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

suivi par un enrichissement à partir des données du Linked Open Data.
Cependant, la fiabilité de l'information est très souvent remise en cause. Notre démarche est de considérer le cycle de vie de la connaissance depuis son acquisition jusqu'à son stockage dans la base de connaissances pour cela, nous identifions trois catégories :
Pré-extraction de la connaissance : il s'agira lors de cette étape de considérer les modalités de publication de l'information à savoir : la date et le lieu de publication, la fiabilité accordée à la source, qu'il s'agisse de l'auteur ou de l'organisme de publication...
Pendant l'extraction de la connaissance : l'incertitude pourra concerner aussi bien l'information véhiculée que la règle d'extraction à appliquer.
Post-extraction de la connaissance : l'incertitude peut intervenir au niveau des règles de mise en cohérence ou bien au niveau du choix de la base de référence.

Le formalisme de représentation de connaissances choisie est le RDF tout en nous basons sur une ontologie développée pour prendre en compte les concepts relatifs à un domaine particulier. Notre approche consiste à considérer l'incertitude comme une connaissance à part entière telle que le décrit l'ontologie suivante.
La classe *Uncertainty*, nous permet de modéliser l'incertitude. Elle est décrite par trois propriétés : *weight* : une propriété littérale pour quantifier l'incertitude identifiée, *hasUncertainProp* : une propriété objet qui servira d'intermédiaire entre le domaine initial de la propriété et la propriété en question, *isUncertain* : propriété objet qui aura pour co-domain le top-concept, cela veut dire que tout concept de l'ontologie pourra être visé par une incertitude. Cette ontologie est indépendante de tout domaine d'application. Dès lors, elle peut être ajoutée à toute autre ontologie voulant prendre en compte l'incertitude.

3 Conclusion et perspectives

Dans cet article nous nous intéressons au traitement de l'information incertaine dans le cadre d'une extraction de connaissances à partir de texte. Le traitement repose sur les technologies du web sémantique pour permettre de faire le lien avec les données du Linked Open Data. Notre démarche consiste à identifier les différentes situations où une incertitude remettant en cause la validité de l'information peut subsister. Nous proposons une ontologie pour modéliser l'information incertaine et la représenter au format RDF.
Nous travaillons actuellement sur développement d'un ensemble de patterns pouvant faciliter l'interrogation du graphe RDF prenant en compte notre représentation de l'incertitude. Nous prévoyons par la suite de développer un raisonneur basé sur le formalisme des logiques possibilistes afin de permettre l'inférence sur les données incertaines.

Summary

The knowledge representation area needs some methods that allow to detect and handle uncertainty. Indeed, a lot of text hold information whose the veracity can be called into question. These information should be managed efficiently in order to represent the knowledge in an explicit way. As first step, we have identified the different forms of uncertainty during a knowledge extraction process, then we have introduce an RDF representation for these kind of knowledge based on an ontologie that we developped for this issue.