TALN 2008, Avignon, 9–13 juin 2008

Apprentissage artificiel de règles d'indexation pour MEDLINE

Aurélie Névéol
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
neveola@nlm.nih.gov

Vincent Claveau
IRISA - CNRS
Campus de Beaulieu
35042 Rennes cedex, France
Vincent.Claveau@irisa.fr

Résumé. L'indexation est une composante importante de tout système de recherche d'information. Dans MEDLINE, la base documentaire de référence pour la littérature du domaine biomédical, le contenu des articles référencés est indexé à l'aide de descripteurs issus du thésaurus MeSH. Avec l'augmentation constante de publications à indexer pour maintenir la base à jour, le besoin d'outils automatiques se fait pressant pour les indexeurs. Dans cet article, nous décrivons l'utilisation et l'adaptation de la Programmation Logique Inductive (PLI) pour découvrir des règles d'indexation permettant de générer automatiquement des recommandations d'indexation pour MEDLINE. Les résultats obtenus par cette approche originale sont très satisfaisants comparés à ceux obtenus à l'aide de règles manuelles lorsque celles-ci existent. Ainsi, les jeux de règles obtenus par PLI devraient être prochainement intégrés au système produisant les recommandations d'indexation automatique pour MEDLINE.

Abstract. Indexing is a crucial step in any information retrieval system. In MEDLINE, a widely used database of the biomedical literature, the indexing process involves the selection of Medical Subject Headings in order to describe the subject matter of articles. The need for automatic tools to assist human indexers in this task is growing with the increasing amount of publications to be referenced in MEDLINE. In this paper, we describe the use and the customization of Inductive Logic Programming (ILP) to infer indexing rules that may be used to produce automatic indexing recommendations for MEDLINE indexers. Our results show that this original ILP-based approach overperforms manual rules when they exist. We expect the sets of ILP rules obtained in this experiment to be integrated in the system producing automatic indexing recommendations for MEDLINE.

Mots-clés : Analyse et Indexation/méthodes ; Medical Subject Headings ; Apprentissage Artificiel ; Programmation Logique Inductive.

Keywords: Abstracting and Indexing/methods ; Medical Subject Headings ; Machine Learning ; Inductive Logic Programming.

1 Introduction

La recherche d'information dans une collection de documents qu'elle soit spécialisée (*e.g.* Legifrance[1], MEDLINE®[2]) ou non (*e.g.* Internet) nécessite une *indexation* des documents. L'index ainsi créé est ensuite utilisé pour évaluer la pertinence de chacun des documents de la collection pour les requêtes des utilisateurs. Dans le cadre de documents textuels, on distingue usuellement deux types d'indexation : l'*indexation libre*, qui permet d'utiliser sans limitation des séquences de mots quelconques, et l'*indexation contrôlée* qui utilise de manière contrainte les concepts répertoriés dans une liste prédéfinie. On parle alors de vocabulaire contrôlé ; c'est dans ce cadre que se situe notre travail.

Dans le domaine biomédical, le thésaurus MeSH® (Medical Subject Headings) développé par la U.S. National Library of Medicine (NLM) est l'outil de prédilection pour indexer la littérature. Ainsi, la base MEDLINE recense plus de 17 millions d'articles scientifiques du domaine et reçoit 3 millions de requêtes par jour. Ces articles sont référencés à l'aide de notices descriptives, ou « citations », contenant notamment une quinzaine de mots-clés représentant les concepts abordés par les auteurs parmi les quelque 24 000 du thésaurus (*e.g. aphasia*, *patient care*, *hand*...), organisés hiérarchiquement (*e.g. thumb* est un descendant de *hand* dans l'arborescence *Anatomy*). Le cas échéant, ces mots-clés doivent ensuite être précisés à l'aide des 83 qualificatifs (*e.g. surgery*, *pharmacology*). Pour chaque mot-clé, le MeSH définit un sous-ensemble de qualificatifs « affiliables », de sorte que seules certaines paires peuvent être formées. Par exemple, les paires *aphasia/metabolism* ou *hand/surgery* sont autorisées, mais pas *hand/metabolism*.

