Solution Proxem d'analyse sémantique verticale : adaptation au domaine des Ressources Humaines

François-Régis Chaumartin[1]
(1) Proxem, 19 bd de Magenta, 75010 Paris
frc@proxem.com

RESUME

Proxem développe depuis 2007 une plate-forme de traitement du langage, Antelope, qui permet de construire rapidement des applications sémantiques verticales (par exemple, pour l'e-réputation, la veille économique ou l'analyse d'avis de consommateurs). Antelope a servi à créer une solution pour les Ressources Humaines, utilisée notamment par l'APEC, permettant (1) d'extraire de l'information à partir d'offres et de CVs et (2) de trouver les offres d'emploi correspondant le mieux à un CV (ou réciproquement). Nous présentons ici l'adaptation d'Antelope à un domaine particulier, en l'occurrence les RH.

ABSTRACT

How to adapt the Proxem semantic analysis engine to the Human Resources field

Proxem develops since 2007 the NLP platform, Antelope, with which one can quickly build vertical semantic applications (for e-reputation, business intelligence or consumer reviews analysis, for instance). Antelope was used to create a Human Resources solution, notably used by APEC, making it possible (1) to extract information from resumes and offers and (2) to find the most relevant jobs matching a given resume (or vice versa). We present here how to adapt Antelope to a particular area, namely HR.

MOTS-CLES : entités nommées, extraction de relations, création d'ontologies, similarité
KEYWORDS : named entities, information extraction, ontologies development, matching

1 La plate-forme de traitement linguistique Antelope

Antelope (Chaumartin, 2008) est une plate-forme de TAL intégrant des composants de traitement syntaxique et sémantique, ainsi qu'un lexique de large couverture pour l'anglais et le français. Elle permet de créer rapidement des applications d'analyse sémantique, en enchaînant plusieurs opérations au sein d'une chaîne de traitement.

La qualité des documents traités étant très variable, une correction orthographique est souvent nécessaire ; néanmoins, cette opération doit être effectuée avec une connaissance du contexte métier ; par exemple, les noms propres (qui ne figurent pas dans le lexique intégré) ne doivent pas être « corrigés » vers un mot proche.

La reconnaissance d'entités nommées vise classiquement à identifier des personnes, lieux et organisation. Dans un contexte d'enseigne de grande distribution, les entités intéressantes à détecter sont plutôt les produits, marques et concurrents cités, ainsi que des concepts liés au métier (risque sanitaire, risque juridique…). Dans le domaine des Ressources Humaines, il s'agira plutôt de reconnaître des métiers, des compétences, des talents, des diplômes…

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 11–12,
Grenoble, 4 au 8 juin 2012.

2 Adaptation d'Antelope au domaine RH

Nous avons développé une nouvelle approche d'acquisition à large échelle d'entités nommées. (a) Une première phase d'extraction terminologique permet d'amorcer la liste des concepts du domaine. (b) Une seconde phase utilise des ressources de large couverture (la Wikipédia et un WordNet pour le français) pour créer des gazettes ; en cas d'ambiguïté (le métier d'*architecte* relève par exemple du BTP de l'informatique), les termes des gazettes sont automatiquement associés à des mots clés activateurs ou inhibiteurs. (c) L'application de ces gazettes permet de constituer un premier corpus annoté selon les entités nommées du domaine. Un apprentissage (par CRF) est alors effectué sur le corpus, pour identifier de nouvelles instances d'entités. Les concepts correspondant aux entités nommées sont organisés sous forme de taxonomie. La figure 1 montre par exemple, sur la partie de gauche, l'organisation des métiers, compétences et talents sous forme d'arborescence. Cette information est utilisée lors de la recherche de documents similaires ; concrètement, elle permet de déterminer que –toutes choses égales par ailleurs– la compétence « développement Java » est plus proche d'une compétence « développement en langage objet » que de « développement COBOL ». Ce point améliore fortement la pertinence des documents trouvés lors d'une recherche.

FIGURE 1 – Une capture d'écran de la solution d'analyse sémantique RH de Proxem.

Remerciements

Le projet SIRE (FEDER) a partiellement financé l'adaptation d'Antelope au domaine RH.

Références

CHAUMARTIN F.-R. (2008). ANTELOPE, une plate-forme industrielle de traitement linguistique. *Traitement Automatique des Langues* 49:2.