Vers un outil de co-construction d'ontologies à partir de textes à l'aide d'un système multi-agent adaptatif

Zied Sellami
IRIT (Institut de Recherche en Informatique de Toulouse)
Université Paul Sabatier, 118 Route de Narbonne F-31062, Toulouse cedex 9
France
sellami@irit.fr

Nathalie Aussenac-Gilles
IRIT (Institut de Recherche en Informatique de Toulouse)
Université Paul Sabatier, 118 Route de Narbonne F-31062, Toulouse cedex 9
France
aussenac@irit.fr

Marie-Pierre Gleizes
IRIT (Institut de Recherche en Informatique de Toulouse)
Université Paul Sabatier, 118 Route de Narbonne F-31062, Toulouse cedex 9
France
gleizes@irit.fr

ABSTRACT

Manual ontology engineering and maintenance is a difficult task that requires significant effort from the ontologist to identify and structure domain knowledge. Automatic ontol-ogy learning makes this task easier, especially through the use of text and natural language processing tools. In this paper, we present DYNAMO, a tool based on an Adaptive Multi-Agent System (AMAS), which aims at assisting on-tologists during ontology design and evolution. This work is carried out in the context of the ANR DYNAMO (Dy-namic Ontology for Information Retrieval) project. DY-NAMO is based on text extracted terms, lexical relations and provides an AMAS based module to support ontol-ogy co-construction. The ontologist interacts with the tool by modifying the ontology (move, add, change concepts, terms and/or relationships). Then Dynamo adapts to these changes and proposes new evolutions to improve the ontol-ogy. After describing the context of this work and the prin-ciples of DYNAMO, we report an experiment where DY-NAMO is used to enrich an ontology from a corpus describ-ing software bug reports. We end by analyzing these results and identifying key issues to improve the tool.

Keywords

Ontology Engineering from Text, Adaptive Multi-Agent System

1. INTRODUCTION

La construction et la maintenance manuelle d'une ontologie est une tâche difficile qui nécessite la mise en place de procédés élaborés afin de rendre exploitable la connaissance d'un domaine, manipulable par les systèmes informatiques et dans certains cas interprétable par les êtres humains [10]. Ces procédés peuvent être très longs et demandent une constante vérification de l'ontologie au regard de l'évolution du domaine qu'on désire modéliser.

La construction et la maintenance d'ontologies à partir de textes est apparue depuis quelques années comme une solution pour faciliter cette tâche [14]. Notre recherche se situe dans cet axe en proposant un outil facilitant la construction et la maintenance d'ontologies à partir de textes par le biais d'une co-construction de l'ontologie. Dans ce sens, le système propose des solutions à l'ontographe (nouveaux concepts, nouvelles relations, déplacements de concepts, nouveaux termes, etc.) et apprend à partir des réponses qu'il reçoit. Une des particularités de ce système est le modèle de donnée de l'ontologie qui accorde une place importance aux termes dans l'ontologie (ce ne sont pas de simple label de concept). Dans ce modèle, les termes extraits des corpus et les concepts sont reliés par des liens de dénotations. Ils servent par la suite à annotés sémantiquement les documents. Ainsi, le lien entre ontologies, textes et annotations sémantiques des textes est plus clarifié.

Dans cet article, nous présentons DYNAMO, un outil basé sur un Système Multi-Agent Adaptatif qui permet la co-construction d'ontologies à partir de textes. Nous présentons dans la section 2 le contexte de ce travail, à savoir le projet DYNAMO et le modèle d'ontologie utilisé. Dans la section 3, nous détaillons les principes de fonctionnement du système DYNAMO. Dans la section 4, nous décrivons une première expérimentation de construction d'ontologie avec ce système. Les résultats obtenus sont ensuite analysés dans la section 5. Dans la section 6, nous analysons un état de l'art des outils de construction d'ontologies à partir de textes. Nous concluons cet article par les perspectives d'amélioration de DYNAMO.

2. CO-CONSTRUCTION D'ONTOLOGIE À PARTIR DE TEXTES

2.1 Contexte

Ce travail est réalisé dans le cadre du projet ANR [1] DYNAMO [2] (DYNAMic Ontology for information retrieval). Son objectif principal est de concevoir une approche méthodologique et un ensemble d'outils logiciels qui gèrent la construction et la maintenance de ressources ontologiques à partir de documents et l'utilisation de ces ressources pour une indexation sémantique facilitant la recherche d'information. L'originalité de DYNAMO réside dans la spécification conjointe du fonctionnement des modules de maintenance d'ontologie et de recherche d'information, de manière à prendre en compte, d'une part, les répercutions d'une évolution du corpus sur les ressources ontologiques et, d'autre part, la dynamique de l'annotation (éventuellement sa remise en cause) en fonction des évolutions constatées dans l'ontologie. Dans ce projet, les corpus documentaires fournis par les partenaires sont issus de trois domaines particuliers : la recherche en archéologie des techniques, le diagnostic de pannes automobiles et le diagnostic de défauts logiciels.

2.2 Le modèle d'ontologie

Dans DYNAMO, nous nous intéressons à la construction de Ressources Termino-Ontologique (RTO). Une RTO est une ressource comportant une composante conceptuelle, l'ontologie, et une composante lexicale, la terminologie. La RTO contient alors non seulement une représentation des concepts du domaine, mais aussi une représentation séparée des termes associés (termes désignant les concepts) qui permettent d'annoter ou indexer des documents dans le cadre d'une annotation sémantique.
Les RTO dans DYNAMO sont représentées selon le méta-modèle OWL proposé dans [16]. Dans ce méta-modèle, les concepts et les termes sont deux *meta-class* adaptées de *owl:class*. La relation *denote* entre la classe *term* et la classe *concept* permet de relier un ou plusieurs concepts avec un ou plusieurs termes. Aussi, la manifestation linguistique de chaque concept dans les documents est ainsi sauvegardée avec l'ontologie.