L'augmentation constante du nombre de publications à indexer dans MEDLINE (environ 3 000 par jour) a conduit au développement d'outils automatiques tels que MTI (Aronson *et al.*, 2004) destinés à aider les indexeurs en leur proposant des mots-clés jugés pertinents pour chacun des documents traités. La complexité de la tâche d'indexation et le volume croissant de données à traiter fait du perfectionnement des outils existant une nécessité. En particulier, en ce qui concerne l'indexation MeSH, un progrès notable repose sur l'extraction par des outils automatiques de paires mot-clé / qualificatif et non seulement des mots-clés isolés. Des travaux récents (Névéol *et al.*, 2007) menés dans le cadre du projet « Indexing 2015 » de la NLM ont montré que l'utilisation de règles d'indexation appliquées à partir des mots-clés isolés est une bonne méthode pour l'extraction de paires. Pour quelques qualificatifs, des règles régissant leur emploi ont été trouvées manuellement. Les performances obtenues à l'aide de ces règles manuelles sont satisfaisantes au niveau de la précision, mais le rappel reste faible et le développement manuel de nouvelles règles est à la fois complexe et coûteux.

L'objectif du travail présenté dans cet article est de dépasser ce cadre manuel en inférant automatiquement de nouvelles règles d'indexation impliquant des qualificatifs pour compléter les résultats de MTI. Pour ce faire, nous adoptons une approche originale en utilisant une méthode d'apprentissage symbolique particulière, la programmation logique inductive (PLI), que nous adaptons aux spécificités des données MEDLINE afin d'obtenir des règles pertinentes permettant de traiter efficacement une grande quantité de données. Ces règles, que nous cherchons à inférer, sont du type :

Si un terme de l'arborescence « Anatomy » ainsi qu'un « Carboxylic Acids » sont recommandés pour l'indexation, ***alors*** la paire « [Carboxylic Acids]/pharmacology » doit également être recommandée.

[1] http://www.legifrance.gouv.fr/

[2] Accessible via PubMed à l'URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

2 Travaux connexes

L'indexation de textes du domaine biomédical à l'aide de descripteurs MeSH a donné lieu à de nombreux travaux aussi bien en anglais que dans d'autres langues européennes. Les méthodes utilisées pour l'indexation relèvent aussi bien du TAL (Névéol *et al.*, 2006) que de l'apprentissage (Lin & Wilbur, 2007; Rak *et al.*, 2007) ou d'une combinaison des deux (Markó *et al.*, 2003; Aronson *et al.*, 2004). Cependant, du fait de la complexité du problème de l'indexation MeSH (nombre de descripteurs, multiples combinaisons possibles...), la plupart de ces travaux, même récents, se focalisent sur l'indexation à l'aide de mots-clés MeSH isolés et ne prennent pas en compte le rôle des qualificatifs. Ainsi, nos travaux prennent appui sur ceux de Névéol *et al.* (2006; 2007) pour proposer une avancée innovante dans ce domaine.

Bien que la formalisation de règles d'indexation MeSH n'ait pas fait l'objet de travaux en indexation automatique, cette question a été abordée dans le cadre de la recherche d'information par Soualmia (2004) qui proposait l'application de règles d'association pour l'expansion de requêtes. Cependant, peu de règles avaient été obtenues du fait de la complexité de la description du problème d'apprentissage par des techniques standard (*cf.* section suivante). La technique d'apprentissage artificiel que nous utilisons pour aborder ce problème, la PLI, doit nous permettre de contourner ces difficultés. La PLI a déjà été employée pour certains travaux en TAL ; par exemple pour inférer des patrons d'extraction (Claveau *et al.*, 2003), des règles d'alignement (Ozdowska & Claveau, 2006), faire de l'étiquetage et d'autres tâches (Cussens & Džeroski, 2000). Dans tous ces travaux, c'est principalement l'expressivité de la PLI qui est exploitée, notamment sa capacité à représenter simplement des problèmes relationnels, et à produire des règles interprétables. Cette capacité à traiter des problèmes impliquant des représentations complexes se traduisent en revanche souvent par des complexités calculatoires prohibitives pour travailler sur de grandes quantités de données. Il est donc important de bien contrôler ces aspects, comme nous le décrivons en section 3.3.

