Annotations sémantiques pour le domaine Biopuces

Khaled Khelif, Rose Dieng-Kuntz

Annotations sémantiques pour le domaine des biopuces

Khaled KHELIF, Rose DIENG-KUNTZ

INRIA, UR Sophia Antipolis projet ACACIA
2004, route des lucioles BP 93, 06902 Sophia Antipolis Cedex, France
(khaled.khelif@sophia.inria.fr, rose.dieng@sophia.inria.fr)

Résumé : Après avoir souligné l'intérêt du web sémantique dans le domaine biomédical et de l'apport des annotations sémantiques dans la recherche d'informations, nous présentons une méthode pour la génération semi-automatique des annotations sémantiques décrivant des articles dans le domaine des biopuces et ce en se basant sur les techniques d'extraction d'informations.

Mots-clés : TAL, ontologie, annotation sémantique, web sémantique, biopuces

1 Introduction

Les documents publiés sur le web constituent une source très importante de connaissances. Ces connaissances sont, entre autres, indispensables pour la vérification, la validation et l'enrichissement d'un travail de recherche. C'est le cas pour la recherche dans le domaine de la biologie moléculaire et plus particulièrement dans le domaine des expériences de puces à ADN.

Ces expériences dites expériences biopuces et visant à trouver de nouvelles fonctionnalités des gènes ainsi que les interactions possibles entre eux, présentent des difficultés pour le biologiste notamment lors de la validation et de l'interprétation des résultats obtenus. Il doit alors faire des recherches dans des bases de documents ou de données génétiques à l'aide de mots clés correspondant aux gènes et au phénomène biologique étudiés dans le but de trouver des résultats qui argumentent, confirment ou infirment les siens. Le biologiste doit alors analyser les documents trouvés par le moteur de recherche classique afin d'identifier les connaissances pertinentes.

L'une des ambitions du web sémantique est de faciliter la tâche de recherche en essayant d'automatiser le traitement de l'information sur le web. Cela peut être effectué en associant à chaque document une annotation dite sémantique basée sur une ontologie décrivant le domaine. Cette annotation décrira alors le contenu sémantique des documents. Dans le cas d'expériences biopuces, il s'agira du type de gènes intervenant dans l'expérience décrite par l'article et les interactions pouvant exister entre eux, avec des composants cellulaires ou des processus biologiques.

Malgré ses avantages, la création d'une annotation sémantique est un processus difficile et coûteux (temps, personnes…) pour les biologistes. L'extraction

automatique d'information peut donc être une alternative pour la génération de ces annotations.

Dans le cadre du projet MEAT nous collaborons avec les biologistes de l'IPMC (Institut de Pharmacologie Moléculaire et Cellulaire) travaillant sur les expériences biopuces, et notre objectif est de leur permettre de construire et exploiter une mémoire des expériences sur les puces à ADN. Notre approche repose sur la génération semi-automatique d'annotations sémantiques sur les articles du domaine des biopuces, ces articles peuvent provenir de sources internes telles que les bases documentaires propres à chaque biologiste comme ils peuvent provenir d'une source externe telles que les bases documentaires en ligne (exp. Medline). Ainsi, nous avons mis au point une méthode qui, à partir d'un texte brut écrit par un biologiste, permet de générer une annotation sémantique structurée, basée sur une ontologie du domaine, et décrivant le contenu sémantique de ce texte.

Après un état de l'art, l'article détaillera les différentes étapes de notre méthode avant de conclure par une discussion sur les résultats obtenus lors de nos tests.

2 Etat de l'art

2.1 Le web sémantique et le domaine biomédical

Le W3C a proposé une extension du web, appelée web sémantique, (Berners-Lee et al, 2001), qui a pour but de structurer le contenu du web et permettre ainsi aux machines de traiter ce contenu automatiquement et de raisonner sur les connaissances représentées dans les pages web pour le compte des humains.

Une des couches les plus importantes dans l'architecture du web sémantique est l'ontologie. Définie par (Gruber, 1993) comme une spécification explicite d'une conceptualisation, celle-ci est constituée d'une hiérarchie de concepts décrivant un domaine particulier et d'une hiérarchie de relations pouvant exister entre les concepts.

