Indexation automatique de ressources de santé à l'aide de *paires* de descripteurs MeSH

Aurélie Névéol[1, 2] Alexandrina Rogozan[1] Stéfan J. Darmoni[1, 2]

[1]Laboratoire PSI – FRE 2645 CNRS Université et INSA de Rouen
{aneveol, arogozan}@insa-rouen.fr
[2]Equipe CISMeF, CHU de Rouen - 1, rue de Germont - 76031 Rouen
stefan.darmoni@univ-rouen.fr

Mots-clés: Indexation Automatique, Terminologie Médicale, Vocabulaire Contrôlé.

Keywords: Automatic Indexing, Medical Terminology, Controlled Vocabulary.

Résumé: Depuis quelques années, médecins et documentalistes doivent faire face à une demande croissante dans le domaine du codage médico-économique et de l'indexation des diverses sources d'information disponibles dans le domaine de la santé. Il est donc nécessaire de développer des outils d'indexation automatique qui réduisent les délais d'indexation et facilitent l'accès aux ressources médicales. Nous proposons deux méthodes d'indexation automatique de ressources de santé à l'aide de paires de descripteurs MeSH. La combinaison de ces deux méthodes permet d'optimiser les résulats en exploitant la complémentarité des approches. Les performances obtenues sont équivalentes à celles des outils de la littérature pour une indexation à l'aide de descripteurs seuls.

Abstract: The increasing number of health documents available in electronic form, and the demand on both practitioners and librarians to encode these documents with controlled vocabulary information calls for automatic tools and methods to help them perform this task efficiently. In this article, we are presenting and comparing two methods for the automatic indexing of heath resources with pairs of MeSH descriptors. A combination of both methods achieves better results by exploiting the complementarity of the approaches. This performance matches the tools described in the litterature for single term indexing.

Introduction

Depuis quelques années, le nombre de documents électroniques augmente de manière exponentielle. Se pose alors la question de l'exploitation efficace de ces ressources. Dans les domaines de la santé et de la bio-médecine, de nombreux travaux ont été entrepris afin de guider les utilisateurs dans leur recherche d'information. Ainsi, la base documentaire MEDLINE® recense 11 millions d'articles scientifiques en langue anglaise indexés à l'aide du thésaurus MeSH® (Medical Subject Headings) développé et maintenu par la NLM

(National Library of Medicine). En Europe, la fondation HON (Health On the Net-http://www.hon.ch/) ou le Catalogue et Index des Sites Médicaux Francophones (CISMeF-http://www.cismef.org) se proposent de guider les internautes vers une information médicale de qualité. Dans ces deux projets européens, la description des ressources[1] s'appuie sur la version française du MeSH. Notre objectif est de formaliser cette démarche et de proposer une méthode d'indexation automatique en adéquation avec les caractéristiques de l'indexation manuelle. Nous présentons deux méthodes répondant à ces critères. Après les avoir évaluées séparément puis combinées, nous discutons de l'intérêt d'utiliser de tels outils dans le cadre de l'aide à l'indexation.

1 Méthodes d'indexation

1.1 Indexation Manuelle

Les principes de l'indexation manuelle à l'aide de descripteurs MeSH (mots clés et qualificatifs) sont clairement exposés dans (Dailland et al., 2003). Les caractéristiques de cette indexation sont: (a) l'utilisation de descripteurs obligatoires (des mots clés particuliers, par exemple <*sujet âgé*>), (b) l'association de qualificatifs pour préciser les mots clés le cas échéant (par exemple, la paire <*diabète/chimiothérapie*> sera utilisée de préférence au mot clé isolé <*diabète*> pour évoquer les traitements médicamenteux du diabète) et (c) l'adaptation du nombre de mots clés (ou de paires) utilisés pour indexer une ressource en fonction de son contenu (par exemple, dans le catalogue CISMeF, les ressources sont indexées avec un nombre de mots clés (ou paires) pouvant aller de zéro à plusieurs dizaines). De manière plus générale, on retiendra également qu'il existe un consensus entre les experts (Anderson et Pérez-Carballo, 2001) selon lequel l'indexation manuelle s'effectue en deux étapes: 1- l'analyse du document, afin d'en retirer le contenu et 2- la traduction de ce contenu dans le langage retenu pour la description (par exemple, le MeSH). En pratique, une troisième étape de relecture et de révision peut également être ajoutée.

