Annotations et inférences de relations dans un réseau lexico-sémantique: application à la radiologie

L Ramadier[1, 2] M Zarrouk[1] M Lafourcade[1] A Micheau[2]
(1) LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
(2) IMAIOS, 34090 Montpellier
lionel.ramadier@lirmm.fr,manel.zarrouk@lirmm.fr, mathieu.lafourcade@lirmm.fr
antoine.micheau@imaios.com

Résumé. Les ontologies spécifiques à un domaine ont une valeur inestimable malgré les nombreux défis liés à leur développement. Dans la plupart des cas, les bases de connaissances spécifiques à un domaine sont construites avec une portée limitée. En effet, elles ne prennent pas en compte les avantages qu'il pourrait y avoir à combiner une ontologie de spécialité à une ontologie générale. En outre, la plupart des ressources existantes manque de méta-informations sur les annotations (informations fréquentielles : de fréquent à rare ; ou des informations de pertinence : pertinent, non pertinent et inférable). Nous présentons dans cet article un réseau lexical dédié à la radiologie construit sur un réseau lexical généraliste (JeuxDeMots). Ce réseau combine poids et annotations sur des relations typées entre des termes et des concepts, un mécanisme d'inférence et de réconciliation dans le but d'améliorer la qualité et la couverture du réseau. Nous étendons ce mécanisme afin de prendre en compte non seulement les relations mais aussi les annotations. Nous décrivons la manière de laquelle les annotations améliorent le réseau en imposant de nouvelles contraintes spécialement celles basées sur la connaissance médicale. Nous présentons par la suite des résultats préliminaires.

Abstract. Relations annotation and inference in a lexical-semantic network : application to radiology Domain specific ontologies are invaluable despite many challenges associated with their development. In most cases, domain knowledge bases are built with very limited scope without considering the benefits of plunging domain knowledge to a general ontology. Furthermore, most existing resources lack meta-information about association strength (weights) and annotations (frequency information like frequent, rare ... or relevance information (pertinent or irrelevant)). In this paper, we are presenting a semantic resource for radiology built over an existing general semantic lexical network (JeuxDeMots). This network combines weight and annotations on typed relations between terms and concepts. Some inference mechanisms are applied to the network to improve its quality and coverage. We extend this mechanism to relation annotation. We describe how annotations are handled and how they improve the network by imposing new constraints especially those founded on medical knowledge. We present then some results

Mots-clés : réseau lexical, inférence, annotation, radiologie.

Keywords: relation inference, lexical semantic network, relation annotation, radiology.

1 Introduction

Depuis environ deux décennies, l'utilisation d'ontologies et de réseaux lexicaux dans le domaine biomédical est devenue très répandue (Bodenreider *et al.*, 2008). Ces ressources sont utilisées pour l'analyse sémantique comme pour la reconnaissance d'entités nommées (par exemple l'identification de noms des gènes) ou bien l'extraction de relations (identification des relations sémantiques entre entités biomédicales (Abacha et Zweigenbaum, 2011) comme pour le cas des interactions entre protéines). Dans le cadre du projet UMLS (Unified Medical Language System), un réseau sémantique a été construit (Lomax et McCray, 2004). Ce réseau est utilisé dans le domaine de la radiologie pour analyser de façon automatique les comptes rendus radiologiques afin d'extraire les recommandations en vue d'améliorer la prise en charge des patients (Yetisgen-Yildiz *et al.*, 2013). La plupart du temps, l'ontologie dédiée à la radiologie est plongée dans une ontologie médicale généraliste qui est trop importante et complexe pour l'utilisateur final. Pour tenter de résoudre ce problème, la société de radiologie Nord-Américaine (RSNA) a crée une ontologie spécifiquement dédiée à la radiologie Radlex (Rubin, 2008), (Mejino Jr *et al.*, 2008). Cependant, la couverture de RadLex n'est pas considérée comme complète (Hong *et al.*, 2012). Il existe une version allemande de RadLex (Gerstmair *et al.*, 2012) mais aucune en français à notre connaissance.

Cependant, un lexique médical unifié en langue française a été réalisé (Zweigenbaum *et al.*, 2005) mais il reste d'ordre général.

Dans le domaine de la radiologie où il peut être intéressant d'extraire des termes pertinents des comptes-rendus et les relier aux images (Napel *et al.*, 2010), les relations pertinentes entre les termes sont cruciales et les modèles taxonomiques ne capturent pas ces informations aussi bien qu'un réseau sémantique, la taxonomie indiquant seulement la hiérarchie entre les termes (relation is-a). Il peut être intéressant pour le médecin de disposer également plus facilement de relations non hiérarchiques. Par exemple, il est pertinent de donner pour une certaine maladie la liste des symptômes, des cibles potentielles, des localisations anatomiques et cela indépendamment de toute hiérarchie. Ceci peut être modélisé de façon plus simple par un réseau sémantique. L'association entre un réseau sémantique général et spécialisé peut jouer un rôle important dans l'analyse des comptes rendus radiologiques. En effet, dans la section *Indication* du rapport de radiologie, le texte est souvent écrit avec des termes courants alors que la section *résultats* comporte des termes très spécialisés. Le but de la construction d'un tel réseau est d'analyser les comptes rendus radiologiques dans leur totalité et d'en extraire les termes importants mais également les relations sémantiques pertinentes. Cette extraction pourra servir à annoter et indexer le texte et indirectement les images médicales afin de faciliter leur recherche et utilisation.