En ce qui concerne le domaine biomédical, plusieurs travaux ont été menés afin de construire des ontologies le décrivant. Nous pouvons citer :

- La Gene Ontology : Utilisée pour annoter les données biologiques telles que les protéines, les gènes et les séquences, elle comprend plus de 13 000 concepts du domaine de la génétique, une relation de partonomie (part of) et une relation de spécialisation (is_a).
- Galen (General Architecture for Language and Nomenclatures): ensemble de 4 000 concepts multilingues décrivant les procédures chirurgicale.
- Menelas : cette ontologie dans le domaine des pathologies coronariennes. comprend 1 800 concepts et 300 relations.

Mais aucune de ces ontologies du domaine biomédical ne couvre tous les domaines traités par les expériences biopuces. Nous nous sommes donc penchés sur UMLS : ce projet, élaboré par la NLM (National Library of Medecine de Bethesda) déjà à l'origine de MeSH et de Medline, propose depuis 1986 de mettre au point un

langage médical unifié. L'UMLS sert à traduire et à conceptualiser une source d'information médicale pour la rendre accessible à une interrogation. Pour ce faire, ce langage se base sur : (1) un métathesaurus qui énumère tout le vocabulaire médical existant et comprend 900 551 concepts et 2.5 millions termes différents ; (2) un réseau sémantique constitué d'une hiérarchie de 134 types sémantiques et d'une hiérarchie de 54 relations entre ces types ; il représente une classification de tous les concepts représentés dans le métathesaurus ainsi que les relations qui peuvent exister entre eux (Schulze-Kremer et al, 2003).

Par analogie, nous considérons, le réseau sémantique de UMLS comme une ontologie : la hiérarchie des types sémantiques est la hiérarchie de concepts et les termes du métathesaurus sont des instances de ces concepts.

2.2 Les annotations basées sur une ontologie

Une annotation, ou métadonnée, désigne une donnée fournissant des informations sur une ressource. En terme de documentation, c'est une information secondaire apposée à une ressource primaire qui est le document. Une annotation « sémantique » fournit, en plus d'informations simples telles que le titre et les auteurs, une description plus précise des connaissances contenues dans le document et de leur sémantique par rapport au domaine. Une annotation sémantique doit avoir une structure prédéfinie et propre au domaine étudié car la génération libre d'une annotation peut créer des problèmes d'ambiguïté dans la définition des termes, de redondance et de réutilisation. De plus, l'utilisation de structures différentes rend l'exploitation de ces annotations par un moteur de recherche sémantique très difficile.

Se baser sur un ensemble déjà défini de concepts, de propriétés et de relations (*i.e.* se baser sur une ontologie) rend l'annotation sémantique plus intéressante d'un point de vue de la structure et du contenu. Ce guidage permet à celui qui annote le document de ne pas se confronter à des problèmes d'ambiguïté et de savoir quelles connaissances contenues dans le document il doit annoter.

Dans (Staab et al, 2001), les auteurs soulignent l'intérêt de l'utilisation d'une ontologie pour la création des annotations sémantiques et (Soo et al, 2003) propose une comparaison des résultats de recherche de deux systèmes, l'un basé sur des annotation générées librement et l'autre sur des annotations basées sur une ontologie.

Cependant, malgré son importance pour la gestion des connaissances d'un domaine, le processus d'annotations reste très lourd et nécessite plusieurs ressources, ce qui a motivé plusieurs travaux sur l'automatisation de cette tâche et ce, en se basant sur des techniques d'extraction d'informations.

2.3 L'extraction d'information

Le texte a toujours été considéré comme le moyen le plus sûr pour stocker et pérenniser l'information ou la connaissance, mais vu sa prolifération incessante dans tous les domaines, se pose le nouveau problème d'extraire ces informations afin de les exploiter. L'extraction d'information consiste à analyser le texte afin d'identifier,

de structurer et d'extraire des données pertinentes dans un domaine particulier. C'est un travail fastidieux effectué le plus souvent à la main.

