Utilisation de méthodes de structuration de terminologies pour la création de groupements de termes de pharmacovigilance

Marie Dupuch
CRC
Université Paris 6
Inserm, UMRS 872
F-75006, Paris, France
marie.dupuch@crc.jussieu.fr

Amandine Périnet
Thierry Hamon
LIM&BIO (EA3969)
Université Paris 13
93017 Bobigny Cedex
France
amandine.perinet@edu.univ-paris13.fr
thierry.hamon@univ-paris13.fr

Natalia Grabar
CNRS UMR 8163 STL
Université Lille 3
59653 Villeneuve d'Ascq
France
natalia.grabar@univ-lille3.fr

Abstract

Dans ce travail, nous cherchons à détecter des termes sémantiquement proches dans le domaine de la pharmacovigilance et de les regrouper. De tels groupements sont en effet indispensables dans la fouille de données de pharmacovigilance et dans la recherche de nouveaux signaux, afin de réguler au mieux la commercialisation des médicaments. Nous proposons d'utiliser pour ceci des méthodes de TAL. Plus particulièrement, nous exploitons les méthodes de structuration de termes avec des relations de synonymie et hiérarchiques. Les groupements obtenus sont comparés avec des groupements de référence.

1 Introduction

La détection de termes synonymes est cruciale dans des contextes de plusieurs applications, comme la recherche et l'extraction d'information, la structuration de terminologies ou l'annotation sémantique des documents. Par exemple, en recherche d'information, il est important de retrouver les documents répondant au mot-clé *muscle ache* lorsque l'utilisateur donne le mot-clé *muscle pain*. Les méthodes dédiées à la détection de variantes morpho-syntaxiques (Jacquemin et al., 1997) ou de relations de synonymie (Hamon et al., 1998) sont alors utilisées. Souvent, l'utilisation de termes équivalents, c'est-à-dire sémantiquement proches ({*asystolic*; *asystole*}, {*hematoma muscle*; *hemorrhage muscle*}, {*muscle ache*; *muscle pain*}, {*localized muscle*; *localized muscle weakness*}) sans pour autant être des synonymes apparaît nécessaire pour augmenter la couverture et la sensibilité d'une application.

Ainsi, dans le contexte de l'identification d'effets indésirables susceptibles d'être dus à un médicament, l'acquisition de termes équivalents permet des regroupements plus larges des déclarations issues des bases de pharmacovigilance. Il est ainsi possible d'agréger plus de données textuelles similaires et, par conséquent, de cerner plus rapidement et plus efficacement un effet indésirable de médicament, et d'améliorer ainsi la surveillance du risque médicamenteux.

Les textes de déclarations d'effets indésirables sont encodés avec les termes contrôlés fournis par la terminologie MedDRA (Medical Dictionnary for Drug Regulatory Activities). Pour regrouper ces déclarations et, de cette manière, avoir une information plus complète sur les effets indésirables, plusieurs informations proposées dans MedDRA sont utilisées : (1) les niveaux hiérarchiques de MedDRA structurant la liste de ses termes, (2) les SMQ (Standartised MedDRA Queries) regroupant les termes associés à des conditions médicales données (*Agranulocytose*, *Insuffisance rénale aigue*). Les SMQ sont issus d'un travail long et méticuleux réalisés par des groupes d'experts et consistant à étudier la structure de MedDRA et à analyser manuellement la littérature scientifique (CIOMS, 2004). Des études (Pearson et al., 2009) ont montrées que les SMQ ne sont pas exhaustifs car plusieurs conditions médicales importantes pour la pharmacovigilance n'ont pas encore été répertoriées.

Nous considérons que des méthodes automatiques peuvent faciliter et systématiser le processus de création de groupements de termes. Les termes pouvant être dans plusieurs SMQ, les méthodes de classification s'avèrent peu adaptées. Nous nous proposons d'exploiter des méth-

odes d'identification de variantes terminologiques (termes synonymes ou plus généralement des termes sémantiquement liés). Pour valider notre approche, nous évaluons les regroupements calculés automatiquement par rapport aux SMQ. La structure hiérarchique de MedDRA est aussi analysée en regard de nos résultats.

