Enrichissement de bases de connaissances par l'annotation sémantique

Florence Amardeilh, Thomas Francart

Enrichissement de bases de connaissances par l'annotation sémantique

Plate-forme Web Sémantique couplée avec des outils linguistiques pour des activités de veille et d'édition

Florence Amardeilh *, — Thomas Francart ****

** Laboratoire Lalicc - Université Paris IV - UMR CNRS*
28, rue Serpente, F-75006 Paris
{nom.prenom]@paris4.sorbonne.fr

*** Mondeca*
3, cité Nollez,F- 75018 Paris
{prenom.nom}@mondeca.com

RÉSUMÉ. *Dans cet article, nous présentons une nouvelle forme de portail Web Sémantique utilisant les technologies liées au traitement automatique du langage naturel (TALN). Notre système fournit les moyens d'annoter semi-automatiquement des documents et d'enrichir une base de connaissance contrainte par l'ontologie du domaine concerné. Ce nouveau système permet donc d'assister les communautés du Web, notamment celles travaillant dans le domaine de la veille scientifique et économique, à mettre en place des portails Sémantique centrés sur le domaine d'application. L'utilisateur final pourra ensuite visualiser les données de la base de connaissance, formuler des requêtes intelligentes et complexes et enfin publier les résultats trouvés. La plate-forme décrite dans la suite de cet article est conforme aux standards et langages du Web Sémantique (XML, XTM, RDF(S) et OWL).*

ABSTRACT. *In this paper we will present a new form of Semantic Web portal using Human Language Technologies (HLT). Our system provides the means to annotate documents with metadata, populate a knowledge base according to the corresponding domain ontology and most of all, update the linguistic resources with all the new extracted information in order to improve the performance of the entire system. As a consequence, it will assist the Web communities, and among them the competitive intelligence workers, to create domain-centric Semantic Web portals. The final user will be able, through the application's interfaces, to visualise the knowledge base data, to formulate intelligent and complex queries and at least, to publish the returned results. We would like to point out the fact that the entire platform is Semantic Web-compliant as it is based on its standards (XML, XTM, RDF(S) and OWL).*

MOTS-CLÉS : *Web Sémantique, ontologies, annotations sémantiques, extraction d'information, enrichissement de bases de connaissance.*

KEYWORDS: *Semantic Web, ontologies, semantic annotations, information extraction, knowledge base enrichment.*

1. Introduction

D'après Tim Berners-Lee (Berners-Lee, 1998), la vision du Web Sémantique consiste à rendre le Web actuel compréhensible et donc exploitable par les machines. Pour cela, les ressources documentaires doivent être annotées et structurées sémantiquement en ajoutant du sens et de la connaissance via l'apposition d'étiquettes sémantiques (Katz *et al.*, 2002). Ces étiquettes agissent comme autant d'indices pour les machines afin qu'elles puissent interpréter, traiter et réutiliser l'information presque comme les humains (Lu *et al.*, 2002). Ces derniers pourront ensuite exploiter ces ressources sémantiquement étiquetées afin de les rechercher, de les partager, d'y accéder ou de les publier et donc de travailler plus efficacement (Laublet *et al.*, 2002).

A cause des problématiques issues de l'annotation de corpus documentaires existants et des besoins en productivité et en qualité des annotations créées, il est essentiel pour le succès du Web Sémantique de formaliser des méthodes permettant la production semi-automatique d'annotations à partir de documents non structurés. Il s'agit alors d'extraire la connaissance d'un domaine d'application et de peupler une base de connaissance, celle-ci étant préalablement implémentée et contrainte via l'ontologie du domaine. Cette ontologie permet de définir les concepts relatifs au domaine de l'application, leurs propriétés et les relations entre ces concepts. La connaissance peut ainsi être gérée et exploitée par les utilisateurs finaux, comme les documentalistes ou les veilleurs.

Pour construire ces méthodes dont le Web Sémantique a besoin, nous pensons tout naturellement aux technologies du Traitement Automatique du Langage Naturel (TALN). Ces technologies linguistiques peuvent avoir un impact majeur sur la gestion de la connaissance et spécifiquement dans les communautés du Web Sémantique afin d'aider à mettre en place des solutions opérationnelles pour les utilisateurs. Seule une forte collaboration entre ces deux domaines de recherche améliorera la compréhension du Web par les machines et deviendra la base d'une future génération d'outils intelligents pour le Web.

Tout d'abord, ce papier présentera les raisons de notre intérêt dans le TALN pour développer un portail Web Sémantique (section 2). Ensuite, nous décrirons l'architecture actuelle de notre système et ses principaux composants (section 3). Dans la partie suivante, nous présenterons notre travail de recherche basé sur la mise en correspondance des outils linguistiques avec l'ontologie du domaine tout en l'illustrant par un exemple issu de l'édition juridique (section 4). Enfin, nous passerons en revue les projets relatifs à notre démarche (section 5) avant de conclure et de discuter de nos futurs travaux (section6).

2. Vers un portail Web Sémantique basé sur le TALN

La recherche sur la représentation des connaissances, développée dans le champ de la gestion des connaissances, possède une forte tradition dans la description de connaissances spécifiques à un domaine. Ces techniques permettent le traitement de cette connaissance par les machines. Par ailleurs, le Web Sémantique se base sur des systèmes de représentation de la connaissance, précisément par l'usage d'ontologies, afin d'aider les machines à comprendre et exploiter les ressources documentaires provenant soit d'Internet, soit de systèmes de gestion de contenu. Ces systèmes utilisent des langages tels RDF, XTM et OWL.

