Analyse et transformation des questions médicales en requêtes SPARQL

Asma Ben Abacha* — **Pierre Zweigenbaum***

LIMSI-CNRS, BP 133 91403 Orsay cedex

abacha@limsi.fr, pz@limsi.fr

RÉSUMÉ. *La conception des systèmes de questions-réponses nécessite une analyse profonde des questions posées. Cette tâche primordiale requiert d'être étudiée et évaluée séparément. Dans cet article, nous nous intéressons à l'analyse de questions en domaine médical. Plus précisément, nous étudions la transformation de questions posées en langage naturel en requêtes basées sur un langage formel. Cette étude examine 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. Notre étude supporte l'hypothèse que SPARQL peut représenter de nombreux types de questions. Les résultats obtenus sur un corpus de questions réelles montrent que les requêtes SPARQL sont correctes pour 62 % des questions. Ces résultats dépendent fortement de la couverture des relations traitées. Si les systèmes d'extraction d'information sont fiables, SPARQL peut représenter correctement 98 % des questions du corpus d'évaluation.*

ABSTRACT. *In this paper we tackle question analysis in the medical domain. More precisely, we study how to translate a natural language question into a machine-readable representation. The underlying transformation process requires determining three key points: (i) What are the main characteristics of medical questions? (ii) Which methods are the most suited to extract these characteristics? and (iii) how to translate the extracted information into a machine-readable representation? We present a complete question analysis approach including medical entity recognition, relation extraction and automatic translation to SPARQL queries. Our study supports the claim that SPARQL can represent a wide range of natural language questions.*

MOTS-CLÉS : *Analyse de question, Requêtes, Extraction d'information, Questions-réponses*

KEYWORDS: *Question Analysis, Queries, Information Extraction, Question Answering*

CORIA 2012, pp. 111–122, Bordeaux, 21-23 mars 2012

1. Introduction

Les systèmes de questions-réponses (SQR) visent à répondre automatiquement à des questions posées en langage naturel en recherchant les réponses dans une collection de documents ou à partir du Web. De tels systèmes permettent de fournir un accès rapide à l'information recherchée, un point crucial en domaine médical pour les praticiens mais aussi les patients. La performance des SQR est souvent évaluée en mesurant la précision et le rappel des réponses retournées. L'évaluation se fait en général en boite noire, sans évaluer chaque composant du système. Plusieurs campagnes d'évaluation ont été menées en domaine ouvert, citons par exemple récemment Quæro [1] sur des textes du web en français et en anglais.

Un SQR nécessite deux entrées : le corpus utilisé pour extraire les réponses et la question elle-même. Chacune de ces deux entrées doit être analysée correctement pour pouvoir trouver le meilleur appariement question-réponse. Ce processus implique de représenter à la fois les questions et les réponses candidates avec une représentation formelle homogène pouvant être traitée par les systèmes d'information. Dans la dernière décennie, plusieurs méta-langages tels que RDF(S) [2] et OWL [3] 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 APIs (p.ex. Jena [4]).

Dans cet article, nous nous intéressons au premier module d'un SQR : *« Analyse de question »*, une étape importante qui mérite d'être analysée et évaluée séparément. Nous discutons les principales caractéristiques des questions médicales et proposons une approche consistant à (i) extraire les informations les plus importantes à partir de la question (p.ex. entités médicales, relations sémantiques) et (ii) transformer les informations extraites en requête(s) SPARQL [5] (SPARQL Protocol and RDF Query Language), 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 définis dans une ontologie de référence.

Dans la section 2 nous présentons un aperçu sur les travaux menés autour de l'analyse de question en domaine ouvert et en domaine médical. Nous étudions dans la section 3 quelques caractéristiques des questions médicales. Nous présentons ensuite notre approche pour l'analyse de questions médicales et la construction de requêtes SPARQL dans la section 4. La section 5 est consacrée aux expérimentations et aux résultats obtenus sur un corpus comportant des questions médicales réelles extraites à partir du Journal of Family Practice [6] (JFP) [7].

