Indexation Automatique Conceptuelle de Thèses Pharmaceutiques Françaises.

Vincent MARY †, Bruno POULIQUEN ‡, Franck LE DUFF †
Stefan J. DARMONI ≀, Alain SEGUI ∝, Pierre LE BEUX †
† *Laboratoire d'informatique médicale, Faculté de Médecine - Rennes, France*
‡ *European Commission, IPSC, Joint Research Centre Ispra, Italie*
≀ *Direction de l'Informatique et des Réseaux, CHU - Rouen, France*
∝ *Laboratoire de mathématiques et de physique pharmaceutique*
Faculté de Pharmacie - Rennes, France

Abstract. French pharmaceutical thesis are seldom refered to. If the main obstacles originate from language or access barriers, proper indexation could also be blaimed. Manually extracted key words dont necessary come from a structured thesaurus. In the following work, this manual indexing method is compared to an automated one, Nomindex, based on UMLS. The automated method is further improved by the addition of a relevance scoring system. The first step consists in downloading, adapting and indexing thesis in electronic format. Results will then be analysed, sorted by relevance, by comparing classic statistical indices (noise/silence/relevance). It was assumed that the manually obtained key words were always relevant. The manual silence is nevertheless high : 7 new key word are proposed by Nomindex, which results are mixed (10 Theses results are promising for a the first experience an pharmaceutical document without dictionary improvement : the indexing, if it is currently insufficient for a direct use, could easily be improve by specific updates of the dictionary

1 Introduction

En France, les mille thèses annuelles de pharmacie représentent un travail universitaire très important. Celui ci peut être bibliographique, de paillasse ou de recherche. Il est encadré et validé par des professeurs habilités à diriger des recherches. Il pourrait donc être repris au même titre qu'un publication classique. Or, bien que la pharmacie touche toutes les branches médicales, les thèses françaises de pharmacie ne sont que très peu citées dans les publications françaises, et à fortiori internationales, et ce pour plusieurs raisons. Outre la barrière de la langue, le catalogue des thèses était jusqu'à peu mis à jour tout les six mois. De plus, leur mise en ligne reste exceptionnelle, un système de prêt entre bibliothèque ayant été mis en place. Nous proposons ici l'utilisation d'une méthode d'indexation automatique, Nomindex [1], qui, outre une mise en ligne au format HTML, couple une indexation automatique conceptuelle, via le métathésaurus UMLS [2] [3], à une simplification de la recherche via le réseau sémantique de ce même métathésaurus, et l'élaboration de score de pertinence.

2 Objectif

Notre objectif principal est de comparer sur un corpus unique de thèses récupérées sur internet, les résultats du moteur automatique et ceux de l'indexation manuelle. Pour ce faire, nous récupèrerons, adapterons et indexerons des thèses au format électronique. Puis, nous comparerons les résultats, via les indices classiques de bruit, silence et pertinence.

3 Etat de l'art

3.1 Docthèse / Sudoc

Le catalogue français des thèses était il y a peu le système Docthèse. Il était mis à jour via un CD ROM tous les six mois, et était disponible dans les bibliothèques universitaires. Parallèlement, et en vu de son remplacement à été développé le Sudoc [1], géré par l'ABES [2]). Disponible en ligne, le Sudoc fournit des informations relatives aux thèses : le titre, l'auteur, le directeur de thèse, et les mots clefs, classiquement au nombre de quatre ou cinq. Les mots clefs ne sont pas forcement issue du MeSH [4] ou même d'un quelconque thésaurus. L'indexation est réalisée manuellement par le thésard, son directeur de thèse, et est validée par un bibliothécaire. La mise à jour du Sudoc se fait au grée de l'édition de nouvelles thèses

3.2 Nomindex

L'objectif du moteur d'indexation est de reconnaître des concepts figurant dans un texte, et d'utiliser ceux ci pour la création d'une base de données permettant de retrouver ces documents. Le moteur utilisé se sert d'un dictionnaire médicale, principalement basé sur l'ADM (Aide au Diagnostique Médical) [5]. Les termes ont été regroupés en concepts, en tenant compte des mots associés, composés, des préfixes et des suffixes. Chaque concept correspond à une notion du MeSH français. On récupère le code CUI de l'UMLS de ce terme MeSH, que l'on attribue finalement à la notion du dictionnaire. A chaque concept d'un document est attribué un score de pertinence appelé TFIDF [6][7], donnant de bons résultats. Ce score donne une importance au concept en fonction de sa fréquence dans le document (TF = Term Frequency) pondérée par la fréquence d'apparition du concept dans tout le corpus (IDF = Inverse Document Frequency). Apres une recherche, il est possible d'étendre celle ci de deux manières :

- aux concepts pères, concepts fils et concepts reliés, grâce au réseau sémantique de l'UMLS.
- aux documents dont les concepts possèdent des scores de pertinences proches. Au delà d'une extension sur quelques mots clefs, le moteur recherche des documents similaires sur l'ensemble des concepts.

