Annotation et interrogation sémantiques de textes médicaux

Asma Ben Abacha¹ & Pierre Zweigenbaum¹

LIMSI - CNRS, B.P. 133 91403 Orsay CEDEX France {asma.benabacha, pz}@limsi.fr

Résumé: Ce travail s'inscrit dans le cadre de la réalisation d'un système de question réponse pour le domaine médical. Dans ce papier, nous présentons notre approche d'annotation sémantique de textes médicaux. Cette approche utilise des ressources sémantiques du domaine ainsi que des méthodes linguistiques pour l'annotation (i) des entités médicales et (ii) des relations sémantiques qui les relient. Ces informations sont sauvegardées sous format RDF et interrogées ultérieurement pour répondre aux requêtes utilisateur. Dans la dernière partie de cet article, nous proposons une approche basée sur l'annotation et l'interrogation sémantiques pour la mise en place d'un système de question réponse.

Mots-clés: Annotation sémantique, RDF, domaine médical, UMLS, recherche sémantique, SPARQL.

1 Introduction

Le domaine médical dispose de plusieurs ressources sémantiques structurées (e.g. le Métathésaurus et le réseau sémantique de l'UMLS). Ces ressources permettent de résoudre certains problèmes comme la grande variabilité des termes médicaux et de structurer et enrichir le contenu des textes médicaux par des annotations sémantiques. La tâche d'annotation sémantique améliore toute exploitation ultérieure des textes annotés (e.g. pour la recherche et l'extraction d'information). Dans cette optique plusieurs travaux ont été réalisés. (Khelif & Dieng-Kuntz, 2004) ont utilisé le réseau sémantique de l'UMLS comme une ontologie du domaine biomédical et les termes du Métathésaurus comme instances possibles de concepts biomédicaux. Ils ont utilisé GATE (Cunningham et al., 2002) et le service Web UMLSKS pour annoter les documents par les différents termes de l'UMLS. Ils ont aussi essayé d'extraire les relations pouvant exister entre ces termes, et particulièrement la relation « play role » en utilisant l'analyseur syntaxique Syntex (Bourigault & Fabre, 2000) et une grammaire développée en Jape. (Jonquet et al., 2009) ont proposé un service web d'annotation appelé Open Biomedical Annotator (OBA) qui permet aux utilisateurs d'exploiter les ontologies biomédicales (e.g. Gene Ontology, SNOMED-CT) pour annoter leurs données automatiquement. L'annotateur OBA a été utilisé pour créer un système d'indexation de ressources biomédicales appelé Open Biomedical Ressources (OBR). Cet index permet (i) de rechercher des

données biomédicales annotées par un concept donné et (ii) de retrouver les annotations correspondant à une ressource ayant un identifiant donné (e.g. PubMedID, GEO datasetID). (Jilani & Amardeilh, 2009) ont proposé une méthode pour construire et peupler une ontologie automatiquement en utilisant des patrons lexico-syntaxiques. Elles ont construit des patrons lexicaux pour déterminer les liens entre une mutation, un gène, un miARN (micro Acides RiboNucléiques) et/ou une maladie en recherchant leurs co-occurrences dans les phrases du corpus. Leur méthode donne une précision de 72% ainsi qu'un rappel de 50%.

Dans cet article nous proposons une approche qui exploite l'UMLS pour annoter (i) les entités médicales présentes dans des textes en anglais (e.g. Maladie, Médicament) et (ii) les relations sémantiques qui les relient (e.g. traiter, prévenir). Ces annotations sont sauvegardées sous format RDF et utilisées ultérieurement pour répondre aux requêtes d'un utilisateur (e.g. «?x Traite Varicelle »). Nous ciblons particulièrement des résumés et des articles médicaux de Medline. Nous commençons par présenter notre méthode d'annotation sémantique des textes médicaux ainsi que les résultats expérimentaux obtenus (section 2). Après la production de ces annotations sémantiques, nous nous intéressons dans une deuxième partie à leur exploitation et nous présentons notre méthode d'interrogation sémantique (section 3). Dans une deuxième partie nous proposons une architecture sémantique pour la mise en place d'un système de question réponse (section 4). Enfin nous concluons et donnons quelques perspectives (section 5).

2 Annotation sémantique

L'annotation sémantique de textes au sens large comprend l'association de sens à l'information contenue dans les documents par des annotations/méta-données, insérées dans le document lui-même ou dans des supports externes. Dans le cadre du Web sémantique, les besoins d'interopérabilité et de passage à l'échelle rendent nécessaire l'automatisation du processus d'annotation et l'utilisation de formalismes standardisés. Dans cette perspective nous utilisons les langages RDF(S) et OWL pour représenter les connaissances du domaine sous forme d'ontologie (conceptualisation partagée d'un domaine). Plus particulièrement, nous utilisons le réseau sémantique de l'UMLS sous format RDFS comme ontologie de référence.

