Extraction et exploitation des annotations contextuelles

Noureddine Mokhtari*, Rose Dieng-Kuntz*

*INRIA
2004 route des lucioles - BP 93
FR-06902 Sophia Antipolis cedex
{Noureddine.Mokhtari, Rose.Dieng}@sophia.inria.fr

Résumé
Dans la perspective d'offrir un web sémantique, des travaux ont cherché à automatiser l'extraction des annotations sémantiques à partir de textes pour représenter au mieux la sémantique que vise à transmettre une page web. Dans cet article nous proposons une approche d'extraction des annotations qui représentent le plus précisément possible le contenu d'un document. Nous proposons de prendre en compte la notion de *contexte* modélisé par des relations contextuelles émanant, à la fois, de la structure et de la sémantique du texte.

1 Introduction

L'annotation sémantique est devenue l'une des approches privilégiées par les travaux sur le web sémantique. Les travaux visant à extraire semi-automatiquement ces annotations, plus particulièrement à partir de textes, ont connu ces dernières années une avancée importante. Dans ce contexte, des outils de traitement automatique de la langue naturelle (TALN) sont proposés. Ces outils reposent en général sur des méthodes linguistiques telles que la projection de patrons morpho-syntaxiques ou des méthodes statistiques (fréquence d'apparition). Les méthodes de TALN peuvent être semi-automatiques (l'intervention de l'expert du domaine est alors requise) ou automatiques (dans ce cas, les approches proposées requièrent une certaine spécialisation dans un domaine particulier (Aussenac-Gilles et al., 2006)). Les approches utilisées jusqu'à présent reposent en général sur l'extraction de termes, certaines permettent également l'extraction de relations entre ces termes, mais en ignorant en général le contexte de leur apparition.

Dans le cadre de cette problématique, nous proposons une approche de modélisation, d'extraction et d'exploitation des annotations, qui prenne en compte leurs contextes. La limite observée, concernant les approches d'extraction des termes pour l'annotation, a été notre principale motivation pour offrir des annotations qui représentent au mieux le contenu d'un document. Nous considérons l'annotation sémantique d'un document comme une image par un annotateur (humain ou programme) du contenu de ce document. Cette annotation sémantique doit être exploitable par la machine et de la qualité de cette image dépend son exploitation par l'application visée. Ce travail s'inscrit dans le cadre du projet *SEVENPRO* qui a comme objectif de développer, en reposant sur des technologies et des outils qui aident à la fouille de connaissances sur un produit, des corpus de textes multimédia et sur la réalité virtuelle 3D enrichie sémantiquement.

Tout d'abord, dans la section 2, nous allons analyser quelques travaux sur l'extraction des annotations à partir du texte. Puis dans la section 3, nous aborderons notre proposition sur la modélisation de la notion du contexte. Dans la section 4, nous proposerons notre approche

d'extraction automatique des annotations contextuelles illustrée par des exemples issus du projet SEVENPRO et nous étudierons les possibilités pour inférer/exploiter les annotations contextuelles. La section 5 présentera nos conclusions.

2 Etat de l'art

Nous nous intéressons plus particulièrement aux travaux qui permettent de produire des annotations d'une manière automatique ou semi-automatique et qui concernent la notion de contexte. Cependant, pour une vue plus approfondie sur l'annotation sémantique, citons (Prié Y. et al., 2004) ou encore (Amardeilh F., 2007).

Par rapport à notre problématique d'utilisation de la notion de contexte lors de l'extraction des annotations, nous avons classé les travaux en deux grandes catégories : extraction d'annotation par le contenu du document et extraction d'annotation à partir de sources externes au document.

2.1 Extraction d'annotation par le contenu du document

On distingue deux types de techniques d'annotation de documents par le contenu (ou indexation) : la technique classique, qui consiste en général à attribuer un ensemble de mots clés (ou termes) à chaque document, et la technique sémantique qui attribue une annotation basée sur des concepts (et non de simples mots-clés) et éventuellement sur les relations entre eux. Les travaux visant à extraire des annotations par le contenu se focalisent généralement sur l'extraction des termes. (Guarino et al., 1999) (Khelif et al., 2005) prennent également en compte les relations sémantiques entre termes. Dans (Guarino et al., 1999), les auteurs décrivent OntoSeek un système de recherche documentaire en ligne pour les « pages jaunes ». Afin de construire automatiquement des résumés, (Berri J., 1996) utilise la méthode d'exploration contextuelle (Desclés J. P. et al., 1994) qui repose sur des critères linguistiques et qui consiste à affecter des étiquettes sémantiques aux phrases contenant des indicateurs pertinents. Dans (Desmontils et al., 2002), est proposée une approche supervisée pour indexer des ressources web en reposant sur le contenu des pages web à l'aide d'une ontologie ; les termes sont pondérés par rapport à leur importance dans la page (titre, paragraphe,…). D'autres travaux utilisent les techniques d'extraction d'information pour l'annotation de textes dans un domaine particulier : par exemple, la génomique (Nédellec C., 2004).