RDF, the Resource Description Framework (Ora *et al.*, 1999), est un formalisme de représentation des connaissances, issu des réseaux sémantiques, dont la syntaxe, la plus couramment employée, utilise XML. Il sert à décrire des ressources, tel un document électronique du Web, par un ensemble de métadonnées documentaires (auteur, date, source, etc) et de descripteurs (termes provenant d'un thesaurus). Ces metadonnées sont constituées sous la forme de triplets : (sujet, verbe, objet) ou (objet 1, relation, objet 2) ou encore (ressource, propriété, valeur) selon le type de description nécessaire.

Les Topic Maps sont un autre formalisme de représentation des connaissances qui dispose aussi d'une syntaxe basé sur XML (Park *et al.*, 2003). Les Topic Maps définissent un ensemble de sujets relatifs à un même domaine avec des interactions entre eux formant ainsi une carte sémantique de la connaissance. Un sujet représente tout ce qui peut être décrit ou pensé par un humain. Il peut participer à une ou plusieurs relations, appelées associations, dans lesquelles il joue un rôle spécifique. Les sujets ont également au moins un nom et des propriétés intrinsèques, appelés occurrences. Ce langage permet une grande flexibilité de représentation des connaissances, particulièrement pour la modélisation de relations sémantiques complexes (n-aires).

OWL, Ontology Web Language (Hendler *et al.*, 2004), permet de formaliser une ontologie (Gruber, 1995), ou plus globalement des ressources terminologiques et ontologiques (Bourrigault *et al.*, 2003), par la définition des concepts utilisés pour décrire et représenter un domaine de connaissance. Ce langage standardise la sémantique de ces concepts par un ensemble de propriétés, de relations et de contraintes. Le formalisme utilisé correspond à ceux de certaines logiques de description. Ces différents formalismes sont utilisés dans nos projets.

Même si de plus en plus de documents sont créés dynamiquement à partir de bases de données, l'information textuelle non structurée est encore prédominante. Or, la communauté du Traitement Automatisé du Langage Naturel est précisément spécialisée dans la représentation des connaissances à partir de documents textuels. Ainsi, il nous paraît naturel d'utiliser les méthodes et outils développés par cette communauté de recherche afin de traiter linguistiquement les ressources textuelles

par des moyens informatisés. Parmi ces technologies linguistiques existantes, nous focaliserons notre recherche sur l'Extraction d'Information (IE).

L'IE est composée de plusieurs méthodes linguistiques pour trouver et extraire de l'information pertinente à partir d'un corpus documentaire représentatif d'un domaine d'application donné. Cette information extraite est principalement composée d'entités nommées comme les noms propres (personnes, organisations, lieux, etc.), les nombres (montants, pourcentages, mesures, etc.) et les dates (absolues et/ou relatives). La notion même d' « entité nommée » a été définie lors des différentes conférences MUC (Message Understanding Conferences[1]). L'objectif de ces conférences était de mesurer la précision et l'efficacité des technologies développées pour extraire des entités nommées et des relations sémantiques (ou scénarios) entre ces entités nommées à partir de documents textuels semi-structurés.

Grâce aux fonctionnalités offertes par les technologies du Traitement Automatique du Langage Naturel, des solutions adaptées aux besoins du Web Sémantique peuvent être implémentées comme :

– La construction semi-automatique de vocabulaires/terminologies d'un domaine à partir d'un corpus documentaire représentatif.

– L'enrichissement semi-automatique de bases de connaissance par les entités nommées et les relations sémantiques extraites des documents textuels.

– L'annotation sémantique de ces documents.

Nous travaillons sur l'architecture d'un portail Web Sémantique intégrant des outils linguistiques existants. D'après les solutions ci-dessus, l'objet de nos recherches porte sur la possibilité d'utiliser les méthodes d'IE afin d'annoter des documents textuels et d'enrichir une base de connaissance existante. Afin d'améliorer la productivité et la qualité de l'annotation humaine, les nouvelles informations extraites et ajoutées à la base de connaissance vont à leur tour enrichir les ressources linguistiques nécessaires aux outils d'IE. Par conséquent, le système est capable de réutiliser les résultats de ses composants afin d'améliorer sa performance globale. L'architecture de ce portail est détaillée dans la section suivante de ce document.

3. Le portail Web Sémantique d'ITM

Notre solution se base sur l'outil « Intelligent Topic Manager™ » (ITM) de la société Mondeca. ITM est une plateforme logicielle pour la gestion de connaissance. ITM intègre un portail sémantique (Amardeilh et al., 2004) fournissant quatre fonctions clefs : gestion de la base de connaissance, gestion de la terminologie, indexation documentaire et recherche d'information avancée, publication de contenus. L'ontologie cliente, formalisée en OWL, impose ses contraintes de modélisation à la base de connaissance (implémentée en XTM Topic Maps), aux

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

interfaces utilisateurs ainsi qu'à toutes les fonctionnalités du portail. Les éléments de la base de connaissance pointe vers les documents, accessibles par URL sur Internet ou dans un système de gestion de contenus.

3.1. *Spécification d'une ontologie dans ITM*

ITM fournit une meta-ontologie, définissant les concepts de base utilisés dans la plateforme comme « Class », « Attribute », ou la relation « Class-Subclass ». Cette meta-ontologie contrôle la structure de toutes les ontologies applicatives qui seront déployées dans le portail. Informellement une ontologie dans ITM définit une hiérarchie de classes avec les types d'attributs et les types de relations possibles pour chacune d'entre elles. Formellement, une ontologie dans ITM est composée de (Maedche *et al.*, 2003) :

- Un ensemble de classes C. C représente les classes possibles des instances du domaine.
- Une hiérarchie H. Les classes sont reliées par des relations irréflectives, acycliques, transitives H, ($H \subset C * C$). H(C1,C2) veut dire que C1 est une sous-classe de C2. Notez qu'une classe peut ne pas faire partie de la hiérarchie.
- Un ensemble de types d'attributs D. D est l'ensemble des types possibles pour des attributs qui peuvent être attachés aux instances du domaine.
- Un ensemble de types d'association A. A est l'ensemble des types possibles pour des associations entre des instances dans le domaine.
- Un ensemble de types de rôles R. R est l'ensemble des types possibles pour des rôles joués par les instances du domaine.
- Une fonction de contrainte d'attribut AC: $C \rightarrow P(D)$, définissant pour chaque classe, la liste des types possibles d'attributs sur les instances de cette classe.
- Un ensemble de contraintes d'association RC. RC est une relation (c,r,a), avec $c \in C$, $r \in R$, $a \in A$. Une relation contrainte (c,r,a) exprime le fait qu'une instance de la classe c peut jouer un type de rôle r dans l'association de type a.

