TALN 2010, Montréal, 19–23 juillet 2010

MeTAE : Plate-forme d'annotation automatique et d'exploration sémantiques pour le domaine médical

Asma Ben Abacha[1] Pierre Zweigenbaum[1]
(1) LIMSI - CNRS, B.P. 133 91403 ORSAY CEDEX FRANCE
{asma.benabacha, pz}@limsi.fr

Résumé. Nous présentons une plate-forme d'annotation sémantique et d'exploration de textes médicaux, appelée « MeTAE ». Le processus d'annotation automatique comporte une première étape de reconnaissance des entités médicales présentes dans les textes suivie d'une étape d'identification des relations sémantiques qui les relient. Cette identification se fonde sur des patrons linguistiques construits manuellement pour chaque type de relation. MeTAE génère des annotations RDF à partir des informations extraites et offre une interface d'exploration des textes annotés avec des requêtes sous forme de formulaire. La plate-forme peut être utilisée pour analyser sémantiquement les textes médicaux ou interroger la base d'annotation disponible pour avoir une/des réponses à une requête donnée (e.g. « ?X prévient maladie d'Alzheimer », équivalent à la question « comment prévenir la maladie d'Alzheimer ? »). Cette application peut être la base d'un système de questions-réponses pour le domaine médical.

Abstract. This paper presents MeTAE, a platform for semantic annotation and exploration of medical texts. The annotation process encompasses medical entity recognition and semantic relationship identification between the retrieved entities. This identification is based on linguistic patterns constructed manually for each type of relation. MeTAE generates RDF annotations from the extracted information and allows semantic exploration of the annotated texts through a form-based interface. The platform can be used to semantically analyze medical texts or to explore the available annotation base through structured queries (e.g. "?X Prevents Alzheimer's disease" for its natural-language equivalent: "how to prevent Alzheimer's disease?"). MeTAE can be a basis for a medical question-answering system.

Mots-clés : Annotation sémantique, interrogation sémantique, domaine médical.

Keywords: Semantic annotation, semantic querying, medical domain.

1 Introduction

Avec la grande quantité d'informations médicales numérisées, le domaine médical a besoin d'outils d'aide à la recherche d'information plus précis que les moteurs de recherche classiques. La recherche d'information précise nécessite une analyse sémantique des documents desquels l'information va être extraite. L'analyse sémantique des textes médicaux passe généralement par une étape de reconnaissance des entités médicales (e.g. cancer, aspirin, blood test). Pour une analyse plus profonde, certains travaux s'intéressent aux relations sémantiques reliant les entités reconnues (e.g. causalité, traitement, diagnostic). Dans cette démonstration, nous présentons une plate-forme d'annotation automatique de textes médicaux en anglais

qui extrait les entités médicales et les relations sémantiques les reliant. Cette plate-forme offre aussi une interface d'exploration des annotations faites pour répondre aux requêtes utilisateur[1].

2 Processus automatique d'annotation sémantique

Notre méthode d'annotation comporte une succession de 3 étapes : (i) la reconnaissances des entités médicales, (ii) l'identification des relations sémantiques entre les entités reconnues et (iii) la transformation des annotations en RDF[2] (Resource Description Framework).

La reconnaissance des entités médicales se fait grâce à :
– un prétraitement des textes : segmentation en phrases (outil LingPipe spécialisé pour les ressources Medline) et extraction de syntagmes nominaux (TreeTagger-chunker) ;
– une identification des termes médicaux (e.g. Doxorubicin, Myocardial scintigraphy) et de leurs types sémantiques (Antibiotic, Diagnostic Procedure) grâce à l'outil MetaMap (Aronson, 2001) ;
– un filtrage des résultats de MetaMap (filtre sur les erreurs et sur certains types sémantiques).

L'extraction des relations sémantiques se fonde sur une approche linguistique utilisant des patrons (à l'instar de (Hearst, 1992)). Ces patrons permettent de typer la relation entre deux entités médicales en prenant comme référence le réseau sémantique de l'UMLS[3] (Unified Medical Language System). Pour chaque relation, un ensemble de patrons linguistiques a été construit manuellement (cf. tableau 1). Afin d'améliorer l'exploration ultérieure des informations extraites nous structurons ces patrons de façon hiérarchisée dans une ontologie. Les relations de spécialisation/généralisation établies entre patrons permettent de calculer un degré de spécificité qui pourra être utilisé pour associer des valeurs de confiance aux relations médicales extraites.