3 Utilisation de la programmation logique inductive

Dans cette section, nous présentons brièvement les principes de la PLI ; nous invitons le lecteur intéressé à consulter (Muggleton & Raedt, 1994) pour une description étendue. Nous détaillons ensuite l'intérêt de cette méthode pour l'indexation MeSH et décrivons comment nous l'avons adaptée pour prendre en compte les spécificités de notre problème.

3.1 Principes de la PLI

La PLI est une technique d'apprentissage symbolique supervisée permettant d'inférer des règles, exprimées sous forme de clauses logiques (clauses Prolog), à partir d'exemples, eux aussi décrits en Prolog. Formellement, étant donnés les ensembles d'exemples E^+ et de contre-exemples E^-, un ensemble d'informations sur le monde B (*background knowledge*), le but d'un programme de PLI est de trouver un ensemble de règles H tel que ($\square$ représente faux) : $B \wedge H \wedge E^- \not\models \square$ et $B \wedge H \models E^+$.

En pratique, ces deux conditions sont assouplies : on cherche un ensemble de règles couvrant (expliquant) la plupart des exemples positifs et rejetant la plupart des exemples négatifs. Cela permet de travailler avec des données imparfaites ou bruitées – ce qui est souvent le cas avec des

problèmes réels – et d'obtenir des règles plus génériques quitte à accepter quelques exceptions.

La plupart des algorithmes de PLI abordent l'inférence de règles comme un problème de recherche dans un espace d'hypothèses (noté $\mathscr{E}_H$). L'algorithme 1 illustre ce type de fonctionnement ; c'est ce même algorithme qui est au cœur d'ALEPH (Srinivasan, 2001), l'outil de PLI que nous utilisons pour nos expériences. L'organisation et le parcours de l'espace $\mathscr{E}_H$ est un point crucial pour l'efficacité de la phase d'inférence (*cf.* section 3.3). La fonction de score Sc prend généralement en compte le nombre d'exemples positifs et négatifs couverts (respectivement P et N) pour juger de la qualité d'une hypothèse ; dans notre cas, $Sc = \frac{P}{P+N}$.

Algorithme 1 Algorithme d'ALEPH

Itération jusqu'à $E^+ = \emptyset$

1. choisir aléatoirement un exemple positif e^+ dans E^+ ;
2. construire la clause la plus spécifique $\perp$ couvrant e^+ ;
3. générer et parcourir l'espace de recherche $\mathscr{E}_H$ basé sur $\perp$ à la recherche de la clause h maximisant une fonction de score Sc ;
4. ajouter h à l'ensemble H et ôter de E^+ les exemples couverts par h.

Fin itération

3.2 Description des exemples

Dans notre cadre, la PLI doit permettre d'inférer des règles d'adjonction de qualificatifs à partir d'exemples d'articles recensés dans MEDLINE et préalablement annotés à l'aide de descripteurs MeSH. Pour un qualificatif donné, on cherche à obtenir les règles indiquant à quel mot-clé il doit être associé selon le contexte de ce mot-clé.

Ce problème peut être vu comme un simple problème de recherche de règles d'association pour lequel il existe des algorithmes qui peuvent sembler moins lourds que la PLI (Agrawal *et al.*, 1993). Cependant, le choix de la PLI pour ce problème n'est pas un hasard. En effet, une technique de recherche de règle d'association (*e.g.* APRIORI) nécessiterait de décrire chaque exemple par un tuple indiquant quel mot-clé est examiné (celui pour lequel on peut vouloir ajouter le qualificatif étudié) et quels autres mots-clés apparaissent dans le reste de la description. Sans autre subtilité de représentation, cela produirait un tuple de dimension 24 000 x 24 000. La recherche de représentations plus légères se heurte tout de même à deux problèmes (Soualmia, 2004) : le nombre variable de mots-clés par article et la prise en compte des relations hiérarchiques entre mots-clés. Ces deux problèmes propres aux méthodes propositionnelles (*i.e.* décrivant les problèmes sous forme de tuples attributs-valeurs) sont naturellement résolus par l'emploi de la PLI.