2 Matériel

Nous exploitons trois types de matériel en langue anglaise : termes MedDRA que nous cherchons à regrouper, ressources lexicales utilisées par les méthodes de TAL et groupements de référence par rapport auxquels nous évaluons nos groupements.

2.1 Terminologie MedDRA

La terminologie MedDRA (Brown et al., 1999) a été spécifiquement conçue pour le codage des effets indésirables. MedDra contient un large spectre de termes (signes et symptômes, diagnostics, examens de laboratoire, procédures médicales et chirurgicales, antécédents), organisés en cinq niveaux hiérarchiques: *System Organ Class* ou SOC (n=26); *High Level Group Terms* ou HLGT (n=332); *High Level Terms* ou HLT (n=1 688); *Prefered Terms* ou PT (n=18 209); *Low Level Terms* ou LLT (n=66 587). Chaque niveau inférieur est subsumé hiérarchiquement par le niveau supérieur. Nous exploitons les 18 209 termes PT, eux-mêmes utilisés pour le codage des effets indésirables et pour la création des SMQ.

2.2 Ressources lexicales

Les ressources lexicales sont des paires de termes ou de mots synonymes. Nous en utilisons plusieurs ensembles: (1) Synonymes médicaux extraits directement d'UMLS (n=228 542) et nettoyés (n=73 093); (2) Synonymes médicaux acquis à partir de trois terminologies biomédicales grâce à l'exploitation de leur compositionnalité (Grabar and Hamon, 2010) (n=28 691); (3) Synonymes de la langue générale fournis par WordNet (Fellbaum, 1998) (n=45 782). Nous exploitons aussi une ressource de variantes morphologiques de mots médicaux (n=90 583) issus du travail précédent (Grabar and Zweigenbaum, 2000).

2.3 Groupements de référence SMQ

Les SMQ (Standartised MedDRA Queries) sont des groupements de termes MedDRA liés à une condition médicale (ou diagnostic), comme par exemple *Acute renal failure, Hepatic disorders* ou *Agranulocytosis*. Les SMQ sont créés pour apporter une aide dans la recherche de cas de pharmacovigilance pertinents. Comme les SMQ contiennent des termes sémantiquement proches, leur utilisation permet de trouver plus de déclarations et de les agréger. Ainsi, une exhaustivité des données plus grande permet d'intensifier le signal de pharmacovigilance et peut mener à une émergence des alertes liées à un médicament ou à un principe actif. Il existe actuellement 84 SMQ. Nous utilisons les SMQ comme les données de référence pour une appréciation des groupements de termes générés automatiquement.

3 Méthodes pour la génération de groupements de termes MedDRA

L'approche que nous proposons pour regrouper les termes MedDRA s'appuie d'une part sur l'analyse de la structure interne des termes (identification de relations synonymiques et hiérarchiques) mais aussi sur le partitionnement du réseau de termes pour l'identification de sous-graphes.

Les termes sont analysés linguistiquement à travers la plateforme de TAL Ogmios (Hamon and Nazarenko, 2008). Leur analyse morpho-syntaxique est réalisée à l'aide de GeniaTagger (Tsuruoka et al., 2005), et leur analyse syntaxique en tête/modifieur avec l'extracteur de termes YaTeA(Aubin and Hamon, 2006).

3.1 Identification des relations synonymiques

La méthode utilisée s'appuie sur les travaux précédents permettant d'inférer des relations de synonymie entre des termes complexes (Hamon et al., 1998). Ces travaux proposent d'appliquer le principe de compositionnalité sémantique (Partee, 1984) et postulent que le processus de composition préserve la synonymie. Ainsi, deux termes complexes sont considérés comme synonymes si au moins un de leur composant dans la même position syntaxique sont synonymes. Par exemple, étant donné la relation de synonymie entre les mots *infection* et *sepsis*, les termes *wound infection* (*infection de blessure*) et *wound sepsis* (*septicité de blessure*) sont identifiés comme synonymes.