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. 1 – Exemples de patrons de relations (* : nombres en cours d'évolution)

Notre méthode d'annotation est détaillée et évaluée pour extraire les relations de type « treats » dans (Abacha & Zweigenbaum, 2010). Actuellement notre plate-forme permet d'annoter les 6 relations sémantiques présentées dans la figure 1 et les termes médicaux associés aux 5 catégories sémantiques (Therapeutic Procedure, Drug, Medical Problem, Medical Test et Sign or Symptom). Chaque catégorie regroupe un ensemble de types sémantiques du réseau sémantique de l'UMLS. Le Tableau 2 contient quelques exemples de types sémantiques associés à certaines catégories présentées dans l'ontologie.

[1]Ce travail a été partiellement financé par OSEO dans le cadre du programme Quaero.
[2]http ://www.w3.org/TR/rdf-primer/
[3]http ://www.nlm.nih.gov/research/umls/

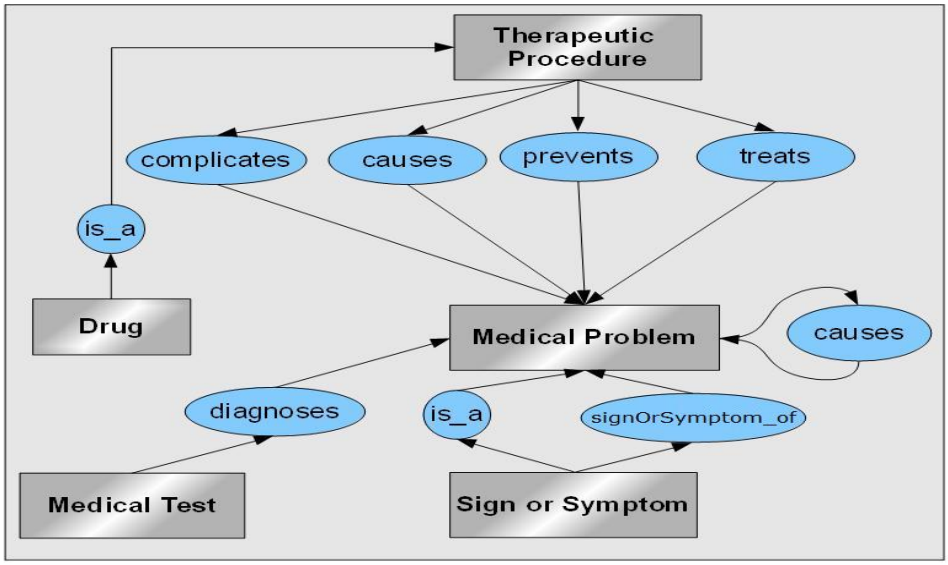

FIG. 1 – Ontologie des relations ciblées

Catégorie	Types sémantiques associés
Medical Test	Diagnostic Procedure, Laboratory Procedure, Pharmacologic Substance, etc.
Medical Problem	Anatomical Abnormality, Injury or Poisoning, Disease or Syndrome, etc.
Therapeutic Procedure	Antibiotic, Pharmacologic Substance, Therapeutic or Preventive Procedure, etc.

TAB. 2 – Exemples de catégories et types sémantiques associés

3 Déroulement de la démonstration

Notre démonstration consiste à :

1. Ajouter des documents/textes médicaux à l'application et visualiser les informations extraites par le système sous forme d'annotations sémantiques. Les entrées consisteront en des fichiers au format Medline ou des textes médicaux.

2. Explorer les textes médicaux via leurs annotations sémantiques. L'utilisateur pourra formuler des requêtes de la forme « ?X relation ?Y » en tapant librement les entités ?X et ?Y et en choisissant la relation dans une liste. Le corpus interrogé sera un corpus de résumés d'articles scientifiques extraits à partir de PubMed et préalablement annotés via la plate-forme.

La figure 2 présente un aperçu de la phase d'exploration de MeTAE. Cet aperçu montre l'exemple de l'exécution de la requête « ?X treats erysipelas » avec dans le premier volet les réponses (phrases) extraites comme étant pertinentes à cette requête. Le deuxième volet contient le document correspondant à la réponse sélectionnée par l'utilisateur.