La construction d'un réseau lexico-sémantique peut être réalisée soit manuellement soit via une analyse de corpus. Par exemple, ConceptNet qui est une base de connaissance générale, est générée automatiquement à partir de 700 000 phrases du Open Mind Common Sense Project (Liu et Singh, 2004). Mais les approches entièrement automatisées sont généralement limitées à la co-occurence des termes car l'extraction des relations sémantiques précises entre termes à partir d'un texte reste difficile. Dans l'optique de la création d'un réseau spécialisé, nous avons décidé d'utiliser JeuxDeMots (JDM) (Lafourcade, 2007) comme base pour le réseau de connaissance générale. Le réseau JeuxDeMots est un réseau lexical construit à partir d'un ensemble de jeux en ligne. Pour la construction du réseau spécialisé, nous avons utilisé Diko un outil contributif proposé par la plateforme JeuxDeMots. La nécessité de ne pas dépendre uniquement de jeux pour construire un réseau lexical dédié à la radiologie vient du fait qu'une partie non négligeable des types de relations de JDM soit sont difficiles à saisir pour un joueur non expert, soit sont peu lexicalisées. Diko utilise par ailleurs des mécanismes d'inférences (Zarrouk *et al.*, 2013b) pour proposer automatiquement de nouvelles relations à partir de celles qui existaient déjà dans le réseau. Cette approche est strictement endogène et ne prend pas en compte des ressources externes.

JDM se fonde sur la peuplonomie (crowdsourcing) pour établir les poids des relations entre les termes. Chaque occurrence de relation est pondérée indiquant la force d'association (elle représente le nombre de joueurs qui ont pensé pour une relation donnée au même terme, à la même position parmi la liste des mots qu'ils ont proposés). Pour certains concepts ou termes, certaines idées viendront spontanément à l'esprit de beaucoup d'utilisateurs. La force d'association sera alors importante. Cette approche consistant à attribuer des poids à une relation est bien adaptée pour les connaissances générales et l'association de termes. Notons qu'il n'y a pas systématiquement une corrélation entre l'importance de la relation pour un domaine considéré et sa force d'association. Pour remédier à ce problème, nous introduisons des annotations entre certaines relations dans le réseau lexico-sémantique. Le but de ces annotations est de guider et d'améliorer le processus d'inférence et d'analyse sémantique.

Dans cet article, nous présentons les principes de construction du réseau lexical et nous l'illustrons grâce au projet JeuxDeMots. Nous discutons aussi de la construction d'un réseau spécialisé en imagerie médicale ainsi qu'un mécanisme d'inférence à savoir le schéma déductif. Ensuite nous détaillons le principe des annotations des relations entre termes médicaux. Dans une dernière partie nous décrivons nos expériences et les premiers résultats obtenus. Nous concluons avec les perspectives et les pistes futures pour la recherche.

2 Réseaux lexicaux et inférences

Un réseau lexico-sémantique est un graphe orienté, pondéré, typé avec des sommets qui représentent les concepts et des arcs les relations entre ces concepts. Il existe plusieurs méthodes pour construire un réseau lexical en tenant compte des facteurs principaux tels que la qualité des données, le coût et le temps de développement. Les approches contributives connaissent une forte popularité car elle se révèlent à la fois peu coûteuses et efficaces en qualité. L'intérêt porté au GWAP (games with purpose ou human-based computation game) comme méthode d'acquisition de ressources variées augmente régulièrement (Thaler *et al.*, 2011).

Le réseau JDM est un réseau lexico-sémantique construit à partir d'un ensemble de jeux en ligne. 6 790 189 relations et 316 983 termes sont présents dans la base. Pour le terme *médecine*, il existe 10 112 relations dans le réseau lexical. Environ 350 relations ont été créées pour le mot *IRM* (figure1).

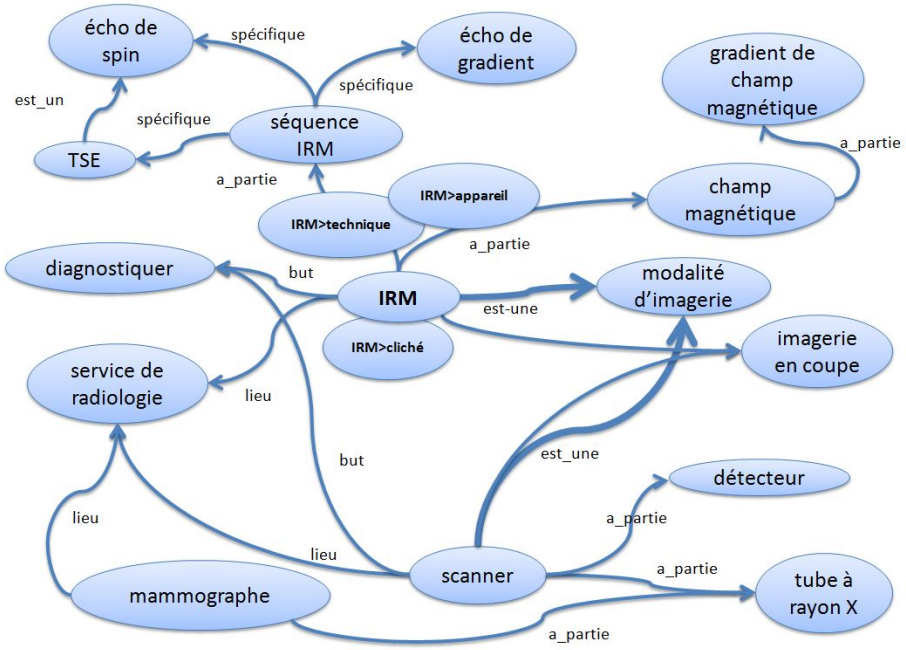

FIGURE 1 – Exemple de sous réseau lexical dédié à la radiologie/médecine : une partie des relations qu'entretient le terme *IRM*.