---

1. http://www.quaero.org/
2. http://www.w3.org/TR/rdf-syntax/
3. http://www.w3.org/TR/owl-ref/
4. http://jena.sourceforge.net/
5. http://www.w3.org/TR/rdf-sparql-query/
6. http://www.jfponline.com/
7. Ce travail a été partiellement soutenu par OSEO dans le cadre du programme Quæro.

2. État de l'art

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. Dans ce cadre, plusieurs travaux ont été menés pour la transformation des questions en requêtes SQL ou autres (p.ex. PQL). Le système PRECISE (Popescu *et al.,* 2003) analyse les questions et essaie de les transformer dans des requêtes SQL correspondantes. En cas d'échec, ils cherchent un appariement de graphes dans lequel les mots de la question sont projetés sur les relations de la base de données, les colonnes et les valeurs. Dernièrement, de plus en plus de travaux utilisent les ontologies et aussi les technologies du Web sémantique dans le cadre des systèmes de questions-réponses. Le système ORAKEL (Cimiano *et al.,* 2008) transforme les questions factuelles ou les questions WH (who, what, which... sauf celles avec why et how) en F-logic (frame logic) ou en requêtes SPARQL. 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 (Fan *et al.,* 2009), d'autres se concentrent sur un type particulier de question, comme les questions why (Verberne, 2006).

L'analyse des questions en domaine médical est différente de celle en domaine ouvert à cause des spécificités de ce domaine de spécialité (p.ex. rôle des pronoms interrogatifs, relations sémantiques spécifiques entre les entités médicales). En effet, en domaine ouvert, le pronom interrogatif *When* par exemple indique une date alors que dans le domaine médical, il peut indiquer une condition (p.ex. *When should you suspect community-acquired MRSA ?*). Aussi, les types de questions diffèrent entre ces deux domaines. Dans le domaine médical, plusieurs taxonomies de questions ont été proposées. (Ely *et al.,* 2000) ont proposé une taxonomie qui comporte les 10 catégories de questions les plus fréquentes à partir de 1396 questions collectées. Nous citons ici les cinq premiers modèles ou catégories (qui représentent 40 % des questions) : (i) *What is the drug of choice for condition x ?*, (ii) *What is the cause of symptom x ?*, (iii) *What test is indicated in situation x ?*, (iv) *What is the dose of drug x ?* et (v) *How should I treat condition x (not limited to drug treatment) ?*.

3. Caractéristiques des questions médicales

Plusieurs caractéristiques des questions ont été plus au moins mises en évidence et/ou utilisées dans différents travaux. D'après notre étude sur des questions réelles, nous présentons ici quelques caractéristiques des questions médicales.

– *Le type de question.* Yes/No (ou booléenne), Définition, Factuelle (la réponse est une entité médicale ou une information précise), Liste ou Complexe (p.ex. Why).

– *Le type de la réponse attendue (TRA), pour les questions WH.* Ce peut être un Traitement, un Test Médical, un Problème médical, etc.

– *Le Focus.* L'entité médicale la plus proche de la réponse attendue (p.ex. *What's the best treatment for pyogenic granuloma ?* a comme focus *pyogenic granuloma*).

– *La relation principale.* Pour les questions WH, nous définissons la relation principale comme étant celle reliant la réponse attendue et le focus. Dans les questions booléennes, cela correspond à la relation la plus importante (objet de la question) entre deux entités médicales (deux focus).

– *Les entités médicales.* Reconnaître ces entités est une étape indispensable qui doit faire face à la grande variation terminologique du domaine médical (synonymes, abréviations, etc.) et à l'évolution continue de cette terminologie (nouveaux concepts médicaux, etc.).

– *Les relations sémantiques.* Elles précisent la sémantique de la question (p.ex. entre un Traitement et un Problème, plusieurs relations sont possibles : traiter, causer, prévenir, ...). Extraire les relations de la question permet d'identifier la relation principale et les relations contextuelles (p.ex. âge du patient, son historique).

