Exploitation de la distance sémantique pour l'adaptation de terminologies biomédicales

Marie Dupuch[1], Ludovic Trinquart[2], Isabelle Colombet[1,3], Marie-Christine Jaulent[1], Natalia Grabar[1,3]

(1) Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872, Paris, F-75006; Université Paris Descartes, UMR_S 872, Paris, F-75006; INSERM, U872, Paris, F-75006 France
(2) AP-HP, Paris France
(3) HEGP AP-HP, 20 rue Leblanc, Paris, F-75015 France

Résumé : Nous présentons une expérience d'adaptation de terminologies du domaine biomédical à des questions cliniques extrêmement précises, ce qui peut être requis dans certains contextes applicatifs comme celui des revues systématiques. L'UMLS, qui réunit une centaine de terminologies, nous fournit le matériel d'origine. Nous exploitons les mesures de distance sémantique pour délimiter, au sein d'UMLS, un espace sémantiquement réduit, homogène et relatif à la question clinique étudiée. Pour ceci, les valeurs symboliques associées aux relations liant deux concepts dans l'UMLS sont transformées en valeurs numériques. Une évaluation effectuée par les experts montre que les concepts extraits sont pertinents. Des filtres supplémentaires sont appliqués et permettent d'augmenter la précision de la ressource terminologique ainsi obtenue.

Mots-clés : Adaptation de terminologies, UMLS, similarité sémantique, biomédecine

1 Introduction

La synthèse d'information sur une question clinique précise nécessite une revue systématique de la littérature (www.cochrane.org). Il s'agit d'un processus long et fastidieux. Dans le cadre du projet ReSyTAL, nous proposons de concevoir et de développer un outil d'aide à la sélection des études pour effectuer les revues systématiques sur de nouvelles questions cliniques. Dans ce contexte, les corpus textuels ne sont pas disponibles et nous proposons de réutiliser les RTO (ressources termino-ontologiques) existantes. En effet, dans le domaine biomédical, il existe plusieurs RTO et l'UMLS (NLM, 2008) en réunit une centaine. Il s'agit souvent des terminologies *généralistes* qui décrivent exhaustivement le domaine médical. Afin de disposer d'une RTO ciblée sur un domaine clinique particulier, nous proposons d'*adapter les terminologies existantes*. Il existe quelques travaux qui ont traité cette question : réduction des ontologies (Wang *et al.*, 2008), réorganisation de top ontologies (Qi *et al.*, 2008) et observation de leur évolution (Guelfi *et al.*, 2007). Nous proposons d'utiliser une méthode basée sur les

mesures de distance sémantique pour l'adaptation de terminologies existantes. En exploitant les graphes terminologiques, la distance sémantique permet d'établir la proximité entre deux termes. Par exemple, cette distance est plus petite entre *péricardite* et *péricarde* qu'entre *péricardite* et *estomac*, car sur le graphe terminologique le nombre de relations est plus petit dans le premier cas. Les approches pour le calcul des distances sémantiques implémentent des notions différentes : longueur du plus court chemin (Rada *et al.*, 1989) ; profondeur et densité des concepts (Sheng *et al.*, 2003) ; profondeur des concepts et types de relations (Sussna, 1993) ; informativité des concepts calculée sur la base de leurs occurrences en corpus (Resnik, 1995) ou de la densité de la structure hiérarchique autour d'eux (Seco *et al.*, 2004). Quelle que soit la mesure choisie, la difficulté principale consiste en l'établissement de poids associés à chaque arête (ou relation). Dans notre travail, nous proposons aussi une approche pour la transformation des valeurs symboliques associées aux arêtes en valeurs numériques afin que l'application des mesures de distance sémantique soit possible.

2 Matériel

UMLS : Unified Medical Language System. L'UMLS est un métathesaurus, proposant la fusion d'une centaine de terminologies biomédicales, dont MeSH (NLM, 2001). Lors de l'inclusion dans l'UMLS, chaque concept est assigné à une hiérarchie. Par exemple, *Colorectal neoplasm* appartient à l'hiérarchie *Neoplastic Process*. L'UMLS propose 16 relations homogénéisées, que nous exploitons dans notre travail :

PAR `père de`, CHD `fils de`, SIB `frère de`
SY `synonyme`, RL `concept similaire ou identique`
RQ `relation apparentée, voire synonyme`
RN `relation étroite`, RO `relation autre que synonyme`
RB `relation éloignée`, RU `relation apparentée mais non-précisée`
QB `peut être qualifié par`, AQ `qualificatif permis`

Requête et descripteurs MeSH. La question clinique traitée dans ce travail, *diagnostic de métastase hépatique d'un cancer colorectal*, correspond à six descripteurs MeSH : *Colorectal Neoplasm ; Liver neoplasm ; Laparoscopy ; Tomography, Emission-Computed ; Magnetic resonance imaging* et *Tomography, X-Ray Computed*.

