Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques

Florence Amardeilh, Philippe Laublet, Jean-Luc Minel

Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques

Florence Amardeilh[1 2], Philippe Laublet[1], Jean-Luc Minel[1]

[1] Laboratoire LaLICC, Université Paris IV,
{prenom.nom}@paris4.sorbonne.fr
[2] Mondeca, Département R&D, Paris
{prenom.nom}@mondeca.com

Résumé : Dans cet article, nous présentons une plate-forme de peuplement semi-automatique d'ontologies à partir de documents textuels. Notre plate-forme fournit un environnement permettant la mise en correspondance des extractions linguistiques avec l'ontologie du domaine de l'application cliente à l'aide de règles d'acquisition de connaissance. Ces règles s'appliquent, pour chaque étiquette linguistique pertinente, aussi bien à un concept, qu'à un de ses attributs ou encore à une relation sémantique entre plusieurs concepts. Elles déclenchent l'instanciation de ces concepts, attributs et relations dans la base de connaissance relative à l'ontologie du domaine. Ce papier détaille le processus et présente les premières expérimentations réalisées à partir d'un cas client provenant de l'édition juridique.

Mots-clés : Méthodologie de population de ressources terminologiques et ontologiques, Modèles de connaissances, Ontologies, Fouille de textes, Web Sémantique.

1 Introduction

Dans les métiers liés au monde documentaire comme l'édition, la documentation, la veille stratégique, etc., les professionnels doivent chaque jour traiter de grands volumes de données provenant de diverses sources documentaires. A partir de l'ensemble des documents qu'ils sont chargés d'étudier, ceux-ci doivent d'abord sélectionner les ressources documentaires pertinentes pour leur travail puis, pour chacune d'entre elles, extraire manuellement l'information pertinente. Cette information sert ensuite à annoter le document par un ensemble de descripteurs (termes du thesaurus et entités nommées comme les noms de personnes) et à enrichir leur base de connaissance (entités nommées, attributs de ces entités nommées et relations sémantiques entre ces entités nommées).

Dans le contexte du Web Sémantique, le contenu d'un document peut être décrit et annoté à l'aide de langages de représentation des connaissances comme RDF, XTM et OWL. RDF, the Resource Description Framework[1], est un formalisme de

[1] ORA L. & SWICK R. (1999). Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation.

représentation des connaissances, issu des réseaux sémantiques, dont la syntaxe utilise XML. Il sert à décrire des ressources documentaires par un ensemble de métadonnées (auteur, date, source, descripteurs, etc.). 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 basée sur XML (Park, 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 , appelées occurrences. Ce langage permet une grande flexibilité de représentation des connaissances, particulièrement pour la modélisation de relations sémantiques complexes (n-ary).

OWL, Ontology Web Language[2], permet de formaliser une ontologie (Gruber, 1993), ou plus globalement des ressources terminologiques et ontologiques (Bourigault, 2004), par la définition des concepts utilisés pour représenter un domaine de connaissance. Ce langage permet de décrire ces concepts par un ensemble de propriétés, de relations et de contraintes. Le formalisme utilisé correspond à ceux de certaines logiques de description.

Dans nos projets, nous utilisons RDF pour décrire le contenu d'une ressource documentaire, OWL pour modéliser l'ontologie qui représentera une vision métier ou applicative du domaine étudié et les Topic Maps pour construire la base de connaissance qui contiendra les instances des concepts, propriétés et relations décrits dans l'ontologie du domaine. Les informations pertinentes au domaine, contenues dans les ressources documentaires, servent à instancier la base de connaissance et à créer les annotations documentaires pour pouvoir être interprétables par les machines qui pourront les partager, les publier, les rechercher et les utiliser de manière générale (Laublet, 2002).

