HAL
open science

Apprentissage de relations entre termes MedDRA dans UMLS pour la détection du signal en pharmacovigilance

Iulian Alecu, Cédric Bousquet, Marie-Christine Jaulent

To cite this version:

Iulian Alecu, Cédric Bousquet, Marie-Christine Jaulent. Apprentissage de relations entre termes MedDRA dans UMLS pour la détection du signal en pharmacovigilance. IC - 16èmes Journées francophones d'Ingénierie des Connaissances, May 2005, Nice, France. pp.13-24. hal-01023696

HAL Id: hal-01023696

https://inria.hal.science/hal-01023696v1

Submitted on 15 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Apprentissage de relations entre termes MedDRA dans UMLS pour la détection du signal en pharmacovigilance

Iulian ALECU[1], Cédric BOUSQUET[1], Marie-Christine JAULENT[1]

[1]INSERM U729, Paris, F-75006, France
{Iulian.Alecu,Cedric.Bousquet,
Marie-Christine.Jaulent}@spim.jussieu.fr

Résumé : Ce travail est motivé par la problématique de recherche d'informations spécifiques au domaine de la pharmacovigilance. Un signal de pharmacovigilance est la relation détectée statistiquement entre un médicament et un groupe d'effets indésirables exprimant des conditions cliniques similaires. Le regroupement pertinent des termes cliniques normalisés rend un signal plus spécifique et sa détection plus sensible. Notre objectif est de trouver les regroupements pertinents des effets indésirables en employant une méthode d'apprentissage automatique sur l'UMLS - un grand metathèsaurus de la médecine. A partir des relations pertinentes regroupant des termes cliniques normalisés et non définies explicitement dans MedDRA, le thesaurus d'origine, nous avons extrait des règles permettant de prédire ces relations à l'intérieur d'UMLS. En appliquant ces règles nous avons réussi à prédire les relations pour deux couples de concepts. Ces résultats démontrent l'intérêt de l'utilisation des méthodes d'apprentissage automatique sur l'UMLS pour l'extraction de relations non définies explicitement dans un système terminologique.

Mots-clés: Terminologies médicales; Appariement terminologique; Réseau sémantique; UMLS; MedDRA.

1 Introduction

Les différents points de vue selon lesquels on peut manipuler de l'information dans un domaine applicatif nécessitent des regroupements différents au sein d'une même organisation de l'information. Dans le cadre du travail présenté, nous nous sommes posés la question du regroupement d'information dans un contexte très particulier, rattaché au domaine médical, celui de la pharmacovigilance. Cette discipline cherche à mettre en évidence des relations statistiques entre des groupes d'effets indésirables (EI) observés et des médicaments, ce que l'on appelle un signal de pharmacovigilance. Une fois découverte, une telle relation peut mener à des études plus étendues afin d'établir la relation de causalité et donc l'imputabilité du médicament pour l'apparition des EI. L'objectif à long terme du projet est d'apporter une réponse à la détection du signal en s'appuyant sur le regroupement des

informations concernant les EI. Dans cet article nous nous intéressons essentiellement à la construction de l'organisation sous-jacente au regroupement des informations.

L'organisation de l'information pour un domaine d'application donné nécessite au moins deux étapes.

Une première étape de *normalisation terminologique* assure qu'une entité du domaine sera décrite au moyen des mêmes termes préférentiels, toutes les fois qu'elle sera indexée. Cette opération donne lieu à un vocabulaire contrôlé.

Ensuite, par un travail consensuel de *normalisation conceptuelle* du domaine (Charlet, 2003), des experts déterminent la position des termes du vocabulaire contrôlé dans une structure relationnelle (souvent hiérarchique). Le thesaurus résultant a l'avantage de regrouper l'information à des niveaux de granularité différents. L'inconvénient majeur d'un thesaurus est son manque de flexibilité. Ainsi, l'intégration d'un terme nouveau (ou l'exclusion d'un terme) dans le thésaurus n'est ni facile ni immédiate. Cette inflexibilité se retrouve aussi dans la structure relationnelle. Généralement, l'utilisation d'un thésaurus pour faire des regroupements d'information basés sur des critères autres que ceux proposés par les auteurs du thésaurus risque fortement de mener à des résultats insatisfaisants.

L'utilisation d'une ontologie formelle pour faire des regroupements d'information résout le problème de l'inflexibilité et s'intègre dans un effort commun pour la mise au point des nouvelles technologies dans l'organisation de l'information.

