Enrichissement d'ontologies grâce à l'annotation sémantique de pages web

Nathalie Aussenac-Gilles*, Davide Buscaldi**, Catherine Comparot*, Mouna Kamel*

*Université de Toulouse, IRIT, CNRS (UMR 5505), F-31062 Toulouse
{aussenac,comparot,kamel}@irit.fr,
http://www.irit.fr/-Equipe-MELODI
**Université Paris 13 Sorbonne Paris Cité, LIPN, CNRS (UMR 7030), F-93430 Villetaneuse
davide.buscaldi@lipn.univ-paris13.fr

Résumé. Nous présentons une approche pour enrichir automatiquement une ontologie à partir d'un ensemble de pages web structurées. Cette approche s'appuie sur un noyau d'ontologie initial. Son originalité est d'exploiter conjointement la structure des documents et des annotations sémantiques produites à l'aide du noyau d'ontologie pour identifier de nouveaux concepts et des spécialisations de relations qui enrichissent l'ontologie. Nous avons implémenté et évalué ce processus en réalisant une ontologie de plantes à partir de fiches de jardinage.

1 Introduction

L'annotation sémantique de documents du web est une des étapes pour assurer le succès des applications du Web Sémantique en tant que support à un accès partagé et unifié aux connaissances et documents, y compris de domaines spécifiques. En faisant appel à une ontologie, on assure un meilleur partage et une meilleure interprétation de ces annotations. La qualité des annotations sémantiques nécessite donc des ontologies de domaine de qualité. Même si des ressources comme WordNet [1] ou Yago [2] offrent des vocabulaires riches et répondent aux besoins de l'annotation de connaissances générales en langue anglaise, les connaissances précises des ontologies de domaine présentent l'avantage de mieux rendre compte de l'information véhiculée par des documents et du sens des formulations linguistiques.

Dans un processus dual, quand les ontologies sont utilisées pour l'annotation sémantique de collections de textes de domaines particuliers, leur construction à partir de ces textes contribue à ce que l'ontologie couvre mieux les concepts et relations nécessaires à la caractérisation du contenu des documents. L'ingénierie d'ontologies à partir de textes a ainsi obtenu des résultats significatifs ces 10 dernières années (Buitelaar et al., 2005) (Maedche, 2002), mais cette activité reste longue et complexe.

Lorsque l'on veut enrichir ou peupler une ontologie existante, les annotations textuelles peuvent fournir des informations pour trouver des indices linguistiques de relations sémantiques. Une première catégorie de systèmes utilise les annotations linguistiques : par exemple,

1. http ://www.w3.org/TR/wordnet-rdf/
2. http ://www.mpi-inf.mpg.de/yago-naga/yago/

RelExt de Schutz et Buitelaar (2005) exploite catégories et dépendances syntaxiques. Un deuxième ensemble d'approches traite des annotations sémantiques en utilisant l'ontologie à enrichir, comme la méthode proposée dans le projet CrossMarc (Valarakos et al., 2004) , ou celle de Navigli et Velardi (2006).

Notre objectif est similaire à ces travaux : partant d'un noyau d'ontologie en adéquation avec une collection de textes, nous cherchons à l'enrichir par analyse automatique de ces textes. La chaîne de traitements que nous proposons enrichit un noyau d'ontologie avec des relations sémantiques définissant des restrictions de relations existant dans l'ontologie afin de décrire des concepts précis. Notre approche est originale au sens où nous introduisons dans ce processus deux types de connaissances sur le texte, jusque là peu utilisés conjointement : des connaissances de niveau sémantique, à savoir les concepts déjà reconnus dans le texte, et des connaissances de niveau discursif, liées à l'architecture du texte au sens de Virbel et Luc (2001). Nous avons évalué ce processus en réalisant une ontologie de plantes à partir de fiches de jardinage dont la structure est explicitée par des étiquettes XML.

2 Annotation Sémantique et Processus d'Enrichissement

2.1 Principes généraux

Nous proposons d'enrichir une ontologie de domaine par des restrictions de relations, et ce à partir d'une collection de documents structurés et d'un noyau d'ontologie existant. Les documents de cette collection possèdent les propriétés d'être conformes à un même modèle (DTD ou schéma XML) représentant une succession de champs, de traiter d'entités d'un même type désigné par concept principal, chacune des entités, appelée concept pivot, étant décrite par un document de la collection. Ces collections de documents sont très fréquentes sur le web. Le noyau d'ontologie est construit à partir d'une analyse manuelle de la structure de ce type de document. Il est composé du concept principal relié aux concepts racines de hiérarchies de concepts issues de l'exploitation des valeurs des champs, la sémantique de ces relations étant fournie par le nom du champ. Les concepts pivots deviennent alors des sous-concepts du concept principal. Par exemple, dans un site web de jardinage, le concept pivot est celui de plante, chaque page décrit une plante spécifique (concept pivot de la page) à l'aide de concepts relatifs à la nature des sols, à l'entretien ou aux parties de la plante.