L'annotation documentaire et le peuplement d'ontologie dépendent fortement des informations extraites des articles par les professionnels. Ce traitement manuel des documents est extrêmement coûteux en temps et en ressources. L'ensemble du processus pose également des problèmes en terme de productivité et de qualité. Pour toutes ces raisons, les entreprises cherchent de plus en plus à mettre en place des solutions basées sur l'utilisation d'outils linguistiques pour extraire (semi-) automatiquement les informations pertinentes des document textuels. Ces outils du traitement automatique du langage naturel devront s'intégrer étroitement aux futures applications du Web Sémantique et seront même essentiels au développement, à l'acceptation et à l'utilisation du Web Sémantique (Bontcheva, 2003). Grâce aux fonctionnalités offertes par les technologies du Traitement du Langage Naturel, et notamment celles de l'Extraction d'Information, des solutions adaptées aux besoins du

[2] HENDLER J., HORROCKS I. et al. (2004). OWL web ontology language reference, W3C Recommendation.

Web Sémantique peuvent être implémentées telles que : la construction semi-automatique de vocabulaires/terminologies d'un domaine à partir d'un corpus documentaire représentatif ainsi que leur maintenance (Bourigault, 2004) ; l'enrichissement semi-automatique de bases de connaissance par les entités nommées et les relations sémantiques extraites des documents textuels après validation (Kyriakov, 2003) ; l'annotation sémantique de ressources documentaires (Kahan, 2001) (Handschuh, 2002) (Vargas-Vera, 2002).

Néanmoins, nous avons constaté dans nos propres projets applicatifs que les outils linguistiques et la modélisation de l'ontologie du domaine client sont implémentés indépendamment l'un de l'autre et non pas l'un pour l'autre comme c'est le cas dans les travaux de recherche cités ci-dessus. C'est pourquoi nous nous sommes intéressés à la mise en place d'une passerelle entre d'un côté les résultats de ces outils linguistiques et de l'autre l'ontologie.

Dans cet article, nous présentons donc une nouvelle plateforme d'annotation documentaire et d'acquisition de connaissances. Dans la prochaine section de ce papier, nous montrerons l'émergence de la problématique actuelle et nous décrirons la mise en place de notre solution. Ensuite, nous présenterons les résultats de premières expérimentations dans le domaine de l'édition juridique. Ce projet permettra aussi d'illustrer notre travail tout au long de cet article. Enfin, nous synthétiserons ces résultats afin d'apporter de nouvelles réflexions pour conclure dans la dernière partie sur les perspectives futures de nos travaux de recherche.

2 Intégration d'outils linguistiques dans un portail Web Sémantique

2.1 Les outils utilisés

Notre solution est basée sur l'outil « Intelligent Topic Manager™ » (ITM) de la société Mondeca. ITM est une plateforme logicielle pour la gestion de connaissance et l'exploitation d'ontologies. ITM intègre un portail sémantique, décrit dans (Amardeilh, 2004), fournissant quatre fonctions clefs : l'Edition, la Recherche, la Navigation et la Publication. L'ontologie cliente, formalisée en OWL, impose ses contraintes de modélisation à la base de connaissance (implémentée en Topic Maps), aux interfaces utilisateurs ainsi qu'à toutes les fonctionnalités du portail. Les éléments de la base de connaissance pointent vers les documents, accessibles par URL sur Internet ou dans un système de gestion de contenus.

L'analyse linguistique est effectuée par l'Insight Discoverer™ Extractor (IDE) développé par la société Temis. Cet outil implémente une méthode d'automates à états finis (Grivel, 2001) s'appuyant 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 (cf. exemple section 2.2.1).

D'une part, le portail d'ITM™ ne permet pas un enrichissement (semi-)automatisé de sa base de connaissance. D'autre part, l'outil d'extraction d'information, l'IDE™, une fois les extractions réalisées sur un corpus documentaire, présente simplement les informations à l'utilisateur dans une interface html, sans les enregistrer dans une base de données ou plus encore dans une base de connaissance, pour être ultérieurement exploitées. Les deux sociétés ont décidé de collaborer sur plusieurs projets client (services de documentation, veille économique ou édition). Néanmoins, le paramétrage de leurs outils pour une application cliente est toujours réalisé indépendamment l'un de l'autre, chacun comportant ses propres contraintes.