3. PRINCIPES DU SYSTÈME DYNAMO

Le système DYNAMO est organisé en différents modules présentés dans la figure 1. L'architecture globale de l'outil comporte un module de traitement des corpus textuels *DYNAMO Corpus Analyser* et un module *DYNAMO MAS* basé sur un Système Multi-Agents adaptatifs (SMA).

Le module de traitement des corpus textuels prend en charge la préparation des entrées du SMA. Celui-ci est formé par un *Extracteur de termes* qui remplit la *Base de candidats termes*. Le mécanisme *d'extraction de relations lexicales* projette sur le corpus des patrons lexico-syntaxiques issus de la *Base de patrons* et les candidats termes issus de la *Base de candidats termes*. Il fournit en sortie des triplets de la forme <Ti, Rel, Tj> où Ti et Tj sont des candidats termes i et j, Rel un type de relation. Chaque triplet a un indice de confiance formé par le couple (Q, I). Q est la qualité maximal des patrons lexico-syntaxiques permettant d'extraire la relation Rel. I est la somme des instances des patrons lexico-syntaxiques de Rel et ayant Q comme qualité. Ces triplets sont les entrées du module SMA.

DYNAMO MAS est formé de deux types d'agents : un *TermAgent* reflète la composante terminologique de l'ontologie et un *ConceptAgent* représente la composante conceptuelle de l'ontologie. Chaque *TermAgent* gère les relations lexicales dont il est source ou cible. De même, chaque *ConceptAgent* gère les relations conceptuelles dont il est source ou cible.

Le processus de co-construction de l'ontologie se traduit alors par des propositions faites par le SMA envers l'ontographe qui peut les accepter ou les refuser. Ces propositions de modification peuvent être l'ajout, la suppression et/ou le déplacement de concepts, de termes et/ou de relations.
En sortie, DYNAMO fournit une ontologie sous la forme d'un fichier OWL respectant le modèle RTO présenté en section 2.2. A ce stade d'avancement de l'étude de DYNAMO, nous ne prenons pas en compte des contraintes de restrictions et/ou de disjonctions de concepts et de relations, etc.

3.1 Choix de l'utilisation d'une approche par patrons lexico-syntaxiques

La construction d'ontologies à partir de textes s'appuie sur des logiciels de traitement du langage naturel et sur des ressources combinant lexique et concepts. L'extraction de concepts fait appel à des extracteurs de termes.

Deux approches différentes existent pour déceler les relations entre concepts. La première approche, dite statistique, décèle des relations entre concepts (co-occurrences de termes, etc.) sans toutefois interpréter ces relations [8]. La deuxième approche est basée sur la définition de patrons lexico-syntaxiques qui établissent une relation entre concepts du domaine [9]. Ces relations ne sont décelées que lorsque les concepts appartiennent à la même phrase. Les patrons lexico-syntaxiques se fondent sur un marqueur, ou pivot (une unité linguistique qui peut être un indice d'une relation lexicale, comme *entre autres* pour la relation d'hyperonymie et un ensemble de contraintes que le contexte lexical ou syntaxique de ce pivot doit remplir. Par exemple, dans le cas de l'hyperonymie et du marqueur *entre autres*, il faut que la forme syntaxique corresponde au patron *DET SN, entre autres SN*. Ce patron permet d'extraire une phrase contenant *Les méningites, entre autres pathologies...*, et de mettre en relation *méningites* et *pathologies*. Cette approche a été présentée dans [9] et mise en oeuvre dans [12] [3] [2].

Les résultats présentés par [11] montrent que les approches statistiques tendent à devenir peu efficaces lorsque le volume du corpus et la redondance de son contenu sont faibles. C'est le cas des données dans le projet DYNAMO. En effet, nous sommes en présence de corpus de faible taille (Corpus ACTIA : 46000 mots, corpus ARTAL : 86000 mots et corpus Arkeotek : 106000 mots), contenant des documents très courts et axés sur un domaine de connaissances très précis. Pour toutes ces raisons, nous adoptons une approche par patrons lexico-syntaxiques pour extraire des relations lexicales entre termes et établir des relations conceptuelles entre les concepts.
Compte tenu de l'avancement du projet, la définition de pa-

[1] http://www.agence-nationale-recherche.fr , Agence Nationale de la Recherche
[2] http://www.irit.fr/DYNAMO ; partenaires : Préhistoire et Technologie (http://www.mae.u-paris10.fr), Actia (http://www.actia.com), Artal (http://www.artal.fr) et l'IRIT(http://www.irit.fr/

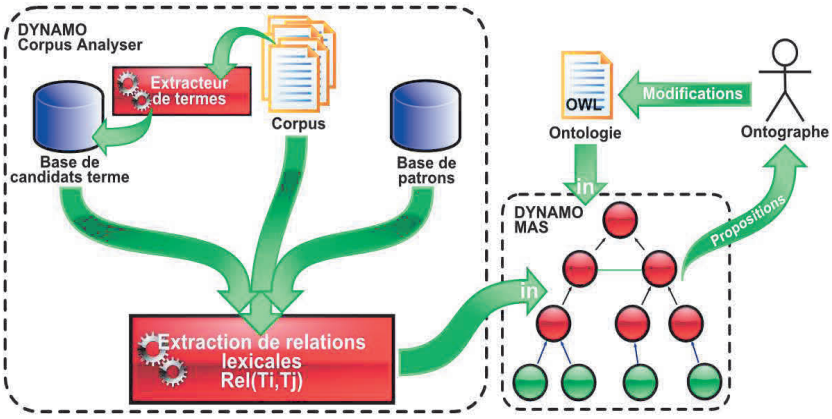

Figure 1: Relation entre les modules *DYNAMO Corpus Analyser*, *DYNAMO MAS* et l'ontographe.

trons lexico-syntaxiques adaptés à chaque corpus n'est pas encore réalisée. Afin de tester le fonctionnement du système et la faisabilité de notre approche nous avons utilisé les mots pivots définis dans TerminoWeb [3] pour détecter les relations d'hypéronymie, de synonymie, de méronymie et transverses. Ces mots pivots sont détaillés dans la section 5.