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 certaines questions et donc la couverture de l'étape d'analyse de questions. Dans la section 4, nous présentons notre approche qui permet aussi le traitement de questions ayant plusieurs focus et/ou plusieurs types de réponses attendues.

4. Transformation de la question en requête(s) SPARQL

La sélection de SPARQL comme langage formel vise à éviter la perte d'expressivité dans la phase de construction de la requête (p.ex. identifier uniquement un focus et/ou un type de réponse attendue alors que la question en langage naturel en contient davantage). De plus, les ontologies sont plus partageables que les schémas de bases de données relationnelles. La sélection nous permet ainsi, par exemple, de répondre à une partie des questions des utilisateurs en accédant aux nombreuses bases de connaissances publiques publiées en ligne et décrites suivant des ontologies de domaine. SPARQL est un langage standard recommandé par le W3C pour interroger des sources de données RDF. Utiliser SPARQL implique d'annoter les corpus textuels contenant les réponses en RDF suivant une ontologie de référence. Pour des raisons d'espace, cet article ne présente pas la phase d'annotation du corpus en détail. Il est néanmoins important de noter que ce processus d'annotation utilise différentes méthodes d'extraction d'information et que les annotations RDF des documents sont sauvegardées dans des fichiers séparés tout en gardant le lien entre chaque phrase et ses annotations. Dans cette section nous décrivons notre algorithme de transformation des questions en langage naturel qui utilise une représentation en logique du premier ordre des informations extraites et des règles logiques afin d'aboutir à une ou plusieurs reformulations SPARQL de la question de l'utilisateur.

4.1. *Description de l'approche*

Nous proposons une méthode en six étapes qui consiste à :

1) Identifier le type de la question (p.ex. Yes/No, Définition, Factuelle, Liste) ;
2) Déterminer le(s) type(s) de réponse(s) attendue(s) pour les questions WH ;
3) Construire la forme affirmative et simplifiée de la question (nouvelle forme) ;
4) Reconnaître 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 4.5).

Le tableau 1 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), avec : TRA = Treatment, - patientHasType(children)	treats(TX,PB)

Tableau 1. *Analyse de questions médicales - Exemples (TRA : Type de réponse attendue, PB : Problème, PA : Patient, TX : Traitement)*

Nous identifions le type de la question (i.e. Yes/No, Définition, Factuelle, Liste ou Complexe) en utilisant un ensemble de patrons construits manuellement pour chaque type de question. Pour les questions factuelles et listes, nous déterminons aussi le type de réponse attendue. 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 (p.ex. Traitement, Médicament, Test médical).

Nous construisons une forme simplifiée et affirmative 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 les bruits 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 correcte pour la phase d'extraction de relation. Par contre, simplifier la question à *ANSWER for headache* permet d'obtenir une extraction de relation plus efficace en identifiant les relations entre seulement l'entité ANSWER (qui est un traitement) et l'entité *headache*.

4.2. *Reconnaissance des entités médicales*

La reconnaissance des entités médicales comporte deux étapes principales : la délimitation ou la détection des frontières des termes médicaux et (ii) la catégorisation ou la classification de ces entités à partir d'une liste de catégories médicales prédéfinies. Par exemple, dans la phrase : *High blood pressure may cause Kidney Disease*, la reconnaissance des entités médicales permet d'identifier *High blood pressure* et *Kidney Disease* comme étant deux problèmes médicaux. Nous travaillons sur sept catégories médicales qui ont été choisies après des analyses des différentes taxonomies de questions. Ces catégories sont : Problème médical, Traitement, Test médical, Signe ou Symptôme, Médicament, Nourriture et Patient. Notre approche pour la reconnaissance d'entités médicales combine deux méthodes différentes : à base de règles, MetaMap Plus, et statistique, BIO-CRF-H (Ben Abacha *et al.,* 2011b).