Le processus d'annotation consiste ensuite en trois étapes que nous détaillons dans cette section : (i) reconnaissance des entités médicales (e.g. maladie, examen médical) présentes dans les phrases du corpus (ii) identification des relations sémantiques qui relient les entités reconnues (e.g. traite, complique, cause) et (iii) transformation des informations extraites en format RDF. Les deux premières étapes seront décrites plus en détail dans un prochain papier.

2.1 Reconnaissance des entités médicales

Plusieurs approches et outils se sont intéressés à l'extraction des entités médicales. Un des outils les plus largement utilisés pour cette tâche est MetaMap. MetaMap identifie les entités et leurs catégories en utilisant le Métathesaurus 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.

Cependant l'étude de l'utilisation simple de MetaMap sur un de nos corpus a révélé qu'il présente certains problèmes résiduels. Il s'agit le plus souvent d'erreurs aux niveaux de la segmentation du corpus en phrases et en syntagmes nominaux, mais aussi au niveau de la catégorisation des termes médicaux trouvés. Dans ce papier nous proposons une approche utilisant des outils efficaces connus dans le domaine du TAL en amont de l'emploi de MetaMap et une méthode de filtrage en aval des résultats obtenus. Cette approche est résumée dans la figure 1.

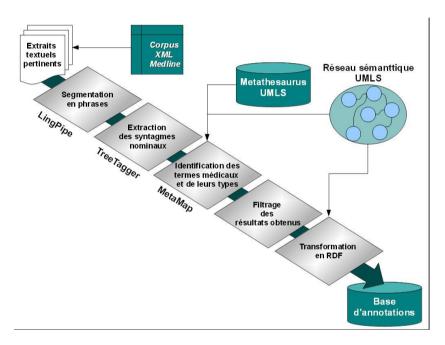


FIG. 1 – Processus de reconnaissance des entités médicales

Nous prenons en entrée des corpus médicaux textuels que nous segmentons en phrases avec les modèles LingPipe spécialisés pour les ressources Medline. Dans une seconde étape, chaque phrase est segmentée en syntagmes nominaux en utilisant TreeTaggerchunker, puis ces syntagmes sont soumis à MetaMap afin d'identifier les entités médicales qu'ils contiennent. Enfin, nous filtrons les résultats obtenus grâce à (i) une liste des erreurs les plus fréquentes/remarquables et (ii) la limitation des types sémantiques utilisés par MetaMap afin d'éviter certains concepts généraux (e.g. Quantitative Concept, Functional Concept, Qualitative Concept).

Nous avons construit un corpus d'évaluation de 20 articles scientifiques anglais variés extraits de PubMedCentral. Nous avons ensuite annoté manuellement les entités médicales correspondant à 16 types sémantiques donnés, sous catégories de « traitant » et « traité », (cf. tableau 1) et les relations sémantiques qui les lient.

Types sémantiques	Antibiotic, Drug Delivery Device, Medical Device, Phar-	
« Traitants »	macologic Substance, Therapeutic or Preventive Procedure	
Types sémantiques	Acquired Abnormality, Anatomical Abnormality, Cell o	
« Traités »	Molecular Dysfunction, Congenital Abnormality, Disease	
	or Syndrome, Experimental Model of Disease, Injury or	
	Poisoning, Mental or Behavioral Dysfunction, Neoplastic	
	Process, Pathologic Function, Sign or Symptom	

TAB. 1 – Types sémantiques sources et cibles de la relation « Traitement »

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. Cette 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 notre 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 :

$$Precision = \frac{C + 0.5 \times B}{Ref} \tag{1}$$

- 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 compare la précision obtenue avec notre méthode (LTS+MetaMap) 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). À titre d'exemple, les résultats de MetaMap comporte 52.27% d'erreurs de localisation pour les entités de type *Disease Or Syndrome* alors que notre méthode n'a conduit qu'à 26.48% d'erreurs de localisation.

	MetaMap			LTS + MetaMap		
	T	В	P	T	В	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%

TAB. 2 – Résultats de la reconnaissance de certaines entités médicales

2.2 Extraction des relations sémantiques

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 déterminer quels types de relations sont plausibles entre deux entités données. 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 dans la phase 1 nous utilisons un ensemble de patrons linguistiques que nous modélisons dans une ontologie liée au réseau sémantique de l'UMLS (cf. figure 2).

