Construction d'ontologies à partir de textes : exploitation des relations verbales

Sylvie Despres[1] et Jérôme Nobécourt[1]

[1] LIM&BIO – EA 3969, Université Paris 13,
sylvie.despres@univ-paris13.fr
j.nobecourt@smbh.univ-paris13.fr

1 Introduction

Les méthodes de construction d'ontologies à partir de textes comportent une phase de conceptualisation [Cimiano, 2006 ; Aussenac-Gilles et *al*., 2008] qui assure la transition du plan du discours au plan ontologique. Dans ce contexte, une ontologie est constituée de concepts, dits ontologiques afin d'éviter toute confusion, organisés dans une hiérarchie avec héritage des propriétés, liés par des relations ontologiques et contraints par des règles et des axiomes. Le passage progressif du texte à l'ontologie prônée par la méthode Terminae est effectué en réalisant des traitements qui relève de la terminologie puis de la modélisation des connaissances et enfin de la représentation des connaissances.

Cette phase de conceptualisation est complexe et difficile à mettre en œuvre manuellement et *a fortiori* automatiquement. Pour pallier ces difficultés un courant de recherche consistant à définir des patrons de conception d'ontologies (ODP) s'est développé et a abouti à la création de bibliothèque d'ODP [NEON Project, 2008]. Dans ce travail, nous nous intéressons aux patrons ontologiques lexico-syntaxiques (LS OPs) qui doivent faciliter la correspondance entre le niveau ontologique et les formulations en langue naturelle. Un LS OPs est une structure linguistique ou un schéma constitué d'un ensemble mots apparaissant dans un ordre précis et qui permet de généraliser et d'extraire certaines conclusions à propos du sens qu'ils expriment. Ces patrons sont conçus pour aider le concepteur dans le choix des propriétés susceptibles de représenter les relations qu'ils ont identifiées [Aguado & al., 2008].

Cette contribution est centrée sur la construction des relations ontologiques liant les concepts de l'ontologie à partir de leurs manifestations *via* des syntagmes verbaux dans les textes du corpus. Ces syntagmes verbaux peuvent traduire des propriétés décrivant l'intension des concepts ou décrire des relations binaires entre ces concepts. La finalité que nous poursuivons est l'automatisation d'une partie du processus de construction des relations ontologiques. L'approche adoptée consiste à extraire les relations verbales sous forme de triplets représentant des relations termino-ontologiques, à regrouper ces triplets à partir des syntagmes verbaux puis à les organiser en fonction des sujets et des compléments qu'ils partagent. On obtient ainsi des classes sémantiques de verbes susceptibles de désigner des relations ontologiques. Enfin nous suggérons de définir des patrons LS OPs afin d'aider l'ontologue dans sa démarche de construction. Dans ce papier, nous abordons essentiellement les

questions posées par une telle approche et nous indiquons quelques éléments relatifs à l'automatisation de la démarche.

2 Rôle des relations verbales dans la phase de conceptualisation

Dans cette partie, nous analysons comment les relations verbales apparaissant dans le texte peuvent contribuer à la construction des relations ontologiques entre les concepts de l'ontologie. Nous nous référons à un exemple dans le domaine de l'accidentologie routière pour illustrer notre propos. Le corpus étudié est constitué d'un ensemble de textes décrivant des scénarios types d'accidents. Un scénario d'accidents est un prototype de déroulement correspondant à un groupe d'accidents présentant des similitudes d'ensemble du point de vue de l'enchaînement des faits et des relations de causalité, dans les différentes phases conduisant à la collision (Brenac & al., 2003). Le corpus est composé de 20 scénarios types. Il contient 4068 mots et 20268 caractères. Les textes décrivant les scénarios ont été rédigés par les chercheurs en accidentologie à des fins de diagnostic de l'insécurité routière. Ils sont écrits en français dans un langage assimilable à un vocabulaire contrôlé. En effet, un soin particulier a été observé dans le choix des mots utilisés pour la rédaction des descriptions associées à chacun des scénarios.