En effet, grâce à l'expressivté de Prolog, la PLI permet de décrire de tels problèmes relationnels simplement. Le tableau 1 présente par exemple la notice d'un article et son encodage en Prolog. Un exemple positif est une occurrence d'un mot-clé associé au qualificatif dont nous cherchons à modéliser l'utilisation. Supposons que l'on s'intéresse aux règles d'indexation impliquant le qualificatif *pharmacology* ; d'après la notice du tableau 1, on ajoute à E^+ l'exemple : qualif(mh16179550_1,"pharmacology"). Les exemples négatifs sont des occurrences de mots-clés auxquels le qualificatif recherché n'est pas associé bien qu'il soit autorisé pour ces mots-clés.

Extrait de la notice d'un article	Extrait de l'encodage de l'article dans B
PMID - 16179550	in_article(pmid16179550,mh16179550_1).
MH - Acrylamide/*pharmacology	hierarchy(mh16179550_1,"Acrylamide").
MH - Animals	in_article(pmid16179550,mh16179550_2).
MH - Astrocytes/drug effects	hierarchy(mh16179550_2,"Animals").
MH - Histidine/physiology	in_article(pmid16179550,mh16179550_3).
MH - Amino Acid Transport	hierarchy(mh16179550_3,"Astrocytes").
Systems/*biosynthesis	in_article(pmid16179550,mh16179550_4).
MH - Cells, Cultured	hierarchy(mh16179550_4,"Histidine").
...	...

TAB. 1 – Description des exemples en PLI

Dans le *background knowledge*, les informations sur le contexte de cette occurrence, c'est-à-dire les autres mots-clés associés au même article sont données comme illustré dans le tableau 1: la première ligne indique qu'un mot-clé identifié par mh16179550_1 est assigné à l'article pmid16179550 à l'aide du prédicat in_article/2 ; la deuxième ligne indique que ce mot-clé correspond au terme MeSH "Acrylamide" avec le prédicat hierarchy/2. Enfin, la hiérarchie MeSH est également transcrite dans le *background knowledge* sous la forme :

```
hierarchy(W,"Neuroglia") :- hierarchy(W,"Astrocytes").
hierarchy(W,"Nervous System") :- hierarchy(W,"Neuroglia").
hierarchy(W,"Anatomy") :- hierarchy(W,"Nervous System").
...
```

Cela permet d'indiquer que le mot-clé *Astrocytes* est un descendant (concept plus spécifique) du mot-clé *Neuroglia* dans la hiérarchie MeSH, lui-même descendant de *Nervous System*... Toutes les informations de hiérarchie du MeSH sont ainsi exploitables par le processus d'inférence.

La phase de construction des exemples, nécessaire pour notre technique supervisée, est donc entièrement automatique. Les règles qui en sont inférées manipulent les différents prédicats que nous venons de décrire. Voici un exemple de règle qui pourrait être inférée à partir de l'exemple qualif(mh16179550_1,"pharmacology"). et qui correspond à la règle donnée en introduction :

```
qualif(W,"pharmacology") :- hierarchy(W," Carboxylic Acids"), in_article(A,W), in_article(A,W), hierarchy(W,"Anatomy").
```

3.3 Efficacité de l'inférence

La contrepartie à l'expressivité de la PLI est la grande complexité de l'inférence des règles. L'espace de recherche $\mathscr{E}_H$ est généralement très grand, voire infini dans certains cas. Par ailleurs, le calcul du score Sc repose sur le nombre d'exemples positifs et négatifs couverts par l'hypothèse examinée, ce qui implique de confronter tous les exemples à la clause. C'est cette partie qui est la plus coûteuse en PLI. Il est donc essentiel de bien contrôler cette phase de recherche pour s'assurer à la fois d'un temps de calcul praticable et de la production de règles pertinentes pour notre problème. Heureusement, $\mathscr{E}_H$ peut être organisé et parcouru hiérarchiquement ; la relation d'ordre usuellement utilisée est la θ-subsomption.

