Extraction et identification d'entités complexes à partir de textes biomédicaux

Julien Lorec*,**, Gérard Ramstein**, Yannick Jacques*

*INSERM U601, Département de Cancérologie, Équipe 3: cytokines et récepteurs
{julien.lorec,yjacques}@nantes.inserm.fr
**LINA, Équipe C.O.D, École polytechnique de l'université de Nantes
gerard.ramstein@polytech.univ-nantes.fr

Résumé. Nous présentons ici un système d'extraction et d'identification d'entités nommées complexes à l'intention des corpus de spécialité biomédicale. Nous avons développé une méthode qui repose sur une approche mixte à base d'ensemble de règles a priori et de dictionnaires contrôlés. Cet article expose les techniques que nous avons mises en place pour éviter ou minimiser les problèmes de synonymie, de variabilité des termes et pour limiter la présence de noms ambigus. Nous décrivons l'intégration de ces méthodes au sein du processus de reconnaissance des entités nommées. L'intérêt de cet outil réside dans la complexité et l'hétérogénéité des entités extraites. Cette méthode ne se limite pas à la détection des noms des gènes ou des protéines, mais s'adapte à d'autres descripteurs biomédicaux. Nous avons expérimenté cette approche en mesurant les performances obtenues sur le corpus de référence GENIA.

1 Introduction

A cette date, de nombreuses méthodes d'étiquettage d'entités biologiques pour les corpus de spécialité ont été proposées ; quelles soient à base de règles (Fukuda et al. (1998)) ou encore réposant sur des techniques d'apprentissage (Collier et al. (2000)). Néanmoins, la simple détection de la présence d'une entité nommée dans un texte ne suffit pas pour l'identifier et l'associer à une instance d'entité biologique particulière. Le couplage des méthodes d'extraction des entités nommées avec l'utilisation de dictionnaires semble être une solution particulièrement adaptée à ce type de problématique (Koike et al. (2003)). De plus, la majorité de ces techniques d'extraction d'entités nommées ont été développées dans le but de ne détecter que quelques types particuliers et spécifiques d'objets biologiques, notamment les gènes et les protéines, et ne peuvent être facilement adaptés à d'autres contextes.
Il existe trois principales difficultés à prendre en compte lors d'une recherche à base de dictionnaire :

- la présence de termes synonymes et la résolution des différentes abréviations et acronymes,
- la variabilité des mots tant au niveau de l'orthographe que de la morphologie et de la syntaxe mais aussi d'un point de vue lexico-sémantique, de la présence d'insertions/déletions et permutations,