Le processus d'enrichissement vise à décrire précisément les concepts pivots à l'aide de propriétés spécifiques. Elles sont exprimées par des relations qui restreignent les relations du noyau d'ontologie ayant pour domaine le concept principal en spécialisant leur co-domaine. Pour cela, le texte et les balises des champs du document sont analysés. Identifier la correspondance entre les balises des champs et les propriétés du *concept pivot* est ce que nous appelons par la suite la *sémantisation* de la structure des documents, un processus à deux étapes :

1. affecter un rôle aux champs de cette structure. Après lecture des fiches, l'ontologue définit F_C , l'ensemble des champs porteurs du concept pivot et F_R , l'ensemble des champs porteurs de relations. Ces derniers contiennent a priori au moins une relation entre le concept pivot et un concept existant dans le noyau d'ontologie et dont on s'attend à trouver une trace linguistique dans le texte présent dans ce champ.
2. formuler les concepts et relations associées à chaque champ. C'est là une originalité de la démarche. La recherche des relations ne s'appuie pas sur la notion de patron lexico-

syntaxique tels qu'ils sont définis dans (Auger et Barriere, 2008). Elle repose sur l'expression déclarative des concepts et relations pouvant être reconnus grâce à chaque balise XML et au texte qu'elles encadrent. Cette expression est le fruit d'une interprétation par l'ontologue de la sémantique des documents. Cette approche généralise celle établie pour d'autres types de documents structurés dans (Laignelet et al., 2011).

2.2 Enrichissement de l'ontologie par les relations

Pour identifier les concepts à relier au concept pivot, nous utilisons l'annotation sémantique de ces textes à l'aide des concepts déjà présents dans l'ontologie. Dans une ontologie $O = (C, R)$, C est l'ensemble des concepts du domaine, et R, l'ensemble des relations liant les concepts. Nous définissons alors :

- $T_{f,d} = \{t_1, t_2, ..., t_n\}$ est l'ensemble des termes extraits du champ f d'un document d, $f \in F_C$ ou $f \in F_R$, par le biais de patrons d'extraction.
- $r(c_i, c_j)$, avec c_i et $c_j \in C$, et $r \in R$ définit une relation autre que la subsomption, respectivement de domaine c_i et de co-domaine c_j.
- $porte\_c(f, c)$, avec $c \in C$ et $f \in F_C$, est un prédicat qui indique que le champ f est porteur du concept c.
- $porte\_r(f, r)$, avec $r \in R$ et $f \in F_R$, est un prédicat qui indique que le champ f est porteur de la relation $r(c_i, c_j)$ où un des concepts c_i ou c_j est le concept principal et l'autre est une super-classe d'un des concepts exprimés dans ce champ.
- $a\_label(c, t)$, avec $c \in C$ et t un terme, indique que c a pour étiquette (rdfs :label) t.

L'algorithme d'extraction de relations produit un ensemble de restrictions sur des relations $r(ci, cj) \in R$, où c_i est le concept pivot de d, et c_j spécialise le co-domaine de r.

2.2.1 Annotation des documents avec les concepts

L'algorithme d'annotation de TextViz [3] (Reymonet et al., 2009) est utilisé pour annoter chacun des documents d de la collection avec le noyau d'ontologie. Il s'appuie sur les labels de concepts, sur les termes présents dans le texte et sur une distance sémantique. Soit le concept $c \in C$, d un document, et $f \in F_R \cup F_C$. Le prédicat $annote(c, f, d)$ indique que c annote le champ f de d. TextViz génère cette annotation si un des labels de c apparaît dans le contenu textuel du champ f du document d. Étant donné $T_{f,d}$, les annotations de TextViz sont équivalentes au résultat de la règle : $\forall t \in T_{f,d}$, $\forall c \in C$, si $a\_label(c, t)$ alors $annote(c, f, d)$.