Pour calculer les relations de synonymie entre les termes MedDRA, nous effectuons plusieurs expériences. Chaque ressource de synonymes

(section 2.2) est d'abord utilisée individuellement et ensuite en combinaison avec WordNet.

3.2 Identification des relations hiérarchiques

Pour identifier des relations hiérarchiques, nous nous appuyons sur l'hypothèse que lorsqu'un terme est inclu lexicalement dans un autre, une relation hiérarchique peut généralement être établie entre le terme court (le père hiérarchique) et le terme long (son fils). Nous calculons trois types d'inclusions lexicales : (1) inclusions calculées à partir des sacs de mots des termes, (2) inclusions calculées à partir de l'analyse syntaxique avec des têtes minimales et (3) inclusions calculées à partir de l'analyse syntaxique avec des têtes maximales. Dans le premier cas, la méthode repère des couples de termes, normalisés à l'aide des variantes morphologiques, où le terme plus long contient tous les mots du terme plus court. Par exemple, le terme *kaolin cephalin clotting time* est considéré comme le fils de *cephalin clotting time*, parce que tous les mots de ce dernier sont inclus dans *kaolin cephalin clotting time*. De plus, aucun contrôle n'est effectué sur le type d'inclusion et le couple {*kaolin cephalin clotting time*; *kaolin*} peut aussi être proposé. Cette approche pouvant être trop permissive, les deux autres approches appliquent la contrainte sur l'identification de relations hiérarchiques en exploitant la décomposition syntaxique des termes en tête/modifieur. Deux stratégies sont alors utilisées. Nous calculons d'abord la tête minimale du terme, qui correspond à la plus petite forme lexicale à laquelle peut se réduire la tête du terme : dans le terme *kaolin cephalin clotting time*, la tête minimale est *time*. Nous calculons ensuite la tête maximale du terme, qui correspond à la forme la plus complète que peut avoir la tête. Dans le même exemple, la tête maximale est *cephalin clotting time*. Dans tous les cas, les termes incluants et inclus doivent être des termes MedDRA. Nous obtenons donc trois listes avec des termes reliés par des inclusions lexicales.

3.3 Des couples de termes vers des groupements

Les trois listes générées par inclusion lexicale proposent des relations hiérarchiques entre termes. Chaque liste est considérée comme un graphe orienté : les termes sont les nœuds du graphe et les relations hiérarchiques sont les arcs orientés. Nous

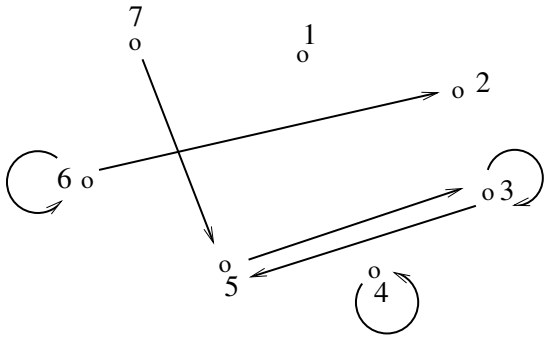
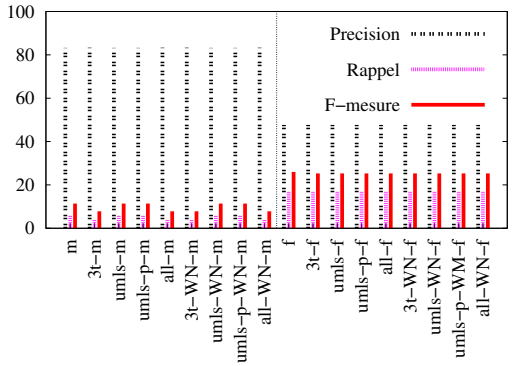


