Un moteur sémantique pour explorer des textes réglementaires

Abdoulaye Guissé, François Lévy, Adeline Nazarenko

Un moteur sémantique pour explorer des textes réglementaires

Abdoulaye Guissé, François Lévy, Adeline Nazarenko

LIPN, UMR 7030 CNRS, Université Paris Nord
A^{e} J.B. Clément 93430 Villetaneuse, France
prenom.nom@lipn.univ-paris13.fr

Résumé : Si on assiste aujourd'hui à un regain d'intérêt pour les systèmes à bases de connaissances, on constate aussi le besoin croissant d'intégrer les sources d'information et la documentation aux systèmes à base de connaissances, pour faciliter l'utilisation de la connaissance accumulée, permettre l'évolution du système et le doter de nouvelles fonctionnalités. Nous montrons ici l'apport d'un moteur sémantique qui vient enrichir la recherche en texte intégral de fonctionnalités sémantiques, souvent très avancées et spécialement mises au point pour servir les besoins des utilisateurs. Nous présentons SemEx, un outil d'exploration sémantique de textes réglementaires que nous avons développé et qui a vocation à être intégré dans des systèmes de gestion de règles métiers.
Mots-clés : Annotation sémantique, navigation, moteur sémantique

1. Introduction

Si on assiste aujourd'hui à un regain d'intérêt pour les systèmes à bases de connaissances, on constate aussi le besoin croissant d'intégrer les sources d'information et la documentation aux systèmes à base de connaissances, pour faciliter l'utilisation de la connaissance accumulée, permettre l'évolution du système et le doter de nouvelles fonctionnalités. Dans le cas des règles métiers, on ne peut dissocier cette tendance d'initiatives comme SBVR (Semantics of Business Vocabulary and Rules) qui vise à développer un standard rendant intelligibles aux gens des métiers les modèles mis en oeuvre par les règles. Un des objectifs affirmés du projet ONTORULE (FP7-ICT) est de « sortir les systèmes des mains des informaticiens », *i.e.* permettre aux res-

0. Ce travail a été réalisé dans le cadre du projet ONTORULE FP7 231875.

ponsables métiers de comprendre, valider et modifier les règles utilisées par le système.

Dans une perspective d'exploration humaine assistée par des outils, nous montrons ici l'apport d'un moteur sémantique qui vient enrichir la recherche en texte intégral de fonctionnalités sémantiques souvent très avancées et spécialement mises au point pour servir les besoins des utilisateurs. Nous présentons SemEx, un outil d'exploration sémantique de textes réglementaires, que nous avons développé et qui a vocation à être intégré dans des systèmes de gestion de règles métiers.

Le traitement que nous proposons consiste à identifier les passages réglementaires, puis à les reformuler en des spécifications semi-formelles sous la forme d'une liste de règles candidates à la fois normalisées, autonomes et élémentaires. Ces formes simplifiées sont ensuite traduites dans un langage de règles exécutables mais nous n'abordons pas cette étape de formalisation ici.

Cette transformation du texte en règles exécutables qui échappent aux gens des métiers, est documentée, ce qui permet d'utiliser les textes sources pendant l'activité du système de décision. Les décisions sont justifiées en revenant aux textes sources de la réglementation d'où dérivent les règles exécutables utilisées. On peut apporter des explications aux incohérences qui éventuellement se produisent. Enfin, les réglementations évoluant, on peut s'appuyer sur les modifications du texte pour circonscrire les changements à apporter à la base de règles.

Nous illustrons notre propos par deux applications réelles, *AAdvantage* un système de bonus destiné aux clients de la compagnie aérienne American Airlines et l'application *Ceintures de sécurité* du constructeur automobile Audi pour les tests de qualité des ceintures utilisées.

La section 2. introduit les annotations sémantiques sur lesquelles repose notre outil. La section 3. décrit les diverses opérations de navigation et de recherche avancée que propose SemEx et la dernière section situe ce travail dans le contexte des travaux actuels sur l'annotation sémantique.