L'ontologie à construire est une ontologie du domaine de l'accidentologie dont la finalité est l'aide à la reconnaissance de procès-verbaux d'accidents. Les concepts ontologiques ont été construits à partir des termes figurant dans les textes et sont organisés hiérarchiquement. Un lien est établi entre les concepts ontologiques et leurs expressions dans le corpus. Une fois les concepts ontologiques établis et structurés hiérarchiquement, l'objectif est de repérer les relations sémantiques liant les termes désignant les concepts ontologiques dans le texte. Nous avons privilégié les relations verbales qui constituent un indice pertinent pour identifier les relations non hiérarchiques entre les concepts du domaine [Sanchez & Moreno, 2008]. Elles traduisent des actions permettant de définir des relations ontologiques entre les concepts établis et mettent également en évidence des propriétés intensionnelles des concepts ontologiques. Dans le domaine de l'accidentologie, elles expriment des liens entre les trois concepts fondamentaux (`Humain`, `Véhicule`, `Environnement`) du domaine. Les exemples relatifs à la situation de conduite (cf. Tab.1) correspondant à deux scénarios (3 et 13) explicitent certaines de ces relations verbales. Il s'agit des verbes « `circuler` » avec les prépositions (`[en]` et `[sur]`), « `s'apprêter` » avec la préposition `[à]` et « `traverser` » avec les prépositions (`[en]` et `[à]`) qui expriment une relation entre les termes (`véhicule`, `piéton`, `agglomération`, `etc.`) désignant des concepts (`Vehicule`, `Pieton`, `Agglomeration`, `etc.`) figurant dans l'ontologie.

Les relations verbales repérées dans le corpus des scénarios sont représentées sous la forme de triplets (sujet, verbe [préposition], complément) par exemple (`véhicule, circuler [en], agglomération`) (cf. figure 1). Les sujets et les compléments sont remplacés par l'étiquette du concept dont ils relèvent dans l'ontologie. Par exemple, les compléments « `voie principale` », « `voie de desserte` », « `voie étroite` », « `voie intermédiaire` », « `voie d'importance intermédiaire` » de la relation

verbale « `circuler sur` » relèvent du concept « `Infrastructure` ». Le triplet (`Vehicule`, **`circuler [sur],`** `Infrastructure`) est une relation termino-ontologique établie entre les concepts `Vehicule` et `Infrastructure` traduisant les relations verbales repérées dans le texte. Les triplets ayant même `concept sujet` et même `concept complément` sont ensuite regroupés. On obtient ainsi des classes sémantiques de verbes susceptibles de désigner des relations ontologiques.

Scénario 3 : Situation de conduite	**Scénario 13 :** Situation de conduite
Un véhicule circule en agglomération (11 cas[1]), **sur une voie principale** (11 cas, dans un autre cas, il s'agit d'une voie d'importance intermédiaire), **le plus souvent bordée de commerces** (7 cas). **Un piéton s'apprête à traverser la chaussée, le plus souvent en intersection** (7 cas) **ou à faible distance** d'une intersection (5 cas). **La circulation est dense à très dense, la chaussée souvent large** (nombre total de voies supérieur à deux – y compris voies spécialisées – dans 8 cas, deux voies de plus de sept mètres hors stationnement dans 1 cas).	**Un véhicule circule sur une voie étroite en centre ville** (3 cas) **ou sur une voie de desserte ou d'importance intermédiaire en périphérie d'agglomération** (3 cas). **Un piéton est arrêté sur la chaussée ou en limite de chaussée, souvent en train de discuter** (avec un conducteur d'un véhicule à l'arrêt dans 4 cas, un autre piéton dans 1 cas).

Tab. 1[2] – Relations verbales entre termes du domaine

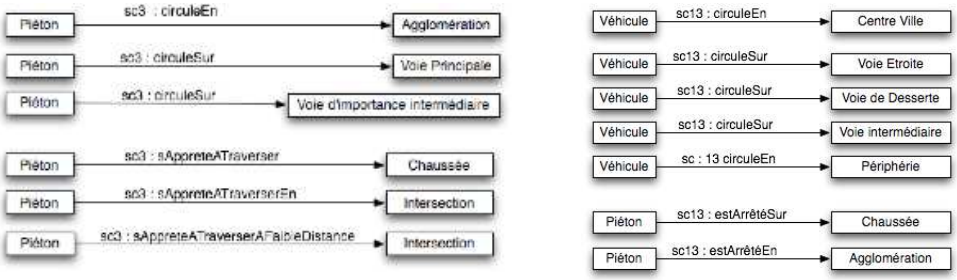

Fig. 1 – Triplets (Sujet, Verbe, Complément)

L'exemple *supra* met en évidence un phénomène de métonymie entre les termes « `piéton` » et « `véhicule` ». Deux relations termino-ontologiques entre les concepts « `Véhicule` » (respectivement « `Piéton` ») et « `Infrastructure` » peuvent être définies et désignées par la relation termino-ontologique « `circuler [sur]` ».