Dans le contexte spécifique des informations sur les effets indésirables, une première étude développée dans notre laboratoire a montré qu'une ontologie formelle des concepts désignant les EI permet d'effectuer des regroupements de cas pertinents et d'améliorer ainsi la détection de signaux (Henegar 2004). En même temps nous nous sommes retrouvés face à des difficultés importantes pour développer cette ontologie qui reste aujourd'hui très partielle. La méthodologie de construction adoptée, qui s'appuyait essentiellement sur les connaissances d'un seul expert, n'était pas adaptée à l'ampleur de la tâche. Nous souhaitons donc proposer une méthodologie originale qui réutilise des connaissances déjà acquises dans le domaine, comme par exemple les connaissances déjà présentes dans les thésaurus médicaux, et en particulier le métathésaurus UMLS (Unified Medical Language System).

L'hypothèse qui sous-tend notre méthodologie consiste à dire qu'il existe à l'intérieur d'UMLS des connaissances implicites qui permettraient de suggérer automatiquement des relations de subsomption entre concepts lors de la construction d'une ontologie dans le domaine médical.

La méthodologie adoptée et présentée dans cet article s'apparente à une méthode d'apprentissage. Grâce à l'ontologie amorcée précédemment, nous avons un certain nombre de couples de concepts qui sont directement liés par une relation de subsomption. C'est notre base d'exemples. La relation de subsomption a été créée à partir des définitions formelles fournies par l'expert et n'existe pas nécessairement dans UMLS. Nous cherchons donc à apprendre si parmi les différentes façons de parcourir UMLS d'un concept à un autre, il y en a certaines qui peuvent suggérer une relation de subsomption en se basant sur les exemples connus. La méthode présentée explicite dans un premier temps le formalisme des exemples qui vont être utilisés lors de l'apprentissage. Ensuite, la méthode d'apprentissage elle-même est fournie. Elle

s'appuie sur une méthode statistique classique (l'odds ratio). Nous présentons ensuite, une étape de vérification sur deux couples exemples qui permet de conforter l'intérêt de l'approche. Notre contribution est aujourd'hui dépendante de la validation dans notre contexte d'application. Néanmoins, des briques sont posées pour une stratégie d'extraction des connaissances à partir d'UMLS pouvant avoir des retombées dans bien d'autres secteurs médicaux que la pharmacovigilance.

2 Contexte et état de l'art

2.1 Les méthodologies de construction des ontologies

Une ontologie est la spécification explicite et formelle d'une conceptualisation partagée, en vue de la réalisation d'une tâche (Bachimont, 2000). Le processus général de modélisation est complexe, mais il s'agit d'abord d'identifier les concepts utilisés dans le domaine pour la tâche, leurs propriétés et les relations qu'ils entretiennent. Plusieurs approches pour la construction d'ontologies ont été publiées (Aussenac 2000). On distingue en général les méthodes descendantes, orientées sur la tâche à réaliser en fonction de laquelle les connaissances du domaine sont choisies et les méthodes ascendantes qui utilise le sens des mots pour organiser les connaissances. Dans le domaine médical, des expériences sont menées qui privilégient les approches ascendantes en donnant beaucoup de place aux rôles respectifs de l'expert et des textes dans la conception de l'ontologie par le choix des primitives (leMoigno 2002). Ces travaux s'appuient sur le fait qu'il existe des sources textuelles importantes comme les comptes rendus médicaux. Par contre, en pharmacovigilance, il n'existe pas de textes reflétant une pratique de la discipline. Les sources textuelles qui sont à notre disposition sont les thésaurus médicaux.

2.2 Les thésaurus médicaux

Les thesaurus en médecine sont utilisés pour enregistrer et d'échanger des informations et des connaissances médicales sous une forme normalisée (Cimino 1996, Zweigenbaum 1999). Ces thésaurus sont construits toujours dans un but précis. Ainsi, le thesaurus MeSH est employé pour l'indexation des connaissances médicales, en particulier les articles scientifiques (NLM 2001). D'autres ont pour vocation l'établissement de statistiques, par exemple de mortalité et de morbidité avec la classification internationale des maladies CIM-10 (OMS 1993). La nomenclature SNOMED Internationale (Côté 1998) est utilisée pour enregistrer des informations cliniques détaillées. S'ils ne sont pas dédiés à cela, ces thésaurus contiennent des informations relatives aux EI.