2.2 Extraction d'annotation à partir de sources externes au document

Les travaux cités dans cette partie font appel à des ressources externes au document. (Njmogue, et al. 2004) propose une approche basée sur un référentiel métier. L'idée principale est que l'indexation d'un document dépend des activités de l'entreprise et non pas des mots clés du document. Cette approche utilise à la fois une analyse linguistique et statistique du document et un traitement sémantique. Nous pouvons souligner que les activités de l'entreprise peuvent être considérées comme un contexte d'utilisation des documents. Dans (Abrouk, 2006), l'auteur propose une approche pour l'annotation semi-automatique de ressources selon les liens de référencement et ce, sans connaissance préalable du contenu du document. D'autres travaux se sont intéressés à modéliser le processus de recherche d'information et la modélisation de l'utilisateur. Nous citons à titre d'exemple (Hernandez,

2005) qui a comme but principal d'associer le thème relatif au document et la tâche de recherche d'information (l'intention), pour offrir un système de recherche d'information.

3 Modélisation des annotations contextuelles

3.1 Définition du contexte

Nous avons constaté, après des observations empiriques, que l'utilisation d'une sémantique quelconque est étroitement liée à son contexte d'apparition. En effet, l'interprétation ou l'inférence d'une sémantique particulière peut produire des incohérences si nous ignorons les autres sémantiques qui *la précédent, la suivent, l'imbriquent,….* Nous considérons dans cet article une sémantique comme une représentation d'une unité linguistique (partie de texte) par un formalisme de représentation des connaissances, sur lequel nous pouvons inférer.

McCarthy définit le contexte (McCarthy, 1993) comme la généralisation d'une collection d'hypothèses. Les contextes sont ainsi formulés comme des objets formels de première classe. McCarthy postule qu'une proposition p est vraie dans un contexte c, où c est supposé capturer tout ce qui n'est pas explicite dans p mais qui est requis pour faire de p un énoncé significatif pour représenter ce qu'il est supposé établir. Une telle relation de base est elle-même toujours donnée dans un contexte. (Brezillon et al. 2003) commente cette définition en soulignant ses conséquences : (a) un contexte est toujours relatif à un autre contexte, (b) les contextes sont de dimension infinie (c) ils ne peuvent donc pas être décrits complètement. (Brezillon et al. 2003) propose alors la définition suivante : « *le contexte est l'information qui caractérise les interactions entre humains, applications et l'environnement* ».

D'une manière plus simple et étant donné que nous manipulons du texte, *nous proposons de modéliser le contexte d'un objet donné (partie de texte et sa sémantique) par l'ensemble des relations sémantiques contextuelles (RC) —spatiales , temporelles et autres— entre cet objet et les objets qui interagissent avec lui.* A la différence des relations entre concepts, proposées pour représenter une connaissance, les relations contextuelles que nous proposons pour modéliser le contexte sont des relations entre objets.

3.2 Type d'objets manipulés

Nous nous intéressons à l'extraction d'annotations contextuelles à partir de textes : par conséquent, les objets que nous manipulons sont de type textuel. Un « Objet Textuel » (OT) est défini comme un élément du texte (mot, terme, phrase, titre, texte mis entre parenthèses, paragraphe, section, partie de phrase,…) qui transmet une sémantique. Nous définissons aussi un « Objet Sémantique » (OS) comme la sémantique transmise par un objet textuel et dont nous cherchons à déterminer le contexte. D'une manière plus simple, un objet sémantique est la représentation sémantique associée à un objet textuel. Nous avons adopté dans notre approche, la représentation de la sémantique par des concepts et des relations provenant d'une ontologie. L'utilisation de la structure du texte, à travers les objets textuels, nous permet de choisir le niveau de détail à étudier ou « granularité ». En effet, nous pouvons étudier, à titre d'exemple, d'une part les « paragraphes » et les RC entre eux, et d'autre part les RC entre « phrases ». La difficulté réside principalement dans le choix de la bonne granularité à étudier. En outre, la notion de granularité vient conforter notre vision sur la récursivité entre les contextes et ainsi assurer la présence du contexte à tous les niveaux d'abstraction.

Par conséquent, nous pouvons étudier les RC, non seulement entre les objets textuels du même niveau d'abstraction, mais entre les objets appartenant à des niveaux différents.

3.3 La portée de validité d'une sémantique

Définir une *portée de validité* d'un objet sémantique donné revient à chercher un ensemble d'objets sémantiques dans lequel il est utilisable pour raisonner sans produire des incohérences sémantiques. Du point de vue des objets sémantiques, la notion de validité permet de distinguer : les objets sémantiques valides quel que soit le contexte et les objets sémantiques valides dans un (ou plusieurs) contexte(s) particulier(s). Pour le contexte, la notion de validité offre la possibilité d'étudier la portabilité entre contextes, c'est-à-dire la validité dans un autre contexte, des objets sémantiques valides dans un contexte donné.

