Disponible en ligne sur
SciVerse ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

IRBM

IRBM 33 (2012) 316–329

Article original

Indexation automatique de documents en santé : évaluation et analyse de sources d'erreurs

Automatic indexing of health documents in French: Evaluating and analysing errors

W. Chebil [a,b,*], L.F. Soualmia [b], B. Dahamna [b], S.J. Darmoni [b]

[a] *Unité de recherche MARS, université de Monastir, Monastir, Tunisie*
[b] *Équipe CISMeF, LITIS-TIBS EA 4108, CHU de Rouen, cours Leschevin, porte 21, 3e étage, 1, rue de Germont, 76031 Rouen cedex, France*

Reçu le 6 juin 2012 ; reçu sous la forme révisée le 13 octobre 2012 ; accepté le 15 octobre 2012

Résumé

Catalogue et index des sites médicaux de langue française (CISMeF) a été développé pour trouver sur Internet l'information médicale utile destinée aux professionnels de santé, les patients et les étudiants en médecine. Les ressources collectées sont indexées manuellement, semi-automatiquement ou automatiquement. Actuellement, la fonction d'indexation automatique de CISMeF indexe qu'une partie des ressources qui sont jugées les moins importantes.

Objectif. – L'objectif de ce travail est d'évaluer la fonction d'indexation automatique de CISMeF et analyser les erreurs générées.

Matériel et méthode. – Nous avons utilisé 500 recommandations pour évaluer la fonction d'indexation basée dès son implémentation sur l'algorithme de sac de mots. L'index automatique généré est comparé à l'indexation manuelle, considérée ici comme le « gold standard ». Nous étudions l'indexation automatique conjointe des titres et des sous-titres courts, l'indexation automatique conjointe des titres et des sous-titres longs, l'indexation automatique conjointe des titres et des sous-titres courts et longs, puis celle des résumés. Les mesures d'évaluation utilisées sont les mesures classiques de précision, rappel et F-mesure.

Résultats. – Les résultats de l'évaluation de l'indexation des titres et des sous-titres courts sont de 0,56 pour la précision et 0,21 pour le rappel. Pour les titres et sous-titres longs, la précision est de 0,39 et le rappel est de 0,27. La précision de l'indexation des résumés est 0,23 et le rappel est de 0,61. Suite à l'analyse des erreurs d'indexation, 13 catégories d'erreurs sont identifiées. L'indexation des titres et sous-titres courts a généré moins d'erreurs qui sont à l'origine de la présence des descripteurs non corrects (0,97 erreurs par titre et sous-titre court). L'indexation des résumés a généré moins d'erreurs qui sont à l'origine de l'absence des descripteurs pertinents (2,52 erreurs par résumé).

Conclusion. – L'évaluation de l'indexation automatique a montré qu'elle n'est applicable telle quelle que pour l'indexation des phrases simples et courtes, vu la précision acceptable de l'indexation des titres et sous-titres courts. Nous visons, suite à l'identification des causes des erreurs, qui représente une étape importante vers l'amélioration de la fonction d'indexation, à proposer et implémenter des solutions ce qui permettra d'indexer automatiquement un plus grand nombre de documents en santé.

Abstract

Catalogue and Index of French Medical Sites (CISMeF) is developed for retrieving the relevant medical information in the Internet for health professionals, the patients and students in medicine. The gathered resources are manually indexed, semi-automatically indexed or automatically indexed. Actually, the function indexing of CISMeF indexes only a part of resources that are judged the less important.

Objectives. – The objective of this work is to evaluate the indexing function developed for CISMeF, and analyse generated errors.

Material and method. – We used 500 clinical guidelines for the evaluation of the indexing function, based since his implementation, on the "bag of words" algorithm. The automatic index generated is compared with the manual one which is considered as the "gold standard". We analyze the automatic indexing of short titles and subtitles associated, the automatic indexing of long titles and subtitles associated, the automatic indexing of long and short titles and subtitles associated and the automatic indexing of abstracts. The measures used for the evaluation are Precision, Recall and F-measure.

* Auteur correspondant.
Adresses e-mail : wiem.chebil@yahoo.fr (W. Chebil), lina.soualmia@chu-rouen.fr (L.F. Soualmia), badisse.dahamna@chu-rouen.fr (B. Dahamna), stefan.darmoni@chu-rouen.fr (S.J. Darmoni).