2.2.2 Identification des relations

Le processus d'identification de relations explore, pour chaque champ f de d tel que $f \in F_R$, les annotations de f et le fichier déclarant les relations associées à f. Étant donnés c_p le concept pivot du document d, un concept c_i qui annote le texte du champ f, le processus vise à identifier la relation $r \in R$ qui peut exister entre c_p et c_i. Cette relation correspond à une des relations de l'ontologie telles que le domaine et co-domaine sont des super-classes de c_p et c_i d'une part, et que le champ f est porteur de la relation. Alors la restriction de r à c_p et c_i

3. plug-in de la plate-forme Protégé-OWL développé dans le cadre du projet Dynamo - projet financé par le programme ANR TechLog. `http://www.irit.fr/DYNAMO`

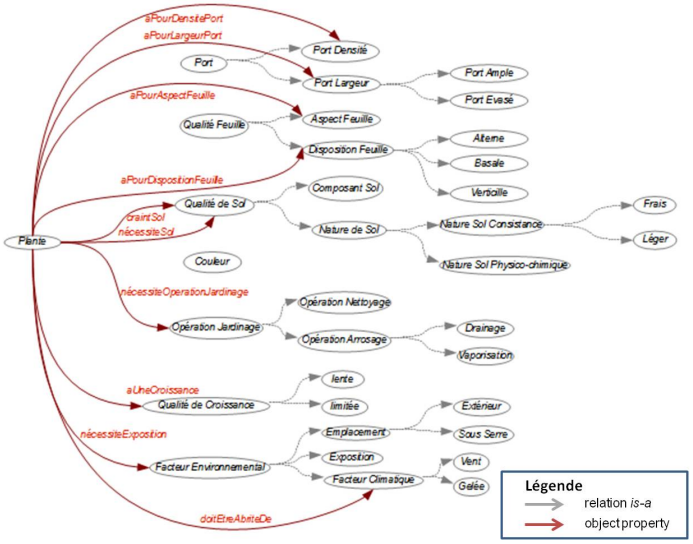

FIG. 1 – *Un extrait du noyau de l'ontologie du jardinage.*

comme domaine et co-domaine est ajoutée à l'ontologie. De façon plus formelle, l'algorithme d'extraction des relations est le suivant : pour $t_i \in T_{f,d}$

pour $c_i \in C_1$ / $a\_label(c_i, t_i)$ et $subsume(c, c_i)$

pour c_p / $est\_pivot(c_p, d)$ (avec $subsume(c_D, c_p)$ où c_D est le concept principal,)

SI $porte\_r(f, r)$ avec $r(c_D, c)$ ALORS

$r(c_p, c_i)$ est ajouté à R ;

fin.

3 Le cas d'étude du jardinage

Dans le projet MOANO [4], une ontologie doit permettre la recherche sémantique d'informations dans un livre de jardinage numérisé [5]. Ce livre, écrit en français, présente une grande collection de plantes et fournit, pour chacune, des informations sur son aspect, des conseils sur son entretien et des soins contre les maladies ou parasites. L'annotation de ce livre doit donc s'appuyer sur une ontologie dédiée aux plantes d'un point de vue non pas de la botanique scientifique, mais du jardinage, qui doit contenir des concepts liés aux plantes, à leur aspect physique, aux conditions de leur développement, ainsi que des conseils pour les entretenir ou prévenir les maladies ou parasites. Pour décrire dans l'ontologie chacune des variétés de plante

4. MOANO est un projet financé par le programme ANR CONTINT, décision ANR-2010-CORD-024-03. Les partenaires sont les laboratoires IRIT, LIUPPA, LIG, LIFL. `http://moano.liuppa.univ-pau.fr`. Nous remercions nos collègues du LIUPPA (M.-N. Bessagnet, A. Royer et C. Sallabery) pour leurs retours lors de l'enrichissement de l'ontologie.

5. http ://www.vilmorin.fr/

du livre, nous n'avons pas pu exploiter directement le livre numérisé à cause de son format. Nous avons choisi un des sites web décrivant des plantes dans une perspective de jardinage [6].

La description des plantes requiert des concepts spécifiques (forme des feuilles, période de floraison, méthode d'entretien, ...) et d'autres concepts (comme saison, couleur ou type de sol). La structure XML des pages web met en valeur des mots clés qui sont à la fois les noms des balises XML et des propriétés de plantes. Pour ces propriétés, la valeur ou les concepts reliés sont exprimés dans le texte balisé. La structure des pages guide ainsi l'identification de relations et de concepts, autres que `Plante`, à définir dans l'ontologie. La figure 1 présente un extrait du noyau d'ontologie ainsi défini.

Dans les textes du site, le champ <Floraison> porte les relations `fleuritEn`, définie entre `Plante` et `Saison`, et `aPourCouleur` qui relie `Plante` à `Couleur`. Dans la fiche du *Gladiolus*, ce champ contient le texte " floraison : printemps ou été ou automne suivant la date de plantation ". Les labels reconnus sont ceux des concepts `Printemps`, `Ete` et `Automne`, sous-classes de `Saison`, et `Plantation`. L'enrichissement consiste à ajouter à l'ontologie `fleuritEn(Gladiolus, Printemps)`, `fleuritEn(Gladiolus, Ete)` et `fleuritEn(Gladiolus, Automne)` en tant que restrictions de la relation `fleuritEn`.

