Les entités nommées : éléments pour la conceptualisation

Nouha Omrane, Adeline Nazarenko, Sylvie Szulman

Les entités nommées : éléments pour la conceptualisation

Nouha Omrane[1], Adeline Nazarenko[1], Sylvie Szulman[1]

LIPN CNRS-UMR 7030, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse
prenom.nom@lipn.univ-paris13.fr

Résumé : Dans une perspective de construction d'ontologie à partir de textes, cet article montre que la prise en compte des entités nommées, souvent destinée à peupler des ontologies, peut aussi enrichir la phase de conceptualisation. Après avoir introduit la notion même d'entité nommée, nous montrons comment ce type d'unités peuvent être intégrées dans la méthode TERMINAE de construction d'ontologie à partir de textes. Nous illustrons notre propos sur une expérience de création d'une ontologie destinée à la formalisation de règles métiers à partir d'un texte réglementaire.

Mots-clés : Construction d'ontologies à partir de textes, Entités nommées, Ontologie de domaine, Conceptualisation

1 Introduction

Ce travail se situe dans le domaine de la construction d'ontologies de domaine à partir de textes. Comme les textes spécialisés reflètent les connaissances partagées par une communauté linguistique, ils peuvent servir de point d'appui pour la construction d'ontologies et remplacent pour partie les entretiens avec les experts qui sont souvent peu disponibles. Le fait de partir de textes permet aussi de lier étroitement le matériau textuel et l'ontologie construite, que des fragments textuels soient utilisés pour documenter les éléments ontologiques ou que l'ontologie soit destinée à annoter sémantiquement de nouveaux corpus textuels.

Il est désormais traditionnel en ingénierie ontologique à partir de textes de distinguer deux processus (Cimiano, 2006) : la construction de la structure de l'ontologie ou de sa partie conceptuelle (ou T-Box), et son peuplement par des instances, qui forment la partie assertionnelle (ou A-Box). L'analyse des entités nommées (EN), à l'instar des noms de personnes, de lieux, de dates et de leurs catégorisations, révèle l'ancrage référentiel du texte et facilite l'identification du contexte dans lequel sont utilisées l'ensemble des termes. A ce titre, les entités nommées ont un rôle à jouer dans le processus la conceptualisation et il paraît dommage de séparer les processus d'élaboration de la T-Box et de la A-Box, la première servant de structure à la seconde.

Ce travail s'inscrit dans un projet plus large de modélisation de règles métier à partir

des documentations d'entreprise [1], où l'ontologie permet de formaliser le vocabulaire conceptuel sur lequel portent les règles métier. Dans ce contexte, nous nous intéressons au corpus AAdvantage qui décrit l'attribution de points de fidélité et autre avantages aux passagers adhérents du programme de fidélité de la compagnie aérienne American Airlines. Le corpus explicite les règles et conditions d'attribution de chaque type d'avantage, les délais de validité des bonus obtenus et des droits de chaque partenaire. Ce corpus de langue américaine contient environ 5300 mots et 250 paragraphes [2].

La section 2 inroduit la notion d'entité nommée et en précise le fonctionnement linguistique. La section 3 montre comment ces unités peuvent être intégrées dans la méthode de conceptualisation TERMINAE. La section 4 illustre l'approche proposée.

2 Des unités textuelles particulières : les entités nommées

2.1 Définition

Aussi utilisée soit-elle, la notion d'entité nommée n'est pas toujours facile à cerner. La terminologie employée est trompeuse (on parle d'"entités nommées" pour désigner des "noms d'entités") mais elle souligne le fonctionnement référentiel (une "entité" est un élément de la référence) de certaines unités textuelles ("noms"). Sans en avoir toujours la forme, ces entités nommées fonctionnent comme des "noms propres".

D'un point de vue linguistique, les noms propres ont un fonctionnement particulier, référentiel, qui en fait tout l'intérêt pour la construction d'ontologies : "désignateur rigide" fixant une référence Kripke (1972), un nom propre renvoie à une entité référentielle unique dans un contexte donné [3], et il y renvoie de manière autonome, *i.e.* par son seul nom et sans l'appui d'élément contextuel (Ehrmann, 2008). En toute rigueur, nous devrions reprendre la distinction introduite dans (LDC, 2004). Les "mentions d'entités nommées" sont des unités textuelles qui renvoient à des "entités" du domaine mais qui peuvent relever de différentes catégories linguistiques : noms propres ("Air France"), pronoms ("elle"), et plus largement descriptions définies ("cette compagnie", "la principale compagnie aérienne française"). Nous mettons cependant l'accent ici sur les seuls noms propres, plus faciles à détecter dans le flux textuel.

