Vers une base de connaissances biographiques : extraction d'information et ontologie

Laurent Kevers* et Cédrick Fairon*

* Cental, Université catholique de Louvain (UCL)
Place Blaise Pascal, 1 - 1348 Louvain-la-Neuve - Belgique
laurent.kevers@uclouvain.be - cedrick.fairon@uclouvain.be

Résumé. Le projet B-Ontology a pour but l'extraction, l'organisation et l'exploitation de connaissances biographiques à partir de dépêches de presse. Sa réalisation requiert l'intégration de diverses technologies, principalement l'extraction d'information, les ontologies et bases de connaissances, les techniques de data mining. Cet article propose un aperçu des choix réalisés dans le cadre du projet. Cette démarche permet également de définir un environnement d'outils utiles pour les applications d'extraction et de gestion de connaissances.

1 Introduction

B-Ontology est un projet de recherche appliquée dont l'objectif est de construire le prototype d'une application capable d'extraire et d'organiser de l'information biographique. Cette information sera exploitée dans le cadre du processus de rédaction d'une agence de presse. L'agence Belga diffuse quotidiennement plus de 250 dépêches en deux langues (français et néerlandais). Cette masse textuelle représente environ 70.000 mots par jour (25 millions de mots en un an) par langue. Dans ce projet, nous nous intéresserons aux informations qui concernent les personnes, les organisations et les événements dans lesquels elles interviennent. Le résultat est stocké dans un ensemble de données structurées facilement consultable. Des systèmes comparables existent déjà (NewsExplorer[1], KIM [2]) mais ne couvrent cependant pas toutes les fonctionnalités désirées ici et sont souvent uniquement adaptés aux textes en anglais.

La première partie exposera les méthodes d'extraction d'information. La deuxième s'attardera sur le choix de l'organisation des données. Une troisième partie, présentera une réalisation concrète, mais limitée, de la base de connaissances et quelques aspects de data mining.

2 Extraction d'information

2.1 Définitions des entités et du formalisme d'annotation

L'extraction d'information passe par l'annotation sémantique du texte. Cette tâche nécessite avant tout une bonne définition des types d'entités recherchées. On définit le concept

[1] http ://press.jrc.it/NewsExplorer/home/en/latest.html
[2] http ://www.ontotext.com/kim/index.html