4 Evaluation

Deux ontologues ont enrichi le noyau d'ontologie à partir de 10 documents du site choisis au hasard. A partir de chacun des documents, ils ont défini des concepts spécifiques de plante, et leurs propriétés, représentations que l'on peut considérer comme les résultats attendus du processus automatique. Nous avons ensuite appliqué l'algorithme d'enrichissement à partir de ces 10 documents, puis calculé le rappel et la précision par rapport aux relations retenues par les ontologues. Le Rappel est évalué à 0.73 et la Précision à 0.8. Ces chiffres sont donc prometteurs mais difficiles à interpréter. Si l'on étudie précisément pourquoi des relations ont été mal interprétées ou n'ont pas été trouvées, on identifie 4 causes de problèmes : l'incomplétude d'une ressource (labels de l'ontologie, liste de champs à analyser), la qualité de l'annotation par les concepts (erreurs de l'algorithme d'annotation de TextViz, annotation des noms de champs), la qualité de l'ontologie (signature ambigüe des relations, absence de labels de concepts), une définition incorrecte de la sémantique des champs, et les défauts de l'analyse linguistique (négations ou intervalles numériques non traités). Les problèmes les plus fréquents sont liés à l'analyse linguistique, en particulier à la gestion des négations, que nous sommes en train de mettre en place pour les corriger.

5 Conclusion

Nous avons montré que dans des collections de documents textuels semi-structurées et thématiquement homogènes, où chaque document décrit un concept d'un même type dans toute la collection, l'exploitation de la structure des documents peut être un atout majeur pour automatiser la modélisation du concept décrit dans chaque document. Nous avons ainsi défini une nouvelle approche pour enrichir une ontologie noyau en exploitant la structure des documents de la collection et leur annotation sémantique par les concepts du noyau. Les résultats obtenus

6. l'encyclopédie botanique du site "Jardin !L'encyclopédie" http ://nature.jardin.free.fr/

sur une petite évaluation sont prometteurs. Cette évaluation permet d'identifier les problèmes à prendre en compte pour améliorer les performances globales de cette approche. Nous sommes en train d'intégrer la gestion des problèmes liés à la langue (négations, références etc.) dans le processus d'extraction des relations. Enfin, nous envisageons de réaliser une évaluation plus intensive sur la totalité de la collection des documents du site analysé.

Références

Auger, A. et C. Barriere (2008). Pattern based approaches to semantic relation extraction: a state-of-the-art. *Terminology 14*(1), 1–19. special Issue on Pattern-based Approaches to Semantic Relation Extraction.

Buitelaar, P., P. Cimiano, et B. Magnini (2005). *Ontology Learning From Text: Methods, Evaluation and Applications*. IOS Press.

Laignelet, M., M. Kamel, et N. Aussenac-Gilles (2011). Enrichir la notion de patron par la prise en compte de la structure textuelle - application à la construction d'ontologie. In M. Lafourcade et V. Prince (Eds.), *Traitement Automatique des Langues Naturelles (TALN 2011)*, Volume 2, pp. 267–272. LIRMM (F).

Maedche, A. (2002). *Ontology Learning for the Semantic Web*, Volume 665. The Kluwer International Series in Engineering and Computer Science.

Navigli, R. et P. Velardi (2006). Ontology enrichment through automatic semantic annotation of online glossaries. In S. Staab et V. Svàtek (Eds.), *15th International Conference EKAW 2006*, Volume LNCS 4248, pp. 126–140. Springer.

Reymonet, A., J. Thomas, et N. Aussenac-Gilles (2009). Ontology based information retrieval : an application to automotive diagnosis. In M. Nyberg, E. Frisk, M. Krisander, et J. Aslund (Eds.), *Workshop on Principles of Diagnosis (DX 2009)*, pp. 9–14.

Schutz, A. et P. Buitelaar (2005). Relext: A tool for relation extraction from text in ontology extension. In *4th International Semantic Web Conference (ISWC 2005)*, Volume 3729, pp. 593–606. Springer: Berlin.

Valarakos, A., G. Paliouras, V. Karkaletsis, et V. G. (2004). A name matching algorithm for supporting ontology enrichment. In *Proceedings of SETN*, Volume 3025, pp. 144–156.

Virbel, J. et C. Luc (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum XXIII*(1), 103–123.

Summary

We present a novel approach to automatically enrich an ontology from a collection of structured web pages. This approach is based on an initial kernel ontology. Its original feature is to jointly exploit the document structure and some semantic annotations obtained using the kernel ontology. Both types of knowledge make it possible to identify new concepts and specializations of semantic relations that enrich the ontology. We have implemented this process and evaluated it by building an ontology of plants from gardening forms.