. D. Fensel, J. A. Hendler, H. Lieberman, and W. Wahlster, Cambridge: MIT Press, p. 171–194.

Meng, F., & Morioka, C., (2015). Automating the generation of lexical patterns for processing free text in clinical documents. Journal of the American Medical Informatics Association, ocv012, 22(5): 980–986.

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", year book of Medical Informatics. p. 44-128.

Middleto, C.,& Baeza-Yates, R.A. (2007). A Comparison of Open Source Search Engines. Middleton2007 ACO, p. 1-26.

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

Mohameth, F., Sylvie, R., Jacky, M. (2012). OBIRS-feedback, une méthode de reformulation utilisant une ontologie de domaine. Conférence en Recherche d'Information et Applications, CORIA 2012, Mar 2012, Bordeaux, France. pp.135-150.

N

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

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout. R.W. (1991). Enabling technology for knowledge sharing. AI Magazine 12(3):36–56.

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

Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P. (2014). The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In Proc BioTextM, Reykjavik.

Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P. (2015). CLEF eHealth Evaluation Lab 2015 Task 1b: clinical named entity recognition. CLEF 2015, Online Working Notes, CEUR-WS 1391.

Névéol, A., Bretonnel Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Grégoire, R., Aude, R., Tannier, X., Zweigenbaum, P. (2016). Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016. CLEF 2016, Online Working Notes, CEUR-WS 1609.

Nirenburg, S., & Raskin, V. 2004. Ontological Semantics. Cambridge, MA: MIT Press, P. 5-332.

P

Paumier, S. (2016). Manuel d'utilisation Unitex 3.1. Université Paris-Est Marne-la-Vallée. http://www-igm.univ-mlv.fr/~unitex. unitex-devel@univ-mlv.fr, p. 1-394.

Picariello, A., & Rinaldi, A. M. (2007). User relevance feedback in semantic information retrieval. International Journal of Intelligent Information Technologies 3(2): 36-50.

Poibeau, T. (2005). Sur le statut référentiel des entités nommées. Conférence Traitement Automatique des Langues, Dourdan, France. Association pour le Traitement Automatique des Langues / LIMSI, p.173-183.

Pragati, B., & Pareek, P. (2014). Novel Approach for Query Expansion Using Genetic Algorithm. International Journal of Information and Computation Technology. (4)3: 239-246.

Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. Multi-source, multilingual information extraction and summarization, Theory and Applications of Natural Language Processing, Book, Springer, p. 23-49.

R

Rashmi, V. (2016). Erasure Coding for Big-data Systems: Theory and Practice. Electrical Engineering and Computer Sciences. University of California at Berkeley, thèse de doctorat.

Ritesh, S., & Suresh, J. (2014). Ontology-based Information Extraction: An Overview and a Study of different Approaches. International Journal of Computer Applications 87(4):6-8.

S

Salton, G. & McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York.

Salton G . (1981). The Smart environment for retrieval system evaluation- advantages and problem areas. In k. Sparck Jones (Ed). Information retrieval experiment. London, Butterworth, p.316-326,.

Santos, R., McCreadie,R., Plachouras, V. (2011). Large-scale Information Retrieval Experimentation with Terrier. 20th ACM Conference on Information and Knowledge Management, p.1-68.

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

Serrano, L., Grilhères, B., Bouzid, M., Charnois, T. (2011). Extraction de connaissances pour le renseignement en sources ouvertes.. Workshop SOS'2011 at EGC'2011, France.

Sobhy, D., El-Sonbaty, Y., Abou Elnasr, M. (2012). MedCloud : Healthcare Cloud Computing System. The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012). London UK, p. 161-166.

Sowa, J. (1984). Conceptual Structures : Information Processing in Mind and Machine, Addison-Wesley.

Spasic, I., Sarafraz, F., Akeane, J., Nenadic, G. (2010). Medication information extraction with linguistic pattern matching and semantic rules, J Am Med Inform Assoc 17(5):532-535.

