Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS

Asma Ben Abacha

UNIVERSITE PARIS-SUD 11
LIMSI-CNRS

Thèse

présentée en première version en vu d'obtenir le grade de Docteur,
spécialité « Informatique »

par

Asma BEN ABACHA

Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS

Thèse soutenue le 28 juin 2012 devant le jury composé de :

M.	Guy Lapalme	Professeur, Université de Montréal	(Rapporteur)
M.	Thierry Poibeau	Directeur de recherche, LATTICE-CNRS	(Rapporteur)
M.	Stéfan Darmoni	Professeur, CHU de Rouen	(Examinateur)
M.	Olivier Ferret	Chercheur, CEA-LIST	(Examinateur)
M.	Joseph Mariani	Directeur de recherche, LIMSI-CNRS	(Examinateur)
M.	Pierre Zweigenbaum	Directeur de recherche, LIMSI-CNRS	(Directeur de thèse)

À mes parents qui m'ont appris à aller de l'avant..
À mon mari qui me donne l'espoir d'aller de l'avant..

Remerciements

Et vient le moment d'écrire les remerciements... Quand je pense à toutes les personnes sans qui ce travail n'aurait probablement pas pu être achevé, je me dis que ce n'est pas pour rien que cette page précède la présentation de tout travail de thèse.

Je voudrais tout d'abord exprimer mes plus profonds remerciements à mon directeur de thèse, Pierre Zweigenbaum, pour le choix du sujet et des axes d'orientation scientifiques de la thèse, pour sa confiance, pour la liberté qu'il m'a accordée et pour la qualité de son encadrement. Malgré ses responsabilités, il a su être présent quand il le fallait. Je voudrais aussi le remercier pour ses qualités personnelles et humaines qui ont aussi beaucoup contribué à la réalisation de ce travail.

Je tiens à remercier mes deux rapporteurs de thèse, Guy Lapalme et Thierry Poibeau pour avoir accepté de relire ce manuscrit et pour leurs remarques et propositions intéressantes. Je voudrais remercier Guy Lapalme pour sa présence le jour de ma soutenance en faisant le déplacement de si loin. Je souhaiterais aussi remercier Thierry Poibeau pour son cours au master qui m'a fait découvrir la recherche et l'extraction d'information et les systèmes de questions-réponses.

Je tiens aussi à remercier les examinateurs de ma thèse : Stéfan Darmoni, Olivier Ferret et Joseph Mariani : merci à vous trois d'avoir accepté d'être parmi le jury de ma thèse. Merci beaucoup pour vos remarques et conseils. Merci également à Stéfan Darmoni pour sa bonne humeur :-) C'est un grand honneur pour moi d'avoir un tel jury à ma soutenance.

Je remercie tout spécialement Anne Vilnat, responsable de l'équipe ILES durant presque tout mon passage au LIMSI, pour sa présence et son dévouement au service de son équipe malgré toutes ses charges.

J'aimerais adresser un remerciement particulier à mes deux collègues de bureau : Aurélien Max et Gabriel Illouz pour leur gentillesse, les cafés, et les discussions toujours intéressantes qu'on a eu.

Je remercie également mes voisins de bureau : Anne-Lyse Minard pour tous les cafés et moments partagés, Béatrice Arnulphy pour ton attention et toutes nos discussions et Cyril Grouin pour tous tes conseils. J'espère que ce bureau restera toujours riche aussi bien par des personnes sympathiques que par des bons gâteaux et chocolats :-)

Je voudrais remercier toutes les personnes avec qui j'ai eu le plaisir de collaborer durant ces années et en particulier Faisal Chowdhury, Aurélien Max, Brigitte Grau, Anne-Laure Ligozat, etc.

Je remercie chaleureusement tous les limsiens qui ont rendu cet espace de travail vivant et souriant, en particulier mes compatriotes Houda Bouamor et Souhir Gahbiche et tous les limsiens que j'ai eu le plaisir de côtoyer durant ces quelques années et que je n'ai pas cités (je pense en particulier à Sophie Rosset et à tous les iliens).

Je terminerai en remerciant de tout cœur tous ceux sans qui cette thèse ne serait pas ce qu'elle est :

Tous les amis et proches, et en particulier les Asma(s), Olfa, Abir, Samia,.. (et la liste est encore longue) : merci d'avoir été toujours présentes.

Mes chers parents : Amel et Hassen, je ne pourrais jamais vous rendre tout ce que vous avez fait et ce que vous faites pour nous.

Ma chère soeur Imen et mes chers frères Anis et Ahmed. Toute ma famille et ma belle-famille et en particulier mes beaux-parents Émna et Ahmed et mon oncle Hamouda.

Mon cher mari Yassine Mrabet : merci pour ta présence, tes qualités et ton grand cœur. Tu n'as jamais hésité à m'aider quand tu le pouvais durant ces années.

Ma chère fille Éya, qui a vécu deux thèses à la maison durant ses trois premières années dans la vie (ça commence bien pour toi :-). Je te souhaite une vie sans stress, avec plein d'amour, de savoir et de travail.

Paris, le 28 juin 2012.

Table des matières

Liste des figures

Introduction générale

1

Sommaire

Ce chapitre présente dans un premier temps les éléments qui ont motivé nos travaux de recherche ainsi que les problématiques étudiées. Dans un second temps, nous décrivons l'approche proposée pour la recherche de réponses à des questions posées en langue naturelle. Cette approche vise à la réalisation d'un système de questions-réponses (SQR) pour le domaine médical. Le plan de ce manuscrit est présenté dans la dernière section de ce chapitre.

1.1 Motivations

Au cours des dernières décennies, la quantité d'information a augmenté de façon exponentielle dans tous les domaines et notamment le domaine médical. En effet, le volume des connaissances médicales double tous les 5 ans (Engelbrecht 1997), voire tous les 2 ans (Hotvedt 1996). Avec la numérisation à large échelle, retrouver automatiquement une information de haute précision est devenu un défi. Dans ce contexte, plusieurs moteurs de recherche spécialisés dans ce domaine ont vu le jour : citons par exemple PubMed [1], CISMeF [2] ou Health On the Net [3]. À une requête donnée, ces moteurs retournent un ensemble de documents et délèguent à l'utilisateur la tâche de trouver l'information cherchée si elle existe dans les documents renvoyés.

Pour faciliter et accélérer la recherche d'information des systèmes plus précis sont mis en œuvre comme les systèmes de questions-réponses. Un système de questions-réponses vise à répondre directement à des questions posées en langue naturelle avec une réponse précise extraite à partir d'une collection de documents, du Web ou d'une base de données. La nature et la complexité de cette tâche varient selon le domaine traité et les types des questions posées par les utilisateurs.

1.2 Objectif

Cette thèse porte sur l'étude de méthodes permettant de rechercher des réponses à des questions médicales dans une base de documents médicaux en anglais. Nous étudions entre autres les questions suivantes :

1. Dans quelle mesure les méthodes utilisées dans des systèmes en domaine ouvert sont-elles transposables à ce domaine ?
2. Les conditions différentes rendent-elles possibles ou nécessaires la conception de méthodes nouvelles ?

Dans ce travail, nous nous intéressons essentiellement à des corpus de la littérature biomédicale (e.g. Medline [4], PubMed [5]) bien que nous examinions aussi l'applicabilité de certaines méthodes à des textes cliniques (e.g. challenge i2b2 2010 [6]).

1.3 Problématique

Plusieurs critères et conditions du domaine ouvert ne sont pas valables pour le domaine médical. Par exemple, le rôle des pronoms interrogatifs : 'When' indique une date en domaine ouvert alors que dans le domaine médical, il peut indiquer une condition (e.g. *When should you suspect community-acquired MRSA ?*). D'autres caractéristiques particulières sont aussi à prendre en compte en domaine médical, citons :

1. http://www.pubmed.com
2. http://www.chu-rouen.fr/cismef/
3. http://www.hon.ch
4. http://www.nlm.nih.gov/pubs/factsheets/medline.html
5. http://www.nlm.nih.gov/pubs/factsheets/pubmed.html
6. https://www.i2b2.org/NLP/Relations/

- les nouveaux types de questions (non nécessairement factuelles),
- les nouveaux types de réponses attendues,
- les entités de domaine autres que les habituelles «entités nommées» des questions factuelles en domaine ouvert (e.g. maladies ou traitements plutôt que noms de personnes connues),
- les relations sémantiques spécialisées (e.g. traiter, prévenir, diagnostiquer),
- la disponibilité de ressources sémantiques et terminologiques spécialisées (e.g. UMLS[7], MeSH[8]),
- les types de documents ciblés (articles scientifiques et leur résumé, recommandations pour la pratique clinique, ouvrages didactiques, etc.) : collections en général «fermées» mais potentiellement très vastes et «fiables».

Le domaine médical a donc besoin de nouvelles approches pour l'extraction d'information à partir des corpus textuels, l'analyse de questions et l'extraction de réponses.

1.4 Existant

De nombreux travaux ont été proposés pour la recherche de réponses à des questions en domaine ouvert (e.g. le système START[9] (Katz 1999, Katz et al. 2002)). D'autres travaux ont porté sur la recherche de réponses en domaine de spécialité (e.g. le système spécialisé ExtrAns (Rinaldi et al. 2004) pour la génomique), en particulier dans le domaine médical ou biomédical (e.g. Le système proposé par (Terol et al. 2007), EpoCare (Niu et Hirst 2004)).

L'existant des SQR médicaux peut être classifié en deux principales approches :

Approche surfacique - syntaxique. Il s'agit de rechercher des passages de textes pouvant contenir la réponse puis l'extraire. La recherche passe généralement par une étape d'indexation qui consiste à identifier les mots ou expressions présentant le mieux le contenu d'un document. La méthode la plus simple pour réaliser cette indexation est de considérer le document comme un « sac de mots » (considérer que les mots sont indépendants et ne pas tenir compte des relations entre les termes). Des traitements linguistiques comme une analyse morphologique et/ou syntaxique des documents peuvent aussi enrichir cette indexation. Dans cette approche, l'extraction de la réponse se fonde sur quelques techniques issues du Traitement Automatique des Langues (TAL) qui n'impliquent pas forcément une analyse sémantique de la question et des documents interrogés.

Approche profonde - sémantique. Cette approche procède typiquement à une analyse sémantique de la question et des documents et produit une représentation formelle de leur sens. Comme exemple de travaux qui se fondent sur cette approche, citons (Rinaldi et al. 2004) qui dérivent des représentations logiques des questions et des documents. Quand un appa-

7. Unified Medical Language System
8. Medical Subject Headings
9. http ://start.csail.mit.edu/

riement se produit entre la représentation logique de la question et celle d'un document, les phrases initiant l'appariement sont extraites comme réponses candidates. (Niu et al. 2003) ont analysé les limitations des ontologies de questions-réponses générales en domaine médical. Ils ont présenté une alternative se fondant sur l'identification des rôles sémantiques dans la question et les textes qui seront utilisés pour chercher et extraire les réponses. Cette identification utilise le format PICO (Sackett et al. 2000) comme une représentation de référence (P : Population/disease, I : Intervention or Variable of Interest, C : Comparison, O : Outcome). Ce type de représentation a été beaucoup utilisé ces dernières années en domaine médical (e.g. (Demner-Fushman et Lin 2007)). Un cas particulier de cette approche, appelé approche à base de modèles (template-based approach), exploite une collection de modèles (ou types génériques) de questions créés manuellement et utilise des techniques sophistiquées de TAL pour la classification des questions. Comme exemple de travaux, (Terol et al. 2007) ont classifié des questions médicales en se basant sur une taxonomie de questions cliniques génériques (Ely et al. 2000).

Nous pensons que les approches profondes - sémantiques sont les plus appropriées au domaine médical. En effet, ces approches permettent de traiter des questions plus complexes et se fondent sur des techniques qui permettent une analyse sémantique des questions et des documents interrogés. Ceci offrira la performance nécessaire pour ce domaine de spécialité où on privilégie souvent la précision sur le rappel.

1.5 Approche proposée

Nous proposons une approche sémantique pour répondre aux questions médicales (cf. figure 1.1) qui combine l'utilisation de connaissances du domaine médical, de techniques de Traitement Automatique de la Langue (TAL) et de technologies du Web Sémantique.

Cette approche effectue une interprétation sémantique des documents et de la question et ramène le problème de questions-réponses en langage naturel à l'interrogation de méta-données structurées suivant une ontologie de domaine. L'idée est d'associer des graphes sémantiques à la question d'une part et aux phrases du corpus d'autre part puis de rechercher les appariements pour trouver l'extrait de document qui répond à la question posée. Ceci permet de traiter toute question, même si elle ne correspond pas à un type de question prédéterminé.

Le processus employé peut être résumé en trois étapes principales :

1. L'annotation en RDF des documents utilisés pour trouver les réponses. Pour ce faire, nous procédons en deux étapes :
   - La reconnaissance des entités médicales présentes dans les textes (e.g. maladie, médicament).
   - L'extraction des relations sémantiques qui les relient (e.g. traiter, prévenir, causer).
2. L'analyse des questions médicales posées en langage naturel et l'extraction des informations importantes (e.g. type de la question, type de la réponse attendue, entités médicales, relations sémantiques,

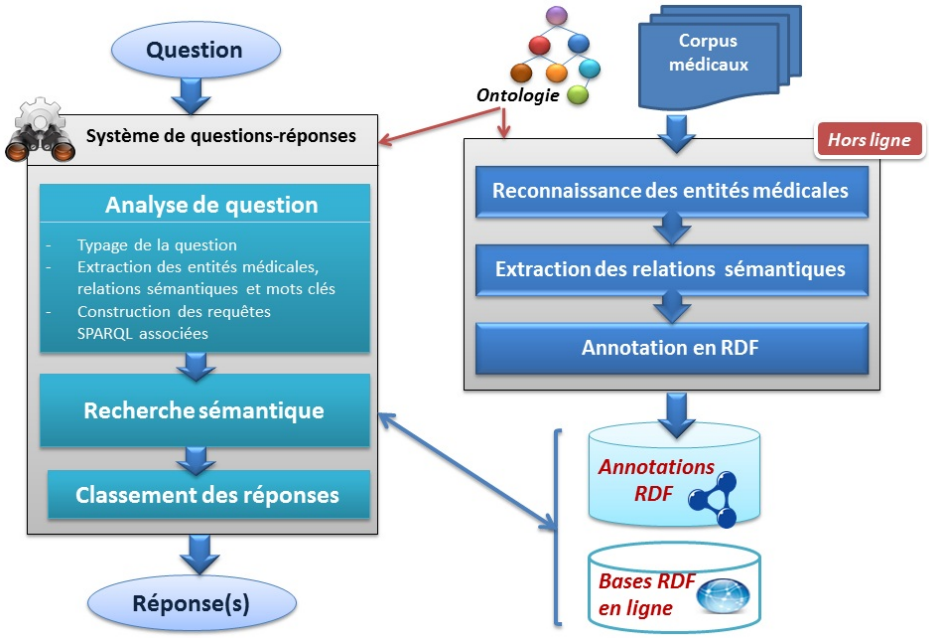


FIGURE 1.1 – *Approche proposée pour la réalisation d'un système de questions-réponses pour le domaine médical*

mots clés). Ces informations seront utilisées pour construire des requêtes SPARQL structurées.

3. La recherche de réponses se fait en interrogeant les annotations RDF avec des requêtes SPARQL construites à partir de la question.

L'utilisation de cette approche et de ces technologies nous permet d'envisager d'exploiter les bases de connaissances RDF en ligne (e.g. BIO2RDF (Belleau et al. 2008), BioGateway [10], DrugBank [11]) comme sources complémentaires de réponses aux annotations de corpus médicaux.

1.6 Organisation du document

Cette thèse est organisée en deux parties. Nous consacrons la première partie à la tâche d'extraction d'information à partir de textes médicaux. Il s'agit d'une tâche importante à la fois pour l'analyse de question et l'annotation des corpus utilisés pour trouver les réponses. Nous nous intéressons en particulier à la reconnaissance des entités médicales (chapitre 2) et à l'extraction de relations sémantiques les reliant (chapitre 3). Nous proposons différentes méthodes (i.e. à base de règles ou patrons, apprentissage et hybrides) et nous discutons leur robustesse, leur portabilité et la manière de les combiner. Le dernier chapitre de la première partie est consacré à l'extraction d'information en d'autres langues par projection des annotations (e.g. entités médicales) d'une langue L1 à une langue L2 (chapitre 4). Ce travail est effectué dans la perspective de réaliser un SQR translingue.

10. http://www.semantic-systems-biology.org/biogateway
11. http://www4.wiwiss.fu-berlin.de/drugbank/

La deuxième partie est dédiée à la tâche Question-Réponse. Nous commençons cette partie par un état de l'art sur les systèmes de questions-réponses en domaine ouvert et en domaine médical (chapitre 5). Nous présentons ensuite notre approche qui comporte deux grandes étapes : l'analyse de question (chapitre 6) et la recherche et l'extraction de réponses (chapitre 7). Nous nous intéressons aussi dans le chapitre 7 à la présentation et l'évaluation du système de questions-réponses proposé (appelé MEANS). Enfin nous concluons et donnons quelques perspectives (chapitre 8).

Première partie

Extraction d'information à partir de textes médicaux

Nous consacrons cette première partie à l'extraction d'information à partir de textes médicaux (cf. figure 1.2). En particulier nous nous intéressons à la reconnaissance des entités médicales et à l'extraction des relations sémantiques les reliant. Ces deux tâches seront utilisées au niveau de l'analyse de question et aussi pour l'annotation des corpus médicaux utilisés pour trouver les réponses.

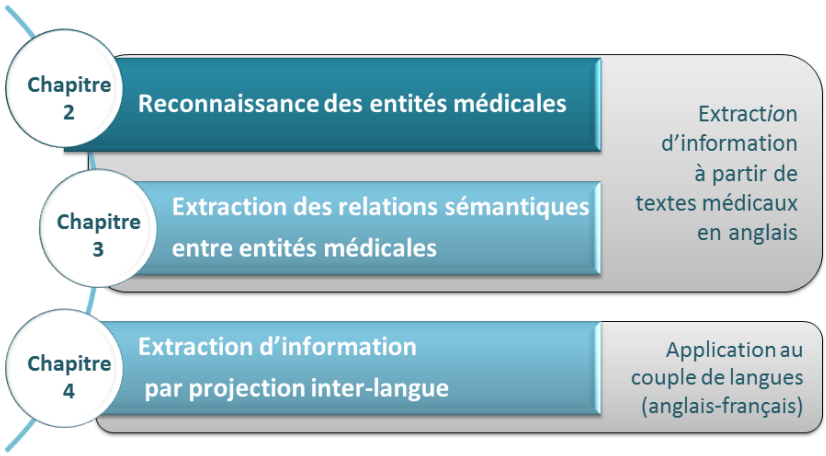


FIGURE 1.2 – *Partie 1 : Extraction d'information à partir de textes médicaux*

Le chapitre 2 traite la reconnaissance des entités médicales à partir de textes médicaux en anglais. Nous proposons différentes méthodes à base de règles, statistiques et hybrides que nous appliquons sur deux corpus médicaux de types différents : résumés d'articles scientifiques et textes cliniques.

Le chapitre 3 s'intéresse à l'identification des relations sémantiques entre les entités médicales extraites. Nous proposons une méthode à base de patrons et une méthode statistique et nous discutons la combinaison de ces deux méthodes en tenant compte de différents paramètres.

Le chapitre 4 traite l'application de ces méthodes d'extraction d'information à d'autres langues. L'idée est d'exploiter des corpus parallèles (e.g. anglais-français) dont on annote la partie anglaise par les méthodes déjà développées pour l'anglais puis où l'on projette ces annotations dans la partie française en utilisant les alignements au niveau des mots.

Reconnaissance des entités médicales à partir de textes médicaux en anglais

2

Sommaire

Une des étapes cruciales dans l'extraction d'information à partir de textes est la reconnaissance des entités nommées. L'expression *entité nommée* est apparue lors de la conférence MUC-6 (Message Understanding Conference). Ce sont les entités qui ont un désignant déterminé (e.g. "EDF", "Jules Verne"). Elles incluent les noms propres ou les expressions telles que les noms d'espèces (e.g. "tigre du Bengale"), de maladies, ou de substances chimiques. Cette définition a aussi été élargie aux expressions temporelles telles que les dates et les heures, ou à des valeurs numériques (e.g. 2.3 g/l).

Par entités médicales, nous désignons les entités nommées spécifiques au domaine médical telles que les maladies, les traitements ou les examens. Détecter de telles entités nécessite de disposer de ressources décrivant le vocabulaire du domaine et/ou de corpus d'entraînement permettant l'apprentissage des caractéristiques communes à ces entités.

De nombreux travaux se sont attaqués à cette tâche. Cependant il n'y a pas eu, à notre connaissance, d'études comparant deux stratégies pour l'extraction des entités médicales : (i) l'extraction en amont des syntagmes nominaux, suivie d'une étape de catégorisation de leur type et (ii) la détermination simultanée des frontières et des types des entités dans les textes. Ce chapitre est dédié à la présentation de cette étude. Nous testons ces deux stratégies en utilisant des méthodes à base de règles et/ou à base d'apprentissage. Nous comparons leur robustesse et aussi leur portabilité en les évaluant sur deux corpus médicaux standards de genres différents.

Nous commençons ce chapitre par un état de l'art sur les différentes techniques utilisées pour la reconnaissance d'entités nommées en domaine ouvert et en domaine médical (section 2.2). Nous consacrons la section 2.3 à la présentation de la tâche de reconnaissance des entités médicales. Nous présentons dans la section 2.4 les méthodes proposées. Les différentes méthodes proposées sont testées sur deux corpus différents. Un premier corpus de textes cliniques construit dans le cadre du challenge international i2b2 2010 (Uzuner et al. 2011). Un deuxième corpus de résumés d'articles scientifiques de MEDLINE proposé par (Rosario et Hearst 2004), appelé corpus de Berkeley. Les résultats obtenus sur ces deux corpus sont présentés dans les sections 2.5.1 et 2.5.2. Nous dédions la section 2.6 à une synthèse des différentes méthodes et directions étudiées et des résultats obtenus.

Les travaux de reconnaissance d'entités médicales présentés dans ce chapitre s'appuient sur les articles (Ben Abacha et Zweigenbaum 2011c; 2012c).

2.1 Introduction

Dans cette section, nous présentons les motivations, la problématique ainsi que nos objectifs.

Motivations

L'analyse profonde des textes médicaux passe obligatoirement par une étape de reconnaissance des entités médicales présentes dans le texte. Pour les systèmes de questions-réponses, la reconnaissance des entités médicales est utilisée au niveau de l'analyse de la question pour déterminer les mots clés et le type de la réponse attendue et aussi au niveau de la recherche des réponses possibles. La reconnaissance des entités médicales est une tâche importante voire nécessaire non seulement pour la recherche d'information et les systèmes de questions-réponses mais aussi pour d'autres tâches comme l'extraction de relations sémantiques entre entités médicales (e.g. (Vintar et al. 2003)), la détermination de la factivité des problèmes médicaux (e.g (Bernhard et Ligozat 2011)) et la résolution de coréférence (e.g. le challenge i2b2/VA 2011 [1]).

Problématique

La reconnaissance des entités médicales est une tâche complexe. Cette complexité réside dans les problèmes classiques rencontrés en domaine ouvert, mais aussi dans les spécificités du domaine médical. En effet, en domaine ouvert, les entités nommées désignent habituellement les noms de personnes, de lieux, d'entreprises, ainsi que les dates et les quantités monétaires. Mais d'autres catégories plus ou moins précises peuvent aussi être incluses (les événements, les fonctions, etc.), ce qui pose la question de la définition de cette tâche. La tâche de reconnaissance des entités nommées est-elle mal définie ? Sa définition dépend-elle du domaine et/ou de l'application ? Ce problème a motivé plusieurs travaux qui se sont intéressés à la définition de la tâche de reconnaissance des entités nommées (e.g. (Ehrmann 2008)). Le même problème se pose en domaine médical, à savoir quelle est la liste des catégories médicales (e.g. *Traitement*, *Examen*, *Problème médical*, etc.) et quelle est la définition exacte de chaque catégorie (e.g. les plantes doivent-elles être rangées parmi les traitements ?).

La polysémie est un deuxième obstacle à la reconnaissance des entités médicales que l'on retrouve en domaine ouvert. Par exemple le mot *drug* a deux sens (médicament ou drogue). À ceci s'ajoute le cas des mots non spécialisés qui, hors contexte, peuvent porter un sens médical particulier. Par exemple le mot *ten* (dix) désigne aussi la maladie *Toxic Epidermal Necrolysis* (le *syndrome de Lyell*). La réciproque est aussi vraie : certains termes médicaux, hors contexte, peuvent prendre d'autres sens. Par exemple, le terme médical *Antimicrobial agents* est aussi le nom d'un journal et d'un site web.

Les spécificités du domaine médical ajoutent aussi une couche de complexité aux défis communs soulevés en domaine ouvert. Ci-dessous, nous

1. https://www.i2b2.org/NLP/Coreference

citons quelques points clés spécifiques à la reconnaissance des entités médicales :

- La grande variation terminologique dans ce domaine de spécialité : chaque concept médical peut être désigné par plusieurs termes synonymes, des abréviations, etc. Par exemple, *Diabetes mellitus type 1*, *Type 1 diabetes*, *IDDM*, ou *juvenile diabetes* désignent le même concept. Il en est de même pour *Papanicolaou test*, *Pap smear*, *Pap test*, *cervical smear* ou encore *cervix smear* qui désignent le même examen médical.
- Certaines entités médicales peuvent avoir des noms différents selon les pays, c'est surtout le cas pour certaines maladies et médicaments. Par exemple, le médicament *Procarbazine* possède aussi les noms commerciaux suivants : *Matulane* (US), *Natulan* (Canada), *Indicarb* (Inde).
- L'évolution rapide de la terminologie médicale (e.g. nouvelles abréviations, noms de nouveaux médicaments ou maladies).

Ces obstacles limitent la généricité des méthodes qui se fondent sur des dictionnaires et des listes d'entités nommées (« gazetteers »). D'un autre côté, le domaine médical dispose de ressources spécialisées intéressantes (e.g. UMLS, voir plus bas). Cependant, ces ressources manquent parfois de précision et doivent être mises à jour d'une façon continue et rapide, ce qui n'est pas toujours le cas. Ceci conduit à l'adoption d'autres méthodes utilisant potentiellement des connaissances du domaine mais aussi exploitant les outils du TAL et les techniques connues en domaine ouvert telles que l'apprentissage.

Objectifs

Notre principal objectif dans ce chapitre est d'étudier et de comparer deux stratégies pour l'extraction des entités médicales. Une première stratégie qui s'appuie sur l'extraction en amont des syntagmes nominaux avec des outils spécialisés, suivie d'une étape de catégorisation et une deuxième stratégie ou les frontières et les catégories des entités médicales sont découvertes simultanément avec des techniques d'apprentissage.

Pour étudier et comparer ces stratégies d'une façon suffisament indépendante des techniques et outils utilisés pour leur implémentation nous proposons quatre méthodes différentes pour l'extraction des entités médicales en anglais (deux méthodes par stratégie) : (i) une méthode à base de règles qui s'appuie sur l'outil de référence MetaMap (Aronson 2001), (ii) une méthode qui extrait les syntagmes nominaux (avec un extracteur robuste, ou « chunker ») puis détecte les entités médicales parmi ces syntagmes par apprentissage supervisé (classifieur SVM), (iii) une méthode qui utilise l'apprentissage supervisé pour déterminer les frontières et les types des entités médicales avec un classifieur CRF et l'encodage B-I-O [2] et (iv) une variante hybride de cette dernière qui combine une méthode statistique et une méthode à base de règles.

Afin d'éviter le biais de l'outil de chunking utilisé pour l'extraction des syntagmes nominaux dans la première stratégie étudiée, nous présentons aussi une étude comparative de la performance de trois outils

2. Le format B-I-O permet de trouver la catégorie ainsi que les frontières des entités en indiquant pour chaque mot s'il correspond au début, à l'intérieur ou l'extérieur de l'entité.

pour l'extraction de syntagmes nominaux à partir de textes médicaux : TreeTagger-chunker, OpenNLP et MetaMap. Cette comparaison permet de sélectionner l'outil le plus performant parmi ces trois outils de référence.

Les différentes approches ont été expérimentées sur deux corpus différents : un premier corpus de textes cliniques (le corpus du challenge international i2b2/VA 2010), et un deuxième corpus de résumés d'articles scientifiques extrait de MEDLINE (le corpus de Berkeley (Rosario et Hearst 2004)).

Enfin, nous faisons une synthèse des deux stratégies étudiées à travers les résultats obtenus par les différentes méthodes sur les deux corpus standards. Cette synthèse permet aussi de mettre en évidence les avantages et inconvénients des méthodes à base de règles, statistiques ou hybrides employées pour l'extraction des entités médicales.

2.2 État de l'art

La reconnaissance des entités nommées, introduite officiellement à la campagne d'évaluation MUC-6 (Grishman et Sundheim 1996), est étudiée depuis une vingtaine d'années en domaine ouvert. On peut distinguer trois types d'approches : (i) les approches linguistiques qui utilisent des listes d'entités nommées et des patrons de reconnaissance écrits manuellement (Poibeau 1999, Elkateb-Gara 2003), (ii) les approches statistiques qui se fondent sur des techniques d'apprentissage à partir de textes annotés (McCallum et Li 2003, Raymond et Wei 2006) et (iii) les approches hybrides qui intègrent les deux premières méthodes (Kosseim et Poibeau 2001, Fourour 2002).

Depuis plus d'une dizaine d'années, des travaux se sont intéressés à la reconnaissance des entités nommées en domaine de spécialité. Dans le domaine biomédical (Rindflesch et al. 2000b) ont développé le système EDGAR qui extrait les informations concernant les médicaments et les gènes liés au cancer à partir de la littérature biomédicale de la base MEDLINE. Le système se fonde sur le metathesaurus et les connaissances lexicales de l'UMLS (Unified Medical Language System) (Bodenreider 2006, Zweigenbaum 2004). Outre la détection de noms de gènes, la détection des noms de protéines a fait l'objet de plusieurs travaux (Liang et Shih 2005, Wang 2007). (Embarek et Ferret 2008) ont utilisé une approche à base de patrons linguistiques et d'entités médicales canoniques pour la reconnaissance de termes médicaux de cinq types. Une autre famille de travaux utilise des outils comme MetaMap (Aronson 2001) qui permettent de reconnaître et de catégoriser les termes médicaux. MetaMap est un outil développé par la NLM (U.S. National Library of Medecine) pour reconnaître les termes médicaux qui désignent des concepts de l'UMLS. MetaMap identifie la plupart des concepts présents dans les titres des articles de la base MEDLINE (Pratt et Yetisgen-Yildiz 2003). Il a été par exemple utilisé par (Shadow et MacDonald 2003) pour extraire des entités médicales à partir de rapports de pathologistes, ces entités pouvant avoir 20 types sémantiques possibles (choisis parmi ceux de l'UMLS). (Meystre et Haug 2005) ont pu obtenir 89,2 % de rappel et 75,3 % de précision avec une approche qui se fonde sur MetaMap et l'algorithme NegEx de détection de négation (Chapman

et al. 2001) pour l'extraction de « problèmes médicaux » (signes, symptômes, diagnostics).

Le domaine médical dispose de plusieurs ressources sémantiques structurées comme le metathesaurus et le réseau sémantique de l'UMLS. L'UMLS est organisé en trois parties (i) le Specialist Lexicon, lexique anglais incluant les termes du domaine ainsi que leurs variations syntaxiques et morphologiques, (ii) le metathesaurus, vocabulaire de plus de deux millions de concepts (un concept « regroupe » des termes synonymes, acronymes et variantes terminologiques) et (iii) le réseau sémantique qui organise les concepts en 135 « types sémantiques » et définit 54 relations entre ces types.

À l'opposé des approches linguistiques qui requièrent plus de connaissances du domaine pour la construction des règles ou patrons, les approches statistiques sont souvent mises en avant pour leur robustesse et leur potentiel de passage à l'échelle. Plusieurs travaux ont utilisé des classifieurs comme les arbres de décision ou les SVM (Isozaki et Kazawa 2002). D'autres se sont basés sur des modèles de Markov comme le modèle de Markov caché ou encore les CRF (He et Kayaalp 2008). La performance des approches fondées sur ces algorithmes supervisés est dépendante de la présence d'un corpus d'entraînement bien annoté et de la conception d'un ensemble pertinent d'attributs.

Les approches hybrides tentent de cumuler les avantages des méthodes à base de règles et des méthodes statistiques tout en éliminant certains de leurs inconvénients (e.g. problème de passage à l'échelle des méthodes à base de règles, performances réduites pour les méthodes statistiques avec des corpus d'entraînement de taille réduite). (Proux et al. 1998) ont construit un système pour la détection des symboles et noms de gènes à partir de textes biomédicaux. Le système traite les mots inconnus avec des règles lexicales afin d'obtenir des catégories candidates qui sont ensuite désambiguïsées en utilisant un modèle de Markov. (Liang et Shih 2005) utilisent à la fois des règles empiriques et une approche statistique pour la reconnaissance des noms de protéines.

Il est aussi important de noter que différents genres de corpus existent dans le domaine biomédical (Zweigenbaum et al. 2001). Parmi les plus récurrents nous pouvons citer les textes cliniques et les articles scientifiques (Friedman et al. 2002). La première catégorie de corpus a intéressé plusieurs travaux (e.g. (Sager et al. 1995), (Meystre et al. 2008)) mais aussi des challenges internationaux comme i2b2 2010 (Uzuner et al. 2011). Les articles scientifiques du domaine biomédical ont aussi fait l'objet de différents travaux (e.g. (Rindflesch et al. 2000a)), en particulier depuis plus de dix ans en génomique (e.g. le challenge BioCreAtIvE (Yeh et al. 2005)).

2.3 Définitions et principes

Nous consacrons cette section à l'introduction de la problématique de reconnaissance des entités médicales, en répondant aux questions suivantes : (i) Qu'est-ce qu'une entité médicale ? et (ii) En quoi consiste le processus de reconnaissance d'une entité médicale ?

2.3.1 Entités médicales

En domaine ouvert, les entités nommées désignent classiquement les noms propres, à savoir les noms de personnes, d'organisations et de lieux (Grishman et Sundheim 1996), mais aussi les dates et les quantités. Certains vont plus loin et cherchent à annoter plus finement leurs corpus, par exemple (Sekine 2004), ou encore (Ehrmann et Jacquet 2006) qui considèrent la reconnaissance d'entités de catégories plus précises (e.g. *professeur* ou *président* plutôt que simplement *personne*).

Nous désignons par « entité médicale » une instance d'un concept médical générique comme « Maladie » (e.g. *l'Alzheimer*) ou « Examen » (e.g. *la laryngoscopie*). Dans le domaine médical, (Khelif et Dieng-Kuntz 2004) ont utilisé le réseau sémantique de l'UMLS comme une ontologie du domaine biomédical et les termes du metathesaurus comme instances possibles de concepts biomédicaux. De même, (Delbecque et al. 2005) ont considéré les types sémantiques du réseau sémantique de l'UMLS comme types d'entités nommées spécifiques au domaine médical (e.g. *Drug*, ou encore *Therapeutic or Preventive Procedure*). Plus récemment, (Embarek et Ferret 2008) se sont intéressés à la reconnaissance de termes médicaux de cinq types (*Maladie*, *Traitement*, *Médicament*, *Examen* et *Symptôme*).

Nos choix. Nous travaillons sur les trois catégories les plus importantes dans le domaine médical, à savoir *Problème* (signes, symptômes, diagnostics, etc.), *Traitement* (y compris médicaments et matériel médical) et *Test* (examens) qui sont aussi les catégories ciblées dans la tâche de reconnaissance d'entités médicales dans le challenge international i2b2 2010 (Uzuner et al. 2011). Le tableau 2.1 présente les types sémantiques de l'UMLS correspondants aux catégories médicales traitées.

Catégorie médicale	**Types sémantiques correspondants dans l'UMLS**
Problème	Virus, Bacterium, Anatomical Abnormality, Congenital Abnormality, Acquired Abnormality, Sign or Symptom, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Cell or Molecular Dysfunction, Injury or Poisoning, Sign or Symptom
Traitement	Medical Device, Drug Delivery Device, Clinical Drug, Steroid, Pharmacologic Substance, Antibiotic, Biomedical or Dental Material, Therapeutic or Preventive Procedure
Test	Laboratory Procedure, Diagnostic Procedure

TABLE 2.1 – *Catégories médicales traitées et types sémantiques correspondants dans l'UMLS*

2.3.2 Principes

La reconnaissance des entités médicales consiste en (i) le repérage des termes médicaux dans les textes (e.g. *beta cell replacement*, *pyogenic liver abscess*, *infection of biliary system*) et (ii) l'identification de la catégorie séman-

tique des termes repérés (e.g. *Maladie*, *Traitement*, *Médicament*, *Examen*). L'exemple suivant montre les résultats de la reconnaissance d'entités médicales dans une phrase extraite d'un texte clinique. Les termes médicaux y sont marqués par les étiquettes *<Treatment>* et *<Disease>*.

<Treatment> Adrenal-sparing surgery </Treatment> is safe and effective , and may become the treatment of choice in patients with <Disease> hereditary phaeochromocytoma </Disease>.

Ces deux étapes amènent à effectuer des choix sur (i) les catégories médicales à traiter (cf. section 2.1) et (ii) les règles de délimitation des frontières des entités médicales dans le texte, telles que :

- inclure ou non les articles (e.g. ***The*** *West Nile Virus* ou juste *West Nile Virus*) et les possessifs (e.g. ***his*** *cancer therapy* ou *cancer therapy*),
- inclure ou non les adjectifs (e.g. ***recurrent*** *or* ***persistent*** *angina*), sachant qu'il est parfois important de conserver les adjectifs pour déterminer correctement l'entité médicale comme par exemple ***Severe Acute Respiratory*** *Syndrome (SARS)*,
- inclure ou non les adverbes (e.g. ***all*** *drugs*),
- inclure ou non les pourcentages, les chiffres (e.g. ***30*** *cancers*) et les doses des médicaments (e.g. *clamoxyl* ***1g***),
- comment annoter les abréviations qui suivent les entités médicales (exp : *Hepatitis A Virus* ***(HAV)***) : annoter les deux ensemble ou chacune à part,
- annoter ou non une entité médicale à l'intérieur d'une autre (e.g. *The BC* ***Cancer*** *Agency*, *Canadian Network For* ***Asthma*** *Care*) (voir par exemple les entités imbriquées de (Grouin et al. 2011)).

Nos choix. Dans ce travail, nous avons effectué les choix suivants : (i) inclure dans les entités médicales : les possessifs, les adjectifs, les adverbes ainsi que les chiffres, (ii) annoter les abréviations séparément et (iii) ne pas annoter une entité médicale qui fait partie d'une autre entité. Ces principes rejoignent les règles qui ont été fixées pour l'annotation du corpus i2b2 (décrit dans la section 2.5.1).

2.4 Méthodes proposées

Dans cette section nous présentons deux stratégies différentes pour la reconnaissance des entités médicales. Nous présentons aussi les méthodes proposées au niveau de chaque stratégie.

2.4.1 Présentation générale

Comme annoncé plus haut, ce travail a principalement deux objectifs. Le premier est d'étudier et comparer deux stratégies différentes pour traiter le problème de reconnaissance des entités : (i) une première reposant sur l'extraction en amont des syntagmes nominaux avec des outils spécialisés, suivie d'une étape de catégorisation et (ii) une deuxième exploitant des techniques d'apprentissage pour déterminer simultanément les frontières et les catégories des entités médicales. La figure 2.1 illustre avec un exemple ces deux stratégies.

Notre second objectif consiste à tester différentes méthodes fondées d'une part sur des règles utilisant les connaissances du domaine (fournies

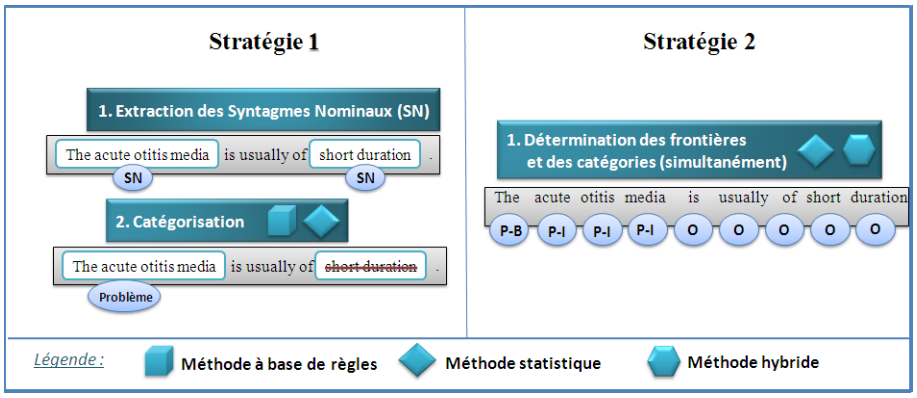


FIGURE 2.1 – *Reconnaissance des entités médicales : Stratégies, étapes et méthodes*

par l'UMLS) et d'autre part sur des techniques d'apprentissage reposant sur deux classifieurs différents. La figure 2.2 présente les quatre méthodes proposées : une méthode à base de règles, deux méthodes différentes à base d'apprentissage supervisé et une méthode hybride. Nous détaillons ces méthodes dans le reste de cette section.

