Patrons sémantiques pour l'extraction de relations entre termes - Application aux comptes rendus radiologiques

Lionel Ramadier[1, 2] Mathieu Lafourcade[1]
(1) LIRMM, 34095 Montpellier, France
(2) IMAIOS, 34000 MONTPELLIER , France
lionel.ramadier@lirmm.fr, mathieu.lafourcade@lirmm.fr

RÉSUMÉ
Dans cet article nous nous intéressons à la tâche d'extraction de relations sémantiques dans les textes médicaux et plus particulièrement dans les comptes rendus radiologiques. L'identification de relations sémantiques est une tâche importante pour plusieurs applications (recherche d'information, génération de résumé, etc). Nous proposons une approche fondée sur l'utilisation de patrons sémantiques vérifiant des contraintes dans une base de connaissances.

ABSTRACT
Here the title in English.

In this paper we tackle semantic relation extraction from radiology reports. The discovery of semantic relations from text becomes an important task for several applications such as Information Extraction, Summarization, etc. We introduce the approach based on semantic patterns that has been tested with a scope of investigating the increase of the performance when using semantic constraints.

MOTS-CLÉS : extraction de relations, TALN, domaine médical, relations sémantiques.

KEYWORDS: semantic relations, extraction of relations, medical domain, BioNLP.

1 Introduction

L'extraction de relations sémantiques à partir de textes peut avoir deux buts principaux. Le premier est d'aider à indexer des documents en vue de réaliser un système de recherche d'information. Le deuxième est l'amélioration de la qualité de bases de connaissances (Lee *et al.*, 2004) que l'on peut ainsi enrichir en nouvelles relations (non présentes dans la base de connaissances) présentes dans les textes. Cet article traite de l'extraction de relations sémantiques avec ajout de contraintes entre des entités, dans le domaine médical et plus particulièrement dans celui de la radiologie. Plusieurs travaux se sont focalisés sur l'analyse sémantique de textes médicaux, soit en proposant une méthode de reconnaissance des entités médicales à l'aide de ressources ontologiques du domaine (Delbecque *et al.*, 2005), soit en s'intéressant à la tâche d'extraction de relations entre les entités médicales (Embarek & Ferret, 2008). Song et al (Song *et al.*, 2015) proposent un système qui combine les deux approches.

Nous proposons ici IMAIOS, un système d'extraction de relations sémantiques à partir de comptes rendus de radiologie en français. Notre corpus est constitué de 35 000 comptes rendus représentant différentes modalités d'imagerie médicale (IRM, scanner, radiographie, échographie, radiologie

interventionnelle). Le système IMAIOS utilise comme base de connaissances un réseau lexico-sémantique de langue française, JeuxDeMots [1] (JDM). Bien que le réseau JDM soit un réseau de connaissances générales, il inclut aussi des connaissances médicales (surtout radiologiques) qui ont été ajoutées dans le cadre du projet IMAIOS (en collaboration avec des médecins radiologues). Notre méthode repose ainsi sur une base de connaissances qui contient une grande quantité d'informations telles que le POS (part of speech), le lemme, des variantes orthographiques, des pluriels ainsi que de nombreuses relations sémantiques (80 types de relations différentes).

Dans cet article, nous présentons une méthode d'extraction des relations sémantiques (cause, conséquences, symptômes, lieu, partie de,...) entre des termes. Dans ce but, nous avons construit manuellement des patrons linguistiques à partir d'un sous-ensemble de notre corpus de comptes rendus radiologiques. Comme certains patrons linguistiques (*de, avec, sous, etc.*) peuvent être trop généraux, des contraintes sémantiques sur les relations ont été ajoutées. Par exemple, dans la phrase *néoplasie du sein*, le système, reconnaissant *néoplasie* comme une maladie et *sein* comme un lieu anatomique, identifie la relation comme étant une relation de lieu : *néoplasie* r_lieu *sein*, grâce à des contraintes sémantiques, bien que le patron linguistique *du* soit très général. Une évaluation de l'ajout de contraintes est proposée.

2 Etat de l'art

La plupart des travaux concernant l'extraction de relations sémantiques se concentrent sur des relations indépendantes du domaine (Snow *et al.*, 2006) ; (Chklovski & Pantel, 2004). Dans le **domaine général** l'extraction de relations sémantiques entre entités utilise soit des approches statistiques (Hindle, 1990) ; (Nazar *et al.*, 2012) et/ou des techniques d'apprentissage automatique ainsi que des approches basées sur l'utilisation de patrons ou des règles d'extraction (Hearst, 1992) ; (Auger & Barrière, 2008) ; (Hogenboom *et al.*, 2012) voire des approches combinant ces deux techniques (Suchanek *et al.*, 2006).