[1] Système Universitaire de Documentation
[2] Agence Bibliographique de l'Enseignement Supérieur : `http://corail.abes.sudoc.fr`

4 Materiel and Methode

4.1 Récupération de thèses, conversion au format HTML et indexation.

Les universités françaises ne récupèrent de la thèse que la version papier. Notre corpus de thèses est donc formé de thèses au format électronique trouvées sur internet (généralement sur des sites personnels) ou part l'intermédiaire de professeurs de facultés de pharmacie. Elles sont dans deux formats : doc (Microsoft Word) ou pdf (portable document format : Adobe), mais le moteur d'indexation ne fonctionne que sur des fichiers au format HTML. La conversion document Word vers HTML fut réalisée directement sous Word 97. Si la conversion n'est pas exempte de tout reproches, le fichier HTML source reste exploitable par le moteur d'indexation. Pour les documents pdf, Adobe propose sur son site deux modules en particulier : "Make accessible plug in" `http://www.adobe.com/support/downloads/detail.jsp?hexID=88de` et "Save as XML plug in" `http://www.adobe.com/support/downloads/detail.jsp?ftpID=1209`.
Les thèses ont été sauvegardées au format HTML 3.2 sans CSS. Ce format conserve les balises H1 et H2 qui sont utilisées par le moteur pour pondérer les concepts. Les thèses sont soumises au moteur, qui en extrait des concepts. Apres l'indexation de toutes les thèses, le score de pertinence est recalculé.
Parallèlement, les mots clefs issus de l'indexation manuelle sont retrouvés sur docthèse et le Sudoc ou, le cas échéant, soit à l'intérieur même des thèses, soit en contactant directement les bibliothèques chargées de leur indexation. Ces mots clefs sont traduits en termes MeSH.

4.2 Comparaison

Nous avons procédé à une comparaison entre les mots clefs MeSH de l'indexation manuelle faite par les documentalistes, et la liste de termes proposés par Nomindex. Arbitrairement, les vingt premiers termes ont été retenus. Nous partons du principe que les mots clefs manuels (a) sont pertinents, ayant été d'une part définis par des personnes compétentes en leur domaine, et d'autre part, validés par des professionnels de l'indexation. Ainsi, en supprimant de ces derniers les mots clefs trouvés conjointement par les deux méthodes d'indexation, nous trouvons le silence du moteur d'indexation (b). Il est égal au nombre de mots clefs "manuels" non trouvés par le moteur. Ensuite, nous sélectionnons les termes réellement pertinents parmi ceux proposés par le moteur, via une lecture de la thèse. Ont été définis comme pertinents les termes MeSH qui n'étaient ni trop fins, ni trop généraux.
Nous définissons ainsi deux critères :

- le silence de l'indexation manuelle (c), qui correspond au nombre de termes pertinents trouvés par le moteur et absents de l'indexation manuelle.
- le bruit de l'indexation automatique (d), égal au nombre de termes non pertinents.

5 Resultats

La distribution du bruit, de la pertinence et du silence est normale (test de Kolmogoroff - Smirnoff).
Le moteur retrouve a - b = 2.90 mots, soit 2.9 / 4.95 = 58.6 % parmi les 20 premiers termes.

Table 1: Silence et bruit.

Nombre moyen de mots clefs	
Manuel (a)	4.95
Silence Nomindex (b)	2.05
Silence manuel (c)	6.95
Bruit (d)	9.76 / 20

75,9 % des termes sont détectés au delà des 20 premiers. Il apporte (c) 6.95 mots nouveaux. Le bruit (20 - (a + c)) et la pertinence sont respectivement de 50.5 % (10.1 / 20) et de 49.5 % (9.9 / 20).
Les thèses furent indexées en quelques minutes chacune et le calcul final de la pertinence en une vingtaine de minutes sur la station Sun de la faculté de médecine de Rennes. Sur un total de 21 thèses, seuls 8 résultats d'indexation furent trouvés grâce au Sudoc. En fait, les termes de l'indexation furent principalement trouvés dans les thèses elles mêmes (9 thèses) ou en contactant les bibliothèques d'origine (4 thèses).

6 Discussion

Nomindex travail à partir d'un dictionnaire médical. Or, la pharmacie est multidisciplinaire, et son champ d'application touche parfois des domaines plus spécifiques comme la physique ou les mathématiques. Mais la normalité de la distribution du bruit et du silence au sein du corpus tend à prouver que cette disparité sémantique est correctement gérée par le moteur.
Le moteur retrouve 58 % des mots clefs manuels parmi les 20 termes proposés (ce qui peut sembler peu), mais plus de 75 % sur tout les termes. On peut expliquer ce taux relativement faible par un bruit important : à peine la moitié des 20 concepts est pertinente. Aussi la notion même de limite semble à posteriori peu intéressante. Il aurait été possible de travailler différemment pour savoir si les mots clefs choisis par le documentaliste, qui sont la référence, sont bien retrouvés.
Par exemple, une évaluation de l'outil d'indexation pourrait permettre de confirmer que l'on retrouve bien les thèses à partir de leur mots clefs, ou encore, il serait possible de rechercher le seuil de bruit acceptable à partir duquel la totalité des mots clefs de la référence apparaît en tête de l'indexation automatique. Ce problème de classement peut s'expliquer par une extraction performante : en moyenne, l'outil apporte près de 7 nouveaux mots clefs pertinents. En réduisant le bruit, ce chiffre serait plus important encore. Néanmoins, le rapport bruit / silence est bon si l'on prend en considération que :