Etapes	Stratégie 1		Stratégie 2	
	Méthode 1 (MetaMapPlus)	Méthode 2 (TT-SVM)	Méthode 3 (CRF-BIO)	Méthode 4 (CRF-BIO-H)
(1) Identification des frontières	Extraction des syntagmes nominaux avec un chunker	Extraction des syntagmes nominaux avec un chunker	▪ Apprentissage supervisé avec un classifieur *CRF* ▪ Utilisation de l'encodage *BIO* ▪ Classification des *mots* ▪ Nbr de classes = *2N+1*	▪ Combinaison des 2 méthodes statistique (CRF-BIO) et à base de règles ▪ Ajout des résultats de MetaMapPlus aux attributs du classifieur CRF
(2) Catégorisation (en N catégories)	▪ Méthode à base de *règles* ▪ Classification des *syntagmes nominaux* ▪ Nbr de classes = *N+1*	▪ Apprentissage supervisé avec un classifieur *SVM* ▪ Classification des *syntagmes nominaux* ▪ Nbr de classes = *N+1*		

FIGURE 2.2 – *Méthodes proposées pour la reconnaissance des entités médicales*

2.4.2 Stratégie 1 ; deux étapes, frontières puis catégories

Extraction des syntagmes nominaux

Malgré l'importance de cette tâche pour l'extraction d'information, il n'y a pas eu beaucoup d'études comparatives des outils disponibles pour le domaine médical. Une étude comparative très récente (Kang et al. 2010) a comparé des outils de segmentation en phrases et en syntagmes nominaux ainsi que l'étiquetage morpho-syntaxique pour le domaine biomédical. Les auteurs ont comparé 6 outils fréquemment utilisés : GATE Chunker, Genia Tagger, Lingpipe, MetaMap, OpenNLP et Yamcha. Les résultats de cette étude montrent que OpenNLP a la meilleure performance : F-mesure de 89,9 % et 95,5 % respectivement pour l'extraction des syntagmes nominaux et verbaux.

Les deux premières méthodes que nous proposons utilisent un chunker pour l'extraction des syntagmes nominaux. Ces syntagmes seront des

entités candidates pour une catégorisation à base de règles pour la première méthode et à base d'apprentissage pour la deuxième. Les performances de ces deux méthodes dépendront de la qualité de l'extraction des syntagmes nominaux et donc de la performance du chunker choisi. Nous avons choisi de comparer un outil médical de référence (MetaMap) et deux outils indépendants du domaine (TreeTagger-chunker [3] et OpenNLP [4]). Nous évaluons ces outils sur un sous-ensemble des syntagmes nominaux (i.e. les syntagmes référant à des entités médicales) de nos deux corpus (décrits dans les sections 2.5.1 et 2.5.2). Nous considérons qu'un syntagme nominal est extrait correctement s'il correspond exactement à une entité médicale annotée. Nous calculons uniquement la valeur du rappel étant donné que les entités retrouvées comportent beaucoup de syntagmes non médicaux (non pertinents pour connaître la performance de ces outils pour la reconnaissance d'entités médicales).

	Corpus de textes cliniques (i2b2)			**Corpus d'articles scientifiques (Berkeley)**		
	MetaMap	**TreeTagger**	**OpenNLP**	**MetaMap**	**TreeTagger**	**OpenNLP**
E1	58115	58115	58115	3371	3371	3371
E2	6532	35314	26862	151	2106	1874
E3	212227	129912	122131	22334	19796	18850
R	11,24 %	**60,76 %**	46,22 %	4,48 %	**62,47 %**	55,59 %

TABLE 2.2 – *Évaluation du rappel de trois chunkers (E1=Entités de référence, E2=Entités correctes, E3=Entités trouvées, R=Rappel=E2/E1)*

Suivant les résultats obtenus et présentés dans le tableau 2.2, TreeTagger-chunker a obtenu les meilleurs résultats et a été choisi pour l'extraction des syntagmes nominaux dans les deux premières méthodes, qui sont présentées ci-dessous.

Catégorisation des syntagmes nominaux : Méthode à base de règles (MetaMapPlus)

En domaine ouvert, ce type de méthode consiste à utiliser des règles écrites manuellement qui exploitent potentiellement des listes de noms (e.g. personnes, organisations, etc). Pour le domaine médical, ce genre d'informations est disponible grâce à des bases de connaissances telles que l'UMLS.

Plusieurs outils se sont intéressés à l'extraction des entités médicales. Un des outils les plus largement utilisés pour cette tâche est MetaMap. L'outil MetaMap (Aronson 2001) permet de segmenter les textes médicaux en phrases et syntagmes nominaux qui correspondent à des termes médicaux. Il identifie les entités médicales et leurs catégories en utilisant le metathesaurus et le réseau sémantique de l'UMLS et fournit potentiellement plusieurs catégories candidates aux entités qu'il retrouve avec des scores de confiance. Plus précisément, ces catégories sont les types du réseau sémantique UMLS jugés comme étant pertinents pour l'entité retrouvée. Le tableau 2.3 montre un exemple de sortie de MetaMap sur une phrase.

3. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
4. http://incubator.apache.org/opennlp

Phrase :	
Adrenal-sparing surgery is safe and effective, and may become the treatment of choice in patients with hereditary phaeochromocytoma.	
Extrait des résultats :	
Phrase : "Adrenal-sparing surgery" `Meta Mapping (825)`: 660 Adrenal [Finding] 589 Sparingly [Intellectual Product,Qualitative Concept] 827 Surgery (Operative Surgical Procedures) [Therapeutic or Preventive Procedure] `Meta Mapping (825)`: 660 Adrenal [Finding] 589 Sparingly [Intellectual Product,Qualitative Concept] 827 Surgery (Surgery specialty) [Biomedical Occupation or Discipline] `Meta Mapping (825)`: 660 Adrenal [Finding] 589 Sparingly [Intellectual Product,Qualitative Concept] 827 surgery (Surgical aspects) [Functional Concept] `Meta Mapping (825)`: 660 Adrenal (Adrenal Glands) [Body Part, Organ, or Organ Component] 589 Sparingly [Intellectual Product,Qualitative Concept] 827 Surgery (Operative Surgical Procedures) [Therapeutic or Preventive Procedure] `Meta Mapping (825)`: 660 Adrenal (Adrenal Glands) [Body Part, Organ, or Organ Component] 589 Sparingly [Intellectual Product,Qualitative Concept] 827 surgery (Surgical aspects) [Functional Concept] (...)	***Phrase : "effective,"*** `Meta Mapping (1000)`: 1000 Effective [Qualitative Concept] ***Phrase : "the treatment"*** `Meta Mapping (1000)`: 1000 Treatment (Administration procedure) [Therapeutic or Preventive Procedure] `Meta Mapping (1000)`: 1000 Treatment (Biomaterial Treatment) [Conceptual Entity] `Meta Mapping (1000)`: 1000 Treatment (Therapeutic procedure) [Therapeutic or Preventive Procedure] `Meta Mapping (1000)`: 1000 Treatment (Treating) [Functional Concept] `Meta Mapping (1000)`: 1000 treatment (therapeutic aspects) [Functional Concept] ***Phrase : "of choice"*** `Meta Mapping (1000)`: 1000 choice (Choice Behavior) [Individual Behavior] ***Phrase : "in patients"*** `Meta Mapping (1000)`: 1000 Patients [Patient or Disabled Group] ***Phrase : "with hereditary phaeochromocytoma."*** `Meta Mapping (888)`: 694 Hereditary [Functional Concept] 861 pheochromocytoma (Benign pheochromocytoma of adrenal gland) [Neoplastic Process] `Meta Mapping (888)`: 694 Hereditary [Functional Concept] 861 PHAEOCHROMOCYTOMA (Pheochromocytoma) [Neoplastic Process]

TABLE 2.3 – *Extrait des résultats de MetaMap pour une phrase donnée. Chaque concept est précédé de son score et suivi [entre crochets] de son type sémantique.*

Cependant l'étude de l'utilisation simple de MetaMap a révélé qu'il présente certains problèmes résiduels, principalement à trois niveaux : (i) la segmentation en syntagmes nominaux n'est pas toujours pertinente et n'est pas du même niveau de performance que d'autres outils connus en TAL, (ii) la détection des entités médicales : MetaMap considère certains mots généraux et certains verbes comme des termes du domaine et (iii) la catégorisation des entités médicales peut rester ambiguë, car MetaMap peut proposer plusieurs concepts pour un même terme ainsi que plusieurs types sémantiques pour un même concept ; donc plusieurs combinaisons « terme-concept-type » sont possibles (cf. tableau 2.3).

Afin de pallier ces problèmes, nous proposons une méthode, nommée MetaMapPlus, qui comprend quatre étapes :

1. Extraire les syntagmes nominaux avec un chunker. Nous utilisons

TreeTagger-chunker qui offre une meilleure segmentation en syntagmes nominaux et permet de diminuer le bruit de la reconnaissance d'entités médicales (cf. tableau 2.2).

2. Filtrer les syntagmes candidats avec une liste de mots vides en amont de MetaMap.
3. Rechercher des termes candidats dans des listes de problèmes médicaux, traitements et tests médicaux obtenus du corpus d'entraînement, de Wikipedia, Health on the Net, et Biomedical Entity Network.
4. Pour les entités candidates qui n'ont pas été détectées dans les listes, déterminer leurs concepts et types sémantiques correspondants dans l'UMLS avec MetaMap, après un filtrage avec (i) une liste des erreurs les plus fréquentes et (ii) la limitation des types sémantiques utilisés par MetaMap afin d'éviter certains concepts trop généraux (e.g. Quantitative Concept, Functional Concept, Qualitative Concept).

Ces améliorations visent à accroître la précision des résultats de MetaMap. Afin d'avoir une première idée sur les modifications proposées et une première évaluation de la méthode MetaMapPlus, nous avons construit un corpus d'évaluation de 20 articles scientifiques anglais variés extraits de PubMedCentral (d'autres évaluations seront présentées dans les sections 2.5.1 et 2.5.2). Nous avons ensuite annoté manuellement les entités médicales correspondant à 16 types sémantiques donnés. Comme il est difficile d'annoter manuellement toutes les entités médicales présentes dans notre corpus, nous avons mesuré uniquement la précision de la reconnaissance d'entités médicales de 16 types sémantiques. La précision dépend de l'exactitude de leurs catégories (types sémantiques) mais aussi de la précision de localisation de ces entités (correcte, avec du bruit, partielle ou fausse). Dans cette évaluation, une erreur liée à la localisation partielle (resp. avec du bruit) d'un terme médical coûte un demi-point, et la précision est calculée selon la formule suivante :

$$Precision = \frac{C + 0.5 \times B}{Ref} \tag{2.1}$$

- C : entités correctes
- B (boundary) : entités avec une catégorie correcte mais une localisation imprécise (partielle ou avec bruit)
- Ref : le nombre total des entités de référence.

Le tableau 2.4 compare la précision obtenue avec notre méthode MetaMapPlus et celle obtenue avec l'utilisation simple de MetaMap sur un sous ensemble de types sémantiques. Les erreurs liées aux types sémantiques sont notées par T, celles liées aux frontières des entités sont notées par B et la précision est notée par P. Notre méthode conduit à une augmentation significative de la précision par rapport à MetaMap (le total indiqué a été calculé sur toutes les occurrences des 16 types sémantiques traités).

Catégorisation des syntagmes nominaux : Méthode statistique (TT-SVM)

La deuxième méthode que nous avons mise en place et testée consiste à extraire les syntagmes nominaux à partir du texte, puis à utiliser un

	MetaMap			MetaMapPlus		
	T	B	P	T	B	P
Disease Or Syndrome	9.09%	52.27%	64.77%	9.81%	26.48%	76.94%
Injury or poisoning	33.33%	34.84%	49.24%	26.19%	35.71%	55.95%
Total	30.24%	34.56%	**54.62%**	12.23%	27.10%	**74.21%**

TABLE 2.4 – *Précision de la reconnaissance des entités médicales de 16 types sémantiques sur un premier corpus de test*

classifieur pour déterminer les syntagmes qui correspondent à des entités médicales et leurs catégories (e.g. Traitement, Test, Maladie). Il s'agit d'une classification des syntagmes nominaux en $n+1$ classes (où n est le nombre des catégories d'entités médicales).

Nous avons choisi d'utiliser un classifieur de type SVM (Machine à vecteurs de support). (Ekbal et Bandyopadhyay 2010) affirment que les classifieurs SVM ont un certain avantage sur les algorithmes d'apprentissage statistique conventionnels, tels que les arbres de décision, les modèles de Markov cachés, les modèles à entropie maximum et cela sur deux aspects : (i) les SVM ont une forte capacité de généralisation indépendante de la dimension des vecteurs d'attributs et (ii) les SVM peuvent effectuer leur entraînement avec toutes les combinaisons des attributs choisis sans augmenter la complexité algorithmique, par l'introduction d'une fonction noyau. La bibliothèque LIBSVM (Chang et Lin 2001a) a été utilisée pour la mise en place du classifieur. La sélection du modèle et plus particulièrement la recherche du paramètre relatif au noyau (gamma) et du paramètre de régularisation (C) a été effectuée automatiquement avec un script disponible en complément de cette bibliothèque. (Keerthi et Sundararajan 2007) ont effectué une expérimentation qui montre que les classifieurs SVM structurés et les CRF sont assez proches en termes de performance si les mêmes attributs sont utilisés pour la classification.

Nous avons sélectionné des attributs lexicaux, orthographiques et morpho-syntaxiques. Le table 2.5 décrit ces attributs.

2.4.3 Stratégie 2 ; une seule étape, frontières et catégories conjointement

Méthode statistique pour la détermination des frontières et des catégories (CRF-BIO)

Cette méthode comporte une seule étape : déterminer les frontières et les types des entités médicales. Le passage par un chunker n'est plus indispensable. Pour ce faire, nous utilisons le format BIO : B (beginning), I (inside) et O (outside) qui permet de représenter un marquage de segments de texte (les entités) par un étiquetage individuel des mots. Une entité de type Problème formée de plusieurs mots (par exemple, *an L5 metastatis*) voit son premier mot (*an*) étiqueté B-P (début de Problème) et ses autres mots (*L5* et *metastatis*) étiquetés I-P (dans un Problème). Une entité de type Problème comprenant un seul mot est étiquetée B-P. Les mots hors entité sont étiquetés O (voir la figure 2.3). Si nous avons n catégories (e.g. Problème, Traitement, Test), nous avons alors n classes de type B-, n classes de type I- (e.g. les classes P-B et P-I associées à la catégorie

Attributs lexicaux	- Mots du syntagme lui-même - Nombre de mots du syntagme - Lemmes des mots du syntagme (en utilisant TreeTagger) - Trois mots avant le syntagme et leurs lemmes - Trois mots après le syntagme et leurs lemmes
Exemples d'attributs orthographiques	- Le premier mot, un mot du syntagme ou tous les mots sont capitalisés - Le premier mot, un mot ou tous les mots sont en majuscules - Le premier mot, un mot ou tous les mots sont en minuscules - Le syntagme contient une abréviation - Le syntagme contient un seul caractère en majuscule, un chiffre, ou un signe spécial (e.g. –, +, & ou /).
Attributs morpho-syntaxiques	les catégories morpho-syntaxiques des mots du syntagme, des trois mots avant le syntagme et des trois mots après le syntagme, en utilisant l'outil TreeTagger.

TABLE 2.5 – *Les attributs lexicaux, orthographiques et morpho-syntaxiques utilisés avec le classifieur SVM pour la classification des syntagmes nominaux*

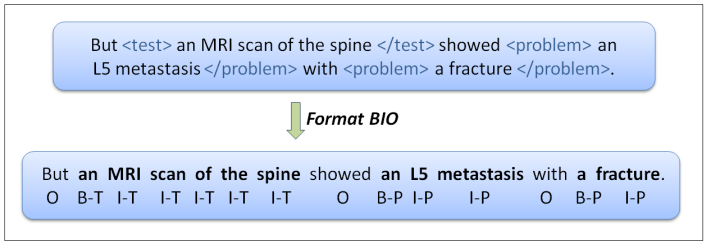


FIGURE 2.3 – *Exemple de phrase au format BIO (T = Test, P = Problème)*

Problème) et une classe de type *O*. Le problème est alors un problème de classification de chaque mot (et non plus chaque syntagme nominal) en $2n + 1$ classes possibles (où n est le nombre de catégories médicales).

Les mots d'une phrase forment une séquence, et la décision sur la catégorie d'un mot peut être influencée par la décision portant sur la catégorie du mot précédent. Cette dépendance est prise en compte dans les modèles séquentiels comme les modèles de Markov cachés (HMM) ou les champs aléatoires conditionnels (Conditional Random Fields, ou CRF). Contrairement aux HMM, l'apprentissage dans les CRF maximise la probabilité conditionnelle des classes relativement aux observations plutôt que leur probabilité conjointe. Cela leur permet d'utiliser un nombre quelconque d'attributs concernant des aspects quelconques de la séquence de mots d'entrée. Ces propriétés expliquent l'intérêt des CRF pour un certain nombre de tâches de traitement automatique des langues, comme l'étiquetage morphosyntaxique, la détection de syntagmes non récursifs (chunks)

ou la reconnaissance d'entités nommées (voir (Tellier et Tommasi 2010) pour une revue).

Nous avons donc testé la détection d'entités médicales avec un CRF. Nous avons utilisé pour cela l'outil CRF++ [5], qui permet de décrire facilement les attributs à utiliser à travers des patrons d'attributs (*feature templates*). Ces attributs sont listés ci-dessous. Nous avons aussi réglé CRF++ pour qu'il utilise la dépendance entre catégories successives (instruction `B` du fichier d'attributs).

Pour chaque mot, nous utilisons un ensemble d'attributs lexicaux, orthographiques et morpho-syntaxiques. Le tableau 2.6 décrit ces attributs.

Attributs lexicaux	- Le mot (M) lui-même, deux mots avant et trois mots après - Les lemmes de ces mots (en utilisant TreeTagger).
Exemples d'attributs orthographiques	- Le mot contient un signe spécial (e.g. +, –, &, /) - Le mot est un chiffre, alphabétique, un signe de ponctuation ou un symbole - Le mot est en majuscules, capitalisé, en minuscules (AA, Aa, aa) - Préfixes de différentes longueurs (de 1 à 4) - Suffixes de différentes longueurs (de 1 à 4)
Attributs morpho-syntaxiques	Les catégories morpho-syntaxiques du mot M lui-même, des deux mots avant et des trois mots après M, en utilisant TreeTagger.
Autres attributs	Premier verbe suivant, premier nom suivant, longueur du mot par rapport à un seuil, etc.

TABLE 2.6 – *Les attributs lexicaux, orthographiques et morpho-syntaxiques utilisés avec le classifieur CRF pour la classification des mots*

Méthode hybride pour la détermination des frontières et des catégories (CRF-BIO-H)

Cette méthode consiste à utiliser les résultats de la méthode à base de règles (MetaMapPlus) comme attributs pour la méthode statistique (i.e. CRF-BIO). Plusieurs choix sont possibles : (i) l'utilisation de la catégorie sémantique associée au mot en appliquant MetaMapPlus, (ii) l'utilisation du type sémantique associé de l'UMLS (en appliquant MetaMapPlus sur le mot) et (iii) la transformation des résultats de MetaMapPlus au format BIO et en les considérant comme attributs pour le classifieur CRF. À chaque mot est associée une classe de type B-problem, I-problem, B-treatment, I-treatment, B-test et I-test (obtenues grâce aux résultats de MetaMapPlus sur le corpus i2b2). Après des tests, nous avons choisi le troisième choix.

5. `http://crfpp.sourceforge.net/`

2.5 Expérimentations

Nous avons mené des expériences de reconnaissance d'entités médicales dans des textes cliniques en anglais (section 2.5.1). Pour tester la robustesse des systèmes créés, nous les avons également appliqués à des résumés d'articles scientifiques extraits de MEDLINE (section 2.5.2). Tous les résultats présentés dans cette section ont été calculés sur la base de la correspondance stricte des frontières.

2.5.1 Expériences sur le corpus i2b2 de textes cliniques

Dans cette section, nous présentons le corpus i2b2 et les résultats des différentes méthodes.

Corpus i2b2 de textes cliniques

Le corpus i2b2 a été construit dans le cadre du challenge i2b2 2010[6]. Ce corpus comporte des entités médicales de trois catégories (Problème, Traitement et Test) annotées dans 76 165 phrases (663 476 mots au total), pour une moyenne de 8,7 mots par phrase. L'exemple suivant montre une phrase annotée du corpus i2b2.

<problem> CAD </problem> s/p <treatment> 3v-CABG </treatment> 2003 and subsequent <treatment> stenting </treatment> of <treatment> SVG </treatment> and LIMA.

L'annotation manuelle du corpus a été conduite suivant un guide d'annotation. Le tableau 2.7 présente l'accord inter-annotateurs pour chaque catégorie médicale.

i2b2	**Accord inter-annotateurs (frontières strictes)**	**Accord inter-annotateurs (frontières non strictes)**
Problème	0,84	0,91
Traitement	0,83	0,90
Test	0,83	0,88
Total	**0,83**	**0,90**

Table 2.7 – *Corpus i2b2 : Accord inter-annotateurs pour chaque catégorie médicale*

Le tableau 2.8 présente le nombre de phrases d'entraînement et de test.

i2b2	**Phrases**	**Mots**
Corpus d'entraînement	31 238	267 304
Corpus de test	44 927	396 172

Table 2.8 – *Nombre de phrases et de mots d'entraînement et de test*

6. http://www.i2b2.org

Configurations et résultats

Nous avons expérimenté les cinq configurations suivantes :

1. MM : MetaMap
2. MM+ : MetaMapPlus
3. TT-SVM : Catégorisation des syntagmes nominaux avec SVM
4. CRF-BIO : Apprentissage format BIO avec CRF
5. CRF-BIO-H : Méthode hybride (CRF-BIO utilisant des attributs sémantiques construits à partir des résultats de MM+)

Le tableau 2.9 présente les résultats obtenus sur le corpus i2b2 avec les mesures classiques de rappel, précision et F-mesure[7]. Le tableau 2.10 détaille les résultats pour chaque catégorie médicale (i.e. Problème, Traitement et Test). Nous avons mis en gras la meilleure performance totale (tableau 2.9) et pour chaque catégorie (tableau 2.10).

Configuration	**Rappel**	**Précision**	**F-mesure**
MM	15,52	16,10	15,80
MM+	48,68	56,46	52,28
TT-SVM	43,65	47,16	45,33
CRF-BIO	70,15	83,31	76,17
CRF-BIO-H	**71,92**	**83,83**	**77,42**

TABLE 2.9 – *Résultats de chaque configuration sur le corpus i2b2 (frontières strictes)*

Configurations	Catégorie	**Précision**	**Rappel**	**F-mesure**
MM+	Problème	60,84	53,04	56,67
	Traitement	51,99	61,93	56,53
	Test	56,67	28,48	37,91
TT-SVM	Problème	48,25	43,16	45,56
	Traitement	42,45	50,86	46,28
	Test	57,37	35,76	44,06
CRF-BIO-H	Problème	**82,05**	**73,14**	**77,34**
	Traitement	**83,18**	**73,33**	**77,95**
	Test	**87,50**	**68,69**	**77,00**

TABLE 2.10 – *Résultats pour chaque catégorie sémantique sur le corpus i2b2*

2.5.2 Expériences sur le corpus Berkeley d'articles scientifiques

Cette section présente les résultats obtenus avec des expériences supplémentaires effectuées sur un corpus de résumés scientifiques extraits de MEDLINE.

7. Lors du challenge i2b2 2010, la meilleure F-mesure (85,23%) a été obtenue par de Bruijn et al. (2011)

Corpus de Berkeley

Le corpus de Berkeley (Rosario et Hearst 2004) est construit à partir de titres et résumés d'articles scientifiques de MEDLINE. L'objectif du corpus était l'étude des relations sémantiques entre les entités médicales de type Maladie et Traitement, qui sont : cures, prevents et side effect. Dans ce travail, nous exploitons les annotations des entités médicales uniquement.

Le corpus contient deux catégories d'entités médicales : Maladies (1660 entités) et Traitements (1179 entités) dans 3654 phrases (74754 mots), donc en moyenne 20,45 mots par phrase. L'exemple suivant montre une phrase annotée du corpus de Berkeley.

We investigated the hypothesis that <TREAT PREV> an antichlamydial macrolide antibiotic , roxithromycin </TREAT PREV> , can prevent or reduce recurrent major ischaemic events in patients with <DIS PREV> unstable angina </DIS PREV>. Plusieurs étiquettes sont utilisées pour chaque catégorie suivant la relation qui les lie. Par exemple, il y a 7 étiquettes différentes pour les maladies (ou problèmes médicaux) : <DIS>, <DIS_NO>, <DIS_VAG>, <DISONLY>, <DIS_PREV>, <DIS_SIDE_EFF>, <DIS_EFF>. Nous avons uniformisé ces étiquettes dans une étape de prétraitement avant les expérimentations finales.

Le tableau 2.11 présente le nombre de phrases d'entraînement et de test.

Berkeley	**Phrases**	**Mots**
Corpus d'entraînement	1462	36642
Corpus de test	2193	38112

TABLE 2.11 – *Nombre de phrases et de mots d'entraînement et de test*

Résultats

Nous avons testé la méthode fondée sur MetaMap (MM+) sur le corpus de Berkeley. Le tableau 2.12 présente les résultats obtenus avec les mesures classiques de précision, rappel et F-mesure.

		Précision	**Rappel**	**F-mesure**
MM	Maladie	5,32	7,63	6,27
	Traitement	6,37	18,84	9,52
	Total	5,35	12,34	7,46
MM+	Maladie	**34,47**	**44,97**	**39,02**
	Traitement	**18,11**	**39,36**	**24,81**
	Total	**23,43**	**42,47**	**30,20**

TABLE 2.12 – *Résultats de la méthode à base de règles (MM+) sur le corpus de Berkeley*

Nous avons aussi testé une méthode statistique sur ce corpus (CRF-BIO). Nous avons construit trois modèles différents pour le classifieur CRF. Un premier modèle est construit à partir du corpus d'entraînement de Berkeley, un deuxième à partir du corpus de i2b2 et un troisième à partir de ces deux corpus d'entraînement (cf. figure 2.4).

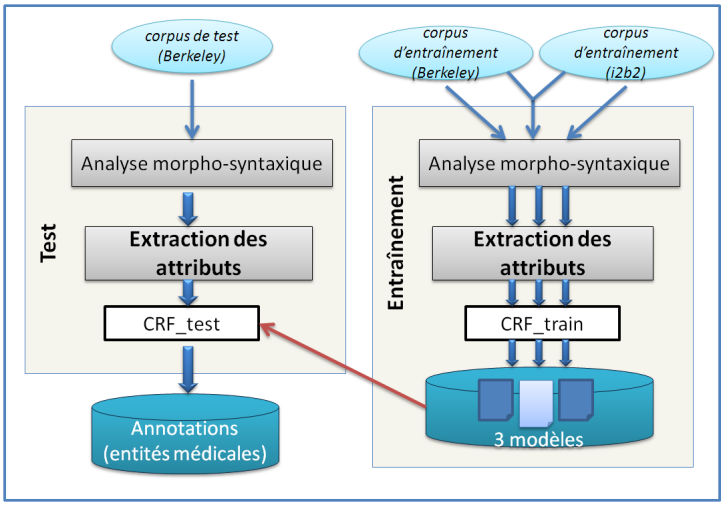


FIGURE 2.4 – *Reconnaissance des entités médicales à partir du corpus de Berkeley : La méthode CRF-BIO et les trois modèles testés*

Nous avons obtenu les meilleurs résultats avec le troisième modèle construit à partir des deux corpus d'entraînement. Le tableau 2.13 décrit les résultats obtenus avec les trois modèles en utilisant les valeurs classiques de rappel, précision et F-mesure. Nous avons obtenu 34,37 % de F-mesure (40,3 % pour Problème et 26,5 % pour Traitement). Ces résultats sont obtenus avec un sous-ensemble d'attributs avec lequel nous avons obtenu 76,17 % sur le corpus i2b2. Ces attributs sont : les mots, les lemmes, les catégories morphosyntaxiques, des attributs orthographiques, les suffixes et préfixes (cf. ensemble A4, tableau 2.14).

Modèles testés	**Rappel**	**Précision**	**F-mesure**
Modèle 1 (Berkeley)	14,64	**53,29**	22,97
Modèle 2 (i2b2)	**36,09**	26,38	30,48
Modèle 3 (Berkeley+i2b2)	32,13	36,94	**34,37**

TABLE 2.13 – *Résultats de la méthode CRF-BIO sur le corpus de Berkeley*

2.6 Discussion

Nous avons présenté quatre méthodes différentes pour la reconnaissance des entités médicales. Dans cette section nous analysons les différences en terme de résultats par rapport aux corpus et aux méthodes.

2.6.1 Expérimentations sur les textes cliniques

Nous avons testé quatre méthodes différentes pour la reconnaissance des entités médicales à partir du corpus i2b2 : MetaMapPlus, TT-SVM, CRF-BIO et CRF-BIO-H. La figure 2.5 compare les résultats de ces méthodes.

À travers nos expérimentations, nous n'avons pas obtenu de bons résultats en appliquant directement l'outil de référence MetaMap, spécialisé

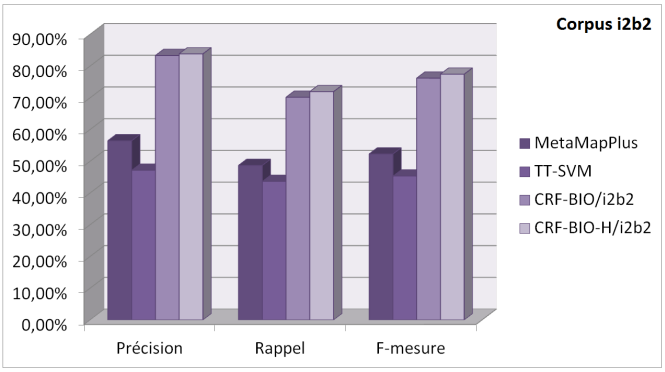


FIGURE 2.5 – *Résultats sur le corpus i2b2*

dans la détection des termes médicaux, sur le corpus i2b2 (15,80 % de F-mesure). Cela était dû essentiellement à deux problèmes :

- le problème de délimitation des frontières des entités : e.g. ***no*** *pericardial effusion* **.** au lieu de *pericardial effusion* et **(** *Warfarin* au lieu de *Warfarin* ;
- les termes généraux considérés par MetaMap comme entités médicales pour des problèmes de polysémie (e.g. *ten*) ou parce que le terme peut être utilisé dans le domaine médical (e.g. *case*, *form*).

Nous avons pu améliorer les résultats de MetaMap en utilisant entre autres (i) un chunker externe qui a été choisi après une étude comparative et (ii) un antidictionnaire contenant les erreurs les plus fréquentes de MetaMap observées lors de tests précédents. Les résultats finals restent limités à cause de la performance du procédé de chunking (comme c'est le cas pour la méthode TT-SVM). La méthode MetaMapPlus a cependant permis de typer correctement 52,28 % des entités sur les 60,76 % d'entités extraites avec de bonnes frontières.

La méthode statistique CRF-BIO détermine les frontières et les catégories des entités médicales simultanément par apprentissage supervisé sans avoir recours à un chunker. Les résultats de cette méthode sur le corpus i2b2 sont encore meilleurs que ceux obtenus avec MetaMapPlus. En effet, les méthodes statistiques dépendent fortement du nombre de données annotées disponibles (ce nombre est important pour le corpus i2b2) et aussi de l'ensemble d'attributs utilisé. Nous présentons l'apport de chaque catégorie d'attributs avec la méthode CRF-BIO dans le tableau 2.14.

Attributs	**Rappel**	**Précision**	**F-mesure**
A1 : Mots/Lemmes/POS	62,81	82.25	71.23
A2 : A1 + attributs orthographiques	63,72	82,19	71,78
A3 : A2 + suffixes	67,91	82,89	74,65
A4 : A3 + préfixes	70,15	83,31	76,17
A5 : A4 + autres attributs	70,22	83,28	76,19

TABLE 2.14 – *Apport de chaque classe d'attributs : méthode CRF-BIO sur le corpus i2b2*

Nous avons essayé de combiner les résultats des deux méthodes à base de règles (MetaMapPlus) et statistique. Plusieurs tests d'attributs sémantiques ont été effectués avec la méthode CRF-BIO. Par exemple, l'utilisa-

tion de la catégorie sémantique associée au mot en appliquant MetaMapPlus diminue les résultats de 76,17 % à 76,01 %. Le type sémantique associé de l'UMLS (en appliquant MetaMapPlus sur le mot) diminue la F-mesure de 76,17 % à 73,55 %. Ceci peut s'expliquer par le nombre de classes (associées aux types sémantiques) qui devient plus important mais aussi par la performance réduite de MetaMap s'il est appliqué au niveau du mot et non au niveau du syntagme ou de la phrase. La meilleure solution a été obtenue en transformant les résultats de MetaMapPlus au format BIO et en les considérant comme attributs pour le classifieur CRF. À chaque mot est associée une classe de type B-problem, I-problem, B-treatment, I-treatment, B-test et I-test (obtenues grâce aux résultats de MetaMapPlus sur le corpus i2b2). Avec ces attributs sémantiques, nous avons pu passer d'une F-mesure de 76,19 % à 77,42 %. Le tableau 2.15 présente l'apport des attributs sémantiques.

Attributs	**Rappel**	**Précision**	**F-mesure**
A5 : Mots, lemmes, POS, attributs orthographiques, suffixes, préfixes et autres	70,22	83,28	76,19
A6 : A5 + attributs sémantiques	**71,92**	**83,83**	**77,42**

TABLE 2.15 – *Apport des attributs sémantiques : méthode CRF-BIO-H sur le corpus i2b2*

2.6.2 Expérimentations sur les articles scientifiques

La figure 2.6 compare les résultats obtenus par les différentes méthodes sur le corpus de Berkeley. Une comparaison avec la figure 2.5 montre que

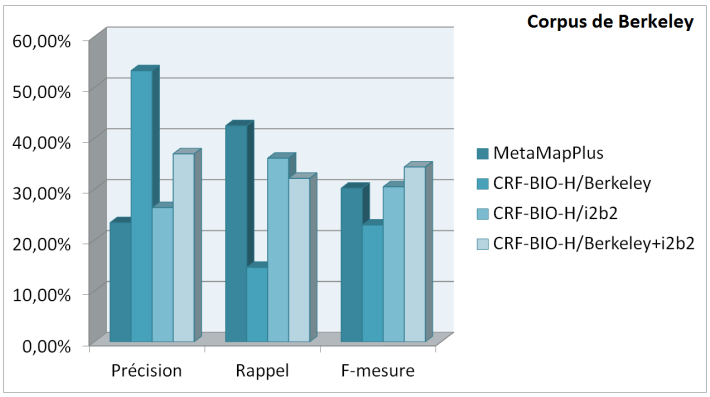


FIGURE 2.6 – *Résultats sur le corpus de Berkeley*

les résultats obtenus pour les deux corpus ne sont pas du même niveau de performance. Deux caractéristiques entrent en jeu :

1. *Les genres différents des deux corpus :* le corpus de i2b2 2010 a un nombre moyen de mots par phrase de 8,7 alors que le corpus de Berkeley a un nombre moyen de mots par phrases de 20,45. Aussi, le corpus de i2b2 utilise un vocabulaire assez spécifique, comme les

abréviations conventionnelles de termes médicaux (e.g. *k/p* pour *kidney pancreas* et *d&c* pour *dilation and curettage*), mais aussi des abréviations de mots généraux (e.g. *w/o* pour *without* et *y/o* pour *year old*). Ceci explique pourquoi nous n'avons pas obtenu de bons résultats en appliquant la méthode CRF-BIO-H entraînée sur le corpus i2b2 : 30,48 % de F-mesure, avec 26,38 % de précision et 36,09 % de rappel (la meilleure valeur obtenue parmi les trois modèles, grâce à la taille importante du corpus d'entraînement de i2b2).

2. *La qualité d'annotation des deux corpus :* le corpus de i2b2 a été annoté avec rigueur étant donné son contexte de challenge et l'utilisation, par les annotateurs, d'un guide qui explicite les principes d'annotation avec des exemples. Le corpus de Berkeley a été annoté d'une manière moins rigoureuse. En effet, aucune règle explicite n'a été énoncée. Par exemple : les articles étaient parfois considérés comme faisant partie de l'entité médicale et d'autres fois exclus de l'entité. Notre évaluation sur un échantillon aléatoire de 200 entités médicales montre un taux d'erreur à l'annotation de 20 %. La qualité d'annotation du corpus de Berkeley et la taille moyenne voire petite du corpus d'entraînement expliquent les résultats obtenus par la méthode CRF-BIO-H entraînée sur le corpus de Berkeley.

2.6.3 Robustesse vs. Portabilité

Les méthodes à base de règles. Elles ont l'avantage d'être reproductibles sur tous types de corpus sans étape de préparation ou apprentissage. Cependant, leur dépendance aux connaissances utilisées fait que leurs performances ne sont pas du même niveau que les processus d'apprentissage, car leurs connaissances restent relativement figées par rapport à la vitesse avec laquelle les approches d'apprentissage permettent d'acquérir de nouveaux critères d'extraction et de catégorisation. Ce type de méthode a aussi l'inconvénient d'être coûteux à mettre en place pour obtenir une couverture satisfaisante (un bon rappel). Leur avantage est cependant d'avoir un consensus sémantique sur les informations extraites (e.g. réseau sémantique de l'UMLS) qui permet potentiellement un traitement plus sophistiqué de l'information extraite. Il est aussi important de noter que des ressources sémantiques (e.g. listes de termes médicaux, lexiques) sont souvent utilisées dans les processus d'apprentissage ce qui rend le développement de telles ressources avantageux à la fois pour les méthodes à base de règles et pour les méthodes statistiques. Inversement, des méthodes statistiques peuvent aussi servir à améliorer les méthodes à base de règles en permettant d'apprendre automatiquement de nouvelles règles d'extraction ou patrons linguistiques pertinents (Hobbs et Riloff 2010).

Les méthodes statistiques. Avec des méthodes à base d'apprentissage supervisé nous avons obtenu de bons résultats sur le corpus i2b2. Mais ce n'est pas le cas pour le corpus de Berkeley. Ces méthodes, bien qu'elles puissent être très robustes, présentent deux inconvénients non négligeables :

1. La dépendance aux données annotées disponibles (cf. résultats de

BIO-CRF-H entraîné et testé sur le corpus de Berkeley), ce qui constitue un obstacle à l'utilisation de ce type de méthode pour des tâches et domaines où l'on ne dispose pas de corpus annotés, d'autant plus que la constitution de ces corpus est une tâche coûteuse.

2. Le problème de portabilité sur des corpus différents de ceux utilisés en entraînement (cf. résultats de BIO-CRF-H entraîné sur i2b2 et testé sur Berkeley) : la dégradation des performances de ces méthodes, appliquées sur des corpus ayant des caractéristiques différentes de ceux utilisés pour l'entraînement, constitue un grand inconvénient pour leur passage à l'échelle.

Cependant, pour un contexte ou un corpus précis, si un bon nombre de données annotées est disponible, les méthodes statistiques peuvent être très robustes et offrir la meilleure solution (cf. 77,42 % de F-mesure obtenue avec la méthode CRF-BIO-H sur le corpus i2b2 vs. 52,28 % obtenue par la méthode MetaMapPlus). Elles offrent en plus la possibilité de découvrir des règles de classification potentiellement indétectables pour l'expert humain. Cela a été démontré dans notre étude où une méthode statistique entraînée sur un corpus médical a permis de délimiter les entités médicales dans les textes plus efficacement qu'un outil spécialisé reposant sur des connaissances et des règles du domaine (MetaMap) et plus efficacement qu'un chunker généraliste exploitant des règles d'extraction linguistiques.

Conclusion du chapitre

Nous avons étudié quatre approches pour la reconnaissance d'entités médicales dans le cadre de deux stratégies différentes pour leur reconnaissance. Les résultats obtenus montrent que l'utilisation d'un chunker pour l'extraction des entités limite la performance finale même si la catégorisation des entités médicales se fait de façon efficace. L'utilisation de procédés d'apprentissage pour l'extraction et la catégorisation simultanées des entités a permis de contourner cette limite. La meilleure performance a ainsi été obtenue par le classifieur CRF en exploitant le format BIO et des attributs lexicaux, orthographiques et morpho-syntaxiques. Nous avons aussi présenté une méthode hybride qui a permis d'améliorer encore ces performances en exploitant des connaissances sémantiques du domaine à travers la méthode à base de règles MetaMapPlus. Les résultats des méthodes statistiques sur le premier corpus de test confirment la robustesse de ce type de méthode. Cependant les résultats obtenus sur le deuxième corpus médical, de type différent, mettent en évidence deux inconvénients des méthodes statistiques, à savoir la dépendance aux données annotées et leur portabilité limitée lorsqu'elles sont appliquées à des corpus ayant des caractéristiques différentes de ceux utilisés pour l'entraînement.