3.2. *Spécification d'une base de connaissance dans ITM*

Toute base de connaissance dans ITM est contrôlée par une ou plusieurs ontologies. Ainsi, les classes d'instances et les types de relations possibles sont ceux définis dans ces ontologies. Informellement, une base de connaissance dans ITM est un ensemble d'instances connectées entre elles par un réseau sémantique d'associations et de rôles. Chaque instance appartient à une seule classe. Chaque association et chaque rôle possède un et un seul type. Par exemple, l'expression « Mr X est le PDG de la société Y » est modélisée par le réseau suivant : « Mr X [classe Personne] joue un rôle de type 'employé' dans une association de type 'emploi', dans laquelle PDG [classe Fonction] joue un rôle de type 'poste' et où la société Y [classe

Société] joue le rôle d'‘employeur' ». Plus formellement, une base de connaissance dans ITM est composée de (Maedche *et al.*, 2003) :

- Un ensemble d'instances I. Tout élément de I possède une classe $c \in C$.
- Un ensemble d'attributs K. Tout élément de K possède un type $d \in D$.
- Une fonction d'attribut Ik: $I \rightarrow P(K)$. Ik définit pour chaque instance $i \in I$ la liste des attributs de i.
- Un ensemble d'associations B. Tout élément de B possède un type $a \in A$.
- Un ensemble de rôles S. Tout élément de S possède un type $r \in R$.
- Une fonction « rôle-instance » Ri: $S \rightarrow I$.
- Une fonction « association-rôle » Ra: $S \rightarrow B$. (Ri et Ra expriment ensemble le fait qu'un rôle connecte une instance à une association).

La base de connaissance peut être enrichie manuellement par l'utilisateur final (documentaliste, veilleur, etc.) à travers des formulaires en ligne. L'utilisateur sélectionne dans l'ontologie la classe de l'élément qu'il veut instancier, et un formulaire est généré avec les attributs et les relations autorisés sur les instances de cette classe. Ainsi l'information requise pour créer le nouvel élément est dépendant de la classe sélectionnée par l'utilisateur. Néanmoins, l'enrichissement manuel d'une base de connaissance a ses limites : coûteux en temps, cause d'erreurs et dépendant fortement de l'utilisateur même si l'ontologie opère un certain contrôle. Ces limites se répercutent également sur la productivité, la qualité et la fréquence de traitement de l'annotation sémantique des ressources documentaires. Pour toutes ces raisons, nous avons décidé d'améliorer le portail sémantique d'ITM en intégrant un outil d'extraction d'information afin d'aider l'utilisateur à annoter les documents et enrichir la base de connaissance.

4. La contribution du TALN au Web Sémantique

L'analyse linguistique est effectuée par l'Insight Discoverer™ Extractor (IDE) développé par la société Temis (Grivel *et al.*, 2001). Cet outil d'extraction d'information implémente une méthode d'automates à états finis basée sur un prétraitement regroupant la segmentation des documents en unités textuelles, la lemmatisation et l'analyse morpho-syntaxique de ces unités textuelles. En sortie, l'IDE™ produit un arbre conceptuel étiqueté. Chaque nœud de l'arbre porte le nom d'une étiquette sémantique attribuée à l'unité textuelle extraite en fonction du domaine traité.

Les outils du TALN peuvent participer à la fois à l'annotation documentaire, i.e. l'ajout d'étiquettes sémantiques à un document, ainsi qu'à l'enrichissement de la base de connaissance, i.e. le peuplement de cette base avec l'information contenue dans le document. Les étiquettes sémantiques et la base de connaissance sont contraintes par l'ontologie du domaine. Par conséquent, la correspondance entre les

concepts de l'ontologie du domaine, les instances de la base de connaissance et les étiquettes de l'arbre conceptuel extrait doit être finement définie à l'aide de règles d'acquisition. Une fois ces règles préalablement définies, le système parcourt le document dans son entier et fournit des suggestions d'annotations à l'utilisateur par l'interprétation de ces règles d'acquisition (cf. Figure 1). L'utilisateur doit ensuite valider l'information extraite permettant l'enrichissement de la base de connaissance. Ce système fait donc appel à un processus semi-automatique.

