Sujet de thèse

Dynamique de l'annotation sémantique : analyse, modélisation et mise en oeuvre

Septembre 2013

Encadrement : Adeline Nazarenko (directrice) et Davide Buscaldi (co-encadrant)
Candidat : Nazanin Firoozeh
Laboratoire : LIPN

Mots clefs

Annotation sémantique, ingénierie des connaissances, web sémantique

Résumé

L'annotation sémantique des documents joue aujourd'hui un rôle clef dans la convergence du web textuel et du web des données ou du web sémantique.

L'annotation de texte consiste à apposer sur le texte des informations ou métadonnées dont la sémantique est portée par un modèle (langage d'indexation, thesaurus, ontologie, par exemple). On associe ainsi au texte une représentation sémantique formelle qui peut être exploitée par des moteurs sémantiques ou des agents logiciels dans le cadre du web sémantique.

Cette thèse a pour but de concevoir les outils et méthodes qui permettent de construire ou mettre à jour dynamiquement le modèle sémantique utilisé en cours d'annotation.

Objectif

L'annotation de texte consiste à apposer sur le texte des informations ou métadonnées dont la sémantique est portée par un modèle (langage d'indexation, thesaurus, ontologie, par exemple). Cette annotation permet d'associer au texte une représentation sémantique dont la granularité dépend des applications visées mais qui est formelle. Cette annotation peut être faite par des outils automatiques ou manuellement dans le cadre de campagnes d'annotation. Les corpus annotés sont ensuite utilisés à des fins d'analyse de contenu avec des opérations (de recherche, comparaison, synthèse, navigation, segmentation, etc.) plus riches

que ce qui peut se faire sur le texte brut. Les corpus annotés manuellement sont généralement utilisés comme données d'apprentissage et d'évaluation.

Des outils existent pour annoter sémantiquement des textes de manière automatique ou pour guider le travail d'annotation manuel, au regard d'un modèle sémantique qui est généralement un thesaurus ou une ontologie. Il existe aussi des méthodes et des outils pour construire des modèles sémantiques à partir de textes, les textes étant des sources d'information précieuses pour l'élicitation des connaissances. Ces deux processus d'acquisition et d'annotation sont cependant généralement considérés comme distincts. Le modèle sémantique utilisé est défini *a priori* puis utilisé tel quel lors de l'annotation sémantique.

Or cette conception statique de la sémantique est peu réaliste. Elle suppose qu'on dispose déjà d'un modèle sémantique adapté et de bonne qualité. En pratique, le modèle sémantique devrait se construire ou être modifié dynamiquement au fil de l'annotation car c'est souvent le processus d'annotation qui met en évidence les limites du modèle utilisé ou des règles d'annotation associées. L'impossibilité d'annoter des textes ou l'insuffisante qualité de l'annotation produite révèle des "trous" dans la couverture sémantique du modèle utilisé ou des incohérences et impose d'enrichir ou de corriger le modèle ou la manière dont on l'utilise.

En pratique dans l'état de l'art, cette mise à jour se limite généralement au peuplement du modèle sémantique : on enrichit le modèle avec les nouvelles instances qui sont mentionnées dans les textes. Pris dans un sens large, l'annotation sémantique considère cependant que les unités textuelles peuvent être annotées avec différents types d'unités sémantiques (des concepts, des instances, des relations, voire règles). Il faut donc repenser le processus de mise à jour dans ce contexte-là.

A l'inverse de l'approche séquentielle et statique traditionnelle, cette thèse vise à concevoir une méthode d'annotation sémantique permettant non seulement de peupler mais aussi de mettre à jour dynamiquement le ou les modèles sémantiques en cours d'annotation. Cela revient à intégrer les processus d'acquisition des connaissances et d'annotation.

Approche

Modéliser un tel processus d'annotation sémantique suppose plusieurs étapes :

1. définir le type d'annotation visé et identifier les outils à utiliser pour le mettre en oeuvre (un nouvel outil d'annotation devra éventuellement être développé) ;
2. identifier et modéliser les conditions qui justifient et déclenchent la mise à jour du modèle utilisé au cours de l'annotation (il peut s'agir, par exemple, d'une mesure de couverture, de la détection d'une incohérence) ;
3. définir et formaliser les mécanismes de mise à jour du modèle sémantique et les opérations qui permettent de les implémenter : les modifications peuvent porter sur le modèle lui-même (ajout/retrait/modification d'un

élément ou restructuration plus globale) mais aussi sur les règles d'annotation qui permettent de projeter le modèle dans le texte ;

4. dans certains cas, réviser l'annotation déjà produite pour tenir compte des changements opérés sur le modèle ;
5. étendre l'approche au cas où plusieurs ressources sémantiques, éventuellement partiellement alignées entre elles, sont exploitées pour l'annotation.