En effet, Mondeca construit l'ontologie du domaine, si elle n'existe pas déjà, en fonction du client, de ses besoins et des données déjà existantes. Temis construit des modules d'extractions linguistiques propres à chaque projet tout en réutilisant, lorsque cela est possible, tout ou partie de modules d'extraction existants (comme celui des entités nommées). Par conséquent, les dénominations des étiquettes linguistiques de l'arbre conceptuel produit par l'IDE sont généralement indépendantes de celles des concepts de l'ontologie même si elles décrivent le même sujet. Il nous faut donc définir un moyen de les faire correspondre et ainsi pouvoir instancier les bons concepts de l'ontologie à partir des extractions linguistiques.

2.2 Intégration ITM/IDE dans le portail Web Sémantique

L'intégration entre les extractions linguistiques de l'IDE et les concepts ontologiques du domaine définis dans ITM doit se faire en plusieurs étapes : 1) parcours de l'arbre conceptuel résultant de l'analyse linguistique ; 2) définition manuelle des règles d'acquisition entre étiquettes linguistiques et concepts ontologiques; 3) déclenchement automatisé des règles d'acquisition sur les textes.

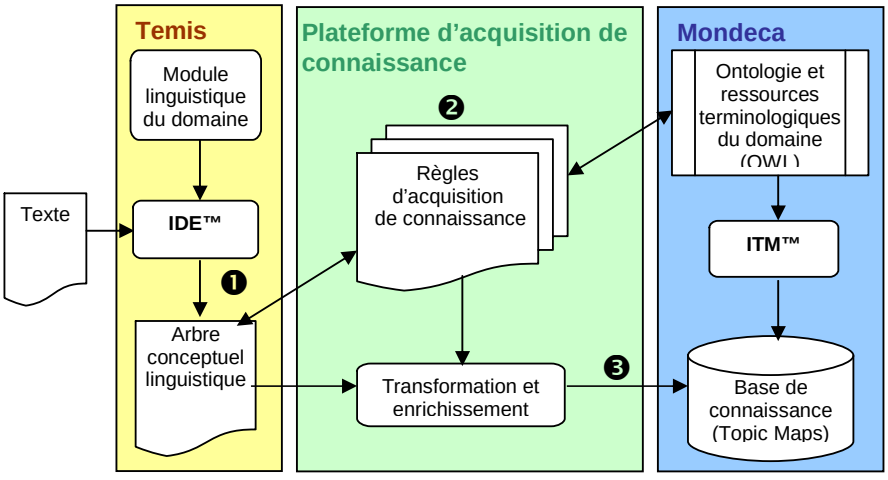

Fig. 1 – Chaîne de traitement de peuplement d'ontologie.

La chaîne de traitement décrite ci-dessus est appliquée à chacun de nos projets client. Nous allons illustrer ce processus à travers un projet concernant le 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 ses propriétés ainsi que les références aux autres textes de lois cités. Le corpus utilisé dans notre exemple est constitué uniquement de comptes rendus de décisions issues de cours de cassation à propos de divorces ou de contrats de travail. Les comptes rendus, cf. **Fig. 2**, se divisent en deux parties : tout d'abord un en-tête 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 »).

CIV. 1 D.S

COUR DE CASSATION

Audience publique du **23 mars 2004** Cassation partielle

M. BOUSCHARAIN, président Arrêt n° 510 F-D

Arrêt n° 510 F-D

Pourvoi n° F 02-19.839

(…) **REPUBLIQUE FRANCAISE**

AU NOM DU PEUPLE FRANCAIS

LA COUR DE CASSATION, PREMIÈRE CHAMBRE CIVILE, a rendu l'arrêt suivant :

Sur le pourvoi formé par Mme X, épouse Y, demeurant xxxxx, 75019 Paris,

(…)