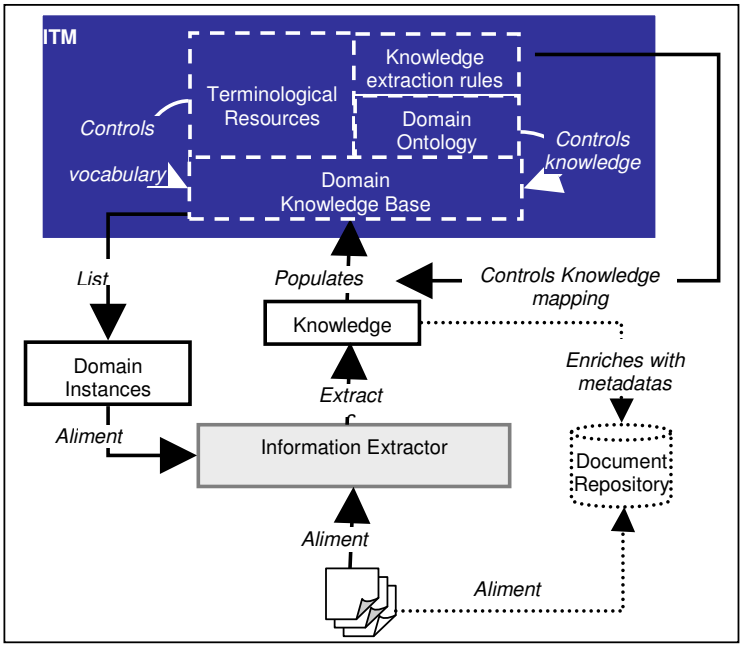

Figure 1. *L'architecture du portail Sémantique avec l'extracteur d'information.*

Nous allons illustrer ce processus à travers un exemple issu d'un de nos projets. Il s'agit du domaine de l'édition juridique : les auteurs d'articles juridiques doivent se tenir informés de l'ensemble des textes de lois et des décisions des cours de justice. Ainsi pour tout texte paru, une référence est enregistrée dans la base de connaissance avec toutes les informations la concernant ainsi que les références aux autres textes de lois cités. Le corpus utilisé dans notre exemple est constitué uniquement de compte rendus de décisions issues de cours de cassation à propos de divorces ou de contrats de travail. Les comptes rendus se divisent en deux parties : tout d'abord un entête semi-structuré représente les informations liées à cette

décision (date, cour de cassation, n° de décision, n° de pourvoi, etc.) et ensuite le corps du document (texte non structuré) décrit dans l'ordre les parties impliquées, les motifs, l'argumentation avec les références aux textes de lois codés (dits « TC », par exemple « Code civil ») et non codés (dits « TNC » comme « Décret du 30 septembre 1953 »).

4.1. Mise en correspondance des étiquettes linguistiques avec les concepts de l'ontologie du domaine

Le processus de mise en correspondance doit être conforme aux exigences suivantes :

- **Facilité d'utilisation**. Le processus de mise en correspondance requiert des experts de différentes branches (expert du domaine d'application, expert en linguistique et expert en représentation des connaissances). Ainsi la solution choisie doit être facilement comprise par ces trois parties et doit permettre un processus itératif.
- **Indépendance entre la structure de l'ontologie et la structure des extractions linguistiques**. Utiliser un traitement naturel du langage pour enrichir une base de connaissance ne doit pas contenir de nouvelles contraintes sur la façon dont l'ontologie est modélisée ou sur le format de l'extraction linguistique.
- **Capacité à évoluer**. Le système doit être capable de prendre en compte les évolutions à la fois de l'ontologie et de l'outil linguistique.
- **Complétude**. Le système doit être capable de retrouver toute information donnée par les extractions linguistiques.
- **Standardisation**. Le système ne doit pas être dépendant de l'outil linguistique utilisé.

L'arbre étiqueté évalué fournit par l'outil linguistique IDE™ est parcouru selon une méthode en profondeur par ordre préfixe (Cf. Figure 2). Ce parcours permet de repérer l'information pertinente afin de la rapprocher d'un concept de l'ontologie, que ce dernier représente une classe $c \in C$, un type d'attribut $d \in D$, un type d'association $a \in A$ ou un type de rôle $r \in R$. Pour cela, nous modélisons des règles d'acquisition de la connaissance qui permettront de déclencher la création d'une instance du concept de l'ontologie à chaque nœud correspondant de l'arbre conceptuel.

Par conséquent, nous devons associer manuellement chacun des nœuds de l'arbre conceptuel à son concept ontologique relatif. Le tableau ci-dessous résume les différents cas possibles :

- Une étiquette peut instancier un seul concept, comme /art num.
- Un concept peut être instancié par plusieurs étiquettes, comme le sujet « Personne ».

– Une étiquette peut servir à instancier plusieurs concepts du même type, comme /MEMBRES COUR.

– Une étiquette peut servir à instancier plusieurs concepts de différents types, comme /REFERENCE.

– Une étiquette peut ne correspondre à aucun concept, comme /MOTIF.

– Un concept peut n'être instanciable par aucune étiquette, comme le rôle « Origine Lien ».

```
/REFERENCE DECISION(cassation 10400510)
   /FORMATION(CIV . 1)
      /Chambre civile(CIV . 1)
   /JURIDICTION(COUR DE CASSATION)
   /DATE SEANCE(Audience publique du 23 mars 2004)
      /DATE(23 mars 2004)
         /MonthDayNumber(23)
         /month(mars)
         /YearNumber(2004)
   /Noms-de-personnes(M. BOUSCHARAIN , président)
      Nom(M. BOUSCHARAIN)
      role(président)
         /Role/Juridique(président)
   /DECISION/ARRET/ARRET DIFFUSE(Arrêt n° 510 F-D)
         num(510 F-D)
   /POURVOI(Pourvoi n° F 02-19.839)
      num(F 02-19.839)
   …
   /REFERENCE(article L. 311-37 du Code de la consommation)
      ref(article L. 311-37 du Code de la consommation)
         /ARTICLE unique(article L. 311-37)
            art num(L. 311-37)
         TEXTE(Code de la consommation)
            /CODE/Code consommation(Code de la consommation)
```

Figure 2. *Extrait d'un arbre conceptuel d'un compte rendu de décision juridique.*

Dans les cas où une étiquette peut instancier plusieurs concepts, notamment de types différents, il faut alors se servir du contexte des autres nœuds, ascendants, descendants ou frères pour résoudre les ambiguïtés. Par exemple, si le nœud « /REFERENCE » a un nœud fils « /article », le concept « Réf Editoriale Législative TNC article » sera instancié, sinon il s'agira du concept « Réf Editoriale Législative TNC ». Pour ces raisons, nous avons choisi, dans une première étape, d'implémenter ces règles d'acquisition en langage XPath[2]. En effet, ce langage permet de parcourir un arbre (document XML, arbre conceptuel, etc.), d'atteindre directement n'importe

2 Site web du W3C : http://www.w3.org/TR/xpath

lequel de ses nœuds et à partir d'un nœud quelconque de sélectionner n'importe lequel de ses ascendants, descendants ou frères.