Figure 1: Exemple de graphe orienté

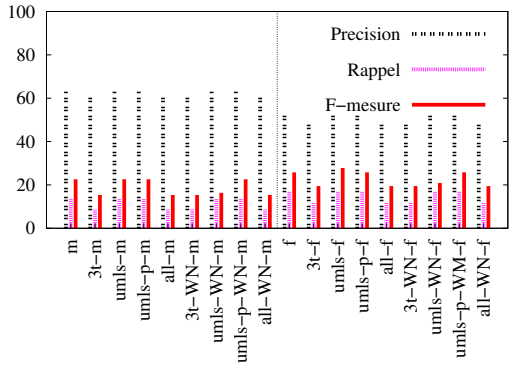
partitionnons ce graphe orienté afin d'identifier les groupes de termes pouvant faire partie du SMQ. Pour cela, nous exploitons la notion de composantes fortement connexes. Il s'agit d'identifier dans le graphe orienté G, les sous-graphes maximaux H de G tel que pour toute couple $\{x, y\}$ de sommets H, il existe un arc orienté de x vers y. Ainsi, le graphe de la figure 1 comporte quatre composantes fortement connexes : {1}, {2}, {2, 6}, {3, 5}, {3, 5, 7} et {4}. Pour améliorer la couverture des groupements correspondant à ces composantes fortement connexes, nous ajoutons les synonymes générés (section 3.1) : si un terme a une relation de synonymie avec un terme du groupement alors il est ajouté à ce groupement. Du point de la théorie des graphes, il s'agit d'augmenter le graphe initial par deux arcs orientés vers et à partir des termes synonymes.

3.4 Évaluation et analyse des groupements

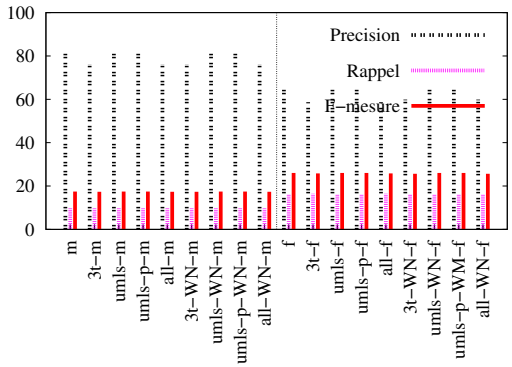
L'évaluation des groupements générés est réalisée grâce à leur comparaison avec les SMQ. Plus précisément, nous considérons neuf SMQ liés aux effets indésirables les plus importants pour la santé publique (*Acute renal failure*, *Agranulocytosis*, *Anaphylactic reaction*, *Cytopenia and haematopoietic disorders affecting more than one type of blood cell*, *Gastrointestinal haemorrhages*, *Peripheral neuropathy*, *Rhabdomyolysis*, *Severe cutaneous adverse reaction*, et *Thrombocytopenia*) et pouvant être à l'origine d'hospitalisations et même de décès. Une évaluation quantitative est effectuée avec trois mesures classiques : précision P (pourcentage de termes pertinents retrouvés rapporté au nombre de termes total groupés), rappel R (pourcentage de termes pertinents retrouvés rapporté au nombre de termes dans un SMQ) et F-mesure F (la moyenne harmonique de P et R).



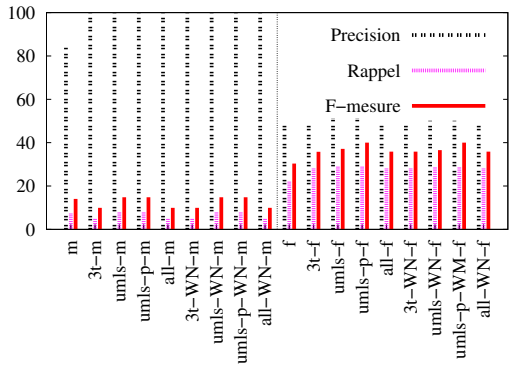
(a) Inclusion lexicale avec des sacs de mots.