Jusqu'à présent, nous avons montré notre vision du contexte et sa modélisation avec les annotations contextuelles. Reste à élaborer une approche permettant d'extraire ces annotations contextuelles.

4 Extraction et exploitation des annotations contextuelles

Avant de proposer l'approche d'extraction des annotations contextuelles, définissons les différentes relations contextuelles structurelles/sémantiques que nous cherchons à extraire.

4.1 Relations contextuelles structurelles/sémantiques

Comme souligné précédemment, plusieurs niveaux de granularité existent, pour les annotations contextuelles, selon les objets textuels que nous voulons étudier (phrase, paragraphe,…). Nous allons proposer une approche basée sur un niveau de granularité particulier que nous jugeons riche sémantiquement. Les objets textuels, que nous manipulerons à ce niveau, ont les « liens logiques » comme délimiteurs. Les liens logiques (Asher et al., 2003) (relation rhétorique, connecteurs logiques ou relations de discours,…) sont définis comme les liens qu'entretiennent les idées entre elles dans un texte argumentatif et servent aussi à assurer les articulations et la progression d'un texte.

Des travaux se sont déjà intéressés à identifier ces relations dans un texte, nous citons (Saito et al., 2006) qui décrit un système permettant d'identifier les relations de discours entre deux phrases qui se suivent en langue japonaise. Aussi, (Marcu et al., 2002) propose une approche non supervisée pour identifier les catégories des relations de discours (*CONTRAST*,…). D'autres (Teufel et al., 2000) proposent une approche qui détecte les actions et les affecte à leurs types en exploitant la structure des documents scientifiques (Aim, Own,…). Une autre approche (Desclés J. P., 2006) va plus loin et propose d'annoter automatiquement des documents en utilisant la « sémantique du discours ». Cette méthode est basée sur un moteur de règles qui permet d'identifier les segments de textes qui contiennent *une définition*, *une causalité*,.... (Laublet et al., 2007) proposent une annotation manuelle basée sur une ontologie du discours énonciatif «OntoDiscours» dans le but de déterminer «qui a parlé, à qui, où et quand ?».

Les liens logiques sont classés en linguistique selon le raisonnement qu'ils veulent transmettre : comparaison (*aussi que,*…) ; d'illustration (*ainsi,*…) ; de condition (*si,*…) ;…. Chaque lien logique possède des arguments (Prasad et al., 2006), généralement deux argu-

ments. Par exemple: [*Jack failed the exam*] *because* [*he was lazy*]. Le premier argument représente ici le fait et le second représente ici la cause.

4.1.1 Relations contextuelles spatiales

Nous définissons une *relation contextuelle spatiale* comme toute relation exprimant la position d'un objet (textuel ou sémantique) par rapport aux autres objets de même granularité ou non. Pour les objets textuels, nous pouvons prendre à titre d'exemple les relations de succession, appartenance,…. Dans le cas des objets sémantiques nous aurons par exemple des relations exprimant un rang (*devant, derrière, après*...) ou un lieu (*dans, chez, sous*...).

4.1.2 Relations contextuelles temporelles

Nous définissons une *relation contextuelle temporelle* comme toute relation exprimant la notion de temps entre un objet (textuel ou sémantique) et d'autres objets du même niveau de granularité ou non. Pour les objets sémantiques, nous pouvons avoir à titre d'exemple comme liens logiques (*avant, depuis, pendant*,...). Dans le cas des objets textuels, nous pouvons considérer le temps des verbes (*présent, futur*,…) comme une relation temporelle qui exprime le moment où la sémantique de ces objets sera manifestée.

4.1.3 Relations contextuelles diverses

Nous appelons *relations contextuelles diverses* des relations exprimant une notion sémantique quelconque (ni spatiale ni temporelle) d'un objet (textuel ou sémantique) par rapport aux autres objets du même niveau de granularité ou non.

Dans le cas d'un objet textuel, un exemple de ce type de relation contextuelle peut être : le degré d'importance entre un paragraphe et son titre. Pour les objets sémantiques, toutes les relations de lien logique qui n'expriment pas l'espace ou le temps peuvent être prises en compte : une addition (*de plus, d'ailleurs*,…), une illustration (*ainsi, comme*,…).

4.2 Approche d'extraction des annotations contextuelles

Nous proposons une approche d'extraction d'annotations contextuelles répartie en deux grandes parties : manipulation textuelle et manipulation sémantique (voir *FIG. 1)*.