Nom de l'étiquette linguistique	Nom du concept ontologique	Type dans la base	Contexte
/nom lex	Personne	classe ∈ C	
/noms lex	Personne	classe ∈ C	
/MEMBRES COUR	Personnalité Juridique	classe ∈ C	∃ Descendant = /Juridique
	Personnalité Politique	classe ∈ C	∃ Descendant = /Politique
/REFERENCE	Réf Editoriale Législative TNC	classe ∈ C	∃! Fils = /article
	Réf Editoriale Législative TNC Article	classe ∈ C	∃ Fils = /article
	Renvoi simple	type d'association ∈ A	∃ Père = /REFERENCE DECISION
	Cible lien	type de rôle ∈ R	∃ Père = /REFERENCE DECISION
/art num	Num Article	type d'attribut ∈ D	
/MOTIF			
	Origine Lien	type de rôle ∈ R	

Tableau 1. *Exemples de concordances entre étiquettes linguistiques et concepts*

Cette solution répond clairement aux exigences décrites ci-dessus, le système étant séparé de l'ontologie, autorisant ainsi une plus grande flexibilité et indépendance. Il est également standard, facile à utiliser et complet puisque n'importe quelle partie d'un arbre peut être atteinte par une expression XPath. A travers cette méthode, non seulement les extractions linguistiques mais aussi d'autres types de documents basés sur des arbres conceptuels peuvent être intégrés dans la base de connaissance contrôlée par l'ontologie, comme les documents structurés ou semi-structurés en XML.

4.2. Déclenchement d'une règle d'acquisition

Une fois que l'extraction a été effectuée, chaque étiquette de l'arbre conceptuel représentant une information pertinente du domaine étudié est comparée à l'ontologie du domaine grâce aux règles d'acquisition de la connaissance définies ci-dessus. A chaque nœud pertinent, l'action d'instanciation de la base de connaissance, associée à toute règle d'acquisition, est déclenchée. Toutefois, afin d'éviter les doublons dans la base de connaissance, un contrôle est effectué avant la création du concept pour vérifier son existence dans la base de connaissance. Une fois le parcours de l'arbre terminé, l'utilisateur peut visualiser toutes les nouvelles instances de la base de connaissance au moyen d'une interface de validation. A partir de cette interface, l'utilisateur peut modifier et/ou supprimer une instance créée, ainsi qu'en ajouter de nouvelles. Grâce à cette interface, l'utilisateur peut contrôler la qualité des annotations et de la base de connaissance ainsi enrichie.

Par ailleurs, l'objectif principal du système étant d'aider les utilisateurs à annoter les documents et à enrichir les bases de connaissance, ce système ne doit pas produire l'effet inverse en surchargeant de travail les utilisateurs par cette tâche de validation des résultats. C'est pourquoi, le système va automatiquement mettre à jour les ressources linguistiques (lexiques, dictionnaires ou autres) avec les nouvelles informations extraites et validées afin qu'elles soient par la suite reconnues et interprétées automatiquement. Le système doit donc sélectionner chaque classe définie dans l'ontologie (comme `Personne`, `Réf Editoriale Jurisprudence`, etc.) où une nouvelle instance peut être trouvée puis il va extraire de la base de connaissance les nouvelles instances validées, i.e. leur nom ainsi que leurs différents alias si ceux-ci existent (par exemple l'instance « Jacques Chirac » et son alias « J. Chirac » de la classe `Personne`). Enfin, le système va exporter ces listes d'instances de classe selon un format XML spécifique afin qu'elles soient importées dans l'outil linguistique et ajoutées aux ressources linguistiques correspondantes.

Nous supposons qu'après un certain laps de temps, toute l'information clef du domaine d'application sera intégré dans la base de connaissance et les ressources linguistiques. L'application cliente convergera vers une base de connaissance fiable. Ainsi, les utilisateurs ne devront pas valider autant d'information qu'au début et pourront passer de plus en plus de temps à utiliser le système pour des recherches d'information ou de la publication. En conséquence, plus le système fonctionnera, plus le gain de productivité sera important.

5. Expérimentations et résultats

5.1. *Méthodologie des expérimentations*

Notre corpus d'expérimentation est constitué de 36 comptes rendus de décisions de cour de cassation. Sur ces 36 comptes rendus que nous avions à disposition, quatre seulement ont servi à définir les règles d'acquisition manuellement. Les 32 documents restants ont été utilisés comme corpus de test. Après réception des patrons d'extraction construits par les linguistes de Temis, nous avons traité l'ensemble du corpus de test et recueilli pour chaque arbre conceptuel. Nous avons comparé les étiquettes linguistiques avec chaque concept repéré et constaté quels étaient ceux correctement créés, incorrectement créés ou non créés dans la base de connaissance.

Afin d'évaluer quantitativement les résultats de ces traitements, nous avons utilisé les mesures de précision et de rappel, définies pour mesurer soit des résultats en recherche d'information (cf. conférences TREC), soit des résultats d'extraction d'information (Cf. conférences MUC). Dans notre cas, nous avons appliqué ces mesures aux extractions linguistiques étiquetées vis à vis des concepts instanciés dans la base de connaissance.

