Extraction et validation par croisement des relations d'une ontologie de domaine

Lobna Karoui*, **
* Laboratoire Epitech de Recherche en Informatique Appliquée "L.E.R.I.A", 24 rue Pasteur 94270 Le Kremlin Bicêtre, France
**Supelec, Plateau de Moulon, Gif-sur-Yvette, France
Lobna.Karoui@supelec.fr

1 Introduction

Face à de grandes quantités de documents web, notre objectif est d'extraire et de valider semi-automatiquement des relations d'un domaine. Dans l'état de l'art, l'extraction des relations a été faite soit par une approche statistique, une approche linguistique ou une approche hybride. De plus, l'intérêt a été toujours porté sur un voire deux types de relations. A contrario, notre objectif est d'extraire des relations de différents types en combinant des analyses de textes et en considérant les caractéristiques des mots. Dans cet article, nous avons défini un algorithme contextuel de découverte de relations qui combine différentes analyses (lexicale, syntaxique et statistique) pour définir des processus complémentaires qui assurent l'extraction de relations variées et pertinentes. Notre algorithme établit des opérations de croisements entre analyses afin de pouvoir valider certaines relations. Les relations valides, comme celles invalides, seront présentées à l'expert du domaine mais séparément.

2 La découverte des relations

La notion de contexte. Pour l'extraction des relations, nous souhaitons trouver les mots qui sont reliés au mot étudié. Donc, nous cherchons des contextes qui contiennent ces mots reliés. Pour cela, nous avons défini différents contextes et nous les avons catégorisés en quatre types: le contexte structurel, le contexte linguistique (centré autour du verbe, globalement syntaxique et lexical), le contexte documentaire (paragraphe) et le contexte fenêtre (avec un degré de proximité). Notre approche utilise toutes ces analyses afin d'extraire de nouvelles relations (en plus de celles existantes dans la hiérarchie) et de les valider automatiquement.
L'algorithme contextuel de découverte des relations. Il applique différents types d'analyses pour extraire et évaluer les relations. Il dépend de certains paramètres comme le degré de confiance (DC), NO est le pourcentage d'occurrences de mots dans le corpus (NO) et FN est la fréquence normalisée des mots dans le corpus (FN). Ces paramètres sont utilisés lors du filtre statistique ainsi que la validation. Le DC doit être défini par l'utilisateur vu qu'il explique sa confiance en l'application. Par contre, NO et FN peuvent être définis soit par l'expert du domaine, soit par le système en les déduisant de la valeur de DC ou par défaut (valeur définie par le concepteur du système). Dans le cas où le système est utilisé pour calculer les valeurs de NO et FN, si la valeur de DC est supérieure à 50% leurs valeurs (par défaut) seront maintenues, sinon elles seront multipliées par deux. Notre algorithme catégorise quatre types de relations extraites : valides, invalides, déduites et étiquetées. Une relation valide est celle qui est récupérée après une opération de croisement entre analyses. Une relation invalide est celle qui n'a pas été retrouvé dans deux analyses.