



Indexation UMLS en français : une expérience

Thierry Delbecque¹, Pierre Zweigenbaum^{1,2,3}

¹INSERM, U729, Paris, France

²Assistance Publique - Hôpitaux de Paris, STIM/DSI, Paris, France

³INALCO, CRIM, Paris, France

Indexation UMLS en français : une expérience

Thierry Delbecque¹, Pierre Zweigenbaum^{1,2,3}

¹INSERM, U729, Paris, France

²Assistance Publique - Hôpitaux de Paris, STIM/DSI, Paris, France

³INALCO, CRIM, Paris, France

1 Introduction

L'indexation de documents médicaux est un prérequis pour de nombreuses tâches d'informatique médicale, qu'elles concernent des documents cliniques [1,2] ou pour des articles scientifiques [3]. Les vocabulaires contrôlés, en particulier leur agrégation au sein du Metathesaurus de l'UMLS [4], sont particulièrement pertinents pour cela. De ce fait, de nombreux travaux se sont intéressés à l'usage du Metathesaurus pour indexer des textes médicaux [3,2].

Dans cet article, nous présentons une méthode et un outil, MetaCoDe, pour indexer des textes médicaux à l'aide de la partie francophone du Metathesaurus de l'UMLS. Au-delà des concepts du Metathesaurus, cela donne accès en même temps aux types sémantiques du réseau sémantique de l'UMLS qui leur sont associés. Cet outil a été testé sur un corpus construit à partir de documents web indexés par le catalogue CISMef [5].

Le problème qui nous a motivés porte sur la faible représentation du français dans le Metathesaurus. Ainsi dans sa version 2002-AA, le français ne couvre qu'à peine 2 % des concepts présents; de même, en termes de synonymie, le français propose en moyenne 1,54 chaînes (termes différents) pour représenter un concept, là où l'anglais en propose 2.

Il n'existe pas actuellement de collection de textes en français préétiquetés par des concepts du Metathesaurus, qui pourrait servir d'étalon de référence pour une évaluation quantitative. Nous avons donc effectué cette évaluation quantitative sur un échantillon de notre corpus. Nous l'avons complétée en employant un autre type de méthode qui permet de réaliser une évaluation qualitative : une analyse factorielle des correspondances entre les types sémantiques trouvés par le programme et les caractéristiques des différentes parties de notre corpus de test. Ceci afin de voir si, malgré son caractère ténu, l'UMLS francophone parvenait encore à doter un corpus d'une structure thématique cohérente, en accord avec le contenu individuel de chaque document.

2 Matériel et méthodes

Nous avons utilisé la version 2002-AA de l'UMLS. Nous avons mis en œuvre l'ensemble des termes français du Metathesaurus et le réseau sémantique.

2.1 Le corpus CISMef-EQueR

Le corpus sur lequel nous avons travaillé est celui de la compétition EQueR¹ de 2004. C'est un sous-ensemble des documents indexés par CISMef, faisant partie des domaines listés dans la figure 1, auxquels viennent se joindre les documents référencés *un lien plus loin*. L'ensemble comporte 5583 documents au format HTML ou PDF, pour un total d'environ 19 millions de mots.

Le corpus est obtenu en effectuant successivement, après le téléchargement:

- la traduction en fichiers textes (iso-latin-9);
- le nettoyage des textes, rendu nécessaire en particulier par le caractère délicat de la conversion depuis le format PDF;
- d'autres traitements tels que le repérage des abréviations et de leurs définitions, ce afin de faciliter le repérage des termes lors de l'indexation.

De tous ces traitements il ne sera pas davantage question par la suite (pour davantage d'information, voir [6]).

CANCER	www.fnclcc.fr www.fnclcc.com - Fédération Nationale des Centres de Lutte Contre le Cancer (294 documents)
DOCFRA	www.ladocfrançaise.gouv.fr www.ladocumentationfrancaise.fr - La documentation française (258 documents)
AFSSAPS	afssaps.sante.fr - Agence Française de Sécurité Sanitaire des Produits de Santé (ex Agence du Médicament) (112 documents)
ANAES	www.anaes.fr - Agence Nationale d'Accréditation et d'Evaluation en Santé (1192 documents)
ORPHA	www.orpha.net - Orphanet, serveur d'informations sur les maladies rares (118 documents)
SENAT	www.senat.fr - Site officiel du Sénat français (1278 documents)
CHUROUE	www.chu-rouen.fr - CHU de Rouen (275 documents)
UROUEN	www.univ-rouen.fr - Université de Rouen, restreinte à sa branche médicale (147 documents)
CANADA	www.hc-sc.gc.ca - Site bilingue de Santé Canada (ministère fédéral de la santé), source d'informations générales sur le santé publique au Canada, de statistiques, de conseils, etc. (1909 documents)

