Extraction d'entités dans des collections évolutives

Thierry Despeyroux*, Eduardo Fraschini*
Anne-Marie Vercoustre*

INRIA - Rocquencourt
Domaine de Voluceau
B.P. 105 - 78153 Le Chesnay Cedex
* prenom.nom@inria.fr
http://www-rocq.inria.fr/who/Prenom.Nom/

Résumé. Nous nous intéressons à l'extraction d'entités nommées avec comme but d'exploiter un ensemble de rapports pour en extraire une liste de partenaires. À partir d'une liste initiale, nous utilisons un premier ensemble de documents pour identifier des schémas de phrase qui sont ensuite validés par apprentissage supervisé sur des documents annotés pour en mesurer l'efficacité avant d'être utilisés sur l'ensemble des documents à explorer. Cette approche est inspirée de celle utilisée pour l'extraction de données dans les documents semi-structurés (wrappers) et ne nécessite pas de ressources linguistiques particulières ni de larges collections de tests. Notre collection de documents évoluant annuellement, nous espérons de plus une amélioration de notre extraction dans le temps.

1 Introduction

La reconnaissance et l'extraction d'entités nommées cherche à localiser et à classer les éléments atomiques d'un texte en catégories prédéfinies telles que noms de personnes, organisations, localisation, dates, quantités, valeurs monétaires, pourcentages etc. Ce domaine de recherche est très actif, bien que des outils commerciaux existent déjà. Citons, par exemple, REX[1] (Rosette® Entity Extractor), Inxight SmartDiscovery[2], Convera-RetrievalWare Entity Extraction[3] and Xerox-Research Entity Extraction system[4]. Le but de ces outils est d'annoter les documents avec des méta données (qui, quoi, où, quand...) permettant une recherche d'information de plus haut niveau.

La plupart de ces systèmes demandent d'importantes ressources linguistiques (listes d'entités, larges corpus de référence) et l'écriture manuelle de règles pour s'adapter à un domaine particulier. Par exemple, le système de Xerox utilise plus de 250 règles manuelles pour extraire des entités biologiques auxquelles s'ajoutent d'autres règles qui sont inférées automatiquement et évaluées sur de gros corpus.

[1] http://www.basistech.com/entity-extraction/
[2] http://www.inxight.com/products/smartdiscovery/ee/
[3] http://www.retrievalware.com/products/retrievalware/entity-extraction.asp
[4] http://www.ipvalue.com/technology/docs/Xerox_entity_extraction_pager.pdf