Construction d'ontologie à partir de corpus de textes

Rokia Bendaoud *, Yannick Toussaint *
Amedeo Napoli *

* LORIA - Campus scientifique BP 239
54506 Vandoeuvre-Lès-Nancy, CEDEX.
{bendaoud,napoli,yannick}@loria.fr

Résumé. Cet article présente une méthode semi-automatique de construction d'ontologie à partir de corpus de textes sur un domaine spécifique. Cette méthode repose en premier lieu sur un analyseur syntaxique partiel et robuste des textes, et en second lieu, sur l'utilisation de l'analyse formelle de concepts "FCA" pour la construction de classes d'objets en un treillis de Galois. La construction de l'ontologie, c'est à dire d'une hiérarchie de concepts et d'instances, est réalisée par une transformation formelle de la structure du treillis. Cette méthode s'applique dans le domaine de l'astronomie.

1 Introduction

Une ontologie est une structure formelle dans laquelle les concepts d'un domaine et les relations entre ces concepts sont définis (Gruber (1993)). Notre ontologie porte sur l'astronomie : dans leurs articles scientifiques, les astronomes identifient manuellement les caractéristiques des objets célestes, afin de les associer ensuite à une catégorie (galaxie, étoile, ...). Les catégories sont pré-définies et l'astronome détermine la classe correspondant le mieux à l'objet étudié. Cette classification a permis de catégoriser 3.751.128 objets célestes. Pourtant, il reste encore des milliards d'objets à classifier et à caractériser de la manière la plus exhaustive possible. L'utilisation des articles scientifiques, très facilement accessibles sous format électronique, permettent de répondre à ces attentes.

Nous proposons une méthode semi-automatique de construction d'une ontologie sur le domaine de l'astronomie. Les concepts de l'ontologie sont des classes dont les instances sont les objets célestes. Les propriétés de chaque classe sont partagées par toutes ses instances. Ces propriétés sont extraites automatiquement des textes par un analyseur syntaxique partiel et robuste "Enju" de Miyao et Tsujii (2005). Objets et propriétés sont classés dans un treillis de Galois selon l'analyse formelle des concepts : FCA présentée dans Ganter (1999). Le résultat de cette méthode est fourni aux astronomes afin d'étiqueter chaque classe d'après les propriétés partagées par les instances de la classe.

Notre méthode présente plusieurs avantages :

– elle peut être appliquée quelque soit le corpus de textes et le domaine spécifique sur lequel elle est utilisée,
– elle est formalisée par la FCA,
– elle est rapide comparée à une ontologie construite manuellement,
– et elle permet d'enrichir l'ontologie résultante par la mise à jour du corpus de textes.

2 Construction de l'ontologie à partir des textes

Notre méthode de construction d'ontologie s'établit à partir d'un corpus de textes. Nous choisissons de caractériser les objets célestes présents dans les textes par les verbes avec lesquels ils apparaissent en tant que sujet ou en tant que complément. Les verbes en effet, nous permettent de définir la nature des objets. Par exemple, tous les objets ne peuvent pas être sujet du verbe "émettre" : un objet émetteur peut être une étoile mais pas une planète.

Tout d'abord, nous extrayons les paires (sujet,verbe) et (complément,verbe) qui représentent l'entrée de la FCA. Ensuite un treillis de Galois est construit avec le context formel $\mathbb{K}= (\mathtt{G}, \mathtt{M}, \mathtt{I})$ tel que : G l'ensemble des objets célestes, M l'ensemble des verbes (propriétés), et il existe une relation I (g,m) ssi : l'objet g est sujet ou complément du verbe m. De là, le treillis est transformé en une ontologie de concepts : les concepts sont représentés par les propriétés (intensions) du treillis et les instances par les objets célestes (extensions) du treillis. Le treillis définit ainsi l'ordre partiel des concepts, d'après l'ensemble des propriétés qu'ils partagent. Enfin, chaque classe d'objets est étiquetée par les experts du domaine.

Cette méthode non supervisée propose aux astronomes une classification des objets célestes pour éviter un "goulot d'étranglement" dans l'acquisition des connaissances. Elle ouvre deux perspectives. D'une part, une analyse plus fine des résultats de l'analyseur syntaxique permettrait d'utiliser des patrons syntaxiques plus précis - au lieu du marqueur général "complément" préciser si c'est un complément d'objet direct, un complément de lieu, etc - ainsi que d'autres types de marqueurs - non seulement les sujets et les compléments sont pris en compte mais aussi les adjectifs, les adverbes, etc. D'autre part, afin de tenir compte des propriétés multivaluées pour obtenir de meilleures classes résultantes, une relation plus riche que la relation binaire dans la FCA devrait être envisagée.

Références

Ganter, B. (1999). *Formal Concept Analysis - Mathematical Foundations*. Springer Verlag.

Gruber, T. (1993). Toward principales for the design of ontologies used for knowledge sharing. *Formal Analysis in Concepttual Analysis and Knowledge Representation.*

Miyao, Y. et J. Tsujii (2005). Probabilistic disambiguation models for wide-coverage hpsg parsing. In *Proceedings of ACL-2005*, pp. 83–90.

Summary

This paper presents a semi-automatic method of building ontology from a textual corpus on a specific domain. This method is based, on the first hand, on a robust and partial syntactic parser and, on the other hand, the use of formal concept analysis for the construction of object class with a Galois lattice. The construction of the ontology from concepts and instances hierarchization is effected in a formal transformation of the lattice structure. The application domain of this method is astronomy.