Figure 1: Sources des documents de travail

2.2 Indexation UMLS : MetaCoDe

MetaCoDe est une plateforme logicielle écrite essentiellement en PERL, initialement mise au point spécialement pour cette expérience d'évaluation, mais que nous comptons enrichir afin de généraliser son utilisation. Pour la partie qui nous concerne ici, cet outil enchaîne les traitements suivants²:

- repérage des termes simples ou composés (syntagmes nominaux, par exemple *inflammation chronique aiguë*, ou *directive ministérielle*);
- projection sur les syntagmes de concepts de l'UMLS, et des types sémantiques

¹ Évaluation de systèmes de Questions-Réponses, projet Technolangue coordonné par ELDA.

² MetaCoDe assure en réalité d'autres traitements tels que l'extraction de prédicats, ou le repérage de relations sémantiques présumées, l'hébergement dans un format relationnel joint à un requêteur [6].

- associés;
- constitution de tableaux de contingence, croisant documents et types sémantiques, afin d'effectuer les analyses exploratoires multidimensionnelles qui nous intéressent.

Le repérage de termes de l'UMLS au sein de textes, et les applications qui en découlent, constituent un sujet fréquent dans la littérature anglophone. Dans ce cadre, l'anglais a la chance de disposer d'outils de traitement automatique de la langue offerts par la NLM conjointement à l'UMLS, (tels que SPECIALIST [4], comptant entre autre un lexique, des méthodes de normalisation de termes, etc.).

Le français ne dispose pas de telles ressources spécialisées; dans notre cas, le repérage des syntagmes nominaux (SN) au sein du corpus est effectué sur la base de patrons de parties du discours (nom, verbe, etc.) que nous avons conçus de manière à ce qu'ils soient suffisamment robustes face aux cas de dégradation de la qualité du corpus fréquemment constatés. Pour cela, l'ensemble du corpus est préalablement étiqueté en parties du discours par TreeTagger [7]³.

Un fois les SN repérés, la procédure d'étiquetage UMLS consiste à chercher, pour chaque mot du syntagme en cours de traitement, les chaînes françaises (termes) de l'UMLS dans lesquelles il intervient. Les chaînes de l'UMLS étant désignées par des identificateurs uniques, les SUI, on obtient, lorsque la recherche individuelle pour chaque mot est achevée, un treillis⁴ de chaînes UMLS, dont on ne conserve que les éléments maximaux. Les types sémantiques (par exemple, *disease or syndrome* ou encore *body part, organ or organ component*) des concepts attachés à ces éléments maximaux constitueront les étiquettes du syntagme. Cette méthode permet par exemple de bien poser le concept *œdème papillaire* sur une occurrence de la chaîne *œdème bilatéral papillaire*, lorsque cette dernière n'est pas une chaîne de l'UMLS.

2.3 Évaluations

Sur la base de l'étiquetage présenté précédemment, nous avons voulu évaluer les aspects suivants.

2.3.1 Mesures quantitatives

En utilisant deux échantillons de 300 syntagmes chacun, nous avons estimé des indicateurs proches du silence et de la précision, et que nous désignerons par la suite par les termes *taux d'étiquettes manquantes* et *taux de faux étiquetages*. La distinction vient du fait que nous avons dû pour chacun de ces indicateurs introduire une troisième classe témoignant soit d'une indécision du juge, soit d'un étiquetage incomplet. Ainsi nous avons estimé dans le cas du *taux d'étiquettes manquantes* les proportions de:

faux négatifs -

syntagmes non étiquetés, mais dont le sens médical aurait justifié qu'ils le soient (*curage ganglionnaire* par exemple);

vrais négatifs -

syntagmes non étiquetés à juste titre (*cadre stratégique national* par exemple);

³ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>.