2.2 Reconnaissance des entités nommées

A la faveur des travaux sur l'extraction d'information et des campagnes d'évaluation, de nombreux outils de reconnaissance d'entités ont été mis au point. Ils reposent généralement sur des règles d'extraction qui expriment un faisceau de contraintes (présence de certains marqueurs ou de certaines catégories syntaxiques, ordre des mots, etc.) couplées avec des dictionnaires ou à des catégories sémantiques. Les règles servent à reconnaître des formes variantes des entités des dictionnaires (*Roland Garros vs. R.*

1. Projet FP7 2009-231875 ONTORULE (http ://ontorule-project.eu/).

2. Ce texte d'une dizaine de pages est publié sur le site de la compagnie (http ://www.aa.com). Sous copyright, il est utilisé avec l'autorisation d'American Airline, que nous remercions.

3. Les cas d'ambiguïté d'entités nommées ne sont pas rares mais, dans un contexte donné, il s'agit essentiellement de métonymie.

Garros) ou à découvrir de nouvelles entités d'un type donné [4]. S'ils peuvent servir à reconnaître les instances d'un domaine, les outils de reconnaissance d'entités nommées ont cependant leurs limites : ils ne reconnaissent que les mentions de types "noms propres", seulement les entités de certains types prédéfinis et avec une part d'erreur.

Nous avons testé trois outils (TagEN [5], OpenCalais [6], Gate [7]) sur notre corpus AAdvantage. Les résultats en termes de précision et de rappel sont bons pour les entités nommées de type DATE ou LIEU et ORGANISATION et plus faibles dans le cas du type PERSONNE. Certaines entités nommées ont un type erroné. Ce sont des termes spécifiques au domaine qui ne peuvent être reconnus correctement par des outils non spécialisés (par exemple l'entité nommée *Ruby* qui désigne une catégorie particulière de passagers a été classée sous le type PERSONNE).

3 Conceptualisation

3.1 Problématique

Par "conceptualisation", nous désignons le processus d'élaboration de la partie conceptuelle de l'ontologie. Cette construction a fait l'objet d'études depuis plusieurs années dans le domaine de l'ingénierie des ontologies. La plupart des méthodes de construction d'ontologies à partir de textes (Aussenac-Gilles *et al.*, 2000, 2008; Cimiano & Volker, 2005) repose sur un cadre méthodologique commun qui repose sur quatre étapes : la constitution d'un corpus de documents reflétant le domaine et l'application visée, l'analyse linguistique de ce corpus, la conceptualisation et l'opérationnalisation de l'ontologie. Nous nous intéressons ici à la phase de conceptualisation, telle qu'elle est mise en oeuvre dans l'outil TERMINAE [8] et prévue dans Dafoe (Charlet *et al.*, 2008). Au cours de cette phase, le cogniticien s'appuie sur les résultats de l'analyse du corpus par des outils de traitement automatique des langues. Cette analyse construit un réseau terminologique constitué de termes et de relations terminologiques. Une étape de désambiguïsation des termes donne ensuite naissance, pour chaque terme reconnu comme pertinent, à un ou plusieurs éléments appelés termino-concepts qui représentent chacun un sens de terme iniatial. Les termino-concepts sont liés à d'autres termino-concepts par des relations termino-conceptuelles, qui peuvent être construites à partir des relations terminologiques. A partir de ces termino-concepts et de leurs relations termino-conceptuelles, des concepts, des instances et des propriétés de concepts sont créés dans l'ontologie.

3.2 Conceptualisation à partir des entités nommées

Nous proposons d'ajouter au niveau linguistique la visualisation les résultats des outils de reconnaissance d'entités nommées sous la forme d'une liste de mentions d'entités

4. Le fragment "Monsieur Roland Garros" permet d'identifier *Roland Garros* comme une personne si celle-ci n'est pas déjà connue.