2.1 Le modèle JeuxDeMots

JeuxDeMots est un GWAP (Lafourcade, 2007) associant les joueurs par paires, et visant à construire un grand réseau lexico-sémantique. Il y a plus de 50 types de relations et chaque occurence de relation est pondérée indiquant une force d'association. Au début d'une partie, une consigne concernant le type de la relation (*idée associé*, *générique*, *caractéristique*, etc) ainsi qu'un terme cible issu du réseau lexical (par exemple donner des *idées associées* aux termes *maladie*) sont présentés au joueur. Ce dernier a un temps limité pour saisir des termes qui lui semblent correspondre à la consigne. Par la suite, ce même couple terme/consigne est proposé à d'autres joueurs. Les réponses communes par paires de joueurs sont insérées dans le réseau lexical ou le renforcent, c'est à dire dans le cas où la relation entre deux termes existe déjà, son poids est augmenté (force d'association). Le jeu est très efficace pour les connaissances générales mais l'est un peu moins pour les connaissances spécialisées. En effet, la majorité des joueurs n'ont pas de connaissance particulière dans le domaine de la radiologie. c'est pourquoi nous utilisons un autre outil fourni par JDM : Diko.

2.2 Diko : un outil contributif pour JDM

Diko est un outil en ligne sur le web permettant de visualiser les informations contenues dans le réseau lexical JDM, mais également constitue un outil contributif et de vérification. Dans le cadre de la construction d'un réseau dédié à un domaine spécifique, nous utilisons Diko comme un outil de développement d'une base de connaissance pour ce domaine. Le principe du processus de la contribution est qu'une proposition faite par un expert en radiologie sera soumise aux votes d'autres experts en imagerie médicale ou en médecine pour un processus de validation/invalidation. Dans le champ de la médecine, nous avons ajouté certains types de relations comme par exemple : *symptômes* ou *diagnostic*. Il nous paraît intéressant dans une base de connaissances dédiées à la radiologie de préciser pour une maladie donnée ses *symptômes* (cliniques), la population cible (*cible*), de même que les moyens de diagnostic (*diagnostic*). Cela peut avoir un intérêt pour l'expansion de requêtes dans l'analyse ou la recherche d'information (recherche d'image pour les patients présentant les symptômes de telle ou telle maladie). La construction d'un réseau spécialisé dans le domaine de l'imagerie médicale a été réalisée à partir d'un corpus de 40 000 compte rendus radiologiques représentant les différentes modalités d'imagerie médicale (imagerie par résonance magnétique, tomodensitométrie à rayon X, artériographie, échographie). La première étape a consisté à réaliser un index inversé de bigrame, trigrame à partir du corpus. Dans un deuxième temps, l'expert a soumis aux autres spécialistes du domaine les termes qu'il a jugés pertinents pour un processus de validation/invalidation. La majorité des concepts génériques a pu être rattachés aux concepts radiologiques(la relation *lieu*, *has part* etc..). Ce

travail contributif est nécessaire pour construire une base de connaissance liée à la radiologie.

