Découverte et extraction d'arguments de relations n-aires corrélés dans les textes

Soumia Lilia Berrahou*,**, Patrice Buche**, Juliette Dibie***, Mathieu Roche*,****

*LIRMM - 860, rue de Saint Priest, 34095 Montpellier, France
berrahou@lirmm.fr,
**INRA - UMR IATE - 2, place Pierre Viala, 34060 Montpellier, France
***AgroParisTech/INRA - UMR MIA - Université Paris-Saclay, 75005 Paris, France
****CIRAD - UMR TETIS - 500, rue J.F. Breton, 34093 Montpellier, France

Résumé. Dans cet article, nous présentons une méthode hybride combinant des approches de fouille de données et des analyses syntaxiques afin de découvrir et extraire automatiquement des informations dans les textes. Ces informations sont modélisées sous forme de relations n-aires représentées dans une Ressource Termino-Ontologique (RTO). La relation n-aire relie un objet étudié (e.g. un emballage) à ses caractéristiques sous forme d'arguments (e.g. son épaisseur). Dans les textes, les arguments de l'objet étudié sont quantitatifs, associés à leurs attributs, une valeur numérique et une unité de mesure, à extraire pour peupler l'ontologie de nouvelles instances. La méthode proposée repose sur la découverte de relations implicites d'expression des arguments dans les textes en utilisant les motifs et règles séquentiels puis, sur l'intégration de relations syntaxiques d'intérêt dans les motifs découverts afin de construire des patrons linguistiques d'identification d'arguments corrélés. Les expérimentations ont été menées sur un corpus du domaine des emballages et consistent à extraire les résultats expérimentaux de perméabilités des emballages alimentaires.

1 Introduction

Les documents disponibles à partir de bibliothèques spécialisées en ligne, sont une source d'information précieuse à exploiter et analyser par les experts du domaine pour, par exemple, paramétrer des modèles d'aide à la décision (Guillard et al., 2015). Le nombre d'articles publiés et disponibles en ligne est toujours grandissant. Aujourd'hui, le défi n'est pas de trouver l'information mais d'être en mesure de l'identifier et l'extraire automatiquement, notamment dans la perspective du développement de l'open access, en prenant en compte la complexité des données textuelles. En effet, identifier et extraire l'information pertinente se révèlent être des tâches complexes car la grande majorité des documents collectés est, en général, partagée en langage naturel. Le langage naturel, du fait de sa richesse et de sa variété est souvent difficile à appréhender. Un même terme revêt souvent plusieurs significations, une même information peut s'exprimer de multiples manières, souvent implicitement, générant des ambiguïtés difficiles à cerner automatiquement par les machines.
Les travaux présentés dans cet article s'inscrivent dans la problématique d'identification et