1959-0318/$ – see front matter © 2012 Elsevier Masson SAS. Tous droits réservés.
http://dx.doi.org/10.1016/j.irbm.2012.10.002

Results. – The results of the evaluation of the short titles and subtitles indexing are 0.56 for the precision, 0.21 for the recall. For the long titles and subtitles the precision is 0.39, the recall is 0.27. The precision of abstracts indexing is 0.23 and the recall is 0.61. Thirteen categories of errors are identified by analysing the indexing function. The short titles and subtitles indexing generated the less errors leading to the presence of wrong descriptors (0.97 errors per short tiles and subtitles). The long titles and subtitles generated the most errors leading to the absence of relevant descriptors (2.52 errors by long titles and subtitles).

Conclusion. – The evaluation of the indexing function showed that it should be used only for short titles and subtitles. We aim, after the identification of the causes of errors, to improve the performance of the automatic indexing function which will allow indexing more medical documents.

1. Introduction

Le Catalogue et index des sites médicaux de langue française (CISMeF) recense, depuis 1995, les principales ressources de santé disponibles en français sur l'Internet [1]. Par choix éditorial, ces ressources sont principalement institutionnelles, en provenance de ministères, agences gouvernementales et universités du monde francophone. Une ressource peut être un site web, ou un document, tout support susceptible de contenir une information de qualité en santé [2]. Ces ressources sont régulièrement mises à jour. Jusqu'en 2006, l'indexation était uniquement réalisée à l'aide des descripteurs du thésaurus Medical Subject Headings (MeSH), développé par la US National Library of Medicine et utilisé notamment pour indexer les articles scientifiques de la base de données bibliographiques MEDLINE. De 1995 à 2005, l'indexation était réalisée exclusivement manuellement par les documentalistes du catalogue. Vu l'augmentation rapide du nombre de ressources ($n = 90\,620$ en 2012 vs $n = 12\,291$ en 2003), depuis 2005 l'indexation des ressources est devenue automatique et semi-automatique. L'indexation automatique est adoptée que pour une partie des ressources qui sont jugées les moins importantes et les moins sensibles. Ces ressources sont indexées uniquement par leurs titres et sous-titres (contrairement à l'indexation manuelle qui est réalisée sur tout le document). En effet, une nouvelle politique éditoriale a dû être définie :

- les recommandations pour la pratique clinique continuaient à être indexées manuellement ;
- les documents pédagogiques nationaux, les brochures pour les patients, les rapports d'agence gouvernementales étaient indexés semi-automatiquement, c'est-à-dire automatiquement puis revus par un(e) documentaliste ;
- les documents pédagogiques facultaires et tous les documents pour la politique de santé étaient indexés automatiquement.

La fonction d'indexation automatique était dès son implémentation dans CISMeF fondée sur l'algorithme du « sac de mots », qui a été utilisé pour améliorer la fonction de recherche d'information dans CISMeF, en indexant les requêtes des utilisateurs [3] par des termes MeSH afin de les apparier avec les documents. Cette fonction a été modifiée régulièrement en ajoutant et en supprimant des fonctionnalités, cela tout en gardant le principe général de cet algorithme. L'algorithme du « sac de mots » considère que tous les mots de la requête de l'utilisateur sont des mots clés, à l'exclusion des mots « vides » qui sont des chaînes fréquentes dans le texte et ne doivent pas être indexés, par exemple : dans, le, à, et ... etc. L'évaluation quantitative et qualitative de l'indexation automatique des documents, ainsi que les sources d'erreurs de l'algorithme est l'objet de cet article.

Plusieurs travaux se sont intéressés à l'indexation automatique des documents médicaux contrôlée par des thésaurus et des terminologies médicales. Les travaux sur le MeSH en anglais ont utilisé le modèle probabiliste [4] et les techniques d'apprentissage automatique telles que le réseau bayésien [5] et les k-plus proches voisins pour la classification des documents [2,6–8]. Aronson et al. [2] exploitent également l'outil MetaMap [9] et la méthode de tri-gram (cette méthode permet de déterminer la similarité entre deux phrases) pour l'extraction des concepts Unified Medical Language System (UMLS[1]) qui sont ensuite restreints aux concepts MeSH. Ruch [10] a proposé deux modèles, l'un est basé sur les expressions régulières et l'autre sur le Space Vector Model [11] et a montré que ces deux méthodes sont complémentaires et mènent à des résultats meilleurs que ceux obtenus en appliquant les deux méthodes séparément. Ces deux méthodes peuvent être appliquées sur des corpus en anglais ou en français. Hliaoutakis et al. [12] ont proposé le modèle Automatic MeSH Term Extraction (AMTx) qui intègre dans une première étape la méthode C/NC-value [13], qui permet l'extraction des termes composés à partir de texte en combinant l'information statistique et linguistique, ensuite classe ces termes selon la valeur de C/NC-value et ne retient que ceux qui correspondent aux descripteurs MeSH. La liste des termes obtenue après classement est enrichie par les termes sémantiquement proches des descripteurs présents dans la liste.

