Extraction automatique de relations sémantiques définies dans une ontologie

Albert Royer, Christian Sallaberry, Annig Le Parc-Lacayrelle, Marie-Noëlle Bessagnet

LIUPPA Laboratoire LIUPPA, France
albert.royer@univ-pau.fr
christian.sallaberry@univ-pau.fr
annig-lacayrelle@univ-pau.fr
marie-noelle.bessagnet@univ-pau.fr

Résumé : Cette contribution se situe dans le domaine de la recherche d'information sémantique et s'intéresse plus particulièrement à la phase d'annotation. Nous proposons une méthode qui permet d'annoter automatiquement des documents textuels sur la base de concepts et de relations sémantiques modélisés dans une ontologie. Cette méthode est générique car elle est capable d'exploiter le contenu sémantique de toute ontologie. Elle a été mise en œuvre sur la plateforme GATE [1].

Abstract : This contribution is in the field of semantic information retrieval from textual documents and more particularly the annotation stage. We propose an automatic annotation method based on concepts and semantic relationships modeled in an ontology. This method is generic as it is able to exploit the semantic content of any ontology. It has been implemented on the GATE platform.

Mots-clés : recherche d'information sémantique, annotation automatique de concept, annotation automatique de relation sémantique, ontologie

1 Introduction

Pour faire face à l'augmentation exponentielle de données désormais disponibles dans des formats numériques, les modèles et technologies supportant les moteurs de recherche ont intégré de nombreuses améliorations (Buscaldi *et al.*, 2013). Toutefois, ces propositions restent souvent limitées par l'usage de mots-clés, qui contraste avec l'idée de recherche « sémantique » où les descripteurs d'une collection de documents ou d'un besoin d'information sont des entités sémantiques (représentant le sens d'un mot ou d'un syntagme). Ce paradigme de recherche, appelé recherche d'information sémantique (RIS), exploite généralement des ressources telles que des thésaurus ou des ontologies dans les phases d'indexation et de recherche.

La difficulté de production de telles ressources sémantiques est réelle. Elle combine souvent des techniques automatiques et l'intervention d'experts de domaines spécifiques. Dans ce travail, nous considérons que ces ressources existent et nos propositions se focalisent sur leur exploitation plutôt que leur construction. Il n'en demeure pas moins que les processus d'annotation automatique de concepts et de relations, en vue de l'extraction de descripteurs sémantiques, sont dépendants de la qualité de ces ressources. Ainsi, les résultats obtenus pour des approches de RIS sont souvent mitigés lorsqu'on les compare à la recherche d'information (RI) classique. Ils sont cependant encourageants pour des domaines spécifiques qui peuvent être décrits plus

1. https://gate.ac.uk/

facilement (Kiryakov *et al.*, 2004). Nous pouvons citer des expériences positives spécifiques au domaine médical (Abasolo & Gomez, 2000; Trieschnigg *et al.*, 2009), par exemple. RIS et RI sont également combinées dans certaines approches (Buscaldi & Zargayouna, 2013; Kara *et al.*, 2012).

Notre proposition se situe dans le contexte de la RIS. Nous avons conçu et mis en œuvre le prototype de RIS ThemaStream (Buscaldi *et al.*, 2013) qui exploite une ontologie de domaine dédiée aux plantes. Comme pour d'autres prototypes (Kara *et al.*, 2012), le processus d'annotation des relations sémantiques s'appuie sur des algorithmes dépendants du domaine et ne traite pas les problèmes d'ambiguïté. Notre contribution est une nouvelle démarche automatique d'annotation de relations sémantiques dans des textes, indépendante du domaine applicatif. Nous proposons un algorithme de recherche de triplets « domaine/relation/codomaine » qui prend la relation comme point de départ et non pas les concepts, comme dans la majorité des approches.

L'originalité de cette proposition est sa généricité. En effet, l'algorithme d'annotation de relations sémantiques est indépendant du domaine. Il exploite les concepts et les relations sémantiques de toute ontologie.

Dans cet article, nous exposons notre processus de reconnaissance, dans des documents textuels, de relations sémantiques préalablement définies dans une ontologie. Après cette partie introductive, une deuxième partie présente succinctement des travaux de recherche en lien avec notre contribution. Une troisième partie décrit le modèle de concept et de relation sur lequel reposent nos propositions. Elle présente deux algorithmes dédiés au marquage de relations sémantiques dans des textes. Une quatrième partie illustre ces propositions à travers des exemples d'ontologie, de texte à analyser et de recherche d'information. Nous terminons par une conclusion et des perspectives.

2 Travaux connexes

L'exploitation de la sémantique est au centre des travaux de chercheurs de différents domaines : la représentation et la gestion des connaissances, le web sémantique (WS) ou web de données et la RI. Quel que soit le domaine, un ensemble de connaissances peut être modélisé sous la forme d'une ontologie exploitée localement ou partagée via le WS (Linked Data ou Linked Open Data).