Au vu des difficultés à déterminer le type de relations entre deux termes à cause de l'ambigüité des patrons linguistiques, (Girju *et al.*, 2003) ont proposé d'ajouter des contraintes sémantiques pour la détection de relations de méronymie. Ils ont déterminé 20 contraintes par un algorithme d'apprentissage et ont obtenu une précision de 83%. D'autres contraintes lexicales et syntaxiques ont été appliquées sur des relations exprimées par des verbes (Fader *et al.*, 2011). Plus de 30% des relations extraites grâce à ces contraintes ont une précision de 80% ou plus. Dans notre corpus il est difficile de se fier aux verbes étant donné qu'ils sont fréquemment omis dans les comptes rendus radiologiques. Concernant l'extraction de relations sémantiques dans le domaine biomédical, il existe quatre principales techniques : l'une se base sur les co-occurences (Jelier *et al.*, 2005), une autre fait intervenir des patrons linguistiques ou de règle (Auger & Barrière, 2008) (Song *et al.*, 2015). Enfin, il existe des techniques basées sur des approches d'apprentissage supervisé à l'aide par exemple de machine à vecteurs de support (SVM) (Rink *et al.*, 2011) ainsi que des approches combinant ces deux techniques (Suchanek *et al.*, 2006), (Chowdhury & Lavelli, 2012). Embarek *et al.* (Embarek & Ferret, 2008) ont proposé un système fondé sur des patrons construits automatiquement en vue de l'extraction de quatre relations entre cinq types d'entités médicales (*maladie-traitement, maladie-médicament, maladie-examen, maladie-symptômes*). Plusieurs travaux (Abacha & Zweigenbaum, 2011), (Lee *et al.*, 2004) se sont intéressés à l'extraction de relations sémantiques reliant deux types d'entités médicales, à savoir une *maladie* et un *traitement* en utilisant une méthode basée sur des patrons lexicaux. Abacha

1. http ://www.jeuxdemots.org/jdm-accueil.php

et al. utilisent une une méthode semi-automatique pour la génération de patrons alors que Lee *et al.* ont utilisé une méthode manuelle. Les principales relations extraites sont les relations "traitement", "remède ("cure" en anglais), "détection", "signe". Les méthodes à base de patrons linguistiques permettent une analyse efficace des relations entre entités, mais certaines relations sont indétectables par ces techniques à cause de l'ambiguïté de certains patrons. Dans ce contexte, nous décrivons une approche qui ajoute des contraintes sémantiques sur les relations afin d'améliorer la précision du système d'extraction.

3 Méthode : patrons sémantiques

Les relations sémantiques extraites ont été choisies suivant les conseils de radiologues. Le système consiste dans un premier temps à identifier les relations sémantiques entre les termes en utilisant des patrons linguistiques. Par exemple, dans la phrase *fracture du plancher orbital passant par le canal infra-orbitaire*, la première étape identifie *fracture du plancher orbital* comme étant une maladie, et canal infra-orbital comme un lieu anatomique. Grâce au patron linguistique *passant par le*, nous pouvons valider la présence d'une relation sémantique entre les deux termes.

Notre technique d'extraction via un filtrage des candidats (issus des patrons linguistiques) par des patrons sémantiques utilise comme base de connaissances le réseau lexico-sémantique JeuxDeMots (JDM) (Lafourcade, 2007). Bien qu'il s'agisse d'un réseau de connaissances générales, il contient des données de plusieurs domaines de spécialités (botanique, ichtyologie, etc.) et en particulier du domaine de la médecine/radiologie (Ramadier *et al.*, 2014) (**hypothèse de non séparation**). L'utilisation de connaissances générales couplées à celles de spécialités permet une amélioration de certains aspects de l'analyse de textes médicaux, en particulier leur indexation.

Le réseau JDM est un graphe lexico-sémantique pour la langue française dont les relations entre termes sont capturées par la combinaison d'un GWAP (Game With A Purpose) (Von Ahn, 2006) avec un outil contributif nommé Diko (contribution manuelle et inférences automatiques avec validation (Zarrouk *et al.*, 2014)). Ainsi, il existe 25 580 relations entrantes dans le réseau lexical JDM pour le terme *anatomie* ainsi que 15 000 relations pour le terme *maladie*.

