Reconnaissance automatique de concepts à partir d'une ontologie

Valentina Ceausu, Sylvie Desprès

Université René Descartes
CRIP5 – Equipe IAA – Groupe SBC
UFR Mathématiques et Informatique
45 rue des Saints-Pères
75006 PARIS
valentina.ceausu@math-info.univ-paris5.fr
sd@math-info.univ-paris5.fr

Résumé Ce papier présente une approche qui s'appuie sur une ontologie pour reconnaître automatiquement des concepts spécifiques à un domaine dans un corpus en langue naturelle. La solution proposée est non-supervisée et peut s'appliquer à tout domaine pour lequel une ontologie a été déjà construite. Un corpus du domaine est utilisé dans lequel les concepts seront reconnus. Dans une première phase, des connaissances sont extraites de ce corpus en faisant appel à des fouilles de textes. Une ontologie du domaine est utilisée pour étiqueter ces connaissance. Le papier donne un aperçu des techniques de fouilles employées et décrit le processus d 'étiquetage. Les résultats d'une première expérimentation dans le domaine de l'accidentologie sont aussi présentés.

1 Introduction

L'important volume de documents disponibles en langue naturelle et leur évolution rapide font émerger la nécessité de définir des approches permettant de retrouver rapidement des informations pertinentes dans ces documents.

Ce papier présente une approche qui utilise une ontologie de domaine pour identifier automatiquement des concepts du domaine dans un corpus en langue naturelle. Cette identification de concepts peut servir dans différents contextes : annotation des documents, indexation d'une collection de documents, etc. L'approche proposée est complètement automatique et non-supervisée, mise à part l'utilisation d'une ontologie de domaine. Etant donnés une ontologie *O* et un corpus *C*, le but est de retrouver dans *C* des termes *w* qui sont l'expression linguistique des concepts de l'ontologie *O*. On peut ainsi étiqueter les termes retrouvés dans le corpus par des concepts de l'ontologie. Cet étiquetage est réalisé en trois étapes : (1) une première étape emploie des techniques de fouille de textes pour identifier des termes du domaine dans le corpus; (2) pour chaque terme *w* retrouvé, le voisinage sémantique *V(w)* est identifié ; (3) en supposant que les relations dans le voisinage du terme *w* soient déjà dans l'ontologie, le positionnement des relations dans l'ontologie et des mesures statistiques sont utilisés pour étiqueter le terme *w*.