Comme perspectives, nous envisageons (i) d'améliorer la précision de la méthode MetaMapPlus en incluant un module d'apprentissage en amont de l'outil MetaMap, pour classifier les syntagmes nominaux en entités médicales ou entités non médicales et (ii) d'élargir la liste des catégories médicales traitées. Aussi, nous envisageons de travailler sur la reconnaissance des entités médicales dans des corpus français. L'inexis-

tence de corpus médicaux annotés en français constitue un obstacle pour l'apprentissage statistique. D'un autre côté, développer des méthodes à base de règles ou annoter manuellement des corpus pour l'apprentissage sont deux approches coûteuses en temps. Nous envisageons d'exploiter les méthodes MetaMapPlus et CRF-BIO-H pour l'extraction automatique d'entités médicales à partir de textes en français. L'idée étant d'utiliser d'une part un corpus médical parallèle (anglais/français) et d'autre part les alignements au niveau des mots pour projeter les annotations de la partie anglaise dans le corpus français (cf. chapitre 4).

3 Extraction de relations sémantiques à partir de textes médicaux en anglais

Sommaire

La tâche d'extraction de relations sémantiques à partir de documents textuels a été abordée pour différents objectifs (questions-réponses, peuplement d'ontologies, etc.) et avec diverses techniques (patrons ou règles, apprentissage supervisé, semi-supervisé ou non supervisé, etc.). Ces différentes méthodes d'extraction d'information présentent chacunes des avantages et des inconvénients (comme cela a été mentionné, en partie, dans le chapitre précédent). Combiner certaines de ces méthodes peut être une solution pour pallier les inconvénients de certaines d'entre elles. Reste à trouver une méthode efficace pour combiner deux approches différentes ou plus.

Dans ce chapitre, nous abordons l'extraction de relations sémantiques à partir de textes médicaux. Nous proposons une approche qui se fonde

sur deux techniques différentes pour extraire les relations ciblées : (i) des patrons de relation basés sur l'expertise humaine et (ii) l'apprentissage supervisé avec un classifieur SVM. L'approche proposée profite des deux techniques, en se basant plus sur des patrons écrits manuellement quand peu d'exemples de relations sont disponibles pour l'apprentissage et plus sur les résultats du classifieur quand un nombre suffisant d'exemples est disponible. Le chapitre commence par un état de l'art sur l'extraction de relations sémantiques (section 3.2). Nous présentons ensuite les différentes techniques que nous avons développées pour cette tâche ainsi que leurs évaluations. La méthode à base de patrons sera présentée dans la section 3.4, la méthode statistique dans la section 3.5 et la section 3.6 sera dédiée à la description de la méthode hybride.

Les travaux d'extraction de relations présentés dans ce chapitre s'appuient sur les articles (Ben Abacha et Zweigenbaum 2011a;b;d).

3.1 Introduction

Les systèmes de questions-réponses doivent être capables d'interpréter correctement (i) les questions posées et (ii) les textes desquels les réponses vont être extraites. Une interprétation efficace de ces deux éléments demande une analyse profonde de leur sémantique. Plusieurs travaux se sont intéressés à la problématique d'analyse sémantique de textes médicaux. Certains ont proposé des solutions pour la reconnaissance des entités médicales moyennant des ressources lexicales/sémantiques du domaine (e.g.(Shadow et MacDonald 2003, Delbecque et al. 2005)). Un plus petit groupe d'approches s'est intéressé à la tâche plus complexe d'extraction de relations entre les entités médicales (e.g.(Embarek et Ferret 2008)). Dans ce contexte, une relation est un lien sémantique entre deux entités médicales données (e.g. dans la phrase « Hypoprothrombinemia should be treated with Vitamin K, intravenously » l'entité médicale "Vitamin K" est lié à l'entité médicale "Hypoprothrombinemia" par une relation de traitement). La complexité de la tâche d'extraction de relations entre entités médicales réside aussi bien dans les difficultés linguistiques connues en domaine ouvert que dans les particularités spécifiques au domaine médical.

Nous proposons une approche hybride pour la détection des relations sémantiques dans les textes médicaux. Cette approche combine : (i) une méthode à base de patrons et (ii) une méthode statistique qui se base sur un classifieur SVM et qui exploite entre autres des ressources sémantiques. Leur fusion se fait en fonction d'un score de confiance associé aux résultats des deux méthodes.

3.2 Etat de l'art sur l'extraction de relations

Extraction de relations en domaine ouvert

En domaine ouvert, l'extraction des relations sémantiques entre entités dans un corpus textuel a fait l'objet de plusieurs approches de différentes catégories.

Un premier type d'approches, dit statistique, se base sur la co-occurrence de termes spécifiques (Hindle 1990) et/ou des techniques d'apprentissage automatique (Wang et al. 2006) pour l'extraction de relations. L'apprentissage automatique consiste dans ce cas à (i) définir un ensemble d'attributs pertinents pour la détection de la relation (e.g. types des entités médicales sources et cibles, mots entre les deux entités médicales et leurs catégories morpho-syntaxique) (ii) entraîner un classifieur sur un corpus d'entraînement ou des relations ont été annotées et (iii) utiliser le modèle ou les règles de classification apprises grâce à ce corpus pour extraire les relations à partir d'autres corpus, dits corpus de test.

Une deuxième catégorie d'approches, dite linguistique, se base sur des patrons pour extraire les relations (Hearst 1992). Un patron peut être vu comme un modèle de phrase qui identifie une forme particulière d'expression de la relation à extraire. Le processus d'extraction, communément appelé "pattern matching", consiste alors à rechercher des correspondances entre les patrons et les phrases du corpus ciblé pour l'extraction.

Une troisième catégorie d'approches, dite hybride, combine les deux premières (e.g. (Suchanek et al. 2006)) pour profiter de leurs avantages respectifs (e.g. précision de l'extraction pour les approches à base de patron, meilleure couverture ou rappel pour les approches d'apprentissage automatique) et éviter certains de leur inconvénients. Cette combinaison peut être effectuée, par exemple, en prenant en compte la correspondance phrase/patron comme étant un attribut d'apprentissage ou, plus simplement, en intégrant les relations extraites par chacune des approches séparément (e.g. sélectionner l'intersection des résultats de chaque approche).

Extraction de relations dans le domaine biomédical

Dans le domaine médical, les mêmes approches existent. Stapley et Benoit (Stapley et Benoit 2000) se sont intéressés à la détection des relations entre gènes en se basant sur des mesures statistiques de cooccurrence entre mots. D'autres approches utilisent des méthodes de forte précision, à base de règles ou patrons écrits manuellement. Cimino et Barnett (Cimino et Barnett 1993) ont utilisé des patrons pour extraire des relations à partir des titres d'articles de Medline. Ils ont exploité les descripteurs MeSH associés à ces articles dans Medline et la cooccurrence de termes cibles dans un même titre pour générer des règles d'extraction de relations sémantiques. Khoo et al. (Khoo et al. 2000) ont abordé l'extraction de relations causales depuis des résumés d'articles médicaux en alignant des patrons de graphe avec des arbres de dépendance syntaxique. Embarek et Ferret (Embarek et Ferret 2008) ont proposé une approche pour l'extraction de quatre relations (*détecte*, *traite*, *signe* et *soigne*) entre cinq types d'entités médicales. L'extraction de ces relations se base sur des patrons construits automatiquement en utilisant une distance d'édition entre deux phrases et un algorithme d'alignement de parties de phrases qui prend en compte plusieurs niveaux d'information sur les mots. SemRep (Rindflesch et al. 2000a) est un outil qui permet d'identifier les relations sémantiques dans des textes biomédicaux en utilisant une approche à base de patrons.

Parallèlement, d'autres travaux utilisent des techniques d'apprentissage automatique pour détecter la relation sémantique reliant deux entités médicales. Xiao et al. (Xiao et al. 2005) ont travaillé sur l'extraction des interactions entre protéines avec une méthode à base d'apprentissage supervisé. Ils ont défini des attributs lexicaux, syntaxiques et sémantiques. Ils ont obtenu un rappel de 93,9% et une précision de 88,0%. Roberts et al. (Roberts et al. 2008) se sont intéressés aux relations sémantiques dans les textes médicaux (e.g. *has finding*, *has indication*, *has location*) et ont proposé une méthode pour l'identification de ces relations en se basant sur un apprentissage supervisé avec des classifieurs SVM (Joachims 1998).

Le challenge i2b2 2010 (Uzuner et al. 2011) a proposé trois tâches différentes d'extraction d'information, parmi lesquelles l'extraction de relations sémantiques de 8 types. Plusieurs systèmes ont participé en proposant différentes méthodes statistiques et hybrides.

Extraction de relations entre les entités Traitement et Maladie

Ces travaux s'intéressent en particulier aux relations sémantiques reliant deux types d'entités médicales, à savoir, une *maladie* et un *traitement*. Cet intérêt s'explique par l'importance de ces deux types d'entités médicales et leur fréquence importante dans les textes médicaux. Pour cette tâche, différentes méthodes ont été utilisées. Lee et al. (Lee et al. 2004) ont appliqué des patrons construits manuellement sur des résumés médicaux dans le domaine du cancer pour l'identification de relations de type « Traitement » entre *médicament* et *maladie*. Appliqué à l'ensemble des phrases de leur jeu de test, leur système a obtenu un rappel de 84.8 % et une précision de 48,1 %.

Parallèlement aux travaux à base de patrons, d'autres ont utilisé l'apprentissage automatique pour identifier les relations entre un *traitement* et une *maladie*. Rosario et Hearst (Rosario et Hearst 2004) se sont intéressés à la désambiguïsation de sept types de relations. Ils ont comparé cinq modèles génératifs et un modèle de réseau de neurones et ont trouvé que le dernier permet d'avoir de meilleurs résultats. Frunza et Inkpen (Frunza et Inkpen 2010) ont travaillé sur le corpus de Rosario et Hearst (Rosario et Hearst 2004) et se sont intéressés à trois types de relations qui sont *cure*, *prevent* et *side effect* entre un Traitement et une Maladie. Ils ont utilisé l'outil Weka (Hall et al. 2009) et ont testé six modèles pour apprendre ces relations et ils ont montré que les modèles probabilistes et linéaires donnent les meilleurs résultats.

Synthèse

Plusieurs méthodologies sont proposées pour l'extraction des relations sémantiques. Certaines approches ont privilégié le rappel alors que d'autres ont mis l'accent sur la précision de l'extraction. Un point commun dans l'extraction des relations sémantiques est cependant le besoin de référentiels du domaine permettant de décrire des relations spécifiques. Les méthodes linguistiques permettent des analyses profondes du contexte d'occurrence de chaque entité médicale et de chaque relation, mais certaines relations sont indétectables avec ce genre de méthodes à cause de la grande variabilité d'expression des relations et en même temps de la structure, parfois très compliquée, de certaines phrases. Aussi les approches qui se fondent sur l'apprentissage ne peuvent garantir un haut degré de précision qu'avec la disponibilité d'un grand nombre d'exemples annotés pour une relation donnée.

3.3 Relations ciblées

Nous avons choisi sept types de relations médicales à partir d'une analyse des taxonomies de questions médicales (cf. chapitre 5) et d'un ensemble de questions médicales réelles (e.g. la collection de questions cliniques [1] contenant 4654 questions). Ces types de relations sont :

- *treats*, un traitement améliore ou traite un problème médical
- *prevents*, un traitement prévient un problème médical

1. http ://clinques.nlm.nih.gov/

- *complicates*, un problème médical (P) ou un traitement empire un problème médical (P')
- *causes*, un problème médical (P) ou un traitement cause un problème médical (P')
- *diagnoses*, un test médical détecte, diagnostique ou évalue un problème médical
- *DhD*, un médicament a une dose
- *P_hSS*, un problème médical a un signe ou un symptôme

La figure 3.1 présente les différentes relations sémantiques ciblées.

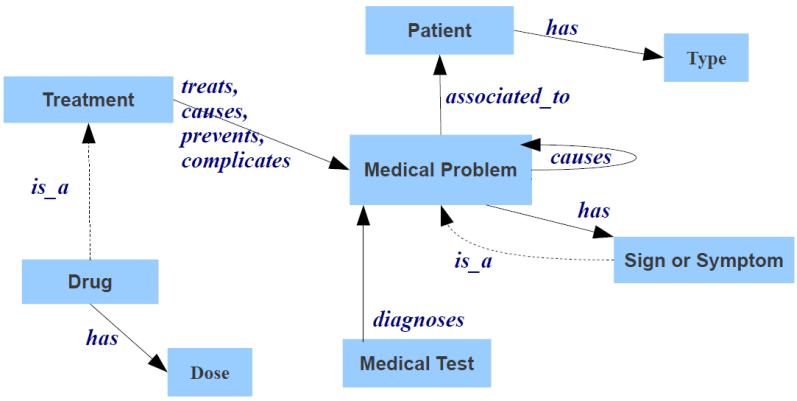


FIGURE 3.1 – *Modélisation des relations sémantiques ciblées*

Nous identifions aussi des relations spécifiques au patient : son sexe, son âge et sa catégorie d'âge (adulte, adolescent, enfant, bébé).

Nous présentons dans les sections suivantes les méthodes que nous proposons pour identifier les relations sémantiques entre entités médicales.

3.4 Méthode à base de patrons

3.4.1 Présentation

Extraire automatiquement des relations sémantiques entre les entités médicales nécessite des connaissances du domaine et une analyse linguistique des phrases du corpus. Les connaissances de domaine sont requises afin de savoir quels types de relations sont plausibles entre deux entités données et/ou ayant des catégories sémantiques connues. Le réseau sémantique de l'UMLS fournit une telle connaissance puisqu'il précise les types sémantiques sources et cibles de chaque relation (e.g. « Antibiotic » et « Disease or Syndrome » sont respectivement une source et une cible possibles pour la relation « treats »).

Afin d'extraire des relations entre les entités médicales identifiées nous utilisons un ensemble de patrons linguistiques que nous modélisons dans une ontologie liée au réseau sémantique de l'UMLS. Nous associons à chaque relation (ciblée) du réseau sémantique de l'UMLS un ensemble de patrons linguistiques. Un patron est une expression régulière décrivant

un modèle de phrase où les entités médicales sont présentes à des emplacements spécifiques (encadrés par des mots ou expressions régulières spécifiques).

Pour chaque couple d'entités médicales, nous commençons par chercher les relations possibles entre leurs types dans le réseau sémantique dans notre ontologie. Si une ou plusieurs relations sont possibles (e.g. entre les types sémantiques *Therapeutic or Preventive Procedure* et *Disease or Syndrome* il existe cinq relations : *treats*, *prevents*, *complicates*, etc.), nous utilisons les patrons associés à chacune d'entre elles pour déterminer la bonne relation.

Les patrons sont construits manuellement à partir de différents articles scientifiques et textes cliniques. Le tableau 3.1 présente le nombre de patrons construits pour chaque relation ainsi que des exemples de patrons.

TABLE 3.1 – *Exemples de patrons de relations sémantiques (DIS=Disease, TREAT=Treatment)*

Relation	Nombre de patrons	Exemples simplifiés
Cure	60	DIS was relieved by TREAT
Prevent	23	TREAT is effective for preventing DIS
Side Effect	51	DIS following administration of TREAT

La tâche de construction manuelle de patrons est coûteuse en temps si nous visons une couverture acceptable. Cette contrainte nous a mené à proposer une méthode semi-automatique pour la construction de patrons de relations sémantiques à partir de corpus sélectionnés sémantiquement (obtenus en interrogeant PubMed Central avec des requêtes MeSH). Cette méthode sera décrite dans la section suivante.

3.4.2 Construction semi-automatique des patrons de relations

La richesse d'expression que peuvent employer les êtres humains rend très difficile l'extraction automatique de relations sémantiques. Les relations de type « traitement » par exemple ne sont pas toujours exprimées avec des mots comme *treat* ou *treatment*. Il est très fréquent que ce type de relation soit exprimé avec des termes comme *reduce*, *respond to*, voire des expressions combinées et complexes. D'où la difficulté de l'élaboration de patrons de phrases pouvant couvrir toutes les expressions possibles. Cependant, l'utilisation des patrons reste une des méthodes les plus efficaces pour l'extraction automatique d'information s'ils sont mis au point de façon précise.

Notre proposons une méthode qui collecte automatiquement des phrases contenant au moins une source et une cible possibles pour une relation R donnée, cette relation devant être définie dans le réseau sémantique de l'UMLS comme une relation possible entre les types sémantiques de la source et de la cible. Cette pré-analyse sémantique réduit l'effort à effectuer pour observer l'expressivité d'une relation donnée, ce qui nous permet d'enrichir les patrons et d'augmenter leur nombre.

Corpus d'acquisition de patrons

Pour une relation ciblée "R", nous construisons un corpus de recherche de patrons. Pour cela, nous extrayons du metathesaurus de l'UMLS tous les couples de concepts reliés par la relation en question. Par exemple, il y a 45145 couples de concepts reliés par la relation « may_treat » (e.g. *Diazoxide may_treat Hypoglycemia* et *Dexamethasone 1 MG/ML Oral Solution may_treat Vomiting*). À partir de ces informations nous construisons des requêtes MeSH pour interroger PubMedCentral [2], une archive libre de plusieurs millions d'articles médicaux. Cette méthode nous donne des garanties quant à l'occurrence et à la variété des patrons dans les textes du corpus.

Méthode d'acquisition de patrons

Une fois le corpus d'acquisition construit pour une relation *R* donnée, nous commençons par récupérer les champs utiles dans chaque article (en format XML à l'origine) : entre autres, le titre, le résumé et le corps de l'article s'ils existent. Nous segmentons chaque texte en phrases, grâce aux modèles de segmentation du projet LingPipe [3], puis nous gardons les phrases qui contiennent un couple de concepts (c1,c2) reliés par la relation *R* dans le metathesaurus. Les phrases obtenues sont les plus aptes à contenir des occurrences de *R*. À partir de ces phrases, nous construisons manuellement des patrons sous forme d'expressions régulières, prenant en compte la présence d'entités médicales à telle ou telle position de la phrase. Cet ensemble contient à la fois des patrons précis (exigeant la présence de mots spécifiques) et des patrons plus génériques permettant de pallier des cas d'hétérogénéité d'expression. Pour la relation « traitement », nous avons construit 45 patrons (voir les exemples du tableau 3.2).

Patron1 : * TX* provide(s) ?* (clinical) ?* benefit(s) ? in* PB* Ex1 : *Adjunctive clonidine provides clinical benefits in patients hospitalized with ascites.*
Patron2 : *TX*-based therapy for | in* PB* Ex2 : *Clinical benefit of lapatinib-based therapy in patients with HER2-positive breast tumors.*
Patron3 : * PB *should be treated with* TX* Ex3 : *Emergent Hyperpyrexia in Children Should Be Treated With Antibiotics.*
Patron4 : *TX *is ((at least as) | as) ? effective(as *) ? in treating* PB* Ex4 : *Xeloda (capecitabine) is at least as effective in treating metastatic colorectal cancer.*

TABLE 3.2 – *Exemples de patrons, R = « traitement ». * : éléments contextuels, TX : Traitant, PB : Problème.*

2. http ://www.ncbi.nlm.nih.gov/pmc/
3. http ://alias-i.com/lingpipe/

Evaluation

Dans cette section nous évaluons notre approche d'extraction de relations sémantiques entre les entités médicales sur un jeu de test.

<relation>
<name>treat**</name>**
<sentence>*A subsequent study of patients with cSSSI also found that daptomycin resulted in faster clinical improvement***</sentence>**
<status>*established-known***</status>**
<source>*daptomycin***</source>**
<target>*cSSSI***</target>**
</relation>

FIGURE 3.2 – *Exemple d'annotation manuelle*

Nous avons construit le corpus d'évaluation en interrogeant PubMed-Central avec des requêtes MeSH (e.g. *Rhinitis, Vasomotor/th[MAJR] AND (Phenylephrine OR Scopolamine OR tetrahydrozoline OR Ipratropium Bromide)*). Nous avons ensuite choisi un sous-ensemble de 20 articles scientifiques anglais variés (e.g. articles de revues, études comparatives). La dernière étape de préparation a été l'annotation manuelle des entités médicales et des relations sémantiques qui les lient dans ces 20 articles (contenant 580 phrases).

La figure 3.2 montre un exemple de phrase annotée.

Par ces annotations nous conservons les entités médicales liées par la relation ciblée mais aussi le statut de cette relation (i.e. *établie-connue*, *établie-nouvelle*, *hypothétique* ou *non spécifié*). Ces statuts seront utilisés ultérieurement pour analyser les résultats et présenter les relations avec les statuts les plus sûrs dans le cadre d'un système de question-réponse.

Nous présentons les résultats obtenus sur notre corpus de 20 articles à deux niveaux : (i) la reconnaissance des entités médicales (cf. Tableau 2.4, page 25) et (ii) l'extraction de relations sémantiques. Pour évaluer les performances de notre approche à extraire des relations sémantiques de type « traitement » nous utilisons les mesures standards de rappel, de précision et la F-mesure. Le rappel est le nombre de relations correctes trouvées divisé par le nombre total de relations. La précision est le nombre de relations correctes trouvées divisé par le nombre de relations trouvées. Les résultats obtenus sont donnés en Tableau 3.3. Nous comparons les performances de notre approche avec celle de (Lee et al. 2004) qui ont aussi ciblé l'extraction de relations sémantiques de type « treats ».

	Rappel	Précision	F-Mesure
Notre approche	60,46 %	75,72 %	67,23 %
Lee & al. 2004	84 %	48,14 %	61,20 %

TABLE 3.3 – *Résultats de l'extraction de relations sémantiques de type « treats » obtenus en utilisant des patrons construits semi-automatiquement*

Discussion

Plusieurs approches d'extraction de relations sémantiques s'intéressent uniquement à la détection des relations sémantiques (e.g., dire que telle

ou telle phrase contient la relation recherchée). Dans le contexte des systèmes de question réponse médicaux, on ne s'intéresse pas uniquement à la reconnaissance de la relation mais aussi à la reconnaissance des entités médicales reliées. On s'intéresse donc à la recherche de triplets <source, relation, cible> tels que la source et la cible soient connues sémantiquement et que la relation soit valide vis à vis des connaissances du domaine et de la linguistique.

Dans ce contexte, une même phrase peut contenir plusieurs triplets <source, relation, cible>. Ce point n'est pas pris en considération par plusieurs approches (e.g. Lee et al. 2004). Dans notre cas, nous nous sommes intéressés à l'extraction précise de ces triplets. Une relation n'a été considérée comme correcte que si elle a été identifiée entre les bonnes entités médicales source et cible.

Nous avons obtenu une précision de 75,72 % sur notre corpus de test. Même en comparaison avec les approches qui ne s'intéressent pas à l'extraction précise de la source et de la cible de chaque relation, nous avons obtenu des résultats encourageants en terme de précision, ce qui rejoint notre objectif initial qui est de conserver le maximum possible de précision pour rendre plus efficaces les systèmes de question-réponse dans le domaine médical.

Une première analyse des faux positifs a montré que les principales causes d'erreur sont : (i) les erreurs d'extraction d'entités médicales (ii) patrons d'une relation *R* pouvant couvrir aussi d'autres types de relations et (iii) des phrases qui contiennent bien de bonnes entités source et cible (suivant les connaissances du domaine) sans qu'elles soient pour autant reliées par la relation recherchée.

Dans la section suivante nous nous intéressons à la qualité des patrons en proposant une approche qui leur associe une valeur de précision relative à la relation qu'ils décrivent et leur degré de spécificité.

3.4.3 Score de spécificité d'un patron et poids associé

Afin d'associer une mesure de confiance aux relations qui seront annotées ultérieurement, les patrons sont classés suivant leur précision. La précision d'un patron est calculée automatiquement en décrémentant les valeurs de précision suivant les relations de généralisation entre patrons. La figure 3.3 décrit l'ontologie représentant les patrons et les relations sémantiques ainsi que les différentes propriétés de généralisation (entre patrons) et de subsomption (entre relations). La propriété *précision* est utilisée pour associer un degré de spécificité au patron. La propriété *rdf:value* indique l'expression régulière correspondante au patron. La propriété *généralise* exprime un lien d'hyperonymie entre deux patrons. Par exemple, le patron ayant l'expression régulière "X for the treatment of Y" est un cas particulier du patron ayant l'expression régulière "X for Y".

La confiance que l'on peut avoir dans une relation extraite par un patron varie selon que le patron spécifie un contexte lexical plus ou moins précis. Nous qualifions de *spécifique* un patron qui précise davantage ce contexte (typiquement en employant davantage de mots). Nous cherchons à prendre en compte cette spécificité en associant à chaque patron un poids. Ce poids sera utilisé pour (i) l'extraction des relations, où il per-

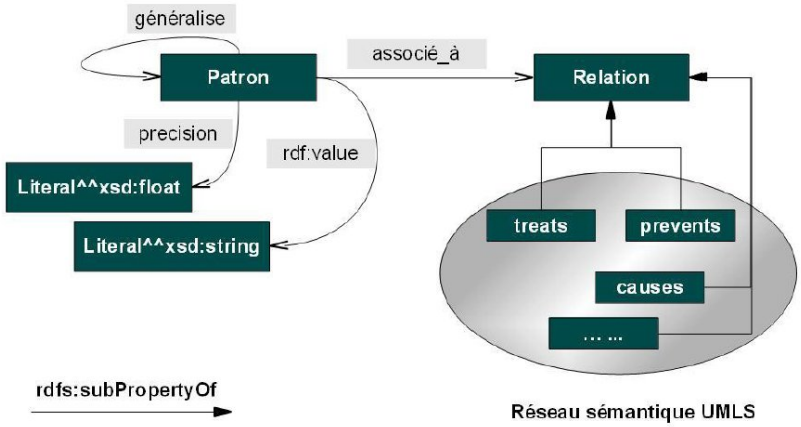


FIGURE 3.3 – *Ontologie modélisant les patrons et les relations sémantiques*

mettra de favoriser la relation obtenue avec le patron ayant le meilleur poids en cas d'ambiguïté (correspondance de plusieurs patrons) et (ii) pour l'approche hybride où il contribuera en tant que facteur pour choisir la relation finale à extraire parmi les relations extraites par les différentes méthodes (cf. section 3.6).

Le poids d'un patron est calculé automatiquement en décrémentant les valeurs à partir de 1 (poids des patrons les plus spécifiques). L'exemple suivant montre les résultats de la propagation des valeurs de poids à partir d'un ensemble de faits $E1$ décrivant trois patrons différents (C est un coefficient entier).

$$E1 = \left\{ \begin{array}{c} < patron1, spécificité, \mathbf{P} > \\ < patron2, généralise, patron1 > \\ < patron3, généralise, patron2 > \end{array} \right\} E2 = \left\{ \begin{array}{c} < patron2, spécificité, \mathbf{P/C} > \\ < patron3, spécificité, \mathbf{P/C/C} > \end{array} \right\}$$

Dans le cas où un patron généralise plusieurs autres patrons différents, nous lui assignons automatiquement le minimum des poids de toutes ses dérivations directes. Les patrons sont ensuite appliqués du plus spécifique au plus général lors du processus d'extraction de relation.

Une relation est identifiée dans une phrase si (i) elle est définie dans l'ontologie entre les types de l'entité source et de l'entité cible et (ii) elle est identifiée par l'application de l'un de ses patrons sur la même phrase. Le tableau 3.4 présente quelques exemples de patrons et les relations qu'ils permettent d'extraire.

TABLE 3.4 – *Exemples de patrons pour la relation* Cure *et leur spécificité*

Patron	Relation	Spécificité	Exemple
TREAT for DIS	Cure	0,50	*Intralesional corticosteroid therapy for primary cutaneous B cell lymphoma.*
TREAT for the treatment of DIS	Cure	0,83	*Cognitive-behavioral group therapy* is an effective intervention for the treatment of *geriatric depression.*

Comme annoncé plus haut, le poids des patrons sert à calculer un indice de confiance pour les relations extraites. Cet indice de confiance

prend aussi en compte le nombre de syntagmes nominaux entre les deux entités médicales concernées. L'idée derrière ce second facteur est que la relation est considérée plus forte s'il n'y a que des verbes et/ou prépositions entre les deux entités médicales en question que s'il y a un ou plusieurs syntagmes nominaux entre les entités.

L'indice de confiance I d'une relation R extraite par un patron P à partir d'une phrase H, entre deux entités médicales $E1$ et $E2$ contenues dans deux syntagmes $S1$ et $S2$ est défini comme :

$$I(R) = \frac{Sp(P)}{e^{Nsynt(H,S1,S2)}} \tag{3.1}$$

- $Sp(P)$: poids du patron P ;
- $Nsynt(H, S1, S2)$: une fonction qui retourne le nombre de syntagmes nominaux entre $S1$ et $S2$ dans la phrase H.

3.5 Méthode statistique

Cette deuxième méthode se fonde sur une technique d'apprentissage automatique supervisé. Étant donné plusieurs catégories définies a priori (ici, les différentes relations à reconnaître, ou l'absence de relation), une telle technique s'appuie sur un ensemble d'exemples d'entraînement pour pouvoir prendre une décision de classification sur les exemples de test. Chaque exemple doit être décrit par un ensemble d'attributs (features) qui doivent être suffisamment discriminants pour assurer une bonne performance de classification.

Dans notre approche, nous utilisons un classifieur linéaire SVM (Joachims 1998) qui est connu pour ses performances en catégorisation de textes. Nous utilisons la librairie LIBSVM (Chang et Lin 2001b).

Le problème à résoudre est modélisé comme suit : étant donné deux entités E1 et E2 dans une phrase, déterminer la relation qui les relie (ou l'absence de relation).

3.5.1 Attributs utilisés

Nous avons choisi 3 types d'attributs pour décrire les données : des attributs lexicaux, morpho-syntaxiques et sémantiques.

Attributs lexicaux

Cette classe couvre les attributs relatifs aux mots : (1) les mots de l'entité source E1, (2) les mots de l'entité cible E2, (3) les mots entre E1 et E2, (4) les mots avant E1, (5) les mots après E2 et (6) les lemmes des mots.

Attributs morpho-syntaxiques

Ce type d'attributs comporte : (1) la catégorie morpho-syntaxique des mots (de E1, de E2, avant E1, avant E2 et entre E1 et E2), (2) les verbes entre E1 et E2, (3) les verbes avant E1 et (4) les verbes après E2. Nous avons utilisé l'outil TreeTagger pour l'analyse morpho-syntaxique du corpus.

Attributs sémantiques

Cette classe regroupe les attributs qui exploitent des ressources sémantiques externes. La ressource la plus importante dans le domaine médical est l'UMLS. La première classe d'attributs exploite le metathesaurus de l'UMLS et comporte : (1) le concept associé à E1, (2) le concept associé à E2, (3) les concepts existant entre E1 et E2. La deuxième classe d'attributs exploite le réseau sémantique de l'UMLS et comporte : (1) le type sémantique de E1, (2) le type sémantique de E2, (3) les types sémantiques des entités médicales entre E1 et E2 et (4) les relations sémantiques possibles entre E1 et E2.

Nous nous intéressons aussi aux types de verbes vu qu'ils sont souvent les premiers indicateurs du type de la relation entre deux termes. Nous exploitons alors les ressources autour des verbes et en particulier : (i) les classes sémantiques de VerbNet et (ii) les classes sémantiques de Levin pour typer les verbes existant entre les entités E1 et E2, avant E1 et après E2.

3.5.2 Evaluation

Nous avons testé notre méthode statistique dans le cadre du challenge DDIE [4]. La tâche était d'identifier une interaction ou non interaction entre deux médicaments. Nous avons adapté ce classifieur à cette tâche en ajoutant deux classes d'attributs : (i) présence d'une négation et (ii) présence d'un mot clé (ou mot déclencheur) à partir d'une liste de mots clés décrivant une interaction. Cette liste est construite manuellement à partir des exemples d'entraînement.

Le tableau 3.5 présente les résultats obtenus par notre méthode statistique à base d'attributs (FBM) et une deuxième méthode à base de noyau (KBM) réalisée par Faisal Chowdhury (Chowdhury et al. 2011). Nous avons testé l'union et l'intersection pour combiner ces deux méthodes.

	FBM	**KBM**	**Union**	**Intersection**
Précision	0.5910	0.4342	0.4218	0.6346
Rappel	0.3640	0.5277	0.6083	0.2821
F_1 Score	0.4505	0.4764	**0.4982**	0.3906

TABLE 3.5 – *Résultats des expérimentations sur 37% des documents originaux avec un entraînement sur les 63% restants*

Le tableau 3.6 présente les résultats obtenus à l'évaluation finale. Le système combinant les résultats avec l'union (FBK+KBM) a obtenu un F-score de 63.98% et été classé deuxième au challenge (sur 10 participants). Notre méthode à base d'attributs FBM a obtenu une bonne précision (70.58%) mais un faible rappel (42.25%) par rapport à la méthode à base de noyaux KBM (67.95%).

4. http ://labda.inf.uc3m.es/DDIExtraction2011

	FBM	**KBM**	**Union (FBM+KBM)**	**Meilleur système**
Précision	**0.7058**	0.5986	0.5859	0.605
Rappel	0.4225	0.6795	**0.7046**	0.719
F_1 Score	0.5286	0.6365	**0.6398**	0.657

TABLE 3.6 – *Résultats obtenus à l'évaluation finale du challenge*

3.6 Méthode hybride

Dans cette section nous présentons notre approche pour la combinaison de nos méthodes à base de règles et statistique.

3.6.1 Approche proposée

Cette méthode combine les deux méthodes précédentes pour extraire des relations en fonction des indices de confiance attribués aux résultats de chaque méthode. L'extraction d'une relation se fait suivant l'influence ou poids accordé à chaque méthode. Nous nous sommes basés sur le nombre d'exemples d'entraînement pour calculer l'influence de l'approche statistique sur la procédure d'extraction (ce poids est noté $\mu_s(R)$).

L'influence de l'approche par règles d'extraction ou patrons est calculée avec 2 poids différents. Un poids global $\mu_p(R)$ qui est le complémentaire de $\mu_s(R)$ pour une relation R donnée : $\mu_p(R) + \mu_s(R) = 1$ et un poids plus fin au niveau de chaque occurrence de relation extraite qui tient compte de l'indice de confiance associée à cette occurrence (cf. section 3.4). La relation extraite par patrons n'a d'influence que si (i) son indice de confiance est supérieur à un seuil donné I_{min} et (ii) son poids global est supérieur ou égal au poids de la méthode statistique pour la même relation : $\mu_p(R) >= \mu_s(R)$

3.6.2 Expérimentations

Dans cette section nous décrivons le corpus utilisé pour l'évaluation ainsi que les résultats obtenus.

Description du corpus

Nous utilisons le corpus de Rosario et Heart (Rosario et Hearst 2004), qui a aussi été utilisé par Frunza et Inkpen (Frunza et Inkpen 2010). Ce corpus a été extrait de Medline 2001 et a été annoté avec 8 types de relations sémantiques entre un traitement (TREAT) et une maladie (DIS). Ces relations (cf. tableau 3.7) sont : *Cure*, *Only DIS* (TREAT n'est pas mentionné), *Only TREAT* (DIS n'est pas mentionné), *Prevent*, *Vague* (la relation n'est pas claire), *Side Effect*, *No Cure*. Les relations *Only DIS* et *Only TREAT* ne correspondent pas à notre objectif d'extraction vu qu'une seule entité est présente dans la phrase. Le nombre d'exemples pour les relations *Vague* et *No Cure* est aussi très petit et ne nous permet pas d'appliquer une approche par apprentissage efficace. Notre choix final s'est donc porté sur les relations : *Cure*, *Prevent* et *Side Effect*.

TABLE 3.7 – *Corpus initial*

Relation nb (train, test)	Définition *Exemple*
Cure 810 (648, 162)	TREAT cures DIS *Intravenous immune globulin for recurrent spontaneous abortion*
Only DIS 616 (492, 124)	TREAT not mentioned *Social ties and susceptibility to the common cold*
Only TREAT 166 (132, 34)	DIS not mentioned *Flucticasone propionate is safe in recommended doses*
Prevent 63 (50, 13)	TREAT prevents the DIS *Statins for prevention of stroke*
Vague 36 (28, 8)	Very unclear relationship *Phenylbutazone and leukemia*
Side Effect 29 (24, 5)	DIS is a result of a TREAT *Malignant mesodermal mixed tumor of the uterus following irradiation*
No Cure 4 (3, 1)	TREAT does not cure DIS *Evidence for double resistance to permethrin and malathion in head lice*
Total relevant : 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	TREAT and DIS not present *Patients were followed up for 6 months*
Total : 3495 (2793, 702)	

Le corpus initial a été divisé en un corpus d'entraînement et un corpus de test de tailles égales pour toutes les relations. Cependant, étant donné que chaque phrase du corpus est annotée par une relation unique même si elle comporte plusieurs relations entre des couples d'entités différents, nous avons étendu les corpus d'entraînement et de test en dupliquant les phrases « multi-relation » en plusieurs phrases « mono-relation » avec des entités source et cible différentes.

Cette démarche nous a aussi permise d'intégrer les phrases annotées comme étant *to see* sous forme de plusieurs phrases avec des relations potentiellement différentes. Par exemple, la phrase suivante : « *<DIS_SIDE_EFF> Progressive multifocal leukoencephalopathy </DIS_SIDE_EFF> following <TREAT> oral fludarabine </TREAT> treatment of <DIS> chronic lymphocytic leukemia </DIS>* » a été réécrite en 2 phrases exprimant respectivement les relations « side_effect » et « cure ». Le nombre de phrases (« mono-relation ») des corpus d'entraînement et de test est présenté dans le tableau 3.8 pour chaque relation. Les nombres très variés d'exemples disponibles pour chaque relation permettent de tester et d'évaluer l'apport de différentes méthodes d'extraction.

TABLE 3.8 – *Notre distribution de données*

Relation	Corpus d'entraînement (nombre de phrases)	Corpus de test (nombre de phrases)
Cure	524	530
Prevent	43	33
Side Effect	44	28

Configurations et résultats

Le corpus d'entraînement a servi à ajouter des patrons supplémentaires et à entraîner la méthode d'apprentissage. Toutes les méthodes ont ensuite été testées sur le corpus de test. Nous récapitulons les différentes configurations que nous avons testées et présentons les résultats obtenus à l'aide des mesures classiques de *précision*, *rappel* et *F-mesure*. La précision est la proportion des relations correctes parmi les relations proposées par le système. Le rappel est la proportion des relations correctes que le système a trouvées. La F-mesure est la moyenne harmonique du rappel et de la précision.

Le tableau 3.9 présente les trois configurations que nous avons testées.

TABLE 3.9 – *Les différentes configurations testées*

P	Patrons
A	Apprentissage
H	Patrons + apprentissage (selon μ_p, μ_s et I)

Le tableau 3.10 présente les résultats obtenus pour chaque type de relation. Le tableau 3.11 présente le rappel, la précision et la F-mesure moyennes en pourcentages pour chaque relation.

TABLE 3.10 – *Précision P, rappel R et F-mesure F de chaque relation pour chaque configuration*

	Cure			Prevent			Side effect		
Config.	P	R	F	P	R	F	P	R	F
P	95,55	32,45	48,44	89,47	51,51	65,37	65,21	53,57	58,63
A	90,44	100	94,98	15,15	15,15	15,15	0	0	0
H	98,51	100	99,24	100	51,51	67,99	93,75	53,57	68,18

Sur les 631 relations trouvées par l'approche hybride, 182 relations ont été retrouvées à la fois par l'approche par patrons et l'approche par apprentissage, dont 7 relations fausses et 175 relations correctes (le chiffre de 631 ainsi que les résultats présentés ne comprennent pas de redondances).

La méthode hybride apporte bien un plus aux deux méthodes séparées. L'apport qualitatif à l'apprentissage est important pour les relations « prevent » et « side effect », l'apport quantitatif (sur l'ensemble des relations) pourrait aussi être sensiblement plus grand en présence d'un grand nombre de relations avec peu d'exemples d'entraînement.

TABLE 3.11 – *Précision P, rappel R et F-mesure F calculées sur les occurrences de toutes les relations*

Configuration	Précision (%)	Rappel (%)	F-mesure (%)
P	91,89	34,51	50,17
A	90,52	90,52	90,52
H	89,38	95,43	92,30

3.7 Discussion

Plusieurs travaux d'extraction de relations sémantiques (e.g. (Frunza et Inkpen 2010)) ciblent uniquement la détection de la présence ou non d'une relation *R* dans une phrase. Dans notre approche nous nous intéressons à l'extraction d'une relation *R* entre deux entités médicales précises. Par exemple dans la phrase *TX treats Pb1 but increases the risk of Pb2* notre objectif est de déterminer qu'une relation de type *Cure* existe dans cette phrase entre *TX* et *PB1*. Frunza et Inkpen (Frunza et Inkpen 2010) ont utilisé les phrases *Only DIS* et *Only TREAT* comme des exemples négatifs alors que dans notre approche, un exemple négatif est un exemple où deux entités de type TREAT et DIS existent, sans être reliée par la bonne relation.

