Vers une approche interactive pour l'annotation sémantique

Sondes Bannour Souihi, Laurent Audibert

Vers une approche interactive pour l'annotation sémantique ⋆

Sondes Bannour, Laurent Audibert

LIPN, UMR 7030 CNRS, Université Paris 13
99 av. J.-B. Clément - F-93430 Villetaneuse, France
prenom.nom@lipn.univ-paris13.fr

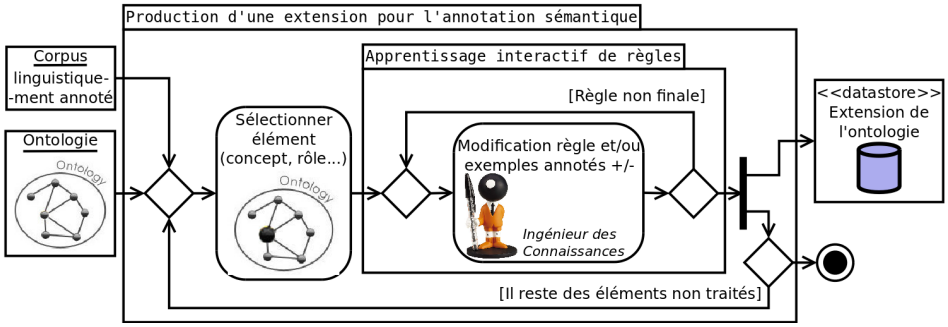
Résumé :
Nous présentons une méthodologie permettant la constitution d'une ressource destinée à l'annotation sémantique de corpus. Notre démarche s'inscrit dans le cadre des plateformes d'annotation linguistique. Elle permet de créer un étage d'annotation sémantique constitué de règles d'annotation qui tirent profit dans leur expression des différents niveaux inférieurs d'annotation linguistique de la plateforme. La particularité de l'approche présentée est d'assister l'utilisateur à travers un processus interactif et itératif où il est possible de travailler de manière duale sur les règles d'annotation ainsi que sur des exemples d'annotation.
Mots-clés : Annotation sémantique, ontologie, règles d'annotation, apprentissage artificiel

L'annotation sémantique se définit comme le processus qui fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite. Les ontologies jouent habituellement un rôle central dans ce type de processus. De nombreux systèmes annotent des textes au regard d'une ontologie. Certains systèmes utilisent des patrons construits manuellement par des experts du domaine (Handschuh, 2002; Kogut, 2001). D'autres systèmes mettent en œuvre des techniques d'apprentissage numérique ou de patrons (Ciravegna, 2000), mais nécessitent l'écriture par un expert d'un guide d'annotation détaillé ainsi que l'annotation manuelle d'une grande quantité d'exemples. L'expérience montre que ces deux approches sont très couteuses en travail manuel.

⋆. Ce travail a été réalisé dans le cadre du programme Quaero, financé par OSEO, agence nationale de valorisation de la recherche.

Pour minimiser ce coût, nous proposons une approche interactive de production de règles d'annotation sémantique où l'utilisateur travaille de manière duale sur les règles d'annotation (intension) ainsi que sur des exemples d'annotation (extension). Ce type d'approches présente l'avantage d'être plus souple et plus rapide qu'une approche uniquement basée sur l'intension (production de règles par un expert) ou sur l'extension (annotation manuelle d'une grande quantité d'exemples). Cette approche est également modulaire et s'intègre bien dans le cadre des plateformes d'annotation linguistique dans la mesure où elle permet d'ajouter un niveau d'annotation sémantique à de telles plateformes. Les règles de cette approche s'appuient sur les étages inférieurs d'annotations linguistiques afin de simplifier au maximum leur expression et forment une extension de l'ontologie au regard de laquelle se fait l'annotation sémantique (Ma *et al.*, 2009). La méthodologie proposée peut tirer parti de techniques d'apprentissage très variées comme l'apprentissage de patrons, l'apprentissage numérique et l'apprentissage actif.

La figure ci-dessous illustre notre méthodologie :



Etant donné une ontologie, les différents éléments qui la composent (concepts, rôles ...) sont traités successivement de manière à venir alimenter de manière incrémentale l'extension de l'ontologie (paquetage *Production d'une extension pour l'annotation sémantique* sur la figure).

Chaque règle d'annotation d'un élément de l'ontologie est construite de manière itérative et interactive, jusqu'à aboutir à une règle couvrant de manière satisfaisante les portions de texte qui doivent être mises en correspondance avec l'élément de l'ontologie traité (paquetage *Apprentissage interactif de règles* sur la figure). Un utilisateur intervient à chaque itération en travaillant de manière duale sur les règles d'annotation ainsi que sur des exemples d'annotation.

La méthodologie proposée permet ainsi d'assister l'utilisateur à constituer une ressource destinée à l'annotation sémantique de corpus à travers

un processus interactif et itératif. Nous pensons que cette méthodologie présente les avantages suivants : (1) L'utilisateur peut travailler de la manière qui lui semble la plus évidente, soit en rédigeant des règles dans un language de règles générique comme TextMarker (Kluegl *et al.*, 2009), soit en validant ou invalidant des exemples. (2) L'expression des règles peut être très concise car elle peut tirer profit de différents niveaux d'annotation linguistique. (3) Le nombre d'exemples à annoter peut être réduit grâce au recours à des techniques d'apprentissage actif. (4) L'utilisation conjointe de plusieurs techniques d'apprentissage devrait permettre d'améliorer les performances. (5) L'approche présentée est très générique et modulaire. (6) Cette approche permet de préserver une articulation claire entre l'annotation linguistique (faite par les outils de TAL) et l'annotation sémantique qui est guidée par l'ontologie.

Références

CIRAVEGNA F. (2000). Learning to tag for information extraction from text. In *Proceedings of the ECAI-2000 Workshop on Machine Learning for Information Extraction*.

HANDSCHUH S. (2002). S-cream - semi-automatic creation of metadata. In *In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*, p. 358–372.

KLUEGL P., ATZMUELLER M. & PUPPE F. (2009). Textmarker : A tool for rule-based information extraction. In *Proceedings of the Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop*, p. 233–240.

KOGUT P. (2001). Aerodaml : Applying information extraction to generate daml annotations from web pages. In *First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation*.

MA Y., AUDIBERT L. & NAZARENKO A. (2009). Ontologies étendues pour l'annotation sémantique. In *20es Journées Francophones d'Ingénierie des Connaissances*, p. 205–216.