Par ailleurs, depuis quelques années sont apparus plusieurs systèmes essayant d'automatiser cette tâche, et souvent basés sur des outils de TAL (Traitement Automatique de la Langue) de plus en plus efficaces et accompagnés par un nombre important de ressources linguistiques en ligne (dictionnaires, thésaurus…).

Dans le contexte du web sémantique, les techniques d'extraction d'informations à partir des textes peuvent être utilisées pour :

- la construction de l'ontologie : en faisant émerger, en coopération avec les experts du domaine, la terminologie du domaine, qui peut être ensuite formalisée pour obtenir l'ontologie (Aussenac-Gilles et al, 2000, 2002) ;
- l'enrichissement d'une ontologie : en permettant de compléter une ontologie existante (Golebiowka et al, 2001, 2002) ;
- l'instanciation d'une ontologie : en produisant la liste des termes qui peuvent être considérés comme une instance d'un concept ou d'une relation ;
- la génération des annotations : en détectant les informations importantes qui peuvent décrire le contenu d'un document.

Dans notre travail nous avons exploité les deux derniers points.

3 Description de la méthode

Afin de réaliser un système de génération semi-automatique d'annotations sémantiques pour les articles du domaine des biopuces, deux points sont à considérer : (1) quelle ontologie pourra être la base des annotations construites et (2) quels outils de TAL permettront de traiter les textes et d'en extraire les informations pertinentes.

Pour le premier point, notre choix s'est donc porté sur le réseau sémantique de UMLS qui couvre presque la totalité du domaine biomédical. Quant aux outils de TAL utilisés, nous les présentons dans la section 3.1.

Les étapes de notre méthode ainsi que l'architecture générale de notre système sont présentés dans la section 3.2.

3.1 Présentation des outils de TAL utilisés

3.1.1 GATE (General Architecture for Text Engineering)

Il s'agit d'une architecture offrant une infrastructure pour construire des applications de traitement linguistique. Cette architecture a été proposée afin de faciliter la tâche des chercheurs travaillant sur le parsing, l'étiquetage et l'analyse morphologique, dans le but de fournir des applications dans les domaine de l'extraction de l'information, la génération des résumés, la traduction…

Gate (Cunningham et al, 2002) comprend trois composantes principales :

- Une base de données qui a pour modèle un modèle orienté objet et qui permet de stocker des informations sur les textes.
- Une interface graphique pour lancer les outils de traitement sur les données et pour visualiser ensuite les résultats obtenus.
- Une collection d'algorithmes qui interagissent avec la base de données et l'interface et qui constituent une collection d'objets réutilisables pour l'ingénierie linguistique.

3.1.2 Syntex

L'analyseur syntaxique, Syntex (Bourigault & Fabre, 2000), permet l'extraction à partir d'une collection de textes, d'une liste de syntagmes nominaux, verbaux et adjectivaux. Le résultat est représenté sous la forme d'un graphe de dépendances, dans lequel chaque syntagme extrait est relié à sa tête et son expansion syntaxique.

Syntex a été utilisé dans plusieurs travaux sur l'extraction d'informations à partir des textes. Dans (Le Moigno et al, 2002), il a fait partie du noyau sur lequel repose une méthode pour la construction d'une ontologie à partir d'un corpus dans le domaine de la réanimation chirurgicale.

3.2 Notre méthode

Pour générer semi-automatiquement des annotations sémantiques pour des articles traitant le domaine des biopuces et ce, en se basant sur l'ontologie UMLS, nous avons élaboré une méthode qui se décompose en quatre étapes décrites ci-dessous.