Environnement requis : Système d'exploitation Linux, JRE 6.0.

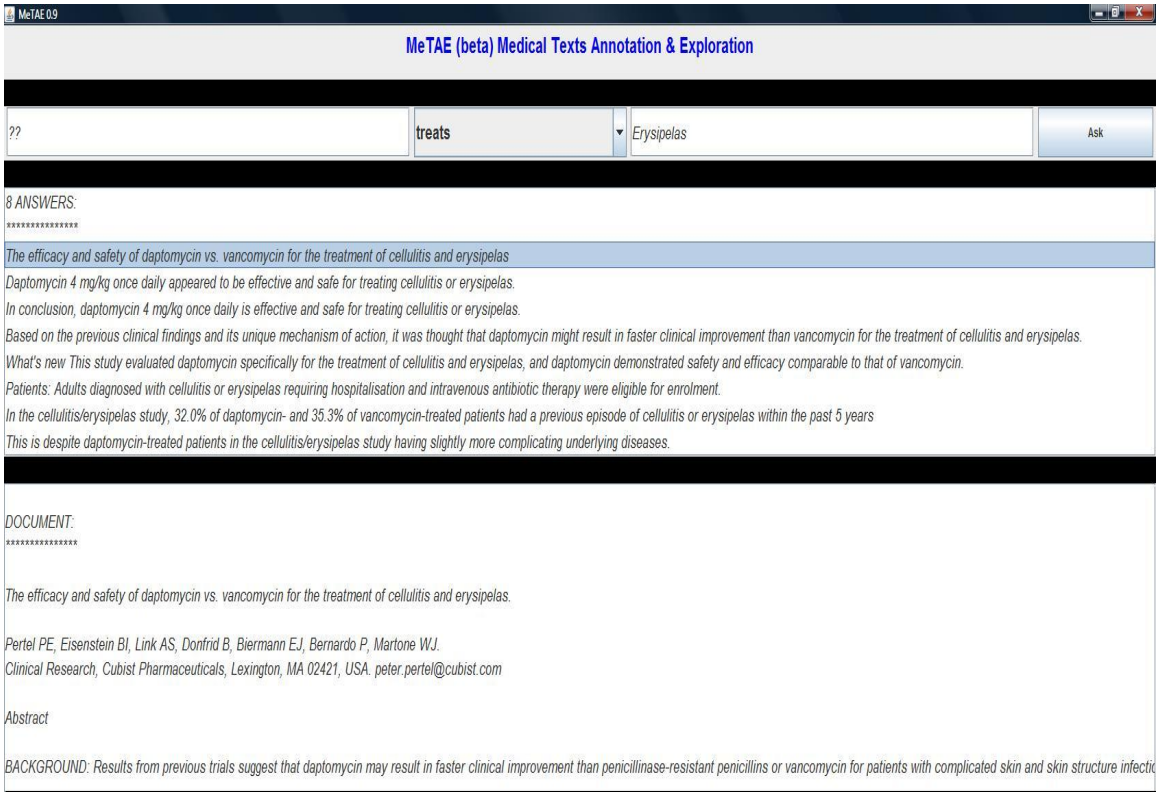

FIG. 2 – Interface d'exploration - MeTAE

4 Conclusion et Perspectives

Nous avons présenté MeTAE, une plate-forme d'annotation et d'exploration de textes médicaux. L'annotation comporte la reconnaissance des entités médicales en utilisant entre autre l'UMLS et l'identification des relations reliant ces termes en appliquant des patrons linguistiques. Une version de la plate-forme est accessible en ligne à `http://www.limsi.fr/Individu/abacha/metae.html`.

Comme perspectives, nous envisageons d'améliorer l'annotation avec l'extraction d'information sur les patients, les médicaments (e.g. dosage), etc. Une seconde et dernière étape pour la mise en œuvre d'un système de questions-réponses médical sera le développement d'un module d'analyse des questions posées en language naturel. Dans le cadre d'un système translingue, nous envisageons aussi de répondre à des questions écrites en français à partir de textes médicaux en anglais.

Références

ABACHA A. B. & ZWEIGENBAUM P. (2010). Annotation et interrogation sémantiques de textes médicaux. In *Atelier Web Sémantique Médical, IC*.

ARONSON A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. In *Proc. AMIA Symp*, p. 17–21.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, p. 539–545.