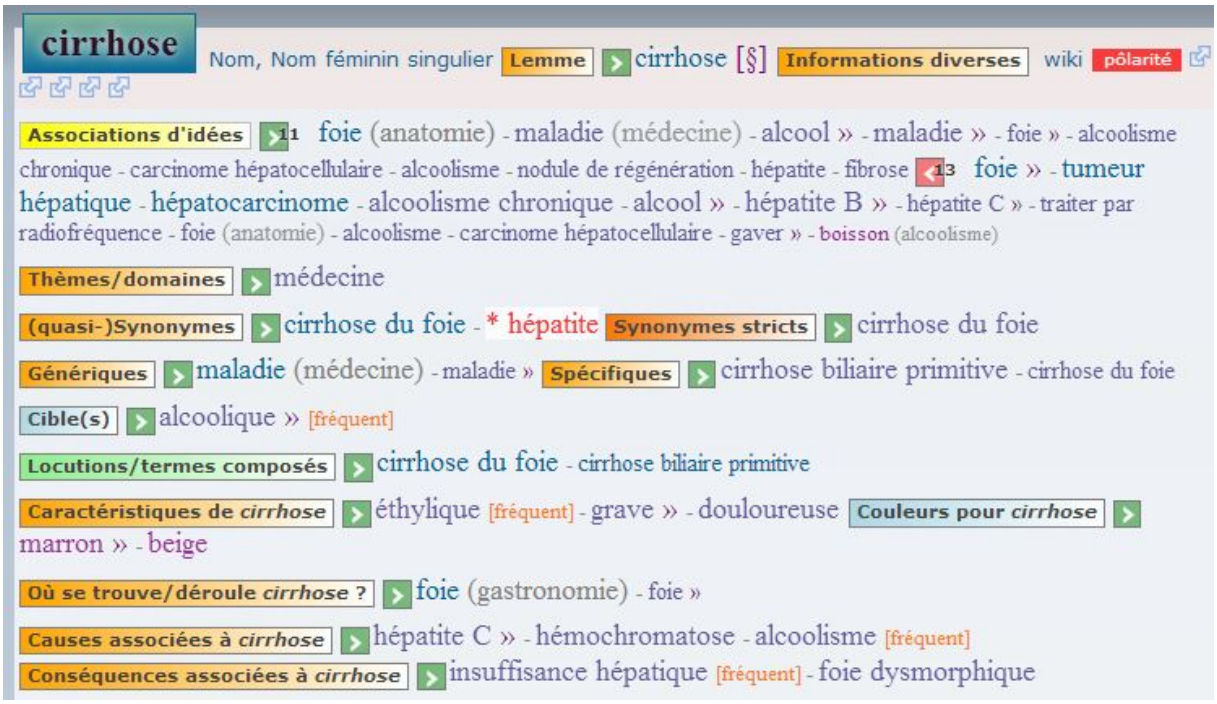

FIGURE 2 – Capture écran de la fenêtre de Diko du terme *cirrhose*, par exemple, *cause alcoolisme* annotée comme *fréquent*.

Dans le but d'améliorer la pertinence des informations sémantiques du réseau, nous ajoutons des annotations à certaines relations entre termes, en particulier pour ce qui nous concerne dans ce travail ceux liés à la médecine. Par exemple, pour la relation suivante *cirrhose* (cause) *alcoolisme* nous ajoutons l'annotation *fréquent* (figure 2). Nous donnons un autre exemple pour le terme *sclérose en plaques* (figure 3). Dans la troisième partie, nous détaillons le concept d'annotations de relations.

Dans le but de formuler de nouvelles conclusions (c'est à dire des relations entre les termes) à partir de prémisses (des relations préexistantes), un moteur d'inférence a été proposé (Zarrouk *et al.*, 2013b). Le moteur d'inférence propose des relations, à l'image d'un contributeur, qui vont être votées par la suite par un autre contributeur et validées par un expert dans le domaine de l'imagerie médicale. Dans le cadre de ce travail nous décrivons un seul type d'inférence : le schéma déductif. Le schéma déductif est basé sur la transitivité de la relation ontologique *is-a* (hyperonyme). Si un terme A est un type de B et B a une relation R avec le terme C, alors on peut proposer que A entretienne la même relation avec C. Le moteur d'inférence est appliqué sur les termes ayant au minimum un hyperonyme. Si un terme T possède un ensemble d'hyperonymes pondérés, le moteur d'inférence déduit un ensemble d'inférences. Ces hyperonymes vont être classés selon un ordre hiérarchique. Le poids d'une inférence proposée est la moyenne géométrique incrémentale de chaque occurrence (c'est à dire que la présence d'un poids négatif suffit à rendre la moyenne invalide). Le schéma présenté ci dessus est très simple, en effet le terme B peut être polysémique, et l'inférence proposée sera probablement fausse. Nous utilisons alors un blocage logique (figure 4). Ce mécanisme a été décrit dans un précédent travail (Zarrouk *et al.*, 2013b)

Dans le cas d'invalidation, un agent réconciliateur est invoqué pour essayer d'évaluer pourquoi la relation a été trouvé fausse : erreur dans les prémisses, polysémie (l'inférence est faite en se basant sur un terme central polysémique) ou une exception. Dans ce qui suit, c'est ce type d'inférence que nous allons considérer. Néanmoins, il existe deux autres types d'inférences : l'induction (du spécifique au général) et l'abduction (imitation par des exemples similaires)... .

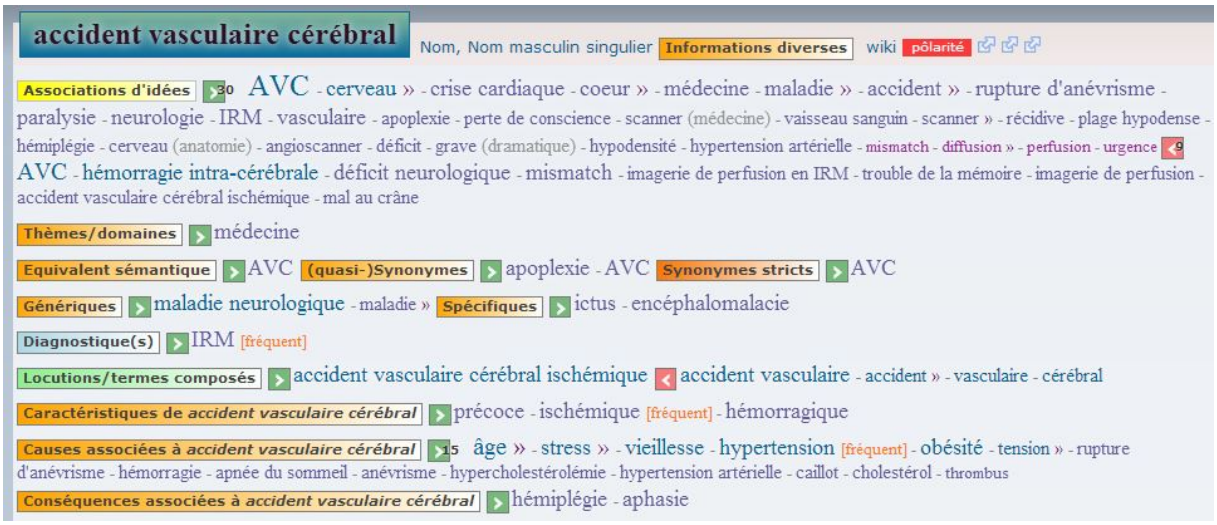

FIGURE 3 – Fenêtre de Diko du terme *accident vasculaire cérébral* dont la *caractéristique ischémique* est *fréquente*.

