Apport des outils de TAL à la construction d'ontologies : propositions au sein de la plateforme DaFOE *

Jean Charlet[1,2], Sylvie Szulman[3], Nathalie Aussenac-Gilles[4], Adeline Nazarenko[3], Nathalie Hernandez[4], Nadia Nadah[5], Éric Sardet[6], Jean Delahousse[7], Guy Pierra[6]
(1) INSERM UMR_S 872, Eq. 20, Paris
(2) Université Pierre et Marie Curie ; AP-HP, Paris
(3) LIPN - UMR 7030, Université Paris 13 - CNRS
(4) CNRS/IRIT et Université de Toulouse
(5) Heudiasyc CNRS/UMR 6599, Université de Technologie de Compiègne
(6) LISI-ENSMA et CRITT-Informatique, Poitiers
(7) MONDECA, Paris
Jean.Charlet@spim.jussieu.fr, Sylvie.Szulman@lipn.univ-paris13.fr

Résumé. La construction d'ontologie à partir de textes fait l'objet d'études depuis plusieurs années dans le domaine de l'ingénierie des ontologies. Un cadre méthodologique en quatre étapes (constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie) est commun à la plupart des méthodes de construction d'ontologies à partir de textes. S'il existe plusieurs plateformes de traitement automatique de la langue (TAL) permettant d'analyser automatiquement les corpus et de les annoter tant du point de vue syntaxique que statistique, il n'existe actuellement aucune procédure généralement acceptée, ni a fortiori aucun ensemble cohérent d'outils supports, permettant de concevoir de façon progressive, explicite et traçable une ontologie de domaine à partir d'un ensemble de ressources informationnelles relevant de ce domaine. Le but de ce court article est de présenter les propositions développées, au sein du projet ANR DaFOE 4app, pour favoriser l'émergence d'un tel ensemble d'outils.

Abstract. The concept of ontologies, appeared in the nineties, constitute a key point to represent and share the meaning carried out by formal symbols. Thus, the building of such an ontology is quite difficult. A way to do so is to use preexistent elements (textual corpus, taxonomies, norms or other ontologies) and operate them as a basis to define the ontology field. However, there is neither accepted process nor set of tools to progressively built ontologies from the available resources in a traceable and explicit way. We report in this paper several propositions developed within the framework of the ANR DaFOE4App project to support emergence of such tools.

Mots-clés : Ontologie, construction d'ontologie, TALN.

Keywords: Ontology, Ontology building, NLP.

∗. Ce travail bénéficie d'un financement ANR (2006 TLOG 10). Nous remercions l'ensemble des partenaires du projet qui ont contribué à cette réflexion.

1 Motivation

Depuis son émergence, au début des années 1990, dans les recherches en modélisation de connaissances, la notion d'ontologie s'est rapidement diffusée dans un grand nombre de domaines de recherche en informatique. Compte tenu du caractère très prometteur de cette notion, de nombreux travaux ont visés à permettre son utilisation dans des domaines aussi divers que le traitement automatique de la langue naturelle, la recherche d'information, le commerce électronique, le web sémantique, la spécification des composants logiciels et l'intégration de système d'information.

L'efficacité de toutes ces approches présuppose néanmoins l'existence d'une ontologie de domaine susceptible d'être développée, ou d'être mise en œuvre, au sein de l'application cible. Or la conception d'une telle ontologie s'avère particulièrement difficile, surtout si l'on souhaite qu'elle fasse l'objet de consensus dans une communauté assez large. Un moyen très largement utilisé pour atteindre cet objectif est de partir d'éléments préexistants dans le domaine : corpus textuels, taxonomies, normes ou fragments d'ontologie préexistants, et de les exploiter comme base pour définir progressivement l'ontologie du domaine. La construction d'ontologie à partir de textes fait l'objet d'études depuis plusieurs années dans le domaine de l'ingénierie des ontologies. Un cadre méthodologique en quatre étapes (constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie) est commun à la plupart des méthodes de construction d'ontologies à partir de textes (TERMINAE [1] (Aussenac-Gilles *et al.*, 2000), Text2Onto (Cimiano & Volker, 2005)). Ces méthodes sont implémentées dans des outils qui se distinguent par leur approche de la phase de conceptualisation plus ou moins automatique (Mondary *et al.*, 2008). Cependant s'il existe des outils largement utilisés, tels que Protégé, pour représenter formellement une ontologie supposée déjà conçue, et s'il existe également plusieurs plateformes de traitement automatique de la langue (TAL) permettant d'analyser automatiquement les corpus et de les annoter tant du point de vue syntaxique que statistique, il n'existe actuellement aucune procédure généralement acceptée, ni a fortiori aucun ensemble cohérent d'outils supports, permettant de concevoir de façon progressive, explicite et traçable une ontologie de domaine à partir d'un ensemble de ressources informationnelle relevant de ce domaine.