Nous obtenons ainsi les deux formules suivantes :

– **Précision** : nombre d'instances correctement acquises / nombre d'instances acquises

– **Rappel** : nombre d'instances correctement acquises / nombre d'instances existantes dans l'arbre conceptuel

5.2. *Analyse des résultats et problèmes soulevés*

Suite à l'analyse des 32 documents du corpus de test, et à partir des mêmes règles d'acquisition définies précédemment, le tableau ci-dessous présente les résultats de l'ensemble des concepts présents dans le corpus des extractions linguistiques. Un ensemble de 1765 concepts de l'ontologie répartis en classes, attributs, associations et rôles sont présents dans les arbres conceptuels du corpus. Parmi ces concepts, 975 ont été correctement instanciés par les règles, 257 incorrectement instanciés et enfin 533 non instanciés. En moyenne, nous obtenons donc un rappel de 0,55 et une précision de 0,79.

Type de concept	Nombre de concepts dans l'arbre (A)	Nombre instanciés corrects (B)	Nombre instanciés incorrects (C)	Nombre non instanciés (D)	**Rappel (B/A)**	**Précision (B/B+C)**
Classes	585	432	139	14	**0,74**	**0,76**
Attributs	798	329	0	469	**0,41**	**1**
Associations	80	69	0	11	**0,93**	**1**
Rôles	302	145	118	39	**0,48**	**0,55**
Total	**1765**	**975**	**257**	**533**	**0,55**	**0,79**

Tableau 2. *Résultats des expérimentations par type de concept*

En résumé, même si la précision est plutôt satisfaisante pour une première expérimentation, nous constatons qu'un nombre important d'unités textuelles, pourtant correctement étiquetées dans l'arbre, ne sont pas instanciées par la suite, surtout en ce qui concerne les attributs et les associations. D'autres concepts sont incorrectement instanciés, notamment les classes. Ceci est principalement dû à un problème de redondance lié à des règles conflictuelles. Ce problème se répercute alors sur les rôles avec le non respect des contraintes modélisées dans l'ontologie, notamment les cardinalités, engendrant pour une même association plusieurs rôles du même type au lieu d'un seul.

5.2.1. Nécessité d'une analyse contextuelle

Nous constatons également que certains problèmes rencontrés sont dus à la nécessité de prendre en compte le contexte des étiquettes générées par l'extraction linguistique. Dans le cas des noeuds ascendants, prenons par exemple l'étiquette « /num » qui permet de créer une occurrence numéro : si le noeud père est « /ARTICLE », l'occurrence créée sera un numéro d'article alors que si ce même noeud est « /POURVOI », l'occurrence sera un numéro de pourvoi. Le contexte des ascendants est notamment primordial pour la création des occurrences des sujets.

```
/Noms-de-personnes(M. BOUSCHARAIN , président)
    Nom(M. BOUSCHARAIN)
    role(président)
       /Role/Juridique(président)
```

Figure 3. *Exemple d'une analyse contextuelle*

Le contexte des nœuds descendants peut également apporter des précisions par rapport à la création d'une classe de sujet ou d'un type d'association. Dans la Figure

3, l'étiquette « /Noms-de-personnes » permet de savoir que le nœud concerne un concept « Personne » dans l'ontologie. Or, cette classe à deux sous-classes, « Personnalité Juridique » et « Personnalité Politique ». Une analyse des descendants du nœud « /Noms-de-personnes », et notamment de la présence de l'un ou l'autre des noeuds « Juridique » ou « Politique », permet de préciser la classe du sujet instancié.

5.2.2.Nécessité d'une gestion de résolution des conflits

Enfin, il existe des règles XPath identiques décrivant différents concepts ontologiques et provoquant des conflits dans l'instanciation de ces concepts dans la base de connaissance. Par exemple, il existe une même règle qui instancie à la fois des sujets de la classe « Ref Editoriale TC » et des sujets de la classe « Ref Editoriale TC Article », ce qui cause les problèmes de redondance. Par contre, cette règle peut indépendamment instancier un type d'association sans pour autant causer une erreur dans la base de connaissance. Par conséquent, les règles d'acquisition peuvent donc être identiques à la condition qu'elles ne permettent pas l'instanciation d'un même type de concept (sujet, associations, rôles ou occurrences).

6. Travaux relatifs

Le projet de recherche présenté dans ce document est innovant par plusieurs aspects et principalement par sa capacité à mettre à jour les ressources linguistiques avec le résultat de l'enrichissement de la base de connaissance. De plus, l'annotation documentaire et le peuplement de la base de connaissance sont validés à travers la même interface, qui fournit une convivialité intéressante et un gain de productivité important pour l'utilisateur. La plupart des autres systèmes de ce type s'intéressent à un seul aspect seulement, soit la problématique de l'annotation soit celle de l'enrichissement au lieu de les combiner.

Plusieurs méthodes ont été proposées pour extraire la terminologie d'un domaine à partir de textes et d'utiliser ces extractions pour construire une ontologie comme OntoLearn (Missikof *et al.*, 2002). Or, notre système n'a pas pour but de créer de nouvelles ontologies mais plutôt d'enrichir une ontologie existante en peuplant sa base de connaissance sous-jacente. Dans OntoKnowledge (Fensel *et al.*, 2002) une base de données RDF enregistre l'information extraite grâce à son outil d'extraction d'information. Mais cette base de données n'est contrainte par aucune ontologie et nous préférons, en ce qui concerne l'implémentation de notre base de connaissance, utiliser les Topic Maps car ainsi que nous l'avons dit, cette formalisation permet de représenter des relations sémantiques complexes de manière plus intuitive que le RDF.