MedDRA est le thésaurus actuellement utilisé pour la normalisation des EI en pharmacovigilance. Les études sur la pertinence de regroupements des termes MedDRA désignant des EI ont mis en évidence des problèmes portant spécialement sur la pauvreté relationnelle et le positionnement des termes qui présentent parfois des granularités différentes sur un même niveau hiérarchique. La structure relationnelle

dans laquelle ces termes sont organisés est d'autant plus importante qu'elle a un impact direct sur la spécificité et la sensibilité des signaux de pharmacovigilance détectés (Yokotsuka 2000 ; Brown, 2002). Nous avons montré dans un article récent que MedDRA ne permet pas le regroupement pertinent d'information pour la détection du signal et que ceci est en partie lié au manque de représentation formelle des termes (Bousquet 2005).

Devant ces constats, nous nous sommes plus particulièrement intéressés à l'Unified Medical Language System® (UMLS) de la National Library of Medicine (NLM) qui présente un statut particulier par rapport à ces thésaurus (Lindberg 1993).

UMLS[1]

L'objectif déclaré de l'UMLS est de mettre ensemble et de rendre interrelationnels la majorité des standards terminologiques utilisés actuellement par les divers domaines de la santé afin d'obtenir une terminologie complète de la médecine. La construction d'UMLS, ne vise pas un domaine spécifique et reste la plus générale possible, laissant aux utilisateurs la possibilité de l'adapter conformément à leurs besoins. L'effort de la NLM se concentre seulement dans l'intégration des terminologies au sein d'une base de connaissances sans apporter de modifications à leurs composants linguistiques ou structurels.

La couverture terminologique du domaine médical offerte par UMLS est impressionnante (60 familles de terminologies médicales, 2 millions de termes et 900 mille concepts, 12 millions de relations entre les concepts), MedDRA y étant incluse. Les thesaurus inclus dans l'UMLS ont été crées dans des contextes sensiblement différents, ayant comme point commun le domaine de la médecine.

Les spécificités à l'intérieur des différents thesaurus présentent deux aspects :

- un aspect *structurel*. On trouve des thesaurus plutôt *profonds* (AOD le thésaurus « Alcool and Other Drugs » - 10 niveaux et environ 2 000 termes par niveau), ainsi que des thesaurus *larges* (MedDRA - 5 niveaux hiérarchiques et environ 15 000 termes par niveau).
- un aspect qualitatif ou *sémantique*, lié aux critères qui ont mené au positionnement taxinomique des termes les uns par rapport aux autres. Par exemple pour un même terme, dans certains thesaurus l'aspect sémiologique du terme est le plus important, dans les autres c'est l'aspect lésionnel. Autrement dit, les relations taxinomiques n'ont pas le même sens d'un thésaurus à l'autre.

Les termes synonymes provenant de plusieurs thésaurus sont associés dans UMLS à un seul concept. De plus, les relations taxinomiques originaires des thesaurus sont conservées entre les concepts. UMLS est donc un réseau de concepts comprenant des relations de subsomption (Petiot 1996).

En conclusion, UMLS par l'entrelacement des thesaurus, facilite beaucoup la comparaison des concepts entre eux ainsi que leurs façons d'être regroupés.

[1] Nous avons utilisé sa version 2003AC (http://umlsks.nlm.nih.gov).

Etant donnée sa forme de réseau, on peut trouver dans UMLS plusieurs chemins entre deux concepts qui peuvent être composés d'un nombre d'arcs différent (Bodenreider, 2003). Tous les arcs ne traduisent pas le même degré de généralisation/spécialisation. Ainsi, un chemin court d'un arc entre deux concepts peut correspondre à un passage « général – spécifique » rapide. En contraignant les trajets à avoir un grand nombre d'arc, on traverse de façon préférentielle des concepts issus de thésaurus présentant une plus grande profondeur avec une organisation conceptuelle du domaine plus précise.

Par ailleurs, UMLS propose un typage sémantique des concepts au sein du réseau sémantique. Ce typage réalise un regroupement très large et général des concepts (e.g. type « syndrome »). Il y a actuellement 135 types sémantiques pour une totalité d'1 million de concepts dans le réseau UMLS.

3 Matériel

La *base de connaissance* UMLS est utilisée dans sa version 2003AC mise à disposition en ligne et en tant que web service par la NLM.

Une *base d'exemples* est constituée. Elle est composée de 190 couples de concepts. Pour un couple donné, les concepts sont liés par la relation « is-a » dans l'ontologie des EI que nous avons à notre disposition mais ne sont pas en relation directe dans MedDRA. L'exemple que nous allons présenter lors des différentes étapes composant la méthode est celui du couple « Purpura » – is_a – « Bleeding Diathesis ».

