Extraction de concepts et relations sémantiques à partir de labels d'ontologies

Thierry Declerck[1] et Piroska Lendvai[2]

[1]Language Technology Lab, DFKI GmbH
declerck@dfki.de
[2] Research Institute for Linguistics, Hungarian Academy of Sciences
piroska.r@gmail.com

Résumé : Dans cet article, nous proposons d'appliquer le traitement automatique du langage aux contenus textuels des labels d'une ontologie. A l'origine notre but était d'améliorer de la sorte l'annotation sémantique de documents textuels, mais l'analyse des labels nous a mené à examiner la possibilité de proposer sur cette base une extension ou une réorganisation de l'ontologie elle-même.

Mots-clés : Traitement automatique du langage, Ingénierie des connaissances

1 Introduction

Les travaux en cours présentés dans cet article ont leur origine dans le projet de recherche Theseus-Medico[1], qui traite, entre autres, de l'indexation sémantique d'images médicales. Dans ce contexte, nous avons été amenés à prendre en considération comme ressources textuelles des rapports radiologiques. Ces rapports décrivant (en partie) ce que le praticien observe dans la radiographie, l'annotation sémantique des textes peut être utilisée pour indexer sémantiquement le contenu de l'image. Le travail d'annotation sémantique du texte se fait à l'aide d'un parseur robuste qui consulte RadLex[2] comme ressource terminologique et ontologique.

Au cours de la mise en œuvre et de l'évaluation de nos outils et ressources notre attention s'est portée sur la possibilité d'extraire à partir des données textuelles des suggestions pour l'amélioration ou l'extension de l'ontologie[3]. Mais avant de nous pencher sur l'annotation sémantique des rapports radiologiques et l'extraction de relations sémantiques à partir de ces mêmes documents, nous avons réfléchi à l'utilité

[1] Voir http://theseus-programm.de/anwendungsszenarien/medico/default.aspx pour plus d'information. Nous avons noté aussi une grande similarité d'intérêt avec le travail proposé par (Mhiri & Després, 2007).

[2] RadLex est une terminologie et une ontologie pour le domaine de la radiographie, qui existe maintenant en une version bilingue pour l'Anglais et l'Allemand (voir http://www.radlex.org/viewer pour la version Anglaise). Nous n'avons pas trouvé un équivalent en français sous une forme électronique, mais nous avons trouvé une Nomenclature Anatomique qui propose une table de correspondances de termes anatomiques pour le latin, le français et l'anglais (voir Doyon et al., 1998).

[3] Nous avons suivi une approche similaire à celle décrite par (Aussenac-Gilles & Jacques, 2008).

de proposer en premier lieu une analyse des contenus textuels des labels de l'ontologie RadLex, afin de faciliter l'établissement des correspondances entre ontologie et textes. Cette analyse des labels nous a mené à examiner la possibilité de proposer sur cette base une extension ou une réorganisation de l'ontologie elle-même.

Notre article est concentré sur cet aspect. Nous décrivons en premier lieu RadLex, puis le type de patrons textuels et linguistiques que nous pouvons reconnaître dans les labels de l'ontologie, pour achever sur des propositions générées automatiquement pouvant mener à la réorganisation de structures ontologiques ou à leur extension.

2 L'Ontologie RadLex

En 2003, la « Radiological Society of North America » (RSNA) a commencé le développement d'un lexique standardisé de termes radiologiques, lequel a pris depuis 2007 la forme d'une ontologie. Ce travail a pour but de supporter la génération de rapports structurés d'observations de radiographies. Nous disposons maintenant de la version 2.0 de cette ontologie pour l'Anglais et l'Allemand[4]. Le lien entre les deux langues est assuré par l'intermédiaire de noms de classes, qui ont la forme « RID + chiffres ». Ainsi, la classe avec le nom « RID1373 » réfère aux labels « right pleural fissure » et « Rechter Interlobärspalt » pour l'Anglais et l'Allemand respectivement.

2.1 La structure de RadLex

La structure générale de l'ontologie (version anglaise 2.0) est représentée dans la figure 1. L'ontologie introduit également une classe nommée « modifier », ce qui réfère au terme linguistique « modification », qui est généralement utilisé pour les adjectifs, les adverbes ou les syntagmes prépositionnels utilisés pour modifier substantifs et verbes qui gouvernent des syntagmes ou des phrases. Les termes regroupés sous la classe «modifier » indiquent souvent des relations. Bien que cette classe « modifier » et sa liste de termes s'avèrent être très utiles pour la détection des « modifications » dans les rapports, nous constatons que dans le stade actuel du développement de l'ontologie RadLex il n'existe pas de restriction formulée sur le type de classes qui peuvent être « modifiées ».