L'annotation de ressources documentaires est un problème abordé par des projets tels que Annotea3 (Kahan *et al.*, 2001). Dans ce projet, les documents sont annotés avec des commentaires et des métadonnées RDF très basiques comme le nom de l'auteur, la date, la source, etc. L'utilisateur doit manuellement créer ses propres annotations. Mais ces systèmes sont coûteux en temps et en ressources et limités quant aux types d'annotations produites. Dans notre système, les annotations sont produites semi-automatiquement et concernent, outre les métadonnées administratives comme la date, la source, l'auteur, etc., des descripteurs de thesaurus ou des entités nommées représentatifs du sujet du document. D'autres outils utilisent des technologies du langage naturel comme S-Cream (Handschuh *et al.*, 2002), MnM (Vargas-Vera *et al.*, 2002), Amilcare[4] (Ciravegna, 2001) et Melita[5] (Dingli, 2003), qui assistent l'utilisateur lors de l'annotation de documents textuels du Web. Ces deux derniers systèmes utilise des méthodes d'apprentissage, l'algorithme d'extraction d'information (LP)², afin d'adapter l'outil d'IE à de nouveaux domaines d'applications et pour généraliser les règles d'induction. L'inconvénient majeur est que l'algorithme a besoin d'un corpus pré-étiquetté en XML pour chaque scénario sélectionné par l'utilisateur. L'utilisateur doit de plus corriger les nouvelles étiquettes insérées jusqu'à obtention d'un seuil de précision prédéfini. Les extractions restent souvent limitées en ce qui concerne la complexité des relations sémantiques contrairement à notre système. Et ces systèmes d'annotation ne traitent pas l'information extraite dans les documents pour simultanément enrichir la base de connaissance.

A la conférence ISWC2003, et plus particulièrement lors de l'atelier « Human Language Technologies for the Semantic Web » (HLT4SW), plusieurs projets ont émergé autour de la problématique de l'utilisation d'outils linguistiques aidant à créer de nouvelles applications Web Sémantique. Un point commun important entre la plupart de ces projets est leur utilisation de la plateforme d'ingénierie linguistique GATE (General Architecture for Text Engineering) (Cunningham *et al.*, 2002). GATE fournit des ressources lexicales, syntaxiques et sémantiques « open-source » pour construire son propre outil linguistique. GATE est particulièrement pertinent pour le développement d'applications d'extraction d'information. C'est pourquoi il est de plus en plus utilisé dans les projets Web Sémantique ayant rapidement besoin d'un outil d'IE facile à mettre en place pour l'annotation sémantique ou la création d'ontologies. Parmi les systèmes du HLT4SW, KIM (Knowledge Information Manager) (Popov *et al.*, 2003) semble être le plus proche de notre approche. Il extrait des entités nommées, leurs attributs et alias ainsi que quelques relations sémantiques simples comme des situations géographiques de personnes et d'organisations. Puis il peuple la base de connaissance avec ces informations

[3] Annotea Project website : http://www.w3.org/2001/Annotea
[4] Amilcare Project website : http://nlp.shef.ac.uk/amilcare
[5] Melita Project website : http://www.dcs.shef.ac.uk/~alexiei/Melita.htm

extraites mais n'annote pas le document avec. Une autre application relative est le projet Artequakt (Alani *et al.*, 2003) qui recherche sur Internet comment répondre à une requête sur des artistes, extrait la connaissance de toutes les pages trouvées et après l'enrichissement de la base de connaissance peut générer des biographies d'artistes grâce à un outil de génération de langage naturel automatique. Comparé à ces deux projets, notre approche est plus complète et plus riche puisque nous annotons les documents et enrichissons la base de connaissance dans un même traitement de la ressource documentaire. Le système est capable d'extraire et d'instancier des relations sémantiques complexes, i.e. des types d'associations comportant plusieurs rôles bien définis par ses classes. Et enfin notre système met à jour les ressources linguistiques de l'outil linguistique avec les nouvelles connaissances validées pour augmenter en performance, fiabilité et productivité. Plus le système travaille, moins les utilisateurs passent de temps à annoter, diminuant ainsi le fardeau de l'annotation documentaire.

7. Conclusions et Travaux futurs

Notre système propose donc une solution innovante à la fois d'annotation de ressources documentaires et d'enrichissement d'une base de connaissance contrainte par l'ontologie du domaine à partir d'extractions linguistiques grâce à la définition de règles d'acquisition. Ces annotations et cet enrichissement semi-automatique de la base de connaissance sont particulièrement pertinents dans le cadre de la veille scientifique et économique. En effet, ils vont proposer des suggestions aux utilisateurs leur permettant de recueillir l'information essentielle à leur domaine d'application à partir de diverses sources de contenus (web, systèmes de gestion de contenu, autres). Une fois ces suggestions validées par un processus simple et convivial, ils vont pouvoir l'utiliser, l'échanger et la publier à travers un portail sémantique dédié.

A partir des problèmes soulevés dans la première implémentation, nous avons défini les priorités suivantes dans nos futurs travaux de recherche :

- Amélioration du parcours de l'arbre conceptuel pour gérer plus de complexité dans les règles par une contextualisation des règles plus riche.
- Détection des conflits dans le paramétrage des règles d'acquisition.
- Amélioration des écrans de validation et d'acquisition de nouvelles instances pour l'utilisateur.
- Vérification de l'existence des concepts à instancier dans la base de connaissance pour éviter les doublons et le non respect de la cardinalité, notamment pour les rôles participant à une association.

La résolution de certains de ces problèmes permettrait d'améliorer rapidement les performances actuelles du système, notamment en ce qui concerne les associations et les rôles. Par contre, d'autres problèmes comme la définition de règles complexes et la gestion des règles conflictuelles devront être mûrement réfléchis en terme de solution. Il reste également le problème de cohérence et de maintenance entre des règles d'acquisition qui deviendront de plus en plus nombreuses, surtout si l'ontologie cliente comporte un nombre important de concepts à instancier. Le paramétrage manuel des règles d'acquisition est susceptible de comporter des erreurs et encore fastidieux pour l'administrateur de ces règles. En effet, si les cartouches linguistiques ou si l'ontologie cliente sont modifiées, alors tout le paramétrage doit être mis à jour. Nous sommes donc en train de développer de nouvelles interfaces permettant une gestion simple et conviviale de ces règles.