3.1 Extraction de relations sémantiques

Notre approche d'extraction de relations a été basée dans un premier temps sur l'utilisation de patrons linguistiques. Pour chaque type de relations (table 1), nous construisons des patrons et nous les comparons avec les phrases pour identifier la relation pertinente.

Certains auteurs (Cimino *et al.*, 1993) ont déjà noté que l'utilisation de patrons est une méthode efficace pour l'extraction automatique d'informations ou de relations s'ils ont été correctement conçus.

3.2 Patrons linguistiques + contraintes = patrons sémantiques

Pour les patrons linguistiques créés (table 2), plusieurs difficultés liées à l'ambiguité sont apparues. Par exemple, pour la *relation de localisation*, nous pouvons distinguer deux types de relations dépendant du patron. Premièrement nous avons la relation *r_lieu* (carcinome hépatocellulaire *au niveau du*

types de relation *R*	signification de A *R* B
caractéristique	A a B comme caractéristiques (adjectifs) typiques possibles. Exemple : *carcinome hépatocellulaire* carac *hypervasculaire*
localisation	A a B comme lieux typiques où peut se trouver le terme/objet en question. Exemple : *lobe caudé* localisation *foie*
cible	A a B comme population affecté par le terme. Exemple : *rougeole* cible *enfant*
holonymie	A a B comme tout. Exemple : *fémur* holo *membre inférieur*
partie de	A a B comme parties typiques. Exemple : *fémur* has part *col du fémur*
signe	A a B pour symptômes/signes. Exemple : *grippe* symptôme *fièvre*
cause	B est une cause de A. Exemple : *cirrhose* cause *alcoolisme*
conséquence	B est une conséquence possible de A. Exemple : *accident vasculaire cérébral* conséquence une *hémiplégie*
traitement	A a B comme traitement médicale adapté. Exemple : *anévrisme cérébral* traitement *embolisation*
accompagnement	A est souvent accompagné par B. Exemple : *luxation* accompagné par *fracture*

TABLE 1 – Liste des relation à détecter

types de relation *R*	exemples de patrons linguistiques pour *R*
caractéristique	$x est caractérisé par $y
localisation	$x au niveau de $y ; $x se trouve dans $y ; $x développé aux dépens de $y ; $x du $y
cible	$x n'affecte que les $y
holonymie	$x fait partie de $y
partie de	$x a comme partie $y ;$x se compose de $y
signe	$x se manisfeste par $y
cause	$x déclenchant $y ; $x peut produire $y
conséquence	$x provoque $y ; $x menant à $y
traitement	$x traité par $y
accompagnement	$x associé à des $y ; $x s'accompagne d'un $y

TABLE 2 – Exemples de patrons linguistiques

foie). La deuxième relation est l'holonymie. Un holonyme A d'un terme B est un terme dont le signifié désigne un ensemble comprenant le signifié de B (fémur *r_holo* membre inférieur). Pour certains connecteurs (*du* dans **lobe caudé du foie**), les deux relations sont correctes (lobe caudé *r_lieu* foie and lobe caudé *r_holo* foie). Nous pouvons noter aussi que nous utilisons la co-occurrence entre entités (Nom + Adj + Adj*) pour la détermination de la relation *caractéristique*. Par exemple, *carcinome hépatocellulaire multifocal* apparaît 5 fois, donc nous considérons *multifocal* comme une caractéristique de *carcinome hépatocellulaire* (carcinome hépatocellulaire *r_caracteristic* multifocal).

3.2.1 Contraintes sur les patrons

Dans notre première approche, nous avons sélectionné seulement 42 patrons sémantiques. Le choix de ces patrons est très contrôlé car pour une indexation des relations, le rappel ne doit pas trop diminué après l'ajout de contraintes. Pour certains patrons linguistiques, il est particulièrement difficile de déterminer précisément le type de relations car le connecteur est très général (par exemple les connecteurs *de* ou *sous* peuvent être associés à plusieurs types de relations). Pour surmonter ce problème, nous avons ajouté quelques contraintes d'ordre sémantique sur les patrons linguistiques. Ces contraintes sont exprimées à travers des règles. Nous en présentons quelques exemples ci-dessous : Le connecteur *du* peut donner plusieurs relations possibles comme par exemple *carcinome du foie* avec une relation de lieu entre *carcinome* et *foie* et une relation d'holonymie dans le groupe nominal suivant : *segment VII du foie* . De même le connecteur *avec* peut correspondre à plusieurs relations (par exemple relation de signe entre *luxation* avec *douleur* et une relation de cible (target) dans l'expression *nourrisson* avec *rougeole*. Ci-dessous, nous présentons des exemples de contraintes sur des connecteurs (en gras) représentant plusieurs relations possibles.