Sur le rapport de Mme G-D, conseiller référendaire, les observations de Me B-H, avocat de Mme X, de la SCP VO, avocat de la société XYZ, les conclusions de Mme P, avocat général, et après en avoir délibéré conformément à la loi ;

Sur le moyen unique, pris en sa seconde branche :

Vu l'article L. 311-37 du Code de la consommation, dans sa rédaction antérieure à la loi n°2001-1168 du 11 décembre 2001 ;

(…)

Fig. 2 – Extrait d'un compte rendu de décision d'une cour de cassation

2.2.1 Arbre conceptuel résultant de l'analyse linguistique

Comme nous l'avons mentionné plus haut, l'IDE produit un arbre conceptuel à partir de chaque analyse linguistique d'un compte rendu (cf. **Fig. 3**). A chaque nœud de cet arbre correspond une étiquette linguistique et sa valeur textuelle issue du compte rendu, rappelée entre parenthèse. Notre solution doit parcourir cet arbre étiqueté évalué afin d'en extraire l'information pertinente et la rapprocher d'un concept de l'ontologie, que ce dernier soit un sujet, un attribut, une association ou un rôle dans la base de connaissance.

```
/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)
```

Fig. 3 – Extrait d'un arbre conceptuel d'un compte rendu de décision juridique

Le parcours de l'arbre est régenté par quelques principes de base : 1) Un arbre possède nécessairement une racine père, représentant ici le plus souvent le document ou le sujet principal (ici la décision elle-même). 2) Le parcours de l'arbre s'effectue en profondeur par ordre *préfixe* : partant de la racine, l'algorithme parcourt d'abord le fils gauche avant de parcourir le fils droit et ainsi de suite récursivement. 3) Deux parcours de l'arbre sont nécessaires : le premier pour acquérir les sujets avec leurs attributs et le second pour acquérir les associations avec les différents rôles joués par les sujets dans celles-ci.

Ces deux parcours sont primordiaux car tous les sujets ne jouent pas nécessairement un rôle dans une association. Ils ne seraient donc pas instanciés si l'arbre était parcouru en n'y repérant que les associations, puis leurs rôles et enfin leurs sujets. C'est notamment le cas des sujets « Personnalités », ayant des attributs « Nom » et « Rôle », qui ne participent à aucune association.

Afin de traiter l'arbre conceptuel, nous avons choisi, dans une première étape, d'implémenter les règles d'acquisition en langage XPath[3]. En effet, ce langage permet de parcourir un arbre (document XML, arbre conceptuel, etc.), d'atteindre directement n'importe lequel de ses nœuds et à partir d'un nœud quelconque de sélectionner n'importe lequel de ses ascendants, descendants ou frères.

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

2.2.2 Définition des règles d'acquisition

Nom de l'étiquette linguistique	Nom du concept ontologique	Type dans la base	Contexte
/nom lex	Personne	Sujet	
/noms lex	Personne	Sujet	
/MEMBRES COUR	Personnalité Juridique	Sujet	∃ Descendant = /Juridique
	Personnalité Politique	Sujet	∃ Descendant = /Politique
/REFERENCE	Réf Editoriale Législative TNC	Sujet	∃! Fils = /article
	Réf Editoriale Législative TNC Article	Sujet	∃ Fils = /article
	Renvoi simple	Association	∃ Père = /REFERENCE DECISION
	Cible lien	Rôle	∃ Père = /REFERENCE DECISION
/art num	Num Article	Occurrence	
/MOTIF			
	Origine Lien	Rôle	

Tableau 1 – Exemples de mise en correspondance d'étiquettes et de concepts

Chaque nœud de l'arbre conceptuel doit manuellement être rapproché de son concept ontologique correspondant, quelque soit son type (sujet, attribut, association et rôle) [4]. Pour cela, nous construisions des règles d'acquisition de la connaissance qui permettront de déclencher la création d'une instance du concept ontologique à chaque nœud correspondant de l'arbre conceptuel. Le tableau ci-dessus résume les différents cas possibles :