3 Analyse des Labels de RadLex

Ainsi que déjà indiqué dans l'introduction, nous nous consacrons tout d'abord à l'analyse des contenus textuels des labels des classes de RadLex. Pour ce faire nous testons deux approches.

---

[4] Une nouvelle version (v3.0) pour l'anglais est disponible depuis Juillet 2009. Elle intègre des liens avec les ressources FMA, UMLS et SNOMED, mais propose également des changements au niveau de la modélisation.

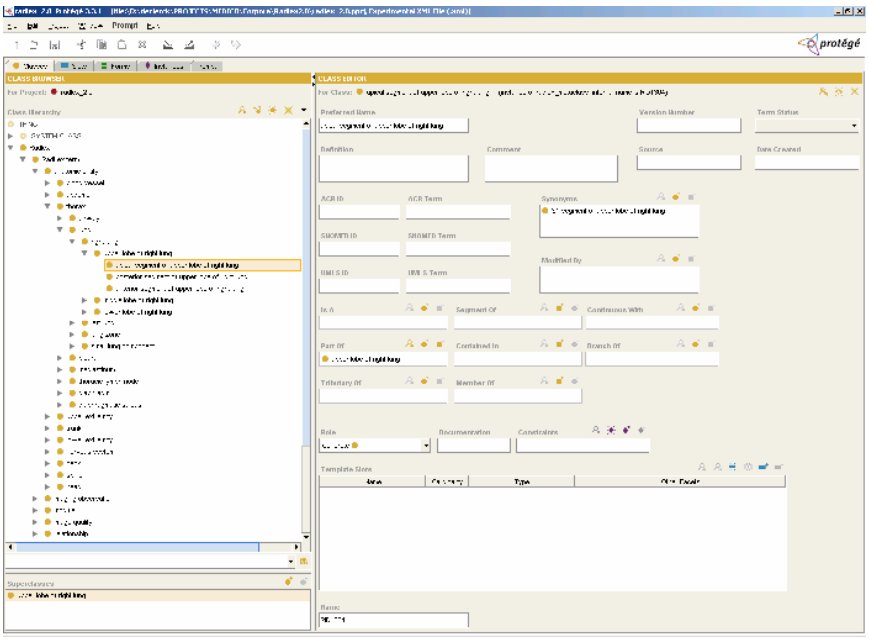

Fig. 1 – La structure générale de RadLex, visualisée par l'outil « protégé ». A gauche on peut voir la hiérarchie des classes, dénotées par leurs « labels ». A droite, on peut voir le genre de relations définies pour les classes.

Nous décrivons d'abord une méthode non-supervisée, appliquée directement aux séquences de mots, nous aidant à identifier sur la base de « string matching »[5] des patrons qui suggèrent des relations de dépendance entre un terme et le reste de la séquence de mots, et ensuite une méthode linguistique qui propose des annotations des labels. Les deux approches, pouvant être combinées, servent de base à la formulation de règles heuristiques pour une extension ou une réorganisation de l'ontologie.

3.1 Reconnaissance non-supervisée de patrons

Nous opérons ici sur la base de résultats d'un algorithme qui induit des grammaires, et qui a déjà prouvé son utilité dans le traitement de textes dans le domaine du patrimoine culturel (Lendvai, 2008)[6]. L'induction de grammaires peut être particulièrement utile pour découvrir des patrons syntaxiques (et même sémantiques) dans le cas de données textuelles brèves et incomplètes, telles que les entrées dans des banques de données. Nous étendons ici cette approche aux labels d'ontologies.

[5] Nous limitons cette approche pour l'instant à l'analyse de l'Anglais, qui ne connaît pas de grandes variations morphologiques. Ceci est encore plus vrai dans le cadre des textes des labels de RadLex.

[6] Plus de détails sur l'induction de grammaires sont donnés dans (Adriaans, P. & Zaanen, M. van., 2004).