Ce travail pourra tirer profit de l'état de l'art existant sur les travaux sur le peuplement d'ontologies [10, 4], l'annotation sémantique [11, 5, 13]), l'évolution des référentiels sémantiques, notamment des ontologies [6, 2, 12], mais devra les étendre et les articuler.

Le doctorant s'appuiera sur les outils développés dans l'équipe RCLN pour l'acquisition des connaissances à partir de textes (TERMINAE [1] et SemEx [7]), de l'expérience acquise en annotation sémantique de corpus, qu'elle soit automatique [9, 8] ou manuelle [3] et des travaux sur la recherche sémantique [14]. Il s'agira dans un premier temps de travailler dans le cadre de modèles classiques (thesaurus, ontologies) pour lesquels il existe des formalismes (SKOS, OWL-DL) et des technologies bien établis mais d'autres modèles sémantique pourront éventuellement être envisagés.

Comme mentionné ci-dessus, ce problème de la dynamique de l'annotation sémantique se pose aussi bien pour les approches automatiques de l'annotation que pour les approches manuelles. La thèse pourra soit se focaliser sur l'annotation automatique soit traiter les deux approches en parallèle.

Contexte

Ce sujet de thèse s'inscrit dans le cadre des travaux de l'équipe RCLN sur l'annotation sémantique (les projets collaboratifs comme Quaero, ONTORULE ou Legilocal ont tous comporté des volets d'annotation sémantique).

Ces questions d'analyse sémantique et d'annotation de corpus constituent également un enjeu important pour le labex "Fondements empirique de la linguistique", auquel participe l'équipe RCLN, et notamment pour l'axe "Analyse sémantique computationnelle" où les thèmes de la sémantique de corpus et l'accès au contenu occupent une place centrale.

A partir de travaux dans le domaine de la gestion de l'information scientifique et technique (collaborations avec l'INRA et l'INIST, notamment), l'équipe RCLN poursuit aujourd'hui ses recherches sur des problématiques liées aux Sciences Humaines et Sociales (notamment dans le champ juridique). Dans le champ des SHS, la taille des textes traités, la structure des collections documentaires et la richesse des interprétations à modéliser montrent les limites de l'approche statique de l'annotation sémantique présentée plus haut et nécessitent de concevoir des méthodes dynamiques à la fois plus robustes et plus puissantes.

Références

[1] Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The TERMINAE Method and Platform for Ontology Engineering from texts. In Paul Buitelaar and Philipp Cimiano, editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 199–223. IOS Press, janvier 2008.

[2] Rim Djedidi and Marie-Aude Aufaure. Ontology change management. In A. Paschke, H. Weigand, W. Behrendt, K. Tochtermann, and T. Pellegrini, editors, *5th International Conference on Semantic Systems (I-Semantics 09), Proceedings of I-KNOW ?09 and I-SEMANTICS ?09*, pages 611–621, Graz, Austria, September 2009. Verlag der Technischen Universitt Graz.

[3] Karën Fort, Adeline Nazarenko, and Sophie Rosset. Modeling the Complexity of Manual Annotation Tasks : a Grid of Analysis. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December 2012.

[4] C. Giuliano and A. Gliozzo. Instance-based ontology population exploiting named-entity substitution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 265–272, Manchester, August 2008.

[5] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2(1) :49–79, 2004.

[6] Pieter De Leenheer and Tom Mens. Ontology evolution : State of the art and future directions. In Martin Hepp, Pieter De Leenheer, Aldo de Moor, and York Sure, editors, *Ontology Management : Semantic Web, Semantic Web Services, and Business Applications*, pages 131–176. Springer, 2007.

[7] François Lévy, Adeline Nazarenko, Abdoulaye Guissé, Nouha Omrane, and Sylvie Szulman. An environment for the joint management of written policies and business rules. In *Proceedings of the International Conference on Tools with Artificial Intelligence (IEEE-ICTAI 10)*, pages 142–149, 2010.

[8] Yue Ma, François Lévy, and Sudeep Ghimire. Reasoning with Annotations of Texts. In *The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 192–197, États-Unis, May 2011.

[9] Yue Ma, Adeline Nazarenko, and Laurent Audibert. Formal description of resources for ontology-based semantic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, Malta, May 2010. ELRA.

[10] Bernardo Magnini, Emanuele Pianta, Octavian Popescu, and Manuela Speranza. Ontology population from textual mentions : Task definition and benchmark. In *Proceedings of the OLP2 workshop on Ontology Population and Learning*, Sidney, Australia, 2006.

[11] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. Kim – a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4) :375–392, 2004.

[12] Zied Sellami, Valérie Camps, and Nathalie Aussenac-Gilles. Dynamo-mas : a multi-agent system for ontology evolution from text. *J. Data Semantics*, 2(2-3) :145–161, 2013.

[13] Victoria Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4, 2006.

[14] Haïfa Zargayouna. *Indexation sémantique de documents XML.* Thèse de doctorat. Université Paris-Sud, Déc. 2005.