Définition 1 (θ-subsomption) *Une clause C_1 θ-subsume une clause C_2 ($C_1 \succeq_\theta C_2$) si et seulement si (ssi) il existe une substitution θ telle que $C_1\theta \subseteq C_2$ (en considérant les clauses comme des ensembles de littéraux).*

Par exemple, la clause $p(a,b) \leftarrow r(b,a)$ est subsumée par la clause $p(Y_1,Y_2) \leftarrow r(Y_2,Y_1)$. En effet, on a bien $\{p(Y_1,Y_2), \neg r(Y_2,Y_1)\}\theta_1 \subseteq \{p(a,b), \neg r(b,a)\}$ avec $\theta_1 = \{Y_1/a, Y_2/b\}$. Grâce à cette relation hiérarchique, l'espace peut être généré et exploré efficacement, par exemple de la clause la plus générale à la plus spécifique selon la θ-subsomption. À partir d'une clause C_1, on peut générer des clauses plus spécifiques qui ont notamment pour propriété de ne couvrir que des sous-ensembles des exemples couverts par C_1. Le calcul du score des clauses plus spécifiques que C_1 s'en trouve grandement accéléré. De même, on peut éviter de générer une clause que l'on considère comme ne couvrant pas assez d'exemples dès lors que son ancêtre ne couvre lui-même pas assez d'exemples.

Cependant, la θ-subsomption n'est pas parfaitement adaptée à notre problème car la hiérarchie entre termes MeSH n'est pas prise en compte. Ainsi, la clause qualif(W1,"genetics") :- in_article(A,W1), in_article(A,W2), hierarchy(W2,"Frameshift Mutation") n'est pas subsumée par qualif(W1,"genetics") :- in_article(A,W1), in_article(A,W2), hierarchy(W2,"Mutation") bien que *Mutation* soit un ancêtre de *Frameshift Mutation* dans la hiérarchie MeSH. Nous proposons donc une autre notion de subsomption en nous inspirant des travaux de Buntine (1988) :

Définition 2 *Une clause C_1 θ_{hier}-subsume une clause C_2 ($C_1 \succeq_{hier} C_2$) relativement au background knowledge B ssi il existe une substitution θ et une fonction f_D telles que $f_D(C)\theta \subseteq D$, où f_D est telle que $\forall l \in C, \quad B, f_D(l) \models l$, avec $f_D(\{l_1, l_2, ..., l_m\})$ signifiant $\{f_D(l_1), f_D(l_2), ..., f_D(l_m)\}$.*

La fonction f permet donc de prendre en compte les informations du *background knowledge*, c'est-à-dire l'organisation hiérarchique des mots-clés du MeSH. Cette subsomption, que nous avons mise en œuvre en modifiant ALEPH, est cohérente avec la notion de couverture. Il est aisé de montrer qu'elle induit un ordre partiel entre les clauses et structure donc l'espace $\mathscr{E}_H$ en treillis, ce qui facilitera son parcours. Parfaitement adaptée à notre problème, elle va nous permettre d'inférer efficacement des règles d'indexation à partir de grandes quantités d'exemples et contre-exemples (*cf.* section 5) en évitant l'impasse calculatoire qui aurait consisté à considérer indépendamment les 24 000 mots-clés du MeSH. Il est intéressant de noter que cette subsomption est en fait adaptée à tous les problèmes de recherche de régularités impliquant des connaissances structurées (*e.g.* ontologies).

4 Contexte expérimental

4.1 Base de comparaison

Comme nous l'avons expliqué, les méthodes simples de recherche de règles d'association ne peuvent pas traiter directement nos données. Pour néanmoins donner une base de comparaison aux performances de notre approche, nous implémentons une simple *baseline*. Celle-ci consiste à former aléatoirement des paires mot-clé/qualificatif en respectant la distribution statistique de l'ensemble des paires dans la base MEDLINE. Ainsi, nous avons estimé que la probabilité d'apparition P d'une paire mot-clé/qualificatif était égale au nombre d'occurrences de cette paire dans MEDLINE divisé par le nombre total d'occurrences du mot-clé dans MEDLINE, qu'il soit affilié à un qualificatif ou non. Le tableau 2 présente par exemple la distribution complète du mot-clé *Irritable Mood* et un extrait de la distribution du mot-clé *Lung*.