2. Annotation sémantique et index

L'idée d'intégrer les documents réglementaires dans les SGRM consiste globalement à annoter sémantiquement les textes réglementaires afin de pouvoir facilement articuler des textes, des règles métiers qui en sont dérivées et une ontologie du domaine qui décrit le vocabulaire conceptuel dans lequel les règles métiers s'expriment. Nous proposons ici une structure riche d'index

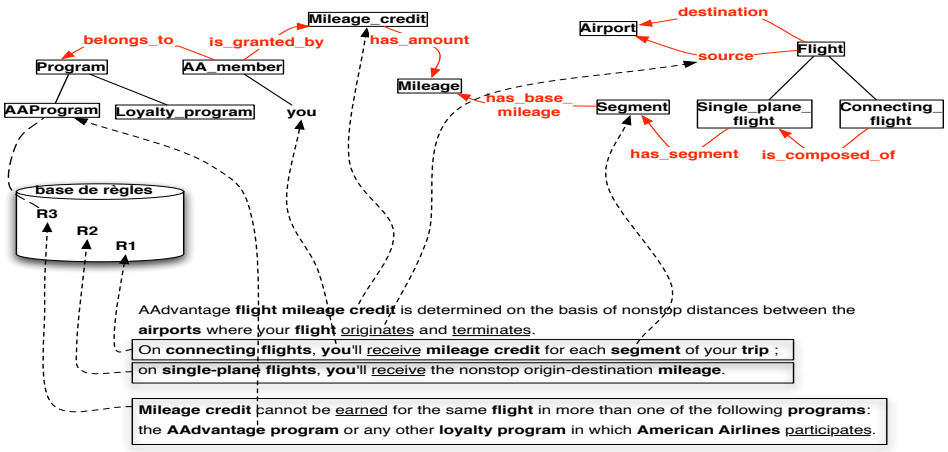

FIGURE 1: Exemple d'annotation

qui permet de passer des concepts de l'ontologie et des règles aux textes et inversement.

Formellement, la structure d'index repose sur deux sous-structures, un modèle documentaire représentant les documents sources et un modèle sémantique composé d'une ontologie et d'une base de règles. Les liens d'indexation sont représentés comme des liens d'annotation inclus dans les documents. L'implémentation de cette structure repose sur des standards de représentation de connaissances W3C du Web sémantique définis autour de RDF afin de faciliter l'exploration et l'interrogation commune des ressources à travers un langage commun de requête comme SPARQL.

Le ***modèle documentaire*** détermine quels fragments ou unités documentaires peuvent être annotés dans un document, un fragment étant défini comme une séquence de caractères pouvant être typée pour distinguer des mots, des syntagmes, des phrases, des paragraphes, des sections, etc. Initialement, il s'agit de documents structurés en format XML mais il sont ensuite représentés en RDFa pour intégrer les liens d'annotations vers l'ontologie et les règles.

Le ***modèle sémantique*** indique quelles unités sémantiques peuvent être associées aux unités documentaires et quelles relations ces unités sémantiques entretiennent entre elles. Une unité sémantique définit soit une entité ontologique (concept, relation ou instance) soit une règle. L'ontologie est représentée de façon classique en OWL. Quant à la base de règles, elle est constituée d'une liste de « règles candidates » (RC) exprimées en langage contrôlé. Chaque règle est modélisée formellement par un graphe RDF associant un

identifiant et un ensemble de propriétés, dont un champ textuel correspondant à l'énoncé de la règle qui est lui même annoté au regard de l'ontologie.