(b) Inclusion lexicale avec l'analyse syntaxique (tête syntaxique minimale).



(c) Inclusion lexicale avec l'analyse syntaxique (tête syntaxique maximale).



(d) Inclusion lexicale: fusion des trois approches (sacs de mots, tête syntaxique minimale et tête syntaxique maximale).

Figure 2: Résultats d'évaluation des groupements de termes obtenus avec des méthodes de structuration de termes (`3t` : synonymes acquis sur trois terminologies biomédicales, `umls` : synonymes extraits d'UMLS, `umls-p` : synonymes extraits d'UMLS et nettoyés, `all` : ensemble des listes de synonymes précédentes, `WN` : synonymes de WordNet. `m` : meilleur regroupement, `f` : n meilleurs regroupements).

Une difficulté de notre évaluation réside dans le choix des SMQ utilisés pour évaluer nos regroupements. Étant donné que les pharmacovigilants favorisent la précision au rappel, l'association entre les SMQ et les groupements est guidée par la précision de ces groupements. Ainsi, le groupement qui a la précision la plus élevée par rapport à un SMQ est celui qui sera associé à ce SMQ. Comme l'association des groupements et des SMQ guidée par la précision favorise les groupements de petite taille, nous étudions également *a posteriori* la qualité d'un ensemble de groupements générés dans la situation où ceux-ci sont le résultat de la fusion des n meilleurs groupements proposés par notre méthode. Nous avons réalisé la fusion de groupements en fonction de seuils de précision compris entre 10 et 90 %, et permettant d'obtenir une F-mesure optimale. Nous présentons les résultats pour une précision supérieure à 30 %. Nous effectuons aussi une évaluation qualitative de nos résultats. D'une part nous étudions le contenu des groupements afin d'identifier les indices permettant d'améliorer la qualité de notre méthode, d'autre part nous évaluons la contribution des approches et des ressources de synonymes.

4 Présentation des résultats

La méthode décrite dans les sections 3.1 et 3.2 a été appliquée aux 18 209 termes PT de MedDRA. Les couples de termes PT avec les relations de synonymie et hiérarchiques ont été groupés selon les principes de la section 3.3. L'évaluation des groupements obtenus est présentée à la figure 2. Chaque graphique présente les résultats obtenus avec les approches différentes d'inclusions lexicales : sacs de mots en 2(a), tête minimale en 2(b), tête maximale en 2(c), et fusion de ces trois ensembles en 2(d). Le premier histogramme correspond aux résultats obtenus uniquement avec les inclusions lexicales, les histogrammes suivants représentent les résultats avec les inclusions lexicales et les différents tests de détection de synonymie entre les termes PT. Globalement, les synonymes ont relativement peu d'influence sur les résultats : ce sont surtout les inclusions lexicales qui alimentent les groupements. Chaque graphique présente deux séries d'expériences (séparées par une impulsion) : sélection d'un meilleur groupement (m) et fusion de n meilleurs groupements (f). Comme nous l'avons annoncé, c'est la précision qui est privilégiée dans ce travail. La précision la plus élevée est obtenue avec les sacs de mots en 2(a) et la moins élevée avec l'analyse syntaxique en tête minimale (2(b)). La fusion des n meilleurs groupements optimise le rappel mais conduit à une perte en précision (deuxième série d'expérience sur chaque graphique). La combinaison des approches d'inclusion lexicale en 2(d) permet d'obtenir la f-mesure la plus optimale lorsque la fusion de n meilleurs groupements est réalisée.