- Une étiquette correspond à un seul concept : « /art num » pour l'attribut « Num Article ».
- Plusieurs étiquettes correspondent au même concept : « /Nom lex » et « /Noms lex » pour le sujet « Personne ».
- Une étiquette correspond à plusieurs concepts du même type : « /MEMBRES COUR » pour les sujets « Personnalité Juridique » et « Personnalité Politique ».
- Une étiquette correspond à plusieurs concepts de types différents : « /REFERENCE » pour les sujet « Réf Editoriale Législative TNC » et « Réf Editoriale Législative TNC Article », l'association « Renvoi simple » et le rôle « Cible lien ».
- Une étiquette ne correspond à aucun concept de l'ontologie : « /MOTIF ».
- Un concept n'a pas d'équivalence dans l'ensemble des étiquettes existantes : le rôle « Origine Lien ».

[4] Rappelons que le vocabulaire utilisé est celui des Topic Maps (Park, 2003).

Dans les cas où une étiquette peut instancier plusieurs concepts, il faut alors utiliser le contexte des nœuds ascendants, descendants ou frères pour résoudre les ambiguïtés. Par exemple, si le nœud « /REFERENCE » a un nœud fils « /article », le sujet « Réf Editoriale Législative TNC Article » sera instancié, sinon il s'agira de « Réf Editoriale Législative TNC ».

La première partie d'un compte rendu, et par conséquent des extractions linguistiques, concerne la décision de la cour de cassation. Elle contient donc les attributs du concept représentant cette décision, i.e. « Réf Editoriale Jurisprudence » marqué par l'étiquette « /REFERENCE DECISION ». Il est donc possible de mettre en relation chacun des noeuds de cette première partie avec les attributs correspondants, telle l'étiquette « /FORMATION » avec « formation » dans **Fig. 4**.

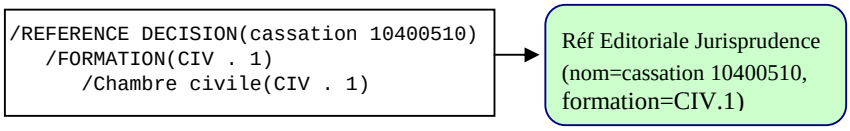

Fig. 4 – Extraction linguistique du Sujet « Réf Editoriale Jurisprudence »

La deuxième partie du document permettra de recueillir d'autres types d'instances de concepts, notamment les personnes, que ce soit des parties ou des personnalités juridiques (avocats, présidents, greffiers, etc.), et les références aux textes juridiques sur lesquels se base l'argumentation des différentes parties. Ces références seront instanciées selon leur concept, texte codé ou non codé, avec leurs attributs (date, type de texte, etc.) puis mises en relation avec la décision à travers l'association « Renvoi simple » et la spécification de leur rôle « Cible lien », cf. **Fig. 5**.

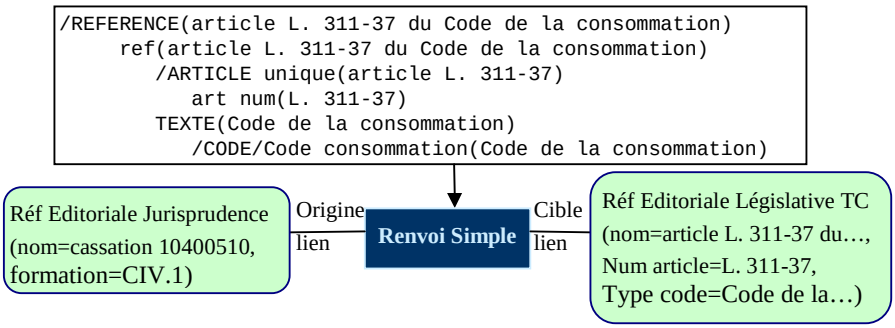

Fig. 5 – Extraction linguistique modélisable en Association « Renvoi simple »

Une fois la mise en correspondance définie, chacune des règles d'acquisition sera formalisée en langage Xpath et ajoutée dans l'ontologie du domaine sur le concept qu'elle va instancier. Par exemple, le concept «Réf Editoriale Législative TC Article » aura dans l'ontologie la règle d'acquisition associée « /REFERENCE DECISION/REFERENCE/ref[ARTICLE and TEXTE] ».