3 Méthode pour l'adaptation des terminologies

La méthode conçue est composée de trois étapes principales : (1) extraction des concepts d'UMLS à partir des descripteurs MeSH ; (2) pondération des arêtes d'UMLS avec des heuristiques ; (3) calcul de la distance sémantique et filtrage des concepts. L'évaluation des résultats est effectuée par les experts. La précision est calculéee, elle correspond au nombre de concepts validés parmi tous les concepts extraits.

3.1 Extraction des concepts à partir des points d'amorce

Nous utilisons les descripteurs MeSH comme points d'amorce pour extraire d'UMLS un ensemble de concepts. Le parcours du graphe UMLS est effectué avec un algorithme

de parcours en largeur : il permet d'atteindre tous les concepts d'une profondeur donnée. La profondeur parcourue est définie expérimentalement.

3.2 Pondération des arêtes d'UMLS avec les heuristiques

La définition des heuristiques pour assigner les poids aux arêtes d'UMLS est l'étape principale de notre travail grâce à laquelle la distance sémantique peut être calculée. Les valeurs des poids sont positionnées sur une échelle de 0 à 100, où 0 correspond à des concepts très proches, alors que 100 correspond à des concepts très éloignés.

Type de relation entre 2 concepts $Trel$. Le principe d'exploitation des relations d'UMLS (sec.2) est le suivant : une relation qui relie des termes sémantiquement proches reçoit un poids faible, tandis qu'avec le relâchement de la sémantique des relations le poids augmente. Par exemple, deux concepts reliés par une relation comme la synonymie ont de nombreux éléments sémantiques communs. La synonymie reçoit ainsi un poids $w_{Trel}(SY)$ égal à 0, tandis qu'une relation comme RB, reliant des concepts distants, reçoit un poids $w_{Trel}(RB)$ égal à 90.

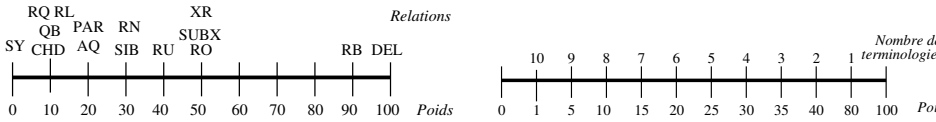
Terminologies source qui proposent une relation donnée $NatT$. Cette heuristique est composée de plusieurs critères et donne lieu à deux poids : w_T (*taille des terminologies*) et w_{Nat} (*thème abordé*, *ancienneté et mises à jour*, et *nombre de publications sur les terminologies*). Par exemple, la terminologie MeSH (tab. 1) est une grande terminologie relativement fiable (ancienneté et mise à jour récente) : elle reçoit un poids w_T égal à 30 et un poids w_{Nat} égal à 20. La moyenne de ces deux poids correspond au poids w_{NatT}, qui est alors égal à 25.

Nombre de terminologies qui proposent une relation donnée $Nsour$. Chaque relation peut être proposée par une ou plusieurs terminologies. Nous considérons que si une relation est proposée par une seule terminologie, elle est moins fiable que lorsqu'elle est proposée par deux ou plus terminologies : elle est d'autant plus fiable que le nombre de terminologies est élevé.

Établissement d'un poids unique $\mathcal{W}$. Le poids unique de chaque arête $\mathcal{W}$ correspond à la moyenne pondérée des trois heuristiques. La pondération est effectuée parce que les heuristiques ne sont pas équivalentes entre elles.

3.3 Calcul de la distance sémantique et filtrage des concepts

Le calcul de la distance sémantique est effectuée avec l'algorithme de Dijkstra (Cormen *et al.*, 2001), car il permet de prendre en compte des valeurs numériques hétérogènes, obtenues grâce aux heuristiques, dans le calcul de la distance sémantique. Cette démarche permet d'extraire d'UMLS un sous-graphe terminologique spécifique et homogène. Cette approche aboutit à une terminologie descriptive parce que la majorité des terminologies source est descriptive. Les terminologies descriptives sont utilisées dans plusieurs contextes applicatifs mais, dans le cadre de revues systématiques diagnostiques, la présence de certains termes peut être gênante. Par exemple, les médicaments ont des chances de privilégier des études consacrées aux essais thérapeutiques et de donner ainsi moins de valeur aux études diagnostiques. Pour répondre à cette problématique, nous appliquons des filtres supplémentaires en exploitant la hiérarchie des termes.