Des liens de proximité sémantique définis par le contexte de l'application entre les relations peuvent également être repérés. Par exemple, dans le contexte de l'accidentologie, une proximité sémantique peut être établie entre les relations

[1] Il s'agit du nombre de cas d'accidents caractérisés par cette action par rapport à l'ensemble des cas étudiés.

[2] Les termes grisés correspondent aux sujets et compléments des syntagmes verbaux (verbe [préposition]) représentés en noir.

verbales (piéton, **est présent [sur]**, chaussée) (piéton, **divague [sur]**, chaussée) et (piéton, **est couché [sur]**, chaussée).

Les relations verbales peuvent également être spécialisées. Dans l'exemple précédent, il est possible d'interpréter la relation verbale (piéton, **est présent sur**, chaussée) comme une relation généralisant les relations (piéton, **divague sur**, chaussée) et (piéton, **est couché sur**, chaussée).

Le verbe « être » en tant que verbe d'état peut suggérer des informations conduisant à la définition d'attributs pour un concept ou suggérer une relation entre classes sémantiques. Dans le scénario 3, il est utilisé pour qualifier la « Circulation » qui est « *dense* » ou « *très dense* ». Lors de la conceptualisation, la question sera de déterminer si l'on associe une échelle de densité au concept « Circulation » ou s'il est nécessaire de définir un concept « Densité » qui sera en relation avec le concept « Circulation » par la relation ontologique « **APourDensité** ».

Dans le contexte de l'accidentologie, le traitement de la négation intervient le plus souvent pour qualifier un état par exemple (piéton, **n'est pas**, alcoolisé). Le triplet sera regroupé avec des triplets du type (piéton, **est**, alcoolisé) ou (piéton, **est fortement**, alcoolisé) ou (piéton, **est fortement diminué par**, alcool). Lors de la conceptualisation, les concepts « Alcoolémie » et « TauxAlcoolémie » seront créés et le concept « TauxAlcoolémie » sera défini par la relation « **aPourTauxAlcoolémie** ».

3 La construction des relations ontologiques

Une fois les relations termino-ontologiques définies entre les classes, elles sont traduites en OWL. L'étude des relations termino-ontologiques autres que hiérarchiques aboutit à la création de deux types de liens : les *propriétés d'objets* qui sont des propriétés entre des instances de classes et *les propriétés de type de données* qui relient des instances à des types de données. Lorsqu'on définit une propriété, il existe plusieurs façons de restreindre la relation. On peut définir un domaine et un codomaine ou définir la propriété comme une spécialisation (sous-propriété) d'une propriété existante. La définition des propriétés d'objets à partir des relations termino-ontologiques nécessite de définir un terme préféré pour désigner la relation, de fournir une définition la qualifiant et de spécialiser la relation en définissant un domaine et un codomaine.

Dans le domaine de l'accidentologie, les relations verbales peuvent être regroupées en catégories relatives aux actions, aux manœuvres, aux mouvements, à la perception, à la décision, etc. Elles sont établies entre les concepts de l'ontologie en cours de construction. Par exemple, les relations de manœuvre désignées par le « **effectueManoeuvreSur** » sont établies entre le concept « Humain » et le concept « Infrastructure».

Dans le corpus constitué des scénarios d'accidents, la définition des propriétés de type de données est généralement effectuée à partir d'une relation verbale exprimée par verbe d'état. Le domaine de la relation est le concept dont relève les éléments de la classe concernée. Par exemple, si le choix est d'associer un concept de « Densite »

à celui de « `Circulation` », l'ensemble des valeurs du concept « densite » défini par la relation « **APourDensite** » pourra être {fluide, …, dense, très dense}. Dans le cas du concept « `TauxAlcoolemie` », l'ensemble des valeurs du concept défini par la relation « **aPourTauxAlcoolemie** » pourra être un ensemble de valeurs numériques exprimées en g/l.

Une fois effectué le choix entre les types de propriétés ontologiques pour représenter les relations termino-ontologiques, la question de leur dénomination est posée. En effet, le choix de l'étiquette associée à la relation ontologique créée aura par la suite un impact important sur l'usage de l'ontologie pour servir à l'annotation des textes ou pour concevoir de nouveaux scénarios.