L'apprentissage se fait sur 188 couples sélectionnés aléatoirement. Les deux couples restants sont utilisés pour vérifier les résultats obtenus à partir de la base d'apprentissage. Une méthode de validation de type « bagging » est envisagée dans le futur qui s'appuiera sur une série de tirages aléatoires de 188 couples. Dans le cadre du travail présenté, un seul tirage a été réalisé.

4 Méthode

Nous appelons « *trajet* » l'information contenue dans un chemin existant entre deux concepts (nœuds) connexes dans l'UMLS (incluant : concepts intermédiaires, informations liées aux relations). L'objectif de la méthode mise au point consiste à trouver des règles d'association pour chaque couple de la base d'apprentissage entre les trajets UMLS existants pour ce couple et la relation « is_a » dans l'ontologie des *EI*. L'idée sous-jacente à cette méthode est de prédire la relation de subsomption à partir des trajets existants entre deux concepts MedDRA dans l'UMLS.

La méthodologie élaborée pour mettre en œuvre et évaluer cette méthode a suivi trois étapes distinctes :

1) Acquisition de tous les trajets UMLS pour chaque couple de la base d'apprentissage.

2) Extraction de modèles de trajets définissant les règles d'association.

3) Validation des résultats ainsi obtenus. Dans l'état d'avancement actuel du travail, cette étape consiste à vérifier que les règles d'association appliquées sur un concept exemple permet de suggérer des concepts candidats à être liés au concept exemple par une relation de subsumption.

4.1 Formalisme pour la représentation des relations et trajets

Le modèle formel de représentation des relations de l'ontologie est un couple du type :

- C_F – R – C_P où R est la relation « is_a_ontologique », C_F le *concept fils* et C_P le *concept père.*

Dans l'UMLS ce couple va prendre la forme $C_F – T_{UMLS} – C_P$, où :

- $T_{UMLS} = \{t_{UMLS} \mid C_F – t_{UMLS} – C_P\}$ est l'ensemble des trajets qui mettent en relation C_F et C_P.
- et $t_{UMLS} = R_{UMLS_1} – C_1 – R_{UMLS_2} – C_2 –…– C_{n-1}– R_{UMLS_n}$, un trajet en UMLS composé de relations et concepts UMLS (R_{UMLS} , C) ;

Les trajets sont décrits par les propriétés suivantes :

- pour les concepts
  - le nom (« arrhythmia », « stomach ulcer »…),
  - le type sémantique attribué à l'intérieur d'UMLS («disease », « finding »).
- pour la relation R_{UMLS} , que nous allons appeler « *arc* »
  - le type. Ce type a été attribué lors de la construction du metathesaurus. Il existe 11 types définis selon plusieurs critères. Ils indiquent la direction pour les relations purement hiérarchiques : CHD (child), PAR (parent), RN (narrower), RB (broader). Les autres types indiquent des relations associatives (part-of, caused-by, …)
  - le nom du thesaurus d'origine.

4.2 Acquisition des données sur les trajets

Dans cette section nous présentons la méthodologie employée pour l'acquisition des trajets, les choix que nous avons fait concernant les informations pertinentes à retenir ainsi que les raisons qui nous ont amené à faire ces choix. Pour l'acquisition des trajets, nous avons implémenté un algorithme classique d'exploration en profondeur de graphes en éliminant les cycles (Bodenreider, 2001).

Les tests que nous avons faits pour la connectivité des concepts au sein de l'UMLS ont mis en avant que tous les types de relations ne sont pas susceptibles de nous aider. Ainsi, nous avons trouvé plusieurs centaines de trajets en deux arcs entre les concepts « Fracture » et « Hépatite ». Cette proximité purement topologique dans le réseau est trompeuse car ces deux concepts ont des sens éloignés. Sur la base de ce constat, nous

avons choisi de considérer exclusivement les relations de type hiérarchique incluses en UMLS. Ces relations sont typées comme : CHD, PAR, RN et RB.

Afin de mieux déterminer le domaine de couverture des concepts pris en compte nous avons restreint la liste des thesaurus source en ne gardant que ceux qui sont en anglais et dont le domaine d'application est la médicine clinique. Une quinzaine de thésaurus a été ainsi écartée.

Le choix du nombre des relations composant un trajet a été extrêmement difficile. Le constat qu'un trajet plus long passerait par des concepts intermédiaires plus proches sémantiquement plaide en faveur d'un choix des trajets longs. En effet, en empruntant un trajet long, on choisit systématiquement les relations appartenant à des thesaurus où la structure est plus profonde et en conséquence plus riche du point de vue relationnel. Dans une telle structure il est évident que les différences sémantiques entre les concepts subsumés et subsumants sont plus fines.