La RI est la tâche de recherche de documents, au sein d'une collection, satisfaisant un besoin d'information (Manning *et al.*, 2008). Les premières propositions relatives à la RIS, au-delà des modèles de recherche basés sur les seuls mots-clés, visent l'exploitation de vocabulaires (WordNet [2], par exemple) qui associent du sens à des termes (Kara *et al.*, 2012). Ces vocabulaires permettent par exemple l'expansion d'index et de requêtes dans des processus de RI. De manière générale, la RIS exploite des descripteurs sémantiques de documents pour répondre à un besoin d'information.

La première difficulté est donc l'annotation sémantique en amont de l'étape de RIS proprement dite. L'annotation de textes fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite (Kiryakov *et al.*, 2004). Elle consiste à associer des descripteurs relatifs au contenu, à la structure, ou encore au contexte des documents textuels.

2. `http://wordnet.princeton.edu/`

(Kara *et al.*, 2012; Fernandez *et al.*, 2011) distinguent les approches d'annotation sémantique automatique basées sur l'exploitation (1) de la langue, (2) de modèles statistiques et (3) d'ontologies.

La première catégorie d'approches est basée sur le traitement automatique de la langue et exploite des patrons linguistiques définis «manuellement» par des experts. Elle nécessite des ressources importantes en termes de capacité de traitement mais aboutit généralement à des taux de rappel et de précision satisfaisants.

La seconde catégorie est basée sur des techniques d'apprentissage. Les approches de cette catégorie peuvent être supervisées. Dans ce cas, les règles d'annotation sont déduites automatiquement à partir de l'analyse d'un échantillon de documents manuellement associés à des classes (catégories) par des experts ; citons SVM (Support Vector Machine) et K-NN (k-Nearest Neighbors). Elles peuvent être aussi non supervisées et associées à des ressources externes, auquel cas les classes sont définies à partir de ces ressources et les règles d'annotation déduites automatiquement de l'analyse d'un échantillon de documents. On parle d'approches ressource-dépendantes : par exemple, la méthode ESA (Explicit Semantic Analysis) est associée à des bases de connaissances telles que Wikipedia. Enfin, ces méthodes peuvent être non supervisées et ne requérir aucune ressource externe. On parle d'approches corpus-dépendantes : les classes sont déduites de l'analyse d'un échantillon de documents. Citons les méthodes LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis), pLSA (probabilistic LSA) et LSI (Latent Semantic Indexing). Moins exigeantes en termes de capacité de traitement, elles aboutissent généralement à des taux de rappel et de précision moins importants que ceux de la première catégorie.

La troisième catégorie d'approches, quant à elle, s'appuie sur l'exploitation de bases de connaissances ontologiques (Ontologie Based Information Extraction - OBIE). L'annotation est guidée par les connaissances modélisées dans l'ontologie : classes, concepts, relations (Wimalasuriya & Dou, 2010; Nebhi, 2012). Au-delà de la simple extraction d'information, il s'agit d'enrichir l'annotation d'un document par des liens vers des connaissances supplémentaires décrites dans une ontologie locale (Wang & Stewart, 2015) ou partagée via le WS (Nebhi, 2012). L'architecture générale d'un système OBIE est présentée dans (Wang & Stewart, 2015). L'ontologie y est considérée comme une ressource qui peut-être soit fournie au système en entrée, soit construite et mise à jour par le système.

Nos travaux se situent dans cette troisième catégorie. Comme la plupart des systèmes OBIE (SOBA, KIM, PANKOW cités dans (Wang & Stewart, 2015)) nous avons adopté l'approche basée sur une ontologie existante fournie en entrée.

La table 1 liste de nombreux prototypes de RIS qui s'appuient sur l'exploitation de bases de connaissances ontologiques. Ces prototypes utilisent des ontologies ou le WS pour annoter des concepts (C) uniquement ou des concepts et des relations (C et R).

La plupart de ces systèmes sont liés à un domaine, comme celui proposé par (Kara *et al.*, 2012) ou encore dans nos précédents travaux relatifs à ThemaStream (Buscaldi *et al.*, 2013). Notre contribution propose une nouvelle démarche automatique d'annotation de relations sémantiques dans des textes. Cette approche est générique et ouvre ainsi la possibilité d'exploiter les concepts et les relations sémantiques définies dans toute ontologie. Ce travail va dans le sens des préconisations de (Lee *et al.*, 2014) qui mettent en exergue l'importance de l'exploitation des relations sémantiques entre concepts dans le processus de RIS.