Concernant l'indexation des documents médicaux en français, contrôlés par le thésaurus MeSH, plusieurs méthodes se sont basées sur le traitement automatique de la langue naturelle (TAL) [8] ou encore des méthodes statistiques, comme Pouliquen et al. [14] qui calculent un poids statistique pour chaque terme extrait initialement à partir du texte et mis en correspondance avec les termes du dictionnaire de l'Aide au Diagnostic Médical (ADM) [15] et les termes MESH. Les méthodes sémantiques sont également utilisées. On peut citer Majdoubi et al. [16] qui intègrent également le modèle vectoriel pour calculer la similarité entre les phrases de textes et les termes de concepts MeSH, ensuite calculent un poids pour garder les termes pertinents. Dinh et al. [17] combinent plusieurs outils d'indexation, comme [2] et [11].

[1] National Library of Medicine. UMLS Meta-thesaurus. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.

D'autres travaux d'indexation automatique se basent sur des terminologies autres que le MeSH, telles que l'UMLS et la Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT).

Pour l'UMLS, les travaux d'indexation comprennent souvent les méthodes sémantiques et statistiques comme Harrathi et al. [18], qui proposent une approche d'indexation indépendante de la langue, ou aussi les méthodes sémantiques et le TAL à la fois [19–21]. Zhou et al. [22] ont proposé d'annoter le texte avec les mots les plus significatifs d'un concept cela en déterminant un score qui calcule le poids d'un mot dans un concept. Cette méthode vise surtout à améliorer le rappel[2]. Cette méthode est testée également aussi en utilisant le méta-thésaurus de l'UMLS.

Pour la nomenclature SNOMED-CT, on peut citer Ruch et al. [23] qui ont suivi le même processus d'indexation présenté dans [10] pour indexer MEDLINE à l'aide des termes SNOMED-CT. Jonquet et al. [24] exploitent plus de 200 ontologies pour l'indexation des ressources médicales, cela en utilisant Mgrep [25] comme outil de reconnaissance de concept, et en calculant un score pour chaque annotation générée en fonction de sa provenance (terme préféré d'un concept, terme synonyme d'un concept. . .).

Les objectifs de ce travail sont l'évaluation quantitative et qualitative de la fonction d'indexation de CISMeF et l'analyse les causes des erreurs générées. Atteindre ces objectifs constitue une étape importante vers l'amélioration de la performance d'indexation automatique.

Le manuscrit est organisé de la manière suivante : nous présentons dans la première section les étapes sur lesquelles est basée la fonction d'indexation ; l'objet de la seconde section est de décrire en détails la procédure de l'évaluation. La troisième section comprend les résultats générés, qui sont ensuite discutés dans la Section 4.

2. Matériel et méthode

2.1. La fonction d'indexation

L'indexation des documents dans CISMeF se base sur l'algorithme « sac de mots » (la fonction d'indexation est détaillée sur la Fig. 1). C'est un algorithme qui consiste à transformer un texte, écrit en langage naturel ensuite prétraité, sous la forme de sacs de mots. En fonction du vocabulaire d'indexation, ces mots peuvent être pertinents ou non pertinents pour l'indexation du document.

Les différentes étapes d'indexation, illustrées par la Fig. 2, sont les suivantes.

Étape 1. La phase de pré-traitement du texte consiste à désaccentuer, enlever les signes de ponctuation et transformer tous les caractères en minuscule.

Étape 2. Le texte est ensuite segmenté en mots (« tokens ») cela à l'aide d'une liste de séparateurs (exemples : «. », «, », « », « * » « ; »,« ! », « ? », « - »).

Étape 3. Les mots vides sont supprimés.

Exemple : Titre « Pertinence et faisabilité, en 2004, d'un programme préventif de réduction du risque de transmission du virus du Nil occidental avec des larvicides ».

Après le passage par les étapes 1 et 2 le titre devient : « "pertinence", "faisabilité", "2004", "programme", "préventif", "réduction", "risque", "transmission", "virus", "nil", "occidental", "larvicides" ».

Étape 4. La suite des mots du texte obtenue à l'étape 3 est découpée en portions. Les portions se chevauchent : les deux premiers mots de chaque portion correspondent aux deux derniers de la portion précédente. Chaque portion est constituée d'un nombre fixe n choisi[3] de mots constituant ainsi le sac de mots (peut être appelée aussi « fenêtre »).