1.2 Indexation Automatique

1.2.1 Méthode de Traitement Automatique de la Langue Naturelle (TALN)

Cette méthode, détaillée dans (Névéol, 2004), suit les étapes de l'indexation manuelle (analyse, traduction et révision). Une analyse de surface utilisant un dictionnaire MeSH et une bibliothèque de graphes permet de reconnaître les différentes formes prises par les concepts médicaux (flexions, synonymes, hyperonymes, …) et de les comptabiliser afin d'attribuer un score à chaque concept. Les informations nécessaires à la traduction des concepts sous forme MeSH sont contenues dans le dictionnaire (pour les mots clés) et dans les graphes (pour les paires). Les relations hiérarchiques entre termes MeSH permettent de réduire la liste des candidats en reportant les occurrences des mots clés pères vers leurs fils afin d'indexer au

[1] Afin de rendre compte de la multiplicité des documents électroniques indexés tant du point de vue du format, du type de document, que des usages auxquels ils sont destinés, nous utiliserons le terme de "ressource".

plus précis. L'application de règles d'indexation dans un deuxième temps permet de réviser la liste des candidats avant d'obtenir l'indexation finale avec la fonction de rupture décrite en 1.2.3. Les règles d'indexation sont de deux types :

1. Des **règles d'indexation issues de la NLM** préconisant l'utilisation d'un mot clé MeSH de préférence à une paire mot clé/qualificatif pour représenter un concept. La règle « $MC_1/Q_1 \Rightarrow MC_2$ » indique qu'il convient de remplacer la paire MC_1/Q_1 par le seul mot clé MC_2. Par exemple:<*cœur/transplantation*> => <*transplantation cardiaque*>

2. Des **règles d'indexation CISMeF** préconisant l'introduction d'un mot clé (ou paire). Ainsi, la règle « $MC_1/Q_1^2 + MC_2/Q_2^2$ » indique qu'il convient d'ajouter la paire MC_2/Q_2 à l'indexation d'une ressource déjà indexée avec la paire MC_1/Q_1. Par exemple : <*appendicectomie*> + <*appendicite/chirurgie*>

1.2.2 Méthode des k plus proches voisins (k-PPV)

La méthode des *k*-PPV est une référence dans le domaine de la classification. Son principe est simple. Soit *C* une collection de ressources étiquetées notées r_i. Soit $r \notin C$ une ressource à étiqueter. On calcule la similarité $s(r, r_i)$ de *r* à chaque ressource r_i de *C*, et on sélectionne ses *k* plus proches voisins, c'est à dire $r_1, \ldots r_k$ tels que $s(r, r_i)$ soit maximum pour $i=1, \ldots k$. La similarité $s(r, r_i)$ entre deux ressources correspond au nombre de mots pleins en commun entre le titre de *r* et celui de r_i. Chaque ressource est représentée par un sac de mots issu du titre après filtrage des mots grammaticaux à l'aide d'un anti-lexique (Salton et Mc Gill, 1983). Dans le cas où on cherche à étiqueter les ressources avec une seule classe, la nouvelle ressource peut être étiquetée avec la classe dominante parmi ses *k* plus proches voisins. Dans le cadre de l'indexation, les classes avec lesquelles nous allons étiqueter les ressources sont des mots clés ou des paires mot clé/qualificatif MeSH. De plus, l'indexation d'une ressource doit être composée d'un nombre *a-priori* inconnu de mots clés (ou paires). Nous associons donc à *r* une liste de candidats MeSH auxquels un score S (compris entre 1 et *k*) est attribué en fonction du vote des *k* voisins. Le choix d'une valeur de *k* est évoqué à la section 2. La fonction de rupture décrite en 1.2.3 permet de sélectionner les candidats retenus pour l'indexation finale.

1.2.3 Fonction de rupture

Soit *N* le nombre de mots clés (ou paires) candidats à l'indexation extraits à l'aide de l'une des méthodes ci-dessus. Soit S_i le score attribué au *i*-ème candidat. On suppose que les candidats sont classés par ordre de scores décroissants, de sorte que $S_1 > \ldots > S_i > \ldots > S_N$. Pour $i = 1, \ldots, N-1$, on calcule $F = \frac{S_i - S_{i+1}}{S_i + S_{i+1}}$. Le seuil retenu sera *i* tel que F soit maximum.

---

[2] On considère ici que les qualificatifs peuvent être « vides », c'est à dire que les règles CISMeF peuvent statuer sur des mots clés seuls ou associés à un qualificatif.