(a) Heuristique $Trel$ et poids w_{Trel} associé

(b) Heuristique $Nsour$ et poids w_{Nsour} associé

FIG. 1 – Établissement des poids en fonction de deux heuristiques : $Trel$ et $Nsour$

RTO	*Nb publi*	*Création*	*Mise à jour*	*Thème*	w_T	w_{Nat}	w_{NatT}
MeSH	—	1954	2009	Vocabulaire clinique	30	20	25
Snomed Int	106	1979	1998	Signes et symptômes	40	25	32.5
Snomed CT	332	2002	2009	Signes et symptômes	30	20	25
ICPC	188	1993	2005	Signes et symptômes	10	10	10
BI	—	1999	—	Signes et symptômes	30	45	37.5
RxNorm	—	2004 ?	2008	Médicaments	20	35	27.5
NCI	—	2004 ?	2008	Pathologies (cancer)	20	30	25

TAB. 1 – Exemple de quelques terminologies avec leur poids w_{NatT}.

4 Résultats et discussion

Une première extraction de concepts d'UMLS est effectuée à partir des 6 descripteurs MeSH et sans contrainte sémantique. Nous avons 2 590 concepts de profondeur 1 et 63 911 de profondeur 2, tandis qu'en profondeur 3 nous obtenons plus d'un million de concepts (ce qui correspond presque à l'intégralité d'UMLS). La profondeur maximale d'extraction est fixée à 2 arêtes, car les concepts provenant de profondeurs supérieures paraîssent non pertinents. Les échelles de la figure 1 indiquent la pondération définie pour les heuristiques $Trel$ et $Nsour$. Quant à l'heuristique $NatT$ (table 1), il s'agit d'une heuristique composée : le poids w_{NatT} correspond à la moyenne des poids w_T et w_{Nat}. Ainsi, la Snomed CT reçoit un poids w_{NatT} égal à 25. La figure 2 montre la pondération des trois heuristiques pour établir un poids unique $\mathcal{W}$.

Ces poids sont ensuite exploités pour poser les contraintes lors du parcours du graphe d'UMLS : les poids de chaque arête du chemin parcouru sont additionnés et le poids total du chemin ne doit pas dépasser le seuil fixé. Nous avons fait les tests avec plusieurs seuils. A la première position de la figure 4, il s'agit des résultats d'une extraction brute. Lorsque le seuil est fixé à 100, il a peu d'influence sur les résultats ; les seuils fixés à 76 et 60 ont une influence intermédiaire ; tandis que les seuils 50 et 49 sont restrictifs. Nous avons décidé de fixer le seuil à 60 : il montre un bon compromis de sélection. Avec ce seuil, notre méthode extrait 2 503 et 2 533 concepts de profondeur 1 et 2 respectivement.

Pour effectuer une évaluation, ces résultats ont été examinés par des experts. Une première étape de l'évaluation a consisté en une validation de deux concepts : *Colorectal Neoplasms* et *Liver neoplasms*. Les deux premières lignes de la table 2 et la colonne *Seuil=60* indiquent les résultats de cette évaluation : les chiffres correspondants sont en gras. La dernière colonne *Experts* indique la sélection effectuée par les experts. La

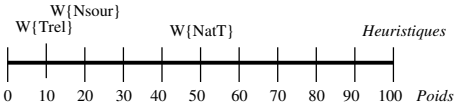


FIG. 2 – Positionnement des heuristiques entre elles pour le calcul d'un poids unique