Les résultats de la reconnaissance des entités médicales sont représentés en logique d'ordre 1 avec les prédicats `category`, `name` et `position`. Par exemple, l'ensemble de prédicats E1 indique que le troisième token de la question (*aspirin*) est une entité médicale et que sa catégorie est TREATMENT : `E1 = {category(#ME1,TREATMENT) ∧ name(#ME1,aspirin) ∧ position(#ME1,3)}.`

4.3. *Extraction de relations sémantiques*

Cette étape est très importante pour une analyse de question efficace. Nous visons sept types de relations choisies après une analyse des taxonomies de questions médicales et aussi de questions médicales réelles : (i) *treats*, un traitement améliore ou traite un problème médical, (ii) *complicates*, un traitement empire un problème médical, (iii) *prevents*, un traitement prévient un problème médical, (iv) *causes*, un traitement cause un problème médical, (v) *diagnoses*, un test médical détecte, diagnostique ou évalue un problème, (vi) *DhD*, un médicament a une dose et (vii) *P_hSS*, un problème a un signe ou un symptôme. Pour extraire ces relations sémantiques, nous utilisons une combinaison de deux méthodes : une méthode à base de patrons et une méthode statistique qui utilise un classifieur SVM (Ben Abacha *et al.,* 2011a). Nous identifions aussi des relations spécifiques au patient : son sexe, son âge et sa catégorie d'âge (adulte, bébé, etc.).

Les résultats de l'extraction de relation sont représentés dans la logique d'ordre 1 avec plusieurs prédicats indiquant le nom de la relation (p.ex. treats, causes). Par exemple, l'expression E2 indique que trois entités médicales sont reliées par deux relations sémantiques *treats* et *patientHasProblem* : `E2 = { treats(#ME1,#ME2) ∧ patientHasProblem(#ME3,#ME2) }`. Le processus d'extraction de relation a des entrées explicites pour les questions de type Yes/No étant donné que toutes les entités médicales y sont complétement 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.

4.4. *Cas de multiples types pour la réponse attendue (TRA)*

Pour prendre en compte les cas de TRA multiples, nous construisons autant de questions que de réponses attendues. Ce processus consiste à :

1. Identifier le type de la question
2. Identifier les m types de réponses attendues (TRA) pour les questions WH
3. Construire la forme simplifiée et affirmative de la question (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

4.5. *Constuction de requête(s) SPARQL*

Une fois les entités médicales et les relations sémantiques extraites de la question en langue naturelle, la dernière étape consiste à construire une requête SPARQL équivalente. SPARQL définit 4 types de requêtes : SELECT, DESCRIBE, ASK et CONSTRUCT [8]. 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). 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 sont considérées comme des questions WH car elles recherchent des morceaux de texte contenant la définition.

Une requête SPARQL a deux composants principaux : un entête et un corps. L'entête indique la forme de la requête et d'autres déclarations (p.ex. préfixes, noms des graphes RDF à interroger, variables à retourner pour la forme SELECT). La construction du corps et de l'entête de la requête requiert des processus différents. Nous présentons dans la section suivante l'ontologie de référence qui est utilisée pour la construction des requêtes SPARQL.

4.5.1. *Ontologie de référence*

Nous définissons l'ontologie MESA (MEdical queStion Answering ontology) afin de représenter les concepts et les relations utilisés pour la construction des requêtes SPARQL. Cette ontologie est aussi utilisée pour annoter le corpus médical à partir duquel les réponses seront extraites. L'ontologie MESA, qui peut être alignée avec l'UMLS, décrit une partie du domaine : elle contient des concepts et des relations décrivant les fragments de texte qui seront retournés comme réponses finales de notre système de question-réponse (p.ex. *text1, text2*, cf. tableau 2). Elle se rapproche donc plus des ontologies d'application que des ontologies de domaine.