5 Discussion et analyse des résultats

5.1 Influence des approches d'inclusion lexicale et des ressources

Les méthodes d'acquisition de relations hiérarchiques par inclusion lexicale exploitant des dépendances syntaxiques, permettent d'obtenir 3 816 relations avec les têtes minimales et 3 366 relations avec les têtes maximales. L'acquisition de relations hiérarchiques avec des sacs de mots est très prolifique et génère 46 053 relations. Comme observé sur les figures 2, lorsque les inclusions lexicales sont calculées avec les sacs de mots, la précision est la plus élevée, tandis que l'analyse syntaxique avec les têtes maximales offre le meilleur compromis entre la précision et le rappel. Dans ce dernier cas, la contrainte syntaxique appliquée mène à une amélioration globale de la qualité des groupements. Par ailleurs, la combinaison de ces trois approches d'inclusion lexicale permet aussi d'améliorer les performances globales : ces approches se complètent mutuellement, ce qui mène à une augmentation du rappel et de la f-mesure.Mais le bruit de chaque approche provoque une diminution de la précision.

Méthode	nb_t	nb_c	P	R	F
sacs de mots	13	6	46	8	13
AS tête min	46	37	80	50	61
AS tête max	45	36	80	49	60

Table 1: Comparaison entre les trois approches d'inclusion lexicale pour le SMQ *Agranulocytosis*.

Quelle que soit l'approche d'inclusion lexicale, l'influence des synonymes (section 2.2) est stable. Pris séparément, la ressource de langue générale WordNet *WN* a un faible impact, mais permet une amélioration de la qualité des regroupements lorsqu'elle est utilisée avec des ressources du domaine médical. Les synonymes obtenus à partir de trois terminologies *3t* génèrent 1 879 couples (1 939 en combinaison avec *WN*). Leur influence lors de l'évaluation est soit nulle soit légèrement négative. Quant aux synonymes extraits d'UMLS, *umls* et *umls-p*, bien que très volumineux, ils génèrent peu de relations (190 paires lorsqu'ils sont projetés seuls et 227 en combinaison avec *WN*). Par contre, les paires générées ont une influence positive sur les résultats. Par ailleurs, nous n'observons pas de différence selon que cette ressource est nettoyée (*umls-p*) ou non (*umls*).

5.2 Analyse détaillée des groupements

Nous avons analysé en détail les groupements relatifs aux SMQ *Agranulocytosis* et *Gastrointestinal haemorrhage* (avec fusion de n meilleurs groupements) pour observer l'impact des trois méthodes d'inclusion lexicale. Nous présentons les résultats de cette analyse pour le SMQ *Agranulocytosis*. Ce SMQ contient 74 termes. La table 1 indique le nombre de termes nb_t dans le groupement calculé, le nombre de termes communs nb_c et les performances obtenues avec les différentes approches (P, R et F). Les meilleurs

résultats sont obtenus avec l'analyse syntaxique (têtes minimales et maximales). La précision est alors élevée et le rappel satisfaisant pour les pharmacovigilants. Les performances sont faibles pour l'approche par sacs de mots. Nous avons réalisé, avec un expert, une étude détaillée du bruit généré.

L'approche par sacs de mots fournit sept termes qui n'appartiennent pas au SMQ : *Nocardia test positive*, *Autoimmune neutropenia*, *Cyclic neutropenia*, *Idiopathic neutropenia*, *Neutropenia neonatal*, *Bubonic plague* et *Pneumonic plague*. Les termes du SMQ *Agranulocytosis* concernent l'agranulocytose mais aussi ses conséquences. Trois termes (*Bubonic plague*, *Pneumonic plague* et *Plague sepsis*) pourraient être inclus dans ce SMQ : les deux premiers sont des manifestations de cette pathologie, et le troisième correspond à une conséquence. Quant aux autres termes (*Autoimmune neutropenia*, *Cyclic neutropenia*, *Idiopathic neutropenia* et *Neutropenia neonatal*), ce ne sont pas des effets indésirables. Leur filtrage peut être basé sur des marqueurs lexicaux comme *autoimmune*, *cyclic*, *idiopathic* ou *neonatal*.