Irritable Mood 5 qualificatifs affiliables		*Lung* (extrait) 26 qualificatifs affiliables	
Qualificatif	***P***	**Qualificatif**	***P***
sans qualificatif	0.758	sans qualificatif	0.047
classification	0.008	classification	0
drug effect	0.129	drug effect	0.082
ethics	0	metabolism	0.117
physiology	0.101	pathology	0.171
radiation effect	0.004	radiation effect	0.011

TAB. 2 – Extrait de la distribution des descripteurs MeSH dans MEDLINE

4.2 Corpus de travail

Pour l'apprentissage des règles par PLI, nous avons utilisé un corpus composé de 100 000 citations choisies aléatoirement parmi celles ajoutées à la base MEDLINE en 2006. Pour la méthode *baseline*, nous avons utilisé les 15 433 668 citations contenues dans la base MEDLINE fin 2005[3] afin de calculer la distribution statistique des paires de descripteurs MeSH. Finalement, nous avons également utilisé un corpus de test composé de 100 000 citations choisies aléatoirement parmi celles ajoutées à la base MEDLINE en 2006. Ce corpus est entièrement disjoint des deux corpus d'apprentissage mentionnés ci-dessus.

4.3 Évaluation des règles obtenues dans un contexte réel

En pratique, on voit d'après l'exemple de règle d'indexation donné en introduction que des mots-clés isolés servent de point de départ à la recommandation de paires. Pour l'anglais, le logiciel MTI (Aronson *et al.*, 2004) développé à la NLM permet de générer automatiquement des recommandations d'indexation pour des mots-clés isolés. Il faut remarquer que les mots-clés recommandés par MTI ont, sur le corpus de test, un rappel de 45 % par rapport aux termes sélectionnés par les indexeurs MEDLINE. Ainsi, un écart est attendu entre les performances théoriques des règles d'indexation obtenues par PLI et les performances obtenues à partir de l'application de ces mêmes règles sur les mots-clés isolés recommandés par MTI, c'est-à-dire dans un environnement pratique d'indexation automatique. Les règles d'indexation obtenues par PLI sont évaluées sur le corpus de test, l'indexation MEDLINE servant de référence. Ces performances sont comparées à celles de règles d'indexation similaires obtenues manuellement par un expert du domaine et à celles de la *baseline* décrite en 4.1.

Pour chacune des méthodes (PLI, manuelle, *baseline*) seules les paires constituées à partir de mots-clés extraits par MTI et figurant dans MEDLINE sont prises en compte. Les mesures de performances utilisées sont la précision, le rappel et la F-mesure. La précision correspond au nombre de paires mot-clé/qualificatif recommandées utilisées dans MEDLINE divisé par le nombre total de paires recommandées. Le rappel correspond au nombre de paires recommandées utilisées dans MEDLINE divisé par le nombre total de paires figurant dans l'indexation MEDLINE. La F-mesure correspond à la moyenne harmonique de la précision et du rappel.

[3]Voir `http://mbr.nlm.nih.gov/Reference/medline_Baseline_Repository_Detail.pdf` pour plus de détails

5 Résultats

Nous ne présentons ici que les résultats obtenus sur quelques qualificatifs, notamment ceux pour lesquels il existe des règles manuelles qui nous permettent d'effectuer une comparaison. Ces résultats sont discutés dans la section suivante.

Le tableau 3 présente les performances théoriques obtenues en appliquant les règles PLI sur les mots clés contenus dans les notices MEDLINE. Nous indiquons également les temps de calcul obtenus sur une machine de bureau (linux Intel Xeon 3 GHz) au regard du nombre d'exemples utilisés.