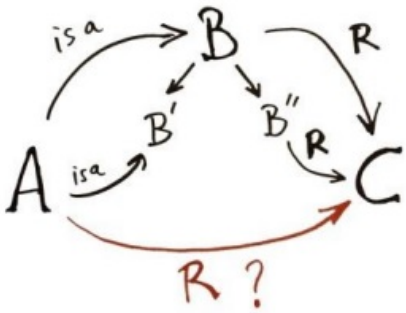

FIGURE 4 – Schéma d'inférence déductive triangulaire avec un blocage logique se basant sur la polysémie du terme B du milieu. Les termes B' et B" sont des raffinements/usages de B.

.

3 Annotation de relations

En général, et surtout dans le domaine des connaissances spécialisées, la corrélation entre la force d'association de la relation et son importance conceptuelle n'est pas toujours assurée. Par exemple, pour le terme *carcinome hépatocellulaire*, la relation *caractéristique* avec *wash-out* est très spécifique à la radiologie, par conséquence le poids de la relation sera faible dans le cadre général de la médecine mais pour le radiologue cette relation est particulièrement importante. Un autre cas est la relation *diagnostic* entre la *sclérose en plaques* et l'*IRM*. Nous avons affaire à une relation, là encore, spécifique au domaine de l'imagerie médicale, qui sera pertinente pour le radiologue. C'est pourquoi, il est apparu intéressant d'introduire des annotations pour certaines relations. Dans le réseau lexical, une relation est représentée par un triplet : $< noeud_{source}$, type de relation / annotation, $noeud_{cible} >$
Dans le champ de la radiologie, les relations les plus utiles pour le radiologue, qui ont été établies par des radiologues suivant leur pratique quotidienne, sont indiquées dans le tableau 1.Parmi les relations pertinentes, seules trois ont été rajoutées(*symptômes*, *diagnostique* et *cibles*), toutes les autres proviennent du domaine général. Dans les ontologies existantes dédiées à la radiologie comme RadLex, il n'existe pas autant de type de relations potentiellement utiles pour l'analyse des comptes rendus. Dans la recherche d'informations radiologiques (comptes-rendus et images), ces annotations peuvent apporter un complément d'information et permettre de classer les réponses par ordre de pertinence. Par exemple cela peut aider les radiologues devant une image anormale pour savoir si une caractéristique est *rare* ou *fréquente* et ainsi leur apporter une aide au diagnostic. D'autres types d'annotations peuvent exister comme par exemple la pertinence ou non d'une relation entre deux termes. Mais ce type d'information est généralement absent d'un réseau ou d'une ontologie. Par exemple, la relation *caractéristique* entre *carcinome hépatocellulaire* et *hypervasculaire* est *fréquent* et cette information

is-a	Hyperonymes du terme. Exemple : *IRM* est une *modalité d'imagerie* (possible)
partie-de	Parties, constituants, éléments du mot cible. Exemple : *foie* a comme partie *segment I* (toujours vrai)
caractéristique	Caractéristiques (adjectifs) possibles,typiques. Exemple : *carcinome hépatocellulaire* carac *hypervasculaire* (fréquent)
localisation	Lieux typiques où peut se trouver le terme/objet en question. Exemple : *sclérose en plaque* loc *système nerveux central*
cible	Population affecté par le terme. Exemple : *rougeole* cible *enfant* (fréquent)
diagnostic	Examen. Exemple : *sclérose en plaque* diag *IRM* (fréquent, crucial)
symptôme	Symptômes d'une maladie, *rougeole* symptôme *fièvre* (fréquent)
cause	B est une cause de A. Exemple : *cirrhose* cause *alcoolisme*
conséquence	B est une conséquence possible de A. Exemple : *accident vasculaire cérébral* peut avoir comme conséquence une *hémiplégie*

TABLE 1 – Relations pertinentes en radiologie pour l'analyse de compte-rendu

sera directement disponible dans le réseau (figure 5). Ces annotations ont par ailleurs, une fonction de filtre dans le schéma d'inférence.

