UNITEX/GRAMLAB: plateforme libre basée sur des lexiques et des grammaires pour le traitement des corpus textuels

Tita Kyriacopoulou*, Claude Martineau*, Cristian Martinez*

*5 boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2
{tita,claude.martineau,cristian.martinez}@univ-mlv.fr

Résumé. L'objectif de notre recherche est de répondre aux besoins croissants et divers d'extraction d'information pertinente exprimés par de nombreuses disciplines. Nous utilisons pour cela l'analyseur multilingue de corpus Unitex/GramLab développé à l'Université Paris-Est Marne-la-Vallée. Il fait appel à une approche symbolique et utilise des ressources linguistiques, dictionnaires électroniques et grammaires locales. Cette présentation ne constitue qu'une prise en main d'Unitex/GramLab et ne reflète que très partiellement les possibilités du logiciel et son champ d'utilisation, notamment pour l'extraction d'information, qui s'étend du monde de la recherche à celui de l'industrie.

1 Introduction

Un des objectifs de notre recherche est l'identification et l'extraction automatique d'information pertinente à partir de données textuelles provenant d'une multiplicité grandissante de domaines (littéraire, journalistique, scientifique, médical,technique, etc) et de sources (bases de données, bibliothèques numériques, blogues, etc). Afin de permettre un accès efficace à l'information et d'en simplifier l'utilisation nous devons prendre en compte le traitement de corpus de grande taille, la réduction du bruit contenu, le multilinguisme, ainsi que plusieurs tâches de traitement par l'ordinateur et l'utilisateur. C'est pourquoi il est nécessaire d'automatiser certains processus, notamment l'extraction d'entités nommées (noms propres, adresses, dates, etc). Cette analyse automatique est effectuée par des outils principalement issus de deux disciplines : l'informatique et le TAL (Traitement Automatique des Langues). Les deux disciplines fondent leurs analyses sur des techniques statistiques et/ou des connaissances et des ressources linguistiques.

UNITEX/GRAMLAB [1] utilise des ressources linguistiques même s'il doit évoluer vers un système hybride. Il est open source, multilingue [2], multiplateforme et permet d'analyser des textes en langue naturelle grâce à des ressources linguistiques telles que des dictionnaires électroniques et des grammaires locales. Ces dernières sont fondées sur la notion d'automate et de manière plus générale de réseau de transitions récursif (RTN) comportant des sorties. Ces grammaires sont représentées sous forme de graphes aisément réalisables grâce à un éditeur intégré.

1. UNITEX/GRAMLAB a été principalement développé par Sébastien Paumier (2001-2012). Son développement se poursuit grâce à une communauté de développeurs et de linguistes.

2. Français, anglais, . . ., grec, russe, arabe (écriture de droite à gauche), thaï et coréen (absence de séparateurs).