⁴ Suivant la relation d'ordre partiel définie par l'inclusion.

indécision -

cas ambigus, pour lesquels il a été jugé préférable de ne pas trancher (*enfermement familial*, ou *entourage ultraviolet*).

et dans le cas du *taux de faux étiquetage* les proportions de:

cas corrects -

syntagmes étiquetés conformément à leur sens;

cas incomplets -

syntagmes dont seule une partie du sens a été étiquetée; ainsi, l'étiquetage du terme *atteinte gastrointestinale* avec le type sémantique T047 (*disease or syndrome*) est incomplet du fait que l'étiquette porte sur *atteinte* seulement;

cas incorrects -

syntagmes dont le sens est clairement différent du sens de l'étiquette.

2.3.2 Aspect qualitatif

Au-delà des mesures très synthétiques précédentes, nous avons voulu nous rendre compte si globalement les étiquettes posées dans le corpus parvenaient à doter celui-ci d'une structure cohérente avec la vocation des documents (donc une structure thématique). Il s'agit donc de voir s'opérer le rapprochement de deux ensembles, celui des types sémantiques de l'UMLS et celui des thèmes des documents. Pour cela, le recours à l'analyse factorielle des correspondances [8] appliquée au tableau de contingence croisant types sémantiques projetés et documents sources a été une solution naturelle.

3 Résultats

3.1 Mesures quantitatives

Le tableau 1a donne la densité globale d'étiquettes posées dans le corpus. Il montre qu'environ 37 % des syntagmes ont été marqués par au moins un type sémantique de l'UMLS. Le *taux d'étiquettes manquantes*, mesuré sur un échantillon aléatoire de 300 syntagmes non étiquetés, et le *taux de faux étiquetage*, mesuré sur un échantillon aléatoire de 300 syntagmes étiquetés, figurent dans les tableaux 1b-c respectivement, et peuvent être brièvement résumés en disant que la moitié environ des syntagmes nominaux ont été bien étiquetés.

Item	Nombre
Occurrences de	4101404
Syntagmes distincts	391966
... étiquetés	147007
... non étiquetés	246959

(a) densité

Item	Taux
Vrais	45 %
Faux négatifs	25 %
Indécisions	30 %

(b) silence

Item	Taux
Corrects	52 %
Incomplet	32 %
Incorrects	16,00%

(c) faux étiquetages

Tableau 1: Evaluation de l'étiquetage : densité, étiquetage manquant et faux étiquetage

3.2 Aspect qualitatif

Les résultats détaillés de l'analyse des correspondances peuvent être consultés dans [6]. L'analyse exhibe 104 valeurs propres non nulles, dont il faut conserver les 11 premières pour capturer la moitié de la variance du tableau de données. L'histogramme des valeurs propres indique que l'étiquetage a bien fait émerger une structure de l'ensemble des documents de travail.

La figure 2a représente la projection des types sémantiques sur le premier plan factoriel (plan 1-2). On voit trois pôles thématiques émerger sur ce diagramme:

au Nord-Ouest

les documents à connotation physique et biologique en général; nous désignons ce pôle par l'étiquette **matière**;

au Nord-Est

les documents à connotation médicale et physiologique; nous désignons ce pôle par l'étiquette **bio-process**;

au Sud

les documents plutôt à sujet administratif, ou sociologique; nous désignons ce pôle par l'étiquette **organisation**.