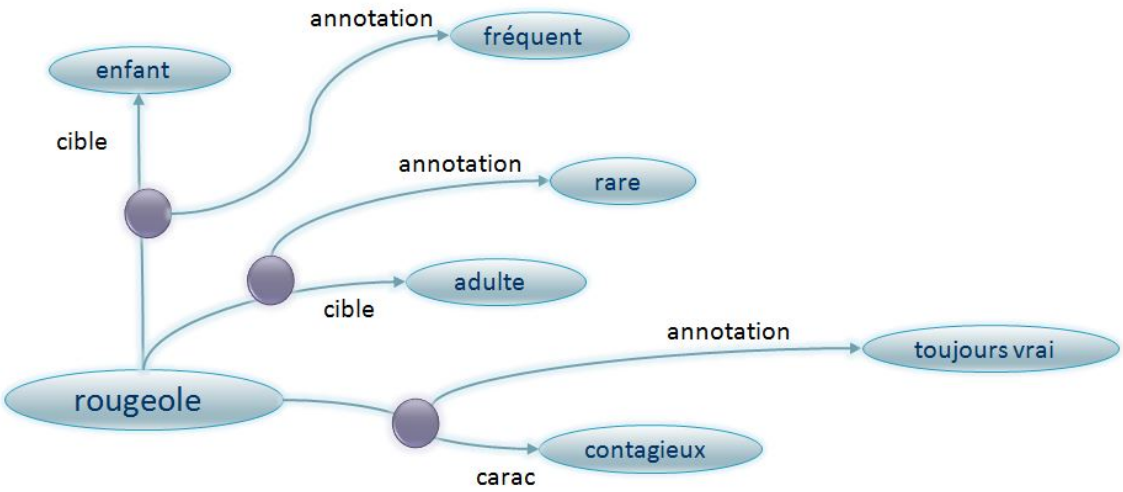

FIGURE 5 – Exemple d'implémentation d'annotation. L'implémentation d'une annotation se fait par réification de la relation à annoter dans le réseau lexical. Le noeud relation ainsi créé peut être associé à d'autres termes. La relation annotation n'est qu'un type de relation parmi d'autres. Les valeurs d'annotation sont des termes standard.

.

Les types d'annotations peuvent être de nature différente (information fréquentielle d'usage ou de pertinence). Ci-dessous, nous présentons les principaux types d'annotation :

- annotations fréquentielles : *très rare*, *rare*, *possible*, *fréquent*, *toujours vrai* ;
- annotations d'usage : *souvent crû vrai*, *abus de langage* ;
- annotations quantitatives : *un nombre (1, 2, 4, ...)*, *beaucoup*, *peu*, etc ;
- annotations d'exception : *exception* ;
- annotations qualitatives : *pertinent*, *non pertinent*, *inférable*.

Un médecin peut utiliser le terme *grippe* au lieu de *virus de la grippe* : c'est un **abus de langage**, le praticien fait simplement un raccourci de langage sans pour autant faire de confusion dans son esprit. Il semble évident pour lui que ces deux expressions sont différentes. L'annotation **souvent crû vrai** s'applique pour une fausse relation (avec un poids négatif) qui est souvent considérée comme vrai, par exemple *araignée* (*is-a/souvent crû vrai*) *insecte*. Les exceptions sont également renseignées et prennent la forme d'une relation ayant un poid négatif. Ce type d'annotation est utilisé pour

bloquer le schéma d'inférence.

L'annotation de nature qualitative est liée au statut inférable de la relation, particulièrement concernant l'inférence. L'annotation **pertinente** se rapporte à un niveau ontologique adéquat pour une relation donnée. Par exemple, *être vivant (carac/pertinent) vivant* ou *être vivant (carac/pertinent) mort*. L'annotation **inférable** est supposée être ajoutée quand une relation est inférable (ou a été inférée) à partir d'une relation existante, par exemple : *chien (carac/inférable) vivant* car *chien (is-a) être vivant*. L'annotation *non pertinent* est ajoutée aux relations vraies mais qui sont très éloignées du niveau pertinent, par exemple *animal (possède/non pertinent) atomes*. Pour avoir l'annotation la plus précise, nous avons besoin d'ordonner les termes centraux du plus spécifique au moins spécifique. Pour le terme *carcinome hépatocellulaire*, la hiérarchie sera :

carcinome hépatocellulaire
< tumeur maligne du foie < tumeur du foie < pathologie hépatique < pathologie

sclérose en plaques
< maladie du système nerveux central < neuropathie < maladie dégénérative < maladie

Pour choisir la bonne annotation de la nouvelle relation inférée, la hiérarchie ontologique joue un rôle important. L'annotation du terme le plus spécifique doit avoir plus d'influence que le moins spécifique. Nous prenons en compte ce fait pour les mécanismes d'inférences avec annotations.

Dans le mécanisme d'inférence, le terme B (le terme central) joue un rôle primordial. Nous inspectons la hiérarchie des termes B selon laquelle une relation spécifique a été inférée plusieurs fois et nous gardons la plus spécifique. Si nous obtenons deux termes ou plus ayant le même niveau sémantique, nous appliquons la règle du maximum aux valeurs correspondant à chaque annotation (toujours vrai : 5, fréquent : 4, possible : 3,rare : 2, très rare :1 ...) (figure 6).

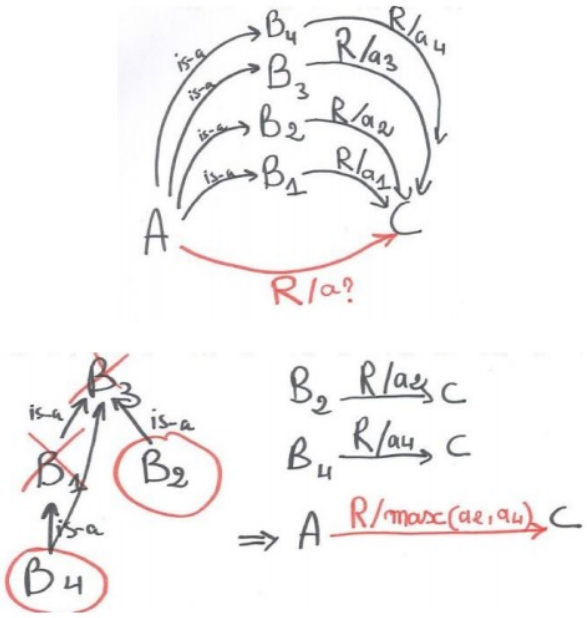

FIGURE 6 – Approche basée sur la hiérarchie utilisée pour choisir l'annotation la plus précise avec plusieurs termes centraux.