1.2.4 Méthodes Mixtes

Afin d'évaluer la complémentarité des méthodes présentées ci-dessus (1.2.1 et 1.2.2), nous avons combiné les indexations obtenues selon deux procédés, l'un prenant en compte le rang des mots clés (ou paires) résultant de l'indexation TALN et k-PPV, l'autre prenant en compte le score attribué aux mots clés (ou paires) par ces méthodes. Ainsi, à chaque mot clé (ou paire) est attribué un nouveau score égal à la somme de ses rangs (resp. scores relatifs) dans les deux méthodes. Si un mot clé n'a été extrait que par une seule méthode, son rang (resp. score relatif) pour la deuxième méthode est considéré comme nul. Les mots clés (ou paires) sont ensuite classés par score décroissant. Dans les deux cas, nous avons placé en tête du classement les mots clé (ou paires) extraits conjointement par les deux méthodes.

2 Expérimentation

Le corpus d'évaluation utilisé est composé de 82 ressources extraites aléatoirement du catalogue CISMeF. Chaque ressource du corpus a été indexée automatiquement à l'aide des méthodes TALN et k-PPV successivement. Les résultats obtenus par les deux méthodes ont ensuite été combinés comme indiqué en 1.2.4. Dans chaque cas, l'indexation automatique obtenue a été comparée à l'indexation manuelle de référence. Les performances sont évaluées avec les mesures habituelles de précision et de rappel, ainsi que la F-measure qui combine ces deux dernières. Un poids équivalent est alors accordé à la précision et au rappel (Manning et Schütze, 1999). La Figure 1 présente la F-measure en fonction du rang pour chacune des méthodes.

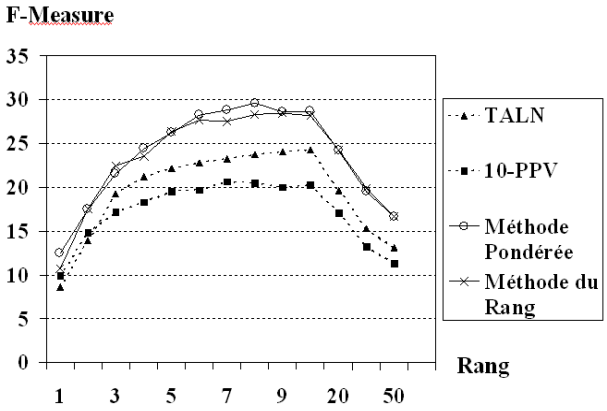

Figure 1 : Courbes F-measure en fonction du rang

Le Tableau 1 présente la précision et le rappel obtenus pour chaque méthode séparément (colonnes 1 à 3), et pour la combinaison des méthodes fondée sur le rang et sur les scores (colonnes 4 et 5). Pour la méthode des k-PPV, plusieurs valeurs de k ont été testées (k=1, 3, 5, 10, 15) et les meilleurs résultats ont été observés avec k=10, bien que pour certaines ressources, il ne soit pas possible de trouver dix voisins (il arrive même qu'aucun voisin ne soit trouvé). Dans ce cas, les valeurs de précision et de rappel de la méthode sur une telle ressource sont considérées comme nulles. Dans la troisième colonne (N=73) nous donnons les performances obtenues pour les soixante treize ressources pour lesquelles au moins dix voisins ont été trouvés.

Rang	TALN	10-PPV (N=82)	10-PPV (N=73)	Mixte Rang	Mixte Score
	P - R	**P - R**	**P - R**	**P - R**	**P - R**
1	36 – 5	51 - 6	57 - 6	51 - 6	49 - 6
4	32 – 16	30 – 13	34 – 15	38 - 17	36 - 18
10	22 – 27	20 – 21	22 – 23	26 - 31	27 - 33
50	8 – 40	7 – 35	8 - 40	10 - 53	10 - 53
T	**27 - 21** (T_{moyen}**=12)**	**32 – 16** (T_{moyen} **=5)**	**36 – 18** (T_{moyen} **=5)**	***34 – 24*** (T_{moyen} ***=9)***	***32 – 25*** (T_{moyen} ***=9)***

Tableau. 1 : Précision et Rappel de chaque méthode sur un corpus de 82 ressources