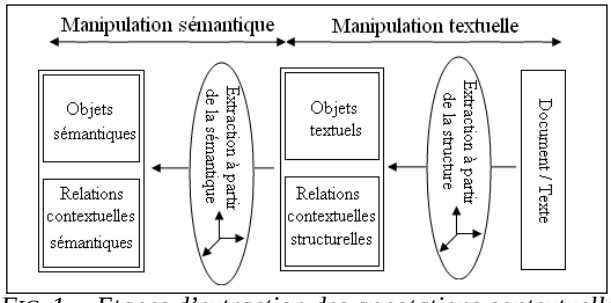

FIG. 1 – Etapes d'extraction des annotations contextuelles

4.2.1 Manipulation textuelle

Cette partie a pour objectif d'obtenir un ensemble d'OT et les RC structurelles entre eux.
Identification des objets textuels. Identifier ces objets revient à identifier les titres, phrases, lien logique, mots,… ainsi que les arguments de chaque lien logique présent dans le texte. Nous avons utilisé dans cette étape la bibliothèque de développement d'ingénierie linguistique GATE (Cunningham al., 2002) qui repose sur l'application successive (chaîne de traitement) de transducteurs[1] aux textes. Suivant une liste de liens logiques, des règles JAPE[2] sont gérées automatiquement pour obtenir la position dans le texte des *liens logiques.* D'autres règles sont construites manuellement pour identifier les marqueurs *des indices numériques qui précèdent certaines phrases.* Les règles JAPE sont appliquées en tant que transducteur dans la chaîne de traitement. D'autres transducteurs fournis par défaut dans GATE nous ont permis d'identifier les phrases et les paragraphes dans le texte. Nous avons exploité, à ce niveau, les marqueurs de position tels que le début et la fin dans le texte (d'une phrase, d'un lien logique,…) pour identifier les arguments des liens logiques dans le texte.
Identification des relations contextuelles structurelles. Afin d'identifier les RC structurelles citées dans (4.1.1), nous avons (a) identifié les titres à l'aide des marqueurs numériques, qui précèdent certaines phrases tels que « *les numérotations de type '6.2.1'*», ainsi que des heuristiques telles que « *une seule phrase existante dans le paragraphe qui contient le titre (cette phrase représente le titre lui même)* ». Nous soulignons que le transducteur fourni par défaut dans GATE pour identifier les *titres,* donne une faible précision, ce qui nous a menés à introduire ces heuristiques pour les identifier automatiquement; (b) calculé la portée des titres et construit les imbrications entre eux (nous aurons ainsi défini quel paragraphe et quel sous-titre appartiennent à quel titre); (c) construit les imbrications entre les paragraphes, les phrases et les arguments en utilisant les marqueurs de position dans le texte. Une fois la structure hiérarchique du texte construite, les RC structurelles peuvent être déduites.

4.2.2 Manipulation sémantique

Nous nous intéressons dans cette partie aux OS et aux RC sémantiques qui les relient.
Identification des objets sémantiques. Identifier ces objets signifie représenter la sémantique des OT, par un formalisme de représentation des connaissances. Nous avons opté pour une représentation avec le standard RDF(S) reposant sur la notion des triplets (ressource, propriété, valeur). Dans notre approche, nous supposons qu'une ontologie est déjà construite. Pour associer les OT à des triplets RDF en se référant à l'ontologie, nous proposons d'identifier les ressources (ou concepts), les propriétés et les valeurs (instances) dans le texte. Pour cela, nous proposons de construire des règles JAPE d'une manière automatique. En effet, l'idée principale est de bénéficier de la propriété *rdfs:label*, dans un schéma RDFS, pour construire des règles JAPE qui détectent les différentes manifestations d'un *concept* (ou *propriété*) dans le texte. La propriété *rdfs:label* représente le nom lisible par un humain d'un *concept* (ou *propriété*) dans le texte. Par ailleurs, les règles JAPE qui détectent les instances sont construites en se référant à un document RDF qui contient la liste d'instances pour cha-

[1] Un transducteur est un automate à états finis qui, pour chaque état parcouru, produit une ou plusieurs informations.
[2] JAPE (Java Annotation Patterns Engine) est un langage d'expression de grammaires pour le TALN (un exemple est donné dans 4.2.3).

que concept. Souvent, la détection des instances est difficile puisque nous ne connaissons pas auparavant les instances et à quel concept elles réfèrent. Le document RDF qui contient les listes des instances est une particularité propre aux données que nous manipulons dans notre expérimentation. Par la suite, les règles JAPE sont introduites dans la chaîne de TALN pour produire des marqueurs, qui permettent de connaître la position des concepts, propriétés et instances dans le texte. Enfin, il reste la construction des triplets RDF associés à chaque OT. Des problèmes d'ambiguïté peuvent surgir dans cette étape. Ils sont dus à l'identification de plusieurs concepts, propriétés ou instances dans un même OT. Nous proposons d'ajouter la prise en compte des contraintes *range* et *domain* pour déterminer quelle propriété correspond à quel concept. Néanmoins, nous avons remarqué que l'ambiguïté est généralement inexistante si la fréquence des liens logiques est grande dans le texte. En effet, plus la fréquence des liens logiques est grande et plus la taille[3] des arguments est relativement petite. Par conséquent, le même argument correspond moins souvent à plusieurs concepts, propriétés ou instances. Quant à l'objet textuel de type « titre » sa taille est généralement petite.