3. Exploration sémantique

3.1. Présentation de SemEx

SemEx est un outil d'exploration sémantique qui repose sur la structure d'index (section 2.) et qui utilise le moteur de recherche sémantique CORESE [1] pour exécuter des requêtes SPARQL sur des données RDF. Deux interfaces dédiées à l'exploration sémantique sont proposées à l'utilisateur (fig. 2). L'interface de navigation permet de visualiser toutes les annotations sémantiques posées sur un document et de naviguer dans la structure d'index. Elle repose sur des requêtes sémantiques générales qui sont prédéfinies. Une seconde interface est dédiée aux requêtes plus particulières : l'utilisateur peut définir ses propres requêtes pour interroger l'index.

3.2. Navigation

La recherche sémantique dans SemEx donne lieu à deux catégories de navigation dans l'index.

Une navigation interne à chacune des ressources permet de passer 1) d'une unité documentaire à une autre suivant des recherches en texte intégral (parcours intra-documentaire pour, ou 2) d'un objet conceptuel à autre, d'une règle à une autre, mais aussi d'une règle aux éléments ontologiques sur lesquels elle pointe et vice versa (parcours sémantique).

Une navigation sémantico-documentaire basée essentiellement sur l'interrogation de l'index permet également de passer des documents au modèle sémantique et inversement. Par exemple, ce type de parcours permet de visualiser les concepts ou règles qui annotent une unité documentaire donnée et, inversement, les fragments de textes où apparaissent un concept ou une règle donné(e).

3.3. Fonctionnalités avancées

Toutefois, l'usage de l'index ne se limite pas à la navigation. La possibilité de tracer les règles dérivées jusqu'à leur texte d'origine est utile dans l'analyse

1. http ://www-sop.inria.fr/edelweiss/software/corese/

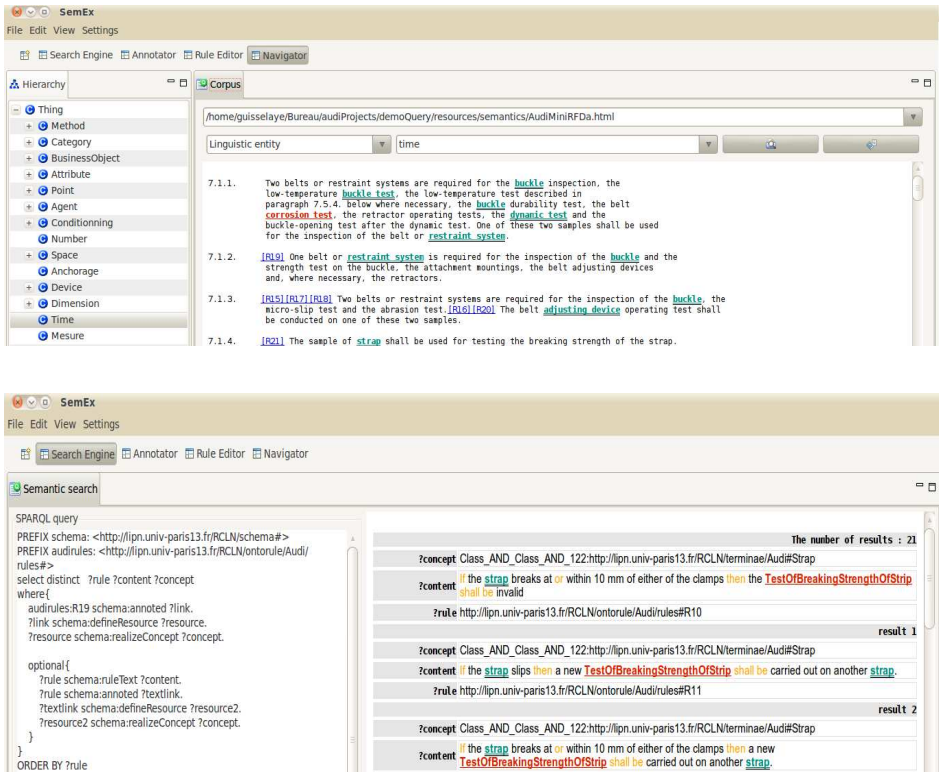

FIGURE 2: SemEx - (*haut*) Interface de navigation - (*bas*) Exemple de requête SPARQL qui montre toutes les règles ayant un concept en commun avec la règle R19.

et la compréhension des règles et des décisions qui en découlent.