- le dictionnaire est axé essentiellement sur le diagnostic médical (et non le domaine médical dans son ensemble).
- c'est une première expérience sur des documents pharmaceutiques, et le dictionnaire n'a jamais été amélioré en ce sens, comme il l'a pu l'être pour la médecine. Son enrichissement en terme spécifique lui permettront de réduire le silence.
- le bruit vient très souvent de quelques concepts qui sont abusivement reconnus. De petites

modifications ponctuelles du dictionnaire élimineraient beaucoup de ces fautes. Dans le cadre d'une indexation par les bibliothécaires, ceux ci pourraient affiner le dictionnaire au gré des erreurs trouvées [8].

Le silence devient très faible si on le compare à celui du Sudoc, qui ne retrouve qu'un tiers de thèses seulement. Il semblerait que les thèses ne soient par mis directement sur le système Sudoc. Certaines thèses de plus de 6 mois n'y figuraient pas encore. En cela, Nomindex pourrait palier à ces insuffisances : il suffirait aux facultés de récupérer la version électronique des thèses en même temps que leur dépôt. Les utilisateurs bénéficieraient d'un moteur de recherche performant, avec une extension de recherche reposant sur le réseau sémantique d'un metathésaurus validé.
D'autant que le temps d'indexation est compatible avec une mise en ligne en temps réel des thèses sur un serveur central à toutes les facultés par exemple. Seul le temps de calcul des TFIDF sera peut être plus problématique quand plusieurs années seront indexées. Dans un premier temps, il serait possible d'utiliser ce moteur au seul abstract.
Notre corpus de thèses pourrait ne pas être représentatif des thèses françaises pour deux raisons. Premièrement, elles sont majoritairement issues de sites personnels. Les auteurs et les sujets traités ne sont pas forcement représentatifs. D'autre part, notre corpus ne représente qu'un quart environ des thèses passées annuellement dans une faculté de taille moyenne. Ceci limite la portée de notre travail, mais nous poursuivrons celui ci sur un nombre plus important de thèses, avec le concours de l'administration des facultés de pharmacie.

7 Conclusion

Il est du devoir des facultés, des professeurs et des bibliothécaires de mettre en valeur le travail réalisé au sein de leur faculté [9]. Au cours de ce travail, nous avons montré que le type d'indexation automatique permettrait une utilisation performante de l'UMLS (metathesaurus et réseau sémantique) et une extraction efficace de la connaissance. L'indexation, si elle est actuellement insuffisante pour une utilisation directe pourrait facilement être améliorée par des mises à jour spécifiques du dictionnaire.
Reste les inconvénients incontournables d'une indexation basée sur le dictionnaire ADM, qui ne distingue pas ”aine” de ”aîné” [10]. Il sera peut être nécessaire de procéder à une analyse syntaxique préliminaire. Par la suite, un perfectionnement soit des techniques d'extension de recherches [11] soit de l'UMLS lui même [12] permettront peut être de mettre à disposition des chercheurs et étudiants une base de connaissance médicale fournie, et simple d'accès.

References

[1] Pouliquen B, Delamarre D and LeBeux P. Indexation de textes médicaux par extraction de concepts et ses utilisations. 6th International Conference on Statistical Analysis of Textual Data 2002. Proceedings to be printed.

[2] Lindberg DA, Humphreys BL and McRay AT. The Unified Medical Language System. Methodes Inf Med 1993; 32: 281–91

[3] Nadkarni P, Chen R and Brandt C. UMLS concept indexing for prodution database: a feasibility study. J Am Med Inform Assoc. 2001; 8: 80–91.

[4] National Library Of Medicine, Medical Subject Headings (1986).

[5] Lenoir P, Michel JR., Frangeul C and Chales G. Réalisation,développement et maintenance de la base de données A.D.M. Médecine informatique. 1981; 6: 51–6

[6] Berrios DC, Automated Indexing for Full Text Information Retrieval. Proc AMIA Symp 2000; 71–5.

[7] Salton and Buckley C. Term weighting approaches in automatic text retrieval. Information Processing and Management 1988; 24: 513–23.

[8] In : Construction de ressources terminologiques. Bourigault D and Jacquemin C eds. Hermes 2000;

[9] In : Bulletin Officiel du ministère de l'Education Nationale et du ministère de la Recherche Numero 34. Ministère de l'éducation. 2000; pp.430–4.

[10] Le Duff F, Burgun A and Pouliquen B et al., Automatic enrichment of the unified medical language system starting from the ADM knowledge base. Stud Health Technol Inform 1999; 68: 881–6.

[11] Hersh W, Price S and Donohoe L. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. Proc AMIA Symp 2000; 344–8

[12] W.T. Hole and S. Srinivasan, Discovering Missed Synonymy in a Large Concpt-Oriented Metathesaurus. Proc AMIA Symp 2000; 354?8