Les expérimentations que nous avons effectuées montrent que les méthodes à base de patrons permettent d'obtenir une bonne précision, mais présentent un inconvénient par rapport à la grande variabilité et la structure complexe de certaines phrases. Les méthodes à base d'apprentissage automatique peuvent quant à elles être très robustes mais nécessitent pour cela un grand nombre d'exemples annotés pour obtenir de bons résultats d'extraction.

La combinaison des méthodes linguistique et statistique permet de tirer parti des avantages des deux méthodes. Notre approche a obtenu de bons résultats en se basant automatiquement sur l'apprentissage dans le cas de la relation *Cure* vu que nous avons beaucoup d'exemples annotés (524 phrases). Pour les autres relations (*Prevent* et *Side Effect*), nous n'avons pas beaucoup d'exemples et donc l'apprentissage n'a pas donné de bons résultats. Ce manque a pu être pallié par l'utilisation simultanée des patrons de phrase construits manuellement pour ces mêmes relations.

Conclusion du chapitre

Nous avons présenté dans ce chapitre une approche hybride pour l'extraction de relations sémantiques. Cette approche se fonde d'une part sur une méthode à base de patrons et d'autre part sur une méthode à base d'apprentissage supervisé qui utilise un ensemble d'attributs lexicaux, morpho-syntaxiques et sémantiques. Nous avons montré que la combinaison des deux méthodes donne de meilleurs résultats que lorsqu'elles sont utilisées séparément. En perspectives, nous envisageons d'inclure plus de paramètres pour la combinaison des deux méthodes. En particulier, nous pouvons exploiter le coefficient de corrélation entre le nombre d'exemples d'entraînement pour une relation *R* et les résultats obtenus pour cette même relation avec la méthode statistique.

4 Extraction d'information à partir de textes médicaux en français

Sommaire

La tâche d'extraction d'information de fine granularité à partir de textes médicaux en français est relativement semblable dans ces grandes lignes à l'extraction d'information depuis des corpus en anglais. Diverses méthodes statistiques, à base d'apprentissage artificiel, ou linguistiques à base de patrons peuvent aussi être appliquées à des textes en français pour retrouver les entités médicales et les relations qui les lient. Cependant, dans la pratique, ces méthodes ne sont pas toujours applicables à cause de l'absence des ressources ou de corpus adéquats pour le français. Dans ce cadre, une solution alternative est de projeter les informations extraites d'un corpus en anglais sur son corpus correspondant en français.

Dans ce chapitre, nous proposons une méthode pour l'annotation automatique de textes médicaux en français par projection inter-langues. Cette recherche est issue de notre volonté de tester de nouvelles méthodes automatiques d'annotation ou d'extraction d'information à partir d'une langue L1 en exploitant des ressources et des outils disponibles pour une

autre langue L2. Cette approche repose sur le passage par un corpus parallèle (L1-L2) aligné au niveau des phrases et des mots. Pour faire face au manque de corpus médicaux français annotés, nous nous intéressons au couple de langues (français-anglais) dans le but d'annoter automatiquement des textes médicaux en français. En particulier, nous nous intéressons dans ce chapitre à la reconnaissance des entités médicales. Nous évaluons dans un premier temps notre méthode de reconnaissance d'entités médicales sur le corpus anglais (L1). Dans un second temps, nous évaluons la reconnaissance des entités médicales du corpus français (L2) par projection des annotations du corpus anglais. Nous abordons également le problème de l'hétérogénéité des données en exploitant un corpus extrait du Web et nous proposons une méthode statistique pour y pallier.

Les travaux présentés dans ce chapitre s'appuient sur l'article (Ben Abacha et al. 2012).

4.1 Introduction

L'extraction d'information vise à extraire automatiquement à partir de textes des informations structurées pertinentes pour une tâche particulière (Poibeau 2003). Il y a essentiellement deux types de méthodes utilisées en extraction d'information : les méthodes où une personne (un « expert ») fournit des connaissances (linguistiques ou sur le domaine) [1], et les méthodes dirigées par les données, où ces connaissances sont construites par apprentissage supervisé. Il existe également des méthodes hybrides combinant ces deux techniques. Ces deux types de méthodes ont certaines limitations ((Bach et Badaskar 2007), (Nadeau et Sekine 2007), (Ben Abacha et Zweigenbaum 2011c)) :

- Les méthodes à base de connaissances expertes sont simples à mettre en place mais coûteuses en temps pour ce qui est de la construction des connaissances. Elles ont aussi un potentiel de couverture réduit comparé aux méthodes statistiques.
- Les méthodes par apprentissage peuvent être très robustes si (i) on dispose d'un bon nombre d'exemples d'entraînement et si (ii) le corpus de test est du même type que le corpus d'entraînement. Ces méthodes sont de fait dépendantes des données et des corpus annotés, ressources qui ne sont pas disponibles pour toutes les langues (par exemple, il n'existe pas de corpus médicaux annotés en français) ni pour toutes les tâches (par exemple, reconnaissance des entités médicales, extraction de relations sémantiques, etc.).

Cette observation s'applique aussi au domaine médical : pour l'anglais, plusieurs outils spécialisés d'extraction d'information existent (tels que MetaMap (Aronson 2001), cTAKES (Savova et al. 2010)), ainsi que des corpus annotés en entités nommées (tels que i2b2 (Uzuner et al. 2011), Berkeley (Rosario et Hearst 2004)). En revanche, peu de ressources sont disponibles en français que ce soit au niveau des outils spécialisés ou au niveau des corpus médicaux annotés.

L'annotation manuelle d'exemples pour l'entraînement peut être une solution pour les méthodes par apprentissage supervisé ou semi-supervisé. Cependant, cette tâche nécessite des experts du domaine ciblé, au moins pour la validation. D'après nos expériences précédentes portant sur l'annotation manuelle de corpus médicaux en français constitués (i) de résumés d'articles scientifiques et (ii) de textes extraits du corpus EQueR (Ayache 2005), plusieurs obstacles ont été mis en évidence. Dans une première phase, nous avons annoté manuellement des textes médicaux avec le concours de deux médecins. L'obstacle principal était le fait que la tâche est longue et fastidieuse. Ensuite, et pour accélérer la tâche d'annotation, nous avons développé une interface pour l'annotation de phrases (et non pas de textes entiers) permettant à davantage de médecins de prendre part à l'annotation. Le premier inconvénient de cette méthode est la perte du contexte des phrases. Un deuxième inconvénient réside dans le fait que, même si le guide d'annotation est très détaillé, les divergences dans les avis des médecins augmentent avec le nombre de médecins intervenant (par exemple dans l'annotation des symptômes et des relations dans des textes dans le domaine psychiatrique). Ces divergences,

1. Méthodes souvent appelées improprement « à base de règles ».

portant par exemple sur les types d'entités médicales et les relations à annoter, peuvent ralentir le processus d'annotation manuelle et le rendre moins fiable.

Dans ce chapitre, nous exploitons un autre type de méthode, la *projection d'annotations* d'une langue à une autre (Yarowsky et Ngai 2001), et testons son application au domaine médical. L'idée générale consiste à transférer des annotations d'une langue L1 (pour laquelle plus de ressources sont disponibles) à une langue L2 en utilisant des corpus parallèles et leur alignement au niveau des mots. Cette approche devrait nous permettre d'exploiter, pour l'annotation automatique de textes en français, les ressources disponibles en anglais ainsi que les méthodes d'extraction d'information développées pour cette même langue. Notre premier objectif, présenté à travers ce chapitre, consiste à annoter automatiquement les entités médicales de textes en français par transfert d'entités détectées dans les textes anglais correspondants par des outils existants de reconnaissance d'entités médicales. La table 4.1 présente un exemple de ce que nous cherchons à obtenir.

Phrase en anglais	The role of carotid endarterectomy in the management of asymptomatic carotid stenosis is much less clear.
Phrase équivalente en français	Le rôle de l'endartériectomie carotidienne dans le traitement d'une sténose carotidienne asymptomatique est beaucoup moins clairement défini.
Alignement au niveau des mots	0-0 1-1 2-2 3-3 3-4 4-3 5-5 6-6 7-7 8-8 9-11 10-8 11-9 11-10 12-12 13-13 14-14 15-15 15-16
Entités médicales (en anglais)	"carotid endarterectomy" 3–4 [treatment] "asymptomatic carotid stenosis" 9–11 [problem]
Résultat final (annotations en français)	"l'endartériectomie carotidienne " 3–4 [treatment] "une sténose carotidienne asymptomatique" 8-11 [problem]

TABLE 4.1 – *Exemple illustratif de l'approche proposée*

Après un rappel des travaux similaires (section 4.2), nous présentons les deux étapes principales de l'approche que nous proposons ici (telle qu'illustrée sur la figure 4.1) :

1. L'extraction d'information à partir de la partie L1 du corpus parallèle, en utilisant des méthodes déjà développées ou des outils disponibles (section 4.3). Pour ce faire, nous utilisons la méthode MetaMapPlus décrite dans le chapitre 2.
2. L'alignement des mots [2] des parties L1 et L2 du corpus (section 4.4.1) et la projection des entités repérées sur L1 vers la partie L2 en utilisant ces alignements (section 4.4.2). Nous mettons en place quelques heuristiques pour réparer certaines erreurs et améliorer la précision de la projection en diminuant le bruit des alignements.

2. L'alignement des phrases a été effectué par Louise Deléger, l'alignement des mots a été effectué par Aurélien Max.

Nous évaluons notre approche (section 4.5) sur une partie du corpus Santé Canada [3] et discutons ses résultats, puis concluons sur des perspectives de travaux futurs.

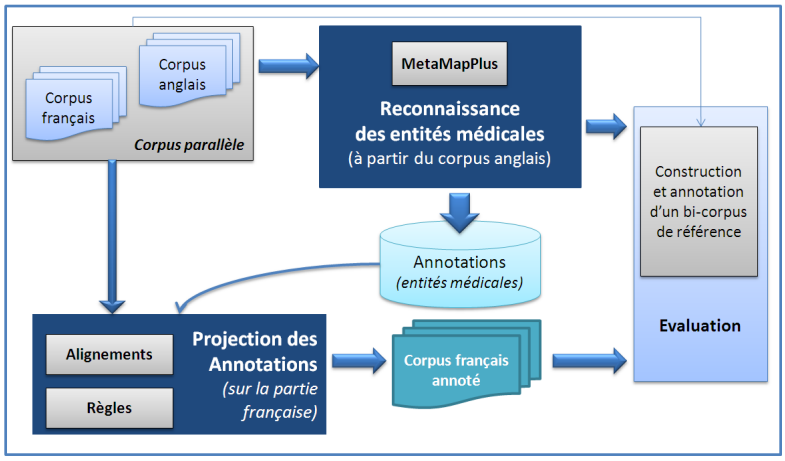


FIGURE 4.1 – *Approche proposée pour l'annotation automatique d'un corpus médical français, utilisant un corpus parallèle et des méthodes d'extraction d'information à partir de textes anglais*

4.2 Travaux similaires

Des travaux sur la projection d'analyses linguistiques ou d'annotations d'une langue à l'autre se sont développés essentiellement à partir des années 2000. (Yarowsky et Ngai 2001) ont proposé d'utiliser des corpus parallèles alignés au niveau des mots pour transférer de façon robuste de l'anglais au français ou au chinois des étiquettes morphosyntaxiques et des frontières de syntagmes nominaux. (Lopez et al. 2002) ont étudié comment transférer un arbre de dépendances de l'anglais vers le chinois. (Lü et al. 2002) se sont également intéressés au transfert d'analyses syntaxiques de l'anglais vers le chinois (Padó et Pitel 2007) ont traité le problème de l'annotation automatique de rôles sémantiques dans une langue ne disposant pas de lexique dans FrameNet [4], en s'intéressant au couple de langues (anglais-français).

Dans le domaine médical, plusieurs travaux se sont attaqués au transfert de connaissances d'une langue à une autre (Névéol et al. 2005, Deléger et al. 2006). En particulier, (Deléger et al. 2009) se sont intéressés à l'acquisition de nouvelles traductions de termes issues de trois terminologies différentes («MeSH», «SNOMED CT» et «MedlinePlus Health Topics»). Ces auteurs se sont basés sur l'alignement des mots à partir d'un corpus parallèle anglais-français.

3. http://www.hc-sc.gc.ca
4. http://framenet.icsi.berkeley.edu/fndrupal/

4.3 Reconnaissance d'entités médicales dans des textes anglais

Nous utilisons la méthode MetaMapPlus (cf. chapitre 2) pour la reconnaissance des entités médicales à partir de textes anglais. Le corpus que nous ciblons ici est une collection de documents extraites du site Web Santé Canada [5] qui regroupe des articles destinés aussi bien au grand public qu'aux praticiens du domaine. Cependant, ce corpus est très hétérogène au niveau du vocabulaire et contient beaucoup de mots généraux qui brouillent l'extraction d'information médicale à cause de leur similitude lexicale avec certaines entités médicales (e.g. extraire "table" comme étant un appareil médical, type sémantique *MedicalDevice* dans l'UMLS, ou "section" comme une substance, type sémantique *Substance* dans l'UMLS).

Ainsi, l'application de la méthode MetaMapPlus, qui se réfère au metathesaurus de l'UMLS, a mis en évidence les problèmes liés à l'ambiguïté lexicale au niveau de certains termes à cause de l'hétérogénité du corpus. Cette ambiguïté peut être divisée en deux catégories principales : (i) les homonymes (tels que « ten », qui désigne « dix » en domaine ouvert et la maladie « Toxic Epidermal Necrolysis » en domaine médical) et (ii) les termes généraux ayant un sens qui se spécialise dans le domaine médical (tels que « case », « form »).

Nous proposons dans cette section une solution pour pallier cette difficulté à travers un filtre statistique utilisé en amont pour identifier si les entités extraites sont des éntités médicales ou non. Cette étape consiste à utiliser un classifieur pour distinguer les termes médicaux et les termes généraux, avant d'appliquer MetaMap. Le but précis de ce module est de :

1. Réduire le bruit lié à l'ambiguïté lexicale, en éliminant les syntagmes nominaux (SN) « généraux » fréquents en domaine ouvert même lorsqu'ils sont utilisés dans le domaine médical (par exemple « table »).
2. Réduire le volume à traiter par la catégorisation *via* MetaMap en éliminant une bonne partie des syntagmes nominaux à classifier, ce qui devrait réduire le temps d'exécution.

Nous proposons une approche d'apprentissage supervisé, appelée Maxent_SNG, pour filtrer les entités extraites suivant leur acception médicale ou générale. Les méthodes statistiques à base d'apprentissage supervisé peuvent être très robustes. Pour différencier les données d'entraînement et offrir une meilleure *adaptabilité* en évitant le sur-apprentissage, nous traitons le problème selon deux dimensions : (i) comment choisir les exemples d'entraînement ? (ii) et quels sont les attributs à utiliser ?

4.3.1 Sélection des données d'apprentissage

À l'instar des travaux sur l'apprentissage actif (Active Learning) (Thompson et al. 1999, Tomanek et Olsson 2009) qui sélectionnent des exemples diversifiés et représentatifs à annoter manuellement, nous avons trouvé utile de sélectionner les exemples à utiliser pour « bien » apprendre. Deux questions clés se posent alors :

5. http://www.hc-sc.gc.ca

– le nombre des exemples *positifs* et *négatifs* à utiliser ;
– le choix de ces exemples qui doivent être représentatifs.

Pour choisir ces exemples, nous proposons d'utiliser :

1. la fréquence des mots/syntagmes nominaux (positifs et négatifs) dans un même corpus ;
2. la présence des mots/syntagmes nominaux (positifs et négatifs) dans des corpus textuels médicaux de genres différents ;
3. le Web pour collecter des données (des exemples positifs et négatifs).

Plus précisément, pour la construction des données d'apprentissage nous utilisons les exemples positifs et négatifs suivants :

1. Exemples positifs : entités médicales (EM)
   – les EM les plus fréquentes dans le corpus i2b2 de textes cliniques ;
   – les EM les plus fréquentes dans le corpus Berkeley d'articles scientifiques (Rosario et Hearst 2004) ;
   – les EM communes aux deux corpus ;
   – des EM extraites du Web (notamment de Wikipedia [6] et HON [7]) ;
2. Exemples négatifs : SN « généraux » (SNG) qui ne contiennent pas des entités médicales :
   – les SNG les plus fréquents dans le corpus i2b2 ;
   – les SNG les plus fréquents dans le corpus de Berkeley ;
   – les SNG les plus fréquents qui existent dans les deux corpus à la fois ;
   – des SNG extraits du Web, à partir de sites thématiquement distant du domaine médical. Nous avons choisi des sites d'histoires pour enfants [8] [9]). Notre motivation est d'utiliser des corpus ne contenant pas ou peu d'entités médicales.

La table 4.2 décrit les types d'exemples positifs et négatifs que nous avons utilisés, selon trois critères : corpus, nombre d'exemples et nombre d'occurrences de chaque exemple.

4.3.2 Traits utilisés par le classifieur

Pour cette tâche, nous utilisons un classifieur à maximum d'entropie [10]. Pour chaque syntagme nominal (médical ou général), les attributs utilisés par le classifieur sont :

– la longueur du SN, son nombre de tokens ;
– le SN est un mot en majuscules / le SN est en majuscules / le SN contient un mot en majuscules ;
– les mots / lemmes / catégories syntaxiques du SN ;
– la présence et la fréquence des mots du SN dans la liste des mots du corpus général BNC [11] ;

6. Différentes listes d'entités médicales ont été extraites à partir de Wikipedia : medical tests, diseases, disorders, treatments, procedures (diagnostiques, therapeutiques, surgicales,..)
7. HON (Health On the Net) : `http://www.hon.ch/`
8. `http://www.goodnightstories.com/read/pnkbook1.htm`
9. `http://www.vtaide.com/png/stories.htm`
10. Nous avons utilisé l'implémentation disponible à : `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.
11. British National Corpus, `http://www.natcorp.ox.ac.uk/`.

	Corpus	Nb d'exemples	Fréquence des exemples
Exemples positifs	extraits du Web (Wikipedia, HON, etc.)	1114	entre 1 et 3
	corpus médical 1 (i2b2 : textes cliniques)	3974 (sur 26187 EM)	>= 3 (allant jusqu'à 347 pour « hypertension »)
	corpus médical 2 (Berkeley : articles scientifiques)	391 (sur 2463 EM)	>=2 (allant jusqu'à 28 pour « chemotherapy »)
	Total	**5479** entités médicales	
Exemples négatifs	extraits du Web (sites d'histoires pour enfants)	2127	1 et 2
	corpus médical 1 (i2b2 : textes cliniques)	2031 (sur 15882 SN)	>= 3 (allant jusqu'à 855 pour « the patient »)
	corpus médical 2 (Berkeley : articles scientifiques)	1639 (sur 10464 SN)	>= 2 (allant jusqu'à 278 pour « patients »)
	Total	**5797** syntagmes nominaux (généraux)	

TABLE 4.2 – *Classification des syntagmes nominaux en termes médicaux et termes généraux, et sélection des exemples positifs et négatifs selon trois critères : corpus, nombre d'exemples et nombre d'occurrences de chaque exemple.*

– la présence des mots du SN dans un dictionnaire général (nous avons utilisé le dictionnaire standard du système d'exploitation Linux).

4.4 Projection des annotations sur des textes en français par alignement

4.4.1 Alignement au niveau des mots

La projection que nous réalisons se fonde sur des alignements calculés au niveau des mots (alignements effectués par Aurélien Max). Pour les obtenir, nous avons utilisé les programmes d'alignement de corpus parallèles du système de traduction statistique MOSES (Koehn et al. 2007), en utilisant le paramétrage par défaut. Celui-ci utilise l'outil GIZA++ (Och et Ney 2004), qui calcule des modèles statistiques d'alignement de mots de complexité croissante. L'alignement est réalisé dans les deux directions puis ses résultats sont *symétrisés*. Finalement, des *tables de traduction*, qui regroupent l'ensemble des bi-segments pouvant être extraits du corpus, sont construites par application d'heuristiques d'extraction de bi-segments *cohérents*, qui imposent que tout mot d'un segment dans une langue doit être aligné avec au moins un mot du segment dans l'autre langue, mais avec aucun mot en dehors de celui-ci.

4.4.2 Projection

La figure 4.3 présente quelques exemples de bi-phrases alignées au niveau des mots.

FIGURE 4.2 – *Exemples : trois bi-phrases alignées au niveau des mots*

Pour projeter les annotations, nous utilisons le principe suivant :

Soit $E_1 = \{m_{11}, \ldots, m_{1n}\}$ l'ensemble des mots constituant une entité médicale dans le corpus anglais et $E_2 = \{m_{21}, \ldots, m_{2p}\}$ l'ensemble des mots constituant la projection de E_1 (i.e. l'union des projections de chaque mot dans E_1). En notant $position(m_i)$ la fonction retournant la position d'un mot dans sa phrase, nous considérons que la projection de l'entité médicale anglaise est la séquence ordonnée de mots $E_3 = \{m_{31}, \ldots, m_{3k}\}$ telle que :

- $position(m_{31}) = Min_{m_{2i} \in E_2}(position(m_{2i}))$
- $position(m_{3k}) = Max_{m_{2i} \in E_2}(position(m_{2i}))$

Le bruit produit par l'alignement et les annotations affecte la qualité des annotations projetées. Pour diminuer ce bruit et améliorer la phase de projection, (i) nous définissons des heuristiques (e.g. fixer une longueur maximale de l'entité trouvée en français par rapport à la longueur de l'entité originale en anglais) et (ii) nous utilisons un *antidictionnaire* [12] pour filtrer les entités médicales obtenues et supprimer les « mots vides ».

4.5 Expérimentations et évaluation

Le corpus utilisé pour les expérimentations a été construit à partir du site bilingue « Santé Canada [13] » aligné au niveau des phrases (Deléger et al. 2009). La table 4.3 présente le corpus parallèle (anglais-français) SantéCanada.

	Corpus anglais	**Corpus français**
Nombre de mots	4465672	5052543
Nombre de caractères	29845733	33901471
Nombre de caractères par mot (en moyenne)	7	7

TABLE 4.3 – *Le corpus parallèle SantéCanada*

4.5.1 Construction et annotation manuelle d'un bi-corpus de référence

Pour évaluer notre approche, nous avons besoin d'un bi-corpus de référence annoté. Deux éléments sont à déterminer : (i) la taille du corpus de référence et (ii) la manière de choisir ce corpus à partir du corpus initial SantéCanada. Nous nous sommes pour cela basés sur des travaux en statistiques.

12. http://members.unine.ch/jacques.savoy/clef/index.html
13. http://www.hc-sc.gc.ca

Choix du corpus de référence : taille et exemples

Taille du corpus. Pour déterminer la taille (acceptable) du corpus de référence à sélectionner, nous utilisons la formule utilisée en statistiques (Sim et Wright 2005) pour déterminer la taille d'un échantillon :

$$N = \frac{T^2 P(1-P)}{E^2}$$

$$\begin{cases} N & = & \text{La taille de l'échantillon attendu.} \\ T & = & \text{Niveau de confiance déduit du taux de confiance (traditionnellement 1,96 pour un taux de confiance de 95 \%).} \\ P & = & \text{Proportion estimée de la « population » présentant la caractéristique ciblée dans l'étude. Lorsque cette proportion est ignorée, une pré-étude peut être réalisée ou sinon p = 0,5 sera retenue.} \\ E & = & \text{Marge d'erreur (traditionnellement fixée à 5 \%).} \end{cases}$$

Avec les valeurs par défaut ($P = 0,5$, $T = 1,96$ et $E = 0,05$), la valeur de N obtenue est de 385.

Sélection du corpus. Différentes méthodes sont possibles, telles que l'échantillonnage aléatoire simple (*simple random sampling*) ou l'échantillonnage stratifié (*stratified sampling*). Nous avons choisi d'utiliser l'échantillonnage aléatoire simple et sélectionné aléatoirement 385 phrases, contenant 4 613 mots.

Annotation manuelle du corpus de référence choisi

Nous avons annoté manuellement les 385 phrases sélectionnées avec trois types de catégories médicales : Traitement, Problème et Maladie. Nous avons annoté les deux parties françaises et anglaises du corpus de référence en utilisant le guide d'annotation de i2b2 2010 (tâche 1) [14].

4.5.2 Évaluation de l'annotation du corpus anglais

Dans cette section, nous évaluons la reconnaissance des entités médicales à partir du corpus anglais. Nous différencions le cas où les entités médicales ont été reconnues avec des frontières précises ou exactes et le cas où les frontières ne sont pas précises (par exemple, «as antimicrobial resistance» au lieu de «antimicrobial resistance», «Pap smear» au lieu de «a Pap smear» ou «the Pap smear» dans le texte). Nous utilisons les mesures standard de rappel, de précision et de F-mesure.

Nous avons effectué un premier test du module Maxent_SNG pour classification des syntagmes nominaux en entités médicales et entités générales en l'entraînant sur le corpus d'entraînement i2b2 puis en l'évaluant sur le corpus de test i2b2. Nous avons obtenu une correction (proportion d'exemples de test correctement classés) de 90,99 % (16169/17769).

Nous avons ensuite entraîné le module Maxent_SNG sur le corpus ciblé pour la projection (385 phrases extraites de Santé Canada). La table 4.4 présente la contribution du module à la méthode MetaMapPlus.

14. https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf

MetaMapPlus						**Maxent_SNG + MetaMapPlus**					
Frontières strictes			Frontières larges			Frontières strictes			Frontières larges		
R	P	F	R	P	F	R	P	F	R	P	F
61,36	22,37	32,79	82,26	28,18	41,97	50,00	40,23	**44,58**	59,26	45,98	**51,78**

TABLE 4.4 – *Résultats de la méthode MetaMapPlus sans et avec le module Maxent_SNG sur le corpus Santé Canada (partie anglaise).*

Comme attendu, ce filtrage améliore sensiblement la précision des entités médicales détectées, et en dépit d'une baisse importante de la valeur de rappel, la F-mesure connaît une nette augmentation.

En vue de la projection vers un autre corpus à des fins de détection d'entités correctes dans la langue cible, il est fortement souhaitable que les entités du corpus source soient correctes : c'est ce que vise à obtenir le filtrage mis en place, qui privilégie la précision par rapport au rappel.

4.5.3 Évaluation de l'annotation du corpus français par projection

Dans cette section, nous évaluons dans un premier temps la qualité de la projection indépendamment des erreurs d'extraction. Pour ce faire, nous étudions la qualité de la projection des entités de référence (i.e. entités annotées manuellement dans le corpus anglais). Dans un second temps, nous évaluons l'ensemble du processus pour la reconnaissance des entités médicales en français (comprenant l'extraction automatique des entités médicales en anglais et leur projection).

Le tableau 4.5 présente les résultats de la projection des entités de référence et les résultats de la projection des entités extraites avec la méthode MetaMapPlus.

Annotation du corpus français : projection des entités médicales extraites						**Annotation du corpus français : projection des entités médicales de référence**					
Frontières strictes			Frontières larges			Frontières strictes			Frontières larges		
R	P	F	R	P	F	R	P	F	R	P	F
22,39	22,90	22,64	43,08	42,75	42,91	44,78	57,69	50,42	67,91	87,50	76,47

TABLE 4.5 – *Évaluation de la projection des entités médicales extraites avec la méthode MetaMapPlus et les entités de référence sur le corpus Santé Canada (partie française).*

Dans les colonnes frontières larges les mesures de rappel (R), précision (P) et F-mesure (F) sont calculées en comptablisant les entités retrouvées partiellement comme correctes. Dans les colonnes frontières strictes, seules les entités retrouvées exactement (délimitation correcte dans le texte) sont considérées correctes. Nous dédions la section suivante à la discussion de ces résultats et à la synthèse de l'approche de projection présentée.

4.5.4 Discussion

Analyse des résultats et des erreurs

La projection des entités de référence (cf. tableau 4.5) annotées manuellement dans le corpus anglais a permis de retrouver les entités médicales dans le corpus français avec un rappel de 44,78% et une précision

de 57,69% avec les frontières strictes. Un premier constat est donc que la projection permet de retrouver en moyenne la moitié des entités médicales en français. Dans ce cadre, les résultats obtenus par la projection des entités médicales en anglais extraites automatiquement (22,64% de F-mesure) étaient attendus, car la précision de l'extraction automatique des entités en anglais est de 40,23% en frontières strictes (cf. tableau 4.4).

Nous avons pu améliorer les résultats de la méthode MetaMapPlus en intégrant le module MaxEnt entraîné sur trois types différents de corpus. Notons que les résultats obtenus en exploitant ces trois types de corpus sont meilleurs que ceux obtenus en entraînant le classifieur sur un ou deux corpus uniquement (ce que nous avons testé mais ne pouvons pas détailler dans ce chapitre). Les résultats sont relativement acceptables (51,78 % de F-mesure) étant donné la complexité de la tâche sur un corpus hétérogène non spécialisé extrait du Web.

Pour la projection, nous avons utilisé les alignements au niveau des mots avec une approche simple qui consiste à prendre l'entité correspondante (projetée) la plus large. Nous avons essayé d'améliorer cette projection en utilisant un antidictionnaire pour filtrer les entités obtenues et quelques heuristiques telles qu'une différence maximale entre la longueur de l'entité initiale et celle de l'entité projetée (cf. table 4.6).

	Frontières larges	**Frontières strictes**
Projection sans filtrage	79,09 %	**47,15 %**
Projection + antidict.	75,52 %	**49,79 %**
Projection + antidict. + heuristiques	76,47 %	**50,42 %**

TABLE 4.6 – *F-mesure de la projection des entités de référence sans et avec filtrage*

Les résultats de la projection des entités médicales extraites sont relativement faibles, mais ceci dépend directement de la performance de la méthode d'extraction d'information (51,78 % de F-mesure, avec frontières larges), qui fixe le plafond de performance atteignable en projetant ses résultats sur le corpus français. Par projection, nous perdons tout de même près de 50 % des extractions correctes dans le corpus anglais. Ceci résulte principalement de la qualité des alignements au niveau des phrases puis des mots. L'alignement au niveau des mots est influencé par la qualité de l'alignement des phrases (cf. les exemples 1 et 2 ci-dessous). En effet, dans certains cas, la phrase correspondante en français n'est pas équivalente à celle en anglais (soit elle est beaucoup plus courte et contient moins d'information, soit elle est beaucoup plus longue), dans d'autres cas elle a un contenu complètement différent ou reste formulée en anglais : cela reflète un problème d'alignement de phrases qu'il nous faudra corriger dans la suite de nos travaux.

Exemple 1 :

- Statement on Immunization for <PB>Lyme Disease</PB>, 2000 (*)
- 0-0 1-1 5-2 4-3 5-3 5-4 5-5 2-7 5-9 5-10 5-11 5-12 6-13 7-14
- Déclaration sur <PB>un schéma révisé pour la vaccination des adolescents contre l'hépatite B</PB>, 2000 (*)

Exemple 2 :

- <PB>Lung Cancer</PB> : Guidelines for processing Specimens and Reporting Tumor Stage (2000)
- 0-0 1-0 2-0 1-1 1-2 3-3 4-6 9-13 8-18 6-23 7-24 10-27
- <PB>Utilisation, aux fins </PB> de la surveillance, des renseignements sur les patients atteints de cancer : Examen systématique des lois, des réglements, des politiques et des lignes directrices (2000)

Il semble que ces deux phrases ne soient pas en relation de traduction, ce qui peut résulter d'un mauvais appariement de documents ou entre phrases.

Les résultats finaux obtenus par la méthode de projection ne sont pas encore satisfaisants, cependant cette première tentative a quand même permis de localiser 43% des entités médicales du corpus français (rappel avec frontières larges), ce qui constitue un pas encourageant pour pousuivre les recherches dans ce sens, sachant qu'aucun outil spécialisé pour le français n'a été utilisé.

Méthodes statistiques vs. passage à l'échelle

En appliquant la méthode de reconnaissance des entités médicales CRF-BIO-H entrainée sur le corpus i2b2 (cf. chapitre 2) sur le corpus SantéCanada, nous nous sommes confrontés à un inconvénient classique des méthodes statistiques : le problème de portabilité. Ce problème pose une question clé :

- Comment éviter le sur-apprentissage et adapter les méthodes statistiques à des nouveaux corpus d'applications différents de ceux utilisés pour l'entraînement ?

Pour savoir si nous avons un problème de sur-apprentissage, nous avons entraîné et testé le classifieur CRF-BIO sur le même corpus d'entraînement de i2b2. Nous avons obtenu 92.20% de F-mesure. Nous pensons qu'il n'y a pas de problème de sur-apprentissage mais que le problème principal est le fait que les données de test sont très différentes des données d'entraînement et qu'une solution adaptative est nécessaire.

Pour résoudre ce problème, nous proposons une approche similaire à celle utilisée en bootstrapping. En effet, au cas où il n'y a pas assez de données annotées pour apprendre, certains travaux essayent d'apprendre à partir de quelques exemples de départ (seed) annotés manuellement pour apprendre de nouveaux exemples puis les ajouter aux données d'entraînement (bootstrapping). D'autres utilisent plusieurs classifieurs qui « coopèrent » ensemble (co-training). La figure 4.3 décrit la méthode que nous proposons pour adapter notre méthode statistique. L'annotation semi-automatique consiste à annoter les exemples avec la méthode MetaMap-Plus puis à filtrer manuellement les résultats.

Cette approche fait partie des perspectives à court terme de nos travaux.

Conclusion du chapitre

Nous avons présenté dans ce chapitre une approche pour l'annotation automatique de textes médicaux en français par projection depuis l'anglais, et présenté nos premières expérimentations en reconnaissance des

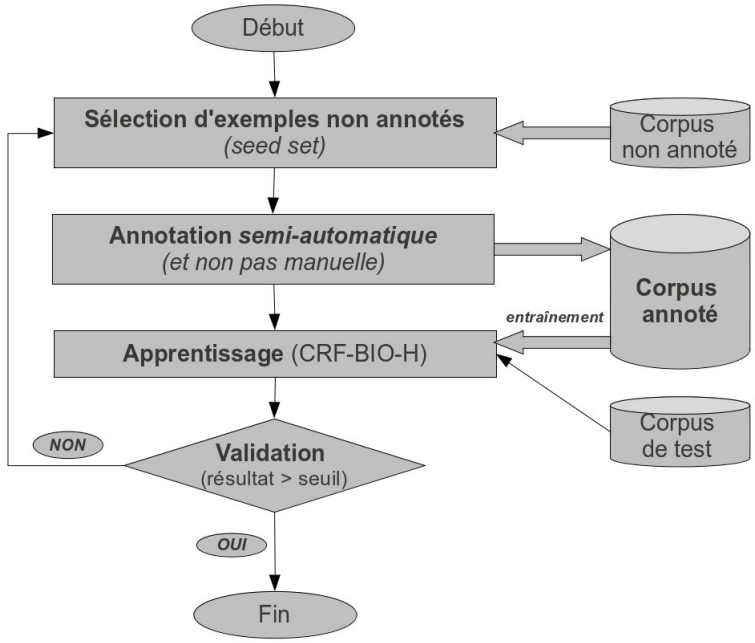


FIGURE 4.3 – *Approche proposée pour l'adaptation de la méthode statistique CRF-BIO-H aux nouveaux types de données*

entités médicales (REM). L'approche présentée utilise un corpus parallèle aligné au niveau des mots pour projeter les annotations obtenues sur la partie anglaise vers la partie française d'un bi-corpus que nous avons construit. L'application de notre méthode de REM sur un grand corpus de données hétérogènes extrait du Web a posé une problématique de passage à l'échelle pour laquelle nous avons proposé une solution qui consiste à intégrer un module de filtrage statistique des entités candidates pour améliorer la précision des entités extraites.

Nous envisageons principalement quatre perspectives à ce travail :

- L'annotation automatique des relations sémantiques dans des textes français en reprenant la méthode présentée.
- L'utilisation ou la construction d'autres corpus médicaux parallèles de meilleure qualité.
- L'exploitation de corpus français annotés pour la mise en place de méthodes statistiques pour l'extraction d'information à partir de textes en français.
- L'intégration de ces méthodes d'extraction d'information dans un système de questions-réponses translingue.

Deuxième partie

Questions-Réponses dans le domaine médical

Nous consacrons cette deuxième à la tâche questions-réponses (cf. figure 4.4). L'approche que nous proposons consiste à analyser une question médicale posée en langage naturel et la transformer en une ou plusieurs requêtes SPARQL représentant l'interprétation de la question avec le vocabulaire de l'ontologie utilisée. Ces requêtes seront utilisées pour interroger des bases de données RDF générées à partir des corpus médicaux utilisées pour trouver les réponses.

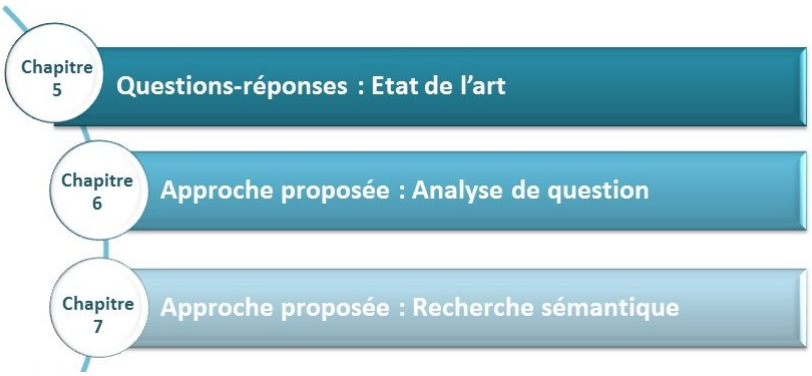


FIGURE 4.4 – *Partie 2 : Questions-réponses dans le domaine médical*

Le chapitre 5 présente un état de l'art sur les systèmes de questions-réponses en domaine ouvert et aussi en domaine médical où nous parcourons les différents travaux et les différentes ressources disponibles pour ce domaine de spécialité.

Le chapitre 6 présente notre méthode d'analyse de questions. Nous proposons une méthode permettant de traiter des questions médicales avec plusieurs focus et/ou type(s) de réponse attendue. Nous évaluons notre approche sur une collection de questions médicales réelles.

Le chapitre 7 sera consacré à la recherche et l'extraction de réponses. Nous présentons la méthode utilisée pour l'annotation en RDF des corpus médicaux utilisés pour trouver les réponses et notre approche pour l'interrogation de ces annotations et le classement des réponses. La dernière partie de ce chapitre sera dédiée à l'évaluation du système de questions-réponses final MEANS.

Etat de l'art sur les systèmes de questions-réponses

5

Sommaire

Dans ce chapitre nous présentons un état de l'art sur les Systèmes de Questions-Réponses (SQR) en domaine ouvert dans un premier temps. Dans un second temps, nous nous intéressons aux travaux antérieurs sur les SQR en domaine médical ainsi qu'aux ressources sémantiques disponibles pour ce domaine.

5.1 Introduction

La recherche de réponses précises à des questions formulées en langue naturelle est un des défis majeurs posés dans le champ de la recherche d'information. Les premiers systèmes de questions-réponses tels que **BASEBALL** (Green et al. 1961) et **LUNAR** (Woods 1973) ont vu le jour dans les années 60. Le système BASEBALL traite les questions qui portent sur les dates, les lieux ou les résultats des matchs de Baseball du championnat américain. LUNAR est un des premiers systèmes de questions-réponses scientifiques. Il a été conçu pour assister l'analyse géologique des pierres retournées par la mission Apollo. Lors de son expérimentation il a pu répondre avec succès à 90 % des questions posées par les utilisateurs. Les deux systèmes BASEBALL et LUNAR exploitaient des bases de connaissances écrites manuellement par des experts de leurs domaines d'application.

Un système de questions-réponses vise à fournir une réponse à une question posée en langage natuel. La réponse visée peut aussi bien être un morceau de texte extrait d'une collection de documents ou provenant du Web qu'une donnée récupérée d'une base de données ou base de connaissances. Dans d'autres cas plus rares, les réponses retournées sont des informations multimédia. C'est le cas par exemple du système START (Katz 1999), introduit en 1993 comme un « serveur d'information », qui permet de répondre à des questions écrites en anglais avec des informations multimédia. Il utilise des techniques issues du TAL pour analyser les phrases et associe des annotations linguistiques à des segments d'information multimédia. Ces annotations sont ensuite exploitées pour retourner les bons segments d'information aux questions posées par les utilisateurs.

