Gestion de l'incertitude dans le cadre d'une extraction des connaissances à partir de texte

Fadhela Kerdjoudj[*,**] Olivier Curé[*]

[*]Université Paris-Est Marne-La-Vallée, LIGM, CNRS UMR 8049, France
fadhela.kerdjoudj@univ-mlv.fr, ocure@univ-mlv.fr
[**]GEOLSemantics 12 rue Raspail, 94250, Gentilly

1 Contexte

La recrudescence des documents textuels disponibles sur le web incite de plus en plus travaux à l'exploitation de ces données de manières automatiques. Pour faire interagir ces données entre elles de manière efficace, il faut développer des moyens basés non seulement sur la ressemblance syntaxique mais également sur la correspondance sémantique.

GEOLSemantics est une entreprise qui propose une solution logicielle de traitement linguistique basée sur une analyse linguistique profonde. Le but est d'extraire automatiquement, d'un ensemble de textes, des connaissances structurées, localisées dans le temps et l'espace. Pour représenter ces connaissances, nous avons opté pour les technologies du web sémantique. Nous représentons nos extractions sous forme de triplets RDF et exploitons une ontologie pour apporter de la cohérence. Cette approche permet de relier les résultats de nos extractions aux connaissances du Linked Open Data, tels que Dbpedia et Geonames.

Lors de l'analyse linguistique, il arrive que l'information traitée contienne des imperfections. Dans notre travail, nous intéressons à l'incertitude. Notre première contribution porte sur une catégorisation de l'incertitude lors des différentes phases d'extraction. Notre seconde contribution se situe au niveau de la représentation de l'incertitude dans le graphe RDF.

2 Acquisition de l'information avec incertitude

L'analyse des textes comporte plusieurs étapes distinctes allant du simple découpage du texte en mots à la représentation de son contenu. Parmi ces étapes, nous retrouvons : *(i) l'analyse syntaxique*, il s'agit de la mise en évidence des structures d'agencement des catégories grammaticales, afin d'en découvrir les relations formelles ou fonctionnelles. *(ii) l'analyse sémantique*, l'objectif principal de cette analyse est de déterminer le sens des mots des phrases. *(iii) l'extraction de connaissances* permet de mettre en évidence des entités nommées et des relations relatives à un concept particulier. Grâce à des déclencheurs qui indiquent qu'une relation relative à un concept peut être présente et extraite. Un déclencheur correspond généralement à un concept présent dans l'ontologie, ce qui permet de guider la règle d'extraction par la suite. *(iv) la mise en cohérence* permet de consolider les connaissances extraites notamment le regroupement des entités nommées, la résolution des dates relatives. Cette étape peut être