2.2.3 Déclenchement d'une règle d'acquisition

Après analyse linguistique, l'arbre conceptuel du document sélectionné par l'utilisateur est parcouru automatiquement par l'ensemble des règles d'acquisition. 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 de la base de connaissance ainsi enrichie.

3 Expérimentations et résultats

Notre corpus d'expérimentation est constitué de 36 comptes rendus de décisions de cour de cassation. Sur ces 36 comptes rendus, 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 et compilés par les linguistes de Temis, nous avons traité l'ensemble du corpus de test et recueilli 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. La **Précision** mesure le nombre d'instances correctement acquises divisé par le nombre d'instances acquises et le **Rappel** mesure le nombre d'instances correctement acquises divisé par le nombre d'instances existantes dans l'arbre conceptuel.

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 sur l'ensemble des concepts présents dans le corpus des extractions linguistiques. Un ensemble de 1765 concepts de l'ontologie répartis en sujets, attributs (ou occurrences) de ces sujets, 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.

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 sujets. 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.

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)**
Sujets	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 sur les 32 documents du corpus de test

Nous constatons également qu'il est nécessaire d'introduire plus de complexité dans le contexte de l'arbre entre les étiquettes générées par l'extraction linguistique. Pour l'instant, nos règles d'acquisition se limitent aux contraintes sur les fils, les pères ou les frères. Or, le contexte des ascendants est particulièrement important pour la création des attributs des sujets. Prenons par exemple l'étiquette « /num » : si le noeud père est « /ARTICLE », l'attribut sera un numéro d'article alors que si ce même noeud est « /POURVOI », l'attribut sera un numéro de pourvoi. Le contexte des nœuds descendants peut également apporter des précisions par rapport à la création d'un sujet ou d'une association. Dans la **Fig. 6**, l'étiquette « /Noms-de-personnes » renvoie au concept « Personne » dans l'ontologie et qui a deux sous-concepts : « Personnalité Juridique » et « Personnalité Politique ». Une analyse des descendants du nœud « /Noms-de-personnes », et notamment la présence d'un nœud « Juridique » ou « Politique », permet de préciser le sous-concept à instancier.

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

Fig. 6 – Exemple d'une analyse contextuelle

4 Conclusion et discussion

Cette plateforme propose donc une solution innovante 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. A notre connaissance, il

n'existe pas d'approche similaire dans le cadre d'applications pour le Web Sémantique. Bien sûr, d'autres systèmes (Kyriakov, 2003) s'intéressent au peuplement d'ontologies grâce à des outils linguistiques mais leurs ontologies sont modélisées en concordance avec les résultats de leurs extractions linguistiques, à un niveau général et sans relation sémantique complexe (n-aire). A contrario, notre approche permet de peupler une ontologie donnée à partir de n'importe quel outil linguistique, du moment que celui-ci extrait sous forme d'arbre conceptuel les informations pertinentes au domaine concerné.

A partir des problèmes soulevés dans la première implémentation, nous avons défini les priorités suivantes : vérification du respect des cardinalités, notamment pour les rôles participant à une association ; amélioration des deux parcours de l'arbre conceptuel pour gérer plus de complexité dans les règles grâce à une contextualisation plus riche ; détection des conflits liés au recouvrement entre les règles d'acquisition . La résolution de ces priorités permettrait d'améliorer rapidement les performances actuelles du système, notamment en ce qui concerne les associations et les rôles. 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. La construction manuelle des règles est assez fastidieuse et susceptible de comporter des erreurs. Et si les ressources linguistiques ou si l'ontologie cliente sont modifiées, alors ces règles doivent être vérifiées et mises à jour par l'administrateur de ces règles.