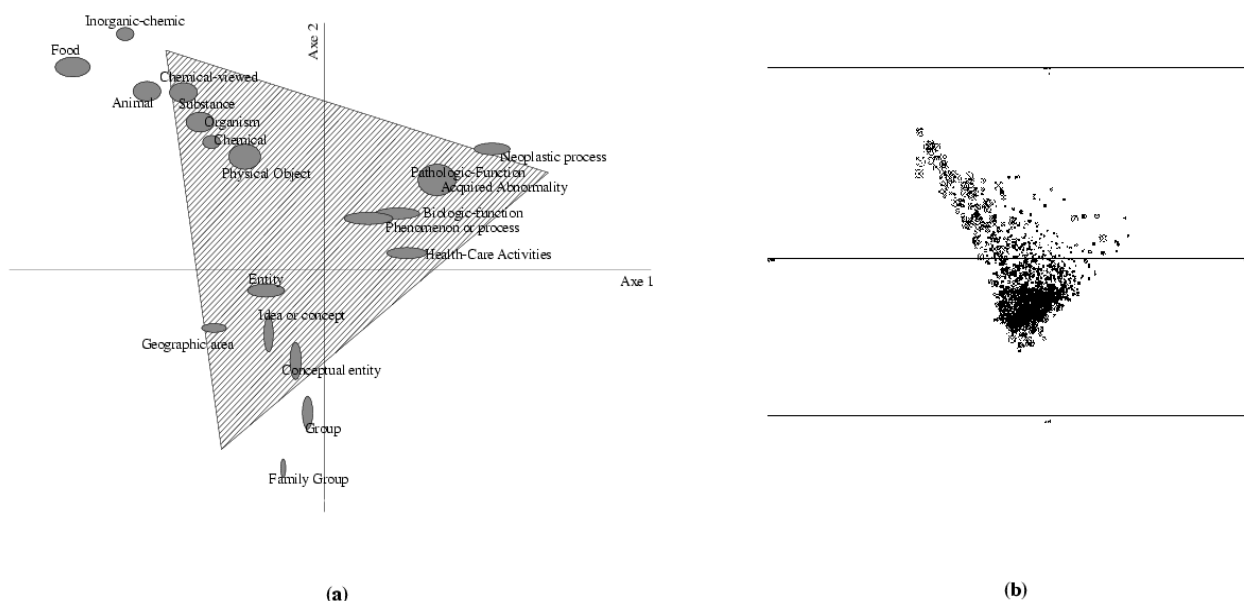


Figure 2: (a) Projection des types sémantiques sur le plan 1-2. (b) Projection des documents de [SENAT] (cercles) et [DOCFRA] (carrés noirs) sur le plan 1-2.

L'examen de la projection des documents sur ce plan, en fonction de leur origine respective, permet de comparer la structure thématique que l'étiquetage semble apporter avec la structure thématique avérée du corpus de travail, puisque chaque source a une vocation distincte. La projection des documents en provenance de [SENAT] et [DOCFRA] est fournie en exemple sur la figure 2b.

Le résultat de l'ensemble des projections des documents se résume ainsi [6]:

- [SENAT] et [DOCFRA] occupent quasi-exclusivement le pôle **organisation**

(Sud), malgré une incursion prononcée (une cinquantaine de documents) dans le pôle **matière** (Nord-Ouest);

- [ORPHA], [CANCER], et [UROUEN] se concentrent sur le pôle **bio-process** (Nord-Est);
- [ANAES] et [AFSSAPS] se partagent entre le pôle **organisation** (Sud) et le pôle **bio-process** (Nord-Est), avec une plus forte présence sur ce dernier;
- [CHUROUEN] et [CANADA] se répartissent sur l'ensemble des pôles.

Ainsi, il apparaît à cette lecture que le jeu d'étiquettes UMLS que nous avons projeté sur le corpus parvient à en faire ressortir une structure thématique, en mettant en exergue ce que nous pouvons considérer *a priori* comme des grandes familles thématiques. La source des documents, projetée sur le plan, semble logique et confirmer cette structuration. [SENAT] et [DOCFRA] hébergent des documents législatifs et politiques, leur positionnement sur **organisation** est donc cohérent; l'incursion en **matière** s'explique par une thématique sanitaire. La présence de [ANAES] et [AFSSAPS] en **organisation** reflète le caractère institutionnel de ces sources, leur spécificité médicale est bien traduite par leur positionnement en **bio-process** également. [ORPHA], [CANCER], et [UROUEN], plus spécialisés sur les aspects pathologiques, sont bien sur le quadrant qui leur convient (**bioproces**). Enfin, [CHUROUEN] et [CANADA], sites médicaux mais ouverts au grand public, concrétisent cette double vocation par leur occupation globale du plan.

4 Discussion et conclusion

L'indexation réalisée arrive ainsi à refléter globalement le contenu thématique du corpus de départ, en préservant les différences entre les sites indexés. L'examen local des syntagmes indexés montre que les indexations erronées sont peu fréquentes (16 %), et que c'est surtout la couverture (25 à 55 % de faux négatifs) et la complétude (32 % d'indexations incomplètes) de l'indexation qui restent à améliorer. Pour cela, nous considérons que trois voies sont à poursuivre.

La première consiste à utiliser des méthodes plus élaborées (voir par exemple [3,9]) mettant en jeu par exemple un apprentissage automatique. La deuxième consiste à employer des connaissances linguistiques qui permettent de reconnaître des mots ou termes sous des formes variantes [10]. MetaCode tient compte des variantes liées aux formes fléchies (pluriel, féminin) à travers son utilisation de TreeTagger qui lemmatise les mots rencontrés. Une extension naturelle est la prise en compte de variantes dérivationnelles (*intestinal* / *intestin*) [11], compositionnelles (*gastrectomie* / *ablation estomac*) [12] et de leur combinaison dans des termes (*virus de la variole* / *virus variolique*) [13].