— $x **du** (connecteur) $y :
si $x r_isa lieu_anatomique & $y r_isa lieu_anatomique => $x r_holo $y
si $x r_isa maladie & $y r_isa lieu_anatomique => $x r_lieu $y

— $x **en** $y
si $x r_isa maladie & $y r_isa lieu_anatomique => $x r_lieu $y

— $x **avec** $y :
si $x r_isa maladie & $y r_isa signe_clinique => $x r_sign $y
si $x r_isa individu & $y r_isa maladie => $y r_cible $x

— $x **au niveau du** $y
si $x r_isa maladie & $y r_isa lieu_anatomique => $x r_lieu $y

— $x **due à** $y
si $x r_isa maladie & $y r_isa microorganisme || r_isa facteur_environnemental
=> $x r_cause $y

4 Résultats et discussion

Pour évaluer les performances de notre système, nous utilisons les mesures classiques, à savoir la précision (P), le rappel (R) et la F-mesure. A partir de notre corpus, nous avons extrait 120 000 relations. Environ 800 de ces relations sont vérifiées manuellement par un médecin et un spécialiste en imagerie médicale pour évaluer la précision. Afin d'évaluer le rappel, nous identifions les relations dans 300 comptes rendus et ensuite nous appliquons notre algorithme pour comparaison.

Globalement, la mesure de la précision s'améliore de façon non négligeable quand nous ajoutons des contraintes sur les relations sémantiques. Nous observons aussi une amélioration de la F1-mesure. Ceci s'explique par le fait que l'ajout de contraintes permet une meilleure caractérisation de la relation (par conséquence une amélioration de la précision) alors que le nombre de relations extraites ne varie pas (donc le rappel n'est pas modifié) du fait que nous n'ajoutons pas de patrons linguistiques. De plus le choix des contraintes (réalisé manuellement) a été réalisé de manière précise et contrôlé. Sinon, l'ajout de contraintes aurait entrainé une baisse du rappel. Certaines relations

type de relations	P	R	F1-mesure	P	R	F1-mesure
cause	74%	60%	66.3%	90%	60%	72%
conséquence	70%	62%	65.7%	89%	62%	73%
localisation	48%	40%	43.6%	83%	40%	54%
traitement	70%	60%	64.6%	88%	60%	71.3%
partie de	32%	30%	31%	75%	30%	42.9%
cible	45%	40%	42.4%	80%	40%	53.3%
caractéristique	60%	58%	58.9%	88%	58%	70%
lieu	45%	39%	41.7%	86%	39%	53.6%
holonymie	50%	50%	50%	65%	50%	56.5%

TABLE 3 – Résultats de l'extraction de relations sémantiques avec patrons linguistiques **sans** (à gauche) et **avec** (à droite) contraintes sémantiques

n'ont pas pu être extraites facilement car elles n'apparaissent pas dans notre corpus. Par exemple, la relation hyperonyme ou synonyme (d'où leur absence des résultats) est rarement présente dans nos comptes rendus ce qui semble normal parce que les radiologues connaissent déjà les informations taxonomiques comme par exemple *carcinome* est un *cancer*. Une autre cause de limitation de notre traitement concerne les relations impliquant des termes n'appartiennent pas à la même phrase. Nous avons aussi appliqué notre méthode sur d'autres types de corpus. Pour un corpus de 45 000 recettes de cuisine, nous avons extrait 245 000 relations avec une précision moyenne de 95%. L'évaluation a été réalisée manuellement sur un échantillon de 755 relations. De plus, nous avons extrait 789 000 relations à partir de pages Wikipedia avec une précision moyenne de 92% (évaluation manuelle sur un échantillon de 1250 relations). L'extraction de la relation d'hyperonymie à partir d'articles de Wikipedia a une précision de 94% (le rappel d'environ 48% bien que faible n'est pas véritablement un problème dans la mesure où l'on ne cherche pas à indexer les articles wikipedia, mais à en extraire la connaissance la plus précise possible).