En ce qui concerne la variabilité en longueur des trajets, le traitement statistique n'est pas pertinent pour des trajets de longueurs différentes. L'acquisition de trajets de taille variable nous aurait obligé à créer des ensembles de trajets de même longueur pour pouvoir soumettre chaque ensemble au même traitement statistique.

Suite à ces constats, nous avons décidé qu'un nombre de 3 relations pour chaque trajet est un compromis avantageux entre la longueur et les ressources disponibles.

A la fin de cette étape nous obtenons pour chaque couple $C_F - R - C_P$ un nombre de trajets du type $C_F - t_{UMLS} - C_P$ qui vérifient les conditions :

- t_{UMLS} est composé uniquement de 3 relations R_{UMLS} *(1)*
- les R_{UMLS} sont d'un type inclus dans l'ensemble {CHD, PAR, RN, RB} et ont comme origine un des thesaurus que nous avons conservés. *(2)*

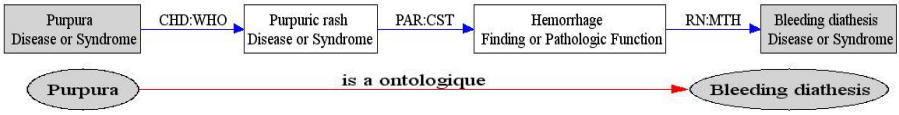

Fig. 1 : Exemple d'un trajet (en haut) pour le couple ontologique « Purpura » – is_a – « Bleeding Diathesis ». WHO, CST, MTH sont des acronymes représentant le thesaurus d'origine de la relation..

4.3 Extraction des modèles de trajets

L'objectif principal de cette étape est de trouver des règles d'association entre la relation « is_a_ontologique », caractérisée par le couple de concepts d'ancrage C_F, C_P et les trajets acquis. Nous définissons dans un premier temps, une classification des trajets. Les classes de trajets sont formées en considérant les types sémantiques pour les nœuds et les types de relation pour les arcs. Pour chaque couple de la base, les trajets sont organisés par groupes sémantiques selon cette classification (Figure 2).

Nous posons comme modèle de couple $C=(T_F, T_P)$, où T_F, T_P sont les types sémantiques des concepts C_F, C_P , et le modèle de trajet $M = R_{UMLS1} - T_1 - R_{UMLS2} - T_2 - R_{UMLS3}$, où R_{UMLS} est un type de relation et T le type sémantique du concept intermédiaire. (Figure 2)

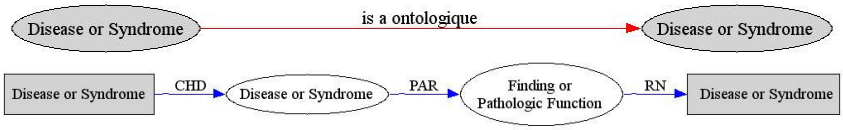

Fig. 2 : Le modèle de trajet de l'exemple du trajet présenté dans la figure 1.

Après la création de modèles de trajet nous avons obtenu comme résultat un tableau qui comprenait pour chaque modèle de couple C le nombre de trajets du modèle M qui ont été trouvés. Pour l'exemple montré en figure 2, 4 trajets du modèle présenté (figure 2 en bas) ont été trouvés pour l'ensemble des couples appartenant au modèle de couple présenté (figure 2 en haut).

Ensuite nous avons calculé le rapport des cotes (*odds ratio*-OR) pour le nombre d'occurrences de chaque modèle de trajet M, afin d'exprimer de façon synthétique la spécificité statistique de chaque association modèle de trajet – modèle de couple (C – M) . Pour chacune des associations on construit le tableau de contingence dans lequel :

Tableau 1 : Tableau de contingence 2x2.

	M_j	$\neg M_j$
C_i	*a*	*b*
$\neg C_i$	*c*	*d*

- *a* est le nombre de trajets des couples du type C_i avec le modèle M_j
- *b* est le nombre de trajets des couples du type C_i avec un modèle autre que M_j
- *c* est le nombre de trajets basés sur le modèle M_j mais qui ne décrivent pas les couples du type C_i
- *d* est le nombre de trajets qui ne sont pas basés sur le modèle M_j et qui ne décrivent pas les couples du type C_i

L'odds ratio est le rapport entre (*a/c*) et (*b/d*).