4 Expérimentation

Dans une précédente expérience menée par (Zarrouk *et al.*, 2013a), le moteur d'inférence déductive a été appliqué à l'ensemble du réseau lexical. Dans notre approche, nous avons lancé l'expérience sur une partie du réseau lexical JDM qui contient toutes les relations *is-a* (sur laquelle est fondé le schéma déductif) et toutes les relations annotées manuellement et ce dans le but de réduire l'espace de recherche.

4.1 Inférence des relations

Pour augmenter la précision des résultats et pour éviter d'inférer certaines relations peu pertinentes mais vraies (*homme a pour partie protons*), nous avons bloqué les inférences sur les relations qui avaient été annotées comme *non pertinent* ou *exceptionnel*. Le moteur d'inférence déductive a été appliqué sur **146 934** relations produisant un total de **1 825 933** relations avec **573 613** distinctes ce qui fait une moyenne de 3 occurrences par relation (table 2). Il est intéressant de constater que l'inférence renforce le niveau de confiance d'une relation déjà existante.

relations existantes	146934
relations inférées	1825933
relations inférées distinctes	573613

TABLE 2 – Nombre de relations inférées à partir de celles déjà existantes.

4.2 Propagation de l'annotation des relations

Le moteur d'inférence d'annotations est appliqué dans la seconde partie de notre système. Il permet d'ajouter des annotations aux relations de façon automatique à partir d'annotations de relations déjà existantes. Il est lancé sur la base des relations déjà enrichies avec le mécanisme d'inférence déductive. Contrairement au moteur d'inférence, nous autorisons la redondance en vue d'améliorer la précision des résultats du système de propagation d'annotation de relations. Prenons un exemple :

Prémisses : *accident vasculaire cérébral(is-a)infarctus cérébral*
& infarctus cérébral(diagnostic/fréquent) IRM
→ **relation inférée** : *accident vasculaire cérébral (diagnostic/possible) IRM* (1)

Prémisses : *accident vasculaire cérébral(is-a) maladie cérébrovasculaire*
& maladie cérébrovasculaire(diagnostic/possible) IRM
→ **relation inférée** : *accident vasculaire cérébral(diagnostic/possible) IRM* (2)

Le système d'annotations produit deux occurrences **(1)** et **(2)** de la même relation *accident vasculaire cérébral (diagnostic) IRM*, avec deux annotations différentes (possible, fréquent), nous décidons de garder celui avec la plus forte valeur (fréquent). Le système d'inférence des annotations appliqué sur la base de relations provenant des résultats du moteur d'inférence déductive, a annoté **10 085** relations à partir d'une amorce de seulement **72** relations annotées (table 3).

Type d'annotation	**annotation existante**	**annotation inférée**
Fréquentiel : toujours vrai	20	8092
Fréquentiel : fréquent	18	1
Fréquentiel : possible	16	150
Fréquentiel : rare et très rare	7	35
Qualitatif : souvent crû vrai	1	7
Qualitatif : non pertinent	5	1604
Quantificateur :	5	178
Total	72	10085

TABLE 3 – Nombre d'annotations inférées après application du système d'annotation des relations sur celles existantes.

Nous nous concentrons essentiellement sur les annotations concernant la fréquence car ces dernières comportent des informations importantes dans le domaine de la radiologie. Le nombre de relations annotées par type d'annotation ne dépend pas du nombre de relations existantes au départ mais simplement du nombre de relations d'hyperonymie existantes pour le terme central. Le schéma d'inférence est le suivant :

A(is-a)**B** et **B**(R/annot)**C**→**A**(R/annot)**C**

Par exemple :

$$\underbrace{\begin{array}{c}\text{cancer du poumon non à petites cellules}\\ \text{carcinome hépatocellulaire}\\ \text{glioblastome}\end{array}}_{\begin{array}{c}\text{(is-a) tumeur maligne}\\ \text{\& tumeur maligne (carac/frequent) mauvais pronostic}\end{array}} \quad (1)$$

Plus le nombre de relations d'hyperonymie vers le terme B (*tumeur maligne*) qui a une relation annotée (tumeur maligne(carac/fréquent)) est important, plus le nombre de relations annotées est élevé. Supposons que le terme *carcinome hépatocellulaire* n'ai pas de relation d'hyperonymie, donc dans ce cas l'annotation *fréquent* ne générera pas d'autre annotation. Ceci peut expliquer la raison pour laquelle il y a peu d'annotations inférées pour le type d'annotation fréquentiel *fréquent*. Notons que l'absence de certaines relations ou certains termes est due à l'aspect de progression continue du réseau qui fait qu'il est possible qu'à un instant précis un terme ou une relation manquent.