2 Méthodes du TAL et spécifications pour DAFOE

Les logiciels de TAL permettent d'extraire des textes des éléments de connaissances à représenter dans une ontologie. Les expériences en la matière sont très diverses et après avoir fait un inventaire des outils existants que nous ne détaillons pas ici [2], nous avons analysé les fonctionnalités les plus importantes [3] et ce qu'elles peuvent apporter à la construction d'ontologies dans le contexte de DaFOE.

Ainsi, les principales fonctionnalités prises en compte dans DaFOE sont : l'extraction de candidats termes, la pondération des candidats termes, la validation des éléments terminologiques, la normalisation des termes, l'extraction de relations conceptuelles et la gestion du multilinguisme.

1. http://www-lipn.univ-paris13.fr/~szulman/logi/index.html
2. Le lecteur pourra se référer aux ouvrages de (Maedche, 2002) et (Buitelaar & Cimiano, 2007)
3. A noter que le problème important de l'extraction des entités nommées n'est pas une priorité pour DaFOE qui met l'accent sur la construction d'ontologie plus que sur son peuplement, même si des extensions sont envisagées.

3 Modèle de données

Un cadre méthodologique a été élaboré durant la définition de la plateforme. Il a été utilisé de deux façons, à savoir comme cadre permettant d'avoir une description commune des processus mis en jeu en même temps que modèle évoluant pour être à même de tenir compte des desiderata de tous les partenaires. Ainsi, la plateforme a différents niveaux d'entrées, correspondant aux différentes ressources, et différents niveaux de sortie correspondant à des produits de plus en plus élaborés (1) des réseaux terminologiques s'organisant durant l'analyse des données (« Réseaux termino-ontologiques » et « Modèle conceptuel »), (2) un niveau termino-ontologique où les concepts sont organisés et (3) un niveau où l'ontologie est formalisée (« Ontologie formelle »). Le modèle de données de la plateforme DaFOE suit le cadre méthodologique ainsi décrit et se décompose en quatre couches (Charlet *et al.*, 2008).

4 Conclusion et perspectives

Pour valider l'architecture de méta-modélisation proposée, un premier prototype est en cours de réalisation par le LISI. Il s'agit d'une application Java s'appuyant sur le SGBD PostgreSQL. Le modèle de données de DaFOE, et en corollaire les fonctionalités, est particulièrement développé sur les niveaux 1 et 2 du schéma précédent car c'est à ces niveaux que se concentrent un certain nombre de difficultés sur la construction d'ontologies selon les méthodes proposées.

Références

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2000). Revisiting ontology design : a methodology based on corpus analysis. In R. DIENG & O. CORBY, Eds., *Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000)*, LNAI 1937, p. 172–188 : Springer-Verlag.

P. BUITELAAR & P. CIMIANO, Eds. (2007). *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. IOS Press.

CHARLET J., SZULMAN S., PIERRA G., NADAH N., TEGUIAK H. V., AUSSENAC-GILLES N. & NAZARENKO A. (2008). Dafoe : A multimodel and multimethod platform for building domain ontologies. In D. BENSLIMANE, Ed., *2^e Journées Francophones sur les Ontologies*, Lyon, France : ACM.

CIMIANO P. & VOLKER J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In A. MONTOYO, R. MUNOZ & E. METAIS, Eds., *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, p. 227–238, Alicante, Spain : Springer.

MAEDCHE A. (2002). *Ontology learning for the Semantic Web*. Kluwer Academic Publisher.

MONDARY T., DESPRES S., NAZARENKO A. & SZULMAN S. (2008). Construction d'ontologies à partir de textes : la phase de conceptualisation. In Y. PRIÉ, Ed., *19^es Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 87–98.