Nous avons choisi pour chaque C toutes les M pour lesquels OR>1 (Li 2003). Nous avons obtenu ainsi un ensemble de règles d'association $C \rightarrow \{(M_1, OR_1); (M_2, OR_2); \ldots ; (M_n, OR_n)\}$ pour chaque modèle de couple.

4.4 Vérification

Nous considérons les deux couples de concepts choisis aléatoirement pour la vérification. Pour un couple donné, nous avons identifié le modèle auquel ce couple (C) appartient, c'est-à-dire les types sémantiques de C_F et C_P. Ensuite, nous avons retenu les modèles de trajets (M) qui ont été associés à ce modèle au cours de l'apprentissage, c'est-à-dire les règles d'apprentissage

Nous avons exploré l'UMLS à partir du C_F en respectant les modèles de trajets définis dans la règle d'association. Nous avons retenu tous les concepts qui se trouvent au bout de chaque trajet qui respecte un des modèles. Nous avons appelé ces concepts, *concepts présumés subsumant* (C_{PS}).

Soit MT l'ensemble des modèles de trajets pour lesquels on trouve au moins un trajet dans le cas d'un couple utilisé pour la vérification. Pour mettre en évidence la pertinence des C_{PS} trouvés nous avons mis au point un système de scores. Ainsi nous avons considéré comme score maximum la somme des OR des éléments de MT. Un score individuel est défini pour chaque C_{PS} comme la somme des OR des modèles de trajets (M) pour lesquelles nous avons trouvé au moins un trajet qui menait à C_{PS}. Ces scores individuels ont été exprimés comme un pourcentage du score maximum que nous avons appelé *niveau de pertinence* du C_{PS}. Une forte valeur du niveau de pertinence signifie que le C_{PS} a été trouvé en suivant des trajets correspondant à des T fortement associés au C modélisant (C_F, C_P).

5 Résultats

5.1 Base de trajets

19234 trajets ont été recueillis pour 128 (68%) couples sur 188. Dans 60 couples (31%) aucun trajet en 3 arcs n'a été trouvé. Le nombre moyen de trajets par couple était de 205.

Dans le tableau suivant nous présentons quelques exemples de trajets entre les concepts « Purpura » (C_F) et « Diathèse hémorragique » (C_P). Pour ce couple on trouve 57 trajets dans l'UMLS.

Tableau 2 : Exemples de trajets entre « Purpura » et « Diathèse hémorragique ».

R_{UMLS1}		C_1		R_{UMLS2}		C_2		R_{UMLS3}
CHD CSP	-	Purpura, Thrombocytopenic *Disease or Syndrome*	-	PAR CST	-	Hemorrhage *Finding|Pathologic Function*	-	RN MTH
CHD SNM	-	Purpura Fulminans *Finding|Disease or Syndrome*	-	PAR CST	-	Hemorrhage *Finding|Pathologic Function*	-	RN MTH
CHD WHO	-	Purpuric rash *Disease or Syndrome*	-	PAR CST	-	Hemorrhage *Finding|PathologicFunction*	-	RN MTH
PAR CST	-	Disease of capillaries *Disease or Syndrome*	-	CHD CST	-	Ecchymosis *Finding|Pathologic Function*	-	PAR MDR

5.2 Traitement statistique

Les 19234 trajets ont produit 488 modèles de trajets. Dans 443 (91%) des modèles de trajets, le nombre de trajets correspondants était inférieur à 100. Les 45 modèles de trajets restants généralisent 12670 trajets, soit 65,8% du total des trajets. Les modèles de couples sont au nombre de 22. Pour 4 modèles de couples on a trouvé un nombre de trajets de 15674, soit 81% du nombre total des trajets. 10 modèles de couples n'ont

aucun modèle de trajet correspondant qui ait un `OR` supérieur à 1. Le nombre moyen des modèles de trajet avec un `OR` supérieur à 1 par modèle de couple est de 9,8.

5.3 Vérification

Nous avons réalisé une vérification sur deux couples aléatoirement choisis parmi les 190 de la base d'exemple. Le premier couple a les caractéristiques suivantes :

- (C_F, C_P) = (« Tachycardia, Paroxysmal », « Tachycardia »)
- le modèle de couple `C` = (« Disease or syndrome », « Finding »)

En explorant l'ensemble de ces modèles de trajets on part du C_F pour aboutir à 43 C_{PS}. La figure 3 montre le niveau de pertinence de chaque C_{PS}.

