Annotation de textes par extraction d'informations lexico-syntaxiques et acquisition de schémas conceptuels de causalité

Laurent Alamarguy*, Rose Dieng-Kuntz*, Catherine Faron-Zucker**

*ACACIA, INRIA Sophia Antipolis
{Laurent.Alamarguy, Rose.Dieng}@sophia.inria.fr
**MAINLINE, I3S, Sophia Antipolis
faron@essi.fr

Résumé. Nous présentons la méthode INSYSE (Interface Syntaxe SEmantique) pour l'annotation de documents textuels. Notre objectif est de construire des annotations sémantiques de ces résumés pour interroger le corpus sur la fonction des gènes et leurs relations de causalité avec certaines maladies. Notre approche est semi-automatique, centrée sur (1) l'extraction d'informations lexico-syntaxiques à partir de certaines phrases du corpus comportant des lexèmes de causation, et (2) l'élaboration de règles basées sur des grammaires d'unification permettant d'acquérir à partir de ces informations des schémas conceptuels instanciés. Ceux-ci sont traduits en annotations RDF(S) sur la base desquelles le corpus de textes peut être interrogé avec le moteur de recherche sémantique Corese.

1 Introduction

Lors de la constitution d'une mémoire de communauté en génomique fonctionnelle, la notion de causalité est centrale pour appréhender certaines corrélations. Dans le cadre du *web sémantique* l'automatisation de cette tâche doit permettre, à partir de données hétérogènes, de détecter et générer de nouvelles représentations conceptuelles traduisant cette notion.

Nous présentons une méthode semi-automatique d'annotation de documents textuels basée sur l'acquisition de schémas conceptuels[1] à partir de l'extraction de structures lexico-syntaxiques ; elle est baptisée INSYSE - pour INterface SYntaxe SEmantique. Cette méthode est appliquée à un corpus de 5000 résumés médicaux issus de Medline et traitant de maladies du système nerveux central et des interactions des gènes dans ces maladies. Notre objectif est de construire des annotations sémantiques de ces résumés qui permettent d'interroger le corpus sur la fonction des gènes et leurs relations de causalité avec certaines maladies pour ainsi constituer une mémoire de communauté.

Nous présentons dans cet article les différentes étapes de la méthode INSYSE : la partie suivante est consacrée à l'extraction d'informations lexico-syntaxiques à partir de certaines phrases comportant des lexèmes de causation ; la partie 3 est dédiée à l'élaboration de règles basées sur des grammaires d'unification qui permettent d'extraire des informations lexico-syntaxiques des schémas conceptuels instanciés. La partie 4 décrit comment ces schémas sont traduits en annotations RDF(S) sur la base desquelles le corpus pourra être interrogé à l'aide du moteur de recherche sémantique Corese (Corby et al. 2004). Nous comparons dans

[1] Un schéma conceptuel non instancié constituant de fait un *template* d'annotation.