La figure suivante schématise l'architecture générale du système reposant sur cette méthode:

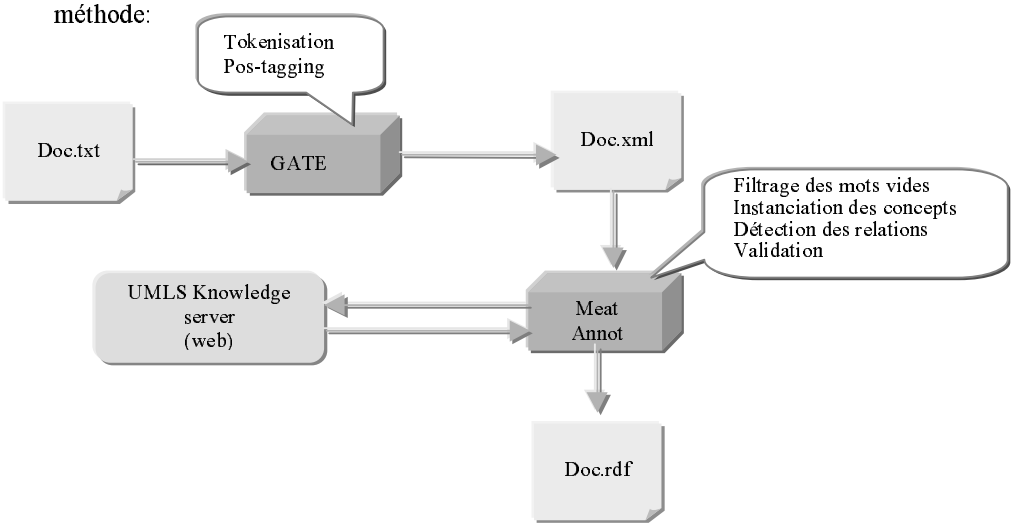

Fig1 : Architecture générale du système proposé

3.2.1 Etape 1 : Extraction des candidats termes

Dans cette étape, nous avons utilisé les modules Tokeniser et PosTagger de GATE. Le Tokeniser permet de découper le texte en unités élémentaires ou de base appelés tokens en distinguant les espaces et les ponctuations, et le PosTagger permet d'affecter à chaque mot une catégorie grammaticale (Nom, verbe …) en prenant en compte son emplacement dans la phrase.

Après ces deux phases, vient l'extraction des candidats termes. Nous avons considéré une fenêtre d'extraction de taille quatre (quatre mots successifs peuvent représenter un terme). Pour chaque candidat terme, nous vérifions s'il fait partie de UMLS (étape 2). Si c'est le cas, nous passons au mot suivant, sinon, la fenêtre d'extraction est diminuée jusqu'à ce qu'elle devienne nulle.

Dans cette phase d'extraction une règle a été suivie : un candidat terme ne peut commencer ni par un verbe ni par un mot vide (article, préposition…). Cela a permis d'augmenter le degré de pertinence des termes extraits et de diminuer considérablement le temps de calcul.

3.2.2 Etape 2 : Interrogation de UMLS

Afin de faciliter le partage et l'accès à UMLS, NLM a créé un serveur de connaissances sur le web appelé UMLSKS. Il permet l'accès à toutes les ressources de UMLS (métathesaurus, réseau sémantique). L'interface fournie offre la possibilité de navigation et d'interrogation de UMLS à distance.

Nous avons utilisé une API java qui permet d'accéder à ce serveur pour valider nos termes. Chaque candidat terme extrait est envoyé vers ce serveur qui nous renvoie une réponse sous format XML. Dans le cas où la réponse est positive, le contenu est traité afin d'en extraire les informations pertinentes telles que les synonymes du terme envoyé et le type sémantique auquel il appartient.

L'utilisation d'un tel serveur peut paraître coûteuse mais nous avons fait ce choix pour plusieurs raisons : (a) l'incomplétude des versions de UMLS mises en ligne, (b) la complexité de la gestion d'une base de données énorme telle que celle de UMLS, (c) le serveur nous offre une analyse de haut niveau en lemmatisant les termes fournis et en traitant quelques variations linguistiques simples (« development of lung » est reconnu comme « lung development »), ce qui complète le travail fait dans la première étape.

3.2.3 Etape 3 : Extraction des relations entre les termes

Pour cette étape, nous avons passé tous les textes à Syntex dans le but de découvrir les syntagmes verbaux fréquemment utilisés par les auteurs des articles et qui pourraient constituer une relation potentielle entre deux concepts.