3.2 Les relations conceptuelles

Dans l'outil DYNAMO, nous nous intéressons à quatre relations lexicales par lesquelles sont déduites les relations conceptuelles entre les concepts de l'ontologie à concevoir.

1. L'hypéronymie exprime une relation de généricité-spécificité entre termes. Cette relation permet de définir ensuite des relations *is_a* entre deux concepts.

2. La méronymie exprime une relation de partitionnement entre deux termes. Ceci permet par la suite de définir une relation de *part_of*, ou une relation de *ingredient_of* ou autre relation de partitionnement spécifique à un domaine particulier telle que la relation *member_of* en biologie.

3. La synonymie permet de rapprocher différents termes sémantiquement proches et de les relier à un même concept par une relation de désignation *denote*.

4. De nombreuses autres relations lexicales (que nous appelons dans la suite *relations transverses*) sont spécifiques à un domaine. Ils peuvent être identifiées pour définir d'autres relations sémantiques entre concepts telles que *causes, relates_to, affects*.

3.3 DYNAMO MAS

L'originalité de DYNAMO est le module de co-construction d'ontologie basé sur la technolgie des SMA. Les SMA sont spécialement pertinents pour la résolution de problèmes complexes et dynamiques tel que la construction et la maintenance d'ontologies. Un SMA est un système ouvert, dynamique dont le traitement est réparti entre les différents agents qui le composent. Ces caractéristiques sont particulièrement intéressantes dans notre contexte, elles facilitent une conception interactive (*la co-construction*) et permettent de considérer l'ontologie *comme un système dynamique réagissant aux évolutions de son environnement.*

Lorsque l'ontographe intervient sur l'ontologie, ses actions sont transmises au SMA. Ce dernier, en fonction des connaissances extraites du corpus (candidats termes et relations lexicales), peut proposer à l'ontographe des évolutions complémentaires (telles que le déplacement, l'élimination d'un concept, l'ajout d'un terme, l'ajout d'une relation, ...). C'est l'évolution du domaine par l'ajout de nouveaux documents au corpus qui fait évoluer l'ontologie proposée par les agents de DYNAMO. En effet, les nouvelles fiches entraînent l'apparition de nouveaux termes, de nouveaux triplets <Ti, Rel, Tj> et la mise à jour des confiances de triplets existants. DYNAMO joue le rôle d'un deuxième ontographe qui confronte les propositions de l'ontographe humain aux données tirées du corpus textuel.

Dans le système, nous considérons une ontologie un SMA. Un *TermAgent* représente sa composante terminologique. Un *ConceptAgent* représente sa composante conceptuelle. Les relations de l'ontologie ne sont pas agentifiées. Elles sont des représentations (ou attributs) des agents. En effet, nous avons priviligié cette solution pour diminuer le nombre d'agents dans le système afin de réduire le temps de calcul. Les relations de l'ontologie sont des liens entre les *ConceptAgent*. Les relations lexicales (non présentent dans l'ontologie) sont des liens entres les *TermAgent*. Aussi, nous avons maintenu cette solution afin de réduire et de simplifier les interactions entre les agents. Avec des agents relations, les interactions deviennent plus nombreuses.

L'état initial du SMA est une organisation d'agents termes. Chaque agent terme est relié avec d'autres agents termes en accord avec les relations lexicales extraites du corpus. Chaque relation établie entre deux agents termes est valuée par un indice de confiance. Cet indice est le couple (Q, I) du triplet <Ti, Rel, Tj>. Ti et Tj sont les deux termes agentifiés. En fonction de cet indice de confiance, chaque *TermAgent* traite ses relations de la plus pertinente (ayant un indice de confiance le plus élevé) à la moins pertinente. Par exemple, une relation de synonymie entre deux agents termes conduit à la création d'un agent concept *ConceptAgent*. Chaque agents termes crée une relation de dénotation avec l'agent concept. La force de cette relation correspond

au couple (Q,I) maximal des relations lexicales de l'agent terme. Une relation d'hypéronymie entre deux agents termes entraîne l'envoi par l'agent terme source d'une demande de traitement de la relation avec sa force (Q, I) au *ConceptAgent* qu'il dénote. Si la demande est prioritaire (force maximal) ce dernier la traite. Il créé une relation *is_a* avec l'agent concept dénoté par l'agent terme cible de la relation d'hypéronymie si la relation *is_a* n'est pas contradictoire avec d'autres relations de l'agent concept. Dans le cas contraire, l'agent concept refuse de traiter la demande et informe son agent terme.