La troisième voie repart du constat du nombre relativement faible de termes en français dans l'UMLS. Pour l'essentiel, jusqu'en 2003, ces termes français sont ceux du thesaurus MeSH (et, depuis 2004, de MedDRA). Il est donc crucial, pour espérer atteindre une couverture plus importante du contenu des textes français du domaine biomédical, d'accroître le nombre de terminologies médicales en version française dans le Metathesaurus de l'UMLS, ainsi que le nombre de termes synonymes présents dans chacune. La situation du français dans l'UMLS est amenée à évoluer dans ce sens, notamment du fait du travail actuel du consortium VuMEF, dont l'objet est précisément d'accroître le vocabulaire français dans le Metathesaurus [14].

L'indexation présentée ici a été utilisée dans un système de recherche de réponses à des questions médicales (système STIM-LIPN) qui a participé à l'évaluation EQueR. Les questions catégorisées comme portant sur un type sémantique donné (par exemple,

diagnostic) déclenchent la sélection d'énoncés indexés par ce type sémantique, aidant ainsi à focaliser la recherche de réponses.

Références

- [1] Huang Y, Lowe HJ, et Hersh WR. A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports. *J Am Med Inform Assoc* 2003;10(6):580-7.
- [2] Zou Q, Chu WW, Morioka C, Leazer GH, et Kangarloo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. In: Musen M, ed, Actes AMIA Annual Fall Symposium 2003, Washington, DC. AMIA, novembre 2003; pp. 763-7.
- [3] Aronson AR, Bodenreider O, Chang F, et al. The NLM indexing initiative. *J Am Med Inform Assoc* 2000;7(suppl):17-21.
- [4] McCray AT et Nelson SJ. The semantics of the UMLS knowledge sources. *Methods Inf Med* 1995;34(1/2).
- [5] Darmoni SJ, Leroy JP, Thirion B, et al. CISMef: a structured health resource guide. *Methods Inf Med* 2000;39(1):30-5.
- [6] Delbecque T. Structuration de corpus médicaux par l'UMLS. utilisabilité comme source d'entités nommées pour les systèmes de questions-réponses. Rapport de DEA, Informatique Médicale, Université Paris 5, 2004.
- [7] Schmid H. Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing, Manchester, UK. 1994; pp. 44-9.
- [8] Benzécri JP. Correspondances, (vol1). Dunod, 1979.
- [9] Ruch P, Baud R, Bouillon P, et Robert G. Minimal commitment and full lexical disambiguation: Balancing rules and hidden markov models. In: Cardie C, Daelemans W, Nedellec C, et Tjong Kim Sang E, eds, Proc CoNLL-2000 and LLL-2000, Lisbon, Portugal. 2000; pp. 111-4.
- [10] Grabar N, Zweigenbaum P, Soualmia L, et Darmoni SJ. Matching controlled vocabulary words. In: Baud R, Fieschi M, Le Beux P, et Ruch P, eds, Actes Medical Informatics Europe, (vol95) of *Studies in Health Technology and Informatics*, Amsterdam. IOS Press, 2003; pp. 445-50.
- [11] Zweigenbaum P, Baud R, Burgun A, et al. A unified medical lexicon for French. *International Journal of Medical Informatics* 2004. *À paraître*.
- [12] Namer F et Zweigenbaum P. Acquiring meaning for French medical terminology: contribution of morphosemantics. In: Fieschi M, Coiera E, et Li YCJ, eds, Actes 10th World Congress on Medical Informatics, San Francisco, Ca. 2004; pp. 535-9.
- [13] Jacquemin C et Tzoukermann E. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In: Strzalkowski T, ed, *Natural language information retrieval*, (vol7) of *Text, speech and language technology*. Kluwer Academic Publishers, Dordrecht & Boston, 1999; pp. 25-74.
- [14] Darmoni SJ, Jarrousse E, Zweigenbaum P, et al. Extending the French part of the UMLS. In: Musen M, ed, Actes AMIA Annual Fall Symposium 2003, Washington, DC. AMIA, novembre 2003. (poster).

Adresse de correspondance

Thierry Delbecque, INSERM U729, Santé Publique et Informatique Médicale (SPIM)

Faculté de Médecine Broussais-Hôtel-Dieu, 15 rue de l'école de médecine, 75006 Paris.

thd@biomath.jussieu.fr