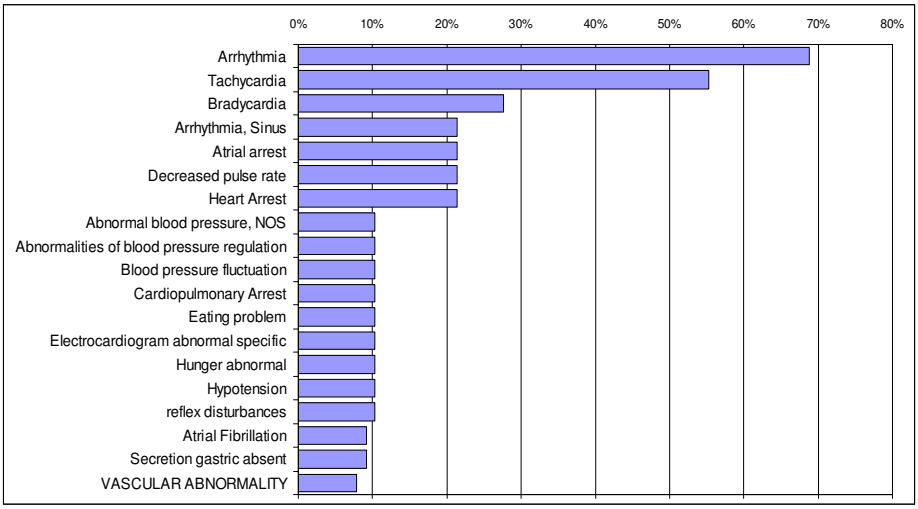

Fig. 3 : Les scores des C_{PS} trouvés pour le concept de validation « Tachycardia, Paroxysmal ». Un nombre de 24 C_{PS} ayant un score de pertinence inferieur à 8% n'ont pas été représentés dans cette figure.

Notre méthode met en évidence une différence nette entre les deux premiers concepts classés et les autres. Le concept que nous recherchions est classé deuxième dans cette liste (« tachycardia »). Le concept « arrhythmia », classé en premier, est le concept qui subsume « tachycardia » dans l'ontologie des effets indésirables. Pour l'autre couple utilisé pour la vérification (« Tachycardia, Supraventricular », « tachycardia ») nous avons obtenu des résultats similaires. Entre les plus pertinents (16) C_{PS} trouvés 5 sont des arythmies et 8 des affections cardiovasculaires entraînant une arythmie.

6 Discussion et conclusion

Plusieurs études se sont attachées à identifier dans l'UMLS des relations en un arc entre des concepts ciblés. Le taux partiel de couverture obtenu grâce à ces relations nous a encouragé à développer une méthode capable d'identifier des relations comportant un nombre d'arcs supérieur à un. Nous avons proposé dans ce travail une méthode originale pour l'apprentissage de nouvelles relations de subsomption dans une classification basée sur une hiérarchie stricte. Notre cible, la terminologie MedDRA, devrait bénéficier d'une telle approche pour retrouver les liens hiérarchiques permettant de regrouper ses concepts de façon pertinente pour la détection du signal. Nous avons identifié des modèles de trajets entre concepts UMLS qui sont spécifiques pour une relation de type *is_a* ontologique. Cette étude montre l'intérêt de l'UMLS pour l'extraction de connaissance et l'aide à la construction d'une ontologie.

Bien que cette méthode soit en grande partie automatisable, l'intervention d'un expert est nécessaire afin de valider les concepts proposés subsumants proposés par le système et éventuellement choisir un ou plusieurs couples parmi les couples qui présentent les niveaux de pertinence les plus élevés.

Une validation plus ample (encore embryonnaire dans le travail accompli) ainsi que l'affinage des paramètres de la méthode d'apprentissage rendront certainement la validation experte moins importante pour le résultat final.

En ce qui concerne la description des trajets nous avons limité cette première étude aux types attachés aux relations taxinomiques (PAR, CHD, etc.). Des informations supplémentaires disponibles dans l'UMLS doivent être prise en compte, comme par exemple les relations associatives issues de la SNOMED CT ou autres. De la même façon la méthodologie d'apprentissage des règles d'association peut encore être optimisée. La prise en compte du facteur de Jaccard, du cosinus, du Laplace et d'autres mesures statistiques, comme des indicateurs de pertinence des règles trouvées seraient des optimisations envisageables (Li 2003).

Nous avons choisi d'une façon expérimentale des trajets en 3 arcs. Des comparaisons de résultats obtenus avec des trajets d'autres longueurs viendraient compléter cette étude et préciser son intérêt.