Par la suite, nous avons utilisé JAPE (Cunningham et al, 2003), un langage basé sur les expressions régulières et qui offre la possibilité de créer des grammaires permettant l'extraction d'information d'un texte déjà traité par GATE. Pour chaque

relation retenue de l'étude faite par Syntex, nous avons créé manuellement une grammaire permettant d'extraire toutes ses occurrences ainsi que les concepts qui jouent le rôle d'acteurs pour cette relation.

Exemple de grammaire utilisée:

```
({Token.string == "play"}                    |
    {Token.string == "plays"})
    {SpaceToken}
    ({Token.string == "a"}                   |
    {Token.string == "an"})?
    ({SpaceToken})?
    ({Token.string == "vital"}|
    {Token.string == "important"}            |
    {Token.string == "critical"}             |
    {Token.string == "some"}          |
    {Token.string == "unexpected"}           |
    {Token.string == "multifaceted"}          |
    {Token.string == "major"})?
    ({SpaceToken})?
    ({Token.string == "role"}                 |
    {Token.string == "roles"})
```

Cette grammaire permet de détecter toute occurrence de la relation « play role ».

3.2.4 Etape 4 : Génération des annotations

Dans l'optique de l'extension du web vers le web sémantique, plusieurs langages tels que RDF(S), DAML-OIL et OWL ont vu le jour afin de représenter les ontologies et les annotations sémantiques, pour décrire les ressources contenues dans le nouveau web sémantique. Parmi ces langages, nous avons opté pour deux langages proposés par le W3C : RDFS et RDF. RDFS a permis la description des ontologies et RDF celle des annotations. Le codage de l'ontologie en RDFS a été fait à l'aide d'un script permettant de traduire automatiquement le réseau sémantique de UMLS de son format textuel vers une ontologie représentée dans le format RDFS :

- Dans une première étape, la hiérarchie des types conceptuels (resp. des relations), liés entre eux par la relation « is-a », a été traduite en une hiérarchie de classes (resp. de propriétés) liées par la relation «SubClassOf » (resp. «SubPropertyOf »), de telle sorte que chaque classe (resp. propriété) hérite les caractéristiques de sa classe (resp. propriété) mère.
- Dans une seconde étape, le domaine d'application de chaque relation (i.e. l'ensemble des classes sur lequel une relation peut être appliquée) a été défini. Généralement, ce travail nécessite une étude laborieuse du réseau sémantique. Notre utilisation de UMLS ayant juste pour but la construction des annotations avec des relations particulières telle que « Play role », nous

nous sommes contentés de lier les classes de plus haut niveau uniquement, ce qui a allégé le processus.

En utilisant l'ontologie ainsi traduite, nous avons développé une interface permettant de collecter toutes les informations fournies par l'outil de TAL et de les représenter à l'utilisateur qui choisit de les valider.

Une fois ces informations validées, l'annotation RDF correspondante est générée automatiquement. Pour mieux comprendre le déroulement de tout le processus, nous allons l'illustrer sur l'exemple d'un document qui parle du développement des poumons et qui contient la phrase :

« *HGF plays an important role in lung development* »

Les informations extraites de cette phrase sont :

- HGF : une instance du concept « Amino Acid, Peptide, or protein »
- Lung development : une instance du concept « organ or tissue function »
- HGF **play role** lung development : une instance de la relation « play role » entre les deux termes.

Et l'annotation RDF correspondante obtenue est la suivante :

```
<rdf:RDF
   xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
   xmlns:m='http://www.inria.fr/acacia/meat#'
   xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'>
<m:Amino_Acid_Peptide_or_Protein rdf:about='HGF#'>
          <m:play_role>
           <m:Organ_or_Tissue_Function rdf:about='lung
            development#'/>
          </m:play_role>
</m:Amino_Acid_Peptide_or_Protein>
</rdf:RDF>
```

4 Validation et discussion

Notre corpus de tests nous a été fourni par une équipe de recherche de l'IPMC travaillant sur les expériences biopuces. Nous disposons donc d'une quinzaine de documents traitant les maladies des poumons. Les tests que nous avons effectués visaient essentiellement la validation de la méthode d'extraction d'instances de concepts (section 4.1), la méthode de l'extraction des relations (section 4.2) et enfin la cohérence des annotations générées avec l'ontologie (section 4.3).