Exemple : Pour $n = 7$, la suite de mots résultante des étapes 1, 2 et 3 dans l'exemple précédent est découpée en deux sacs de mots :

- sac de mots 1 : « "pertinence", "faisabilité", "2004", "programme", "préventif", "réduction", "risque" » ;
- sac de mots 2 : « "réduction", "risque", "transmission", "virus", "nil", "occidental", "larvicides" ».

À partir de l'étape 5, nous allons considérer qu'un seul sac de mots (le premier sac de mots du texte). Le reste des sacs de mots suivent les mêmes étapes de transformation.

Étape 5. Les mots constituant le sac de mots sont triés par ordre alphabétique. Ainsi, deux expressions contenant les mêmes mots dans un ordre différent aboutiront au même sac de mots.

Les descripteurs MeSH et leurs synonymes sont également traités de la même manière que le texte cela en passant par les étapes 1 jusqu'à l'étape 5 et enregistrés dans la base de données sous trois formes : sac de mots initial sans aucune transformation, sacs de mots dé-suffixés et sacs de mots phonémisés (la dé-suffixation et la phonémisation sont expliquées aux étapes 9 et 10).

Étape 6. Le sac de mots du texte est apparié aux sacs de mots des descripteurs français, et aux sacs de mots des synonymes des descripteurs français. S'il y a une correspondance alors le descripteur français trouvé (correspondant au sac de mots de descripteur français) est ajouté à l'index automatique.

Étape 7. S'il n'y a aucune correspondance trouvée à l'étape 6, le sac de mots du texte est apparié aux sacs de mots des descripteurs anglais et aux sacs de mots des synonymes des descripteurs anglais.

Étape 8. S'il n'y a aucune correspondance, alors les mots (constituant le sac de mots du texte obtenu à l'étape 5) sont réduits à leur racine. Par exemple, les mots « Altérer » « Altérée » « Altérés » « Altérées » « Altérant » « Altération » sont tous regroupés sous la racine « alter ».

Étape 9. Le sac de mots dé-suffixé (résultant de l'étape 8) est apparié aux sacs de mots dé-suffixés des descripteurs français et de leurs synonymes français.

[2] Cette mesure est définie dans la Section 2.3.