Une nouvelle organisation se construit formée d'agents termes et d'agents concepts. Le rôle des agents concepts (*ConceptAgent*) est de simplifier et d'optimiser cette organisation. Par exemple, chaque concept a un label préféré (l'agent terme ayant la force de dénotation maximal). Si deux agents concepts sont créés avec un même label préféré, ces derniers doivent traiter ce problème (par exemple un des deux agents choisi un autre label). Dans DYNAMO MAS, seuls les triplets ayant dépassés un seuil de confiance sont agentifiés. La mise à jour des confiances des triplets et l'apparition de nouveaux triplets entraînent la création de nouveaux agents, l'augmentation des indices de confiance des relations existantes entre agents et/ou la création de nouvelles relations entre agents. Ainsi, les évolutions du domaine ou les changements effectués par l'ontographe sur l'ontologie provoquent des perturbations dans le SMA. Ces perturbations provoquent des échanges entre agents termes et agents concepts jusqu'à parvenir à un état stable. Le SMA dans son état stabilisé propose alors une nouvelle RTO à l'ontographe. Les principes de fonctionnement du SMA et les mécanismes des agents sont détaillés dans [17].

4. SCÉNARIO D'UTILISATION DE DYNAMO

Nous illustrons le processus d'enrichissement d'ontologie par DYNAMO à l'aide d'un scénario basé sur le corpus et l'ontologie ARTAL. Ce scénario explicite comment utiliser le système DYNAMO et son originalité. A ce stade d'avancement du projet, nous avons déjà un prototype dont les résultats nous permettent de montrer la faisabilité de notre démarche ainsi que de définir les points à améliorer dans l'outil.
L'ontologie ARTAL décrit des *défauts* qui surviennent sur des *Fonctions logicielles* affectant des *composants logiciels* suite à des *événements*. L'ontologie contient 552 concepts et 693 termes. La figure 2 montre le noyau de l'ontologie ARTAL. Les concepts *Trigger_Event, Default, Function* et *Component* sont les éléments de couleur foncée et les termes *_Trigger_Event, _Default, _Function* et *_Component* sont les éléments de couleur claire. Chaque terme est relié à un concept par la relation *Denote*. Sur la figure, nous pouvons aussi voir des relations transverses telles que *Concerns_Function , Causes* ou *Affects_Component*.

Afin d'enrichir l'ontologie par de nouvelles relations, nous avons exploiter le corpus d'ARTAL formé de 792 fiches de pannes informatiques. Chaque fiche est formée d'une seule phrase qui décrit la panne et quelques fois sa localisation et/ou l'événement déclenchent, par exemple *No position reported to the client Station. MU Connection is not accepted by the server*. Le bug informatique présenté dans cette fiche

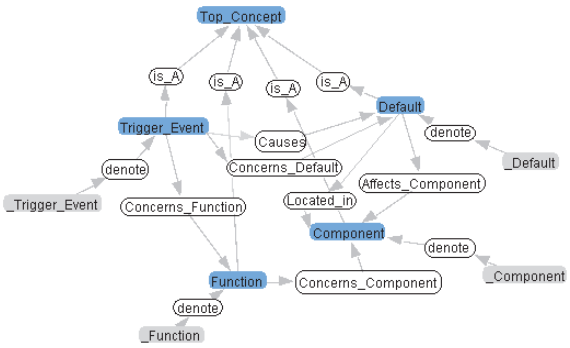

Figure 2: Noyau de l'ontologie ARTAL.

est l'incapacité de l'utilisateur à suivre le déplacement d'un véhicule sur l'IHM (ClientStation) de contrôle (la position du véhicule n'est pas visible sur la carte). La source du problème est la connexion du capteur de position MU sur le véhicule qui est refusée par le serveur (d'où l'incapacité d'envoyer ses positions).
Le volume global du corpus et de 86395 mots. La base des candidats termes est alimentée par les 693 termes de l'ontologie comme *Socket Exception, BlackList Alarm*. Nous avons aussi utilisé un extracteur de termes, *Termostat*, disponible sur le Web [3] pour alimenter cette base par de nouveaux termes. Dans ce scénario, nous nous focalisons sur l'enrichissement de l'ontologie par des relations proposées par le système. Pour cela, nous avons utilisé les mots pivots définis dans TerminoWeb [3] permettant la détection d'une relation lexicale dans une phrase. Par exemple :

- *another term for, also called, synonym* permettent de détecter des relations de synonymies ;
- *in, while, on, to, during* permettent de détecter des relations transverses ;
- *is a part of, element of, represents* permettent de détecter des relations de méronymies ;
- *such as, sort of, kind of, type of* permettent de détecter des relations d'hypéronymies.

Afin d'extraire des relations transverses temporelles (qui sont pertinentes pour l'ontologie étudiée), nous avons utilisé des mots pivots spécifiques tels que *when, if, at, on, before, after*. Le tableau 1 résume le nombre de mots pivots employés par type de relation lexicale.

Relations lexicales	Nombre de mots pivots
Synonymie	12
Hypéronymie	21
Méronymie	23
temporelles	16

Table 1: Nombre de mots pivots employés pour extraire les relations lexicales

5. ANALYSES

La projection des mots pivots et de l'ensemble des termes sur le corpus permet d'extraire les relations lexicales présentées dans le tableau 2.

[3]http://olst.ling.umontreal.ca/ drouinp/termostat_web/

Relation lexicales	Nombre de relations retrouvé
Synonymie	2
Hypéronymie	68
Méronymie	36
temporelles	370

Table 2: Nombre de relations lexicales extraites du corpus ARTAL.

Nous remarquons que le nombre de relations de synonymie, d'hypéronymie et de méronymie est faible par rapport au nombre de relations temporelles. Ceci est dû à l'utilisation de mots pivots génériques pour les trois premièrs types de relations. Le nombre de relations temporelles retrouvées est le plus important car nous avons séléctionné dans le corpus les mots pivots signalant la présence d'une relation termporelle entre termes. Par ailleurs, dans cette expérimentation, seule la forme exacte des termes est recherchée dans le corpus. En conséquence, de nombreuses occurrences de relations (celles où les termes ont une forme différente) sont ignorées par le système d'extraction utilisé.