L'approche avec les têtes minimales fournit neuf termes qui n'appartiennent pas au SMQ : *Group b streptococcus neonatal sepsis*, *Herpes sepsis*, *Fungal sepsis*, *Burkholderia cepacia complex sepsis*, *Anthrax sepsis*, *Autoimmune pancytopenia*, *Viral tonsillitis*, *Chronic tonsillitis* et *Candida sepsis*. Les termes *Herpes sepsis*, *Fungal sepsis* et *Candida sepsis* sont une conséquence de l'agranulocytose et pourraient aussi être inclus dans ce SMQ. Le même type de marqueurs peut être utilisé pour filtrer les termes qui ne sont pas les effets indésirables (*Autoimmune pancytopenia*, *Viral tonsillitis* et *Chronic tonsillitis*). Les trois termes qui restent (*Group b streptococcus neonatal sepsis*, *Burkholderia cepacia complex sepsis* et *Anthrax sepsis*) sont des faux positifs.

L'approche avec les têtes maximales génère neuf termes qui ne sont pas dans le SMQ : *Congenital aplastic anaemia*, *Anthrax sepsis*, *Herpes sepsis*, *Fungal sepsis*, *Phlebitis*, *Viral tonsillitis*, *Chronic tonsillitis*, *Injection site phlebitis*, *Candida sepsis*. *Congenital aplastic anaemia* pourrait être inclus dans le SMQ car il correspond à une conséquence de l'agranulocytose bien que rare. Par rapport aux analyses précédentes, nous avons deux termes de plus (*Phlebitis* et *Injection site phlebitis*) qui sont des erreurs et ne devraient pas être groupés. Pour ces cas, il est nécessaire d'établir d'autres stratégies pour éliminer le bruit.

5.3 Structure de MedDRA

Lors de l'analyse des résultats obtenus, nous avons été très surpris d'observer un apport aussi important des inclusions lexicales dans la génération des groupements. En effet, nous avons travaillé avec les termes MedDRA qui proviennent du même niveau hiérarchique (PT). Logiquement, ces termes devraient être assez équivalents et peu de relations hiérarchiques devraient être générées. En réalité nous détectons deux, trois et même plus de niveaux hiérarchiques au sein des groupements. Cela veut dire que la structuration hiérarchique actuelle de MedDRA pourrait être affinée et des niveaux hiérarchiques intermédiaires créés.

5.4 Travaux similaires

D'autres travaux s'intéressent au groupement de termes de pharmacovigilance mais ceux-ci proviennent de la terminologie WHO-ART, progressivement abandonnée au profit de MedDRA. Les méthodes utilisées sont la distance sémantique (Bousquet et al., 2005; Iavindrasana et al., 2006) et la subsomption hiérarchique (Jaulent and Alecu, 2009). La distance sémantique a été appliquée à un sous-ensemble de termes WHO-ART : les groupements obtenus montrent plusieurs types de relations (synonymie, antonymie, fonctions physiologiques, symptômes associés, tests de laboratoire anormaux, pathologies et leurs causes, ainsi que des groupements hétérogènes), et n'ont pas été comparés aux SMQ. Les groupements obtenus par subsomption hiérarchique ont été confrontés aux SMQ : le rappel est en moyenne de 82, et la précision n'a pas été évaluée. Nos résultats, évalués en termes de précision et de rappel, semblent être supérieurs au travail cité. Plus récemment, des groupements de termes MedDRA ont été réalisés avec la distance sémantique (Dupuch et al., 2011) : la comparaison avec les SMQ montre une précision très élevée (souvent entre 80 et 100 %) et un rappel pouvant alors atteindre 23 %. Nos résultats semblent être complémentaires avec ceux générés par la distance sémantique.

Les groupements générés par la méthode proposée ici se distinguent surtout par une bonne précision. Nous pensons qu'ils peuvent être utilisés d'une part pour constituer des composantes qui al-

imentent la création des SMQ et d'autre part pour affiner la structure des termes au sein des SMQ. Car il a été en effet observé qu'actuellement les SMQ visent surtout à privilégier le rappel (Pearson et al., 2009; Mozzicato, 2007) au détriment de la précision : très souvent jusqu'à 95 % de cas collectés avec les SMQ sont rejetés par les pharmacovigilants, ce qui correspond à un filtrage manuel très lourd. Si des groupements plus précis et fins sont disponibles au sein des SMQ, cela peut faciliter le travail des pharmacovigilants.