4.1 Extraction des instances

Le traitement du corpus par notre outil MEAT Annot nous a permis d'extraire la majorité des termes appartenant au metathésaurus de UMLS. Les quelques exceptions remarquées sont dues, le plus souvent, aux fautes d'orthographe faites par les auteurs lors de la rédaction de l'article, aux abréviations utilisées et qui ne figurent pas dans UMLS et enfin à l'utilisation de caractères spéciaux tels que les caractères latins.

Ces termes sont automatiquement reliés aux concepts auxquels ils font référence, et sont ensuite utilisés pour la génération de l'annotation (voir fig3).

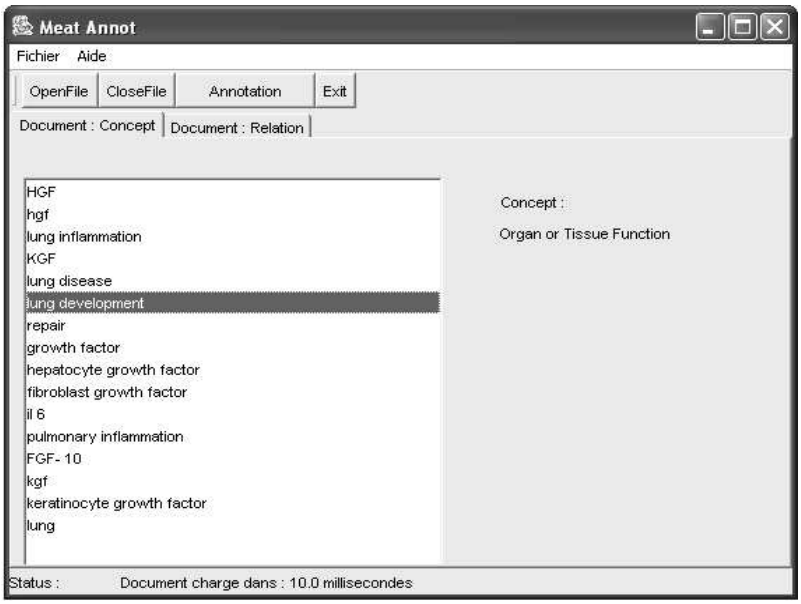

Fig2 : exemple d'une liste des termes extraits pour un document

4.2 Extraction des relations

Les tests d'extraction des relations ont été réalisés sur la relation « Play role ». La grammaire définie dans la section 3.2.3 nous a permis de détecter 35 occurrences de la relation « Play role » parmi les 49 extraites par Syntex. Ce pourcentage nous paraît intéressant vu la simplicité de la grammaire utilisée. Par ailleurs, l'étude que nous avons menée sur le corpus nous a montré que cette différence est généralement due, soit aux différentes formes du verbe « play » (playing, played…), soit aux variations linguistiques existant dans le texte, comme dans l'exemple «*the key role that endogenous KGF has been shown to play in wound healing in the skin* ».

Une fois cette extraction terminée, l'outil essaie alors de détecter les termes reliés par chaque instance de relation et ce en se basant sur les informations (instances) récupérées dans la première phase (section 4.1) afin de proposer l'annotation adéquate à chaque relation extraite. Cette phase dépend de la qualité d'extraction des termes, car le système ne propose rien s'il ne dispose pas d'informations suffisantes sur les termes se trouvant à gauche et à droite de la relation extraite.