Dans un cadre général, un SQR peut être décomposé en trois tâches principales : (i) Analyse de la question utilisateur, (ii) Analyse du corpus à partir duquel les réponses vont être extraites, et (iii) recherche et extraction des réponses en sélectionnant le meilleur appariement question-réponse. La deuxième tâche ne sera pas requise pour les systèmes qui extraient leurs réponses à partir de bases de données ou de bases de connaissances structurées. Les méthodes employées pour l'analyse de questions et/ou de corpus peuvent être surfaciques (e.g. fondées sur l'indexation des mots clés de la question et du corpus), sémantiques (i.e. formalisation explicite de tout ou partie du sens de la question et du corpus) ou hybrides (combinant les deux approches).

Cet état de l'art n'est pas dédié aux techniques d'extraction d'information (i.e. extraction d'entités nommées et de relations) à partir des questions ou des corpus textuels : ce point a été abordé dans la première partie de cette thèse. Nous nous plaçons ici au niveau du processus complet et décrivons quelques uns des travaux les plus significatifs sur les systèmes de questions-réponses « sémantiques ». La première partie de cet état de l'art est dédiée à la classification des systèmes de questions-réponses. La deuxième partie est dédiée aux systèmes de questions-réponses en domaine ouvert et la troisième partie porte sur les SQR proposés pour le domaine médical.

5.2 Classification des systèmes de questions-réponses

Plusieurs classifications ont été proposées pour les systèmes de questions-réponses. Par exemple, Moldovan et al. (2003) se sont appuyés sur la complexité des questions traitées et la difficulté de la tâche d'extraction de réponses pour proposer cinq classes de systèmes avec une complexité croissante :

1. Systèmes capables de répondre à des questions factuelles.
2. Systèmes employant des processus de raisonnement simples.
3. Systèmes qui constituent les réponses à partir de plusieurs sources à la fois.
4. Systèmes qui proposent un dialogue interactif avec l'utilisateur.
5. Systèmes capables d'effectuer un raisonnement analogique.

La première catégorie de systèmes proposée couvre un grand nombre de systèmes traitant les questions WH (What, Who, Where, etc.). Dans la classification proposée cette catégorie regroupe les approches qui extraient des réponses constituées de morceaux de textes provenant d'un ou de plusieurs documents. La deuxième catégorie de systèmes est une sous-classe de la première catégorie où l'extraction de réponses nécessite d'effectuer des inférences logiques. Dans la troisième classe de systèmes proposée, la réponse est éparpillée dans plusieurs documents et une fusion est nécessaire pour la constituer. Le quatrième type de systèmes exploite des interactions précédentes avec l'utilisateur pour extraire les réponses. Enfin la dernière catégorie, jugée plus complexe, consiste à extraire des réponses qui ne sont pas explicites dans les documents. C'est le cas, par exemple, des questions de type prédiction (e.g. « Is the Fed going to raise interests at their next meeting ? »). Afin de répondre à de telles questions, les SQR doivent effectuer un raisonnement par analogie complété potentiellement par un raisonnement temporel, spatial ou conditionnel.

Cependant cette approche n'utilise pas les entrées/sorties comme critère de classification des SQR car elle s'est fixé un cadre avec toujours le même type d'entrée (question en langage naturel) et le même type de sorties (extraits textuels). Elle divise aussi les méthodes exploitant des inférences/raisonnement en raisonnement simple et raisonnement avancé (par analogie), ce qui constitue un critère de classification de précision élevée qui ne peut pas être évalué pour tous les systèmes. Aussi les types de techniques utilisées pour l'analyse de la question et/ou des corpus et l'appariement questions/réponses ne sont pas exploités comme critère de classification. Enfin, cette classification n'explicite pas non plus les sources desquelles les réponses sont collectées ou extraites.

D'autres travaux proposés dans la littérature explicitent les sources des réponses et les entrées/sorties des SQR comme critères de classification. Par exemple, Lopez et al. (2011) proposent un état de l'art sur une sous-catégorie des systèmes de questions-réponses qui exploitent des métadonnées sémantiques présentes dans des bases de connaissances structurées et des ontologies. Ils proposent pour cela deux classifications des systèmes à base d'ontologies. Une première classification en trois catégories :

- interfaces en langage naturel pour les bases de données,
- questions-réponses à partir de documents textuels,

– questions-réponses avec des données/textes/langages propriétaires,
et une deuxième classification suivant les sources des réponses :
– bases de données structurées,
– textes non structurés,
– bases de connaissances sémantiques précompilées.

Athenikos et Han (2010) ont proposé une classification des techniques de questions-réponses basées sur des connaissances sémantiques en trois catégories :
– SQR sémantiques,
– SQR basés sur les inférences,
– SQR fondés sur des représentations logiques.

Cette classification reste cependant ambiguë car les trois catégories ont beaucoup d'intersections ce qui ne permet pas de trier efficacement les différents systèmes. Athenikos et Han (2010) proposent aussi une classification des SQR médicaux qu'ils partagent en deux catégories : (i) Approches basées sur des connaissances sémantiques et (ii) Approches non sémantiques.

Sur ces trois travaux, notre observation est que l'objectif des auteurs n'était pas de définir des critères qui permettraient de trier/étiqueter efficacement les SQR mais plutôt de donner quelques grandes directions de recherche et synthétiser les challenges actuels et ceux à venir pour les systèmes de questions-réponses. Ainsi, bien que ces travaux fournissent des descriptions et des résumés pertinents pour chaque SQR individuellement, les trois classifications de SQR proposées présentent certaines ambiguïtés. Par exemple, le type de représentation de la question (e.g. SQL, XQuery, SPARQL, formules logiques) et le type de représentation des informations extraites des corpus (e.g. XML, relationnel, RDF, textuel) ne sont pas explicités comme critères de classification. Les techniques d'extraction d'information appliquées sur la question en langage naturel ne sont pas précisées dans la catégorisation des systèmes (e.g. patrons utilisant les catégories morpho-syntaxiques des mots, les arbres de dépendance, techniques d'apprentissage artificiel). Pour les SQR sémantiques, les éventuelles techniques de désambiguïsation du sens des mots ne sont pas non plus prises en compte comme critère de classification.

5.3 Travaux existants en domaine ouvert

Les approches sémantiques en domaine ouvert ciblent des entités nommées qui ne sont pas spécifiques à un domaine particulier (e.g. personne, lieu, organisation, date). Les relations représentées et recherchées entre ces entités peuvent aussi bien être syntaxiques (e.g. sujet-verbe-objet), sémantiques (e.g. <entité-1> est le président de <entité-2>) ou combinées (e.g. rechercher les triplets sujet-verbe-objet dont le sujet est une entité de type *Personne*, l'objet est une entité de type *Organisation* et le verbe est un synonyme de *présider*).

Plusieurs approches sémantiques ont été proposées dans la littérature pour concevoir des systèmes de questions-réponses en domaine ouvert. Ces systèmes diffèrent par les méthodes employées pour l'analyse de la question et l'extraction d'information à partir des corpus mais aussi par la

nature et les sources des réponses et les représentations formelles utilisées pour l'interprétation des informations extraites.

Dans une première partie de cette section nous nous intéressons aux SQR en domaine ouvert qui extraient des réponses à partir d'une collection de documents (ou du Web). Dans une seconde partie, nous nous intéressons aux travaux qui extraient les réponses à partir de bases de données ou de bases de connaissances.

5.3.1 Extraction de réponses à partir d'une collection de documents

Dans cette catégorie de SQR, les méthodes d'extraction d'information sont communes à la tâche d'analyse de question et de corpus. Il s'agit le plus souvent de détecter les entités nommées et potentiellement les relations syntaxiques et/ou sémantiques qui les lient dans les phrases du corpus et dans la question de l'utilisateur. L'analyse de question comprend, en plus de ces éléments, les tâches classiques de typage de la question et de détermination du type de la réponse attendue (e.g. évènement, lieu).

Par exemple, le système LASSO (Moldovan et al. 1999) répond aux questions d'utilisateurs en retournant les morceaux de texte susceptibles de contenir les réponses. Pour cela le système analyse la question en quatre étapes :

1. détection du type de la question (what, why, how, where)
2. détection du type de la réponse (personne, lieu, etc.)
3. détection du focus de la question (information principale requise par l'interrogation)
4. extraction des mots-clés pertinents de la question

Cependant tous les mots de la question ne sont pas forcément contenus dans les réponses. Par exemple, si la question recherche le deuxième mois de l'année chinoise, la réponse ne contient pas forcément l'expression « mois de l'année ». LASSO propose une solution à cette problématique en employant des heuristiques de reconnaissance d'entités nommées pour extraire les réponses avec des entités proches de celles citées dans la question. Les questions sont aussi analysées en se référant à une hiérarchie de types de questions construite à partir des données d'entraînement du challenge TREC-8 [1] en se focalisant sur le type de réponse attendue.

D'autres systèmes utilisent des techniques de Traitement Automatique de la Langue (TAL) pour extraire des réponses aux questions factuelles en domaine ouvert. Par exemple, QALC (Ferret et al. 2002) est un SQR qui répond à des questions formulées en anglais en effectuant une analyse syntaxique de la question et du corpus. Il détecte les paraphrases dans le corpus en utilisant le « focus » de la question comme pivot afin d'atteindre des formulations différentes de la réponse. Dans une première phase, les documents pertinents pour la question sont extraits en recherchant ceux contenant les termes de la question et leurs variations sémantiques. Les réponses sont ensuite extraites des phrases candidates en fonction des paraphrases possibles. Le système QALC a aussi été la base des systèmes FRASQUES (Grau et al. 2006b) (adaptation de QALC pour le français),

1. La piste QA de TREC a été un challenge annuel pour les systèmes de questions-réponses en domaine ouvert de 1999 à 2007.

MUSQAT (questions en français et réponses extraites de textes en anglais) et MUSCLEF (combinaison des systèmes MUSQAT et QALC). Plus de détails sur ces trois systèmes sont décrits dans (Grau et Chevallet 2007).

FIDJI (Moriceau et Tannier 2010) est un SQR conçu pour le français qui combine l'analyse syntaxique, l'extraction des entités nommées et la pondération des termes. Il indexe le corpus avec un moteur de recherche classique et valide les réponses sur plusieurs documents à la fois. Des techniques différentes sont ensuite employées suivant le type de la question (e.g. booléenne, liste).

Indépendamment de la méthode employée (e.g. statistique, à base de patrons, mots-clés), l'avantage de l'extraction de réponses à partir d'un ensemble de document est que la réponse est automatiquement fournie avec sa justification (les entités et les relations étant extraites à partir des textes). Cependant, se focaliser sur des corpus particuliers réduit la capacité de passage à l'échelle de ces approches (e.g. sur-apprentissage, patrons linguistiques de couverture restreinte). D'un autre côté, plusieurs approches qui effectuent l'extraction de réponses à partir d'une collection de documents ne proposent pas de stratégies pour l'acquisition cumulative des connaissances et n'utilisent pas de référentiels communs (e.g. base de connaissances, thesaurus) qui permettraient de résoudre les éventuelles coréférences en les associant à un même élément du référentiel commun (e.g. concept d'un thésaurus, individu d'une base de connaissances).

5.3.2 Extraction de réponses à partir de bases de données

La problématique des questions-réponses a aussi été abordée dans le cadre des bases structurées où il s'agit de répondre à une question posée en langage naturel à partir d'une base de données relationnelle ou d'une base de connaissances.

Ainsi, plusieurs travaux ont été menés pour la transformation des questions en requêtes SQL. Par exemple, le système PRECISE (Popescu et al. 2003) analyse les questions et essaie de les convertir en des requêtes SQL. Il cherche un appariement de graphes dans lequel les mots de la question sont appariés avec les relations de la base de données, leurs colonnes et leurs valeurs. Ils définissent pour cela l'ensemble des questions interprétables par leur système, qu'ils appellent « semantically tractable questions ». Dans une première phase hors ligne, ils associent chaque attribut de la base de données à une valeur 'wh' (what, where, etc.). Pendant l'interrogation, ils utilisent une base lexicale (lexicon) pour aligner les termes de la question avec des noms d'attributs des relations de la base de données. Les éléments ainsi récupérés de la base de données sont utilisés pour construire une requête SQL. Si plusieurs requêtes sont possibles pour la question, le système demande à l'utilisateur de choisir parmi les interprétations possibles.

Cependant, le système PRECISE pose la contrainte que les mots de la question doivent tous être distincts pour pouvoir retrouver un appariement avec la base de données. Les questions avec des mots inconnus ne sont pas « traçables sémantiquement » et ne sont pas gérées, cela fait que le système n'est pas capable de répondre à des questions employant des mots non répertoriés dans sa base lexicale. Les expérimentations de

PRECISE sur un ensemble de questions utilisateur ont abouti à une précision de 80 %. Les 20 % d'erreur sont dues au fait qu'un cinquième des questions n'est pas représentable par le système.

D'autres approches se basent sur le sens des mots dans la requête pour rechercher des réponses à partir de bases de connaissances préconstruites. Ces bases de connaissances servent aussi à désambiguïser le sens des mots de la question et leurs relations.

Par exemple, START (Katz et al. 2002) répond aux questions portant sur la géographie avec une précision de 67 % sur 326 000 requêtes. Il utilise des bases de connaissances enrichies manuellement pour retrouver des triplets de la forme sujet-relation-objet. L'extraction de la réponse se fait en comparant directement la question de l'utilisateur avec les annotations de la base de connaissances. Cependant, toutes les questions ne peuvent pas être représentées avec des relations binaires. L'acquisition de nouvelles connaissances est aussi une limitation du système car seuls des experts du domaine peuvent ajouter des connaissances et augmenter la couverture.

AquaLog (Lopez et al. 2007) permet à ses utilisateurs de choisir une ontologie de domaine puis de poser des questions en langage naturel dans son vocabulaire. Le système reste cependant relativement générique car sa personnalisation pour une ontologie donnée se fait automatiquement en un temps négligeable. AquaLog effectue un appariement lexical pour transformer la question en une représentation sous forme de triplets puis essaie de faire correspondre les triplets avec les relations appropriées de l'ontologie utilisée. Le système détecte simultanément les concepts, les instances de concept et les relations de la question avec un processus d'apprentissage interactif dans lequel l'utilisateur intervient si une ambiguïté n'a pas pu être résolue automatiquement grâce au contexte des termes. Pour extraire les réponses, AquaLog utilise un moteur d'inférence pour rechercher la représentation en triplets de la question dans la base de connaissances. Cependant le traitement de questions avec plusieurs relations (interprétables en plusieurs triplets) n'est pas étudié et le type d'inférence utilisé n'est pas mentionné.

FREyA (Damljanović et al. 2010) exploite l'arbre de dépendances syntaxiques extrait de la question de l'utilisateur, obtenu avec le Stanford parser, pour identifier le type de la réponse attendue et extraire des réponses pertinentes. Le système FREyA assiste l'utilisateur pendant la formulation de sa question avec plusieurs boîtes de dialogues dédiées à la clarification des ambiguïtés. Les choix effectués par l'utilisateur pendant cette interaction sont sauvegardés puis utilisés pour entraîner davantage les processus d'extraction d'information et de désambiguïsation. FREyA emploie aussi des ontologies de domaine pour apprendre des règles génériques par inférence (e.g. associer une même règle à d'autres classes/concepts par subsomption). Ainsi, les informations extraites de la question utilisateur ont la forme d'annotations référant à une ontologie de domaine. Ces annotations sont générées grâce à un « gazetteer » associé à l'ontologie, qui est le seul garant de leur qualité. Les suggestions proposées à l'utilisateur via les boîtes de dialogue sont triées selon qu'elles sont synonymes ou non des termes employés par l'utilisateur et selon leur similarité lexicale avec ces termes. Dès qu'une suggestion est choisie par l'utilisateur, le système apprend à mieux trier les propositions correctes, pour les placer au dé-

but de la liste des suggestions lors de questions futures similaires. Dans une phase finale (i) des modèles de triplets (patrons de triplets RDF) sont construits à partir de la question annotée en prenant en compte les domaines et les co-domaines des relations déclarées dans l'ontologie et (ii) une requête SPARQL est construite en combinant les modèles de triplets générés.

QACID (Ferrández et al. 2009) est aussi un SQR qui exploite une ontologie comme référence. Sa particularité est qu'il essaie d'associer les nouvelles questions des utilisateurs à des clusters de questions. Ces clusters sont construits en analysant et en groupant les questions existantes. Chaque cluster regroupe différentes formulations de la même question et leur associe manuellement une ou plusieurs requêtes SPARQL. Dans QACID, les questions utilisateurs sont considérées comme des « sacs de mots » et une correspondance est effectuée entre les mots de la question et les instances de la base de connaissances avec des mesures de similarité lexicales et un lexique ontologique (ensemble de termes associés aux concepts et relations de l'ontologie). .

ORAKEL (Cimiano et al. 2008) est un SQR construit pour les domaines de spécialité. Son application à un domaine donné nécessite l'intervention d'experts du domaine pour construire/ajouter un lexique adapté et le rattacher à une ontologie pertinente. ORAKEL transforme les questions des utilisateurs en des clauses formulées en logique du premier ordre en exploitant les prédicats définis par les experts de domaine. Ces clauses sont ensuite réécrites dans le langage d'interrogation de la base de données ou de la base de connaissances du domaine. Cette écriture/transformation utilise des règles écrites manuellement en Prolog pour chaque domaine ciblé.

Dans un ordre plus général, la performance des méthodes d'extraction de réponses à partir de bases de données reste variable suivant les ressources disponibles et faible pour les domaines ayant une faible couverture/description préalable. Elles présentent cependant le grand avantage de pouvoir intégrer facilement de nouvelles données/connaissances grâce à leur format standardisé. D'un autre côté, le défi reste posé quant à pouvoir apporter des justifications à ces réponses. L'intérêt de ces justifications est d'autant plus primordial que certains formats de données ne sont pas lisibles/compréhensibles pour le grand public (e.g. triplets RDF, tuples d'une base de données relationelle).

5.4 Travaux existants en domaine biomédical

Les SQR spécialisés doivent s'adapter à leur domaine d'application et/ou profiter de ses caractéristiques. Le domaine médical possède plusieurs traits plus ou moins spécifiques, citons par exemple :

- les ressources sémantiques utilisées (e.g. les terminologies spécialisées),
- les questions traitées (complexes, simples, avec plusieurs focus, plusieurs TRA),
- les entités médicales traitées,
- les relations médicales ,
- référentiels (e.g. ontologies, taxinomies, réseaux sémantiques),

– les types de documents ciblés.

Les ressources disponibles pour le domaine médical peuvent servir à pallier certaines caractéristiques telles que la polysémie ou la synonymie du vocabulaire médical, voire être utilisées directement pour l'extraction d'information à partir de corpus ou de questions médicales. Les principales méthodes et techniques d'extraction d'information en domaine médical ont été présentées aux chapitres 2, 3 et 4.

Dans cette section, nous présentons dans un premier temps une comparaison entre les caractéristiques du domaine ouvert et les caractéristiques du domaine médical. Dans un second temps, nous présentons les taxinomies (classifications) de questions proposées pour le domaine médical. Enfin, nous décrivons quelques uns des travaux les plus significatifs pour les SQR médicaux.

5.4.1 Domaine ouvert vs domaine médical

Pour mettre en évidence la problématique des SQR en domaine médical, nous examinons les différences entre un SQR en domaine ouvert et un SQR spécialisé pour des questions médicales. Pour ce faire, nous avons testé le SQR généraliste FIDJI (Moriceau et Tannier 2010), développé au LIMSI, sur une collection de questions médicales (la liste EQueR médicale) et un corpus médical (le corpus EQueR médical). . Nous avons ensuite analysé les résultats fournis, les questions et des extraits du corpus. En se basant sur ces analyses, nous avons relevé plusieurs différences entre le domaine ouvert et le domaine médical que nous classons en quatre catégories.

Entités nommées (EN). Il s'agit d'unités textuelles désignant un objet précis. Les entités nommées classiques désignent les noms de personnes, de lieux, d'organisations, les dates et les quantités.. Pour le domaine médical, les EN correspondent aux termes du domaine, à savoir les noms de maladies, de médicaments, etc. La reconnaissance des EN, qui consiste en leur identification ainsi que leur catégorisation, est une étape essentielle lors de l'analyse de la question et de l'extraction de la réponse.

Classification de la question et détermination du type de la réponse attendue. En domaine ouvert, la classification se base principalement sur le pronom interrogatif (qui, quand, etc.) et sur les EN de la question. Elle permet de déterminer le type de la réponse attendue. Par exemple, la question « Quand se déroule le bac 2012 ? » attend une réponse de type Date. Cependant, une classification des questions basée sur le pronom interrogatif n'est pas appropriée au domaine médical. Une question avec Quand, par exemple, peut avoir plusieurs types de réponse attendue. Dans la question « Quand devrais-je prendre de l'Hiconcil ? », le pronom Quand réfère à des conditions d'utilisation d'un médicament et dans « Quand devrais-je prendre mon antibiotique ? », Quand réfère à un temps relatif (une heure après le repas, etc.). D'autre part, les questions médicales portent souvent sur des connaissances médicales : des diagnostics, des symptômes, etc. Par exemple, le type de la réponse attendue pour la question « Comment traiter l'obésité chez les enfants ? »

est Traitement et celui de la question « Comment diagnostiquer la maladie d'Alzheimer ? » est Examen.

Informations contextuelles. Les questions médicales contiennent souvent des informations relatives au patient (e.g. « Comment traiter la bronchiolite chez les enfants de moins de 2 ans ? »). Il est indispensable de tenir compte de ces informations pour trouver la bonne réponse. En domaine ouvert, d'autres types d'informations contextuelles sont importants à repérer dans une question, comme le contexte temporel (e.g. « Quels rois ont régné en France entre le XV^{e} et le XVI^{e} siècle ? »).

Recherche de documents et analyse sémantique. En domaine ouvert, un SQR peut se baser sur la fréquence d'occurrence de l'information dans des grandes bases de données (ou sur le web) et sur des méthodes classiques de recherche d'information. En domaine médical, des sources d'information plus fiables et plus crédibles sont nécessaires. D'un autre côté, se baser sur la redondance des informations n'est pas suffisant vu que les questions médicales nécessitent des interprétations sémantiques plus profondes. Par exemple, dans une question comportant les deux termes T_1 de type médicament et T_2 de type maladie, il est important de connaître la relation sémantique entre T_1 et T_2 car chercher si T_1 complique T_2 est différent de chercher si T_1 traite T_2.

5.4.2 Taxinomies de questions médicales

Plusieurs travaux dans la littérature biomédicale ont proposé des taxinomies pour les questions médicales. Par exemple, Ely et al. (2000) ont proposé une taxinomie qui comporte les 10 catégories de questions les plus fréquentes dans le domaine médical à partir de 1396 questions collectées. Nous citons ici les cinq premiers modèles ou catégories (qui représentent 40 % des questions) de leur étude :

1. What is the drug of choice for condition X ?
2. What is the cause of symptom X ?
3. What test is indicated in situation X ?
4. What is the dose of drug X ?
5. How should I treat condition X (not limited to drug treatment) ?

Ely et al. (2002) ont proposé une autre taxinomie qui classifie les questions en : *Clinical vs Non-Clinical*, *General vs Specific*, *Evidence vs No Evidence*, et *Intervention vs No Intervention*. Yu et al. (2005) ont utilisé la taxinomie de (Ely et al. 2002) pour la classification automatique des questions médicales en se basant sur l'apprentissage supervisé. Leur approche a montré que l'utilisation de la taxinomie de questions avec un classifieur SVM et le metathesaurus de l'UMLS donne les meilleurs résultats parmi plusieurs autres configurations testées dans leur évaluation.

Jacquemart et Zweigenbaum (2003) ont collecté cent questions cliniques en chirurgie buccale et les ont utilisées pour proposer une autre taxinomie de modèles de questions médicales. Trois modèles en particulier ont permis de couvrir 90 % des questions. Ces modèles sont :

1. Quel [X]-(r)-[B] ou [A]-(r)-[quel Y]
2. Est-ce que [A]-(r)-[B]
3. Pourquoi [A]-(r)-[B]

où (r) désigne une relation et A, B, X et Y désignent des entités médicales données (A, B) ou recherchées (X, Y). Ces taxinomies de questions ont quelques limites au niveau de l'expressivité. Par exemple, la taxinomie de (Ely et al. 2000) fournit uniquement quelques formes particulières d'expressions pour certains types de questions, alors qu'en réalité d'autres formes existent pour ces mêmes types. Dans la taxinomie proposée par (Jacquemart et Zweigenbaum 2003)) les relations sémantiques ne sont pas complètement exprimées. Cela permet de couvrir un plus grand nombre de questions, ce qui correspond au but de modélisation, mais nécessite un travail supplémentaire pour réaliser l'appariement questions-réponses avec des relations particulières dans le cadre des systèmes de questions-réponses.

5.4.3 Systèmes de questions-réponses en domaine médical

Plusieurs approches ont été proposées pour la réalisation de SQR dans le domaine biomédical. Par exemple, le système ExtrAns est un système conçu pour s'adapter aux domaines restreints qui a été expérimenté sur des articles de recherche en génomique (Rinaldi et al. 2004). Les documents de ce corpus médical sont analysés hors ligne avec des techniques de TAL puis transformés en une représentation sémantique, appelée « Minimal Logical Form », sauvegardée dans une base de connaissances. Pendant la phase d'interrogation, la question de l'utilisateur est analysée avec la même méthode utilisée pour l'analyse du corpus. Le système recherche ensuite un appariement entre la représentation sémantique de la question utilisateur et la base de connaissances. Si un appariement est découvert, les phrases qui ont permis cet appariement sont retournées comme réponses possibles à la question.

Terol et al. (2007) utilisent la taxinomie de questions proposée par (Ely et al. 2000) et ciblent uniquement les dix catégories de questions les plus fréquentes dans leur SQR médical. Ce système procède en quatre étapes : (i) analyse de question, (ii) recherche de documents en ligne et/ou en local, (iii) sélection de passages pertinents et (iv) extraction de réponses. L'analyse de la question et des documents est effectuée en dérivant des formes logiques à partir des relations de dépendance entre les mots de la phrase/question en utilisant la codification de la version étendue de WordNet (i.e. les prédicats de la forme logique ne relient pas les mots mais les ensembles de synonymes auxquels ils appartiennent). Les réponses sont extraites parmi les phrases des passages sélectionnés en comparant la forme logique de la question à la forme logique de la phrase.

Embarek (2008) a travaillé sur la mise en place d'un système de questions-réponses pour le domaine médical. Il a proposé une approche comportant deux modules adaptés à ce domaine de spécialité. Le premier module, Analyse de question, utilise un ensemble de règles construites manuellement pour déterminer la relation principale de la question, en partant de l'hypothèse qu'une part significative des questions médicales

peuvent être modélisées sous la forme d'une relation unissant une ou plusieurs entités explicitement instanciées dans la question et une entité absente correspondant à la réponse cherchée (ce qui est conforme au premier modèle de (Jacquemart et Zweigenbaum 2003)). Ces règles exploitent plusieurs informations extraites à partir de la question comme le pronom interrogatif, les lemmes et les catégories morpho-syntaxiques des mots de la question ainsi que les catégories sémantiques des entités médicales présentes dans la question. Le deuxième module cherche la réponse dans une ontologie médicale, construite à partir de ressources sémantiques disponibles, qui représente les entités et les relations les plus communément manipulées en médecine. Si la réponse n'est pas trouvée, une autre méthode est appliquée qui se base sur des patrons linguistiques appris de façon supervisée pour repérer et extraire la réponse à partir d'une collection de documents. Ces patrons, appelés patrons multi-niveaux, sont des expressions régulières pouvant faire référence à trois niveaux d'information sur les mots : la forme fléchie, la catégorie morpho-syntaxique et le lemme. Cette approche a été utilisée pour adapter un SQR généraliste existant (Œdipe) au domaine médical et mettre en place le SQR Esculape spécialisé dans le domaine médical.

Le système EPoCare (Niu et Hirst 2004) effectue une recherche par mots-clés dans une base de documents XML pour répondre aux questions utilisateurs posées en langage naturel. Plus précisément la question de l'utilisateur est traduite par le système en une requête formée de mots-clés combinant les termes recherchés et les catégories médicales dans le format PICO. Ce format distribue les caractéristiques d'une question dans quatre champs : les problèmes médicaux et les patients (P), les interventions (I), les relations de comparaison (C) et les résultats d'interventions (outcome : O). Le processus d'extraction de réponses utilise un ensemble de patrons de réponses préconstruits qui consistent en des chemins XML pour chacune des quatre catégories définies dans le format PICO. Le système identifie les chemins XML qui contiennent tous les mots-clés de la question tout en filtrant les chemins qui ne correspondent pas aux catégories PICO associées à ces mots-clés.

Dans un travail qui précède EPoCare, Niu et al. (2003) ont proposé une approche pour répondre à des questions médicales reposant sur l'identification des catégories PICO. Le processus employé consiste à déterminer les frontières textuelles de chaque catégorie puis à identifier les relations entre les différentes catégories détectées dans la phrase/question. Les relations traitées étaient restreintes aux relations thérapeutiques (traitement).

Demner-Fushman et Lin (2005) ont aussi proposé une approche qui utilise le format PICO pour décrire leurs composants d'extraction de connaissances sémantiques dans le cadre d'un SQR médical. Ils identifient les éléments correspondant aux catégories PICO dans les résumés d'articles de MEDLINE.

Demner-Fushman et Lin (2006) ont proposé une approche hybride pour les SQR en domaine médical qui repose sur des techniques issues de la recherche d'information et du résumé automatique de textes. Ils ont ciblé une classe fréquente de questions de la forme « What is the best drug treatment for X ? ». Étant donné un ensemble initial de résumés de MEDLINE retrouvés avec PubMed, le système identifie d'abord

les médicaments en cours d'étude en utilisant le composant d'extraction d'interventions proposé dans (Demner-Fushman et Lin 2005). Il regroupe ensuite dans des clusters sémantiques les résumés MEDLINE retrouvés en exploitant (i) les interventions principales identifiées dans le texte du résumé et (ii) un algorithme de clustering agglomératif hiérarchique afin de calculer les similarités entre les interventions. Pour chaque résumé de MEDLINE, le système produit un résumé plus court composé de l'intervention principale, du titre du résumé et de la meilleure phrase citant un résultat d'intervention, calculée par le composant d'extraction de résultats d'interventions présenté dans (Demner-Fushman et Lin 2005).

Cependant, comme observé par Sackett et al. (2000), la représentation PICO ne permet pas d'exprimer la sémantique complète de la question en langage naturel. Huang et al. (2006) ont examiné davantage le format PICO en étudiant sa compatibilité avec les questions cliniques posées en langage naturel en classifiant manuellement un ensemble de cent questions. Les principales observations qu'ils ont effectuées sur l'adéquation du format avec les questions évaluées sont que la représentation PICO est mieux adaptée pour traiter les questions de type *therapy* (traitement) et moins adéquate pour les questions de type *diagnosis* (diagnostic), *etiology* (cause) et *prognosis* (pronostic). Par ailleurs, le format PICO ne permet pas de reconstruire la question originale à partir des informations extraites (*P*, *I*, *C* et *O*). Ceci est dû principalement à l'incapacité d'exprimer des relations précises entre les entités médicales (e.g. est-ce que `[Problem: hypomagnesemia, Intervention: ?]` correspond à « What is the most effective treatment for hypomagnesemia ? » ou « What are the causes of hypomagnesemia ? »). D'autres limitations incluent l'ambigüité de ce format (e.g. il représente le problème médical et aussi la population par le même élément *P*) et il n'est pas capable d'identifier les relations anatomiques.

5.5 Synthèse et positionnement

Nous proposons à travers cette thèse un système de questions-réponses sémantique pour le domaine médical. Nous pensons que les approches sémantiques sont les plus appropriées pour ce domaine d'application car elles permettent de suivre son évolution en termes de connaissances, notamment parce qu'elles permettent d'interroger des bases de connaissances provenant de sources différentes ou construites de multiples façons. En effet, ces bases peuvent aussi bien être construites à partir de bases de données publiques ou propriétaires (e.g. BIO2RDF (Belleau et al. 2008)) que par l'annotation de corpus médicaux. Aussi, dans le cadre du projet Linked Open Data [2] plusieurs bases de connaissances biomédicales ont été publiées en ligne (e.g. BioGateway [3], DrugBank [4]). L'avantage ici étant que l'intégration se fait avec des langages standards qui donnent une formalisation unique aux liens entre les différentes bases et facilite ainsi leur partage et leur accès (e.g. les axiomes standards OWL [5] sameAs pour la ré-

2. http://linkeddata.org
3. http://www.semantic-systems-biology.org/biogateway
4. http://www4.wiwiss.fu-berlin.de/drugbank/
5. http://www.w3.org/2004/OWL/

conciliation d'individus ou equivalentClass pour l'alignement de concepts ontologiques).

Cependant, bien que l'interrogation directe de ces bases permette de répondre aux questions des utilisateurs avec des faits (triplets), les éléments retournés en réponses (e.g. URIs de concepts ontologiques ou d'instances de ces concepts) ne sont pas forcément pertinents ou parlants pour les utilisateurs. Ainsi, le fait de pouvoir justifier ou retrouver une réponse grâce à un corpus textuel devient important. Dans notre vision, ces deux types de réponses ne sont pas disjoints mais complémentaires. Un SQR sémantique efficace devrait pouvoir rechercher des réponses à partir des annotations sémantiques de corpus cibles et compléter les informations manquantes dans ces annotations en interrogeant les bases de connaissances disponibles. L'exploitation conjointe des bases de données et des corpus annotés est aussi employée par le système START, cependant, l'ordre y est inversé car les réponses sont recherchées d'abord dans une base de données (donc sans justifications) puis dans les documents textuels.

Sur un deuxième plan, une des observations principales qui peut être faite sur les systèmes existants est que les processus d'analyse de questions se fixent souvent un cadre dans lequel une question ne peut avoir qu'un seul focus, un seul type de réponse attendu et une seule relation principale, ce qui ne permet pas de saisir complètement les questions posées par les utilisateurs.

L'approche de questions-réponses que nous proposons à travers cette thèse prend en compte ces aspects au moment de la formalisation/interprétation des questions de l'utilisateur en construisant une ou plusieurs requêtes structurées pour une question posée en langage naturel.

Conclusion du chapitre

Dans ce chapitre nous avons présenté un état de l'art sur les systèmes de questions-réponses en domaine ouvert permettant de rechercher des réponses à partir (i) d'une collection de documents ou du web et (ii) à partir d'une base de données relationnelle ou d'une base de connaissances. Nous avons ensuite présenté un aperçu des spécificités du domaine médical puis des travaux proposant des taxinomies de questions médicales que nous avons exploité pour fixer les entités médicales et les relations sémantiques les plus fréquentes et les plus importantes. Nous avons enfin présenté des travaux portant sur la réalisation de systèmes de questions-réponses sémantiques dédiés à ce domaine de spécialité. Dans ce qui suit, nous proposons une approche « sémantique » pour répondre automatiquement aux questions médicales en transformant les questions médicales en des requêtes structurées qui peuvent aussi bien interroger les annotations d'une « collection de documents » que les bases de connaissances potentiellement disponibles. Le chapitre suivant est dédié à la présentation de notre méthode pour l'analyse des questions médicales et la construction des requêtes structurées correpondantes.

Choix de représentation et analyse de question

6

Sommaire

Un SQR nécessite deux entrées : le corpus utilisé pour extraire les réponses pertinentes et la question elle-même. Chacune de ces deux entrées doit être analysée correctement pour pouvoir trouver le meilleur appariement question-réponse. Dans une perspective de réutilisation, il est aussi important de représenter à la fois les questions et les réponses candidates avec une représentation formelle homogène pouvant être traitée par les systèmes d'information.

La section 6.1 de ce chapitre sera dédiée à la discussion des représentations formelles choisies dans la littérature et à notre choix de représentation. Nous consacrons la section 6.4 à notre approche d'analyse de question, une tâche primordiale qui est étudiée et évaluée séparément. Nous étudions la transformation de questions posées en langage naturel en requêtes basées sur un langage formel. Cette étude pose trois points clés : (i) Quelles sont les caractéristiques d'une question médicale, (ii) Quelles sont les méthodes les mieux adaptées pour l'extraction des informations utiles et (iii) Comment transformer les informations extraites en une représentation formelle. Nous présentons une approche « sémantique » comportant la reconnaissance des entités médicales, l'extraction de

relations sémantiques et la transformation automatique de la question en requête(s) SPARQL.

Les travaux présentés dans ce chapitre s'appuient sur les articles (Ben Abacha et Zweigenbaum 2012b;a).

6.1 Choix de représentation

Dans cette section nous commençons par comparer trois types de représentation/interrogation des informations extraites et des questions dans le cadre des systèmes de questions-réponses :

1. les langages propriétaires ou adhoc
2. la représentation relationnelle et l'interrogation avec SQL
3. la représentation « sémantique » avec les langages du web sémantique

Nous introduisons ensuite les langages du web sémantique que nous avons choisis pour notre approche en donnant les définitions préliminaires qui seront exploitées dans la présentation du système de questions-réponses réalisé, que nous avons appelé *MEANS*.

6.1.1 Formalisation de la question

Plusieurs systèmes de questions-réponses en domaine ouvert ou en domaine médical ont opté pour un format de représentation adhoc des informations extraites du corpus et de la question une fois analysée. C'est le cas, par exemple, du format PICO qui a été utilisé par (Sackett et al. 2000) pour représenter les informations extraites des corpus/questions dans le domaine médical. Bien que ce format ait été à l'origine conçu pour aider à formuler des recherches bibliographiques pour des humains, il a été exploité par plusieurs approches pour représenter des informations médicales extraites automatiquement.

La représentation avec un langage propriétéiaire ou adhoc limite la couverture du domaine d'une part et l'interopérabilité des systèmes de questions-réponses d'autre part. Par exemple, il ne sera pas possible pour un système basé sur le format PICO d'exploiter directement les résultats d'extraction d'information exprimés dans d'autres langages, ce qui implique que pour exploiter de nouveaux corpus ou de nouvelles méthodes d'extraction il sera nécessaire soit de modifier les outils d'extraction, soit de construire des adaptateurs spécifiques qui permettront de faire le lien entre les différents langages.

Sachant que les connaissances dans le domaine médical évoluent très rapidement, il sera de moins en moins concevable pour un système de questions-réponses d'adopter une telle stratégie car elle ne permet pas, à l'échelle mondiale, de construire, d'extraire ou de modifier les connaissances de manière cumulative.

D'autres systèmes de questions-réponses représentent les informations extraites par des tables relationnelles et transforment les questions des utilisateurs en requêtes SQL (Popescu et al. 2003). L'avantage de cette approche est qu'elle permet d'interroger à la fois les informations extraites à partir d'un corpus de documents mais aussi les informations préexistant dans les bases de données relationnelles métier utilisées par une institution ou une entreprise donnée. Cependant, pour ce faire, le même schéma de base de données doit être utilisé. Ceci est le principal inconvénient de la représentation relationnelle car l'intégration de données provenant de sources différentes requiert le développement d'adaptateurs adhoc pour intégrer les différents schémas de données.

Dans la dernière décennie, plusieurs méta-langages tels que RDF(S) [1] et OWL [2] ont été normalisés par le W3C afin de formaliser la représentation du sens sur le web. Ces langages fournissent un niveau élevé d'expressivité et sont de plus en plus utilisés dans des applications sémantiques et soutenus par des systèmes de stockage efficaces ainsi que des API (e.g. Sesame [3], Jena [4]) qui facilitent la lecture et l'interrogation des données. Mis à part l'avantage d'accessibilité, rendue uniforme grâce au langage d'interrogation SPARQL [5], les schémas de données dans le cadre du web sémantique sont des ontologies qui sont plus facilement partageables que les schémas de bases de données relationnelles. En effet, l'intégration de données dans le cadre du Web sémantique peut être automatisée grâce aux axiomes par défaut OWL qui ont standardisé la façon avec laquelle les alignements entre différentes ontologies sont exprimés et interprétés.

Notre choix s'est porté sur les langages du web sémantique OWL/RDF(S)/SPARQL car notre vision est que la conception de systèmes de questions-réponses efficaces dans le futur va requérir la disponibilité de bases de connaissances universelles accessibles qui peuvent être interrogées et mises à jour avec une interopérabilité aussi bien technique que sémantique.

6.2 Langages du web sémantique

Les langages du web sémantiques forment une pile avec des éléments ayant différents rôle et/ou différents degré d'expressivité (cf. figure 6.1). Dans le cadre de notre approche, nous utilisons le langage RDF(S) pour définir l'ontologie de domaine de référence et le langage SPARQL qui permet d'interroger des données au format RDF. OWL est aussi un langage du web sémantique qui permet de représenter des ontologies. Il est basé sur RDF(S) et permet de définir des classes et des relations avec des primitives plus évoluées (e.g. intersection, restriction) mais augmente le temps de calcul (raisonnement) nécessaire pour réaliser les inférences conséquentes lors de la manipulation des données.