Identification des relations contextuelles sémantiques. Cette étape consiste à attribuer le rôle sémantique aux liens logiques déjà détectés. Des travaux (Sporleder et al., 2005) (Marcu et al., 2002) identifient ces rôles d'une manière automatique : par exemple, une addition pour les liens logiques (*de plus, d'ailleurs,…*). Néanmoins, des problèmes d'ambiguïté persistent dans certains liens logiques plus complexes. Nous nous sommes contentés, à ce stade de notre travail, des liens logiques qui ne représentent pas des ambiguïtés sémantiques tels que les liens logiques «*because, compared to, except,…*».

3 Description of the mill internal elements

3.1 Inlet Headliners

This new design is composed of 3 thicker bolted rings, compared to the original design of 2 rings. The liners have a thickness of 70mm. except for the area of most wear (R/2-R/2+R/3), where the thickness is 85mm.

FIG. 2 – Exemple de partie de texte dans un document

```
<title Id="OT1">
 Description of the mill internal elements
 <title Id="OT12"> Inlet Headliners
  <paragraph>
   <sentence>
    <contextualRelation Id="OT121" type="comparedTo">
      <argument1 Id="OT1211"> This new design is composed
          of 3 thicker bolted rings,
      </argument1>
      <argument2 Id="OT1212">the original design of 2 rings
      </argument2>
    </contextualRelation>
   </sentence>
   <sentence>
    <contextualRelation Id="OT122" type="except">
     <argument1 Id="OT1221">
          The liners have a thickness of70mm.
     </argument1>
     <argument2 Id="OT1222">
          for the area of most wear (R/2-R/2+R/3),
      <contextualRelation Id="OT12221" type="where">
       <argument Id="OT122211"> the thickness is 85mm.
       </argument>
      </contextualRelation>
     </argument2>
    </contextualRelation>
   </sentence>
  </paragraph>
 </title>
</title>
```

FIG. 3 – Arborescence des objets textuels

[3] La taille est la longueur de la chaîne de caractères.

4.2.3 Déroulement de l'approche proposée

Nous soulignons que la langue que nous prenons en compte actuellement est l'*anglais*. L'approche que nous proposons est expérimentée sur un document texte (3768 mots et 1862 autres unités linguistiques tel que les : chiffres, virgules, parenthèses,…) issu des partenaires industriels dans le cadre du projet SEVENPRO.. Nous exposons dans ce qui suit le déroulement des étapes de l'approche sur une partie de ce document *FIG. 2*.

```
<rdfs:Class rdf:ID="RingOfDiaphragmMill">
 <rdfs:subClassOf rdf:resource="#Item"/>
 <rdfs:label xml:lang="en">Ring of mill diaphragm
   </rdfs:label>
 <rdfs:label xml:lang="en">Ring of diaphragm
   </rdfs:label>
 <rdfs:label xml:lang="en">Ring</rdfs:label>
 <rdfs:label xml:lang="fr">Anneau diaphragme
  </rdfs:label>
 <rdfs:comment xml:lang="en">denotes a part
   of a diaphragm mill (i.e. ring).
 </rdfs:comment>
</rdfs:Class>
<rdf:Property rdf:ID="hasPart">
 <rdf:type rdf:resource="http://www.w3.org/
    2002/07/owl#TransitiveProperty"/>
 <owl:inverseOf rdf:resource
   ="http://www.sevenpro.org/
   ontologies/2006/estanda#partOf"/>
 <rdfs:domain rdf:resource="#Item"/>
 <rdfs:range rdf:resource="#Item"/>
 <rdfs:label xml:lang="en">has part</rdfs:label>
 <rdfs:label xml:lang="en">is composed of
   </rdfs:label>
 <rdfs:comment xml:lang="en">
   hasPart is transitive and also reflexive,
   and anti-symmetrical.
 </rdfs:comment>
</rdf:Property>
```

```
phase: first
options: control = appelt

Rule:JRuleRingOfDiaphragmMill (
({Token.lemma
=="Ring"}({SpaceToken})?{Token.lemma
=="of"}({SpaceToken})?{Token.lemma
=="mill"}({SpaceToken})?
{Token.lemma=="diaphragm"}({SpaceToken})?)|
({Token.lemma=="Ring"}({SpaceToken})?
{Token.lemma=="of"}
({SpaceToken})?{Token.lemma=="diaphragm"}
({SpaceToken})? )|
({Token.lemma =="Ring"}({SpaceToken})? )
):RingOfDiaphragmMill  -->: RingOfDia-
phragmMill.Concept = {kind
="RingOfDiaphragmMill",
rule=JRuleRingOfDiaphragmMill }

Rule: JRulehasPart (
({Token.lemma=="has"}({SpaceToken})?
{Token.lemma =="part"})|
({Token.lemma=="is"}({SpaceToken})?
 {Token.lemma=="composed"}
({SpaceToken})?{Token.lemma=="of"}
({SpaceToken})?)|
):hasPart -->:hasPart.property = {kind
="hasPart", rule=JRulehasPart}
```

FIG. 4 – Représentation rdfs d'un Concept et d'une relation dans l'ontologie et les règles JAPE permettant de les identifier dans le texte

En appliquant l'étape « manipulation textuelle » et ses différentes identifications et constructions en utilisant les marqueurs extraits à partir des règles JAPE, nous obtenons un ensemble d'OT, selon une structure arborescente sous format XML (*FIG. 2*).