Sugam, S., Udoyara, S.T., Wong, J.S., Gadia, S., Sharma, S. (2014). A Brief Review on Leading Big Data Models. Data Science Journal, 13(4):138-157.

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.

Sunil, K. S & Ashish, A. (2016). Recurrent neural network models for disease name recognition using domain invariant features. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, p. 2216–2225.

Sure, Y., & Studer, R. (2003). A methodology for ontology-based knowledge management. In Towards the Semantic Web. Ontology-driven knowledge management, (eds.) J. Davies, D. Fensel, and F. van Harmelen,. Chichester: Wiley, p.33–46.

T

Tamine, L. (2000). Optimisation de requêtes dans un système de recherche d'information, Université Paul Sabatier de Toulouse, France, thèse de doctorat.

Tannier, X., (2005). Extraction d'Information et Structure de Documents. Reconnaissance des entités nommées. Master 2 Pro. Université Paris SUD.

Thenmalar, S., Balaji, T., Geetha, T.V. (2015). Semi-supervised Bootstrapping approach for Named Entity Recognition. arXiv:1511.06833 [cs.CL].

U

Uschold, M., &King, M. (1995). Towards a methodology for building ontologies. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at IJCAI'95, (ed.) D. Skuce, Montreal.

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.

V

Van Heijst, G. (1995). The role of ontologies in knowledge engineering. Social Science Informatics, University of Amsterdam, thèse de doctorat.

Van Mulligen, E., Afzal, Z., Akhondi, SA., Vo, D., Kors, JA. (2016). Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts. CLEF 2016 Online Working Notes, CEUR-WS.

Van Rijsbergen C. J.(1979). Information Retrieval, Second Edition, Butterworths.

Vandecasteele, A., Napoli, A,. (2012). An Enhanced Spatial Reasoning Ontology for gMaritime Anomaly Detection. 7th International Conference on System Of Systems Engineering - IEEE SOSE 2012, p. 247-252.

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

Vinciarelli, A., & Favre, S .(2007). Broadcast news story segmentation using social network analysis and hidden markov models. September 2007 MM '07: Proceedings of the 15th ACM international conference on Multimedia.

W

Wang, X., & Zhai, C. (2008). Mining term association patterns from search logs for effective query reformulation. In Proceeding of CIKM, California, USA, p. 479-488.

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(5):883-887.

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.

Wu,Y., Jiang, M., Lei, J., Xu, H. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. Stud Health Technol Inform. 2015; Vol. 216; p. 624-628.

X

Xavier, L.(2005). Introduction à owl, un langage xml d'ontologies web, Aout 2005.

Xinbo, L., Guan, Y., Yang, J., and Jiawei, W. (2016). Clinical Relation Extraction with Deep Learning, International Journal of Hybrid Information Technology, 9(7): 237-248.

Xu, Y., Jones, G. J. F., & Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, p. 59–66.

Y

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.

Yangyang., K, Jianqiang., Li, Jijiang., Y, Qing., W, Zhihua., S. (2017). Semantic Analysis for Enhanced Medical Retrieval. IEEE International Conference on Systems, Man, and Cybernetics (SMC) Banff Center, Banff, Canada.

Z

Zaidi–Ayad, S. (2012). Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran), thèse de doctorat.

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics, 46(6): 1088-1098.

Zribi, I; Mezghani Hammami, S;and Belguith Hadrich, L. (2010). L'apport d'une Approche Hybride pour la Reconnaissance des Entités Nommées en Langue Arabe. 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN), Montréal, p. 1-6.

Zulkarnain, N., Meziane, F., Croft, G. (2016). A Methodology for Biomedical Ontology Reuse. 21st international conference on applications of natural language to information systems, NLDB, Salford, UK, p. 3-14.