6 Conclusion et perspectives

Le travail présenté ici exploite les méthodes de structuration de termes et permet de générer des groupements de termes MedDRA qui montrent une très bonne précision. Une analyse détaillée des résultats indique que certains des termes absents des SMQ pourraient y être également inclus. Avec une précision élevée, ces groupements peuvent être utilisés pour constituer des composantes qui alimentent la création des SMQ mais aussi pour affiner la structure des termes au sein des SMQ. Une première analyse effectuée montre que les résultats obtenus dans cette expérience sont complémentaires aux méthodes qui exploitent la distance sémantique. Nous prévoyons ainsi de combiner ces deux types de méthodes pour optimiser les groupements. Nous allons également analyser les types de relations sémantiques qui existent entre les termes au sein des SMQ pour ouvrir des pistes vers d'autres méthodes de clustering et de groupements de termes. Par exemple, nous prévoyons d'exploiter des corpus pour détecter des relations sémantiques transversales entre termes, comme par exemple les causes d'une pathologie ou les résultats d'examens biologiques anormaux et relevant d'une pathologie. Une attention particulière est portée à l'évaluation et à l'association entre les SMQ et les groupements.

References

S Aubin and T Hamon. 2006. Improving term extraction with terminological resources. In *FinTAL 2006*, number 4139 in LNAI, pages 380–387. Springer.

C Bousquet, C Henegar, A Lillo-Le Louët, P Degoulet, and MC Jaulent. 2005. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform*, 74(7-8):563–71.

EG Brown, L Wood, and S Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109–17.

CIOMS. 2004. Development and rational use of standardised MedDRA queries (SMQs): Retrieving adverse drug reactions with MedDRA. Technical report, CIOMS.

M Dupuch, M Lerch, A Jamet, MC Jaulent, R Fescharek, and N Grabar. 2011. Grouping pharmacovigilance terms with semantic distance. In *MIE*.

C Fellbaum. 1998. A semantic network of english: the mother of all WordNets. *Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network*, 32(2-3):209–220.

N Grabar and T Hamon. 2010. Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In *MEDINFO 2010*, pages 1015–9.

N Grabar and P Zweigenbaum. 2000. A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, pages 310–314.

T Hamon and A Nazarenko. 2008. Le développement d'une plate-forme pour l'annotation spécialisée de documents web: retour d'expérience. *TAL*, 49(2):127–154.

T Hamon, A Nazarenko, and C Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In *COLING-ACL'98*, pages 498–504.

J Iavindrasana, C Bousquet, P Degoulet, and MC Jaulent. 2006. Clustering who-art terms using semantic distance and machine algorithms. In *AMIA Annu Symp Proc*, pages 369–73.

C Jacquemin, JL Klavans, and E Tzoukerman. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *ACL/EACL 97)*, pages 24–31, Barcelona, Spain.

MC Jaulent and I Alecu. 2009. Evaluation of an ontological resource for pharmacovigilance. In *Stud Health Technol Inform*, pages 522–6.

P Mozzicato. 2007. Standardised MedDRA queries: their role in signal detection. *Drug Saf*, 30(7):617–9.

BH Partee, 1984. *Compositionality*. F Landman and F Veltman.

RK Pearson, M Hauben, DI Goldsmith, AL Gould, D Madigan, DJ O'Hara, SJ Reisinger, and AM Hochberg. 2009. Influence of the MedDRA hierarchy on pharmacovigilance data mining results. *Int J Med Inform*, 78(12):97–103.

Y Tsuruoka, Y Tateishi, JD Kim, T Ohta, J McNaught, S Ananiadou, and J Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *LNCS*, 3746:382–392.