Les exemples de RC structurelles pouvant être extraites à ce niveau sont : OT1 *imbrique* OT12; OT122 *succède* OT121; OT12 *plus Important que*[4] OT121; …
Les étapes de la manipulation textuelle sont implémentées sous forme de requêtes XQuery[5].

Les objets textuels étant identifiés, il faut leur associer des objets sémantiques gérés à l'étape de manipulation sémantique. Selon l'ontologie associée[6], un ensemble de règles JAPE est construit automatiquement. La *FIG. 4* montre une description RDFS du concept «*RingOfDiaphragmMill*» et la relation «*hasPart*» dans l'ontologie ainsi que les règles JAPE associées permettant de les identifier dans le texte. Nous avons obtenu 65 règles gérées automatiquement et qui correspondent à 65 concepts dans l'ontologie. Cependant, comme l'ontologie est amenée à évoluer, il faudra relancer l'opération de génération pour avoir des règles qui reflètent l'état de l'ontologie. L'ontologie étant en cours de construction, le nom-

[4] Un titre est plus important qu'un paragraphe.
[5] Langage de requête pour les documents XML: *http://www.w3.org/TR/xquery/*
[6] L'ontologie utilisée dans le cadre du projet SEVENPRO concerne la description et la composition de produits industriels (ingénierie des moulins).

bre de propriétés/instances construit n'est pas conséquent. De ce fait, les règles associées aux propriétés et instances ne sont pas générées dans l'expérimentation.

L'utilisation de ces règles dans la chaîne de TALN permet de marquer tous les concepts et propriétés. Nous utilisons « Token.*lemma* » pour faire référence à toute variante d'un mot donné. Les concepts, propriétés et instances identifiés dans l'exemple de la *FIG. 2* sont affichés par un fond gris. Le résultat de la construction des triplets de la *FIG. 5* montre les objets sémantiques correspondant aux objets textuels *OT1211* et *OT1212* de la *FIG. 3* Aussi, les RC entre objets sémantiques ayant été identifiées, sont soulignées dans la *FIG. 2*.

```
<rdf:Description rdf:about="#NewDesign">
 <hasPart>
  <rdf:Description rdf:about="#RingOfDiaphragmMill">
  <quantity>3</quantity>
   </rdf:Description>
 </hasPart>
</rdf:Description>
```

```
<rdf:Description rdf:about="#OriginalDesign">
 <hasPart>
  <rdf:Description rdf:about="#RingOfDiphragmMill">
  <quantity>2</quantity>
   </rdf:Description>
 </hasPart>
</rdf:Description>
```

FIG. 5 – Objets sémantiques représentés en RDF

Bonne ou Mauvaise	Titres	Paragraphes/phrases	RC	Arguments	Concepts
B. identification	26	204/283	72	99	406
M. identification	0	0/101	3	6	25
Total présent	26	204/313	95	142	452
Précision (%)	100	-	96	94,28	94,19
Rappel (%)	100	-	75,78	69,71	89,82

TAB. 1 – Evaluation des étapes d'extraction

Les résultats d'évaluation partielle des étapes de l'extraction (*TAB. 1*) sont très satisfaisants. Cependant la construction des triplets RDF (*FIG. 5*) n'a pas été expérimentée encore vu que l'ontologie est en cours de construction (pas assez de propriétés et instances).

A la différence des simples relations prises en compte par le formalisme RDF reliant des concepts, les RC que nous proposons sont des relations entre OS. Par conséquent, cela se traduit en relations entre triplets RDF (c'est-à-dire relations entre des annotations sémantiques), or RDF ne permet pas la représentation de telles relations. Afin de pallier à ce problème de représentation, nous attribuons des identifiants aux objets sémantiques. Ces derniers sont mis dans un document RDF à part. Nous remplaçons les OT dans le fichier XML, qui représente la structure du document, par une référence (avec les identifiants) vers les OS.

4.3 Exploitation des annotations contextuelles

Dans cette partie, nous allons donner quelques exemples d'utilisation d'annotations contextuelles ainsi que des orientations de nos travaux futurs sur leur exploitation. L'exploitation des annotations contextuelles repose essentiellement sur la notion de portée de validité. La portée de validité d'un objet sémantique donné est calculée suivant un contexte donné modélisé par les RC structurelles et sémantiques. Prenant l'exemple de la *FIG. 3*, la portée de validité de l'objet sémantique associé à l'objet textuel «*OT1*» est l'ensemble d'objets sémantiques associés aux objets textuels «*OT12, OT121, OT122, OT1211, OT1212,*

OT1221, OT1222, OT12221, OT122211». Les conséquences de cette portée sont que : nous pouvons déduire, par exemple, que le concept «*newDesign*» cité dans l'objet «*OT1211*» correspond à la nouvelle conception du concept «*mill*» citée dans l'objet «*OT1*». Par ailleurs, la notion de portée de validité nous permettra d'étudier la portabilité entre contextes, ce qui conduit à une complexité supplémentaire dans l'inférence.