ملخص

عمل هذه الأطروحة يدور حول ثلاثة محاور: (1) استخراج المعلومات لجمع وبناء الأنطولوجيا الطبية (2) البحث عن المعلومات الدلالية في النظام الطبي و (3)اقتراح إنشاء نظام بحث طبي واسع النطاق .في مساهمتنا الأولى، نقترح استخدام طريقة قائمة على قواعد لاستخراج الكيانات الطبية والعلاقات التي تربط هذه الكيانات الموجودة في التقارير الطبية باللغة الفرنسية .استخدمت القواعد المحلية من جهة (1) لتمثيل قواعد لتحديد الكيانات والعلاقات، ومن ناحية أخرى (2) لتصنيف الكيانات والعلاقات .هدفنا في هذا العمل هو بناء الأنطولوجيا الطبية بطريقة شبه آلية. في مساهمتنا الثانية، قمنا بدراسة تأثير استخراج المعلومات لاسترجاعها، اقترحنا توسيع طلب بحث المستخدم في نظام البحوث الطبية باستخدام علم الأنطولوجيا التي بنيت في مساهمتنا الأولى .لذلك اقترحنا ثلاثة تقنيات توسع: (1) توسع الكيانات الطبية، (2) التوسع عن طريق استخراج العلاقات الدلالية و (3) التوسع عن طريق استخراج العلاقات وإعادة صياغة منطقية للاستعلام .يتم اختبار هذه التقنيات الثلاث ضد نظام البحث التقليدي .في مساهمتنا الثالثة، للتعامل مع مشاكل التخزين والتحليل نقترح استخدام الحوسبة السحابية التي لها قابلية عالية لتمثيل مؤشر نظامنا من البحوث وعلم الأنماط في بيئة موزعة.

الكلمات المفتاحية: استخراج المعلومات، التعرف على الكيانات الطبية، استخراج العلاقات الطبية، الأنتولوجيا، استرجاع معلومات، توسيع الاستعلام، نطاق واسع.

Abstract

The work of this thesis revolves around three axes: (1) the extraction of information for the collection and construction of a medical ontology (2) the search for semantic information in a medical system and (3) the proposal for a large-scale medical research system. In our first contribution, we propose to use a rule-based method to extract the medical entities and relationships linking these entities present in the French clinical reports. Local grammars are used on the one hand to (i) represent the rules for the identification of entities and relationships and on the other hand to (ii) categorize entities and relationships. Our goal in this work is to semi-automatically build a medical ontology. In our second contribution, we studied the impact of information retrieval for information retrieval, for which we propose to extend the user's query in a medical research system using ontology built in our first contribution. Three expansion techniques are proposed: (i) expansion of the medical entities, (ii) expansion by extraction of semantic relations and (iii) expansion by extraction of relations and Boolean reformulation of the query. These three techniques are tested against a traditional search system. In our third contribution, to deal with the problems of storage and analysis we propose to use the Cloud Computing which will allow a high scalability for the representation of the index of our system of research and the ontology in a distributed environment.

Keywords: Information Extraction, Medical Entities Recognition, Medical Relations Extraction, Ontology, Information Retrieval, Query Expansion, Large Scale.

Résumé

Cette thèse s'intéresse à l'étude de l'extraction d'information et son impact dans un système de recherche d'information médicale à large échelle. Pour cela, nous avons subdivisé le travail de cette recherche en trois étapes principales : (1) l'extraction d'information (entités et relations) en vue de la construction d'une ontologie médicale (2) la recherche d'information médicale à partir des rapports médicaux (3) la proposition d'un système de recherche médicale à large échelle. Dans notre première contribution, nous proposons d'utilisé une méthode à base de règles pour extraire les entités médicales et les relations reliant ces entités présentes dans les rapports cliniques français. Les grammaires locales sont utilisées d'une part pour (i) représenter les règles pour l'identification des entités et des relations et d'autre part pour (ii) catégoriser les entités et les relations. Notre objectif, dans ce travail est de construire semi-automatiquement une ontologie médicale. Dans notre deuxième contribution, nous avons étudié l'impact de l'extraction d'information pour la recherche d'information, pour cela nous proposons d'étendre la requête de l'utilisateur dans un système de recherche médical par utilisation de l'ontologie construite dans notre première contribution. Trois techniques d'expansion sont proposées: (i) expansion des entités médical, (ii) expansion par extraction de relations sémantiques et (iii) expansion par extraction de relations et reformulation booléenne de la requête. Ces trois techniques sont testées par rapport à un système de recherche traditionnel. Dans notre troisième contribution, pour faire face aux problèmes de stockage et d'analyse nous proposons d'utiliser le Cloud Computing ce qui permettra une haute évolutivité (scalabilité) pour la représentation de l'index de notre système de recherche et l'ontologie dans un environnement distribué.