4.3 Cohérence des annotations

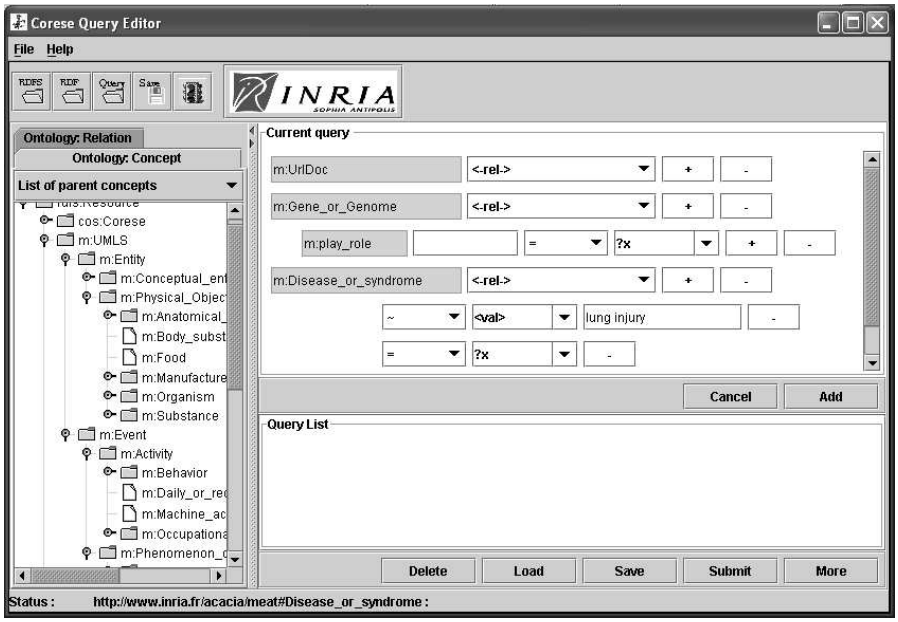

Fig3 : Interface de CORESE montrant la structure de l'ontologie

Afin de vérifier la structure de l'ontologie et la cohérence des annotations générées, nous avons utilisé le moteur de recherche sémantique CORESE (Corby & Faron-Zucker, 2002), qui à partir d'une ontologie en format RDFS et d'un ensemble d'annotations en format RDF permet de faire des recherches sur la base d'annotations en se basant sur l'ontologie. Nous avons donc utilisé l'interface de CORESE afin d'interroger l'ensemble des documents, ce test nous a permis de vérifier la cohérence des annotations avec l'ontologie ainsi que leur conformité par rapport au langage RDF. La figure Fig4 montre la hiérarchie des concepts de l'ontologie UMLS à gauche ainsi qu'un exemple de requête permettant de trouver tous les articles exprimant qu'un gène particulier joue un rôle dans les maladies des poumons.

5 Conclusion

Nos travaux peuvent se comparer avec (a) les travaux exploitant l'extraction d'information dans le domaine de la biologie (Staab, 2002), (b) ceux sur la génération d'annotations sémantiques pour le web sémantique (Handschuh et al, 2002).

Dans le domaine de la fouille d'articles en biologie, (Shatkay et al, 2002) propose des techniques statistiques et des algorithmes d'apprentissage pour découvrir des interactions entre gènes à partir de résumés d'articles en biologie dans la base PubMed. Reposant sur des outils linguistiques, notre approche diffère d'une part, de l'approche proposée par (Blaschke et al, 2002), qui utilise une méthode mixte (statistique et linguistique) pour l'extraction des relations entre les gènes, et d'autre part de celle de (Nédellec, 2002), qui à la suite d'une phase d'apprentissage sur un grand corpus définit des règles d'extraction de noms de gènes et de protéines ainsi que les relations entre eux.

Concernant la génération automatique d'annotations RDF, notre approche diffère de (Cao et al, 2003) qui repose sur la généralisation à partir d'un exemple d'annotation manuelle sur des documents web structurés. Elle diffère aussi de celle proposée dans (Golebiowska, 2002) où l'annotation générée sur un document consistait en des instances des concepts de l'ontologie qui avaient justement été générés à partir de ce document : la génération d'annotation était donc liée à l'enrichissement de l'ontologie. Or notre approche permet de créer des annotations consistant non seulement en des instances de concepts, mais aussi en *des instances de relations*, et le tout en reposant sur une ontologie déjà existante.

Ce travail se situe dans le cadre d'une collaboration avec des biologistes travaillant sur les biopuces, afin de leur apporter l'aide nécessaire pour la validation et l'interprétation de leurs expériences. L'utilisation des annotations sémantiques ayant un impact positif sur la pertinence des résultats d'une recherche d'information, cette méthode devrait faciliter la tâche des biologistes.