Prototype de RIS	Domaine	Annotation
TextViz (Dudognon *et al.*, 2010)	Ontologie de domaine (Mécanique)	C
(Fernandez *et al.*, 2011)	WS Ontologie de domaine (Football)	C C
(Kara *et al.*, 2012)	Ontologie de domaine (Football)	C et R
ThemaStream (Buscaldi *et al.*, 2013)	Ontologie de domaine (Plantes)	C et R
YaSemIR (Buscaldi & Zargayouna, 2013)	Ontologie (domaine indifférent)	C
Broccoli (Bast *et al.*, 2014)	WS (Plantes sur Wikipedia et FreeBase)	C et R
(Lee *et al.*, 2014)	Ontologie de domaine (Bibliographie)	C et R
(Berlanga *et al.*, 2015)	WS (Bioinformatique sur UMLS et WikiNet)	C
Mimir (Tablan *et al.*, 2015)	WS (Innondations sur DBpedia et Geonames) Ontologie de domaine (médical)	C et R C et R
(Wang & Stewart, 2015)	Ontologie de domaine (Catastrophes naturelles)	C et R

TABLE 1 – Prototypes de RIS

3 Annotation de concepts et de relations sémantiques

Cette partie présente notre démarche pour l'annotation de concepts et de relations sémantiques dans un texte. La méthode présentée a pour origine les travaux menés dans le cadre des projets ANR DYNAMO (Dudognon *et al.*, 2010) et MOANO (Bessagnet *et al.*, 2013; Buscaldi *et al.*, 2013). Nous étendons ces travaux de manière à prendre en compte l'annotation automatique des relations quelle que soit l'ontologie.

Nous définissons une **ontologie** O par l'ensemble C des concepts et par l'ensemble R des relations entre concepts : $O = (C, R)$. On note $R = \{r_\nu\}$ avec $r_\nu = (\nu, \delta, \rho)$ où la relation r_ν de nom ν a pour domaine de classe δ et pour co-domaine de classe ρ. Dans l'ontologie, à chaque concept est associée une liste de termes qui *dénotent* le concept ; on note T l'ensemble des termes pouvant dénoter un concept (c'est-à-dire, des termes dont la présence dans un texte implique automatiquement la présence du concept dénoté). Rappelons que la rédaction de la liste associée à chaque concept de l'ontologie est du ressort d'un spécialiste du domaine. Par exemple, dans une ontologie pour le domaine Topo-carto de l'IGN, on trouve les concepts *commune*, *bâtiment remarquable* et *château*. En plus des relations hiérarchiques classiques (*is-a*) permettant de modéliser que le concept *château* est un sous-concept de *bâtiment remarquable*, il est possible de modéliser des relations sémantiques comme la relation *embellir* entre les concepts *château* et *commune*.

Le processus d'annotation que nous proposons s'appuie sur quatre prédicats.

Pour les **concepts** :

$subsumes(c_i, c_j)$ où $c_i \in C$ et $c_j \in C$, indique que c_i est un ascendant de c_j ou bien $c_i = c_j$;
$has\_label\_c(c, t)$ où $c \in C$ et $t \in T$, indique que c a pour label le terme t.

Pour les **relations sémantiques** :

$has\_label\_r(r, t)$ où $r \in R$ et $t \in T$, indique que r a pour label le terme t ;
$relation(t, c_\delta, c_\rho)$ où $t \in T, c_\delta \in C$ et $c_\rho \in C$, indique la relation révélée par t entre c_δ et c_ρ.

Le **corpus** est un ensemble D de documents. Chaque document d est composé de plusieurs champs f_i. L'ensemble des n champs d'un document $d \in D$ est noté $F_d = \{f_0, \ldots, f_n\}$. L'unité de traitement du texte est le champ : une partie de phrase, une phrase, un paragraphe, voire un

document (par défaut, la portée du champ correspond à la phrase). Plus le champ est large plus le risque d'ambiguïté augmente.

Ainsi, un champ f contient un concept c si et seulement si un terme t dénotant un concept c' existe et si le concept c' est un descendant de c ou si $c = c'$ (c'est à dire, $has\_label\_c(c', t)$ et $subsumes(c, c')$). De plus, un champ f contient une relation r si trois termes du champ dénotent, l'un la relation et les deux autres les concepts du domaine c_δ et du co-domaine c_ρ correspondants à cette relation.

L'annotation automatique de concepts et de relations sémantiques dans un champ f se déroule en deux temps : une première phase procède à l'identification des concepts et des relations potentielles puis, une seconde phase valide ou non ces relations (voir figure 1).

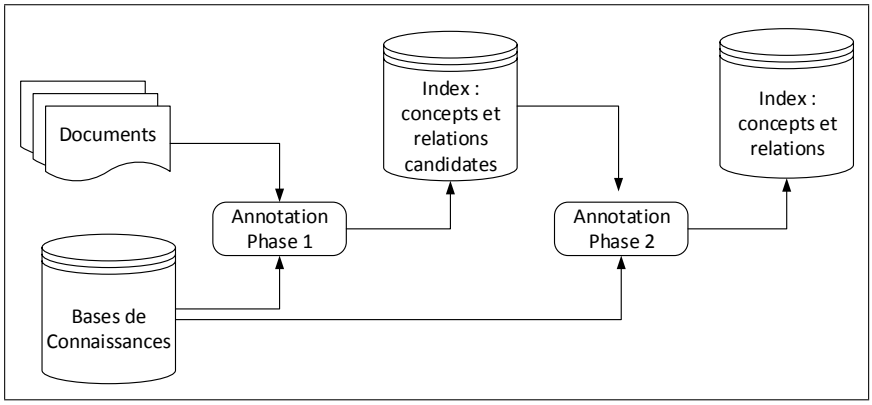

FIGURE 1 – Principe général d'annotation

3.1 Phase 1 : Identification de concepts et de relations

La première phase construit, pour le champ f du document analysé et une ontologie O, les trois ensembles suivants :