En résumé, nous sommes convaincus que l'utilisation des outils linguistiques du TALN va grandement faciliter le développement d'applications Web Sémantique, notamment dans le cadre de la veille scientifique et économique et de l'édition. Les outils linguistiques deviendront des composants obligatoires du succès de ces applications (Bontcheva *et al.*, 203). Pour le moment, notre solution intègre uniquement un outil d'extraction d'information mais d'autres outils linguistiques peuvent également fournir d'intéressants et pertinents services pour les applications, comme la Catégorisation pour l'annotation documentaire, le Résumé de textes ou même la Génération de langage naturel pour la présentation des résultats de requêtes, etc.

8. Bibliographie

Alani H., Kim S., Millard D. *et al.*, « Automatic Extraction of Knowledge from Web Documents », in *Proceedings of the Second International Semantic Web Conference*, Workshop on Human Language Technology for the Semantic Web and Web Services, Floride, octobre 2003, pp. 77-88.

Amardeilh F. & Francart T. « A Semantic Web Portal with HLT Capabilities », In *Actes du colloque « Veille Stratégique Scientifique et Technologique »* (VSST2004), Toulouse, octobre 2004, vol. 2, p 481-492.

Berners-Lee T., *Weaving the Web*, Eds Harper, San Francisco, 1998, 226 p.

Bontcheva K. et Cunningham H., « The Semantic Web: A New Opportunity and Challenge for Human Language Technology », in *Proceedings of the Second International Semantic Web Conference*, Workshop on Human Language Technology for the Semantic Web and Web Services, Floride, octobre 2003, pp. 89-96.

Bourigault D, Aussenac-Gilles N. et Charlet J., « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », *Revue d'Intelligence Artificielle*, 18(4), 2003, 24 p.

Ciravegna F., « Adaptive Information Extraction from Text by Rule Induction and Generalisation », In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, août 2001, 6 pp.

Cunningham H., Maynard D., Bontcheva K. *et al.*, « GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications », In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphie, 2002, 8 pp.

Dingli A., « Next Generation Annotation Interfaces for Adaptive Information Extraction », In *Proceedings of the 6th Annual Computer Linguists UK Colloquium (CLUK03)*, Edinburgh, 6-7 janvier 2003, 5 pp.

Fensel D., Bussler C., Ding Y. *et al.*, « Semantic Web Application Areas », In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems*, Stockholm, 27-28 juin 2002, 14 pp.

Grivel L., Guillemin-Lanne S., Lautier C. *et al.*, « La construction de composants de connaissance pour l'extraction et le filtrage de l'information sur les réseaux », In *Filtrage et résumé automatique de l'information sur les réseaux*, 3ème congrès du Chapitre français de l'ISKO International Society for Knowledge Organization, 5-6 juillet 2001, 9 pp.

Gruber T., « A Translation approach to portable ontology specifications », In *Knowledge Acquisition*, 5(2), 1995, p. 199-220.

Handschuh S., Staab S. et Ciravegna F., « S-CREAM – Semi-automatic CREAtion of Metadata », In *Proceedings of the 13th International Conference on Knowledge Engineering and Management (EKAW 2002)*, Espagne, 1-4 octobre 2002, pp. 358-372.

Hendler J., Horrocks I. *et al.*, *OWL web ontology language reference* W3C Recommendation, 2004.

Kahan J., Koivunen M., Prud'Hommeaux E. *et al.*, « Annotea: An Open RDF Infrastructure for Shared Web Annotations », In *Proceedings of the WWW10 International Conference*, Hong Kong, mai 2001, pp. 623-632.

Katz B., Lin J. et Quan D., « Natural Language Annotations for the Semantic Web », In *Proceedings of the ODBASE 2002*, Irvine, California, octobre 2002, 15 pp.

Laublet P., Reynaud C. et Charlet J., « Sur Quelques Aspects du Web Sémantique », *Assises du GDR I3*, Editions Cépadues, Nancy, décembre 2002, 20 pp.

Lu S., Dong M. and Fotouhi F., « The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications », *Information Research*, Special Issue on the Semantic Web, 7(4), 2002, 12 pp.

Maedche A, Staab S., Stojanovic N. *et al.*, « SEmantic portal : The SEAL Approach », In *Spinning the semantic web : bringing the world wide web to its full potential*, by D. FENSEL *et al.*, The MIT Press, 2003, pp. 317-359.

Missikof M., Navigli R. and Velardi P., « The Usable Ontology: an Environment for Building and Assessing a Domain Ontology », In *Proceedings of the First International Semantic Web Conference*, Sardaigne, juin 2002, pp. 39-53.

Ora L., Swick R., *Resource Description Framework (RDF) Model and Syntax Specification*, W3C Recommendation, 1999.

Park J., Hunting S., *XTM Topic Maps : Creating and using Topic Maps for the Web*, Addison Wesley Eds, Boston, 2003, p. 81-101.

Popov B., Kiryakov A., Manov D. *et al.*, « Towards Semantic Web Information Extraction », In *Proceedings of the Second International Semantic Web Conference*, Workshop on Human Language Technology for the Semantic Web and Web Services, Floride, octobre 2003, pp. 1-22.

Vargas-Vera M., Motta E., Domingue J. *et al.*, « MnM : Ontology Driven Tool for Semantic Markup », In *Proceedings of the Workshop Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002)*, Lyon, 22-23 juillet 2002, p. 43-47.