Le module d'extraction de relations lexicales affecte une valeur de criticité à chaque relation retrouvée entre deux termes. Cette valeur est calculée en fonction de la fréquence d'instanciation de la relation dans le corpus et du nombre total de relations instanciées. Cette criticité permet de juger l'importance et la qualité de la relation retrouvée. Elle permet aussi de filtrer les triplets <Ti, Rel, Tj> en entrée du module *DYNAMO MAS* : uniquement les relations ayant dépassé un seuil de criticité (fixé ici à 0,005) seront prises en compte par le module. Le choix de ce seuil a permis d'éliminer un grande nombre de propositions non pertinentes, ce qui a facilité la comparaison entre l'ontologie initiale et l'ontologie modifiée.

DYNAMO MAS présente ensuite à l'ontographe la nouvelle ontologie, qui est l'ontologie initiale enrichie et modifiée. L'ontographe valide, supprime et/ou modifie les relations proposées par le système grâce à l'interface graphique (figure 3). Il peut aussi visualiser un ensemble d'informations sur les concepts, termes et relations de l'ontologie sous la partie propriété (*Properties* sur la figure 3). La figure 3 montre une partie de l'ontologie ARTAL sur DYNAMO. Notre outil a été implémenté comme un plug-in de l'éditeur d'ontologie Protégé [4].

Les résultats d'enrichissement de l'ontologie ont été présentés à l'expert du domaine pour validation. Le tableau 3 résume la pertinence des nouvelles relations proposées par notre outil. NTR est le nombre total de relations; NNRP est le nombre des nouvelles relations proposées par le système; NNRPV est le nombre des nouvelles relations proposées par le système et validées par l'ontographe.

L'ontographe a jugé qu'aucune des relations *is_a* n'est pertinente pour l'ontologie. Dans l'expérimentation, le système a relié tous les concepts au Top_Concept de l'ontologie. Ce résultat est dû en grande partie à la généricité des mots pivots utilisés pour détecter les relations d'hypéronymie.
Le nombre nul de relation *part_of* est dû au seuil de criticité fixé à 0.005 mais aussi à la généricité des mots pivots utilisés pour les relations de méronymie. Nous avons réalisé un autre test en baissant le seuil à 0.001. Le système a proposé 3 relations est_partie_de dont une a été jugée pertinente par l'ontographe (*Message a_part_of Exception*).
Le meilleur résultat obtenu concerne les relations transverses. En effet, parmi les 38 relations trouvées, 23 ont été validées par l'ontographe. Par exemple, DYNAMO a proposé les relations suivantes :

[4]http://protege.stanford.edu/

Relation conceptuelle	NTR	NNRP	NNRPV
is_a	61	61	0
part_of	0	0	0
temporal	38	38	23

Table 3: Pertinence des relations conceptuelles proposés par DYNAMO.

- entre les concepts *View* et *IllegalArgumentException*, une relation libellée ensuite *located_in* par l'ontographe.
- entre les concepts *Tree* et *Auto Collapse Node*, une relation libellée ensuite *applied_in par l'ontographe.*
- entre les concepts *List* et *ArrayIndexOutOfBoundsException*, une relation libellée ensuite *affected_by par l'ontographe.*

Actuellement, compte tenu de l'avancement du projet, le système n'affecte aux relations conceptuelles qu'un seul des labels suivants : *is_a, a_part_of, function ou denote.* L'ontographe a la possibilité ensuite de le renommer. Pour affiner et enrichir ces relations, nous prévoyons de définir des patrons lexico-syntaxiques spécifiques au corpus pour avoir des relations conceptuelles plus précises.

Aussi, compte tenu du nombre élevé des relations proposées, nous avons présenté à l'ontographe deux ontologies: une construite par notre système et l'ontologie d'origine ARTAL. Les nouvelles relations proposées par le système sont ensuite rajoutées par l'ontographe. En effet, le nombre élevé de relations éronnées *is_a* rend la supression manuelle lourde. Cette limite du système sera traitée en utilisant un système de gestion des propositions. Uniquement les propositions prioritaires seront rajoutées à l'ontologie.

Dans ce scénario, nous n'avons pas montré un cas de supression ou de déplacement de concept ou de terme car ses cas ne sont pas encore implémenté dans le protoype. La supression d'un concept sera traité par le SMA de la façon suivante. L'agent concept correspondant sera détruit. Les agents termes ne seront plus dans la RTO et chercheront à se relier à un nouveau concept plus proche (par exemple, le concept père ou le concept frère). L'agent choisi est celui qui maximise la valeur de criticité de la relation entre les deux agents.

6. INGÉNIERIE D'ONTOLOGIES À PARTIR DE TEXTES

L'exploitation des textes comme sources de connaissances est une alternative au recours aux entretiens menés auprès d'experts. Cette manière de construire des ontologies se développe pour plusieurs raisons. Les experts sont peu disponibles et les mobiliser revient très cher. Faire appel à un