- T_f l'ensemble des termes présents dans le champ f du document analysé,
- $AC_f = \{c_0, \ldots, c_m\}$ l'annotation d'un champ f avec les concepts $c_0, \ldots, c_m$ de l'ontologie O,
- $ARP_f = \{r_0, \ldots, r_n\}$ l'annotation d'un champ f avec les relations $r_0, \ldots, r_n$ de l'ontologie O, qui sont potentiellement pertinentes pour le champ f.

Le processus de la phase 1 est décrit par l'algorithme 1 qui, pour chaque terme d'un champ, recherche les éventuels concepts à partir des labels correspondants dans l'ontologie. De la même manière, l'itération suivante de l'algorithme recherche les relations candidates en comparant chaque terme aux labels des relations décrites dans l'ontologie. Il y a ambiguïté lorsqu'un même terme dénote plusieurs relations. L'analyse du domaine et du codomaine d'une relation, dans une seconde phase, permet la différenciation et la validation des relations candidates.

Au terme de cette phase, chaque terme annoté correspond à un ou plusieurs concepts, une ou plusieurs relations, ou encore, un ou plusieurs concepts et relations.

Données :
f : champ étudié,
O : ontologie
Résultat :
T_f : ensemble de termes annotés,
AC_f : ensemble de concepts trouvés,
ARP_f : ensemble de relations sémantiques candidates
début
$T_f \longleftarrow \{\}$
$AC_f \longleftarrow \{\}$
$ARP_f \longleftarrow \{\}$
pour chaque *terme* $t \in f$ **faire**
pour chaque *concept* $c \in C$ **faire**
si $has\_label\_c(c, t)$ **alors**
si $t \notin T_f$ **alors**
$T_f \longleftarrow T_f \cup \{t\}$
$AC_f \longleftarrow AC_f \cup \{c\}$
pour chaque *relation* $r = (\nu, c_d, c_{cd}) \in R$ **faire**
si $has\_label\_r(r, t)$ **alors**
si $t \notin T_f$ **alors**
$T_f \longleftarrow T_f \cup \{t\}$
$ARP_f \longleftarrow ARP_f \cup \{(\nu, c_d, c_{cd})\}$

Algorithme 1 : Phase 1 d'identification de concepts et de relations

3.2 Phase 2 : Validation des relations candidates

La seconde phase a pour objectif de construire l'ensemble de relations sémantiques AR_f à partir du processus de validation appliqué aux relations potentielles contenues dans ARP_f :

– $AR_f = \{r_0, \ldots, r_p\}$ où $r_0, \ldots, r_p$ appartiennent à ARP_f et sont validées pour le champ f.

Ce processus est détaillé par l'algorithme 2. Pour chaque relation détectée lors de la phase 1, il s'agit désormais de valider et d'annoter les triplets « domaine/relation/codomaine ». Ainsi, pour chaque relation candidate, on consulte l'ontologie pour récupérer le concept de *domaine* et le concept de *codomaine* correspondants. Ensuite, on recherche dans le champ, chacun de ces deux concepts parmi les concepts annotés ou leur ascendance.

L'algorithme 2 d'annotation de triplets, qui s'appuie sur la relation comme point de départ, lève la majorité des ambigüités : pour le cas des relations r et r' de mêmes *domaine* et *codomaine*, il validera les deux relations si les vocabulaires associés à r et r' ne sont pas disjoints.

Données :
AC_f : ensemble de concepts trouvés,
ARP_f : ensemble de relations sémantiques candidates
O : ontologie
Résultat :
AR_f : ensemble de relations sémantiques validées
début
  $AR_f \longleftarrow \{\}$
  pour chaque *relation* $r = (\nu, c_d, c_{cd}) \in ARP_f$ **faire**
    si $(\exists c_i \in AC_f, subsumes(c_d, c_i)) \wedge (\exists c_j \in AC_f, subsumes(c_{cd}, c_j))$ **alors**
      $AR_f \longleftarrow AR_f \cup \{(\nu, c_d, c_{cd})\}$

Algorithme 2 : Phase 2 de validation des relations candidates

4 Expérimentation

Nous avons mis en œuvre ces propositions dans une chaîne de traitements dont nous présentons les caractéristiques et des exemples d'expérimentation ci-après.