Nous avons évalué le nombre d'annotations inférées, et il apparaît que 87% d'entre elles ont été évaluées "correctes", 5% comme "incorrectes" et le reste (8%) comme "discutable" (les experts discuteront non pas leur validité mais plutôt leur valeurs fréquentielles pour savoir si elles doivent être modifiées). Dans cette expérience, nous avons appliqué le système relation/annotation une seule fois sur l'ensemble du réseau lexical. Évidemment, comme le réseau est en construction permanente, et que le partie consacrée à la radiologie n'en est qu'à ses débuts, de nouveaux termes ainsi que de nouvelles annotations seront rajoutées. Le système d'inférences et d'annotations tournent à présent en continu dans le but de consolider notre réseau lexico-sémantique.

5 Conclusion

Dans cet article, nous avons présenté quelques éléments pour la construction d'une base de connaissance spécialisée (en radiologie) dans un réseau lexical général et en particulier un modèle d'inférence et d'annotation de relations. Pour améliorer la qualité du réseau et sa couverture, nous avons proposé une approche de consolidation basée sur un moteur d'inférence réalisé sur des relations annotées. Le système d'annotation décrit dans cet article peut être vu comme un complément du système de consolidation de réseau lexico-sémantique. Ce système propage, grâce à la procédure d'annotation, des informations sémantiques ou d'usages importants qui peuvent être utilisées non seulement dans le domaine de la radiologie comme illustré dans cet article mais aussi dans d'autres domaines de spécialités. Il nous semble intéressant de développer des bases de connaissances dans des domaines spécialisés plongée dans un réseau lexical de sens commun. De futures recherches doivent également viser à améliorer la diffusion des annotations de relations à travers le réseau mais aussi améliorer le lexique spécialisé en radiologie à l'aide non pas seulement des experts mais aussi de non experts. Ce réseau lexico-sémantique nous sert pour les analyses sémantiques et d'indexation de comptes-rendus radiologiques. Cette analyse sémantique nous permet d'extraire des relations entre des concepts et des termes médicaux. Elle pourra être combinée avec la recherche d'image par le contenu (Content Based Image Retrieval ou CBIR) qui constitue une piste de recherche pour nos prochain travaux. En effet cette dernière technique pourra être combinée avec une recherche sémantique dans le but d'améliorer la recherche d'information dans le domaine de l'imagerie médicale.

Références

ABACHA, A. B. et ZWEIGENBAUM, P. (2011). A hybrid approach for the extraction of semantic relations from medline abstracts. *Computational Linguistics and Intelligent Text Processing*, pages 139–150.

BODENREIDER, O. *et al.* (2008). Biomedical ontologies in action : role in knowledge management, data integration and decision support. *Yearb Med Inform*, 47:67–79.

GERSTMAIR, A., DAUMKE, P., SIMON, K., LANGER, M. et KOTTER, E. (2012). Intelligent image retrieval based on radiology reports. *European radiology*, 22(12):2750–2758.

HONG, Y., ZHANG, J., HEILBRUN, M. E. et KAHN JR, C. E. (2012). Analysis of radlex coverage and term co-occurrence in radiology reporting templates. *Journal of Digital Imaging*, 25(1):56–62.

LAFOURCADE, M. (2007). Making people play for lexical acquisition with the jeuxdemots prototype. *SNLP'07 : 7th International Symposium on Natural Language Processing. Pattaya, Thailande*, page 8p.

LIU, H. et SINGH, P. (2004). Conceptnet - a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4): 211–226.

LOMAX, J. et MCCRAY, A. T. (2004). Mapping the gene ontology into the unified medical language system. *Comparative and functional genomics*, 5(4):354–361.

MEJINO JR, J. L., RUBIN, D. L. et BRINKLEY, J. F. (2008). Fma-radlex : An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. *AMIA Annual Symposium Proceedings*, 2008:465.

NAPEL, S. A., BEAULIEU, C. F., RODRIGUEZ, C., CUI, J., XU, J., GUPTA, A., KORENBLUM, D., GREENSPAN, H., MA, Y. et RUBIN, D. L. (2010). Automated retrieval of ct images of liver lesions on the basis of image similarity : method and preliminary results. *Radiology*, 256(1):243.

RUBIN, D. L. (2008). Creating and curating a terminology for radiology : ontology modeling and analysis. *Journal of digital imaging*, 21(4):355–362.

THALER, S., SIORPAES, K., SIMPERL, E. et HOFER, C. (2011). A survey on games for knowledge acquisition. *Rapport technique, STI*, page 26.

YETISGEN-YILDIZ, M., GUNN, M. L., XIA, F. et PAYNE, T. H. (2013). A text processing pipeline to extract recommendations from radiology reports. *Journal of biomedical informatics*, 46(2):354–362.

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013a). Inductive and deductive inferences in a crowdsourced lexical-semantic network. *9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, page 6p.

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013b). Inference and reconciliation in a lexical-semantic network. *14th International Conference on Intelligent Text Processing and Computational Linguistic (CICLING-2013)*, page 13p.

ZWEIGENBAUM, P., BAUD, R., BURGUN, A., NAMER, F., JARROUSSE, É., GRABAR, N., RUCH, P., LE DUFF, F., FORGET, J.-F., DOUYERE, M. *et al.* (2005). Umlf : a unified medical lexicon for french. *International Journal of Medical Informatics*, 74(2):119–124.