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

4.5.2. *Transformation des questions WH*

La forme SELECT des requêtes SPARQL est la plus appropriée pour la représentation des questions WH. L'entête 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 extraites de la question originale. Nous formalisons le processus de construction du corps de la requête dans une fonction en logique du premier ordre qui transforme 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 (p.ex. treats(ANSWER,flu), patientAgeGroup(patient,Infant)). Chaque prédicat binaire p(s,o) est transformé en un triplet RDF <s,p,o> où s est le sujet, p est la propriété RDF et o est l'objet. L'IRI[9] de la propriété p est obtenue en concaténant l'espace de nommage de l'ontologie et le nom du prédicat. 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:name` 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 ?arg2
?arg2 mesa:name ?name     FILTER(?name='Flu') }
```

Dans le cas où nous avons le TRA (p.ex. 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 l'IRI correspondant au concept indiquant ces catégories. Cette IRI est plus précisément obtenue en concaténant le nom de la catégorie médicale avec l'espace de nommage de l'ontologie MESA. L'exemple suivant représente la transformation de la catégorisation de la réponse attendue category(ANSWER,TEST) : `SELECT ?answer WHERE {?answer mesa:category mesa:TEST}`. La représentation finale de la question en langue naturelle consiste en une ou plusieurs requêtes SPARQL (dans le cas de multiples TRA, cf. section 4.4). 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 2 montre un exemple de transformation d'une question WH.

4.5.3. *Transformation des questions Yes/No*

La forme de requête SPARQL ASK est la plus appropriée pour les questions de type Yes/No. L'entête 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

9. Internationalized Resource Identifier : `http://www.ietf.org/rfc/rfc3987.txt`

Graphe sémantique simplifié

Requête SPARQL 1	**Requête SPARQL 2**

`Select ?answer1 ?text1 where {`
`?answer1 mesa:category`
`<Treatment>`
`?answer1 mesa:treats ?focus`
`?focus mesa:name` *`intestinal perforation`*
`?focus mesa:category <Problem>`
`?patient mesa:hasProblem ?focus`
`?patient mesa:category <Infant>`
 	
`OPTIONAL{`
`?text1 mesa:contains ?answer1`
`?text1 mesa:contains ?patient`
`?text1 mesa:contains ?focus } }`
	`Select ?answer2 ?text2 where {`
	`?answer2 mesa:category <Test>`
	
	`?answer2 mesa:diagnoses ?focus`
	`?focus mesa:name` *`intestinal perforation`*
	`?focus mesa:category <Problem>`
	`?patient mesa:hasProblem ?focus`
	`?patient mesa:category`
	`<Infant>`
	`OPTIONAL{`
	`?text2 mesa:contains ?answer2`
	`?text2 mesa:contains ?patient`
	`?text2 mesa:contains ?focus } }`

Tableau 2. *Requêtes SPARQL associées à la question :* What are the current treatment and monitoring recommendations for intestinal perforation in infants ?

similaire à la transformation des questions WH sauf qu'il n'y a pas de réponse ni 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 la transformation d'une question Yes/No en un patron de graphe RDF.

Can being on Metformin cause Headache ?

