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

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 informationnelle 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 en nous focalisant sur les réflexions menées autour des outils de TAL.

1 Introduction

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. Définie comme la représentation formelle, et consensuelle au sein d'une communauté d'utilisateurs, des concepts propres à un domaine et des relations qui les relient Gruber (1993), la notion d'ontologie apparaît ainsi comme la pierre philosophale permettant de représenter explicitement, et de partager,

la signification dénotée par des symboles formels. 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), Aussenac-Gilles *et al.* (2008), 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. Le but de cet article est ainsi de présenter les propositions développées, au sein du projet ANR DaFOE 4app, en mettant l'accent sur l'utilisation des résultats des outils de TAL.

Nous présentons, section 2, l'état de l'art dans le domaine des outils de conception d'ontologie et identifions les besoins à satisfaire. La section 3 propose une brève description des étapes de représentation par lesquelles peut passer une démarche de conception d'ontologie puis spécifie, pour chaque niveau de représentation, la structure de modélisation proposée. Enfin, en conclusion, nous donnons l'état de développement de la plateforme.

2 Méthodes du TAL et spécifications pour DAFOE

L'état de l'art effectué à ce sujet a été conduit dans le cadre de l'étude des besoins de la plateforme DaFOE pour expliciter les besoins des utilisateurs au regard des outils existants. Nous avons analysé trois types d'outils (les logiciels de traitement automatique des langues (TAL), les plateformes de construction d'ontologies à partir de texte et

1. `http://www-lipn.univ-paris13.fr/~szulman/logi/index.html`

les éditeurs d'ontologies). Les problèmes d'intégration et d'interopérabilité entre outils sont rapidement apparu, ce qui justifiait le projet de création d'une nouvelle plateforme.

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 nous ne faisons pas ici un inventaire des outils existants [2]. Nous analysons plutôt ce que les fonctionnalités les plus importantes [3] peuvent apporter à la construction d'ontologies et nous justifions les objectifs de DaFOE.

2.1 Extraction de candidats termes

L'analyse terminologique d'un corpus permet d'identifier des syntagmes qui semblent avoir un fonctionnement terminologique, c'est-à-dire qui relèvent d'un vocabulaire spécialisé doté d'une sémantique relativement stable et consensuelle au sein d'une situation de communication déterminée. Dans le processus de construction d'ontologie, l'extraction terminologique permet de repérer les concepts qui sont mentionnés par le corpus. Elle sert donc au repérage du vocabulaire conceptuel.

Les outils d'extraction existants sont nombreux et variés. Dans la mesure où l'on manque de recul et de méthodes pour évaluer et comparer leurs approches, nous ne faisons pas le choix exclusif d'un extracteur contre un autre : nous prévoyons de pouvoir intégrer dans la plateforme DaFOE des résultats de différents extracteurs.

2.2 Pondération des candidats termes

Une fois recensés les candidats termes du corpus, il est important de leur associer un poids et de les trier. Cela permet de focaliser le travail de construction d'ontologie sur un petit nombre de termes et de commencer par les éléments les plus signifiants. En effet, l'extraction produit généralement des listes très importantes de candidats termes (jusqu'à quelques dizaines de milliers de candidats pour un corpus de 100 000 mots).

Certains extracteurs intègrent la pondération dans leur résultat mais il paraît cependant plus pertinent de dissocier les deux fonctionnalités dans DaFOE par souci de modularité. Cela permet de combiner différents outils de TAL et de choisir les critères de saillance en fonction des applications visées.

2.3 Validation des éléments terminologiques

La pondération des termes peut servir à préparer la validation des termes par un expert mais il doit néanmoins retravailler manuellement les résultats. Dans certains cas, on se contente d'explorer les résultats sans passer par une phase systématique de validation. Comme ces tâches sont longues et subjectives, il est important de penser la présentation des résultats, les fonctionnalités et l'ergonomie des interfaces de manière à accélérer

2. Le lecteur pourra se référer aux ouvrages de Maedche Maedche (2002), Buitelaar *et al.* Buitelaar & Cimiano (2007) et Cimiano 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.

le travail de validation, faciliter la navigation et en assurer la cohérence. DaFOE devra tenir compte des possibilités que suscitent les interfaces existantes comme le traitement des candidats termes à travers la navigation *via* les têtes ou les expansions que permet un outil comme TERMONTO.