4.1 La chaîne de traitement

La chaîne de traitement, illustrée sur la figure 2, a été mise en œuvre sur la plateforme GATE (Cunningham *et al.*, 1995; Bontcheva *et al.*, 2004). Elle s'applique à des collections de documents textuels vus comme une suite de phrases. Le premier module intitulé « Traitement de la langue » intègre notamment l'analyseur morphosyntaxique Treetagger (Schmid, 1994) et prend en charge la lemmatisation en langue française afin de tenir compte de variations syntaxiques. Le deuxième module « Annotation d'entités nommées », qui n'est pas instancié systématiquement, permet de détecter des entités nommées décrites dans des bases de connaissances. Le troisième module « Annotation de concepts et de relations » correspond à la mise en œuvre de l'algorithme n° 1 (décrit précédemment). Il s'agit de l'annotation des termes correspondants à des labels définis dans l'ontologie. Chaque annotation comporte les détails suivants : le terme

original, le lemme correspondant, le label identifié, le nom du concept ou de la relation, le type de l'objet reconnu (instance, classe ou relation).

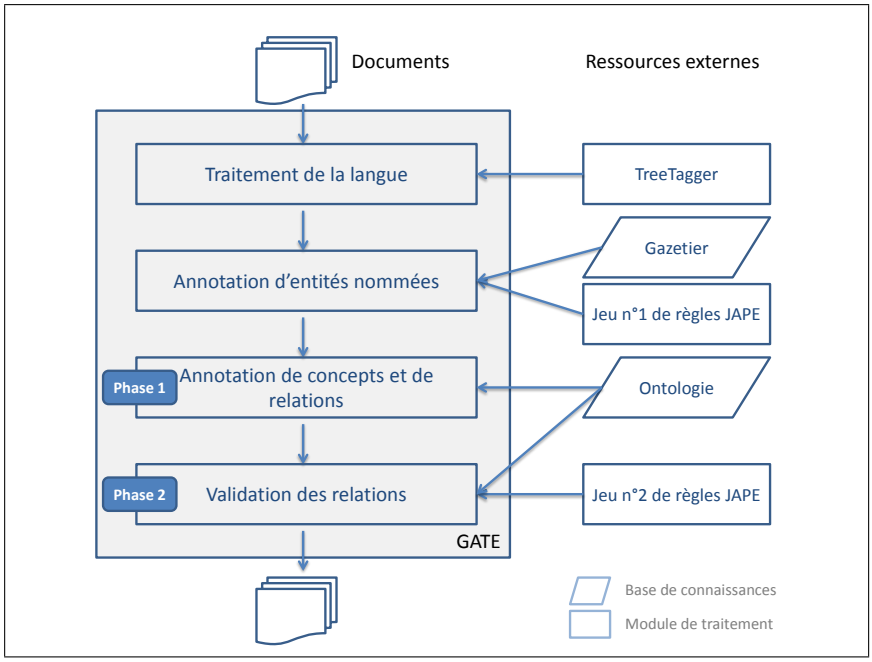

FIGURE 2 – Architecture générale du processus d'annotation sémantique sur la plateforme GATE

Le quatrième module « Validation des relations », quant à lui, correspond à la mise en œuvre de l'algorithme n° 2. Des règles sont définies dans le langage JAPE à partir d'expressions régulières combinant concepts de domaine, de codomaine et relations. Ce module prend, en entrée, les marquages de la phase précédente et s'appuie sur l'ontologie pour valider l'annotation des relations potentielles : à chacune correspond un triplet « domaine/relation/codomaine » défini dans l'ontologie qu'il s'agit de valider à partir des éléments annotés dans le texte. Par exemple, sur la figure 3, la règle CRC0 a pour expression régulière : $C1\ P\ C2$ où $C1$ et $C2$ sont des concepts et P est une relation. $Lookup$ est le résultat d'annotation de la phase précédente comportant notamment un délimiteur de champ. Chaque occurrence de ce triplet $C1\ P\ C2$ est traitée par des instructions Java pour la validation de la relation potentielle. Ainsi, c'est par la présence de concepts relatifs au domaine et au codomaine que la validation permet de lever d'éventuelles ambiguïtés liées aux relations (un même terme pouvant être label de plusieurs relations).

4.2 Trois cas d'étude

Nous avons expérimenté cette chaine de traitement générique avec trois bases de connaissances différentes : les ontologies GEONTO, MOANO et ONTOTHAU, respectivement.

```
Phase: PhaseRelationDetectCRC

Input: Lookup
Options: control =    all

Rule: CRC0
// règle classe relation classe
(
 ({Lookup.type==class}):C1
 ({Lookup.type==property}):P
 ({Lookup.type==class}):C2
 ({Lookup.type==fieldDelimiter})
)
:tripletDPR
-->
:tripletDPR

{
//instructions java calculant et générant les annotations
}
```

FIGURE 3 – Règle Jape pour la validation de relations sémantiques

4.2.1 Ontologie GEONTO

Le projet ANR GEONTO (Mustière *et al.*, 2011) a permis de construire une ontologie de concepts topographiques. Cette ontologie a été conçue par enrichissement d'une première taxonomie de termes, et ce grâce à l'analyse de deux catégories de documents textuels : des spécifications techniques de bases de données de l'IGN et des récits de voyage. Comme décrit dans (Kergosien *et al.*, 2009), l'ontologie GEONTO est une hiérarchie de concepts géographiques et de labels associés. Nous avons enrichi GEONTO avec des relations sémantiques à des fins expérimentales.