3.3.1. Gestion de la cohérence de la base de règle

L'index est important pour gérer la cohérence des règles. Il permet d'analyser des règles qui apparaissent conflictuelles à l'exécution. On peut en effet revenir aux textes sources dont elles sont dérivées pour vérifier si l'incohérence provient du texte source lui-même ou de l'interprétation qui en a été faite. L'index permet aussi de détecter des incohérences au niveau de la base de règles, avant même sa traduction en règles exécutables.

Par exemple, dans l'application *Audi*, avec une traduction trop directe des règlements on obtient une contradiction pour l'entrée (ceinture 3 points, airbag, test de choc dynamique, déplacement au niveau de la poitrine = 350mm,

vitesse mesurée au niveau de la poitrine à 300mm = 12km/h) car deux règles produisent deux réponses contradictoires (*acceptable* et *non acceptable*). En analysant les textes à l'origine de ce résultat [2], on découvre que l'erreur est due à une difficulté d'interprétation fréquente dans ce type de document. Chaque fragment de texte est analysé de manière fidèle mais isolée. Or les prémisses ne sont pas exclusives et la contradiction survient lorsque le cas de figure considéré se trouve dans leur intersection. En fait, la seconde règle est une exception au cas général exprimé par la première phrase et, comme il arrive très souvent, l'exception est introduite après le cas général sans être déclarée comme telle. On peut repérer le problème avant l'exécution recherchant *via* l'index, toutes les règles faisant référence aux concepts correspondant aux notions de ceinture 3 points, test de choc dynamique et déplacement au niveau de la poitrine. On peut alors reformuler les conditions de la première règle de façon plus restrictive (absence d'Airbag) et ajouter à la seconde règle candidate une référence vers la première pour faciliter la maintenance.

3.3.2. Maintenance

L'index est également précieux pour la maintenance des réglementations et des bases de règles associées. Dans la pratique, les règlements des organisations changent constamment et cela peut avoir des conséquences directes sur le vocabulaire conceptuel à utiliser et sur la base de règles. Lorsqu'un texte est modifié (ajout ou suppression de fragment), l'index permet de mettre à jour en conséquence les concepts de l'ontologie ou des règles qui en dérivent tout en sachant que cette première mise à jour peut en amener d'autres.

Par exemple, dans le texte AAdvantage, on trouve une phrase [3] qui s'analyse en deux règles, applicables aux bonus obtenus avant et après le 1er juillet 1989. Avec le temps, la première règle deviendra sûrement obsolète. Pour la supprimer de façon cohérente, il est nécessaire de vérifier l'existence de règles textuelles qui mentionnent l'expiration des bonus dans leur conclusion ou leur prémisse.

---

2. Phrase 1 : *In the case of three points belts, the forward displacement shall be between 80 and 200 mm at pelvic level and between 100 and 300 mm at chest level.*
Phrase 2 : *In the case of a safety-belt intended to be used in a seating position protected by an airbag in front of it, the displacement of the chest reference point may exceed 300mm. if its speed at this value does not exceed 24 km/h.*

3. *"If your account has no qualifying activity in any 18-month period, all miles in the account will expire except for those miles earned prior to July 1, 1989 in accounts established prior to January 1, 1989 whose mileage credit will not expire.*

4. Autres travaux

Avec l'essor des systèmes de gestion de règles métier, l'acquisition des règles métier devient un enjeu important. La question a été abordée jusqu'ici sous deux angles différents. Le premier est celui du langage contrôlé : les règles sont d'emblée écrites de manière normée, ce qui facilite leur traduction en langage formel (Brodie *et al.*, 2006). La seconde approche vise à analyser automatiquement les textes réglementaires mais se trouve confrontée à toute la richesse et la complexité des langues naturelles (Dinesh *et al.*, 2006). Notre approche constitue une troisième voie : elle prend en compte des textes réellement écrits en langue naturelle mais propose des outils pour aider l'analyse manuelle et un dispositif d'exploration sémantique, sans viser une analyse automatique profonde des textes.