Le langage RDF [6] représente les données comme un ensemble de triplets (sujet, prédicat, objet). Le sujet et le prédicat d'un triplet sont des ressources RDF (rdfs:Resource) définies par des IRI (Internationalized Resource Identifiers) [7]. L'objet d'un triplet peut aussi bien être une ressource RDF qu'un littéral. Une ressource RDF conforme à une ontologie est soit une propriété (rdf:Property), une classe (rdfs:Class) ou une instance de classe i vérifiant (i rdf:type C) avec C la classe dont i est l'instance. Les prédicats sont toujours des propriétés RDF. Un ensemble de données RDF forme un graphe dirigé dont les arcs sont les prédicats et les nœuds sont des ressources RDF ou des littéraux.

Le langage SPARQL (SPARQL Protocol and RDF Query Language)

1. http://www.w3.org/TR/rdf-syntax/
2. http://www.w3.org/TR/owl-ref/
3. http://www.openrdf.org
4. http://jena.sourceforge.net/
5. http://www.w3.org/TR/rdf-sparql-query/
6. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
7. http://tools.ietf.org/html/rfc3987

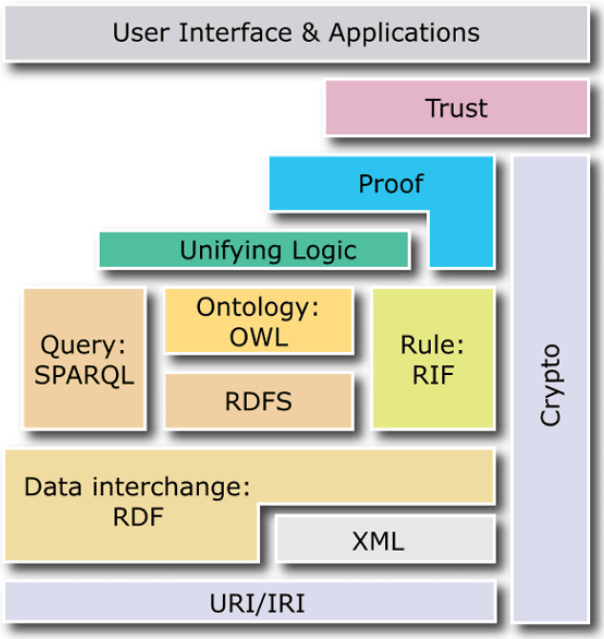


FIGURE 6.1 – *Pile de langages du web sémantique (www.w3.org, 2007)*

permet d'interroger des graphes RDF. Il permet d'exprimer des requêtes recherchant des patrons de graphes RDF requis ou optionnels ainsi que leur conjonction ou leur disjonction. SPARQL inclut plusieurs fonctionnalités ou filtres pour tester les valeurs de littéraux ou d'IRI et permet de spécifier le graphe RDF source à interroger. Les résultats des requêtes SPARQL peuvent aussi bien être des ensembles de réponses (une réponse étant un ensemble d'appariements entre variables et valeurs) que des graphes RDF.

6.3 Questions médicales : caractéristiques et types

Répondre automatiquement à des questions médicales requiert une analyse de question différente du traitement des questions en domaine ouvert à cause des spécificités de ce domaine de spécialité. Dans cette section, nous étudions les caractéristiques des questions médicales, que nous essayons ensuite de classer.

6.3.1 Caractéristiques des questions médicales

Plusieurs caractéristiques des questions médicales ont été plus ou moins mises en évidence et/ou utilisées dans les différents travaux menés sur l'analyse de questions. Certaines de ces caractéristiques sont communes avec les questions en domaine ouvert, d'autres sont plus spécifiques. D'après notre étude sur des questions médicales réelles (e.g. questions cliniques posées par des médecins[8], questions posées par des pa-

8. http ://clinques.nlm.nih.gov/

tients [9]), nous présentons ici quelques caractéristiques des questions médicales.

1. Le type de question. Différencier entre les questions Y/N (ou booléennes) et les questions WH qui peuvent être des questions de définition (e.g. What is Depression ?), des questions listes (e.g. What are the symptoms of blood cancer ?) ou des questions complexes (e.g. avec Why ou How) qui nécessitent des réponses détaillées et non pas uniquement des entités médicales.
2. Le type de la réponse attendue (TRA). Pour les questions WH, le type de la réponse attendue est introduit par le pronom ou l'adjectif interrogatif. Dans le domaine médical, ce type peut être un Traitement (e.g. What is the best treatment for Psoriasis ?), un Examen Médical (aussi appelé Test ; e.g. Colon Cancer : Which screening test should I have ?), etc.
3. Le Focus. Le focus de la question est l'entité médicale la plus proche de la réponse attendue. Par exemple, "pyogenic granuloma" est le focus de la question "What's the best treatment for pyogenic granuloma ?".
4. La relation principale. Pour les questions factuelles, nous définissons la relation principale comme étant celle reliant la réponse attendue et le focus. Dans les questions booléennes, cette relation correspond à la relation la plus importante (objet de la question) entre deux entités médicales (deux focus).
5. Les entités médicales. Reconnaître les entités médicales (e.g. headache) et leurs catégories (e.g. Medical Problem) est une étape importante qui permet de déterminer le focus et les autres entités médicales indispensables pour trouver la réponse exacte. La reconnaissance des entités médicales traite plusieurs problèmes comme la grande variation terminologique du domaine médical (un terme médical peut avoir plusieurs termes synonymes, des abréviations, etc.) et aussi l'évolution continue de cette terminologie (nouveaux termes médicaux, nouvelles maladies, etc.).
6. Les relations sémantiques. Elles apportent plus de précision sur la sémantique de la question (e.g entre une réponse attendue de type Traitement et un focus de type Problème (PB), cherche-t-on un traitement qui traite PB, le prévient, le cause ou le complique). Extraire les relations sémantiques de la question permet d'identifier la relation principale mais aussi les relations contextuelles (e.g. des relations autour du patient : son âge, son historique familial, etc.), point clé pour une analyse de question efficace.

Une observation clé dans notre travail est que la définition d'un seul focus ou d'un seul type de réponse attendue limite le traitement de certains types de questions et donc la couverture de l'analyse des questions.

9. www.askthedoctor.com/topics-a-z/topics-a-z-all.html ?view=category&id=8

6.3.2 Classification des questions médicales

Nous proposons dans le tableau 6.1 une classification des questions médicales en 6 catégories.

Des cas particuliers de ces catégories peuvent aussi être rencontrés. Par exemple, certaines questions factuelles recherchent une liste et non une réponse unique (e.g. « What are the symptoms of Alport's Syndrome ? », « What are the causes of hypomagnesemia ? », « What are the clinical features and prognosis of post-concussion syndrome ? »). Les questions des différentes catégories peuvent aussi être chaînées (e.g. « What is serum sickness and what are the causes ? », « What is cerebral palsy ? How do you diagnose and treat it ? What is the etiology ? What is the pathogenesis ? ») ou inclure une description des patients concernés (e.g. « 74-year-old female with memory loss and a negative workup. What is the dose of Aricept ? »).

Dans le cadre de notre approche nous nous intéressons aux questions booléennes et aux questions factuelles de type 1. Nous présentons dans la section suivante notre méthode pour l'analyse de ces questions, qui permet aussi le traitement de celles ayant plusieurs focus et/ou plusieurs types de réponses attendues. Cette approche se fonde sur l'extraction d'information à partir de la question, plus précisément la reconnaissance des entités médicales et l'identification des relations sémantiques. Précisons que notre approche de transformation des questions en requêtes SPARQL n'est pas dépendante des méthodes d'extraction d'information présentées dans la première partie de cette thèse et que d'autres outils peuvent être utilisés pour l'extraction d'information à partir de la question.

6.4 Analyse et transformation des questions médicales en requêtes SPARQL

La conception des systèmes de questions-réponses nécessite une analyse profonde des questions posées. Cette tâche primordiale requiert d'être étudiée et évaluée séparément.

En questions-réponses, l'analyse et la transformation de la question en une représentation structurée n'est pas une tâche triviale. Cette tâche a été mise en évidence, entre autres, dans le cadre des interfaces en langage naturel pour bases de données (NLIs ou Natural Language Interfaces to databases). Dans ce cadre, plusieurs travaux ont été menés pour la transformation des questions en requêtes SQL, SPARQL ou autres ((Popescu et al. 2003), (Lopez et Motta 2004), (Cimiano et al. 2008)). D'un autre côté, d'autres types de travaux spécifiques à la tâche d'analyse de questions existent. Certains traitent les questions ayant plusieurs types à la fois (e.g. (Fan et al. 2009)), d'autres se concentrent sur un type particulier de questions (e.g. les questions why (Verberne 2006)). Des travaux ont aussi été menés pour l'étude des questions en langage naturel en dehors du cadre des SQR. Par exemple, Duan et al. (2008) se sont intéressés à la détection de questions similaires en se basant sur l'objet et le focus de la question.

Nous proposons une approche d'analyse de question en langage naturel consistant à (i) extraire les informations les plus importantes à par-

Catégorie	Exemple	
Yes/No : questions booléennes	« Can Group B streptococcus cause urinary tract infections in adults ? »	
Explication/Raison : questions "why"	« Why do phenobarbital and Dilantin counteract each other ? »	
Condition/cas : la majorité des questions "when"	« When would you use gemfibrozil rather than an HMG (3-hydroxy-3-methylglutaryl) coenzyme A inhibitor ? »	
Manière : questions "how"	« How are homocysteine and folic acid related to hyperlipidemia ? / How can you do a screening motor exam of the hand ? »	
Définition	« What is seronegative spondyloarthropathy ? »	
Factuelle	Le type de réponse attendue est une entité médicale, une entité nommée ou plus généralement une information spécifique.	
	Type 1	Les réponses attendues correspondent à des entités médicales (correspond au cas le plus fréquent et au cas que nous traitons). Questions exprimées généralement avec les pronoms "what", "which" and "how" (e.g. « What is the complete workup for a meconium plug in a newborn ? », « How should you treat osteoporosis in a man, caused by chronic steroid use ? », « Which test (culdocentesis or pelvic ultrasound) would be best for diagnosis of ovarian cyst in this case ? », « Which medication is causing neutropenia in this baby ? »
	Type 2	Autres types de réponses attendues. Questions exprimées généralement avec "when" (recherche de temps dans ce cas), where, who, et quelques unes avec "how" (e.g. « When will inhaled insulin be available ? / Where do I go to work up a jaw mass ? / Where would a stroke be located that involved a right facial palsy and dysarthria ? », « How much Delsym cough syrup do I give ? », « How often should someone have tonometry as a screen for glaucoma ? », « How old should a patient be before I quit doing prostate specific antigens (PSA's) ? »

TABLE 6.1 – *Classification des question médicales en 6 catégories*

tir de la question (e.g. type de la question, type de la réponse attendue, entités médicales, relations sémantiques) et (ii) transformer les informations extraites en requête(s) SPARQL, le langage de requête standard pour les données RDF. RDF et SPARQL permettent une grande expressivité en représentant et en interrogeant les données comme des instances de concepts et de relations définies dans une ontologie de référence.

6.4.1 Description générale

Pour l'analyse des questions médicales, nous proposons une méthode en six étapes qui consiste à :

1. Identifier le type de la question (e.g. WH, Yes/No, Définition).
2. Déterminer le(s) type(s) de réponse attendue(s) pour les questions Wh.
3. Construire la forme affirmative et simplifiée de la question (nouvelle forme) afin d'éliminer les bruits potentiels dans les étapes suivantes.
4. Reconnaitre les entités médicales dans la nouvelle forme de la question.
5. Extraire les relations sémantiques à partir de la nouvelle forme de la question.
6. Construire la/les requête(s) SPARQL correspondante(s) (cf. Section 6.4.3).

Le tableau 6.2 présente la sortie de chaque étape sur deux exemples de questions.

Analyse de question (Q) -> Extraction d'information	**Exemples (Questions WH vs. Y/N)**	
	Question WH	*Question Y/N*
	What treatment works best for constipation in children ?	Does spinal manipulation relieve back pain ?
Identification du TRA	TRA = Treatment	——
Simplification et transformation de Q en forme Affirmative	new_Q = What treatment ———— ***ANSWER*** *works best for constipation in children.*	new_Q = ~~Does~~ *spinal manipulation relieve back pain.*
Reconnaissance des entités médicales (en utilisant new_Q)	ANSWER works best for <PB> constipation </PB> in <PA> children </PA>.	<TX> spinal manipulation </TX> relieve <PB> back pain </PB>.
Extraction des relations sémantiques (en utilisant new_Q)	treats(ANSWER,PB), with TRA = Treatment	treats(TX,PB)
Construction des requêtes SPARQL	Requête (R1)	Requête (R2)

TABLE 6.2 – *Analyse de questions médicales - Exemples (TRA : Type de réponse attendue, PB : Problème, PA : Patient, TX : Traitement)*

Ces étapes aboutissent à la construction des requêtes SPARQL suivantes :

Requête (R1) :

```
SELECT ?value3 ?umlsConcept3 ?file ?line WHERE {
?concept1 mesa:file ?file .
?concept1 mesa:line ?line .
?concept1 mesa:value “constipation” .
?concept1 mesa:umls_concept “Constipation” .
?concept1 mesa:umls_semanticType “Sign or Symptom” .
?concept1 mesa:category “sign_or_symptom” .
?concept2 mesa:file ?file .
?concept2 mesa:line ?line .
?concept2 mesa:value “children” .
?concept2 mesa:umls_concept “Child” .
?concept2 mesa:umls_semanticType “Age Group” .
?concept2 mesa:category “patient” .
?concept3 mesa:file ?file .
?concept3 mesa:line ?line .
?concept3 mesa:value ?value3 .
?concept3 mesa:umls_concept ?umlsConcept3 .
?concept3 mesa:category “treatment” .
?concept3 mesa:treats ?concept1 .
} GROUP BY ?umlsConcept3 ?value3
```

Requête (R2) :

```
ASK {
?concept1 mesa:value “spinal manipulation” .
?concept1 mesa:umls_concept “Manipulation of spine” .
?concept1 mesa:umls_semanticType “Therapeutic or Preventive Procedure” .
?concept1 mesa:category “treatment” .
?concept2 mesa:value “back pain” .
?concept2 mesa:umls_concept “Back Pain” .
?concept2 mesa:umls_semanticType “Sign or Symptom” .
?concept2 mesa:category “sign_or_symptom” .
?concept1 mesa:treats ?concept2 .
}
```

6.4.2 Identification des caractéristiques de la question

Identification du type de la question

Nous identifions le type de la question (e.g. WH, Yes/No, Définition) en appliquant un ensemble de règles simples sur les questions des utilisateurs (e.g. premierMotDe(Q) = (How | What | Which | When) indique que la question Q est de type Wh).

Identification du type de la réponse attendue

Pour les questions WH, nous déterminons le type de la réponse attendue en projetant des patrons lexicaux construits manuellement sur le texte original de la question. Un ensemble de patrons est construit pour chaque

type de question. Ces patrons utilisent des pronoms interrogatifs, des marqueurs syntaxiques et des mots génériques et identifient un ensemble de questions compatibles. Cependant, il arrive qu'une question ait plus d'un type de réponse attendue. Dans ce cas, nous gardons tous les patrons qui ont pu être projetés sur la question, même s'ils sont associés à des types de réponse différents (e.g. Traitement, Médicament, Test médical).

Construction de la forme affirmative de la question

Dans une deuxième étape, nous construisons la forme affirmative de la question en remplaçant les pronoms interrogatifs par le mot clé "ANSWER". Cette forme sera utilisée à l'étape d'extraction de relations. Nous construisons aussi une forme simplifiée de la question où la séquence de mots indiquant le type de réponse attendue est remplacée par le mot clé "ANSWER". Cela permet d'éviter du bruit au moment de l'extraction des entités médicales.

Par exemple, dans la question : « What is the best treatment for headache ? » le système de reconnaissance des entités médicales retourne "treatment" et "headache" comme entités, ce qui ne constitue pas une entrée précise pour la phase d'extraction de relation, car "treatment" est la catégorie de la réponse attendue et non une entité. La forme affirmative de la question « `ANSWER` is the best treatment for headache ? » permet d'abord d'expliciter les deux entités médicales de la question. Cependant l'erreur consistant à extraire "treatment" comme entité médicale est toujours susceptible de se produire. Simplifier la question à « `ANSWER` for headache » permet d'expliciter l'entité médicale recherchée et d'obtenir une extraction de relations plus efficace en identifiant les relations entre seulement l'entité `ANSWER` (qui est étiquetée comme étant de type "treatment") et l'entité « headache ». L'intuition derrière cette approche est que les indicateurs du type de la réponse attendue peuvent constituer un bruit pour l'extraction de relation qui peut être évité en simplifiant la question.

L'extraction de relations sémantiques est la dernière étape avant la construction de la requête SPARQL. Pour les questions de type Yes/No, le processus d'extraction de relation a des entrées explicites étant donné que toutes les entités médicales y sont complètement identifiées. Par contre, pour les questions de type Wh nous pouvons avoir plusieurs types de réponse attendues. Dans un tel cas, une méthode plus générique est nécessaire.

Reconnaissance des entités médicales

Les résultats de la reconnaissance des entités médicales sont représentés en logique du premier ordre avec les prédicats `category`, `value` et `position`. Par exemple, la formule E1 (tableau 6.3) indique que le troisième token de la question (« aspirin ») est une entité médicale et que sa catégorie est TREATMENT.

Extraction de relations sémantiques

Les résultats de l'extraction de relations sont représentés en logique du premier ordre avec plusieurs prédicats indiquant le nom de la relation

```
E1 :
{category(#ME1,TREATMENT)
∧ value(#ME1,aspirin)
∧ position(#ME1,3)}
```

TABLE 6.3

(e.g. treats, causes). Par exemple, la formule E2 (tableau 6.4) indique que trois entités médicales sont reliées par deux relations sémantiques *treats* et *patientHasProblem*.

```
E2 :
{ treats(#ME1,#ME2)
∧ patientHasProblem(#ME3,#ME2) }
```

TABLE 6.4

Cas où l'on a plusieurs types de réponses attendues (TRA)

Pour prendre en compte le cas de multiples TRA, nous construisons autant de questions que de réponses attendues. Ce processus complet est décrit dans le tableau 6.5.

1. Identifier le type de la question
2. Identifier les types de réponses attendues (TRA) pour les questions WH (m TRA)
3. Construire le forme simplifiée et affirmative des questions (nouvelle forme)
4. Identifier les entités médicales dans la nouvelle forme de la question - *Pour* ($x = 1$, x++, $x \leq m$)
5:x. Extraire les relations sémantiques *[Entrées : (i) le* TRA_x*, (ii) les entités médicales, et (iii) la nouvelle forme de la question)]*
6:x. Construire la requête SPARQL x

TABLE 6.5 – *Processus complet en présence de TRA multiples*

Pourquoi construire plusieurs requêtes ? Notons d'abord qu'il est possible d'utiliser le langage SPARQL pour représenter la sémantique de la question et retrouver toutes les réponses attendues en construisant un seul patron de graphe RDF et une requête unique pour une question donnée. Cependant, cette méthode contraindrait le système à trouver une justification commune dans le corpus textuel pour les différents TRAs. Par exemple, si la question est « How to diagnose and treat pressure ulcers ? », construire un seul patron de graphe RDF mènerait à rechercher un examen et un traitement qui permettent de diagnostiquer et traiter la *même* entité médicale E. Alors que construire deux requêtes, pour cet exemple, permet de rechercher un traitement pour une entité médicale $E1$ de type *pressure ulcers* et un examen pour une entité médicale $E2$ qui est aussi de type *pressure ulcers*. Dans ce cas $E1$ peut aussi bien être égale à $E2$ (même entité dans le corpus) que non, ce qui évite la restriction posée

par l'utilisation d'un seul patron de graphe. Une autre solution potentielle à ce point est d'utiliser la clause *UNION* pour rechercher deux patrons de graphes RDF disjoints dans une même requête, nous avons cependant choisi d'écrire plusieurs requêtes différentes dans un souci de clarté.

6.4.3 Construction de requête(s) SPARQL

Représentation des requêtes et des documents

Nous représentons les entités médicales détectées dans les phrases du corpus, les relations découvertes entre ces entités médicales, les concepts UMLS associés aux entités, les termes UMLS associés et la classe de l'ontologie associée (type de l'entité) avec des annotations RDF . Dans le tableau 6.6 nous résumons l'ensemble des données associées aux entités médicales, leur format de représentation précis et une justification pour les choix de format.

Donnée associée à l'entité	Format de représentation RDF	Explication
Classe ontologique	IRI	les classes RDF sont obligatoirement identifiées par des IRI
Concept UMLS	littéral	le choix d'une IRI aurait été possible techniquement mais dans la perspective de l'intégration de bases de connaissances externes référant à l'UMLS, ces IRIs n'auraient pas permis de réconcilier les données car elles sont écrites d'une manière adhoc suivant le choix de l'annotateur, par contre utiliser un litéral permet de réconcilier plus facilement nos données avec d'autres données utilisant l'UMLS grâce à des similarités lexicales.
Terme UMLS	littéral	même explication que pour les concepts UMLS

TABLE 6.6 – *Format de représentation des données associées aux entités médicales*

Usage de SPARQL

Une fois les entités médicales et les relations sémantiques extraites de la question en langage naturel, la dernière étape consiste à construire une requête SPARQL équivalente. SPARQL définit 4 types différents de requêtes : CONSTRUCT, DESCRIBE, ASK et SELECT [10]. La forme CONSTRUCT vise à générer de nouveaux graphes RDF à partir des données RDF disponibles. La forme DESCRIBE est juste informative (i.e. elle retourne une sélection aléatoire des réponses). La forme ASK vérifie si un patron de graphe RDF a des correspondances dans les graphes RDF interrogés. La forme SELECT permet de faire correspondre un patron de

10. http://www.w3.org/TR/rdf-sparql-query/#initDefinitions

graphe RDF et de retourner des valeurs de variables correspondant à des nœuds de ce graphe.

Dans notre approche de construction de requêtes SPARQL, nous utilisons les formes ASK et SELECT afin de représenter (respectivement) les questions en langage naturel de type Yes/No et de type Wh. Les questions de type définition seront considérées comme des questions Wh étant donné qu'elles recherchent des morceaux de texte contenant la définition, sauf que les variables à retourner ne sont pas les mêmes dans ce cas. Les conditions d'égalité sur les littéraux des requêtes peuvent être exprimées par l'insertion de la valeur dans un patron de triplet ou en récupérant la valeur du littéral dans une variable afin de la tester avec les clauses FILTER et les fonctions de comparaison de chaînes de caractères.

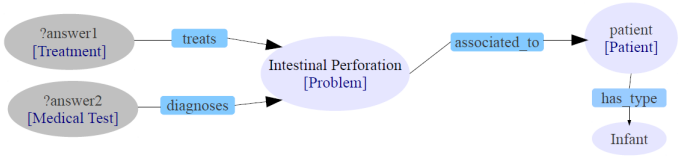
Une requête SPARQL a deux composants principaux : un en-tête et un corps. L'en-tête indique la forme de la requête et d'autres déclarations (e.g. préfixes, noms des graphes RDF à interroger, variables à retourner pour la forme SELECT). La construction du corps et de l'en-tête de la requête requièrent des processus différents.

Transformation des questions WH

La forme SELECT des requêtes SPARQL est la plus appropriée pour la représentation des questions WH. L'en-tête de ces requêtes contient principalement le mot clé SELECT et les noms de variables à retourner. Le corps des requêtes SELECT contient le patron de graphe RDF qui a été construit en utilisant les entités médicales et les relations sémantiques extraits de la question originale. Nous formalisons le processus de construction du corps de la requête comme une traduction d'une formule en logique du premier ordre (représentant la sortie des processus d'extraction d'information) en un patron de graphe RDF basique. Les relations sémantiques extraites peuvent être définies entre deux entités médicales ou entre la réponse attendue et une entité médicale (e.g. treats(ANSWER,flu), patientAgeGroup(patient,infant)). Chaque prédicat binaire est transformé en un triplet RDF <s,p,o> où s est le sujet, p est la propriété RDF et o est l'objet. Le sujet et l'objet sont définis comme des variables représentant les arguments du prédicat en logique du premier ordre. Des triplets additionnels sont aussi générés pour indiquer la catégorie et/ou le nom précis des entités médicales entrant en jeu, et cela avec l'utilisation des propriétés RDF *mesa:value* et *mesa:category* définies dans l'ontologie MESA. L'exemple suivant représente l'équivalent en requête SPARQL du prédicat *causes(ANSWER, Flu)* :

```
SELECT ?answer WHERE {
?answer mesa:causes ?arg
?arg mesa:value 'Flu'
```

Dans le cas où nous avons le TRA (e.g. category(ANSWER, TREATMENT)) ou si l'étape d'extraction des entités médicales nous fournit les catégories des entités médicales (e.g. category(Flu,PROBLEM)), un triplet équivalent est construit en récupérant les IRI correspondant aux concepts représentant ces catégories. L'exemple suivant représente

<table>
<tr><th colspan="2">Question médicale</th></tr>
<tr><td colspan="2">What are the current treatment and monitoring recommendations
for intestinal perforation in infants ?</td></tr>
<tr><th colspan="2">Graphe sémantique simplifié</th></tr>
<tr><td colspan="2"> 
</td></tr>
<tr><th>Requête SPARQL 1</th><th>Requête SPARQL 2</th></tr>
<tr><td>Select ?answer1 ?text1 where {
?answer1 mesa:category <Treatment>
?answer1 mesa:treats ?focus

?focus mesa:value 'intestinal perforation'

?focus mesa:category <Problem>

?patient mesa:hasProblem ?focus

OPTIONAL{
?text1 mesa:contains ?answer1

?text1 mesa:contains ?patient

?text1 mesa:contains ?focus } }</td><td>Select ?answer1 ?text1 where {
?answer1 mesa:category <Test>

?answer1 mesa:diagnoses ?focus

?focus mesa:value 'intestinal perforation'

?focus mesa:category <Problem>

?patient mesa:hasProblem ?focus
OPTIONAL{

?text1 mesa:contains ?answer1

?text1 mesa:contains ?patient

?text1 mesa:contains ?focus } }</td></tr>
</table>

TABLE 6.7 – *Transformation d'une question médicale : requêtes SPARQL associées*

la transformation de la catégorisation de la réponse attendue category(ANSWER,TEST) :

SELECT ?answer WHERE {?answer mesa:category mesa:TEST}.

La transformation finale de la question en langage naturel consiste en une ou (dans le cas de multiples TRA, cf. section 6.4.2) plusieurs requêtes SPARQL. Ces requêtes sont construites en assemblant les transformations unitaires des prédicats obtenus en sortie des étapes d'extraction d'entités médicales et de relations sémantiques. Le tableau 6.7 montre un exemple entier de transformation d'une question WH.

Transformation des questions Yes/No

La forme de requête SPARQL ASK est la plus appropriée pour les questions de type Yes/No. L'en-tête de telles requêtes contient principalement le mot clé ASK. Le corps des requêtes ASK construites contient le patron de graphe RDF représentant la sémantique de la question. Le processus de transformation des questions Yes/No est similaire à la transformation des questions WH sauf que nous n'avons pas de réponses ou de TRA à prendre en compte. En conséquence, la construction du patron de graphe RDF final consiste uniquement en la conversion des résultats de l'extraction des entités médicales et des relations sémantiques en un format de patron de triplet RDF comme décrit dans la section précédente. L'exemple suivant montre le résultat de la transformation d'une question Yes/No en un patron de graphe RDF.

```
 Can being on prednisone cause a high serum
iron?
ASK{
 ?e1 mesa:causes ?e2
?e1 mesa:value ?val1
?e2 mesa:value ?val2
FILTER(?val1='prednisone')
FILTER(?val2='serum iron')}
```

6.4.4 Evaluation de l'analyse de question

Protocole expérimental

Pour évaluer notre approche, nous utilisons 100 questions extraites à partir du Journal of Family Practice (JFP [11]). Nous avons choisi les dernières questions du 1/11/2008 au 1/4/2011. Cet ensemble contient 64 questions WH et 36 questions Yes/No. Le tableau 6.8 présente quelques exemples de questions de ce corpus d'évaluation.

Question	Exemple
Yes/No	Should patients with acute DVT limit activity ?
WH	What is the best approach to benign paroxysmal positional vertigo in the elderly ?
complexe	When should you consider implanted nerve stimulators for lower back pain ?
TRA:Traitement	Childhood alopecia areata : What treatment works best ?
TRA:Médicament	Which drugs are best when aggressive Alzheimer's patients need medication ?
TRA:Test	What is the best noninvasive diagnostic test for women with suspected CAD ?
2 TRA ou plus	When should you suspect community-acquired MRSA ? How should you treat it ?

TABLE 6.8 – *Quelques exemples de questions de notre corpus d'évaluation (TRA = Type de la réponse attendue)*

L'expérimentation consiste à appliquer notre approche pour l'analyse de question à cette collection de questions en langage naturel et à récupérer les requêtes SPARQL retournées. L'évaluation consiste ensuite à vérifier manuellement pour chaque question si la requête SPARQL est une représentation correcte de la question en langage naturel. Une représentation correcte étant une requête SPARQL qui regroupe toutes les entités médicales et les relations de la question et qui recherche/retourne la réponse attendue. Dans le cas de questions avec plusieurs types de réponses attendues, une représentation correcte est un ensemble de requêtes SPARQL reflétant chacune correctement la sémantique de la question en langage naturel pour le type de réponse attendue qu'elle représente.

Résultats

Nous avons testé nos deux méthodes de reconnaissance d'entités médicales sur notre corpus d'évaluation et aussi sur le corpus i2b2 de textes

11. http://www.jfponline.com

Méthode	**corpus i2b2**			**corpus JFP**		
	P	R	F	P	R	F
MM+	56.5	48.7	52.3	66.66	**84.55**	**74.54**
BIO-CRF-H	84.0	72.3	**77.7**	**77.03**	46.34	57.87

TABLE 6.9 – *Reconnaissance des entités médicales (catégories : Traitement, Problème et Test) : résultats sans la simplification des questions*

	F-mesure
Extraction de relations	71%
Construction de requêtes	62%
Construction de requêtes (entités et relations valides)	98%

TABLE 6.10 – *Résultats de l'extraction de relations et la construction de requêtes à partir des questions médicales*

cliniques [12] construit dans le cadre du challenge i2b2/VA 2010 (pour comparer les résultats). Nous avons utilisé le corpus d'entraînement de i2b2 pour entraîner la méthode BIO-CRF-H (le corpus d'entraînement de i2b2 contient 31238 phrases et le corpus de test contient 44927). Le tableau 6.9 présente les résultats sans simplification de la question pour les trois catégories : Traitement, Problème et Test. C'est important de noter que les résultats de la méthode BIO-CRF-H sur le corpus JFP ne sont pas du même niveau que ceux obtenus sur le corpus i2B2. Ceci est dÃ» principalement au fait que cette méthode a été entraînée sur un type particulier de corpus (i2b2) et que les deux corpus i2b2 et JFP ont des types différents (un inconvénient classique des méthodes statistiques).

Nous avons évalué aussi l'influence de la simplification de questions. Cette dernière améliore les résultats de la reconnaissance des entités médicales et spécialement les résultats de la méthode MetaMapPlus, avec une précision de 75.91%, un rappel de 84.55% et donc une F-mesure de 79.99% pour la reconnaissance des trois catégories Traitement, Problème et Test sur le corpus JFP.

Nous avons évalué aussi l'extraction de relations et la construction des requêtes SPARQL (cf. tableau 6.10). Sur 100 questions, l'extraction de relations a échoué pour 29 questions. Le résultat final montre que 62 transformations sont correctes et 38 incorrectes ou incomplètes. Si nous étudions le processus de transformation des questions sur le sous-ensemble de questions pour lesquelles les entités médicales et les relations sémantiques ont été extraites correctement, nous observons que 98% des requêtes SPARQL sont correctes ce qui conforte notre hypothèse sur la robustesse de ce langage structuré pour la formalisation de questions en langage naturel. Cependant, dans les applications réelles, la performance de l'analyse de question et du SQR en général dépendra fortement des performances des systèmes d'extraction d'information utilisés.

Analyse des erreurs. Nous avons obtenu 62 requêtes correctes parmi les 100 questions d'évaluation. Parmi les 38 requêtes incorrectes, 29 sont principalement dues à des erreurs au niveau de l'extraction de relations

12. http://www.i2b2.org/NLP/Relations/

et 8 sont dues à des erreurs au niveau de l'identification du type de la réponse attendue. Plus précisément, les principales causes d'erreurs sont :

1. les relations non encore définies dans notre ontologie ou non encore traitées par notre système d'extraction de relations (e.g. « How does pentoxifylline affect survival of patients with alcoholic hepatitis ? », relation : "affects").
2. les questions complexes et leurs types de réponses attendues non encore traités par notre système. Par exemple, pour la question « How accurate is an MRI at diagnosing injured knee ligaments ? » même si on détermine correctement les entités médicales, et la relation sémantique (ici, *diagnoses*), la requête n'est pas correcte car le type de la réponse attendue est non encore traité par notre système. D'autres types de réponses attendues ne sont pas encore pris en compte non plus (e.g.« Which women should we screen for gestational diabetes mellitus ? », TRA : Patient).

Les résultats obtenus peuvent être largement améliorés en augmentant les performances des systèmes d'extraction d'information avec plus de types de relations sémantiques et d'entités médicales et en traitant les questions complexes.

Conclusion du chapitre

Dans ce chapitre, nous nous sommes attaquée à l'analyse automatique de questions médicales et nous avons présenté une approche originale pour la transformation de questions médicales en requêtes SPARQL. Ce travail a trois principales caractéristiques :

- L'approche proposée permet de traiter différents types de questions, parmi lesquels les questions avec deux types de réponses attendues ou plus et/ou deux focus ou plus.
- Elle permet une analyse profonde des questions, en utilisant plusieurs méthodes d'extraction d'information fondées sur (i) les connaissances du domaine (e.g. UMLS) et (ii) des techniques de TAL (e.g. utilisation de patrons ou règles, apprentissage automatique). Ces méthodes visent l'extraction des entités médicales, des relations sémantiques et aussi des informations supplémentaires ou contextuelles (e.g. informations sur les patients : âge, sexe, etc.).
- Notre approche est basée sur les technologies du web sémantique qui offrent plus d'expressivité par rapport à d'autres langages de formalisation des connaissances (e.g. logique du premier ordre, tables relationnelles/SQL) et fournissent des langages formels standards afin d'augmenter la portabilité des annotations relatives aux questions et aux corpus utilisés pour l'extraction des réponses.

Afin d'atteindre complètement notre objectif de mise en place d'une méthode générique pour l'analyse de questions médicales posées en langage naturel davantage de traitements sont encore requis pour : (i) les questions complexes (e.g. why, when) et (ii) les types de relations sémantiques non encore définis dans notre ontologie et non encore traités par notre système d'extraction de relations. Pour ce dernier point et pour améliorer la couverture ou la portabilité de notre système, nous envisageons

dans le futur de tester la contribution de l'analyse syntaxique à l'analyse de question pour deux tâches : (i) confirmer les relations sémantiques déjà extraites et (ii) détecter les relations sémantiques inconnues (ou non encore traitées par notre système) en se basant sur les dépendances syntaxiques (Sujet-Verbe-Objet) qui peuvent remplacer les triplets (Entité1-Relation-Entité2).

7 Recherche de réponses et évaluation du système MEANS

Dans ce chapitre nous présentons notre approche pour répondre automatiquement à des questions médicales posées en langage naturel (cf. 7.1). Cette approche a été implémentée au sein du système de questions-réponses *MEANS* qui a été évalué sur des questions médicales réelles (cf. section 7.2).

7.1 Recherche de réponses aux questions médicales

Dans cette section nous présentons dans un premier temps la description générale de notre approche. Dans un second temps, nous présentons l'ontologie de référence qui regroupe les concepts et les relations utilisés pour formaliser les questions de l'utilisateur d'une part et les informations extraites des corpus médicaux d'autre part. Enfin, nous détaillons les étapes de recherche et d'extraction de réponses.

7.1.1 Approche générale

Nous avons proposé une approche sémantique pour répondre automatiquement aux questions médicales posées en langage naturel (cf. figure 7.1).

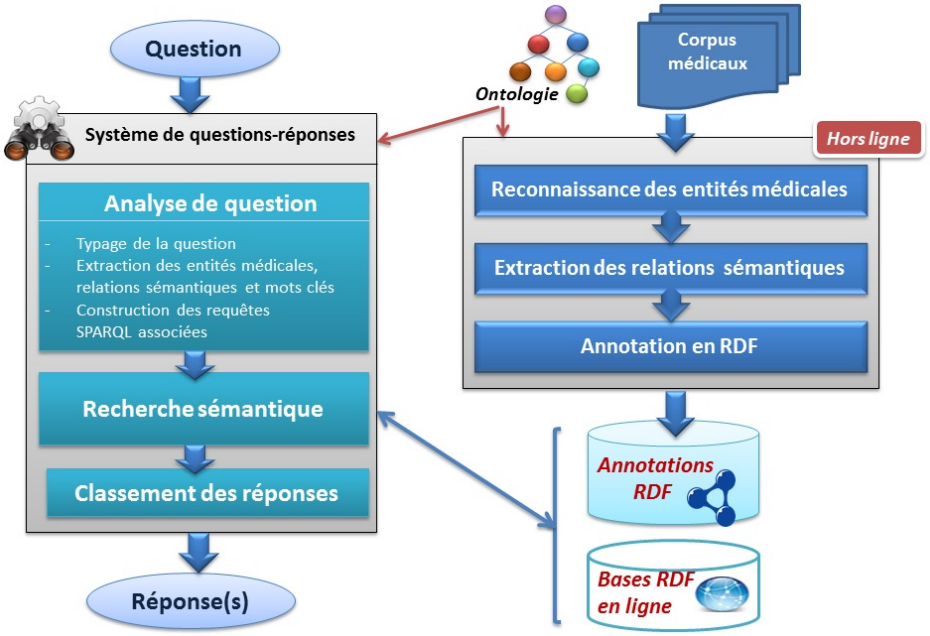


Figure 7.1 – *Architecture du SQR MEANS*

Cette approche comporte les quatre composantes suivantes :

1. **Analyse et annotation hors ligne des documents utilisés pour trouver les réponses.** Il s'agit d'annoter les corpus qui seront utilisés pour trouver les réponses. Cette annotation se fait hors ligne et utilise les méthodes d'extraction d'information présentées dans la première partie de ce manuscrit. Elle fournit en sortie des triplets RDF qui seront interrogés pour trouver les réponses (cf. section 7.1.3).
2. **Analyse des questions en anglais.** Cette étape consiste à analyser la question et à extraire ses caractéristiques (i.e. type de la question, type de la réponse attendue, entités médicales, etc.), puis à construire à partir de ces informations extraites une ou plusieurs requêtes SPARQL avec des degrés de précision différents. Cette tâche d'analyse des questions en anglais a été décrite dans le chapitre précédent (section 6.4, page 95).

3. **Recherche sémantique.** Cette étape se fonde sur un moteur de recherche RDF pour effectuer l'appariement requêtes-documents et trouver les phrases ou entités médicales pertinentes (cf. section 7.1.4).
4. **Extraction et classement des réponses.** Cette étape utilise les informations déduites de la question pour extraire la réponse. Le classement des réponses est effectué en fonction du degré de précision de la requête SPARQL utilisée (cf. section 7.1.4) et du nombre de justifications retrouvées pour la même réponse.

Cette approche a été implémentée au sein du système de questions-réponses *MEANS*.

7.1.2 Ontologie de référence

Pour formaliser la question de l'utilisateur et les informations extraites des corpus médicaux nous exploitons l'ontologie de référence MESA (cf. figure 7.2) que nous avons construite à cet effet. L'ontologie *MESA* définit des concepts et des relations décrivant les fragments de texte qui seront retournés comme réponses finales de notre système de questions-réponses aussi bien que des concepts et des relations du domaine médical.