[3] La valeur de n est fixée expérimentalement.

```
Fonction Indexation
Entrées
titre_sous_titres, sac_mots_descrip_syno_français, sac_mots_descrip_syno_anglais, sac_mots_descrip_syno_désuffixés,
sac_mots_descrip_syno_phonémisés /* ces tableaux contiennent respectivement les sacs des mots : des descripteurs Français et leur synonymes, descripteurs anglais et leurs
synonymes, des descripteurs dé-suffixés et phonétisés*/
Sorties
Index_auto_text
Début
Texte_désaccentué←Désaccentuer (titre_sous_titres)
Texte_sans_ponctuation←Enlever_ponctuation (Texte_désaccentué)
mots←Segmenter_mot (Texte_sans_ponctuation)
Mots_pleins←Enlever_mots_vides (mots)
Sacs_mots_texte ←Découper_en_portion(Mots_pleins, n) /* Cette fonction découpe les mots de texte en portions de n mots (n est le nombre de mots dans un sac de mots) et
retourne le tableau Sacs_mots_texte contenant tous les sacs de mots du texte */
t←0 ;
Pour f de 1 à nsmot faire // nsmot est le nombre de sacs de mots du texte
        Sac_mots_txt ←Sacs_mots_texte[f]
        Sac_mots_texte←Classer_mots_ordre_alphabétique(Sac_mots_txt)
        u←1
        descripteur_français ←« »
        sac_mot_et_sous_ensembles←générer_sous_ensemble_sac_mots(Sac_mots_texte)
/*la fonction générer_sous_ensemble_sac_mots génère les sous-ensembles de sacs de mots et retourne le tableau sac_mot_et_sous_ensemble qui contient le sac de mots (le premier élément du
tableau) et ses sous-ensembles triés par ordre décroissant selon leurs tailles */
  Tant que (descripteur_français = « ») et (u≤nsesm) faire
        /* nsesm est la taille du tableau sac_mot_et_sous_ensemble*/
        Pour h de 1 à nbsmdf faire /*nbsmdf désigne le nombre de sacs de mots des descripteurs français et leurs synonymes */
                Si (sac_mot_et_sous_ensembles[u] = sac_mots_descrip_syno_français[h]) alors
                        descripteur_français←Trouver_descripteur_français(sac_mots_descrip_syno_français[h])
                        // la fonction Trouver_descripteur_français permet de retourner le descripteur français d'un sac de mot
                        Index_auto_text[t] ← descripteur_français
                        t←t+1
                Fin si
        Fin pour
        Si (descripteur_français = « ») alors
                Pour k de 1 à nbsmda faire /* nbsmda désigne le nombre de sacs de mots des descripteurs anglais et leurs synonymes */
                        Si (sac_mot_et_sous_ensemble[u] = sac_mots_descrip_syno_anglais[k]) alors
                                descripteur_français ← Trouver_descripteur_français(sac_mots_descrip_syno_anglais [k])
                                Index_auto_text[t] ← descripteur_français
                                t←t+1
                        Fin si
                Fin pour
                Fin si
        Si (descripteur_français = « ») alors
                Sac_de_mots_texte_désuffixé←Désuffixation_sac_mots_texte(sac_mot_et_sous_ensembles[u])
                Pour j de 1 à nbsmds faire /* nbsmds désigne le nombre de sacs de mots des descripteurs et leurs synonymes dé-suffixés*/
                        Si (Sac_de_mots_texte_désuffixé = sac_mots_descrip_syno_désuffixés [j]) alors
                                descripteur_français← Trouver_descripteur_français(sac_mots_descrip_syno_désuffixés[j])
                                Index_auto_text[t] ← descripteur_français
                                t←t+1
                        Fin si
                Fin pour
        Fin si
        Si (descripteur_français = « ») alors
                Sac_mots_texte_phonémisé←Phonemisation(sac_mot_et_sous_ensemble[u])
                Pour i de 1 à nbsmdf faire /* nbsmds désigne le nombre de sacs de mots des descripteurs et leurs synonymes phonémisés*/
                        Si (Sac_mots_texte_phonémisé = sac_mots_descrip_syno_phonémisés[i]) alors
                                descripteur_français← Trouver_descripteur_français(sac_mots_descrip_syno_phonémisés[i])
                                index_auto_text[t] ← descripteur_français
                                t←t+1
                        Fin si
                Fin pour
        Fin si
        u←u+1
        Fin tant que
Fin pour
Retourne (Index_auto_text)
Fin Indexation
```

Fig. 1. La fonction d'indexation.