Notre réflexion porte actuellement sur la caractérisation de nouvelles classes de verbes (mouvements, action, perception) susceptibles d'aider à cette étape de la conceptualisation. Une fois cette approche évaluée, il conviendra de s'interroger sur l'opportunité d'augmenter la bibliothèque des patrons LS Ops existants.

4 Mise en œuvre de l'approche

La partie implantée de l'approche concerne le travail sur le texte. Nous utilisons une liste de syntagmes verbaux du domaine de l'accidentologie validée par les chercheurs en accidentologie. Le sujet et les compléments sont extraits à l'aide d'une fenêtre définie autour du syntagme verbal. Les candidats triplets retenus sont filtrés en utilisant les étiquettes désignant les concepts de l'ontologie. Cette démarche est valide dans la mesure où le vocabulaire du corpus est stabilisé. Une fois filtrés, les triplets sont exprimés au format RDF.

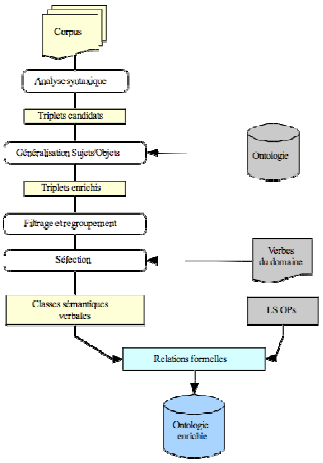

Fig. 2 – Présentation schématique de l'approche

Un algorithme de regroupement est à l'étude pour obtenir les classes sémantiques correspondant aux sujets et compléments et les verbes les liants. Le choix d'un terme préféré pour désigner la relation est effectué et les verbes synonymes lui sont associés.

Les classes de triplets ainsi obtenues seront exprimées au format RDFS. Le recours à de nouveaux patrons LS Ops sera ensuite intégré à notre outil afin d'aider le concepteur dans ses choix pour formaliser les relations obtenues.

5 Conclusion

Dans cette contribution, nous avons analysé le rôle des relations verbales dans la phase de conceptualisation d'une ontologie. Cette analyse a permis de définir une approche qui gère le passage progressif du texte à l'ontologie. Le but est d'apporter une aide au concepteur en lui fournissant des classes sémantiques associées aux verbes afin de l'aider dans la formalisation des relations non hiérarchiques entre les concepts de l'ontologie. Une première étape consiste à regrouper les relations verbales exprimées sous forme de triplets en classes de sujets et de compléments liés par un verbe. Ces classes sont ensuite mises en relation avec les concepts de l'ontologie en cours de construction. Enfin les relations sont traduites par des propriétés de type d'objets ou de type de données. Enfin, pour aider le concepteur le recours aux patrons LS Os est actuellement testé.

Références

AGUADO DE CEA G. & GOMEZ-PEREZ A. & MONTIEL-PONSOLA E. & SUAREZ-FIGUEROA M.C. (2008), Natural Language-based Approach for Helping in the Reuse of Ontology Design Patterns, LNCS, vol.5268/2008 Knowledge Engineering : Practice and Patterns.

AUSSENAC-GILLES N. & DESPRES S. & SZULMAN S. (2008), The TERMINAE Method and Platform for Ontology Engineering from Texts, *Bridging the Gap between Text and Knowledge: Selected Contributions to Ontology learning from Text*. P. Buitelar, P. Cimiano (Eds), IOS Press, P. 199-223.

BRENAC T., NATCHTERGAËLLE C., REGNER H. (2003), Scénarios types d'accidents impliquant des piétons. Rapport N°256. Les collections de l'INRETS.

CIMIANO P. (2006), Ontologies on Demand? A description of the State of the Art Applications, Challenges and Trends for Ontology Learning from Text. In Information, Wissenschaft und Praxis 57 (6-7) : 315-320.

KAMEL M. & AUSSENAC N. (2009) Construction automatique d'ontologies à partir de spécifications de bases de données. In Actes IC'2009.

NEON PROJECT (2008), D2.5.1 : A Library of Ontology design Patterns : reusable solutions for collaborative design of networked ontologies.

SANCHEZ D., MORENO A. (2008), Learning non-taxonomic relationships from web documents for ontology domain construction. In Data & Knowledge Engineering.