Mots-clés : Extraction d'Information, Reconnaissance des Entités Médicales, Extraction de Relations Médicales, Ontologie, Recherche d'Information, Expansion de Requête, Large Échelle.

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 prespecified 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.

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

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: $F - score = \frac{\left(B^2 + 1\right)PR}{B^2P + R}$. 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 rep-

resents these types of relations; which are used to organize concepts as a tree, like in thesaurus of Medical Subject Headings (MeSH) or meta-thesaurus 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.

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.

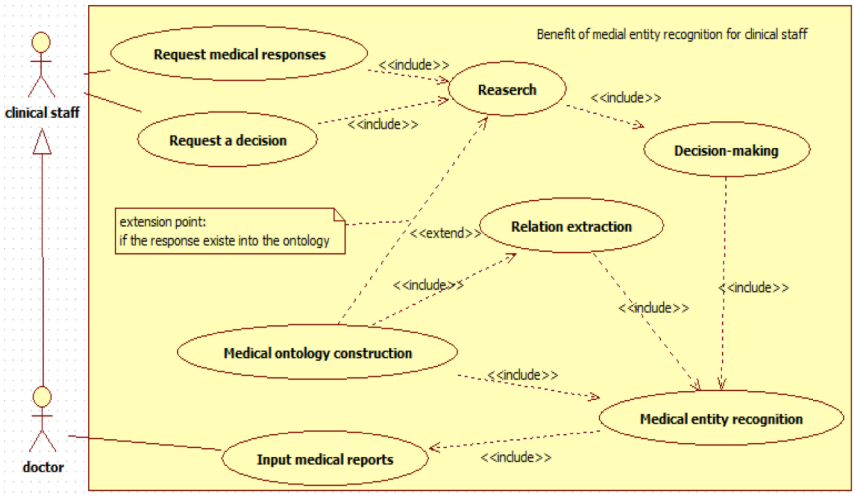

Fig.1. Medical entity recognition's benefits

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

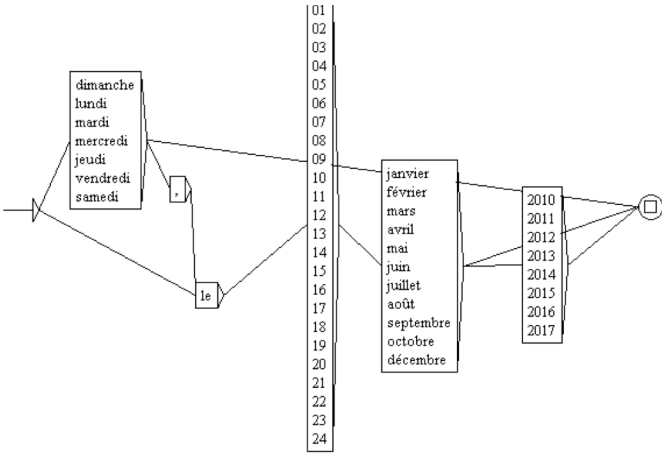

Fig.2. Example of a local grammar for French date expressions

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:

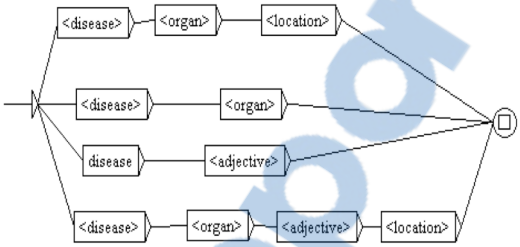

Fig.3. Example of the local grammars of disease entity

- 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

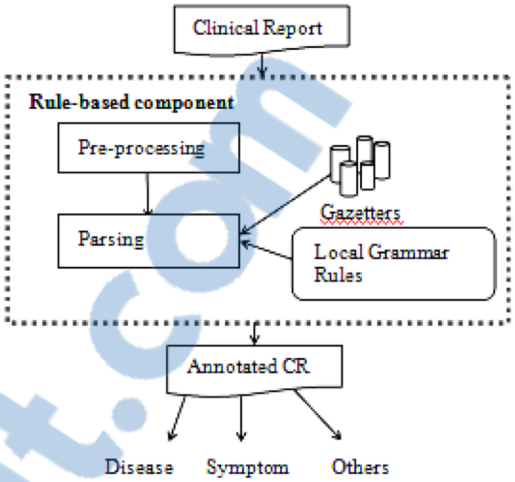

Fig.4. Architecture of the system

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.

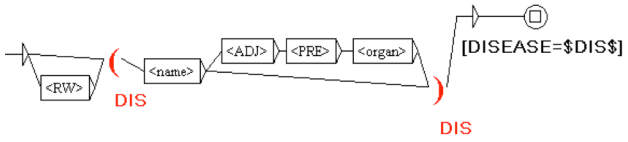

Fig.5. Example of Local Grammar in Unitex

(*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.

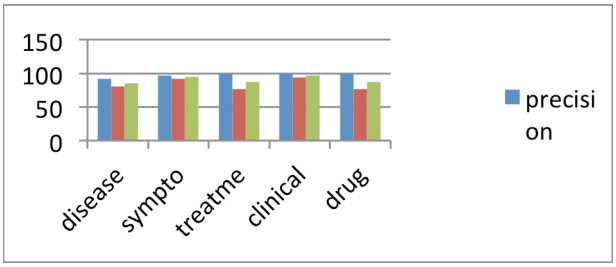

Fig.6. Performance system

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.

Résumé

Cette thèse s'intéresse à l'étude de l'extraction d'information et son impact dans un système de recherche d'information médicale à large échelle. Pour cela, nous avons subdivisé le travail de cette recherche en trois étapes principales :- (1) l'extraction d'information (entités et relations) en vue de la construction d'une ontologie médicale.- (2) la recherche d'information médicale à partir des rapports médicaux.-(3) la proposition d'un système de recherche médicale à large échelle. Pour la première étape, notre contribution consiste à proposer une méthode à base de règles pour extraire les entités médicales et aussi les relations qui les relient à partir des rapports médicaux français. La méthode d'extraction s'appuie sur l'utilisation des grammaires locales d'une part pour (i) représenter les règles pour l'identification des entités et des relations et d'autre part pour (ii) catégoriser les entités et les relations. L'objectif de cette étape est la construction semi-automatique d'une ontologie médicale. Dans la deuxième étape, et dans un objectif d'amélioration de la qualité d'un système de recherche médicale, nous avons étudié l'impact de l'extraction d'information pour la recherche d'information médicale. Notre contribution consiste à proposer des techniques d'expansion de requêtes qui se basent sur l'utilisation d'une ontologie auparavant construite à partir des entités et des relations extraites des rapports médicaux. Trois techniques d'expansion sont proposées: (i) expansion des entités médicales présentes dans la requête par synonymes et descendants, (ii) expansion par extraction de relations sémantiques et (iii) expansion par extraction de relations et reformulation booléenne de la requête. Pour faire face aux problèmes de stockage et d'analyse à grande échelle, nous proposons dans la troisième étape d'étendre le système déjà proposé dans l'étape deux dans un environnement distribué et exploiter les avantages du Cloud Computing pour permettre une haute évolutivité (scalabilité) dans le cas de la représentation de l'index du système de recherche et aussi l'ontologie.

Mots-clés :

Extraction d'Information; Reconnaissance des Entités Médicales; Extraction de Relations Médicales; Rapport clinique; Entité médicale; Ontologie; Recherche d'Information, Lucene; Expansion de Requête; Large Échelle.