5 Conclusion

Dans cet article, nous proposons une méthode pour extraire des relations sémantiques entre des entités dans des textes médicaux et plus particulièrement dans des comptes rendus de radiologie. Notre méthode est basée sur l'utilisation des patrons linguistiques avec contraintes sémantiques qui peuvent être vérifiées grâce au réseau lexical JDM. Nous avons montré que l'ajout de contraintes améliore de façon significative la précision, sans avoir recours à un analyseur syntaxique ou à un étiquetage morphosyntaxique. Comme perspective à court terme, nous envisageons la construction de nouveaux patrons de relations. Une piste future de notre travail est une meilleure couverture de notre système par la découverte/détection des patrons linguistiques de façon automatique (Meng & Morioka, 2015). Dans notre travail les contraintes ont été réalisées manuellement. Dans le futur nous souhaitons implémenter une méthode automatique ou semi-automatique de détection des contraintes. (Girju *et al.*, 2003) a proposé une méthode par apprentissage pour découvrir les contraintes appliquées sur les variables participant à une relation de méronymie. Cela permettra de créer de nouvelles contraintes afin d' améliorer la précision de notre système.

Références

ABACHA A. B. & ZWEIGENBAUM P. (2011). Automatic extraction of semantic relations between medical entities : a rule based approach. *J. Biomedical Semantics*, **2**(S-5), S4.

AUGER A. & BARRIÈRE C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, **14**(1), 1–19.

CHKLOVSKI T. & PANTEL P. (2004). Verbocean : Mining the web for fine-grained semantic verb relations. In *EMNLP*, volume 4, p. 33–40.

CHOWDHURY M. F. M. & LAVELLI A. (2012). Combining tree structures, flat features and patterns for biomedical relation extraction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 420–429 : Association for Computational Linguistics.

CIMINO J., BARNETT G. *et al.* (1993). Automatic knowledge acquisition from medline. *Methods of information in medicine*, **32**, 120–120.

DELBECQUE T., JACQUEMART P. & ZWEIGENBAUM P. (2005). Utilisation du réseau sémantique de l'umls pour la définition de types d'entités nommées médicales. In *CORIA*, p. 101–118.

EMBAREK M. & FERRET O. (2008). Learning patterns for building resources about semantic relations in the medical domain. In *LREC*.

FADER A., SODERLAND S. & ETZIONI O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1535–1545 : Association for Computational Linguistics.

GIRJU R., BADULESCU A. & MOLDOVAN D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, p. 1–8 : Association for Computational Linguistics.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 539–545 : Association for Computational Linguistics.

HINDLE D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, p. 268–275 : Association for Computational Linguistics.

HOGENBOOM F., IJNTEMA W. & FRASINCAR F. (2012). Text-based information extraction using lexico-semantic patterns. In *24th Benelux Conference on Artificial Intelligence (BNAIC 2012)*, p. 293–294.

JELIER R., JENSTER G., DORSSERS L. C., VAN DER EIJK C. C., VAN MULLIGEN E. M., MONS B. & KORS J. A. (2005). Co-occurrence based meta-analysis of scientific texts : retrieving biological relationships between genes. *Bioinformatics*, **21**(9), 2049–2058.

LAFOURCADE M. (2007). Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07 : 7th international symposium on natural language processing*, p.7̃.

LEE C.-H., KHOO C. S. & NA J.-C. (2004). Automatic identification of treatment relations for medical ontology learning : An exploratory study.

MENG F. & MORIOKA C. (2015). Automating the generation of lexical patterns for processing free text in clinical documents. *Journal of the American Medical Informatics Association*, p. ocv012.

NAZAR R., VIVALDI J. & WANNER L. (2012). Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del lenguaje natural*, **49**, 67–74.

RAMADIER L., ZARROUK M., LAFOURCADE M. & MICHEAU A. (2014). Spreading relation annotations in a lexical semantic network applied to radiology. In *Computational Linguistics and Intelligent Text Processing*, p. 40–51. Springer.

RINK B., HARABAGIU S. & ROBERTS K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, **18**(5), 594–600.

SNOW R., JURAFSKY D. & NG A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 801–808 : Association for Computational Linguistics.

SONG M., KIM W. C., LEE D., HEO G. E. & KANG K. Y. (2015). Pkde4j : Entity and relation extraction for public knowledge discovery. *Journal of biomedical informatics*, **57**, 320–332.

SUCHANEK F. M., IFRIM G. & WEIKUM G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 712–717 : ACM.

VON AHN L. (2006). Games with a purpose. *Computer*, **39**(6), 92–94.

ZARROUK M., LAFOURCADE M. & JOUBERT A. (2014). About inferences in a crowdsourced lexical-semantic network. *EACL 2014*, p. 174.