L'algorithme consiste à d'abord aligner les phrases dans le corpus d'entrée[7]. Si plusieurs phrases contiennent des (séquences de) mots identiques, les parties non-identiques de ces phrases peuvent être interprétées comme « clusters » d'un même type. Les séquences identiques des phrases alignées peuvent être considérées comme le « head » des phrases. Cette approche semble être très appropriée dans le cas de RadLex (ou d'autres ontologies), vu que les labels que l'on trouve au sein d'une hiérarchie de classes (par exemple pour la relation « Is_A_Segment_Of » dans RadLex) contiennent souvent des (séquences de) mots identiques, comme l'exemple ci-dessous le montre :

« proximal left anterior descending artery » Is_A_Segment_Of « left anterior descending artery »

Dans ce simple cas, et sur la base des résultats de l'algorithme d'induction de grammaires, nous pouvons considérer « proximal » comme un modifiant de la séquence de mots identiques aux deux labels. Les « clusters » peuvent être proposés aux experts, dans les cas où ils seraient absents de l'ontologie. Nous donnons ci-dessous un exemple (partiel) montrant comment les patrons générés par l'algorithme peuvent être à la base d'une suggestion de la réorganisation de l'ontologie :

(modification à droite)
@@ interphalangeal ^ joint ^ @@
@@ interphalangeal joint ^ of ^ @@
@@ proximal interphalangeal joint of ^ toe ^ @@
@@ proximal interphalangeal joint of ^ finger ^ @@
@@ distal interphalangeal joint of ^ finger ^ @@

(modification à gauche)
@@ ^ capsule ^ of proximal interphalangeal joint of finger @@
@@ ^ capsule ^ of distal interphalangeal joint of finger @@
@@ ^ collateral ^ ligament of proximal interphalangeal joint of finger @@
@@ ^ collateral ^ ligament of distal interphalangeal joint of finger @@
@@ ^ collateral ^ ligament of proximal interphalangeal joint of toe @@
@@ ^ collateral ^ ligament of distal interphalangeal joint of toe @@ @@

L'algorithme reconnait les (séquences de) mots identiques dans les entrées et propose les « catégories » *joint* et *of,* en sus des deux « heads » *interphalangeal* et *interphalangeal joint* (les deux premiers exemples). Nous pouvons observer ici que la proposition « of » peut être éliminée sur la base d'une liste de « stop words »[8]. Le terme « joint » est déjà présent dans l'ontologie, donc il ne sera pas proposé à l'expert.

---

[7] Ici il s'agit tout simplement des contenus textuels des labels de RadLex, qui sont constitués en très grande majorité de syntagmes nominaux.

[8] Mais d'un autre côté nous préférons ici garder ce genre de mots et leur attribuer une catégorie linguistique (préposition), car il joue un rôle important dans le cadre de la classification des relations entre concepts/classes.

Mais que faire de « intraphalangeal » ? Nous observons que ce terme n'est pas présent en tant que tel dans l'ontologie (on pourrait le voir par exemple repris sous la rubrique « modifier »). Mais nous observons aussi que ce terme apparait souvent dans l'ontologie, mais toujours en combinaison avec le terme « joint ». Donc notre heuristique tend à ne pas proposer le terme « intraphalangeal » au spécialiste, car il semble faire part intégrante d'un terme et donc ne pas posséder de « vertu modificatrice ». Mais le terme « interphalangeal joint » n'est pas présent non plus en tant que tel. Nous observons de surcroît une variation de contextes dans lesquels ce terme potentiel apparait. Par exemple : « distal interphalangeal joint » et « proximal interphalangeal joint ». Notre heuristique consisterait donc à proposer une nouvelle classe « interphalangeal joint »[9]. Nous observons également, que « interphalangeal joint » qualifie au moins deux termes différents, à savoir « interphalangeal joint of finger » et « interphalangeal joint of toe ». Prenant en ligne de comptes que l'ontologie RadLex est très pauvre en attributs (*properties*) attachés aux classes, une suggestion pourrait être d'introduire en fin de compte une nouvelle classe « interphalangeal joint » avec les attributs correspondants (et comprenant des informations sur leur domaine et leur portée).