Dans la même direction l'affinage de l'apprentissage pourra porter sur l'utilisation de regroupements proposés par l'UMLS en s'appuyant sur des relations existantes entre les types sémantiques. De la même façon, une relation qui relie les mêmes concepts dans plusieurs terminologies peut être considérée plus pertinente.

A notre connaissance, il n'existe aucune méthode disponible, qui pour un nombre de concepts et un domaine d'interprétation établis, soit capable de filtrer les relations tout en conservant la sémantique que l'on souhaite leur donner. La méthode que nous proposons a comme objectif à long terme de mettre à disposition une telle méthode, qui devrait être suffisamment sensible pour détecter les relations dont nous avons besoin et avoir une bonne spécificité afin d'éliminer les nombreuses relations contenues dans l'UMLS qui n'apportent rien à la modélisation du domaine.

Par rapport aux critères d'évaluation de notre approche, cette méthode basée sur UMLS est une réutilisation effective d'une connaissance déjà acquise. Cette méthode

est actuellement utilisée pour enrichir l'ontologie existante, elle facilite le travail de l'expert et aboutit à un coût de développement plus réduit. Le coût de construction et la reproductibilité dans des contextes différents restent encore à estimer.

Références

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus . Actes des journées francophones d'Ingénierie des connaissances, Toulouse, 2000; p 93-104

BACHIMONT B.(2000) engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances. *In : Ingénierie des connaissances*.

BODENREIDER O. (2001), Circular hierarchical relationships in the UMLS : Etiology, Diagnosis, Treatment, Complications and Preventions, *Proc. of AMIA 2001*, p.57-61

BODENREIDER O. (2003), Strength in numbers: Exploring redundancy in hierarchical relations across biomedical terminologies, *AMIA 2003 Symposium Proceedings,*101.

BOUSQUET C, LAGIER G., LILLO-LE LOUËT A.,LE BELLER C.,VENOT A., JAULENT M.C., (2005) Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reaction. *Drug Saf 2005*; 28(1):19-34.

BROWN EG. (2002) Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART. *Drug Saf* 2002; 25(6):445-52.

CHARLET J. (2003), L'Ingénierie des connaissances, Développements, Résultats et Perspectives pour la gestion des connaissances médicales, *Mémoire d'habilitation à diriger des recherches*, 2003

CIMINO J. J., « Coding Systems in Health Care », VAN BEMMEL J. H., MCCRAY A. T., Eds., *Yearbook of Medical Informatics '95 — The Computer-based Patient Record*, p. 71– 85, Schattauer, Stuttgart, 1996.

COTÉ R.A., ROTHWELL D.J., PALOTAY J.L., BECKET R.S., BROCHU L., SNOMED International. College of American Pathologists (1993)

HENEGAR C, BOUSQUET C, LILLO-LE LOUËT A.,DEGOULET P., JAULENT M.C. (2004) A knowledge based approach for automated signal generation, *Medinfo* 2004; 2004:626-30.

LE MOIGNO S., CHARLET J., BOURIGAULT D., DEGOULET P., JAULENT M-C., Terminology extraction from text to build an ontology in surgical intensive car,. Proc AMIA Symp., 430-4 (2002)

LI J., ZHANG Y. (2003) Direct interesting rule generation, *Proceedings of The Third IEEE International Conference on Data Mining (ICDM)*, 2003, 155 – 162, IEEE computer society.

LINDBERG D.A., HUMPHREYS B.L., MCCRAY A.T., « The Unified Medical Language System », Methods Inf Med, vol. 32, n° 4, 1993, p. 281-91.

NATIONAL LIBRARY OF MEDICINE, Bethesda, Maryland, « Medical Subject Headings », 2001,disponible à *www.nlm.nih.gov/mesh/meshhome.html*.

ORGANISATION MONDIALE DE LA SANTE, Genève, « Classification statistique internationale des maladies et des problèmes de santé connexes— Dixième révision », 1993.

PETIOT D., BURGUN A., LE BEUX P. (1996) Modelisation of a criterion of proximity : aplication to medical thesauri. Brender J. *Medical Informatics Europe 1996*. IOS Press,149-53.

YOKOTSUKA M, AOYAMA M, KUBOTA K. (2000) The use of a medical dictionary for regulatory activities terminology (MedDRA) in prescription-event monitoring in Japan (J-PEM). Int J *Med Inf 2000*; 57(2-3):139-53.

ZWEIGENBAUM P., « Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances », *Innovation Stratégique en Information de Santé*, no 2–3, 1999, p. 27–47.