Le challenge qui reste à surmonter réside dans la difficulté du raisonnement sur les RC spatiales et temporelles. Nous proposons de nous inspirer des travaux sur les SIG (systèmes d'informations géographiques) pour raisonner sur l'aspect spatial et temporel. En effet, nous pouvons utiliser le raisonnement dit « *de propagation de contraintes* », et plus particulièrement en utilisant les relations (ou intervalles) d'Allen (Allen, 1984). Les relations d'Allen (*FIG. 6*) peuvent être utilisées pour représenter les liens logiques exprimant des RC temporelles. Par exemple les liens logiques *avant*, et *durant* peuvent être représentés respectivement par les relations (A *before* B ou A *meets* B), et (A *during* B). Ainsi nous pouvons faciliter l'inférence sur l'aspect temporel. Nous pouvons exploiter les ontologies déjà construites et qui modélisent l'aspect temporel (Santos et al., 2003). Cependant, nous proposons de simplifier ces propositions et de les adapter à la manipulation du texte. Les limites de RDF concernant la représentation des relations entre triplets, ouvrent des perspectives de recherche sur une extension de ce formalisme et ceci nous pousse à utiliser un formalisme de description plus puissant (OWL), notamment pour représenter le rôle sémantique des RC. La *FIG.* 7 donne un exemple d'une représentation OWL de la relation temporelle *before*.

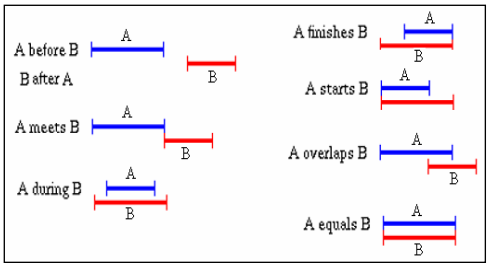

FIG. 6 – Les relations d'Allen

```
<rdf:Property rdf:ID="Before">
 <rdfs:subPropertyOf rdf:resource="#TemporalProperty"/>
  <rdf:type rdf:resource="&owl;TransitiveProperty"/>
  <owl:inverseOf rdf:resource="&TempSchema;After"/>
  <rdfs:domain rdf:resource="#SemanticObject"/>
  <rdfs:range rdf:resource="#SemanticObject"/>
</rdf:Property>
```

FIG. 7 – La relation temporelle «Before»

De la même manière, nous proposons d'exploiter les relations d'Egenhofer (Egenhofer et al., 1991) pour représenter les RC spatiales (4.1.1). Les annotations contextuelles que nous proposons ouvrent des perspectives de raisonnement sur le texte qui auparavant n'étaient pas envisageables.

5 Conclusions

Nous avons proposé, dans cet article, une approche qui modélise le contexte à partir de sources textuelles, en prenant en compte les différents types de relations structurelles/sémantiques (spatiale, temporelle, et diverse). Dans l'approche d'extraction des annotations contextuelles que nous proposons, les étapes automatisées sont : la détection des liens logiques, des titres et leurs portées, des imbrications entre « titre, paragraphe, phrase et argument », des concepts. Les étapes qui restent à automatiser sont : l'identification des propriétés et des instances ainsi que la construction des triplets. Un prototype est implémenté pour évaluer partiellement les différentes étapes de l'extraction. Les résultats de l'évaluation

sont très satisfaisants. Cependant, l'ontologie à laquelle nous nous référons est incomplète et a besoin d'être enrichie. Aussi l'attribution des rôles pour les RC reste à généraliser sur des liens logiques plus complexes. Par conséquent, nous envisageons d'introduire l'inférence spatiale et temporelle pour mieux exploiter la richesse sémantique considérable des liens logiques. Afin de valoriser ce travail, un outil qui regroupe toutes les étapes d'extraction est en cours d'élaboration. Aussi une ontologie regroupant les relations contextuelles (temporelles, spatiales et autres) est en cours de construction, pour faciliter leur réutilisation. L'approche d'extraction et d'exploitation des annotations contextuelles sémantiques offre des perspectives prometteuses dans des domaines d'extraction de connaissances à partir du texte. Néanmoins, i) l'utilisation des relations contextuelles pour déduire les dépendances d'inférence (portée de validité) pourra peser lourd sur le temps d'exécution de l'inférence ; ii) aussi, nous avons constaté des redondances d'annotations contextuelles dans un même contenu textuel. Pour ces deux problèmes, nous envisageons de proposer des solutions techniques pour optimiser l'inférence et réduire la taille des annotations.