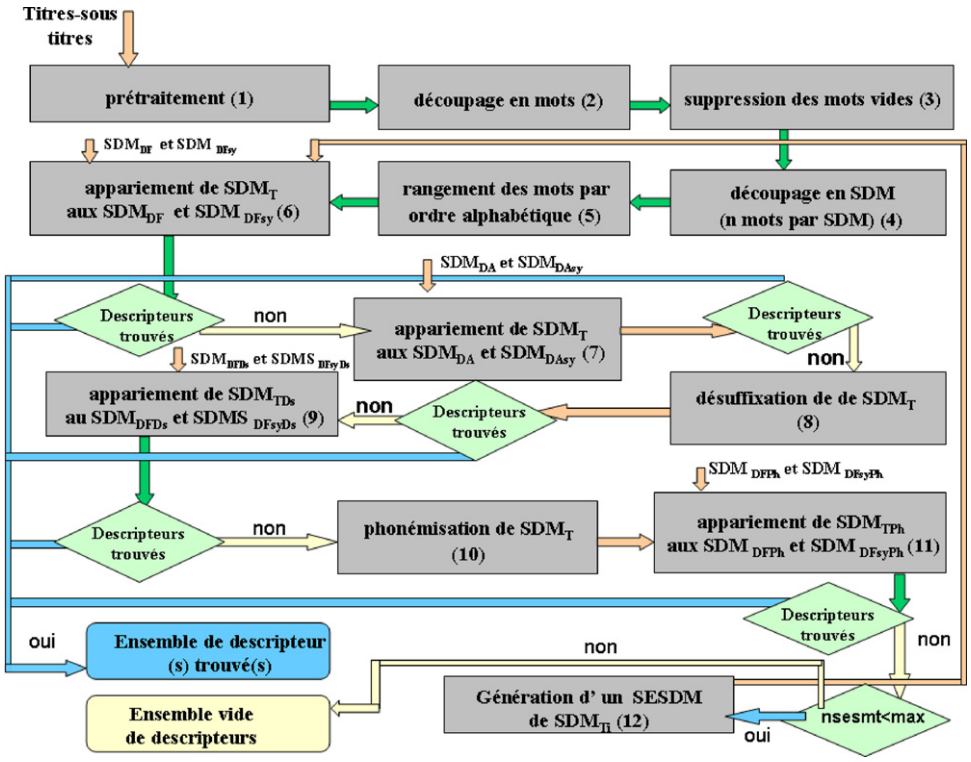

Fig. 2. Processus de l'indexation automatique utilisé dans CISMeF. SDM : sac de mots ; SDM_T : sac de mots du texte ; SDM_{DF} : sac de mots des descripteurs français ; SDM_{DFsy} : sac de mots des synonymes de descripteurs français ; SDM_{DA} : sac de mots des descripteurs anglais ; SDM_{DAsy} : sac de mots de synonymes de descripteurs anglais ; SDM_{TDs} : sac de mots du texte dé-suffixé ; SDM_{DFDs} : sac de mots des descripteurs français dé-suffixés ; SDM_{DFsyDs} : sac de mots de synonymes des descripteurs français dé-suffixés ; SDM_{DFPh} : sac de mots des descripteurs phonémisés français ; SDM_{DFsyPh} : sac de mots des synonymes des descripteurs phonémisés français ; SESDM : sous-ensemble de sac de mots ; SDM_{Ti} : sac de mots du texte initial ; nsesmt : nombre de sous-ensembles de sacs de mots traités ; Max : est le nombre maximal de sous-ensembles d'un sac de mots initial du texte.

Étape 10. Si aucun descripteur n'est trouvé à l'étape 8, les mots (constituant le sac de mots de texte obtenu à la phase 5) sont phonémisés : chaque mot est remplacé par son code phonétique grâce à la fonction décrite en détails dans [3].

Étape 11. Le sac de mots phonémisé est ensuite apparié aux sacs de mots phonémisés des descripteurs français et de leurs synonymes français.

Étape 12. Si à l'étape 11 le sac de mots n'a pas été apparié à un descripteur, les sous-ensembles de sac de mots initial (obtenu à l'étape 5) sont générés et traités l'un après l'autre par ordre décroissant selon leurs tailles, en passant par l'étape 5 jusqu'à 12. Le processus s'arrête dès qu'un descripteur est identifié.

Si tous les sous-ensembles de sacs de mots initiaux sont traités et aucun descripteur n'a pas été trouvé alors le sac de mots du texte en entrée n'a pas de descripteur correspondant.

2.2. Corpus de test

Le corpus utilisé pour l'évaluation de la fonction d'indexation décrite est composé de 500 documents en langue française de type recommandation sélectionnés aléatoirement à partir de 5 741 ressources indexées manuellement avant l'évaluation actuelle de la fonction d'indexation. Il faut préciser que l'indexation manuelle, qui représente le « gold standard » selon la littérature [2], est réalisée sur l'ensemble du document et pas seulement sur son titre et sous-titres ou sur son résumé.

Nous avons indexé automatiquement les titres et sous-titres concaténés[4] des 500 ressources ainsi que les résumés de 200 ressources sélectionnées aléatoirement parmi les 500.

Dans CISMeF sont indexés essentiellement trois types de ressources : ressources de type documentation pour le patient, ressources de type recommandation pour le professionnel et ressources de type enseignement pour les étudiants en médecine. Nous n'avons pas choisi le type de documents enseignement, puisque l'évaluation de l'indexation des titres de ces documents a été réalisée antérieurement dans [26]. De plus, Névéol et al. [26] ont indiqué que le type de ressources enseignement est caractérisé par des titres courts ce qui pouvait expliquer les résultats satisfaisants trouvés. L'évaluation de l'indexation des documents de type recommandation permettra de compléter l'étude réalisée dans [26], puisque les documents que nous étudions dans cet article sont caractérisés par des titres courts et longs à la fois.

Les ressources de type documentation pour le patient vont être l'objet d'une autre évaluation.

Afin d'étudier la performance de la fonction d'indexation en fonction de la longueur des titres et des sous-titres (courts ou longs), nous avons classé les titres et les sous-titres en titres et sous-titres longs et titres et sous-titres courts, cela en considérant

[4] Dans les exemples qui vont suivre les titres et les sous-titres sont séparés par un tiret.

Tableau 1
Caractéristiques du corpus de test.