4.2.2 Ontologie MOANO

Le projet ANR MOANO (Aussenac-Gilles *et al.*, 2013) a permis de construite une ontologie à partir d'une collection de documents web structurés portant sur le domaine des plantes, non d'un point de vue botanique ou scientifique mais plutôt du point de vue du jardinage. Cette ontologie a été construite de manière automatique par raffinement successifs et contient des relations entre concepts.

4.2.3 Ontologie ONTOTHAU

Le projet CNRS MASTODONS ANIMITEX (Roche *et al.*, 2014) a permis de construite une ontologie, dédiée au bassin de Thau et à l'aménagement du territoire, à partir d'un vocabulaire défini par des experts géographes. Comme pour GEONTO, cette ontologie a été enrichie avec des relations sémantiques à des fins expérimentales.

4.3 Focus sur le cas d'étude GEONTO

Nous illustrons ici la mise en œuvre de notre approche sur le cas d'étude GEONTO.

4.3.1 Exemple de relations sémantiques

La table 2 illustre les exemples des relations Crue, Équipement, Patrimoine et Proximité dont les labels sont des verbes à l'infinitif et les domaines et codomaines sont des concepts.

Relation	Labels	Domaine	Codomaine
Crue	inonder, envahir, recouvrir, emporter...	Cours d'eau	Entité à vocation résidentielle
Équipement	situer, exister, être disponible...	Équipement de loisir	Commune
Patrimoine	situer, exister, embellir, enrichir, mettre en valeur, appartient, ériger, bâtir, construire, comporter, compter...	Élément du patrimoine	Commune
Proximité	croiser, longer, traverser, passer sous, passer sur, surplomber...	Route	Route

TABLE 2 – Exemple de relations sémantiques

Notons ici que les relations *Équipement* et *Patrimoine* pourront être vecteur d'ambiguïté puisqu'elles sont décrites par des ensembles de labels non disjoints.

4.3.2 Exemple d'annotations

La figure 4 illustre un exemple de marquage de la relation sémantique *Patrimoine*, définie dans l'ontologie GEONTO, sur un texte tiré de Wikipedia [3].

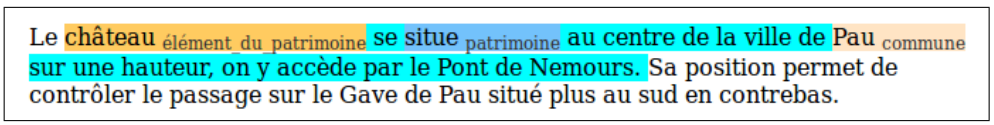

FIGURE 4 – Exemple d'annotation de la relation Patrimoine

L'analyse de ce texte permet d'instancier « château », « situer », « Pau », « Pont de Nemours », « passage », « Gave de Pau »... Or, le verbe « situer » dénote deux relations *Équipement* et *Patrimoine* définies dans l'ontologie. L'ambiguïté est levée par la présence de concepts relatifs au domaine (« château » instancie le concept *Élément du patrimoine*) et au codomaine (« Pau » instancie le concept *Commune*) de la relation *Patrimoine*.

4.3.3 Exemple d'exploitation en RI

Ce marquage nous permet ensuite d'envisager différentes stratégies de RI. Imaginons que nous recherchions tous les documents évoquant des châteaux de la commune de Pau. À cette fin, nous utilisons l'environnement GATE pour mettre en œuvre deux stratégies.

3. http ://fr.wikipedia.org/wiki/Château_de_Pau

L'exemple de la figure 5 illustre une RI basée sur des concepts. Dans ce cas, nous cherchons les concepts « Élément du patrimoine » et « commune ». Deux documents mentionnant « châteaux » et « Pau » sont retournés via la requête. Toutefois, celui mentionnant « le château de Franqueville, visible aisément depuis le boulevard des Pyrénées à Pau » n'est pas pertinent, bien qu'il fasse référence aux concepts « Élément du patrimoine » (château) et « Commune » (Pau).

L'exemple de la figure 6, quant à lui, illustre une RI basée sur des concepts et des relations. Ici, nous affinons la recherche en ciblant la relation Patrimoine. Dans ce cas, un seul document est retourné et il bien pertinent.