Références

Abrouk L., (2006), Annotation *de documents par le contexte de citation basée sur une ontologie*, Thèse de doctorat en informatique, Montpellier.

Allen J.F., (1984), *Towards a general theory of action and time*, Artificial Intelligence.

Amardeilh F., (2007), *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*, Thèse de doctorat en informatique, PARIS X.

Asher N. et Lascarides A., *Logics of Conversation*, Cambridge University Press, 2003.

Aussenac-Gilles N. et J. Marie-Paule, (2006), *Designing and Evaluating Patterns for Ontology Enrichment from Texts*, EKAW'2006, LNAI 4248, pp. 158 – 165, Tchèque.

Berri J., (1996), *Mise en œuvre de la méthode d'exploration contextuelle pour le résumé automatique de textes : Implémentation du système SERAPHIN*. CLIM'96, Canada.

Brézillon P., (2003), *Représentation de pratiques dans le formalisme des graphes contextuels.* In: J.M.C. Bastien (Ed.), EPIQUE'2003, pp. 3-14, Rocquencourt Inria, France.

Cunningham H., Maynard D., Bontcheva K. et Tablan V., (2002), *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, ACL, pp. 235-238, Budapest, Hungary.

Desclés J. P., (2006), *Contextual Exploration Processing for Discourse Automatic Annotation for Texts*, FLAIRS2006, AAAI Press, Florida, pp. 281-284.

Desclés J.-P. et Minel J.-L.,(1994), *L'exploration contextuelle*, *In Le résumé par exploration contextuelle*, rapport interne du CAMS n°95/1, Nancy, pp. 3-17.

Desmontils E. et Jacquin C., (2002), *Indexing a web site with a terminology oriented ontology*, The Emerging Semantic Web, IOS Press, 181-197.

Egenhofer M. et Franzosa R., (1991), *Point-Set Topological Spatial Relations*, International Journal of Geographical Information Systems, vol. 5, no. 2, pp. 161-174.

Guarino N., Masolo C. et Vetere G., (1999), *Ontoseek: Content-based access to the web*, IEEE Intelligent Systems, May-June 4-5:70–80.

Hernandez N., (2005), *Ontologies de domaine pour la modélisation du contexte en recherche d'information*, Thèse de doctorat en informatique, Toulouse.

Khelif K., Dieng-Kuntz R., Barbry P., (2005), *Semantic web technologies for interpreting DNA microarray analyses: the MEAT system*. Proc. of WISE'05, 20-22/11, New York,.

Jackiewicz A. et Laublet P., (2007), *Web sémantique et linguistique du discours OntoDiscours : un tournant énonciatif*, Atelier Ontologie et Texte, TIA2007, Sophia-Antipolis.

Marcu D. et Echihabi A., (2002), *An Unsupervised Approach to Recognizing Discourse Relations*, Proc. of the 40th Annual Meeting of the ACL, pp. 368-375, Philadelphia.

McCarthy J., (1993), *Notes on formalizing context*, Proc. 13th IJCAI, pp.555–560, California.

Nédellec C., (2004), *Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives*, In Text Mining and its Applications: Results of the NEMIS Launch Conference Series: Studies in Fuzziness and Soft Computing, Sirmakessis, Spiros (Ed.), Springer Verlag.

Njmogue W., Fontaine D. et Fontaine P., (2004), *Identification des thèmes d'un document relativement à un référentiel métier*, In Proceedings of MAJECSTIC'04, Calais, France.

Prasad R., Dinesh N., Lee A., Joshi A. and Webber B., (2006), *Attribution and its Annotation in the Penn Discourse TreeBank*, Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 31–38, Sydney.

Prié Y. et Garlatti S., (2004), *Annotations et métadonnées dans le Web sémantique*, in Revue I3 Information-Interaction - Intelligence, Numéro Hors-série Web sémantique, 24 pp.

Saito M., Yamamoto K. et Sekine S., (2006), *Using Phrasal Patterns to Identify Discourse Relations*, Proc. of the HLTCNA Chapter of the ACL, pp. 133–136, New York.

Santos J. et Staab S., (2003), *FONTE: Factorizing Ontology Engineering Complexity*, International Conference On Knowledge Capture, K-CAP'2003, ACM Press, pp. 146-153, Sanibel Island, FL, USA.

Sporleder C. et Lascarides A., (2005), *Exploiting linguistic cues to classify rhetorical relations*, In Proceedings of RANLP-05, pp. 532-539 Bulgaria.

Teufel S. et Moens M., (2000), *What's yours and what's mine: Determining Intellectual Attribution in Scientific Text*. SIGDAT (EMNLP/VLC-2000), 9 – 17, Hong Kong.

Summary

In prospect to offer a semantic Web, some works strived to extract automatically semantic annotations from texts, thus aiming at better representing the semantics a web page conveys. In this paper we propose an approach for extracting annotations that represent the content of a document more precisely. We consider the *context* that we model by contextual relations built up from both the structure and the semantics of the text.