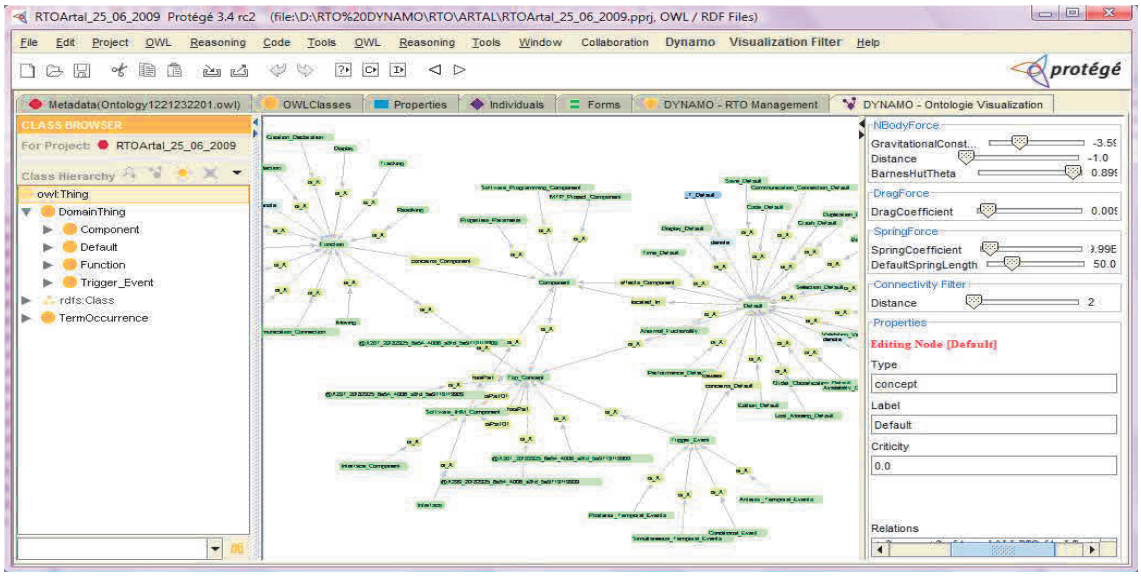

Figure 3: Capture d'écran de l'outil DYNAMO

expert passe par processus de construction manuel, long et coûteux. D'un autre côté, les avancées récentes en matière de Traitement Automatique du Langage naturel (TAL) et en apprentissage laissent présager un repérage de plus en plus automatique des composants d'une ontologie à partir de textes. Le point de départ d'un processus de construction d'une ontologie à partir de textes est le *corpus textuel*, ensemble de documents sur lequel repose cette élaboration.

Différents outils permettant la construction semi-automatique d'ontologies à partir de textes ont été présentés dans la littérature [18]. Les premiers environnements ont vu le jour à la fin des années 90 [1]. Ces derniers fournissent des outils qui aident des utilisateurs à effectuer certaines des tâches principales du processus de développement d'ontologie, telles que : la conceptualisation, l'implémentation, le contrôle de cohérence et la documentation.

Un premier protoype DYNAMO V1 a été proposé pour construire une ontologie à l'aide des SMA [13]. Ce prototype utilise des résultats d'analyses syntaxiques et terminologiques de textes. Il exploite différents critères basés sur des statistiques calculées sur les contextes linguistiques des termes pour créer et positionner les concepts. Le coeur du comportement des agents du SMA est inspiré de la classification. Un algorithme de clusterisation distribué sur les agents permet de rapprocher des termes similaires et d'organiser le SMA en une hiérarchie. L'avantage de cette distribution est de permettre de reprendre, remettre en question et recalculer la classification. En sortie, ce SMA fournit à l'ontographe une organisation hiérarchique de concepts (le SMA lui-même) qui peut être validée, raffinée ou modifiée, jusqu'à obtenir un état satisfaisant du réseau sémantique.
Comme nous l'avons évoqué dans la section 3.1, les approches statistiques ne sont pas efficaces lorsque le volume de données dans les corpus est faible. Les tests réalisés avec le premier prototype de DYNAMO ont confirmé ce constat et c'est pour cela que nous avons adopté une approche par patrons lexico-syntaxiques. Le nouveau système DYNAMO n'est pas une évolution de ce protoype. En effet, nous n'utilisons pas les techniques de classification. De plus, nous avons définis deux types d'agents (agent terme et agent concept) alors que le premier prototype ne contient que des agents concepts. Le premier prototype ne gère pas l'évolution des corpus (l'ontologie est recalculée à chaque modification du corpus) alors que c'est un des objectif majeur du projet. Enfin, dans DYNAMO V1 lorsqu'une modification est apportée par l'ontogrpahe à l'ontolgie, toute la structure de l'ontologie est modifiée. Ceci rend ces modifications illisibles et le temps de réponse du système long. Dans le nouveau DYNAMO, seuls les modifications importantes sont proposées à l'ontographe.

Text Onto Miner (TOM) [7] se base sur les techniques de fouille de données dans les textes pour l'extraction d'informations et la construction d'ontologies. Le corpus est analysé à différents niveaux de granularité (document, paragraphe, phrase) pour en extraire des termes simples, des termes composés, des relations de synonymie, d'homonymie, etc. TOM s'inspire de la plate-forme de TAL GATE : plusieurs processus de traitement de la langue naturelle organisés en pipeline permettent de construire une ontologie. Des techniques de clusering sont employées pour construire une hiérarchie de concepts en calculant des relations de synonymie et des relations d'homonymie entre candidats termes. Cette plateforme est efficace uniquement pour des corpus de grand volume, et ne peut convenir pour des corpus de faible volume comme ceux du projet DYNAMO.