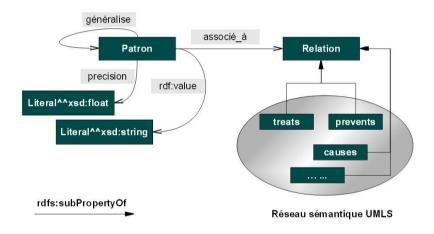


FIG. 2 – Ontologie modélisant les patrons et les relations sémantiques

Nous associons à chaque relation (ciblée) du réseau sémantique de l'UMLS un ensemble de patrons linguistiques ayant une valeur littérale correspondant à une expression régulière donnée et sauvegardée par la propriété *rdf:value* (cf. Tableau 3). Les patrons sont construits manuellement à partir d'un corpus sélectionné sémantiquement (obtenu en interrogeant PubMed Central avec des requêtes MeSH).

Relation	Nombre de patrons*	Exemples de patrons
causes	28	E1 may trigger E2
diagnoses	12	E1 is the best test for (the diagnoses of)? E2
treats	46	E1 was found to reduce E2
prevents	13	E1 for prophylaxis against E2

TAB. 3 – Exemples de patrons de relations

^{*:} nombres en cours d'évolution

Aussi, afin d'associer une mesure de confiance aux relations qui seront annotées ultérieurement, les patrons sont classés suivant leur précision avec la propriété du même nom dans l'ontologie. La précision d'un patron est calculée automatiquement en décrémentant les valeurs de précision suivant les relations de généralisation.

Par exemple, si nous avons les patrons suivants :

- <patron1, précision, P>
- <patron2, généralise, patron1>
- <patron3, généralise, patron2>

alors les annotations suivantes sont générées automatiquement :

- <patron2, précision, P/C>
- <patron3, précision, P/C/C>, avec C un coefficient entier.

Dans le cas où un patron généralise plusieurs patrons différents nous considérons la moyenne de précision de toutes ses dérivations directes divisée par C.

Une relation est identifiée dans une phrase si (i) elle est définie dans le réseau sémantique 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. Les patrons sont appliqués du plus spécifique au plus général. Le tableau 4 présente quelques exemples de patrons.

Patron	Relation	Précision	Exemple
TX for PB	Traitement	0,5	Intralesional corticosteroid therapy
			for primary cutaneous B cell lym-
			phoma.
TX for the treatment	Traitement	0.75	Cognitive-behavioral group therapy
of PB			is an effective intervention for the
			treatment of geriatric depression.
TX for preventing PB	Prévention	0.75	Influenza vaccines for preventing
			coronary heart disease.

TAB. 4 – Exemples de Patrons de relations sémantiques

Nous associons un poids à la relation extraite qui prend en compte la précision du patron employé et le nombre de syntagmes nominaux entre les deux entités. L'idée derrière le second point est que la relation est considérée plus forte s'il n'y a que des verbes et/ou prépositions entre les deux entités dans la phrase que s'il y a plusieurs autres syntagmes nominaux entre les syntagmes contenant les entités médicales en question.

Le poids d'une relation R extraite moyennant un patron P à partir d'une phrase H, entre deux entités médicales E1, E2 contenues dans deux syntagmes S1 et S2 est donc :

$$W(R) = \frac{dsp(P)}{exp(nsynt(H, S1, S2))}$$
(2)

- dsp(P): degré de spécialisation du patron P
- nsynt(H,S1,S2): une fonction qui retourne le nombre de syntagmes nominaux entre S1 et S2 dans la phrase H.

Une fois que les phrases contenant la relation ciblée sont extraites, nous utilisons des filtres pour éliminer les phrases négatives (e.g. *In general, TPE alone is not suggested*

for the treatment of progressive MS cases.) et les phrases interrogatives (e.g. Can statins treat infection?). La méthode complète d'extraction de relations sémantiques est décrite dans la figure 3.

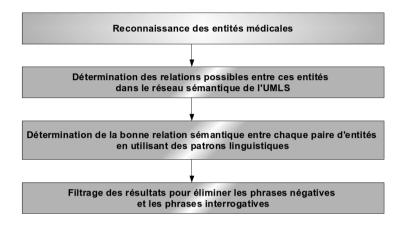


FIG. 3 – Processus d'extraction de relations sémantiques

Notre méthode d'extraction de relations sémantiques a été évaluée sur notre corpus d'évaluation de 20 articles extraits de PubMedCentral. Les résultats de cette évaluation sont présentés dans le tableau 5.

Relation : Traitement			
Rappel	Précision	F-Mesure	
60,46 %	75,72 %	67,23 %	

TAB. 5 – Résultats d'extraction de relations sémantiques de type « Traitement »

2.3 Génération des triplets RDF

L'objectif de base de cette étape est d'écrire les annotations sous forme de triplets RDF conformes à l'ontologie représentant le réseau sémantique de l'UMLS. Nous annotons aussi les poids associés aux relations extraites en utilisant le principe de réification qui permet de considérer un triplet comme un élément dans un autre triplet. La figure 4 présente un exemple d'annotation.