	Nombre total de mots	Nombre moyen de mots
Titres et sous-titres courts	715	6,62
Titres et sous-titres longs	7885	20,11
Total des titres et sous-titres	8600	17,2
Résumés	28 800	144

que le nombre maximal de mots d'un titre et sous-titre court est égal au nombre minimal de mots d'un titre et sous-titre long qui est égale à neuf. Ce choix approximatif de seuil, nous a permis de fixer le nombre de titres et sous-titres longs (392 titres et sous-titres longs), qui représente 78,4 % de corpus de test, et le nombre de titre et sous-titres courts (108 titres et sous-titres courts). Le nombre moyen de mots d'un titre et sous-titre court (6,62) et le nombre moyen de mots d'un titre et sous-titre long (20,11) trouvés nous ont paru convenables, ce qui nous ont poussés à ne pas changer le seuil choisi.

Le nombre total des mots et le nombre moyen de mots du corpus de test sont détaillé dans le Tableau 1. Les Tableaux 2–4 présentent des exemples de l'indexation automatique et manuelle d'un titre court, titre long et un résumé.

2.3. Mesures d'évaluation

Pour évaluer la performance de la fonction d'indexation automatique nous avons utilisé les mesures d'évaluation connues qui sont la précision, le rappel et F-mesure.

La mesure F-mesure combine le rappel et la précision avec un poids égal [27].

$$\text{Précision} = \frac{\text{Nombre de descripteurs corrects extraits par la fonction d'indexation}}{\text{Nombre total de descripteurs extraits la fonction d'indexation}}$$

$$\text{Rappel} = \frac{\text{Nombre de descripteurs corrects extraits par la fonction d'indexation}}{\text{Nombre total de descripteurs extraits manuellement}}$$

$$\text{F-mesure} = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision+rappel}}$$

Pour analyser les erreurs identifiées, nous avons calculé la fréquence absolue de chaque catégorie d'erreur (les catégories des erreurs sont détaillées dans la Section 3.2) par titre et sous-titre et par résumé :

$$\text{FACE} = \frac{\text{Nombre d'occurence d'une caté gorie d'erreurs dans l'ensemble de documents indexés}}{\text{Nombre total de documents indexés}}$$

FACE : Fréquence absolue d'une catégorie d'erreurs.

Dans le but aussi de comparer les fréquences des catégories d'erreurs dans les titres et sous-titres et dans les résumés, sans que cette comparaison soit biaisée par la longueur du texte indexé, nous avons calculé la fréquence relative des catégories d'erreurs dans l'ensemble de documents indexés.

$$\text{FRCE} = \frac{\text{Nombre d'occurence d'une caté gorie d'erreurs dans l'ensemble de documents indexés}}{\text{Nombre total de mots dans l'ensemble de documents indexés}}$$

FRCE : Fréquence relative d'une catégorie d'erreurs.

Dans la formule de calcul de FACE ou de FRCE, l'ensemble de documents indexés peut être celui des titres et sous-titres courts, des titres et sous-titres longs, l'ensemble des titres et sous-titres ou des résumés.

2.4. Processus d'évaluation

Pour évaluer la fonction d'indexation, nous avons réalisé les étapes suivantes, qui représentent les étapes de l'évaluation d'une seule ressource (illustrées par la Fig. 3).

Étape 1. Nous avons comparé l'indexation automatique du titre et sous-titre et l'indexation automatique du résumé à l'indexation manuelle de l'ensemble de document, vu que l'indexation manuelle des titres et sous-titres seuls et des résumés seuls ne sont pas disponibles.

Étape 2. Si un Descripteur extrait Automatiquement (DA) présent dans l'index manuel (IM), alors incrémenter le nombre de descripteurs corrects.

Étape 3. Déterminer le terme[5] du texte qui a généré DA si ce dernier n'est pas présent dans IM. Cette étape comprend des sous-étapes qui sont comme suit :

- étape 3.1. Supprimer le mot i du texte (i est le numéro du mot à supprimer, initialisé à 1) ;
- étape 3.2. Relancer la fonction d'indexation ;
- étape 3.3. Vérifier si le descripteur non correct (DA) existe toujours dans IA. Si oui alors le mot supprimé n'est pas la source d'erreur. Les sous-étapes de l'étape 3 sont alors ré-exécutées à partir de l'étape 3.1 cela en supprimant le mot suivant du texte initial (s'il n'est pas le dernier mot) et en incrémentant de 1 la valeur de i ;
- étape 3.4. Si DA n'est plus présent dans IA, alors le mot supprimé est ajouté à la liste des mots constituant le terme considéré comme la source d'erreur. Ce terme est appelé Terme source d'erreur (TSE) ;
- étape 3.5. Si la valeur de i atteint nombre de mots du texte (nbrmt) alors TSE est ajouté à la liste des termes source d'erreurs de l'indexation de la ressource.

[5] Un terme peut être constitué d'un ou plusieurs mots.