5. TagEN Taggeur d'Entités Nommées

6. http ://www.opencalais.com/

7. http ://gate.ac.uk/

8. http ://www-lipn.univ-paris13.fr/szulman/logi/index.html

nommées associées à leur type sémantique. Suite aux travaux sur le peuplement d'ontologies, on pourrait considérer que les entités nommées ont vocation à devenir des instances. Il nous semble cependant que l'interprétation de ces entités nommées et le processus global de conceptualisation à partir des textes sont plus complexes. La correspondance terme/concept et entité-nommée/instance peut être opératoire à grande échelle dans une perspective de peuplement mais elle est trop simplificatrice dans la phase délicate de conceptualisation. Même si les termes extraits, qui ne sont généralement pas déterminés (*participant vs. ce/le participant*) à la différence des entités nommées, ont une valeur sémantique générique qui pousse à les interpréter comme concept, ils peuvent, en contexte, faire référence à des entités particulières du domaine. De la même manière les entités nommées peuvent conduire à la création de concepts plutôt que d'instances de concepts. Enfin, les types sémantiques qui sont associés aux entités nommées reconnues sont également des éléments utiles pour la conceptualisation.

Plus fondamentalement, représenter quelque chose comme un concept ou comme une instance relève d'un choix de modélisation qui ne peut pas être automatisé. Le cogniticien doit faire ce choix en fonction du contexte applicatif pour lequel l'ontologie est construite et de la granularité de description qu'il veut fournir.

3.3 Méthode

Le choix de modélisation se fait au niveau termino-conceptuel qui constitue un passage obligé entre les unités linguistiques (termes et entités nommées) et les unités conceptuelles (concepts et instances). La conceptualisation passe ainsi par deux étapes essentielles. La première consiste à créer des termino-concepts pour chaque sens de terme ou d'entité nommée jugé pertinent pour le domaine. De même qu'un terme, une entité nommée ambiguë (*Roland Garros*) est associée à plusieurs termino-concepts (***Roland Garros_personne*** et ***Roland Garros_lieu***). Mais les entités nommées sont aussi associées à des types sémantiques qui constituent donc des termino-concepts de rattachement naturels : une entité nommée pertinente u de type T peut permettre de créer deux termino-concepts (***u*** et ***t***), et une relation hiérarchique entre eux. La deuxième étape consiste à élaborer l'ontologie à partir du niveau termino-conceptuel et c'est là que l'on peut choisir de transformer un termino-concept en concept ou en instance [9]. La mise en oeuvre de cette méthode est intégrée à l'outil TERMINAE.

4 Illustration

Une première ontologie a été construite pour la modélisation des règles d'attribution de bonus à partir du corpus AAdvantage mais en s'appuyant sur la seule analyse terminologique. Nous avons repris l'expérience pour apprécier l'apport des entités nommées à ce processus de conceptualisation.

Nous avons rencontré différents cas de figure. L'entité nommée *Central America* traduit une connaissance importante du domaine. Elle fait naturellement référence à un ensemble de pays (*Costa Rica, El Salvador*, etc.) mais elle désigne ici moins une zone

9. En réalité, on peut même choisir de modéliser un termino-concept par un rôle.

géographique qu'un ensemble d'aéroports, certaines règles d'attribution étant liées spécifiquement aux aéroports de cette zone géographique. Cette entité nommée était importante à prendre en compte mais nous avons décidé de la modéliser comme concept pour pouvoir lui attacher des instances (des aéroports particuliers). Ont également été reconnus des noms de compagnies aériennes, comme *American Connection*, *American Eagle*, etc., qui participent au programme de fidélité AAdvantage. Comme nous avions déjà créé un concept **AAdvantage_airline_participant** à partir du terme *AAdvantage airline participant*, nous avons intégré les compagnies aériennes comme instances de ce concept. Ces dernières étant associées au type sémantique ORGANIZATION, nous avons également ajouté un concept **Organization** comme père du concept **AAdvantage_airline_participant**.