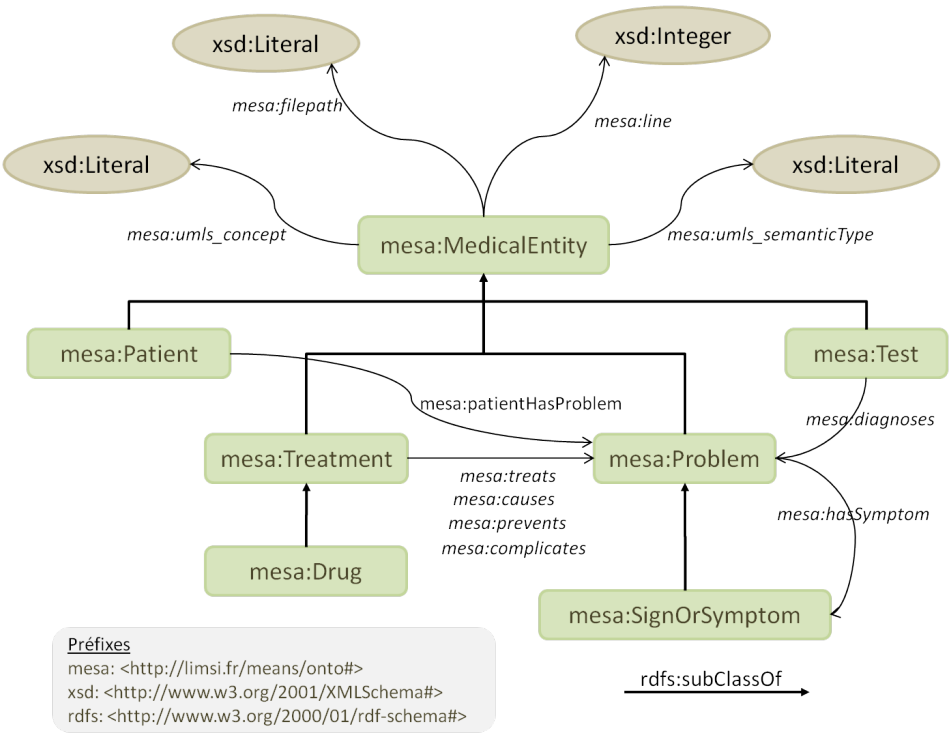


FIGURE 7.2 – *MESA : ontologie de référence pour le système MEANS*

L'ontologie *MESA* représente actuellement six catégories/classes d'entités médicales organisées hierarchiquement par la relation *rdfs:subClassOf* et sept relations de domaine pouvant être exprimées entre les entités médicales appartenant aux catégories représentées. À chaque entité médicale est associé son concpet UMLS (propriété *mesa:umls_concept*), son type sémantique UMLS (propriété *mesa:umls_semanticType*), le chemin de fichier qui contient l'entité médicale (propriété *mesa:filepath*) et le numéro de la

ligne contenant l'entité dans le fichier (propriété *mesa*:*line*). Les fichiers sources du corpus sont structurés de façon à contenir une phrase par ligne.

7.1.3 Annotation RDF hors ligne des corpus médicaux

Les performances du système de questions-réponses final dépendront fortement de la qualité de l'annotation des collections de textes utilisées pour trouver les réponses. Pour cette raison, nous avons travaillé sur la reconnaissance des entités médicales mais aussi sur l'extraction des relations sémantiques reliant les entités reconnues. Pour ces deux tâches nous avons testé et mis au point des méthodes à base de règles ou patrons et des méthodes statistiques et hybrides (cf. chapitres 2 et 3). L'analyse et annotation finale des corpus comporte trois étapes principales, que nous détaillons brièvement ci-dessous :

- La reconnaissance des entités médicales.
- L'identification des relations sémantiques reliant les entités médicales reconnues.
- L'annotation en RDF des entités et relations extraites ainsi que d'autres informations (e.g. les termes MeSH des articles scientifiques).

Reconnaissance des entités médicales

Cette étape, décrite dans le chapitre 2, consiste à identifier les termes médicaux et déterminer leurs catégories (e.g. Traitement, Médicament, Problème médical, Examen ou test médical, signe ou symptôme). Nous déterminons également les concepts de l'UMLS associés à ces termes médicaux. Utiliser les concepts UMLS permettra de retrouver des réponses au moment de l'interrogation dans les cas de synonymie. Par exemple, si l'utilsateur recherche les derniers traitements developpés pour "Type 2 Diabetes", utiliser les concepts UMLS permettra de retrouver des réponses exprimant la même maladie avec d'autres termes (e.g. "Diabetes mellitus type 2" ou "non-insulin-dependent diabetes mellitus").

Extraction de relations sémantiques

Après avoir identifié les termes médicaux, nous nous intéressons aux relations sémantiques qui les relient. Nous cherchons à reconnaître sept relations médicales (cf. chapitre 3).

Annotation en RDF des informations extraites

L'objectif de base de cette étape est d'écrire les annotations ou les informations extraites (i.e. entités médicales et relations sémantiques) sous forme de triplets RDF conformes à l'ontologie utilisée. La figure 7.3 présente un exemple d'annotation.

7.1.4 Recherche sémantique et classement des réponses

Les questions des utilisateurs, posées en langage naturel, sont analysées suivant le procédé indiqué dans le chapitre 6 afin de construire une ou

```
<rdf:Description rdf:about="sentence83">
        <rdf:value>The study results show that ferumoxytol is a potential
        new treatment for iron deficiency anemia.</rdf:value>
        <onto:contains rdf:resource="#entity20"/>
        <onto:contains rdf:resource="#entity21"/>
</rdf:Description>
<rdf:Description rdf:about="entity20">
        <rdf:value>ferumoxytol</rdf:value>
        <rdf:type rdf:resource="&onto;PharmacolgicSubstance"/>
</rdf:Description>
<rdf:Description rdf:about="entity21">
        <rdf:value>iron deficiency anemia</rdf:value>
        <rdf:type rdf:resource="&onto;DiseaseOrSyndrome"/>
</rdf:Description>
<rdf:Description rdf:about="statement7809">
        <rdf:subject rdf:resource="#entity20"/>
        <rdf:predicate rdf:resource="&onto;Treats"/>
        <rdf:object rdf:rsesource="#entity21"/>
        <onto:precision>1</onto:precision>
</rdf:Description>
```

FIGURE 7.3 – *Exemple d'annotation RDF*

plusieurs représentations SPARQL initiales de la question selon le nombre de types de réponses attendues. Ces requêtes SPARQL sont ensuite relaxées en éliminant au fur et à mesure certaines informations.

Par exemple, pour la question « What is the best treatment for oral thrush in healthy infants ? » [1], la requête SPARQL la plus précise est celle décrite dans la figure 7.4a. Ensuite d'autres requêtes sont générées automatiquement. La figure 7.4b présente la requête la moins précise.

Nous définissons trois niveaux de relaxation pour la requête initiale, chacun contenant une ou plusieurs requêtes. Les requêtes sont triées dans chaque niveau suivant leur précision :

1. **Niveau 1** : ce niveau regroupe d'abord la requête initiale elle même, puis une forme relaxée de cette requête obtenue en supprimant les valeurs des entités médicales (ici : ?value1 ; ?value2 et ?value3). Par exemple, en plus de chercher une phrase qui contient *healthy infants*, on cherchera une phrase qui contient une entité médicale ayant comme concept *Infant* sans restriction sur la valeur exacte. Cette relaxation peut diminuer légèrement la précision (le fait de garder le concept UMLS garantit de chercher la bonne entité médicale) mais elle augmente beaucoup le rappel.

1. Cette question fait partie des questions utilisées pour l'évaluation. Les requêtes SPARQL présentées font partie de celles générées automatiquement par le système MEANS au cours de l'évaluation.

```
SELECT ?concept1 ?file ?line WHERE {

?concept1 <http://limsi.fr/means#file> ?file .
?concept1 <http://limsi.fr/means#line> ?line .

?concept2 <http://limsi.fr/means#file> ?file .
?concept2 <http://limsi.fr/means#line> ?line .

?concept3 <http://limsi.fr/means#file> ?file .
?concept3 <http://limsi.fr/means#line> ?line .

?concept1 <http://limsi.fr/means#concept_category> "Treatment" .

?concept2 <http://limsi.fr/means#concept_category> "Problem" .
?concept2 <http://limsi.fr/means#umls_semanticType> "Disease or Syndrome" .
?concept2 <http://limsi.fr/means#umls_concept> "Oral candidiasis" .
?concept2 <http://limsi.fr/means#value> "oral thrush" .

?concept3 <http://limsi.fr/means#concept_category> "Patient" .
?concept3 <http://limsi.fr/means#umls_semanticType> "Age Group" .
?concept3 <http://limsi.fr/means#umls_concept> "Infant" .
?concept3 <http://limsi.fr/means#value> "healthy infants" .

?concept1 <http://limsi.fr/means#treats> ?concept2 .
?concept3 <http://limsi.fr/means#patient_has> ?concept2 .

}
```

(a) *La requête la plus précise*

```
SELECT ?concept1 ?file ?line WHERE {
?concept1 <http://limsi.fr/means#file> ?file .
?concept1 <http://limsi.fr/means#line> ?line .

?concept2 <http://limsi.fr/means#file> ?file .
?concept2 <http://limsi.fr/means#line> ?line .

?concept1 <http://limsi.fr/means#concept_category> "Treatment" .

?concept2 <http://limsi.fr/means#concept_category> "Problem" .
?concept2 <http://limsi.fr/means#umls_concept> "Oral candidiasis" .
}
```

(b) *La requête la moins précise*

FIGURE 7.4 – *Requêtes SPARQL construites automatiquement pour la question : « What is the best treatment for oral thrush in healthy infants ? »*

2. **Niveau 2** : Supprimer les entités médicales une à une (en gardant la réponse attendue et le focus [2])
3. **Niveau 3** : Supprimer la/les relation(s) principale [3].

Ces requêtes SPARQL sont ensuite exécutées dans l'ordre pour interroger les triplets RDF générés dans la phase d'annotation de corpus. Les réponses sont ainsi triées par le tri des requêtes. Un tri supplémentaire est effectué sur les réponses dans le cas des questions factuelles. Pour ces questions, une réponse à une requête est priviliégiée si elle a plus de justifications que les autres. Deux réponses sont jugées identiques si les entités retournées ont le même CUI [4] UMLS (e.g. the corticosteroid injection, corticosteroid injections : C2095490). Les justifications sont ensuite groupées par entité/CUI différent et comptées.

Le tableau 7.1 présente quelques exemples de modèles de questions et de leurs réponses.

Question	**Exemple**	**Forme simplifiée**	**Ex. de réponse**
Definition	What is X	X	X is ..
Yes/No	Can X REL Y	X REL Y	Yes/No
Factuelle ou liste : TRA et REL connus	How can you REL Y ou What is X REL Y	ANSWER ? REL Y (on supprime X)	X' ? REL Y avec X' is a X
Factuelle ou liste : TRA et REL inconnus	What is X REL Y	X REL Y || ANSWER ? REL Y	X' ? REL Y avec X' is a X

TABLE 7.1 – *Questions vs. Réponses*

La nature des réponses retournées par le système *MEANS* varie suivant le type de la question :

- Question Définition : Réponse = une phrase
- Question booléenne : Réponse = Oui / non
- Question Factuelle : Réponse = Entité médicale (e.g. un traitement) ou une information précise (e.g. la dose d'un médicament)
- Question liste : Réponse = une liste d'entités médicales (e.g. liste de symptômes)

En plus de cette réponse, un extrait de quatre phrases entourant la réponse est précisé. Nous utilisons le moteur de recherche sémantique Jena [5] pour la lecture et l'interrogation des annotations RDF.

7.2 Evaluation du système de questions-réponses MEANS

Dans cette section, nous nous intéressons à l'évaluation du système de questions-réponses MEANS. Nous présentons les critères qui entrent en jeu pour évaluer les SQR en domaine ouvert et en domaine médical. Ensuite nous présentons les données d'évaluation ainsi que les résultats obtenus.

2. Un focus est identifié comme étant une entité médicale reliée à la réponse attendue par une relation

3. Une relation principale est identifiée comme étant une relation ayant la réponse attendue pour objet ou sujet

4. Concept Unique Identifier

5. `http://jena.sourceforge.net/`

7.2.1 Critères et mesures d'évaluation : performances vs. rapidité

Plusieurs campagnes d'évaluation de SQR ont été menées en domaine ouvert. Citons par exemple Quaero [6] (français, anglais), TREC [7] (anglais), CLEF [8] (multilingue) et NTCIR [9] (japonais). Pour le domaine médical, les challenges de SQR anglais sont rares. La tâche Genomics du challenge TREC peut être citée comme une piste exploitable pour les tâches de QR en anglais, bien que non officiellement introduite comme telle. Nous avons néanmoins pu collecter un corpus «standard» préparé par des experts du domaine pour la tâche de questions-réponses en anglais [10] (cf. section 7.2.2).

La performance des systèmes de questions-réponses est souvent évaluée en mesurant le MRR (Mean Reciprocal Rank) et potentiellement la précision et le rappel des réponses retournées, en supposant qu'un système renvoie une liste ordonnée de réponses ; éventuellement avec un nombre maximum fixé de réponses.

- Le MRR : la Moyenne des Réciproques du Rang de la première bonne réponse (e.g. 1 si la première réponse est correcte, 0,5 si la seconde est la première correcte, etc.)
- Le rappel : mesure la proportion des réponses correctes retournées par le système parmi l'ensemble des réponses correctes.
- La précision : mesure la proportion de réponses correctes trouvées parmi toutes les réponses trouvées par le système de questions-réponses.

Ces mesures demandent de bien définir : qu'est-ce qu'une réponse « correcte » ? De quels critères la réponse doit-elle tenir compte ?

La réponse à ces questions peut passer par les deux points clés suivants :

1. **La granularité de la réponse.** En domaine ouvert : à une question comme « Où se trouve le musée de Louvre ? », plusieurs réponses sont possibles :
   - *Le musée du Louvre se trouve en France.*
   - *Le Musée du Louvre se situe en plein centre de Paris.*
   - *Le Musée du Louvre se trouve dans un ancien palais royal construit par François Ier sur l'ancien château fort du roi Philippe Auguste.*
   - *Le Musée du Louvre se trouve du côté opposé de l'avenue de l'Opéra et on l'atteint après une marche d'une dizaine de minutes depuis l'hôtel Horset Opéra.*

   On peut ainsi distinguer plusieurs évaluations des réponses qui peuvent être jugées, par exemple, correctes, complètes, incomplètes ou fausses. Le point d'ambiguïté dans cet exemple étant surtout que le type de réponse attendue n'est pas précis, il peut aussi bien s'agir du nom d'un quartier, d'une ville, d'un pays ou d'une adresse complète. En domaine médical, le problème se pose aussi pour les ques-

6. http://www.quaero.org
7. http://trec.nist.gov
8. http://clef.isti.cnr.it
9. http://research.nii.ac.jp/ntcir
10. Nous remercions Dina Demner-Fushman pour nous avoir indiqué cette piste.

tions ouvertes (e.g. « To what extent can we say that cell phones are harmful ? » mais est plus restreint pour les questions fermées (e.g. factuelles, liste) où le type de réponse attendue est explicite ou pour les questions booléennes.

2. **La justification de la réponse.** La justification est l'extrait textuel (e.g. phrase) contenant la réponse extraite.
   - La réponse peut être correcte avec une justification fausse.
   - Mais aussi, la réponse peut être fausse aujourd'hui même si la justification est correcte. Par exemple « Qui est le président de la France » ? la réponse « Jacques Chirac » est fausse même si la justification « Discours de M. Jacques Chirac, Président de la République française devant l'université de Beïda. » est correcte.

En plus de la performance, un autre critère est important dans le cadre des systèmes de questions-réponses : **la rapidité**. Ainsi, en domaine médical, (Takeshita et al. 2002) ont montré que les médecins [11] nécessitent d'accéder à l'information en moins de 30 secondes et abandonnent la recherche au-delà.

7.2.2 Données d'évaluation

Pour évaluer notre système, nous utilisons le corpus de questions-réponses construit par (Mollá 2010, Mollá et Santiago-Martínez 2011). Ce corpus (Corpus for Evidence Based Medicine Summarisation) est une collection de résumés d'articles regroupés par question et provenant de la section des questions cliniques du Journal of Family Practice. Il est écrit au format XML et annoté avec les éléments suivants :

- La question clinique
- La/Les réponse(s) à la question
- Passages justifiant les réponses extraits du Journal of Family Practice
- Référence aux articles par leurs identifiants PubMed

Nous évaluons une séquence de 50 questions médicales de la collection de (Mollá 2010, Mollá et Santiago-Martínez 2011). Cette séquence correspond aux 20 questions booléennes et aux 19 questions factuelles qui ont une sémantique qui peut être exprimée avec notre ontologie. Les 11 questions restantes emploient des entités médicales et/ou des relations qui ne sont pas couvertes par notre système. Les réponses (extraits de texte), qui n'existent pas dans les articles et qui étaient préparées manuellement à partir des différentes justifications trouvées dans les articles, nous ont servi pour évaluer manuellement les réponses retournées par le système MEANS. Une évaluation automatique n'est pas possible car il faut évaluer les justifications retournées par le système (les phrases qui contiennent les réponses) manuellement. Une réponse est considérée correcte si l'entité médicale ou la valeur booléenne retournée est correcte et la justification est correcte aussi.

11. Cette étude a été effectuée avec un groupe de 5 médecins de famille, 5 résidents et 5 médecins internes

7.2.3 Questions booléennes

Deux éléments entrent en jeu pour l'évaluation des questions booléennes : (i) la valeur de la réponse : oui ou non et (ii) sa justification. Nous mesurons donc la précision de notre SQR pour ces questions par rapport à la valeur de la réponse (oui/non) et par rapport à la pertinence de sa justification.

Le tableau 7.2 présente les résultats obtenus pour les 20 questions booléennes traitées. $N1$, $N2$ et $N3$ désignent les niveaux de relaxation employés. Pour les questions booléennes, nous avons évalué les réponses obtenues par le niveau 1 puis celles obtenus par les trois niveaux en même temps.

Nous désignons par "autres" les questions booléennes dont les relations principales sont différentes de celles que nous traitons, par exemple *have a role* (Does routine amniotomy have a role in normal labor ?), *tolerated* (Are any oral iron formulations better tolerated than ferrous sulfate ?).

La précision et le rappel sur les questions booléennes sont ainsi de **45%** si on applique uniquement le niveau 1 de relaxation et de **60%** si on applique le niveau 3. La précision et le rappel ont la même valeur ici car (i) si le système retourne 'non' à une question et que la réponse correcte était 'oui' il n'y a pas de justifications à évaluer et cela compte pour une erreur et (ii) dans le cas ou le système retourne correctement la valeur 'oui' nous n'avons pas eu de justifications erronées.

Nous distinguons quatre principaux types d'erreurs :

- **T1** : la réponse n'existe pas dans le corpus d'articles dont nous disposons (quelques fichiers ne contiennent parfois que le titre de l'article).
- **T2** : Les réponses/justifications sont sur deux phrases ou plus (e.g. description des résultats d'une expérimentation)
- **T3** : Une entité ou une relation importante n'a pas été reconnue. C'est le cas, par exemple, des questions suivantes :
  - *Does reducing smoking in the home protect children from the effects of second-hand smoke ?*
  - *What is the appropriate use of sunscreen for infants and children ?*
  - *Do preparticipation clinical exams reduce morbidity and mortality for athletes ?*
- **T4** : Questions qui nécessitent des connaissances externes ou des inférences, par exemple :
  - *Does heat or cold work better for acute muscle strain ?* (cold n'est pas un problème ici, mais plutôt "Cold Therapy" ou "Cryotherapy", pareil pour heat ("heat therapy")
  - *Does psychiatric treatment help patients with intractable chronic pain ?* (les traitements cités dans les articles peuvent être des sous types de "psychiatric treatment ", par exemple : "Cognitive therapy")
  - *Do antiarrhythmics prevent sudden death in patients with heart failure ?* (les médicaments cités dans les articles peuvent être des sous-types, par exemple : "beta blockers" sont des "antiarrhythmic drugs")

Pour certaines questions la réponse est indiquée comme existante mais les articles qui la contiennent sont fournis uniquement avec leur titre sans

Type	**nbr de questions**	**N1**	**N1,N2,N3**	**Types d'erreurs**
Treats	9	5	5	T1, T2, T4
Prevents	4	0	3	T3, T4
Diagnoses	1	1	1	–
Autres	6	3	4	T3, T4
Total	20	9	12	–

TABLE 7.2 – *Nombre de questions booléennes répondues correctement et types d'erreurs par catégorie*

leur contenu. Ce problème a influencé les résultats des questions de type "treats" même si elles étaient, en général, correctement analysées. Aussi, pour certaines questions de type "treats", le problème était que la réponse est étalée sur plusieurs lignes. Par exemple pour la question "Does yoga speed healing for patients with low back pain ?", la justification [12] dans le corpus est : *"In a case series, 21 women aged >60 years (mean age, 75) with hyperkyphosis, participated in twice-weekly 1-hour sessions of hatha yoga for 12 weeks. Measured height increased by a mean of 0.52 cm, forward curvature diminished, patients were able to get out of chairs faster, and they had longer functional reach. Eleven patients (48%) reported increased postural awareness/improvement and improved well-being ; 58% perceived improvement in their physical functioning."*.

Pour les erreurs de type T3, la relaxation des requêtes a amélioré les résultats (cas des questions de type prevents ou autres). En effet, elle a permis de trouver davantage d'éléments qui répondaient aux requêtes SPARQL et qui menaient de ce fait à des réponses positives (*oui*). Ces réponses positives se trouvaient être correctes alors que sans relaxation le système répondait *non* par absence d'information. Nous n'avons pas eu de bruit car la relaxation conserve au moins deux entités de la question (une hypothèse qui est peut être un peu stricte pour certaines questions mais qui garantit un minimum de précision).

Dans certains cas, même si la question est analysée correctement et la requête qui lui est associée est correcte, le système n'a pas trouvé la bonne réponse car des inférences ou des connaissances externes étaient nécessaires. Par exemple, pour la question "Do antiarrhythmics prevent sudden death in patients with heart failure ?", les réponses étaient par exemple : (i) *Beta-blockers to reduce mortality in patients with systolic dysfunction : a meta-analysis*, (ii) *Beta-blockers are particularly effective in people with a high sympathetic drive (i.e., high pulse rates) to lower blood pressure and reduce cardiovascular risk*. Ainsi, pour retrouver la réponse (i) il fallait inférer que "heart failure" est un type de "systolic dysfunction", que les "Beta-blockers" sont des "antiarrhythmics" et que les relations "reduce mortality" et "prevent (sudden) death" sont suffisamment similaires.

12. Cette justification a été prise du fichier de référence qui contient les réponses et les justifications, pour nous, le fichier en question (12356608) comporte uniquement : "Yoga for women with hyperkyphosis : results of a pilot study."

7.2.4 Questions factuelles

Une réponse à une question factuelle est jugée correcte si la bonne entité médicale est retournée (CUI correct) avec une justification correcte. Ainsi, les entités médicales incomplètes (e.g. un mot en moins) ou qui comportent du bruit ont été considérées comme fausses et les réponses correctes avec de fausses justifications ont été considérées comme fausses. Le tableau 7.3 présente les résultats obtenus sur 19 questions factuelles du corpus de référence. *N*1, *N*2 et *N*3 désignent les niveaux de relaxation 1, 2 et 3.

Type	**nbr qs**	N1		N1+N2+N3	
		MRR	P@5	MRR	P@5
Treats	8	0,625	70,58	1	62,5
Prevents	1	0	–	1	60
Diagnoses	5	0	–	0,432	25
Causes	1	0	–	1	20
Manages	3	0,66	100	0,5	80
2 TRAs ou plus	1	0	–	1	66,66
Total	39	**0,42**	**85,71**	**0,77**	**57,47**

TABLE 7.3 – *Questions factuelles : MRR et précision à 5 réponses (en %)*

La précision finale sur toutes les questions est de 85,71% sans relaxation avec un MMR de 0,42. Le MRR a augmenté de 0,35 avec la relaxation qui a aussi permis de retrouver plus de réponses. Cependant la précision s'est dégradée avec la relaxation, ce qui était attendu, mais la perte de 0.28 points de précision au total reste acceptable au vu de l'augmentation du nombre de réponses et surtout au vu de l'augmentation du MRR. Les résultats obtenus pour chaque question factuelle sont détaillés dans le tableau 7.4.

Pour un exemple concret, nous présentons les résultats obtenus pour la question Q1 du tableau 7.4 : « What is the best treatment for oral thrush in healthy infants ? ». Sans relaxation, deux réponses correctes sont obtenues pour cette question « Nystatin » et « Nystatin Suspension » [13]. Avec le niveau 1 de relaxation (suppression des valeurs textuelles exactes) nous obtenons une réponse supplémentaire et correcte « Fluconazole ». Le niveau 2 de relaxation (ici, suppression de l'entité "healthy infants") poursuit dans le même sens en permettant d'obtenir une nouvelle réponse correcte « Gentian Violet ». Le niveau 3 de relaxation (suppression de la relation principale, ici *treats*) apporte 5 nouvelles réponses dont une seule correcte « Miconazole Gel ». Notons que le tableau détaillé 7.4 s'arrête à 5 réponses et ne montre donc pas toutes les réponses correspondant à cet exemple, mais uniquement les 5 premières qui sont ici obtenues dès la première réponse du niveau 3 de relaxation, qui est ici une réponse correcte.

13. Nystatin et Nystation suspensions ont deux CUI UMLS différents, donc il s'agit bien de deux entités médicales différentes.

Catégorie	Q	P@5 (x)		RR		P@5/catégorie		MRR/catégorie	
		*N*1	*N*1, 2, 3	*N*1	*N*1, 2, 3	*N*1	*N*1, 2, 3	*N*1	*N*1, 2, 3
Treats	Q1	100 (3)	100 (5)	1	1	70,58	62,5	0,625	1
	Q2	na (0)	40 (5)	1	1				
	Q3	60 (5)	60 (5)	1	1				
	Q4	na (0)	100 (5)	0	1				
	Q5	na (0)	60 (5)	0	1				
	Q6	100 (2)	60 (5)	1	1				
	Q7	80 (5)	80 (5)	1	1				
	Q8	na (0)	100 (5)	0	1				
Diagnoses	Q9	na (0)	20 (5)	0	0,33	na (0)	0,25	0	0,432
	Q10	na (0)	0 (5)	0	0				
	Q11	na (0)	40 (5)	0	0,5				
	Q12	na (0)	20 (5)	0	1				
	Q13	na (0)	50 (4)	0	0,33				
Prevents	Q14	na (0)	60 (5)	0	1	na (0)	60	0	1
Manages	Q15	na (0)	na (0)	0	0	100	80	0,66	0,5
	Q16	100 (3)	80 (5)	1	1				
	Q17	100 (3)	80 (5)	1	0,5				
Causes	Q18	na (0)	20 (5)	0	1	na (0)	20	0	1
2 TRA	Q19	na (0)	66,66 (3)	0	1	na (0)	66,66	0	1

TABLE 7.4 – *Résultats détaillées par question et catégories*
P@5 (x) : précision à 5 réponses et x nombre de réponses trouvées (nombre maximum fixé à 5)
na : "no answers", la précision ne peut pas être calculée
Ni : niveau de relaxation i

7.2.5 Discussion

Les résultats montrent l'intérêt de la relaxation pour l'amélioration des performances du système *MEANS* : augmentation de la précision pour les questions booléennes et du nombre de réponses et du MRR pour les questions factuelles. La précision a augmenté pour les questions booléennes car l'absence d'information due à la forme initiale, fortement contrainte, de la requête est interprétée comme une réponse négative (non) alors que les réponses positives et correctes sont présentes dans le corpus et accessibles avec la relaxation.

Les premières observations des cas d'erreur pour les questions booléennes mettent en évidence qu'il est nécessaire de prendre en compte plus efficacement la présence de **négations** dans les phrases car leur impact est décisif pour les question booléennes (e.g. *"Another Cochrane review found no added benefit in function from combining deep transverse friction massage with ultrasound or a placebo ointment"*). Aussi déterminer le **niveau de certitude** joue un rôle important dans la sélection des bonnes justifications. Par exemple, l'annotation de la phrase *"There's insufficient evidence to support specific physiotherapy methods or orthoses (braces), shock wave therapy, ultrasound, or deep friction massage"* a mené à une mauvaise justification car le niveau de certitude n'a pas été pris en compte.

Les principales observations pour les questions factuelles sont les défaillances détectées dans les cas de comparaison qui ne sont pas encore pris en compte par le système *MEANS*. Par exemple, la question « What are the most effective treatments for *PB* ? » qui cherche les meilleurs traitements à une maladie *PB* ne peut pas avoir comme réponse un traitement *T*1 avec la justification "T1 is less effective than T2". *T*1 peut être accepté comme réponse correcte si des justifications contraires existent car parfois des études aboutissent à des résultats contradictoires (c'est un cas que nous avons eu dans notre évaluation).

Conclusion du chapitre

Dans ce chapitre, nous avons présenté le SQR médical MEANS et ses résultats sur des questions médicales extraites du corpus de questions-réponses proposé par (Mollá 2010, Mollá et Santiago-Martínez 2011). Les résultats obtenus sont encourageants aussi bien en termes de précision que de MRR. Le système se fonde sur une analyse sémantique des corpus et des questions. Il construit plusieurs requêtes SPARQL pour représenter la question de l'utilisateur et les trie par ordre de précision (spécificité) décroissante avant de les exécuter. Les résultats obtenus montrent que ce tri au niveau des requêtes permet d'ordonner efficacement les réponses retournées. Les résultats montrent aussi que la diminution de la précision pour les questions factuelles due à la relaxation reste acceptable comme petite contrepartie de l'apport de nouvelles réponses et de l'amélioration du MRR.

Cependant, malgré la généricité de la méthode d'analyse de questions proposée, il reste des questions auxquelles nous n'avons pas répondu (11 questions sur 50) car elles ont des types de réponses attendues non encore traités. Dans ce cadre, nous envisageons d'améliorer la couverture du

système en traitant de nouveaux types de questions et aussi en étudiant la combinaison de notre SQR avec des systèmes de recherche d'information classiques basés sur les mots clés ou sur l'analyse des dépendances syntaxiques.

Conclusion et perspectives

8

8.1 Bilan

Au cours de cette thèse, nous avons proposé une approche sémantique pour la recherche de réponses à des questions médicales :

- L'approche proposée permet de traiter différents types de questions, parmi lesquelles, les questions avec plus d'un type de réponse attendue et/ou plus d'un focus.
- Elle permet une analyse profonde des questions et des textes utilisés pour trouver la réponse, en utilisant plusieurs méthodes d'extraction d'information fondées sur (i) les connaissances du domaine (e.g. UMLS) et (ii) des techniques de TAL (e.g. utilisation de patrons ou règles, apprentissage automatique). Ces méthodes ont permis l'extraction des entités médicales, des relations sémantiques et aussi des informations supplémentaires ou contextuelles (e.g. informations sur les patients : âge, sexe, etc.). Ces méthodes ont été testées sur des corpus de genres différents pour étudier les problèmes de portabilité.
- Notre approche est basée sur les technologies du web sémantique qui offrent plus d'expressivité et d'interopérabilité ainsi que des langages formels standards et ouvrent des perspectives pour l'exploitation des bases de connaissances publiées sur le web.

8.2 Perspectives

Dans la continuité directe de notre travail de thèse, nous envisageons plusieurs perspectives à différents niveaux.

8.2.1 Analyse des questions médicales

Il s'agit du premier module d'un système de questions-réponses et la base de toute la recherche de réponses. Se tromper sur le type de la question ou de la réponse attendue, ne pas détecter un terme médical affecte tout le processus de recherche de réponses. Au niveau de cette tâche, nous envisageons de :

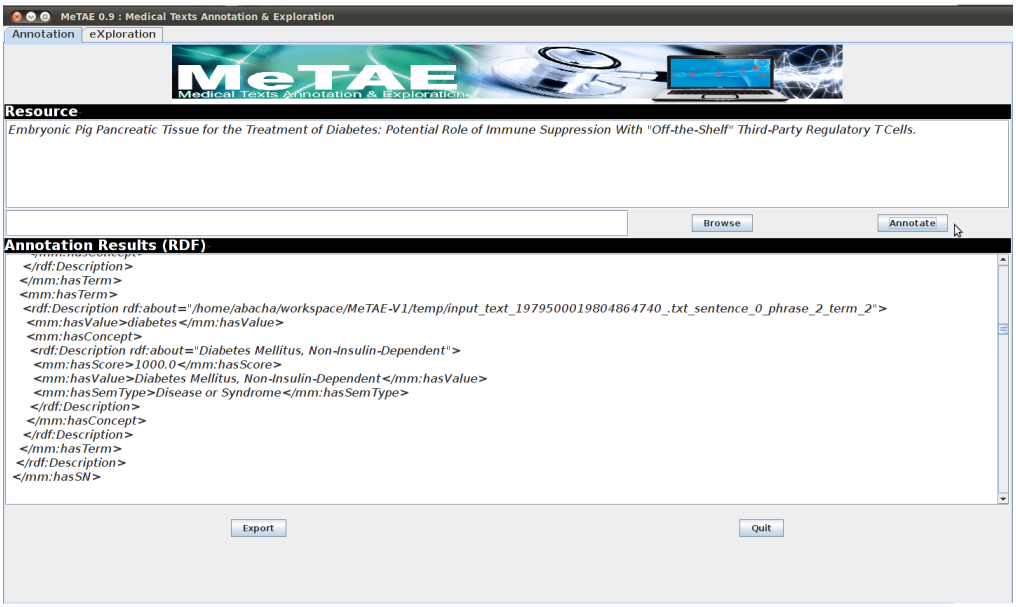
- Améliorer la méthode actuelle d'analyse de questions en ajoutant d'autres types de questions (e.g. when, why) et d'autres patrons pour chaque type.

- Proposer des approches statistiques ou hybrides pour le typage de la question et pour la construction des formes affirmatives et simplifiées des questions.
- Tester un autre type de méthodes pour construire les requêtes SPARQL à partir des questions médicales qui consiste à apprendre automatiquement à traduire les questions posées en langage naturel en requêtes structurées.

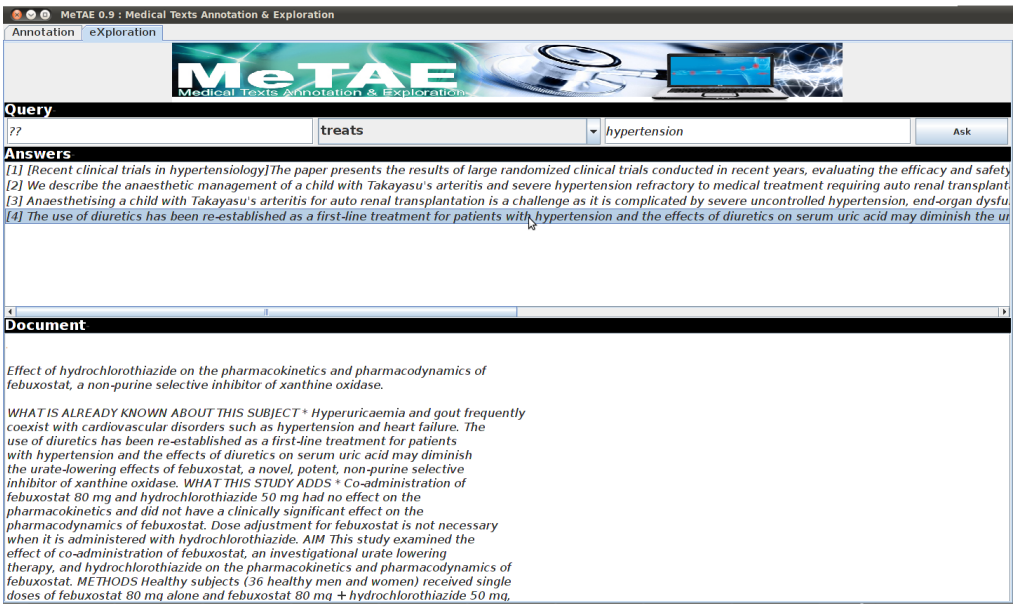
8.2.2 Analyse des documents médicaux

L'annotation des textes médicaux est une tâche très importante pour diverses applications. Nous avons proposé plusieurs méthodes d'extraction d'information pour la reconnaissance des entités médicales et des relations sémantiques qui les lient. Dans un stade intermédiaire du travail sur le SQR MEANS nous avons aussi proposé **la plateforme MeTAE** (Medical Texts Annotation and Exploration) pour permettre à des utilisateurs de soumettre des textes de leur choix, de lancer leur annotation avec les méthodes d'extraction d'information présentées dans ce manuscrit, puis d'explorer ces annotations en posant directement des requêtes SPARQL restreintes à travers un formulaire. Plus précisément, MeTAE permet de rechercher une ou deux entités médicales reliées par une relation sémantique ou par une simple relation de co-occurrence. De ce fait, la partie interrogation de la plateforme MeTAE n'est pas assimilable à un SQR car elle ne permet pas de poser des questions en langage naturel. MeTAE reste cependant une plateforme utile pour l'analyse de documents médicaux cibles soumis par les utilisateurs. Nous donnons brièvement dans ce qui suit de plus amples détails sur les deux modules de cette plateforme.

1. **Annotation sémantique.** La plateforme MeTAE utilise des ressources sémantiques du domaine ainsi que les méthodes d'extraction d'information (cf. chapitres 2 et 3) développées pour l'annotation (i) des entités médicales et (ii) des relations sémantiques qui les relient (cf. figure 8.1a). Ces informations sont sauvegardées à la volée sous format RDF.
2. **Interrogation sémantique.** La plateforme MeTAE permet aussi d'interroger les annotations sémantiques. Les triplets RDF générés dans la phase d'annotation sont interrogés avec des requêtes SPARQL formulées suivant les concepts et les relations de l'ontologie $MESA$. Ces requêtes sont obtenues après la traduction de la requête utilisateur, composée à travers un formulaire qui fixe les relations sémantiques à utiliser mais laisse à l'utilisateur le choix des arguments à rechercher (cf. figure 8.1b). Par exemple, l'utilisateur peut chercher le traitement (inconnu) d'une maladie dont il écrit le nom ou juste la validation d'un fait donné (e.g. « La pénicilline est-elle bien un traitement pour le « S. aureus » »). Les termes entrés par l'utilisateur sont analysés par nos méthodes de reconnaissance d'entités médicales (cf. chapitre 2) afin de connaître les termes médicaux concernés par la requête. La plateforme permet aussi de rechercher une entité médicale donnée ou deux entités médicales liées par un simple lien de co-occurrence ou une relation sémantique du domaine.



(a) *MeTAE-Interface d'annotation*



(b) *MeTAE-Interface d'interrogation*

FIGURE 8.1 – *La plateforme MeTAE pour l'annotation et l'interrogation sémantique des textes médicaux*

Dans le cadre de l'analyse des textes médicaux en anglais, nous envisageons les évolutions suivantes :

- Améliorer les méthodes de reconnaissance des entités médicales en entraînant un chunker sur des textes médicaux.
- Proposer des méthodes et/ou exploiter des outils pour la résolution de l'anaphore pour améliorer l'extraction de relations.
- Rechercher des dépendances syntaxiques au lieu de relations sémantiques du domaine si ces dernières sont inexistantes dans les annotations.
- Inclure toutes ces améliorations au sein du système de questions-réponses *MEANS* et de la plateforme MeTAE.

8.2.3 Interroger des ressources externes disponibles

SPARQL permet l'interrogation des annotations RDF des corpus médicaux mais aussi l'interrogation des bases de connaissances externes écrites en RDF. Pouvoir interroger des ressources sémantiques externes nous permet d'avoir accès à plus d'informations qui sont susceptibles d'être mises à jour fréquemment. Cela permet aussi d'effectuer des inférences par des règles et des connaissances du domaine. Par exemple, si on cherche un traitement à une maladie $M1$, il est utile de savoir si $M1$ est un sous type d'une maladie $M2$ surtout si on sait que des traitements T_i sont adaptés pour $M2$ (et peuvent donc être présentés comme candidats possibles pour le traitement de $M1$).

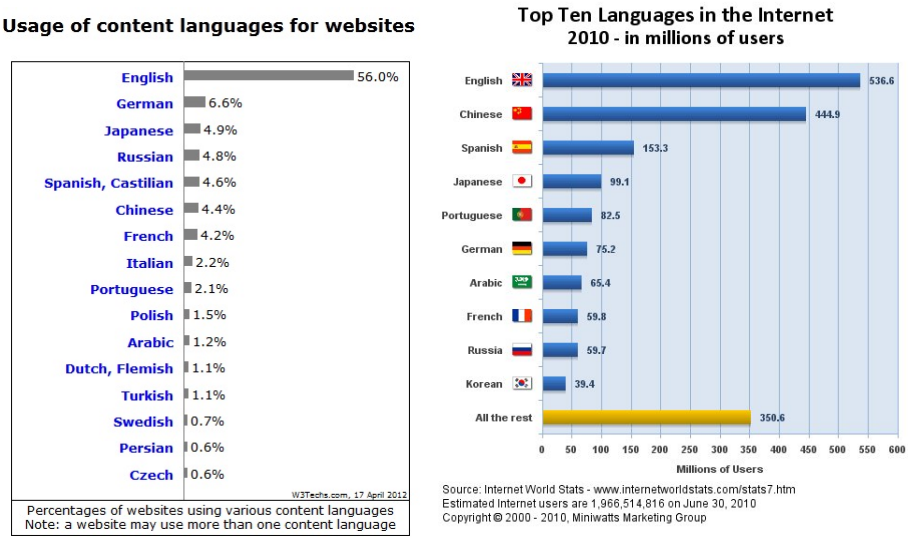
Dernièrement, de plus en plus de serveurs SPARQL (ou SPARQL Endpoint) sont mis à disposition sur le Web et notamment pour le domaine (bio)médical (e.g. la plateforme NCBO [1] permettant d'interroger les bases de connaissances regroupées par BioPortal telles que SNOMED CT ou NCI). Dans ce cadre nous envisageons d'aligner l'ontologie exploitée par le SQR *MEANS* avec les ontologies des bases de connaissances existantes mais aussi de lier les annotations directement aux concepts définis dans ces bases de connaissances afin d'avoir des descriptions sémantiques supplémentaires des entités médicales extraites du corpus. Cela permettra, par exemple, de rechercher une partie de l'information requise par l'utilisateur dans les bases de connaissances publiées en ligne dans le cas où ces informations n'ont pas été retrouvées par l'annotation des corpus médicaux.