2.4 Normalisation des termes

Les termes étant des entités textuelles, ils subissent différentes modifications de surface qui nuisent à leur repérage en corpus et à leur visualisation. Il est donc important, ne serait-ce que pour mesurer leur fréquence et pour les visualiser, de pouvoir regrouper les différentes formes d'un même terme sous une forme canonique commune (une sorte de "lemme" de terme). Cette normalisation n'est pas toujours prise en charge par les extracteurs et elle peut l'être de différentes manières. Dans DaFOE, nous ne faisons pas de présupposition sur la nature de la normalisation mais nous prévoyons de différencier la forme canonique de ses variantes.

2.5 Extraction de relations conceptuelles

Les textes expriment également des informations sur les relations sémantiques que les termes entretiennent en eux. Si l'on considère que les termes représentent les concepts de l'ontologie à construire, les relations sémantiques qu'ils entretiennent peuvent être considérées comme le reflet de relations conceptuelles. Repérer ces relations aide à structurer l'ontologie. L'extraction des relations terminologiques est une tâche complexe du fait de la diversité des relations sémantiques à prendre en compte et de la diversité des méthodes mises en œuvre. La mise en relation des termes est une des étapes sur laquelle il est prévu de travailler dans le cadre de DaFOE. L'idée est de construire un greffon qui permette d'appeler des méthodes classiques à partir de règles contextuelles d'extraction – *e.g.* "patrons" qui sont la plupart du temps spécifiques des méthodes.

2.6 Regroupement de classes sémantiques

De nombreuses recherches visent à construire des classes sémantiques de mots à partir de l'analyse de la distribution des mots en corpus. L'approche est d'autant plus fructueuse que le corpus de départ est plus homogène. Les méthodes diffèrent par la définition du contexte sur lequel elles s'appuient (simple fenêtre de mot, contexte syntaxique, etc.), par la mesure de similarité qu'elles prennent en compte et par la démarche de classification qu'elles adoptent (supervisée ou non, nombre de classes visés, etc.). Les approches numériques et symboliques étant complémentaires, le défi consiste à déterminer comment tirer profit de cette complémentarité dans le processus très concret de construction d'ontologies (méthode et outils) tel que le proposera la plateforme DaFOE.

2.7 Gestion du multilinguisme

Il est important de prendre en compte le multilinguisme dans la construction d'ontologies. Deux approches sont possibles : 1) faire le même travail de construction d'ontologies en parallèle en prenant appui sur des corpus relevant du même domaine dans

les différentes langues cibles puis intégrer les ontologies résultantes ou 2) partir de corpus parallèles, créer une première ontologie à partir du sous-corpus d'une langue et trouver des correspondances traductionnelles des termes associés aux concepts de la première ontologie pour enrichir le niveau lexical (on parle alors *localisation*). Dans les deux cas, c'est une tâche lourde et difficile. Dans un premier temps, nous retenons la première solution.

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 ("Ontologie") et (3) un niveau où l'ontologie est formalisée ("Ontologie formelle").

Le modèle de données de la plateforme DaFOE suit le cadre méthodologique. Il se décompose en quatre couches (voir figure 1).

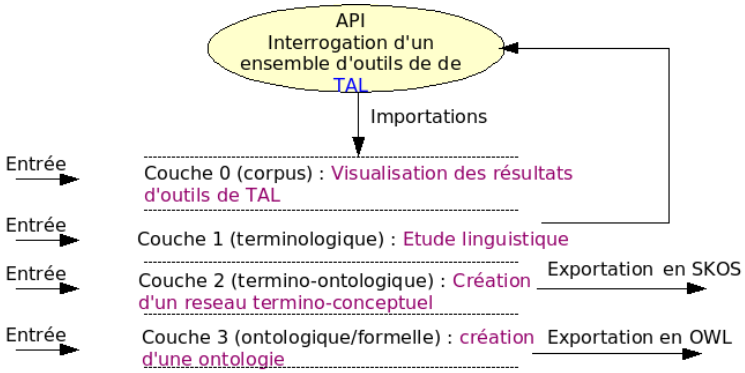

FIGURE 1 – Structure en couches du modèle de données de DaFOE

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. Nous présenterons, durant la conférence IC, la plateforme DaFOE dans laquelle les données d'un outil de TAL seront chargées et nous montrerons comme les différents niveaux de modélisation s'articulent.

Remerciements

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

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.

AUSSENAC-GILLES N., DESPRES S. & SZULMAN S. (2008). *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*, chapter The Terminae Method and Platform for Ontology Engineering from Texts. IOS Press.

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

CIMIANO P. (2007). *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. IOS Press.

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.

GRUBER T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**, 199–220.

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., *19es Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 87–98.