Processus d'acquisition d'un dictionnaire de sigles et de leurs définitions à partir d'un corpus

Vladislav Matviico, Nicolas Muret, Mathieu Roche

LIRMM, Université Montpellier 2 - CNRS UMR5506,
mroche@lirmm.fr

Résumé. Le logiciel présenté dans cet article s'appuie sur une approche d'acquisition de sigles à partir de données textuelles.

1 Introduction

De nombreux domaines comme la biologie ou la médecine voient naître chaque jour de nouveaux termes et abréviations, notamment des sigles. Un sigle est un ensemble de lettres initiales servant d'abréviation, par exemple "RATP" peut être associé à la définition (aussi appelée expansion) "Régie Autonome des Transports Parisiens". Nos travaux ont consisté à développer un logiciel afin de faciliter l'acquisition ou l'enrichissement de dictionnaires en extrayant automatiquement, à partir de diverses sources, les sigles et leur(s) définition(s). Une fois ces dictionnaires constitués, l'approche $AcroDef$ que nous avons proposée dans (Roche et Prince (2007)) consiste à établir la définition pertinente d'un sigle présent dans un document. Dans ces documents, la définition n'est pas toujours présente d'où la difficulté du traitement. Dans ce contexte, il est donc essentiel d'avoir à disposition un dictionnaire adapté, ce qui justifie les travaux présentés dans cet article.

De nombreuses méthodes pour extraire les sigles et leur(s) définition(s) ont été développées (Larkey et al. (2000); Okazaki et Ananiadou (2006)). La plupart des approches de détection de sigles dans les textes s'appuient sur l'utilisation de marqueurs spécifiques associés à des heuristiques adaptées. Certains travaux récents (Okazaki et Ananiadou (2006)) consistent à associer ces approches à des mesures statistiques spécifiques pour améliorer la qualité des méthodes d'acquisition de dictionnaires. L'approche que nous avons développée se compose de deux étapes successives qui sont détaillées dans la section 2.

2 Acquisition d'un dictionnaire sigles/définitions

Notre méthode qui consiste à extraire les candidats sigles/définitions s'appuie sur la présence de marqueurs (parenthèses, crochets). Deux situations peuvent alors être considérées :
1. Le sigle se situe avant la définition qui se trouve entre les marqueurs (les parenthèses dans le cas le plus courant). Exemple : "`... S.I.G. (Solde Intermédiaire de Gestion) ...`"
2. La définition se trouve avant le sigle qui se trouve entre les marqueurs. Exemple : "`... les Systèmes d'Informations Géographiques (SIG) ...`". Dans ce cas, la taille de la définition est pour le moment indéterminable. Il est ainsi nécessaire de la définir arbitrairement en fonction du nombre de lettres composant le sigle. Nous avons expérimentalement fixé cette taille à trois fois le nombre de lettres composant le sigle.