3 Interrogation sémantique

Dans cette section, nous présentons notre méthode d'interrogation sémantique. Les triplets RDF générés dans la phase d'annotation sont interrogés avec des requêtes SPARQL formulées suivant les concepts et les relations du réseau sémantique de l'UMLS.

```
<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:objectrdf:rsesource="#entity21"/>
      <onto:degree of confidence >1</onto:degree of confidence>
</rdf:Description>
```

FIG. 4 – Exemple d'annotation RDF

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 5).

Par exemple, l'utilisateur peut chercher le traitement (inconnu) d'une maladie dont il écrit le nom (e.g. ?X traite « Erysipelas ») ou juste une validation d'un fait donné (e.g. la pénicilline est-elle bien un traitement pour la « S. aureus infection »). Les termes entrés par l'utilisateur sont analysés par MetaMap afin de connaître les types sémantiques concernés par la requête. Dans le cas où aucun type connu ne correspond aux termes recherchés par l'utilisateur, les solutions de remplacement les plus proches syntaxiquement sont proposés à l'utilisateur.

Une dernière étape dans le processus de transformation de la requête est de faire apparaître les poids des triplets recherchés/employés et de trier les réponses suivant ces valeurs.

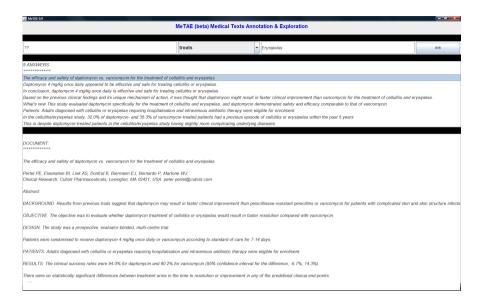


FIG. 5 – Interface d'interrogation

4 Vers un système de question réponse

L'architecture classique d'un système de question réponse comporte 3 modules : analyse de la question, recherche de documents et extraction de la réponse. L'analyse de la question s'intéresse à déduire à partir de la question toutes les informations utiles pour les modules qui suivent comme les mots clés de la question et le type de la réponse attendue. Le deuxième module fouille une collection de documents pour extraire les documents ou les passages susceptibles de contenir la réponse. Enfin, le troisième module utilise les informations déduites de l'analyse de la question pour extraire, depuis les documents retrouvés, la ou les réponses.

En se basant sur notre approche d'annotation et une partie de notre méthode d'interrogation nous proposons une architecture différente où l'accès aux extraits (phrases) pertinents se fait directement en interrogeant la base d'annotation. Un module d'analyse de la question est utilisé en amont pour transformer la question de l'utilisateur en requête structurée écrite en SPARQL (cf. figure 6).

5 Conclusion

Nous avons présenté dans cet article notre méthode d'annotation sémantique de textes médicaux. Cette approche permet d'annoter automatiquement les entités médicales présentes dans les textes et les relations sémantiques qui les relient en combinant l'utilisation d'outils et ressources sémantiques du domaine (UMLS) et des techniques de TAL pour l'extraction d'information. Ces annotations sont exploitées avec un moteur de re-

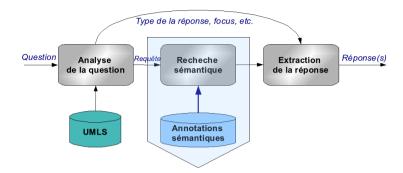


FIG. 6 – Architecture proposée pour un SQR

cherche sémantique pour répondre aux requêtes utilisateur. L'approche présentée dans cet article rentre dans le cadre du développement d'un système de question réponse pour le domaine médical et constitue une grande partie de ce qui est nécessaire pour la réalisation d'un tel système.

Comme perspectives à court terme, nous envisageons d'expérimenter notre approche d'interrogation sur un corpus extrait de Medline et d'améliorer le module d'annotation en traitant d'autres types de relations sémantiques avant de nous intéresser à la mise au point d'un système de question réponse pour le domaine médical. Nous envisageons aussi d'explorer la dimension translingue en répondant à des questions écrites en français à partir de corpus anglais.

Références

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. In *Cahiers de grammaire*, 25, 131-151.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. & TABLAN V. (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. In *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, p. 168-175, Philadelphia.

JILANI I. & AMARDEILH F. (2009). Enrichissement automatique d'une base de connaissances biologiques à l'aide des outils du web sémantique. In *IC2009 - 20èmes journées francophones d'Ingénierie des Connaissances, Hammamet : Tunisie.*

JONQUET C., SHAH N. & MUSEN M. A. (2009). Un service web pour l'annotation sémantique de données biomédicales avec des ontologies. In *13èmes Journées Francophones d'Informatique Médicale*, *JFIM2009*.

KHELIF K. & DIENG-KUNTZ R. (2004). Web sémantique et mémoire d'expériences sur les biopuces. In Web Sémantique Médical (WSM'2004), Rouen.