Qualificatif	$\|E^+\|$	$\|E^-\|$	**Temps de calcul**	**Précision (%)**	**Rappel (%)**	**F-mesure (%)**
Administration & dosage	5 300	40 000	75 mn	41.4	53	46.5
Metabolism	4 500	21 000	37 mn	42.4	60.2	49.7
Pharmacology	5 000	22 000	45 mn	48.8	53.9	51.2
Physiology	5 200	34 000	46 mn	41.4	41.5	41.5

TAB. 3 – Performance de l'inférence de règles par PLI sur les notices complètes

Le tableau 4 présente les performances sur le corpus de test des règles obtenues par apprentissage, comparées aux règles manuelles et à la *baseline* dans le cas où les mots-clés sont extraits automatiquement par MTI.

Qualificatif	**Méthode**	**Nb. règles**	**Précision (%)**	**Rappel(%)**	**F-mesure(%)**
Administration & dosage	PLI	166	38	**29**	**33**
	Manuelle	1	**54**	1	1
	Baseline	-	26	9	13
Metabolism	PLI	134	49	**38**	**43**
	Manuelle	61	**58**	20	30
	Baseline	-	37	12	18
Pharmacology	PLI	217	47	**28**	**35**
	Manuelle	7	**67**	3	5
	Baseline	-	28	12	17
Physiology	PLI	70	**46**	**24**	**32**
	Manuelle	0	-	-	-
	Baseline	-	28	10	15

TAB. 4 – Performance des méthodes sur le corpus de test en conjonction avec MTI

6 Discussion

Performances. Comme attendu, l'utilisation de l'outil MTI pour produire dans un premier temps des recommandations de mots-clés fait que le rappel en situation réelle des règles inférées par PLI est inférieur à celui mesuré sur les notices. Malgré cela, les performances des règles PLI sont bien supérieures à celles de la *baseline* et produisent toujours la meilleure F-mesure

avec une différence de précision modérée par rapport aux règles manuelles. Ces dernières sont beaucoup plus ciblées, ce qui produit de très bons taux de précision mais des rappels généralement faibles, inférieurs non seulement à la PLI mais aussi à la *baseline* (sauf dans le cas de *metabolism*). Dans le cadre d'une utilisation combinée avec d'autres méthodes d'indexation, il est probable que le gain significatif en rappel apporté par les règles PLI contribue à une amélioration globale du rappel des recommandations sans perte de précision. Par ailleurs, grâce à notre notion de subsomption adaptée au problème, les temps de calcul de la phase d'inférence sont tout à fait raisonnables, notamment au regard des temps de développement de règles manuelles. Sans la modification de subsomption, des expériences préliminaires ont montré que la PLI ne pouvait guère manipuler ces quantités d'information (temps de calcul dépassant plusieurs jours, dépassement mémoire).

Règles PLI *vs.* manuelles. Avec les modifications apportées, la PLI permet donc d'obtenir rapidement un grand nombre de règles. En examinant ces règles, il apparaît parfois que des règles un peu différentes obtenant des performances proches semblent sémantiquement meilleures pour un expert du domaine. De même, certaines règles PLI peuvent permettre à un expert de créer manuellement une nouvelle règle qui n'aurait pas pu être inférée à partir du corpus d'entraînement. Ainsi, il pourrait être possible d'optimiser la production de règles d'indexation en envisageant une relecture des règles PLI par un expert afin d'améliorer la lisibilité et les performances des règles tout en minimisant le temps passé à la production des règles.

MTI. Outre la baisse globale de rappel observée avec l'application des règles PLI sur les mots-clés extraits par MTI, on constate en observant les performances individuelles de chaque règle que les variations sont hétérogènes en fonction des mots-clés mis en jeu ; ce qui reflète les performances de MTI pour l'extraction des mots-clés. Ainsi, il est envisageable de filtrer automatiquement les règles les moins performantes dans notre cadre applicatif précis.