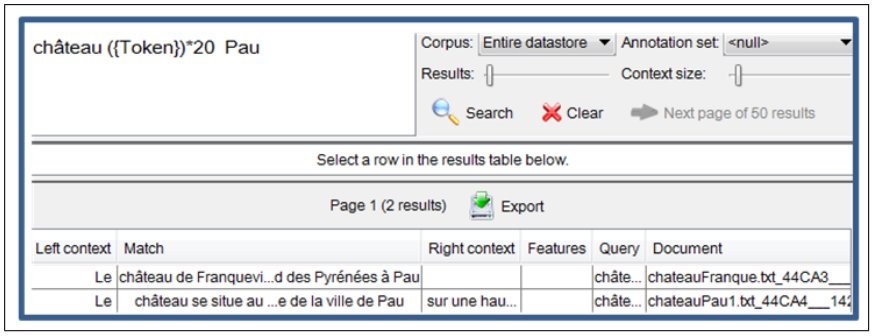

FIGURE 5 – Exemple de RI ciblant des concepts « Élément du patrimoine » et « Commune »

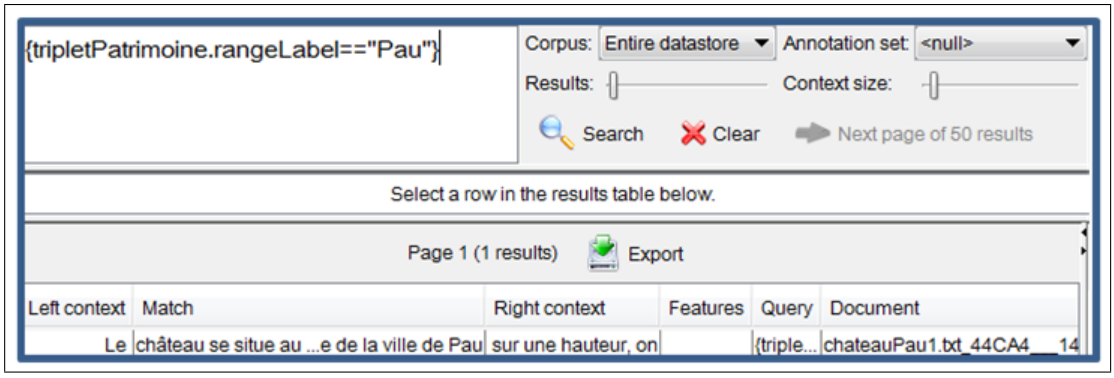

FIGURE 6 – Exemple de RI ciblant des relations « Patrimoine »

5 Conclusion et perspectives

Nous avons décrit une méthode qui vise la reconnaissance automatique, dans des textes, de relations sémantiques (décrites dans une ontologie) entre entités thématiques. Elle s'appuie sur un algorithme de recherche de triplets « domaine/relation/codomaine » qui prend pour point de départ la relation et non les concepts (comme dans la majorité des approches). Notre proposition est générique dans le sens où elle est indépendante du domaine.

Cette proposition est une première étape d'un processus plus large visant la recherche d'information sémantique (RIS). Comme montré précédemment, le processus de RI que nous concevons s'appuie sur les concepts et les relations annotées préalablement. Les travaux de (Maynard & Greenwood, 2012; Buscaldi & Zargayouna, 2013) confirment l'intérêt de telles approches : ils proposent d'indexer les concepts de collections de textes à partir d'ontologies de domaine puis combinent la RI classique de type sac de mots et la RIS de type graphe de concepts. L'approche de RIS que nous proposons s'appuie sur les concepts présents dans les corpus mais aussi sur les relations entre ces concepts (Buscaldi *et al.*, 2013; Bessagnet *et al.*, 2013).

La prochaine étape consistera à évaluer cette approche sur différents corpus de textes et à comparer ces résultats à ceux de nos propositions précédentes ainsi qu'à ceux d'autres systèmes de RIS. Il s'agira de trouver ou de définir un cadre d'évaluation ainsi que des systèmes de RIS ouverts.

Références

ABASOLO J. M. & GOMEZ M. (2000). Melisa. an ontology-based agent for information retrieval in medicine. In *In : Proceedings of the First International Workshop on the Semantic Web (SemWeb2000*, p. 73–82.

AUSSENAC-GILLES N., BUSCALDI D., COMPAROT C. & KAMEL M. (2013). Enrichissement d'ontologies grâce à l'annotation sémantique de pages web. In C. VRAIN, A. PÉNINOU & F. SÈDES, Eds., *Extraction et gestion des connaissances (EGC'2013), Actes, 29 janvier - 01 février 2013, Toulouse, France*, volume RNTI-E-24 of *Revue des Nouvelles Technologies de l'Information*, p. 229–234 : Hermann-Éditions.

BAST H., BÄURLE F., BUCHHOLD B. & HAUSSMANN E. (2014). Semantic full-text search with broccoli. In S. GEVA, A. TROTMAN, P. BRUZA, C. L. A. CLARKE & K. JÄRVELIN, Eds., *The 37th International ACM SIGIR Conference on Research and development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, p. 1265–1266 : ACM.