ASIUM [6] permet l'identification de concepts et la proposition d'une taxonomie ensuite validée par un expert. ASIUM est basé sur l'utilisation de filtres lexico-syntaxiques, où l'on tente de catégoriser des entités intéressantes dans le texte à l'aide de schémas de sous-catégorisation. Tout d'abord, ASIUM commence par appliquer des filtres lexicosyntaxiques très génériques au texte pour grouper les termes en fonction de leur correspondance. Les sous-catégorisations engendrées sont des clusters de base formés par la tête des mots qui se

présentent avec un même verbe et après une même préposition (ou avec le même rôle syntaxique). Ce processus est répété pour construire une ontologie de manière ascendante. L'approche ASIUM est centralisée : dans le cas où une classe est mal construite, il faut trouver l'étape du raisonnement qui a engendré ce résultat erroné et modifier la classe correspondante manuellement. Malheureusement, dans ce cas, toutes les étapes consécutives à la création de la classe modifiée sont perdues et doivent être recalculées en tenant compte de la modification. ASIUM [6] présente à l'utilisateur les classes créées à chaque étape du raisonnement, pour essayer de gommer ce problème grâce à une collaboration système-utilisateur. De plus, la hiérarchie complète n'est visible qu'à la fin du processus. Ainsi, l'ontographe peut ne constater l'introduction d'une erreur que trop tard. Dans DYNAMO, la distribution du traitement sur les différents agents permet de faciliter les révisions effectuées par l'ontographe sur l'ontologie.

Comme ASIUM, OntoLearn [19] fait partie également des systèmes employant l'agrégation de termes. Tout d'abord, OntoLearn conserve des statistiques sur l'occurrence des différents termes tout en préservant l'information linguistique sur chacun de ceux-ci. L'agrégation des termes se base sur l'appariement des concepts avec une ontologie déjà existante telle que WordNet. On obtient donc une ontologie qui spécialise une ontologie beaucoup plus générique. OntoLearn enrichit et adapte seulement des ontologies existantes à partir de textes. Dans DYNAMO, nous traitons deux problèmes : celui de la construction d'ontologies et celui de l'enrichissement et de l'évolution d'ontologies sans systèmatiquement les rattacher à une ontologie générique.

DOGMA [15] utilise une approche similaire à celle d'ASIUM. DOGMA reprend l'idée de grouper les termes selon leur similarité syntaxique mais cette fois-ci, en se basant principalement sur les verbes rencontrés dans le texte. DOGMA construit ainsi une taxonomie de termes en utilisant une approche statistique pour identifier des classes parmi des candidats termes en fonction des verbes qu'ils partagent.

CIAULA [4] permet le regroupement de verbes grâce à un algorithme de classification. CIAULA intègre une série d'outils de traitement de la langue dont le but est de construire une hiérarchie de verbes dont les clusters sont sémantiquement proches. L'utilisation de WordNet permet par la suite de sélectionner pour chaque cluster le meilleur label parmi les noms des concepts proposés dans WordNet.

Selon le même principe que OntoLearn, Text-To-Onto [5] permet la construction itérative d'une ontologie à partir de textes. Les auteurs de Text-To-Onto supposent que les documents d'un domaine contiennent les relations et les concepts qui seront à inclure dans l'ontologie. Une ontologie fondamentale telle que WordNet est enrichie avec les concepts trouvés dans les documents. Les nouveaux concepts génériques et spécifiques sont trouvés à l'aide de filtres (*patterns*) ou à partir des regroupements de concepts. Par la suite, ces concepts seront liés par la relation *sorte_de* (*a_kind_of*). Les relations entre concepts sont découvertes par apprentissage automatique à partir d'exemples de filtres linguistiques et de règles d'associations. Finalement, l'ontologie est élaguée en fonction de mesures statistiques et présentée à l'expert pour une évaluation.

Text-To-Onto et DYNAMO utilisent tous deux une approche linguistique pour construire une ontologie. Dans DYNAMO, nous insistons sur la notion de co-construction d'ontologie, c'est-à-dire que l'ontographe observe, valide ou annule les modifications proposées par le système. Dans Text-to-Onto, seule l'ontologie finale est proposée à l'ontographe.

En résumé, DYNAMO se démarque des autres outils présentés par les points suivants :

1. DYNAMO se base sur la technologie SMA pour co-construire une ontologie : Le SMA est lui-même un sur-ensemble de l'ontologie.
2. DYNAMO traite à la fois le problème de construction et d'enrichissement d'ontologies : il offre la possibilité de *construire incrémentalement* en prenant en compte de nouvelles données (tirées des analyses textuelles et de l'interaction avec l'ontographe) en évitant ainsi de reprendre le processus de construction de zéro.
3. La dynamique de DYNAMO permet à l'ontographe d'effectuer des corrections sur l'ontologie qui seront prises en compte en temps réel par le système. La propagation des modifications sur l'ontologie est rendue explicite à l'ontographe, ce qui lui permet d'annuler son action ou d'accepter de nouvelles modifications proposées par DYNAMO.
4. Lorsque l'ontographe est satisfait de l'ontologie obtenu, DYNAMO permet de l'exporter au format OWL en vue de l'utiliser ou de la partager.

7. CONCLUSION ET PERSPECTIVES

Nous avons présenté un outil de co-construction d'ontologies à partir de textes, DYNAMO qui utilise les résultats du traitement automatique de texte d'un domaine.
L'originalité de DYNAMO est son SMA d'une part, et la possibilité donnée à l'ontographe de modifier, ajouter et supprimer des éléments de l'ontologie d'autre part. Le rôle du SMA est de co-construire l'ontologie avec l'ontographe en lui faisant des propositions de modification. Un scénario d'enrichissement d'une ontologie avec DYNAMO nous a permis de montrer les résultats obtenus ainsi que les limites actuelles du système. Afin d'améliorer la qualité des résultats de DYNAMO, nous avons prévus la réalisation des points suivants :
Améliorer les entrées du système :

- En implémentant un mécanisme d'extraction de relations lexicales à l'aides d'expressions régulières. Ce mécanisme de projection est plus efficace que la projection de mots pivots pour rechercher des relations lexicales;
- En définissant des patrons lexico-syntaxiques spécifiques au corpus étudié au lieu d'utiliser des mots pivots génériques, afin d'améliorer la pertinence des relations lexicales extraites.