```
ASK{ ?e1 mesa:causes ?e2
?e1 mesa:name ?name1     ?e1 mesa:category <Drug>
?e2 mesa:name ?name2     ?e2 mesa:category <Problem>
FILTER(?name1='Metformin')
FILTER(?name2='Headache')}
```

5. Evaluation

Pour évaluer notre approche, nous utilisons 100 questions extraites du site du Journal of Family Practice (JFP). 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 3 présente quelques exemples de questions de ce corpus d'évaluation.

Type de 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 ?
plus d'un TRA	When should you suspect community-acquired MRSA ? How should you treat it ?

Tableau 3. *Quelques exemples de notre corpus de questions d'évaluation JFP (TRA = Type de la réponse attendue)*

Résultats. Nous avons testé nos deux méthodes de reconnaissance d'entités médicales sur notre corpus de questions d'évaluation et aussi, pour comparer les résultats, sur le corpus i2b2 de textes cliniques [10] construit dans le cadre du challenge i2b2/VA 2010. 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 en contient 44927). Le tableau 4 présente les résultats sans simplification de la question pour les trois catégories : Traitement, Problème et Test. Il 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 dû 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 : c'est un inconvénient classique des méthodes à apprentissage. Nous avons évalué de plus 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 de questions JFP. Nous avons aussi évalué l'extraction de relations et la construction des requêtes SPARQL. 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 langue naturelle. Cependant, dans les applications réelles, la performance de l'analyse de question et du SQR en général

10. `http://www.i2b2.org/NLP/Relations/`

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

Tableau 4. *Reconnaissance des entités médicales (catégories : Traitement, Problème et Test) : résultats sans la simplification des questions*

dépendra fortement des performances des systèmes d'extraction d'information utilisés.

Analyse des erreurs. Nous avons obtenu 62 requêtes correctes parmi les 100 questions de l'évaluation. Parmi les 38 requêtes incorrectes, 29 sont principalement dues à des erreurs au niveau de l'extraction de 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 (p.ex. *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 l'on détermine correctement les entités médicales et la relation sémantique (ici, *diagnoses*), la requête n'est pas correcte car le type de réponse attendue n'est pas encore traité par notre système. D'autres types de réponses attendues ne sont pas encore pris en compte non plus (p.ex. *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.

6. Conclusion

Dans cet article, nous avons abordé l'analyse automatique de questions médicales et présenté une approche originale pour la transformation de questions médicales en requêtes SPARQL. Ce travail a trois caractéristiques principales :

– L'approche proposée permet de traiter différents types de questions, parmi lesquels les questions avec plus d'un focus et/ou d'une réponse attendue.

– Elle permet une analyse profonde des questions en utilisant des méthodes fondées sur les connaissances du domaine (p.ex. UMLS) et des techniques de traitement automatique des langues (p.ex. patrons, apprentissage) pour l'extraction des entités médicales, des relations sémantiques et des informations contextuelles (p.ex. rel-

atives aux patients).

– Elle est basée sur les technologies du web sémantiques qui offrent plus d'expressivité, des langages formels standards et augmentent la portabilité des annotations relatives aux questions et aux corpus utilisés pour l'extraction des réponses.

Bien que notre premier objectif ait été la mise en place d'une méthode générique (par rapport aux différents types de questions médicales possibles), plus de traitements spécifiques sont encore requis dans ce cadre, en particulier pour : (i) les questions complexes (p.ex. 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 de tester la contribution de l'analyse syntaxique au niveau de 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). Notre objectif final est la mise en place d'un système de questions-réponses pour le domaine médical.

7. Bibliographie

Ben Abacha A., Zweigenbaum P., « A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts », *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, vol. 6608 of *Lecture Notes in Computer Science*, Tokyo, Japan, p. 139-150, 2011a.

Ben Abacha A., Zweigenbaum P., « Medical Entity Recognition : A Comparison of Semantic and Statistical Methods », *Proceedings of BioNLP 2011 Workshop*, Association for Computational Linguistics, Portland, Oregon, USA, p. 56-64, 2011b.

Cimiano P., Haase P., Heizmann J., Mantel M., Studer R., « Towards portable natural language interfaces to knowledge bases : The Case of the ORAKEL system », *Data Knowledge Engineering (DKE), 65(2)*, p. 325-354, 2008.

Ely J. W., Osheroff J. A., Gorman P. N., Ebell M. H., Chambliss M. L., Pifer E. A., Stavri P. Z., « A taxonomy of generic clinical questions : classification study », *British Medical Journal*, vol. 321, p. 429-432, 2000.

Fan S., Wang X., Wang X., Zhang Y., « A new question analysis approach for community question answering system », *International Journal on Asian Language Processing*, vol. 19, n° 3, p. 95-108, 2009.

Popescu A., Etzioni O., Kautz H., « Towards a theory of natural language interfaces to databases », *Proceedings of the International Conference on Intelligent User Interfaces (IUI'03)*, p. 149-157, 2003.

Verberne S., « Developing an approach for why-question answering », *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics : Student Research Workshop*, EACL'06, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 39-46, 2006.