3.2 Analyse linguistique des labels

Cette approche associe des catégories linguistiques au contenu textuel des labels. Sur la base de cette information, des généralisations peuvent être proposées et des heuristiques mènent à des suggestions de changements de l'organisation de l'ontologie adressées aux experts du domaine ou de l'ingénierie ontologique. Nous présentons juste un exemple concret. Les labels allemands de base qui nous ont menés à une suggestion de réorganisation sont :

> « Ligamentum des Handgelenks » (*ligament of wrist joint*), « Handgelenk » (*wrist joint*)
> « Ligamentum des Ellenbogengelenks » (*ligament of elbow joint*), « Ellenbogengelenk » (*elbow joint*)

Les labels ont été annotés avec des informations catégorielles et de dépendances. Le syntagme « Ligamentum des Handgelenks « est annoté comme étant un syntagme nominal avec la tête « *Ligamentum* » et la modification génétive post-nominale « *des Handgelenks* ». Ce label annoté est mis en relation avec le label annoté « *Handgelenk* », car ils contiennent la même tête d'un syntagme nominal[10]. Nous observons que dans l'ontologie ces deux concepts sont reliés par une relation de type

---

[9] Cette suggestion est renforcée par le fait que les deux mots „proximal" et „distal" sont repris comme termes sous la classe „modifier". Une nouvelle classe « interphalangeal joint » pourrait donc être enrichie d'attributs (*properties*), telles que « distal » et « proximal ». RadLex ne fait pour l'instant presque pas usage d'attributs.

[10] Deux remarques ici: On peut voir que pour l'Allemand il est impératif de travailler au moins sur la base des lèmes, car un « string matching » aurait échoué : la premiére occurrence du mot « Handfgelenk » est au génitif et prend un « s ». La deuxième remarque : on peut observer qu'il semble aisé de transformé l'algorithme non-supervisé pour le laisser « travailler » sur des données textuelles annotées. Nous allons faire des expériences sur ce thème.

« Is_A ». Mais nous observons également que pour le deuxième exemple, qui a les mêmes caractéristiques linguistiques, et dont la classe contenant la tête « Ellenbogengelenk » comme terme fait partie de la même super-classe que « Handgelenk », nous avons une relation « Part_Of » entre le syntagme contenant la modification et la classe contenant seulement la tête

Nous partons ici du principe selon lequel « linguistic regularities always characterise the same kind of knowledge, such as semantic relations » (Aussenac Gilles et Jacques, 2008), et nous nous demandons donc si une relation n'a pas été encodée de manière erronée. Le spécialiste de l'ontologie nous a confirmé le problème, en nous signalant que l'ontologie de référence, la version anglaise, était en train de subir une réorganisation et que la classification des relations « Is_A » et « Part_Of » était soumise à une révision globale.

4 Conclusion

Nous avons présenté deux approches pour l'analyse des contenus textuels des labels de l'ontologie « RadLex » et montré le potentiel que ce genre d'analyse peut avoir sur la réorganisation de l'ontologie elle-même.
Comme prochaine étape, nous voulons essayer de combiner les approches. Aussi nous aimerions étendre notre travail à d'autres langues, comme le Français. Dans ce cas nous pensons à établir une version française de l'ontologie RadLex à partir d'une terminologie, et ce sur la base des régularités statistiques linguistiques que nous pourrions trouver dans la terminologie.

Références

ADRIAANS, P. & ZAANEN, M. VAN. (2004) Computational Grammar Induction for Linguists. *In Grammars; special issue with the theme "Grammar Induction", 7*, p, 57-68.

AUSSENAC-GILLES N. & JACQUES M-P. (2800). Designing and evaluating patterns for relation acquisition from texts with Caméléon. *In Terminology, 14:1* p. 45-73. John Benjamins.

BUITELAAR P. & DECLERCK T. (2003). Linguistic annotation for the semantic web. In S. HANDSCHUH & S. STAAB Eds, *Annotation for the Semantic Web*, p. 93–111. IOS Press.

COLLINS M. (1999). Head-Driven Statistical Models for Natural Language Parsing. *Ph.D. thesis*. University of Pennsylvania.

DOYON D., DOMENGIE F., FRANCKE J.-P. & VEZINA G. (1998). Nomenclature anatomique radiologique internationale. *Collection des Abrégés d'Imagerie Radiologique*. Paris.

LENDVAI P. (2008). Alignment-based expansion of textual database fields. In A. GELBUKH Ed. *CICLing 2008*. Berlin/Heidelberg.

MHIRI S. & DESPRES S. (2007). Towards an Ontology Visualization Tool for Indexing DICOM Structured Reporting Documents. *In 10th Intl. Protégé Conference*. Budapest

.