Dans le champ des travaux sur l'annotation sémantique, notre travail présente une double originalité. L'annotation consiste en général à mettre en correspondance des éléments de l'ontologie avec des fragments de textes mais la plupart des systèmes d'annotation mettent l'accent sur l'étiquetage des entités nommées (Kiryakov *et al.*, 2004; Amardeilh *et al.*, 2005) alors que nous considérons que n'importe quelle entité ontologique, fût-elle concept ou relation, peut être la cible d'un lien d'annotation. Nous considérons même que le texte peut être annoté au regard de différents modèles sémantiques (ontologie et base de règles, dans notre cas), ce qui enrichit d'autant le modèle d'annotation.

Le modèle sémantique étant plus riche, les fonctionnalités d'accès au contenu sont elles aussi plus diverses. Au-delà de la simple recherche d'information (Popov *et al.*, 2004) , nous proposons un véritable outil d'exploration de corpus adapté à la gestion des règles métiers et reposant sur des technologies du W3C. On trouve des approches similaires dans d'autres domaines applicatifs, comme pour l'exploration des textes médicaux (Ben Abacha & Zweigenbaum, 2010), la gestion des connaissances (Uren *et al.*, 2006) ou le travail de linguistes (Widlocher & Mathet, 2009) mais l'originalité de SemEx tient au fait qu'il repose sur un véritable moteur sémantique.

5. Conclusion

Cet article présente SemEx un outil conçu pour l'exploration sémantique d'un corpus de textes et plus particulièrement d'une base réglementaire. L'annotation sémantique permet d'associer à un texte écrit en langue naturelle et

a priori difficile à analyser en profondeur une structure formelle (en l'occurrence, une ontologie couplée à une base de règles). On peut ainsi interroger le contenu textuel soit par des méthodes d'interrogation en texte intégral traditionnelles soit par des requête sémantiques *via* la structure sémantique et les liens d'indexation qui lui sont associés. Un tel dispositif permet d'intégrer complètement la documentation réglementaire aux systèmes d'aide à la décision qui reposent sur des bases de règles métiers : SemEx sert à l'acquisition, la documentation, la vérification de cohérence et la mise à jour des règles et du modèle métier sur lesquels repose ce type de systèmes.

Références

AMARDEILH F., LAUBLET P. & MINEL J.-L. (2005). Document annotation and ontology population from linguistic extractions. In *Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)*, p. 161–168, New York, NY, USA : ACM.

BEN ABACHA A. & ZWEIGENBAUM P. (2010). Metae : Plate-forme d'annotation automatique et d'exploration sémantiques pour le domaine médical. In *Démonstrations à TALN 2010*, Montréal, Canada.

BRODIE C., KARAT C.-M. & KARAT J. (2006). An empirical study of natural language parsing of privacy policyrules using the sparcle policy workbench. In *SOUPS 06*, p. 8–19.

DINESH N., JOSHI A., LEE I. & WEBBER B. (2006). Extracting formal specifications from natural language regulatory documents. In *ICoS-5*, p. 1–10, Buxton, England.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *J. Web Sem.*, **2**(1), 49–79.

POPOV B., KIRYAKOV A., OGNYANOFF D., MANOV D. & KIRILOV A. (2004). Kim a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, **10**, 375–392.

UREN V., CIMIANO P., IRIA J., HANDSCHUH S., VARGAS-VERA M., MOTTA E. & CIRAVEGNA F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Journal of Web Semantics*, **4**, 14–28.

WIDLOCHER A. & MATHET Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009*, Senlis, France.