Le fait que les outils de reconnaissance d'entités nommées aient extrait des unités comme *Sapphire* et *Ruby* a permis d'en mieux mesurer l'importance dans le domaine considéré. Ces entités nommées représentent les statuts que peut avoir un passager dans la terminologie propre d'American Airlines. La découverte de ces entités nommées et l'analyse de leurs contextes d'utilisation a conduit à enrichir une partie de l'ontologie initiale avec les concepts suivants :

- **Elite member** représente une classe de voyageurs jouissant d'un statut particulier pour l'attribution des bonus ;
- **Benefit** regroupe l'ensemble des avantages relatifs à chaque type de statut (bonus, réduction du prix de billet, séjour à l'hôtel, etc.) ;
- **Numerical_quantity** représente les quantités de miles, points ou segments utilisées pour l'attribution de bonus aux voyageurs privilégiés (**Elite member**) ;
- **Account** a été introduit pour représenter le compte où chaque voyageur privilégié accumule des points de fidélité. Des restrictions de rôles permettent de lier le montant du compte et la catégorie du voyageur : par exemple, pour un passager ayant comme statut *Ruby* doit avoir au minimum 25 000 miles ou bien 25000 points ou encore 30 segments.

L'identification de ces entités nommées a également permis de créer des rôles : par exemple, le rôle **earns** qui relit le concept **Elite member** au concept **Benefit**, le rôle **possesses** qui relit le concept **Elite member** au concept **Account** ou le rôle **has_value** reliant le concept **Account** au concept **Numerical_quantity** traduisant le nombre de miles attribués à chaque statut (restriction de rôle suivant le statut).

Les entités nommées ainsi identifiées ont permis à la fois d'enrichir la T-Box et de peupler la A-Box. Nous avons ajouté 7 concepts et 45 instances de concepts aux 130 concepts de l'ontologie initiale. Nous avons exclu les entités nommées représentant les villes qui présentent peu d'intérêt pour le domaine étudié mais toutes les autres entités extraites (76 au total) ont été conservées d'une manière ou d'une autre. Par ailleurs, l'analyse de certaines entités nommées (spécialement celles relatives aux statuts de voyageurs) a amené la redéfinition de 15 des 130 concepts initiaux.

5 Conclusion

Ce travail nous a permis de tester l'hypothèse selon laquelle la prise en compte des entités nommées dans la création d'ontologies à partir de textes peut enrichir la phase

de conceptualisation. Nous avons proposé de médiatiser le lien entre la mention d'entité nommée et l'entité ontologique qui la modélise (concept ou instance) de la même manière que pour les termes et les concepts, c'est-à-dire par des termino-concepts. Nous avons adapté la phase termino-conceptuelle de TERMINAE pour permettre de travailler aussi bien sur les entités nommées que sur les termes.

Les entités nommées reflètent souvent des instances mais l'existence de ces instances guide le travail de modélisation au niveau conceptuel. Il arrive par ailleurs que les entités nommées soient modélisées sous la forme de concepts, notamment quand on a besoin de leur attribuer des instances. Enfin, même quand les entités nommées ne sont pas assez pertinentes pour être conservées dans l'ontologie finale, leur type sémantique peut traduire un concept pertinent pour le domaine.

Remerciements

Ce travail a été partiellement financé par le projet EU-IST Integrated Project 2009-231875 ONTORULE et a bénéficié de nombreuses discussions avec nos collègues F. Lévy et A. Guissé (LIPN) et John Hall (Model Systems, UK).

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. Proceedings of the 12th International Conference (EKAW'2000)*, LNAI 1937, p. 172–188 : Springer-Verlag.

AUSSENAC-GILLES N., DESPRES S. & SZULMAN S. (2008). The terminae method and platform for ontology engineering from texts. In P. BUITELAAR & P. CIMIANO, Eds., *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*. IOS Press.

CHARLET J., SZULMAN S., PIERRA G., NADAH N., TEGUIAK V., AUSSENAC-GILLES N. & NAZARENKO A. (2008). Dafoe : A multimodel and multimethod platform for building domain ontologies. In *Actes des 2'emes Journées Francophones sur les Ontologies*, p. 66, 77, Lyon France.

CIMIANO P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation And Applications*. New York, USA : Springer.

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.

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Thèse de linguistique. Université de Paris VII.

KRIPKE S. (1972). *La logique des noms propres*. Paris, France : Les Editions de Minuit.

LDC (2004). *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. Livrable version 5.6.1 2005.05.23, Linguistic Data Consortium.