8.2.4 Multilinguisme : questions ou documents dans d'autres langues

Interroger des documents en anglais est important étant donné la position de cette langue en tant que véhicule principal des articles scientifiques mais aussi de la plupart des sites et des informations sur le Web (cf. figure 8.2a). Il reste cependant très intéressant d'interroger des documents dans d'autres langues et notamment en français pour permettre de traiter les différents et nombreux documents médicaux comme les dossiers hospitaliers des patients et les recommandations pour la pratique clinique.

Deux approches principales peuvent être citées pour le passage d'une langue à une autre dans le cadre des systèmes de questions réponses et

1. http://alphasparql.bioontology.org/



(a) *Classement de point de vue sites web* (b) *Classemement de point de vue utilisateurs*

FIGURE 8.2 – *Les Langues sur Internet*

pour répondre à des questions dans une langue L1 à partir de documents dans une langue L2 :

- Analyser la question dans la langue originale (L1) puis traduire les mots clés pour interroger les documents en L2. L'avantage est que l'analyse de question est plus efficace car la question est analysée en langue d'origine (donc correcte syntaxiquement et sémantiquement). L'inconvénient est que les termes sont traduits hors contexte (les termes sont traduits un à un sans utiliser la forme complète de la question originale).
- Traduire la question de la langue L1 à la langue L2 puis l'analyser et interroger la collection de documents en L2. L'avantage est que l'analyse de la question est effectuée sur sa forme traduite ce qui permet d'utiliser les méthodes d'extraction de la langue L2, aussi, les termes sont traduits comme éléments d'une phrase/question posant un contexte précis.

Combiner ces deux approches peut être une solution afin de tirer profit des avantages des deux méthodes. Par exemple, (Grau et al. 2006a) ont appliqué ces deux stratégies en parallèle et ont proposé le système de questions-réponses MUSCLEF. Pour chercher des réponses à des questions en français à partir de documents en anglais, MUSCLEF se fonde sur (i) un premier sous-système appelé MUSQUAT qui traduit les mots clés issus de l'analyse de la question en français en utilisant un traducteur automatique FR-EN et recherche les réponses dans une collection de documents en anglais et (ii) un second sous-système qui utilise la traduction de la question puis se base sur le système de questions-réponses monolingue anglais QALC (Ferret et al. 2002) pour trouver la réponse.

Cette approche nécessite un premier système qui traduit les questions (nous utilisons Google Translate), mais aussi un système qui permet l'analyse des questions dans la langue d'origine et donc nécessite la disponibilité d'outils permettant d'analyser les questions en français. Nous avons

commencé à travailler sur des méthodes permettant d'annoter des textes médicaux en français par projection (cf. chapitre 4) ; notre objectif étant d'annoter des corpus médicaux français avec des concepts et des relations et utiliser ces corpus annotés pour développer des méthodes statistiques pour l'extraction d'information à partir de textes médicaux ou de questions.

Aussi, chercher des réponses à des questions en français ou en anglais à partir de documents en français nécessite de développer de telles méthodes. Dans ce cadre, nous envisageons d'améliorer et continuer à travailler sur l'extraction d'information (ie. extraction des concepts et des relations) à partir de textes médicaux en français par projection.

Annexes

A

Sommaire

A.1 La liste des questions utilisées dans l'évaluation

1. How effective are nasal steroids combined with nonsedating antihistamines for seasonal allergic rhinitis ?
2. Is screening for lead poisoning justified ?
3. Do nasal decongestants relieve symptoms ?
4. Do antiarrhythmics prevent sudden death in patients with heart failure ?
5. What are the most effective ways you can help patients stop smoking ?
6. Does antepartum perineal massage reduce intrapartum lacerations ?
7. What is the best way to screen for breast cancer in women with implants ?
8. What is the best treatment for oral thrush in healthy infants ?
9. How can you prevent migraines during pregnancy ?
10. What is the appropriate use of sunscreen for infants and children ?
11. Does treatment with donepezil improve memory for patients with mild cognitive impairment ?
12. What is the appropriate evaluation and treatment of children Who are "toe walkers" ?
13. Does furosemide decrease morbidity or mortality for patients with diastolic or systolic dysfunction ?
14. What interventions reduce the risk of contrast nephropathy for high-risk patients ?
15. Can counseling prevent or treat postpartum depression ?
16. What is the best way to manage phantom limb pain ?
17. What is appropriate fetal surveillance for women with diet-controlled gestational diabetes ?
18. Howshould patients with Barrett's esophagus be monitored ?
19. Does reducing smoking in the home protect children from the effects of second-hand smoke ?
20. What treatment works best for tennis elbow ?
21. Does routine amniotomy have a role in normal labor ?
22. Does psychiatric treatment help patients with intractable chronic pain ?
23. Howdoes pentoxifylline affect survival of patients with alcoholic hepatitis ?
24. Does heat or cold work better for acute muscle strain ?
25. Prophylactic oxytocin : Before or after placental delivery ?
26. Howaccurate is stress radionuclide imaging for diagnosis of CAD ?
27. What causes a low TSH level with a normal free T4 level ?
28. Who should get hepatitis A vaccination ?

29. What is the best way to distinguish type 1 and 2 diabetes ?
30. Do preparticipation clinical exams reduce morbidity and mortality for athletes ?
31. What are effective medical treatments for adults with acute migraine ?
32. What are the indications for bariatric surgery ?
33. Howbest to manage the patient in term labor Whose group B strep status is unknown ?
34. Howoften should you follow up on a patient with newly diagnosed hypothyroidism ?
35. What is the best diagnostic approach to postmenopausal vaginal bleeding in women taking hormone replacement therapy ?
36. What are the indications for treatment with angiotensin-converting enzyme ACE inhibitors in patients with diabetes ?
37. What is the value of screening for heart disease with an exercise stress test (EST) in an asymptomatic person ?
38. What are the most effective treatments for bacterial vaginosis in nonpregnant women ?
39. Which tool is most useful in diagnosing bipolar disorder in children ?
40. Does screening for diabetes in at-risk patients improve long-term outcomes ?
41. What is the best treatment for diabetic neuropathy ?
42. Do inhaled beta-agonists control cough in URIs or acute bronchitis ?
43. Does yoga speed healing for patients with low back pain ?
44. Is methylphenidate useful for treating adolescents with ADHD ?
45. Should we screen women for hypothyroidism ?
46. What is the best approach for managing recurrent bacterial vaginosis ?
47. What are effective treatments for panic disorder ?
48. Is there a role for theophylline in treating patients with asthma ?
49. Are any oral iron formulations better tolerated than ferrous sulfate ?
50. Other than anticoagulation, What is the best therapy for those with atrial fibrillation ?

Bibliographie

Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus : the MetaMap program. Dans *AMIA Annu Symp Proc*, pages 17–21, 2001. (Cité pages 16, 17, 22 et 57.)

Sofia J. Athenikos et Hyoil Han. Biomedical question answering : A survey. *Computer Methods and Programs in Biomedicine*, 99(1) :1–24, 2010. (Cité page 77.)

Christelle Ayache. Campagne EVALDA/EQueR – Évaluation en question-réponse, rapport final. Rapport technique, ELDA, Paris, 2005. `http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf`. (Cité page 57.)

Nguyen Bach et Sameer Badaskar. A review of relation extraction. URL `http://www.cs.cmu.edu/~{}nbach/papers/A-survey-on-Relation-Extraction.pdf`. 2007. (Cité page 57.)

François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, et Jean Morissette. Bio2rdf : Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5) :706 – 716, 2008. (Cité pages 6 et 86.)

Asma Ben Abacha et Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities : a rule based approach. *Journal of Biomedical Semantics*, 2(Suppl 5) :S4, 2011a. (Cité page 38.)

Asma Ben Abacha et Pierre Zweigenbaum. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. Dans *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, volume 6608 de *Lecture Notes in Computer Science*, pages 139–150, Tokyo, Japan, 2011b. ISBN 978-3-642-19436-8. URL `http://dx.doi.org/10.1007/978-3-642-19400-9`. (Cité page 38.)

Asma Ben Abacha et Pierre Zweigenbaum. Medical entity recognition : A comparison of semantic and statistical methods. Dans *BioNLP 2011 Workshop*, pages 56–64, Portland, Oregon, USA, 2011c. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0207`. (Cité pages 14 et 57.)

Asma Ben Abacha et Pierre Zweigenbaum. Une approche hybride pour la détection automatique des relations sémantiques entre entités médicales. Dans *Journées francophones d'informatique médicale (JFIM)*, Tunis, Tunisie, 2011d. (Cité page 38.)

Asma Ben Abacha et Pierre Zweigenbaum. Analyse et transformation des questions médicales en requêtes sparql. Dans *CORIA (COnférence en Recherche d'Informations et Applications)*, Bordeaux, 2012a. (Cité page 90.)

Asma Ben Abacha et Pierre Zweigenbaum. Medical question answering : Translating medical questions into sparql queries. Dans *ACM SIGHIT International Health Informatics Symposium (IHI 2012)*, Miami, FL, USA, January 2012b. (Cité page 90.)

Asma Ben Abacha et Pierre Zweigenbaum. Une étude comparative empirique sur la reconnaissance des entités médicales. *Traitement Automatique des Langues (TAL)*, 53(1), 2012c. (Cité page 14.)

Asma Ben Abacha, Pierre Zweigenbaum, et Aurélien Max. Extraction d'information automatique en domaine médical par projection interlangue : vers un passage à l'échelle. Dans *Proceedings of TALN 2012 (Traitement automatique des langues naturelles)*, Grenoble, 2012. (Cité page 56.)

Delphine Bernhard et Anne-Laure Ligozat. Analyse automatique de la modalité et du niveau de certitude : application au domaine médical. Dans *Proceedings of TALN'11*, Montpellier, 2011. (Cité page 15.)

Olivier Bodenreider. Lexical, terminological and ontological resources for biological text mining. Dans Sophia Ananiadou et John McNaught, éditeurs, *Text mining for biology and biomedicine*, pages 43–66. Artech House, Boston, Massachusetts, 2006. (Cité page 17.)

Chih-Chung Chang et Chih-Jen Lin. *LIBSVM : a library for support vector machines*, 2001a. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. (Cité page 25.)

Chih-Chung Chang et Chih-Jen Lin. *LIBSVM : a library for support vector machines*, 2001b. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. (Cité page 48.)

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, et Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *JBI*, 34(5) :301–310, 2001. (Cité page 17.)

Faisal Mahbub Chowdhury, Asma Ben Abacha, Alberto Lavelli, et Pierre Zweigenbaum. Two different machine learning techniques for drug-drug interaction extraction. Dans *Isabel Segura-Bedmar, Paloma Martinez, and Daniel Sanchez-Cisneros, editors, Proceedings DDIExtraction2011, First Challenge Task on Drug-Drug Interaction Extraction 2011 (SEPLN 2011 satellite workshop), volume 761 of CEUR Workshop Proceedings*, pages 19–26, Huelva, Spain, 2011. Association for Computational Linguistics. (Cité page 49.)

P. Cimiano, P. Haase, J. Heizmann, M. Mantel, et R. Studer. Towards portable natural language interfaces to knowledge bases : The case of the orakel system. Dans *Data Knowledge Engineering (DKE), 65(2)*, pages 325–354, 2008. (Cité pages 81 et 95.)

J.J. Cimino et G.O. Barnett. Automatic knowledge acquisition from MEDLINE. *Methods Inf Med*, 32(2) :120–130, 1993. (Cité page 40.)

Danica Damljanović, Milan Agatonovic, et Hamish Cunningham. Natural language interface to ontologies : combining syntactic analysis and ontology-based lookup through the user interaction. Dans *7th Extended Semantic Web Conference (ESWC2010)*, June 2010. (Cité page 80.)

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel D. Martin, et Xiaodan Zhu. Machine-learned solutions for three stages of clinical in-

formation extraction : the state of the art at i2b2 2010. *JAMIA*, 18(5) : 557–562, 2011. (Cité page 29.)

Thierry Delbecque, Pierre Jacquemart, et Pierre Zweigenbaum. Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales dans un système de questions-réponses : impact de la source des documents explorés. Dans *CORIA*, pages 101–115, Grenoble, 2005. CLIPS. (Cité pages 19 et 39.)

Louise Deléger, Magnus Merkel, et Pierre Zweigenbaum. Contribution to terminology internationalization by word alignment in parallel corpora. Dans *AMIA Annu Symp Proc.*, pages 185–189, Washington, DC, NOV 2006. (Cité page 59.)

Louise Deléger, Magnus Merkel, et Pierre Zweigenbaum. Translating medical terminologies through word alignment in parallel text corpora. *JBI*, 42(4) :692–701, 2009. URL `http://dx.doi.org/10.1016/j.jbi.2009.03.002`. Epub 2009 Mar 9. (Cité pages 59 et 63.)

Dina Demner-Fushman et Jimmy Lin. Knowledge extraction for clinical question answering : Preliminary results. Dans *AAAI 2005 Workshop on Question Answering in Restricted Domains*. AAAI, 2005. (Cité pages 85 et 86.)

Dina Demner-Fushman et Jimmy J. Lin. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. Dans *ACL*, 2006. (Cité page 85.)

Dina Demner-Fushman et Jimmy J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, pages 63–103, 2007. (Cité page 5.)

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, et Yong Yu. Searching questions by identifying question topic and question focus. Dans *Proceedings of ACL-08 : HLT*, pages 156–164, Columbus, Ohio, June 2008. Association for Computational Linguistics. (Cité page 95.)

Maud Ehrmann. *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation.* Thèse de doctorat, Université Paris 7, Juin 2008. (Cité page 15.)

Maud Ehrmann et Guillaume Jacquet. Vers une double annotation des entités nommées. *Traitement Automatique des Langues*, 47(3) :63–88, 2006. (Cité page 19.)

Asif Ekbal et Sivaji Bandyopadhyay. Named entity recognition using support vector machine : A language independent approach. *International Journal of Electrical and Electronics Engineering*, 4(2) :155–170, 2010. (Cité page 25.)

Faiza Elkateb-Gara. Extraction d'entités nommées pour la recherche d'informations précises. Dans *4e Congrès ISKO France, Grenoble*, 2003. (Cité page 17.)

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, M. Lee Chambliss, Daniel C Vinson, James J Stevermer, et Eric A. Pifer. Obstacles to answering doctors' questions about patient care with evidence : qualitative study. *British Medical Journal*, 324 :710, 2002. (Cité page 83.)

John W. Ely, Jerome A. Osheroff, Paul N. Gorman, Mark H. Ebell, M. Lee Chambliss, Eric A. Pifer, et P. Zoe Stavri. A taxonomy of generic clinical questions : classification study. *British Medical Journal*, 321 :429–432, 2000. (Cité pages 5, 83 et 84.)

Mehdi Embarek. *Un système de question-réponse dans le domaine médical : le système Esculape*. PhD thesis, Université Paris-Est, 2008. (Cité page 84.)

Mehdi Embarek et Olivier Ferret. Learning patterns for building resources about semantic relations in the medical domain. Dans *LREC'08*, May 2008. (Cité pages 17, 19, 39 et 40.)

R. Engelbrecht. Expert systems for medicine—functions and developments. *Zentralbl Gynakol*, 119(9) :428–434, 1997. (Cité page 3.)

Shixi Fan, Xuan Wang, Xiaolong Wang, et Yaoyun Zhang. A new question analysis approach for community question answering system. *International Journal on Asian Language Processing*, 19(3) :95–108, 2009. (Cité page 95.)

Óscar Ferrández, Rubén Izquierdo, Sergio Ferrández, et José Luis Vicedo González. Addressing ontology-based question answering with collections of user queries. *Inf. Process. Manage.*, 45(2) :175–188, 2009. (Cité page 81.)

Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Christian Jacquemin, Laura Monceaux, Isabelle Robba, et Anne Vilnat. How nlp can improve question answering. *Knowledge Organization Journal, Vol. 29, N3-4*, pages 135–155, 2002. (Cité pages 78 et 129.)

Nordine Fourour. Nemesis : un systme de reconnaissance incrémentielle des entités nommées pour le français. Dans *TALN 2002*, pages 255–264, 2002. (Cité page 17.)

Carol Friedman, Pauline Kra, et Andrey Rzhetsky. Two biomedical sublanguages : a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35 :222–235, 2002. (Cité page 18.)

Oana Frunza et Diana Inkpen. Extraction of disease-treatment semantic relations from biomedical sentences. Dans *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 91–98, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W10-1912`. (Cité pages 41, 50 et 53.)

B. Grau et J.P. Chevallet. *La Recherche d'informations précises : traitement automatique de la langue, apprentissage et connaissances pour les systèmes de questions-réponses*. Collection Traité IC2, Hermès-Lavoisier, 2007. (Cité page 79.)

Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat, Michael Bagur, et Kevin Séjourné. The bilingual system musclef at qa@clef 2006. Dans *CLEF*, pages 454–462, 2006a. (Cité page 129.)

Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat, et Laura Monceaux. Frasques : A question-answering system in the equer evaluation campaign. Dans *LREC'06*, Gênes, Italie, 2006b. (Cité page 78.)

Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, et Kenneth Laughery. Baseball : an automatic question-answerer. Dans *Papers presented at the May*

9-11, 1961, western joint IRE-AIEE-ACM computer conference, IRE-AIEE-ACM '61 (Western), pages 219–224, New York, NY, USA, 1961. ACM. (Cité page 75.)

Ralph Grishman et Beth Sundheim. Message Understanding Conference - 6 : A brief history. Dans *Proc. of COLING*, pages 466–471, Copenhagen, Denmark, AUG 1996. (Cité pages 17 et 19.)

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, et Ludovic Quintard. Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. Dans *Proc. of the Fifth Linguistic Annotation Workshop (LAW-V)*, Portland, OR, Juin 2011. Association for Computational Linguistics. (Cité page 20.)

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, et Ian H. Witten. The WEKA data mining software : An update. *SIGKDD Explorations*, 11(1), 2009. (Cité page 41.)

Ying He et Mehmet Kayaalp. Biological entity recognition with Conditional Random Fields. Dans *AMIA Annu Symp Proc*, pages 293–297, 2008. (Cité page 18.)

Marti Hearst. Automatic acquisition of hyponyms from large text corpora. Dans *Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992)*, pages 539–545, 1992. (Cité page 39.)

Donald Hindle. Noun classification from predicate argument structures. Dans *Proc. 28th Annual Meeting of the Association for Computational Linguistics (ACL'90)*, Berkeley, USA, 1990. (Cité page 39.)

Jerry R. Hobbs et Ellen Riloff. Information extraction. Dans Nitin Indurkhya et Fred J. Damerau, éditeurs, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921. (Cité page 34.)

Martyn O. Hotvedt. Continuing medical education : actually learning rather than simply listening. *JAMA, 275 :1638*, 275(21) :1637–1638, 1996. (Cité page 3.)

Xiaoli Huang, Jimmy Lin, et Dina Demner-Fushman. Evaluation of pico as a knowledge representation for clinical questions. Dans *AMIA Annu Symp Proc*, pages 359–363, 2006. (Cité page 86.)

Hideki Isozaki et Hideto Kazawa. Efficient support vector classifiers for named entity recognition. Dans *Proceedings of COLING-2002*, pages 390–396, 2002. (Cité page 18.)

Pierre Jacquemart et Pierre Zweigenbaum. Towards a medical question-answering system : a feasibility study. Dans Robert Baud, Marius Fieschi, Pierre Le Beux, et Patrick Ruch, éditeurs, *Medical Informatics Europe*, volume 95 de *Studies in Health Technology and Informatics*, pages 463–468, Amsterdam, 2003. IOS Press. (Cité pages 83, 84 et 85.)

Thorsten Joachims. Text categorization with support vector machines : Learning with many relevant features. Dans *ECML-98, 10th European Conference on Machine Learning*, 1998. (Cité pages 40 et 48.)

N Kang, EM van Mulligen, et JA Kors. Comparing and combining chunkers of biomedical text. *J Biomed Inform*, 44(2) :354–360, nov 2010. (Cité page 21.)

Boris Katz. From sentence processing to information access on the world wide web. Dans *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 1999. (Cité pages 4 et 75.)

Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy J. Lin, Gregory Marton, Alton Jerome McFarland, et Baris Temelkuran. Omnibase : Uniform access to heterogeneous data for question answering. Dans *NLDB*, pages 230–234, 2002. (Cité pages 4 et 80.)

S. Keerthi et S. Sundararajan. CRF versus SVM-struct for sequence labeling. Dans *Yahoo Research Technical Report*, 2007. (Cité page 25.)

Khaled Khelif et Rose Dieng-Kuntz. Web sémantique et mémoire d'expériences sur les biopuces. Dans *Web Sémantique Médical (WSM'2004)*, Rouen, 2004. (Cité page 19.)

Christopher S. G. Khoo, Syin Chan, et Yun Niu. Extracting causal knowledge from a medical database using graphical patterns. Dans *Proc. 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 336–343, 2000. (Cité page 40.)

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, et Evan Herbst. Moses : Open source toolkit for statistical machine translation. Dans *Proceedings of ACL*, Czech Republic, 2007. (Cité page 62.)

Leila Kosseim et Thierry Poibeau. Extraction de noms propres à partir de textes variés : problématique et enjeux. Dans *TALN 2001*, pages 365–371, 2001. (Cité page 17.)

C.H. Lee, C. Khoo, et J.C. Na. Automatic identification of treatment relations for medical ontology learning : An exploratory study. Dans I.C. McIlwaine, éditeur, *Knowledge Organization and the Global Information Society : Proceedings of the Eighth International ISKO Conference*, 2004. (Cité pages 41, 45 et 46.)

Tyne Liang et Ping-Ke Shih. Empirical textual mining to protein entities recognition from PubMed corpus. Dans *NLDB'05*, pages 56–66, 2005. (Cité pages 17 et 18.)

Adam Lopez, Mike Nossal, Rebecca Hwa, et Philip Resnik. Word-level alignment for multilingual resource acquisition. Dans *LREC Workshop on Linguistic Knowledge Acquisition and Representation : Bootstrapping Annotated Data*, Las Palmas, Spain, MAY 2002. ELRA. (Cité page 59.)

V. Lopez et E Motta. Aqualog : An ontology-portable question answering system for the semantic web. Dans *Proceedings of the International Conference on Natural Language for Information Systems (NLDB)*, pages 89–102, 2004. (Cité page 95.)

Vanessa Lopez, Victoria S. Uren, Enrico Motta, et Michele Pasin. Aqualog : An ontology-driven question answering system for organizational semantic intranets. *J. Web Sem.*, 5(2) :72–105, 2007. (Cité page 80.)

Vanessa Lopez, Victoria S. Uren, Marta Sabou, et Enrico Motta. Is question answering fit for the semantic web ? : A survey. *Semantic Web*, 2(2) : 125–155, 2011. (Cité page 76.)

Yajuan Lü, Sheng Li, Tiejun Zhao, et Muyun Yang. Learning Chinese bracketing knowledge based on a bilingual language model. Dans *Proceedings of COLING-2002*, pages 591–598, 2002. (Cité page 59.)

Andrew McCallum et Wei Li. Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons. Dans *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, 2003. (Cité page 17.)

S M Meystre, G K Savova, K C Kipper-Schuler, et J F Hurdle. Extracting information from textual documents in the electronic health record : a review of recent research. *Yearb Med Inform*, 35 :128–44, 2008. (Cité page 18.)

Stéphane M. Meystre et Peter J. Haug. Comparing natural language processing tools to extract medical problems from narrative text. Dans *AMIA Annu Symp Proc*, pages 525–529, 2005. (Cité page 17.)

Dan Moldovan, Marius Paca, Sanda Harabagiu, A Harabagiu, et Mihai Surdeanu. Performance issues and error analysis in an open-domain question answering system. Dans *ACM Trans. Information Systems, 21(2)*, pages 133–154, 2003. (Cité page 76.)

Dan I. Moldovan, Sanda M. Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Girju, et Vasile Rus. Lasso : A tool for surfing the answer net. Dans *Proceedings of the eighth Text REtrieval Conference (TREC-8)*, 1999. (Cité page 78.)

Diego Mollá. A corpus for evidence based medicine summarisation. Dans *Proc. ALTA 2010*, pages 76–80, Melbourne, 2010. (Cité pages 118 et 123.)

Diego Mollá et María Elena Santiago-Martínez. Development of a corpus for evidence medicine summarisation. Dans *Australasian Language Technology Workshop (ALTA 2011)*, Canberra, Australia, 2011. (Cité pages 118 et 123.)

Véronique Moriceau et Xavier Tannier. Fidji : Using syntax for validating answers in multiple documents. *Information Retrieval, Special Issue on Focused Information Retrieval*, 13(5) :507–533, 2010. (Cité pages 79 et 82.)

David Nadeau et Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1) :3–26, January 2007. URL `http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002`. Publisher : John Benjamins Publishing Company. (Cité page 57.)

A Névéol, JG Mork, AR Aronson, et SJ Darmoni. Evaluation of French and English MeSH indexing systems with a parallel corpus. Dans *AMIA Annu Symp Proc.*, pages 565–9, Washington, DC, Novembre 2005. (Cité page 59.)

Yun Niu et Graeme Hirst. Analysis of semantic classes in medical text for question answering. Dans *Proceedings of the ACL-2004 Workshop Question Answering in Restricted Domains*, 2004. (Cité pages 4 et 85.)

Yun Niu, Graeme Hirst, Gregory McArthur, et Patricia Rodriguez-Gianolli. Answering clinical questions with role identification. Dans

Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine - Volume 13, BioMed '03, pages 73–80, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1118958.1118968. (Cité pages 5 et 85.)

Franz Josef Och et Herman Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4), 2004. (Cité page 62.)

Sebastian Padó et Guillaume Pitel. Annotation précise du français en sémantique de rôles par projection cross-linguistique. Dans *Proceedings of TALN 2007, Toulouse, France*, 2007. (Cité page 59.)

Thierry Poibeau. Le repérage des entités nommées, un enjeu pour les systèmes de veille. *Terminologies Nouvelles (actes du colloque Terminologie et Intelligence Artificielle, TIA'99, Nantes)*, (19) :43–51, 1999. (Cité page 17.)

Thierry Poibeau. *Extraction automatique d'information : du texte brut au web sémantique*. Hermès science publications, 2003. ISBN 9782746206106. URL http://books.google.com/books?id=SrUaAAAACAAJ. (Cité page 57.)

A. Popescu, O. Etzioni, et H Kautz. Towards a theory of natural language interfaces to databases. Dans *Proceedings of the International Conference on Intelligent User Interfaces (IUI'03)*, pages 149–157, 2003. (Cité pages 79, 91 et 95.)

Wanda Pratt et Meliha Yetisgen-Yildiz. Concept identification : MetaMap vs. people. Dans *AMIA Annu Symp Proc*, 2003. (Cité page 17.)

Denys Proux, François Rechenmann, Laurent Julliard, Violaine Pillet, et Bernard Jacq. Detecting gene symbols and names in biological texts : A first step toward pertinent information extraction. Dans *Proceedings of Genome Informatics*, pages 72–80, Tokyo, Japan : Universal Academy Press, 1998. (Cité page 18.)

Chiong Raymond et Wang Wei. Named entity recognition using hybrid machine learning approach. Dans *IEEE ICCI*, pages 578–583, 2006. (Cité page 17.)

Fabio Rinaldi, James Dowdall, et Gerold Schneider. Answering questions in the genomics domain. Dans *Proc. ACL04 Workshop on Question Answering in Restricted Domains*, 2004. (Cité pages 4 et 84.)

Thomas C. Rindflesch, Carol A. Bean, et Charles A. Sneiderman. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Dans *AMIA Annu Symp Proc*, pages 704–708, 2000a. (Cité pages 18 et 40.)

Thomas C. Rindflesch, Lorraine Tanabe, John N. Weinstein, et Lawrence Hunter. Edgar : Extraction of drugs, genes and relations from the biomedical literature. Dans *Proceedings of Pacific Symposium on Biocomputing*, pages 517–528, 2000b. (Cité page 17.)

Angus Roberts, Robert Gaizauskas, et Mark Hepple. Extracting clinical relationships from patient narratives. Dans *BioNLP 2008*, 2008. (Cité page 40.)

Barbara Rosario et Marti A. Hearst. Classifying semantic relations in bioscience text. Dans *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, July 2004. (Cité pages 14, 17, 30, 41, 50, 57 et 61.)

David L. Sackett, Sharon E. Straus, W. Scott Richardson, William Rosenberg, et R. Brian Haynes. *Evidence-Based Medicine : How to Practice and Teach EBM*. Churchill Livingstone, Edinburgh, 2000. (Cité pages 5, 86 et 91.)

N Sager, M Lyman, N T Nhàn, et L J Tick. Medical language processing : applications to patient data representation and automatic encoding. *Meth Inform Med*, 34(1–2) :140–6, 1995. (Cité page 18.)

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, et Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes) : architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17 :507–513, 2010. (Cité page 57.)

Satoshi Sekine. Definition, dictionaries and tagger of extended named entity hierarchy. Dans *Proc. of LREC*, Lisbon, Portugal, 2004. (Cité page 19.)

G Shadow et C MacDonald. Extracting structured information from free text pathology reports. Dans *AMIA Annu Symp Proc*, Washington, DC, 2003. (Cité pages 17 et 39.)

Julilus Sim et Chris C. Wright. The kappa statistic in reliability studies : Use, interpretation, and sample size requirements. *Physical Therapy*, March 2005. (Cité page 64.)

B Stapley et G Benoit. Biobibliometrics : information retrieval and visualization from co-occurrences of gene names in medline abstracts. Dans *Proceedings of the Pacific Symposium on Biocomputing*, pages 529–540, Hawaii, USA, 2000. (Cité page 40.)

Fabian M. Suchanek, Georgiana Ifrim, et Gerhard Weikum. Combining linguistic and statistical analysis to extract relations from Web documents. Dans *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Apr 2006. (Cité page 40.)

Harumi Takeshita, Dianne Davis, et Sharon E. Straus. Clinical evidence at the point of care in acute medicine : a handheld usability case study. Dans *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, pages 1409–1413, 2002. (Cité page 118.)

Isabelle Tellier et Marc Tommasi. Champs Markoviens Conditionnels pour l'extraction d'information. Dans Éric Gaussier et François Yvon, éditeurs, *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès, Paris, 2010. (Cité page 27.)

Rafael M Terol, Patricio Martínez-Barco, et Manuel Palomar. A knowledge based method for the medical question answering problem. *Computers in Biology and Medicine*, 37(10) :1511–1521, 2007. (Cité pages 4, 5 et 84.)

Cynthia A. Thompson, Mary Elaine Califf, et Raymond J. Mooney. Active learning for natural language parsing and information extraction. Dans *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pages 406–414, Bled, Slovenia, June 1999. (Cité page 60.)

Katrin Tomanek et Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. Dans *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W09-1906. (Cité page 60.)

Özlem Uzuner, Brett R. South, S. Shen, et Scott L Duvall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5) :552–556, 2011. Epub 2011 Jun 16. (Cité pages 14, 18, 19, 40 et 57.)

Suzan Verberne. Developing an approach for why-question answering. Dans *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics : Student Research Workshop*, EACL'06, pages 39–46, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1609039.1609044. (Cité page 95.)

Spela Vintar, Paul Buitelaar, et Martin Volk. Semantic relations in concept-based cross-language medical information retrieval. Dans *ECML/PKDD Workshop on adaptive text extraction and mining (ATEM)*, Cavtat-Dubrovnik, 2003. (Cité page 15.)

Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, et Ji Wang. Automatic extraction of hierarchical relations from text. Dans *Proceedings of the Third European Semantic Web Conference (ESWC 2006)*, 2006. (Cité page 39.)

Xinglong Wang. Rule-based protein term identification with help from automatic species tagging. Dans *Proceedings of CICLING 2007*, pages 288–298, 2007. (Cité page 17.)

W. A. Woods. Progress in natural language understanding : an application to lunar geology. Dans *Proceedings of the June 4-8, 1973, national computer conference and exposition*, AFIPS '73, pages 441–450, New York, NY, USA, 1973. ACM. (Cité page 75.)

J. Xiao, J. Su, G. Zhou, et C. Tan. Protein-protein interaction extraction : a supervised learning approach. Dans *Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine (SMBM)*, 2005. (Cité page 40.)

David Yarowsky et Grace Ngai. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. Dans *NAACL 2001*, 2001. (Cité pages 58 et 59.)

Alexander Yeh, Alexander Morgan, Marc Colosimo, et Lynette Hirschman. BioCreAtIvE task 1A : gene mention finding evaluation. *BMC Bioinformatics*, 6 Suppl 1, 2005. ISSN 1471-2105. URL http://dx.doi.org/10.1186/1471-2105-6-S1-S2. (Cité page 18.)

Hong Yu, Carl Sable, et Hai Ran Zhu. Classifying medical questions based on an evidence taxonomy. Dans *Proc. AAAI'05 Workshop on Question Answering in Restricted Domains*, 2005. URL http://www.uwm.edu/~hongyu/publications.html. (Cité page 83.)

Pierre Zweigenbaum. L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical. *RIA*, 18 :111–137, 2004. (Cité page 17.)

Pierre Zweigenbaum, Pierre Jacquemart, Natalia Grabar, et Benoît Habert. Building a text corpus for representing the variety of medical language. Dans V. L. Patel, R. Rogers, et R. Haux, éditeurs, *Proceedings of the 10th World Congress on Medical Informatics*, pages 290–294, Londres, 2001. (Cité page 18.)

Publications

Plusieurs idées et résultats présentés dans cette thèse ont déjà été publiés dans les articles suivants, disponibles sur `http://sites.google.com/site/asmabenabacha/publications` :

Ben Abacha, A. (2009). Questions-réponses dans le domaine médical : une approche sémantique. *Dans MajecSTIC 2009*, Avignon.

Ben Abacha, A. et Zweigenbaum, P. (2010a). Annotation et interrogation sémantiques de textes médicaux. *Dans Actes Atelier Web Sémantique Médical 2010 à IC 2010*, pages 61–70, Nîmes.

Ben Abacha, A. et Zweigenbaum, P. (2010b). Automatic extraction of semantic relations between medical entities : Application to the treatment relation. *Dans* Collier, N. et Hahn, U., éditeurs : *Proceedings of the Fourth International Symposium on Semantic Mining in Biomedicine (SMBM)*, pages 4–11, Hinxton, Cambridgeshire, UK.

Ben Abacha, A. et Zweigenbaum, P. (2010c). Metae : Plate-forme d'annotation automatique et d'exploration sémantiques pour le domaine médical. *Dans Démonstrations à TALN 2010, 4 pages*, Montréal, Canada.

Ben Abacha, A. et Zweigenbaum, P. (2011a). Une approche hybride pour la détection automatique des relations sémantiques entre entités médicales. *Dans Journées francophones d'informatique médicale (JFIM)*, Tunis, Tunisie. (Cité page 38.)

Ben Abacha, A. et Zweigenbaum, P. (2011b). Automatic extraction of semantic relations between medical entities : a rule based approach. *Journal of Biomedical Semantics*, 2(Suppl 5) :S4. (Cité page 38.)

Ben Abacha, A. et Zweigenbaum, P. (2011c). A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. *Dans Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, volume 6608 de *Lecture Notes in Computer Science*, pages 139–150, Tokyo, Japan. (Cité page 38.)

Ben Abacha, A. et Zweigenbaum, P. (2011d). Medical entity recognition : A comparison of semantic and statistical methods. *Dans BioNLP 2011 Workshop*, pages 56–64, Portland, Oregon, USA. Association for Computational Linguistics. (Cité pages 14 et 57.)

Ben Abacha, A. et Zweigenbaum, P. (2012a). Analyse et transformation des questions médicales en requêtes sparql. *Dans CORIA (COnférence en Recherche d'Informations et Applications)*, Bordeaux. (Cité page 90.)

Ben Abacha, A. et Zweigenbaum, P. (2012b). Medical question answering : Translating medical questions into sparql queries. *Dans ACM SIGHIT International Health Informatics Symposium (IHI 2012)*, Miami, FL, USA. (Cité page 90.)

Ben Abacha, A. et Zweigenbaum, P. (2012c). Une étude comparative empirique sur la reconnaissance des entités médicales. *Traitement Automatique des Langues (TAL)*, 53(1). (Cité page 14.)

Ben Abacha, A., Zweigenbaum, P. et Max, A. (2012). Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle. *Dans Proceedings of TALN 2012 (Traitement automatique des langues naturelles)*, Grenoble. (Cité page 56.)

Chowdhury, F. M., Ben Abacha, A., Lavelli, A. et Zweigenbaum, P. (2011). Two different machine learning techniques for drug-drug interaction extraction. *Dans Isabel Segura-Bedmar, Paloma Martinez, and Daniel Sanchez-Cisneros, editors, Proceedings DDIExtraction2011, First Challenge Task on Drug-Drug Interaction Extraction 2011 (SEPLN 2011 satellite workshop), volume 761 of CEUR Workshop Proceedings*, pages 19–26, Huelva, Spain. Association for Computational Linguistics. (Cité page 49.)

Grouin, C., BenAbacha, A., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Ligozat, A.-L., Minard, A.-L., Rosset, S. et Zweigenbaum, P. (2010). CARAMBA : Concept, assertion, and relation annotation using machine-learning based approaches. *Dans* Özlem Uzuner et al., éditeur : *i2b2 Medication Extraction Challenge Workshop*. :i2b22010 :i2b22010 :i2b22010 :i2b22010

Minard, A.-L., Ligozat, A.-L., Ben Abacha, A., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zweigenbaum, P. et Grouin, C. (2011). Hybrid methods for improving information access in clinical documents : Concept, assertion, and relation identification. *Journal of the American Medical Informatics Association (JAMIA)*, 18(5):588–593.

Titre Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS

Résumé La recherche de réponses précises à des questions formulées en langue naturelle renouvelle le champ de la recherche d'information. De nombreux travaux ont eu lieu sur la recherche de réponses à des questions factuelles en domaine ouvert. Moins de travaux ont porté sur la recherche de réponses en domaine de spécialité, en particulier dans le domaine médical ou biomédical. Plusieurs conditions différentes sont rencontrées en domaine de spécialité comme les lexiques et terminologies spécialisés, les types particuliers de questions, entités et relations du domaine ou les caractéristiques des documents ciblés. Dans une première partie, nous étudions les méthodes permettant d'analyser sémantiquement les questions posées par l'utilisateur ainsi que les textes utilisés pour trouver les réponses. Pour ce faire nous utilisons des méthodes hybrides pour deux tâches principales : (i) la reconnaissance des entités médicales et (ii) l'extraction de relations sémantiques. Ces méthodes combinent des règles et patrons construits manuellement, des connaissances du domaine et des techniques d'apprentissage statistique utilisant différents classifieurs. Ces méthodes hybrides, expérimentées sur différents corpus, permettent de pallier les inconvénients des deux types de méthodes d'extraction d'information, à savoir le manque de couverture potentiel des méthodes à base de règles et la dépendance aux données annotées des méthodes statistiques. Dans une seconde partie, nous étudions l'apport des technologies du Web sémantique pour la portabilité et l'expressivité des systèmes de questions-réponses. Dans le cadre de notre approche, nous exploitons les technologies du Web sémantique pour annoter les informations extraites en premier lieu et pour interroger sémantiquement ces annotations en second lieu. Enfin, nous présentons notre système de questions-réponses, appelé MEANS, qui utilise à la fois des techniques de TAL, des connaissances du domaine et les technologies du Web sémantique pour répondre automatiquement aux questions médicales.

Mots-clés Questions-réponses, extraction d'information, domaine médical, entités nommées, relations sémantiques, apprentissage, patrons.

Title Finding precise answers to medical questions : the question-answering system MEANS

Abstract With the dramatic growth of digital information, finding precise answers to natural language questions is more and more essential for retrieving domain knowledge in real time. Many research works tackled answer retrieval for factual questions in open domain. Less works were performed for domain-specific question answering such as the medical domain. Compared to the open domain, several different conditions are met in the medical domain such as specialized vocabularies, specific types of questions, different kinds of domain entities and relations. Document characteristics are also a matter of importance, as, for example, clinical texts may tend to use a lot of technical abbreviations while forum pages

may use long "approximate" terms. We focus on finding precise answers to natural language questions in the medical field. A key process for this task is to analyze the questions and the source documents semantically and to use standard formalisms to represent the obtained annotations. We propose a medical question-answering approach based on : (i) NLP methods combing domain knowledge, rule-based methods and statistical ones to extract relevant information from questions and documents and (ii) Semantic Web technologies to represent and interrogate the extracted information.

Keywords Question answering, information extraction, medical domain, named entities, semantic relations, machine learning, patterns.