Améliorer le mécanisme de co-construction :

- En améliorant les algorithmes présents dans les agents termes et les agents concepts du SMA. A ce stade,

seulement une partie des algorithmes des agents termes a été étudiées et implémentées;

- En implémentant les autres fonctionnalités de modifications de l'ontologie à travers l'interface graphique. Une typologie de changements et leurs effets sur l'évolution de la RTO est à l'étude.

- En étudiant et implémentant un SMA capable d'apprendre des patrons lexico-syntaxiques en fonction des patrons prédéfinis et des nouvelles fiches rajoutées au corpus. En effet, les entrées fournies par le module *Corpus Analyser* au module *DYNAMO MAS* sont fortement dépendantes des résultats de projection de patrons lexico-syntaxiques. Or, en phase d'évolution, lorque de nouvelles fiches s'ajoutent au corpus, il se peut qu'aucun patron ne donne de résultat. Dans ce cas, aucune proposition n'est faite par le système.

8. ADDITIONAL AUTHORS

Sylvain Rougemaille (UPETEC (Emergence Technologies for Unsolved Problems) 10 avenue de l'Europe, Ramonville 31520 France, `sylvain.rougemaille@upetec.fr`)
Mohamed Mbarki (Artal Technologies, Rue Pierre Gilles de Gennes Ensemble "La rue" - Bât. 9, BP 38138, 31681 LABEGE Cedex, `mohamed.mbarki@artal.fr`)

9. REFERENCES

[1] N. Aussenac-Gilles, S. Despres, and S. Szulman. The TERMINAE Method and Platform for Ontology Engineering from texts. In P. Buitelaar, P. Cimiano, , and , editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 199–223. IOS Press, http://www.iospress.nl/, janvier 2008.

[2] N. Aussenac-Gilles and P. Séguéla. Les relations sémantiques : du linguistique au formel. *Cahiers de Grammaire. "Sémantique et Corpus". Textes réunis par A. Condamines*, 25:175–198, décembre 2000.

[3] C. Barrière and A. Akakpo. Terminoweb: A software environment for term study in rich contexts. In *Proceedings of International Conference on Terminology, Standardization and Technology Transfer*, pages 103–113, Beijing, August, 25-26 2006. Encyclopedia of China Publishing House.

[4] R. Basili, M. T. Pazienza, T. Vergata, T. Vergata, and P. Velardi. Integrating general-purpose and corpus-based verb classification. *Computational Linguistics*, 22, 1996.

[5] P. Cimiano and J. Völker. Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz, and E. Metais, editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain, JUN 2005. Springer.

[6] D. Faure, C. Nédellec, and C. Rouveirol. Acquisition of semantic knowledge using machine learning methods: The system "asium". Technical report, Universite Paris Sud, 1998.

[7] P. Gawrysiak, G. Protaziuk, H. Rybinski, and A. Delteil. Text onto miner - a semi automated ontology building system. In *the 17th International Symposium on Methodologies for Intelligent Systems (ISMIS), York University, Toronto (Canada), May 20-23*, pages 563–573, 2008.

[8] Z. S. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.

[9] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *14th International Conference on Computational Linguistics (COLING), Nantes (France), August 23-28*, pages 539–545, 1992.

[10] A. Maedche. *Ontology learning for the Semantic Web*, volume 665. Kluwer Academic Publisher, 2002.

[11] V. Malaise. *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels*. PhD thesis, Paris 7 Denis Diderot University, 2005.

[12] E. Morin. Using lexico-syntactic patterns to extract semantic relations between terms from technical corpus. In *5th International Congress on Terminology and Knowledge Engineering (TKE)*, pages 268–278, Innsbruck, Austria, 1999. TermNet.

[13] K. Ottens, N. Hernandez, M.-P. Gleizes, and N. Aussenac-Gilles. A Multi-Agent System for Dynamic Ontologies. *Journal of Logic and Computation, Special Issue on Ontology Dynamics*, 19:1–28, 2008.

[14] B. M. Paul Buitelaar, Philipp Cimiano. *Ontology Learning from Text: Methods, Evaluation and Applications*. Frontiers in Artificial Intelligence and Applications Series. IOS Press, Amsterdam, 7 2005.

[15] M.-L. Reinberger and P. Spyns. Discovering knowledge in texts for the learning of dogma-inspired ontologies. In *Proceedings of the workshop Ontology Learning and Population*, ECAI04, pages 19–24, Valencia, Spain, August 2004.

[16] A. Reymonet, J. Thomas, and N. Aussenac-Gilles. Modelling ontological and terminological resources in OWL DL. In P. Buitelaar, K.-S. Choi, A. Gangemi, and C.-R. Huang, editors, *OntoLex07 - From Text to Knowledge: The Lexicon/Ontology Interface Workshop at ISWC07 6th International Semantic Web Conference, Busan (South Korea), 11/11/07*, 2007.

[17] Z. Sellami, M.-P. Gleizes, N. Aussenac-Gilles, and S. Rougemaille. Dynamic ontology co-construction based on adaptive multi-agent technology. In *International Conference on Knowledge Engineering and Ontology Development (KEOD), Madeira (Portugal), 06/10/09-08/10/09*, 2009.

[18] M. Shamsfard and A. Abdollahzadeh Barforoush. The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review*, 18(4):293–316, 2003.

[19] P. Velardi, R. Navigli, A. Cucchiarelli, and F. Neri. *Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies. In Buitelaar, P., O. Cimiano & B. Magnini (eds.)*. IOS Press, Amsterdam, 2005.