7 Conclusion et perspectives

Nous avons montré qu'il était possible d'exploiter des régularités dans les descriptions des articles biomédicaux indexés dans MEDLINE pour proposer automatiquement des qualificatifs à affilier aux mots-clés lors du traitement d'un nouvel article. L'automatisation de cette tâche jusqu'alors manuelle — problème peu abordé auparavant — a été possible par l'utilisation de la PLI et sa capacité à gérer des problèmes impliquant des représentations complexes. Il a néanmoins été nécessaire d'implémenter une notion de subsomption entre clauses permettant de prendre en compte les spécificités de la tâche ; cela nous a permis d'obtenir des temps de calcul praticables tout en traitant un nombre de données suffisamment important pour obtenir des résultats fiables. Cette approche est d'ailleurs transposable à d'autres problèmes impliquant des recherche de régularités dans des connaissances organisées (thésaurus, ontologies...). Enfin, ces résultats sont d'autant plus remarquables qu'ils n'exploitent que des régularités implicites des descriptions et non le contenu de l'article et permettent ainsi d'envisager l'utilisation d'autres stratégies de recommandations de qualificatifs basées cette fois sur le corps de l'article à indexer.

Plusieurs perspectives sont envisagées à la suite de ce travail. Tout d'abord, il sera nécessaire de l'étendre à l'ensemble des 83 qualificatifs afin d'obtenir un jeu complet de règles d'indexation. Par ailleurs, nous envisageons une amélioration des performances et de la lisibilité grâce à un filtrage automatique ou à une relecture manuelle comme indiqué en 6. Enfin, nous anticipons l'intégration du jeu de règles obtenues au module d'affiliation de qualificatifs de MTI.

Remerciements

Ce travail a été effectué dans le cadre de la participation d'A. Névéol au programme de recherche postdoctorale de la National Library of Medicine administré par ORISE. Les auteurs remercient A. Aronson, J. Mork et S. Shooshan pour de nombreuses discussions sur les aspects théoriques et techniques de cette étude.

Références

AGRAWAL R., IMIELINSKI T. & SWAMI A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the SIGMOD Conference*, Washington, États-Unis.

ARONSON A. R., MORK J. G., GAY C. W., HUMPHREY S. M. & ROGERS W. J. (2004). The nlm indexing initiative's medical text indexer. In *Proc. Medinfo 2004*, p. 268–272.

BUNTINE W. L. (1988). Generalized Subsumption and its Application to Induction and Redundancy. *Artificial Intelligence*, **36**, 375–399.

CLAVEAU V., SÉBILLOT P., FABRE C. & BOUILLON P. (2003). Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming. *Journal of Machine Learning Research (JMLR), special issue on ILP*, **4**, 493–525.

J. CUSSENS & S. DŽEROSKI, Eds. (2000). *Learning Language in Logic*. Lecture Notes in Artificial Intelligence. Sringer Verlag.

LIN J. & WILBUR W. J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, **8**(1), 423.

MARKÓ K., DAUMKE P., SCHULZ S. & HAHN U. (2003). Cross-language MeSH indexing using morpho-semantic normalization. In *Actes de AMIA Symp. 2003*, p. 425–429.

MUGGLETON S. & RAEDT L. D. (1994). Inductive logic programming: Theory and methods. *Journal of Logic Programming*, **19/20**, 629–679.

NÉVÉOL A., ROGOZAN A. & DARMONI S. J. (2006). Automatic indexing of online health resources for a french quality controlled gateway. *Information Processing and Management*, p. 695–709.

NÉVÉOL A., SHOOSHAN S. E., RINDFLESH T. C. & ARONSON A. R. (2007). Multiple approaches to fine-grained indexing of the biomedical literature. In *Actes de PSB 2007*, p. 292–303.

OZDOWSKA S. & CLAVEAU V. (2006). Inférence de règles de propagation syntaxique pour l'alignement de mots. *TAL (Traitement Automatique des Langues)*, **47**(1), 167–186.

RAK R., KURGAN L. A. & REFORMAT M. (2007). Multilabel associative classification categorization of MEDLINE articles into MeSH keywords. *IEEE Eng Med Biol Mag*, **2**(26), 47–55.

SOUALMIA L. F. (2004). *Étude et évaluation d'approches multiples de projection de requêtes pour une recherche d'information intelligente. Application au domaine de la Santé sur l'Internet*. Thèse de doctorat, INSA de Rouen.

SRINIVASAN A. (2001). *The* ALEPH *manual*. Machine Learning at the Computing Laboratory, Oxford University.