3 Discussion

3.1 Performances des différentes méthodes

D'après le Tableau 1 et la Figure 1, on peut remarquer que la combinaison des deux méthodes d'indexation présentées offre une précision supérieure ou égale à chacune des méthodes. Rappel et F-measure sont également meilleurs. Cela rejoint les conclusions de (Aronson et al. 2004) qui observent une amélioration des performances du système MTI lorsqu'un module statistique vient compléter le module de traitement linguistique. Les résultats de la combinaison des deux méthodes sont équivalents à ceux des extracteurs MeSH francophones de la littérature (Névéol et al. 2005). Cependant, le système que nous présentons a l'avantage d'indexer à l'aide de *paires* mot clé/qualificatifs alors que les autres sytèmes existants – par exemple, (Pouliquen, 2002) ou (Gaudinat et al., 2002), extraient des termes isolés. La fonction de rupture est efficace dans la mesure où, la précision au seuil T est généralement plus élevée que la précision obtenue au rang fixe équivalent (par exemple, la précision est de 32 à T=5 contre 29 au rang fixe 5 pour la méthode k-PPV). Pour les méthodes combinées, les chiffres sur le seuil (en italique) sont donnés à titre indicatif, et ont été obtenus en effectuant une moyenne entre précision et rappel aux seuils pour les méthodes TALN et k-PPV.

3.2 Complémentarité des méthodes

Le dictionnaire MeSH utilisé par la méthode TALN permet de ne pas limiter l'indexation aux termes MeSH déjà présents dans la base de ressources indexées (~11.000 termes MeSH sur près de 23.000). L'approche k-PPV utilise une indexation manuelle de référence, et permet ainsi de proposer une indexation cohérente avec la base existante. De plus, elle peut extraire des mots clés qui sont pour l'instant difficilement repérables avec la méthode TALN, comme <*étude comparative*>. En effet, ces termes apparaissent rarement dans le texte des ressources et doivent être déduits d'une analyse globale de la ressource. Par exemple, un article faisant état d'une même étude épidémiologique réalisée en France et au Canada pourra être indexé avec le mot clé <*étude comparative*> bien que celui-ci ne soit jamais explicitement utilisé. Ainsi, dès les premiers rangs, le rappel obtenu avec la fusion des méthodes est supérieur au rappel obtenu avec l'une des méthodes utilisée séparément. Cependant, la méthode des k-PPV

est fondée sur le titre des ressources, ce qui peut poser problème si le titre n'est pas suffisamment explicite, ou si la base comporte trop peu de ressources aux titres similaires. Ainsi, on peut observer un silence de la méthode (11% des cas sur le corpus utilisé) ou bien une indexation approximative. L'utilisation conjointe de la méthode TALN permet de minimiser l'impact de ces erreurs. Le fait de placer les termes communs aux deux méthodes en tête de l'indexation combinée permet de mettre en avant des termes qui avaient obtenu un score bas avec les deux méthodes.

4 Conclusion et perspectives

Nous avons présenté deux méthodes d'indexation automatique de documents médicaux inovantes dans la mesure où elles proposent une indexation à l'aide de *paires* de descripteurs MeSH. Ces approches ont été évaluées séparément, puis combinées sur un corpus de 82 ressources. Il apparaît que la fusion des deux approches offre des performances supérieures à chacune des méthodes seules. Pour la poursuite de ce travail, nous préparons une évaluation auprès d'indexeurs professionnels afin de déterminer si la révision de l'indexation automatique proposée par les systèmes combinés permet un gain de temps et/ou une réduction du silence de l'indexation manuelle.

Références

ANDERSON, JD., PÉREZ-CARBALLO, J. (2001) The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I. *IPM* 37(2), 231-254.

ARONSON AR., MORK JG., GAY CW., HUMPHREY SM., ROGERS WJ (2004). The NLM Indexing Initiative's Medical Text Indexer. Actes de *MEDINFO* 2004; 268-71.

DAILLAND, F. LEUTHEREAU, A. AND VALLEE, H. (2003). Aide mémoire d'indexation MeSH et F-MeSH pour le catalogage. Rapport Technique de l'INSERM. Bibliothèque de la Faculté de Médecine de Paris XI.

GAUDINAT, A., BOYER, C., BAUJARD, V., RUCH P. (2002) Evaluation de l'extraction de termes MeSH pour les systèmes de recherche d'information dans le domaine médical. Actes des JFIM.

MANNING, C., SHÜTZE, H. (1999) Foundations of Statistical NLP, MIT Press.

NEVEOL, A. (2004) Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. Actes de *RECITAL*. 2004;105-114.

NÉVÉOL, A., MARY V, GAUDINAT, A., ROGOZAN A., DARMONI SJ. (2005) Benchmark evaluation of the French MeSH indexing systems; soumis à AIME 2005.

POULIQUEN B. (2002) Indexation de textes médicaux par indexation de concepts, et ses utilisations, Thèse de Doctorat, Université Rennes 1.

SALTON, G., MC GILL, M.J. (1983) Introduction to Modern Information Retrieval. New York : McGraw-Hill.