Titre :

Conception d'un prototype d'annotateur sémantique biomédical francophone.

Information :

Encadrant : Clement Jonquet & Mathieu Roche (LIRMM, UM2) – {jonquet,mroche}@lirmm.fr

Spécialités : IFPRU (DIWEB, I2A, GL)

Nombre d'étudiants : 2-3

Mots clés :

Ontologies, Text mining, Semantic Annotation, Natural Language processing, Web application, Knowledge representation, Service-oriented architecture, Web services, biomedical data.

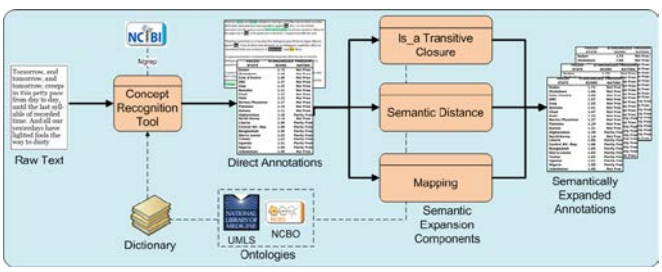
Résumé (francais) :

Les terminologies et ontologies biomédicales jouent un rôle clé dans l'interopérabilité sémantique des données biomédicales en servant de dénominateur commun. Il existe un challenge qui consiste à produire, pour les descriptions textuelles des ressources de données, des annotations (ou labels, tags) qui utilisent des termes d'ontologies et faciliteront la recherche et l'indexation de ces données ainsi que leur intégration. L'université de Stanford a développé un service d'annotation sémantique qui permet aux chercheurs d'utiliser les ontologies biomédicales pour annoter leurs données automatiquement. L'annotateur traite les métadonnées textuelles brutes pour les taguer avec des concepts définis dans des ontologies biomédicales et utilise la connaissance représentée dans les ontologies pour étendre ces annotations. Dans un premier temps, le travail consiste, à partir de l'outil de Stanford, à prototyper un service équivalent qui traitera des données textuelles francophones en utilisant des terminologies francophones (fournies). Ensuite, nous nous intéresserons à l'amélioration du processus d'annotation grâce aux techniques du traitement automatique de la langue.

Présentation du contexte :

Public biomedical data is already enormous and is expanding very fast. Biomedical data integration and interoperability is necessary to enable cross data search and query as well new scientific discoveries [5]. These data are often unstructured and are available in different formats (database, documents, etc.) preventing discoveries that could be made by merging them. To address this problem, the biomedical community has turned to ontologies and terminologies to describe their data and turn them into structured and formalized knowledge [1, 8]. Ontologies formalize the knowledge of a domain by means of concepts, relations and rules that apply to that domain [3].

It is important that different service providers work to converge toward unified set of tools and formats to process biomedical data in the design of clinical, research or industrial applications related to biomedicine. The *Stanford Center for Biomedical Informatics Research* (BMIR) group at Stanford University (http://bmir.stanford.edu) has developed the NCBO Annotator [4] (http://bioportal.bioontology.org/annotator), a Web service which provides a mechanism to employ ontology-based annotation in curation, data integration, and indexing workflows—using any of the several hundred public ontologies in the BioPortal repository [6] (cf. figure).



In a first step the user submitted text is given as input to a concept recognition tool along with a dictionary. The concept recognizer does fast and efficient string matching against the dictionary terms to recognize concepts. In a second step, semantic expansion components use the ontology structure to create additional annotations.

Présentation du sujet :

Using the NCBO Annotator (available open source), the project aims to prototype a French version of the service using a French dictionary that will be provided. This will consist in deploying a REST Web service on a LIRMM's server using the following technological stack: MySQL / Java / Hibernate / Spring / RESTLet / XML.

In a second time, we will look into using natural language processing techniques to improve the concept recognition step and customize the workflow for French language. Indeed, a core component of the NCBO Annotator indexing workflow is the concept recognition tool. The NCBO Annotator uses Mgrep [2], a concept recognizer appropriate for fast string matching in English. We will work at the creation of an ontology-based concept recognition tool that will specifically handle French text. We will proceed to a review of existing approaches. Then, the appropriate natural language processing algorithms (especially using the pattern approach applied in a French context) will be used to design and implement a fast and scalable tool, that can be used with the available French ontologies and that can be integrated in the rest of the workflow.

If the project is successful and time allows, we will proceed to performance evaluation against tools developed by the *Catalogue et Index des Sites Médicaux de langue Française* (CISMeF) group at Rouen University Hospital (http://www.cismef.org): (i) French Multi-Terminology Indexer [7] and (ii) Extracteur de Concepts Multi-Terminologique – http://ecmt.chu-rouen.fr).

Références :

[1] Olivier Bodenreider and Robert Stevens. Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, 7(3):256–274, August 2006.

[2] Manhong Dai, Nigam H. Shah, Wei Xuan, Mark A. Musen, Stanley J. Watson, Brian D. Athey, and Fan Meng. An Efficient Solution for Mapping Free Text to Ontology Terms. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'08*, San Francisco, CA, USA, March 2008.

[3] Tom R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, June 1993.

[4] Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09*, pages 56–60, San Francisco, CA, USA, March 2009.

[5] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In L. Popa, editor, *21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS'02*, pages 233–246, Madison, WI, USA, June 2002.

[6] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37((web server)):170–173, May 2009.

[7] Suzanne Pereira, Saoussen Sakji, Aurélie Névéol, Ivan Kergourlay, Gaétan Kerdelhué, Elisabeth Serrot, Michel Joubert, and Stéfan J. Darmoni. Multi-terminology indexing for the assignment of MeSH descriptors to medical abstracts in French. In *American Medical Informatics Association Annual Symposium, AMIA'09*, pages 521–525, Washington DC, USA, November 2009.

[8] Daniel L. Rubin, Nigam H. Shah, and Natalya F. Noy. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90, 2008.