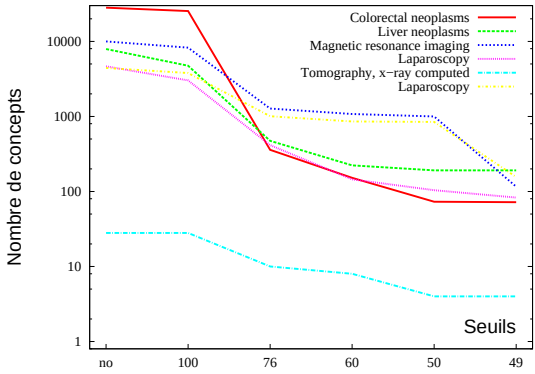


FIG. 3 – Nombre de concepts extraits en fonction des seuils fixés

précision $\mathcal{P}$ est faible : 10,5 et 4,5 %. Un si faible taux de précision est causé par la différence qui existe entre une terminologie descriptive et une terminologie réduite à une question clinique. Pour mieux cerner cette sélection, un filtre supplémentaire, exploitant la hiérarchie des concepts d'UMLS, est appliqué sur les résultats obtenus avec *Seuil=60*. La colonne *Filtre* indique les résultats alors obtenus : la précision augmente et, pour trois concepts, elle reproduit la (non) sélection des experts. Une analyse de cette sélection montre que les 102 concepts validés proviennent en grande partie des relations CHD, SIB et PAR, les relations RQ et RN apparaîssent très peu, et la relation RO n'apparaît qu'une seule fois. D'autres relations (AQ, QB et RB), ne figurent pas dans cette sélection. Ces 102 concepts proviennent de différentes terminologies, dont les principales sont WHO-ART, MEDCIN, NCI, RCD, MSH, SNOMED CT, ce qui justifie le choix d'exploiter l'UMLS et non une seule terminologie.

5 Conclusion et Perspectives

Nous avons proposé une méthode pour l'adaptation de terminologies existantes grâce à l'exploitation de la distance sémantique et à l'extraction de concepts sémantiquement homogènes. Nous avons aussi exploité les informations sémantiques symboliques associées aux arêtes d'UMLS et les ont transformées en valeurs numériques. L'évaluation des résultats par les experts montre que la couverture de la ressource terminologique obtenue est satisfaisante. En ce qui concerne la précision, des filtres supplémentaires ont

Termes	*Seuil=60*	$\mathcal{P}$	*Filtre*	$\mathcal{P}$	*Experts*
Colorectal neoplasms	**152**	10,5	37	43,2	16
Liver neoplasms	**223**	4,5	131	7,6	10
Laparoscopy	145	6,2	44	20,5	9
Laparoscopy	44	–	0	–	0
Tomography, Emission-Computed	8	–	0	–	0
Magnetic Resonance Imaging	1 078	1,8	318	6,0	19
Magnetic Resonance Imaging	2	–	0	–	0
Tomography, X-Ray Computed	857	5,6	247	19,4	48
TOTAL	2 509	3,575	777	49,587	102

TAB. 2 – Filtrage et validation des concepts et la précision $\mathcal{P}$ obtenue.

été ajoutés afin de satisfaire les particularités liées à l'application. La méthode doit être testée dans d'autres contextes et applications afin de prouver sa validité. Elle pourrait aussi être appliquée à d'autres RTO. Les perspectives principales concernent l'étude de l'apport individuel de chaque heuristique et la définition d'autres heuristiques en exploitant les informations contenues dans l'UMLS.

Références

CORMEN T. H., LEISERSON C. E., RIVEST R. L. & STEIN C. (2001). *Introduction to Algorithms*. MIT Press and McGraw-Hill, seconde edition.

GUELFI N., PRUSKI C. & REYNAUD C. (2007). Undestanding and supporting ontology evolution by observing the WWW conference. In *ESOE*.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. `http ://www.nlm.nih.gov/mesh/meshhome.html-`.

NLM (2008). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

QI G., HAASE P., HUANG Z., JI Q., PAN J. Z. & VOLKER J. (2008). A kernel revision operator for terminologies - algorithms and evaluation. In *International Conference on The Semantic Web*, p. 419–34.

RADA R., MILI H., BICKNELL E. & BLETTNER M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on systems, man and cybernetics*, **19**(1), 17–30.

RESNIK P. (1995). Disambiguating noun groupings with respect to wordnet senses.

SECO N., VEALE T. & HAYES J. (2004). An intrinsic information content metric for seman- tic similarity in wordnet. In *Proceedings of ECAI 2004, the 16th European Conference on Artificial Intelligence*.

SHENG Z., CHENG L. & WING H. W. (2003). ChipInfo : software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res*, **31**(13), 3483–3486.

SUSSNA M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Conference on Information and Knowledge Management*.

WANG Z., WANG K., TOPOR R. & PAN J. Z. (2008). Forgetting concepts in DL-Lite. In S. VERLAG, Ed., *WWTP*, p. 245–57.