C'est pourquoi nous proposons de développer un langage formel de description des connaissances mobilisées pour peupler une ontologie à partir d'un arbre conceptuel. Ce langage s'inspirera de LangText (Crispino, 2003), développé pour modéliser les connaissances linguistiques dans le cadre de l'exploration contextuelle (Desclès, 1991), (Minel & al., 2001). L'un des apports de ce langage est de formaliser de manière déclarative la notion d'espace de recherche, d'indicateur et d'annotation d'une unité textuelle. Nous devons néanmoins adapter ce langage au parcours d'un arbre conceptuel et non d'un texte. Ce langage permettra une meilleure maintenance des connaissances, une plus grande efficacité dans la construction des règles d'acquisition grâce à la gestion des conflits potentiels, et un gain de productivité pour l'utilisateur. En effet, à partir de ce langage nous pourrons générer automatiquement les règles XPath associées et utiliser une feuille de style afin de transformer l'arbre conceptuel au format TM de la base de connaissance ou au format OWL de l'ontologie, selon les besoins de l'utilisateur. Ce langage est en cours d'élaboration.

Enfin, il faut souligner que ce système doit rester assez générique afin de pouvoir définir et appliquer les règles d'acquisition à n'importe quel domaine d'application. De plus, les implémentations des modules linguistiques d'extraction et de l'ontologie cliente peuvent rester indépendantes et c'est aux règles d'acquisition de permettre la transformation d'une étiquette linguistique à un concept instancié de l'ontologie.

Références

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

BONTCHEVA K. & CUNNINGHAM H. (2003). 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, Florida, 20-23 October 2003, p. 89-96.

BOURIGAULT D., AUSSENAC-GILLES N. et CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, Revue d'Intelligence Artificielle, 18(4), 24 pp.

CRISPINO G. (2003). Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes, Thèse sous la direction de Jean-Pierre Desclés, Université Paris-Sorbonne, Décembre 2003, 241 pp.

DESCLES J.-P., JOUIS C., OH H-G., MAIRE REPPERT D. (1991). Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte, Knowledge modeling and expertise transfer, Amsterdam, p. 371-400.

GRISHAM R. & SUNDHEIM B. (1996). Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, p. 466-471.

GRIVEL L., GUILLEMIN-LANNE S., LAUTIER C. et al. (2001). 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, Paris, 5-6 July 2001, 9 pp.

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

HANDSCHUH S., STAAB S., CIRAVEGNA F. (2002). 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, Springer Verlag, p. 358-372.

KAHAN J., KOIVUNEN M., PRUD'HOMMEAUX E. et al. (2001). Annotea: An Open RDF Infrastructure for Shared Web Annotations, In *Proceedings of the WWW10 International Conference*, Hong Kong, Mai 2001, p. 623-632.

KIRYAKOV A., POPOV B., OGNYANOFF D., MANOV D., KIRILOV A., GORANOV M. (2003). Semantic Annotation, Indexing, and Retrieval, In *Proceedings of the 2nd International Semantic Web Conference* (ISWC2003), Florida, 20-23 Octobre 2003, p. 484-499.

LAUBLET P., REYNAUD C. et CHARLET J. (2002). Sur quelques aspects du Web Sémantique, In *Assises du GDR I3*, Eds Cépadues, Nancy, 20 pp.

MINEL J.-L., J-P. DESCLES, E. CARTIER, G. CRISPINO, S. BEN HAZEZ, A. JACKIEWICZ. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText, Revue Technique et Science informatiques, 20(3), Hermès, p. 369-395.

PARK J. & HUNTING S. (2003). XTM Topic Maps : Creating and using Topic Maps for the Web, Addison Wesley Eds, Boston, p. 81-101.

VARGAS-VERA M., MOTTA E., DOMINGUE J. (2002). MnM : Ontology Driven Tool for Semantic Markup, In *Proceedings of the Workshop Semantic Authoring, Annotation & Knowledge Markup* (SAAKM 2002), Lyon (France), 22-23 Juillet 2002, p. 43-47.