Certaines améliorations de notre prototype seraient intéressantes : (a) offrir à l'utilisateur la possibilité de rajouter une instance d'un concept n'existant pas dans le metathésaurus de UMLS, ce qui permettra d'enrichir ce dernier ; (b) développer des heuristiques pour proposer tous les termes dans le voisinage d'une relation et appartenant au domaine de son application afin de créer de nouvelles instances qui ne sont pas détectées automatiquement.

Enfin, notons que la méthode présentée pour la génération semi-automatique des annotations sémantiques est générique, elle est indépendante des outils linguistiques utilisés et peut se baser sur n'importe quelle ontologie.

Remerciements : Nous remercions Pascal Barbry et son équipe pour leur expertise du domaine des biopuces, Didier Bourigault pour nous avoir fourni les résultats de son outil Syntex sur notre corpus, Laurent Alamarguy pour son aide dans le domaine linguistique, et la région PACA qui finance ces travaux par une bourse régionale.

Références

Aussenac-Gilles N., Biébow B. & Szulman S., (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In IC 2000, Toulouse

Aussenac-Gilles N., Biébow B. & Szulman S., (2003). D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. In TIA 2003, p. 41-53, Strasbourg

Berners-Lee T., Hendler J. & Lassila O (2001). The Semantic Web, Scientific American, 84(5) p. 34-43

Blaschke C. & Valencia A., (2002). Molecular biology nomenclature thwarts information-extraction progress. IEEE Intelligent Systems & their Applications, p. 73-76, Mai/June.

Bourigault D. & Fabre C. (2000), Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de grammaire, Vol.25, pp.131-151

Cao T-D., Gandon F., Dieng R., (2003), Intégration de sources extérieures dans un Web sémantique d'entreprise géré par un système multi-agents. In IC 2003, Laval

Corby O. & Faron-Zucker C. (2002), Corese: A Corporate Semantic Web Engine. WWW11 Workshop on Real World RDF and Semantic Web Applications, Hawaii

Cunningham H., Maynard D., Bontcheva K. & Tablan V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ACL'02.

Cunningham H. & Maynard D. & Bontcheva K. & Tablan V. & Ursu C. & Dimitrov M. (2003), Developing Language Processing Components with GATE (User Guide)

Gruber T. (1993), A Translation Approach to Portable Ontology Specifications ,Knowledge Acquisition , p. 19-220.

Golebiowska J. & Dieng-Kuntz R. & Corby O., & Mousseau D. (2001), Building and Exploiting Ontologies for an Automobile Project Memory. In K-CAP, Victoria.

Golebiowska J. (2002), Exploitation des ontologies pour la mémoire d'un projet véhicule, Méthode et outil Samovar, thèse de doctorat en informatique, UNSA

Golebiowska J. , Dieng R., Corby O. & Mousseau D. (2002). Ontologies au service de la mémoire d'un projet-véhicule et de la recherche d'information, Doc. Numérique, p.173-192.

Handschuh S. , Koivunen M., Dieng R. & Staab S. (2003), eds, Proc. of KCAP'2003 Workshop on Knowledge Markup and Semantic Annotation, Sanibel, Florida, October 26.

Le Moigno S., Charlet J., Bourigault D. & Jaulent M. (2002), Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale, In IC 2003, Rouen

Nédellec C., (2002), Bibliographical Information Extraction in Genomics. IEEE Intelligent Systems & their Applications, p. 76-80, Mai/June.

Schulze-Kremer S. & Smith B. & Kumar A.(2002). Revising the UMLS Semantic Network

Soo V., Lee C., Li C., Chen S. & Chen C. (2003). Automated Semantic Annotation and Retrieval Based on Sharable Ontology and Case-based Learning Techniques, JCDL'03, Texas.

Shatkay H., Edwards S. & Boguski M., (2002), Information Retrieval Meets Gene Analysis. IEEE Intelligent Systems & their Applications, p. 45-53

Staab S., Maedche A. & Handschuh S (2001), An annotation framework for the semantic web, In Proceedings of the First Workshop on Multimedia Annotation, Japon

Staab S. (2002), Mining Information for Functional Genomics. IEEE Intelligent Systems & their Applications, p. 66-80, March-April.