BERLANGA R., NEBOT V. & PÉREZ M. (2015). Tailored semantic annotation for semantic search. *Web Semantics : Science, Services and Agents on the World Wide Web*, **30**(0), 69 – 81. Semantic Search.

BESSAGNET M.-N., BUSCALDI D., ROYER A. & SALLABERRY C. (2013). Une approche basée sur des relations pour la RI sémantique. In *Atelier Recherche d'Information Sémantique RISE, associé à la conférence IC , édité par Catherine Roussey*, p. 19–33.

BONTCHEVA K., TABLAN V., MAYNARD D. & CUNNINGHAM H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, **10**(3/4), 349–373.

BUSCALDI D., BESSAGNET M.-N., ROYER A. & SALLABERRY C. (2013). Using the semantics of texts for information retrieval : A concept- and domain relation-based approach. In B. CATANIA, T. CERQUITELLI, S. CHIUSANO, G. GUERRINI, M. KÄMPF, A. KEMPER, B. NOVIKOV, T. PALPANAS, J. POKORNÝ & A. VAKALI, Eds., *ADBIS (2)*, volume 241 of *Advances in Intelligent Systems and Computing*, p. 257–266 : Springer.

BUSCALDI D. & ZARGAYOUNA H. (2013). Yasemir : Yet another semantic information retrieval system. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '13, p. 13–16, New York, NY, USA : ACM.

CUNNINGHAM H., GAIZAUSKAS R. & WILKS Y. (1995). *A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering R&D*. Rapport interne CS – 95 – 21, Department of Computer Science, University of Sheffield. `http://xxx.lanl.gov/-abs/cs.CL/9601009`.

DUDOGNON D., HUBERT G., MARCO J., MOTHE J., RALALASON B., THOMAS J., REYMONET A., MAUREL H., MBARKI M., LAUBLET P. & ROUX V. (2010). Dynamic ontology for information retrieval. In *RIAO*, p. 213–215 : CID - Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

FERNANDEZ M., CANTADOR I., LOPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : An ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, **9**(4), 434 – 452. {JWS} special issue on Semantic Search.

KARA S., ALAN O., SABUNCU O., AKPINAR S., CICEKLI N. K. & ALPASLAN F. N. (2012). An ontology-based retrieval system using semantic indexing. *Inf. Syst.*, **37**(4), 294–305.

KERGOSIEN E., KAMEL M., SALLABERRY C., BESSAGNET M.-N., AUSSENAC-GILLES N. & GAIO M. (2009). Construction et enrichissement automatique d'ontologie à partir de ressources externes. In *Journées Francophones sur les Ontologies (JFO'2009)*.

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

LEE J., MIN J.-K., OH A. & CHUNG C.-W. (2014). Effective ranking and search techniques for web resources considering semantic relationships. *Information Processing and Management*, **50**(1), 132 – 155.

MANNING C. D., RAGHAVAN P. & SCHÜTZE H. (2008). *Introduction to information retrieval*. Cambridge University Press.

MAYNARD D. & GREENWOOD M. A. (2012). Large scale semantic annotation, indexing and search at the national archives. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Eds., *LREC*, p. 3487–3494 : European Language Resources Association (ELRA).

MUSTIÈRE S., ABADIE N., AUSSENAC-GILLES N., BESSAGNET M.-N., KAMEL M., KERGOSIEN E., REYNAUD C., SAFAR B. & SALLABERRY C. (2011). Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique. *Revue Internationale de Géomatique*, **21**(2), 155–179.

NEBHI K. (2012). Ontology-based information extraction for french newspaper articles. In B. GLIMM & A. KRÜGER, Eds., *KI 2012 : Advances in Artificial Intelligence - 35th Annual German Conference on AI, Saarbrücken, Germany, September 24-27, 2012. Proceedings*, volume 7526 of *Lecture Notes in Computer Science*, p. 237–240 : Springer.

ROCHE M., TEISSEIRE M., CRÉMILLEUX B., GANCARSKI P. & SALLABERRY C. (2014). ANIMITEX. analyse d'images fondée sur des informations textuelles. *Ingénierie des Systèmes d'Information*, **19**(3), 163–167.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, p. 44–49.

TABLAN V., BONTCHEVA K., ROBERTS I. & CUNNINGHAM H. (2015). Mimir : An open-source semantic search framework for interactive information seeking and discovery. *Web Semantics : Science, Services and Agents on the World Wide Web*, **30**(0), 52 – 68. Semantic Search.

TRIESCHNIGG R., PEZIK P., LEE V., DE JONG F., KRAAIJ W. & REBHOLZ-SCHUHMANN D. (2009). Mesh up : effective mesh text classification for improved document retrieval. *Bioinformatics*, **25**(11), 1412–1418.

WANG W. & STEWART K. (2015). Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Computers, Environment and Urban Systems*, **50**(0), 30 – 40.

WIMALASURIYA D. C. & DOU D. (2010). Ontology-based information extraction : An introduction and a survey of current approaches. *J. Information Science*, **36**(3), 306–323.