Extraction et exploitation des annotations contextuelles

Noureddine Mokhtari*, Rose Dieng-Kuntz*

*INRIA
2004 route des lucioles - BP 93
FR-06902 Sophia Antipolis cedex
{Noureddine.Mokhtari, Rose.Dieng}@sophia.inria.fr

Résumé
Dans la perspective d'offrir un web sémantique, des travaux ont cherché à automatiser l'extraction des annotations sémantiques à partir de textes pour représenter au mieux la sémantique que vise à transmettre une page web. Dans cet article nous proposons une approche d'extraction des annotations qui représentent le plus précisément possible le contenu d'un document. Nous proposons de prendre en compte la notion de *contexte* modélisé par des relations contextuelles émanant, à la fois, de la structure et de la sémantique du texte.

1 Introduction

L'annotation sémantique est devenue l'une des approches privilégiées par les travaux sur le web sémantique. Les travaux visant à extraire semi-automatiquement ces annotations, plus particulièrement à partir de textes, ont connu ces dernières années une avancée importante. Dans ce contexte, des outils de traitement automatique de la langue naturelle (TALN) sont proposés. Ces outils reposent en général sur des méthodes linguistiques telles que la projection de patrons morpho-syntaxiques ou des méthodes statistiques (fréquence d'apparition). Les méthodes de TALN peuvent être semi-automatiques (l'intervention de l'expert du domaine est alors requise) ou automatiques (dans ce cas, les approches proposées requièrent une certaine spécialisation dans un domaine particulier (Aussenac-Gilles et al., 2006)). Les approches utilisées jusqu'à présent reposent en général sur l'extraction de termes, certaines permettent également l'extraction de relations entre ces termes, mais en ignorant en général le contexte de leur apparition.

Dans le cadre de cette problématique, nous proposons une approche de modélisation, d'extraction et d'exploitation des annotations, qui prenne en compte leurs contextes. La limite observée, concernant les approches d'extraction des termes pour l'annotation, a été notre principale motivation pour offrir des annotations qui représentent au mieux le contenu d'un document. Nous considérons l'annotation sémantique d'un document comme une image par un annotateur (humain ou programme) du contenu de ce document. Cette annotation sémantique doit être exploitable par la machine et de la qualité de cette image dépend son exploitation par l'application visée. Ce travail s'inscrit dans le cadre du projet *SEVENPRO* qui a comme objectif de développer, en reposant sur des technologies et des outils qui aident à la fouille de connaissances sur un produit, des corpus de textes multimédia et sur la réalité virtuelle 3D enrichie sémantiquement.

Tout d'abord, dans la section 2, nous allons analyser quelques travaux sur l'extraction des annotations à partir du texte. Puis dans la section 3, nous aborderons notre proposition sur la modélisation de la notion du contexte. Dans la section 4, nous proposerons notre approche