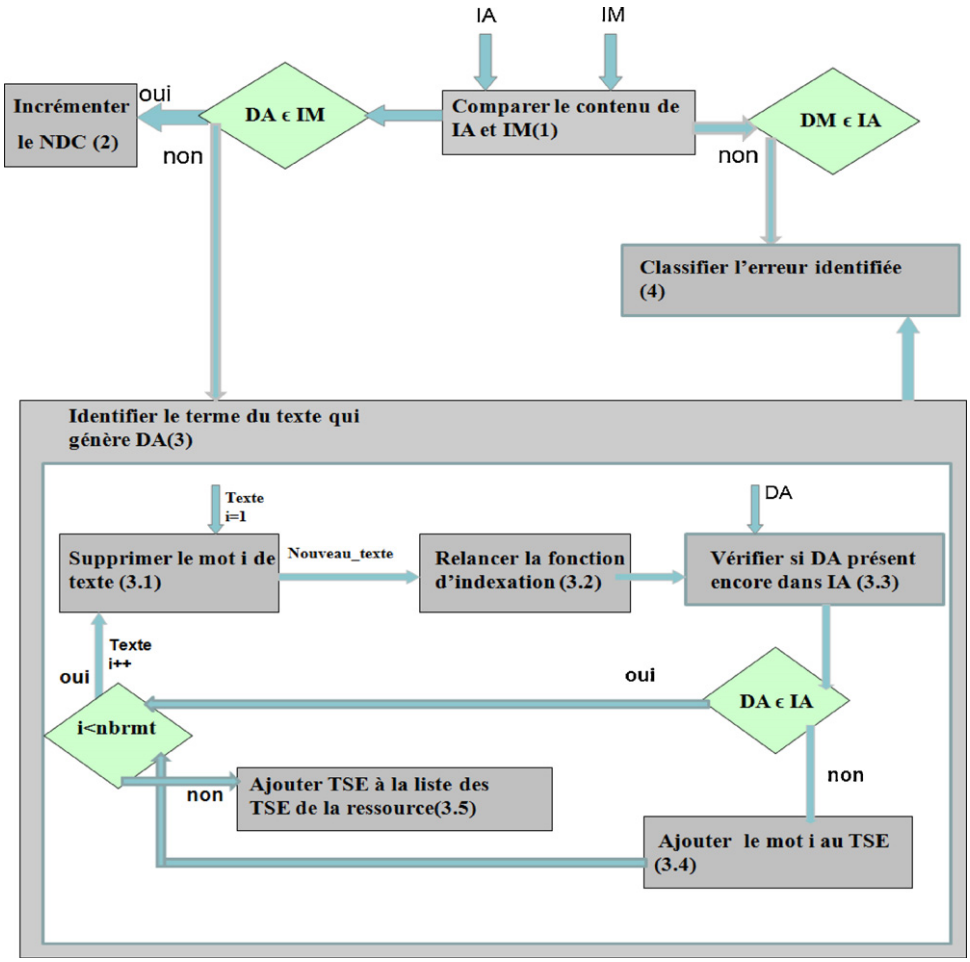

Fig. 3. Processus d'évaluation de l'indexation automatique d'un texte (titre et sous-titres ou résumé). IM : index manuel ; IA : index automatique ; DM : un descripteur extrait manuellement ; DA : un descripteur extrait automatiquement ; NDC : nombre de descripteurs corrects ; TSE : Terme source d'erreur ; i : le numéro du mot supprimé ; Texte : le texte à indexer ; nbrmt : le nombre de mots du texte ; Nouveau_texte : est le texte après la suppression d'un mot.

L'étape 3 est traduite vers une fonction appelée « Liste_Termes_Source_Erreurs » (Fig. 4) qui retourne les termes qui sont à l'origine des erreurs générées dans chaque ressource.

Toute les étapes, de 1 jusqu'à 3, sont réalisées automatiquement.

Étape 4. Au cours de cette étape, les TSE et les descripteurs manquants (présents dans l'IM et absent dans l'index automatique) vont être classifiés selon 13 catégories d'erreurs. Un TSE peut être classifié manuellement selon neuf catégories (les catégories 1 à 9).

Un descripteur manquant peut être classifié selon quatre catégories. Parmi ces quatre catégories, trois sont identifiées automatiquement cela en exécutant une fonction qui parcourt le texte et retourne :

Catégorie 10 : si tous les mots de descripteur manquant sont identifiés dans le contenu de document.

Catégorie 11 : si aucun mot de descripteur manquant n'est identifié dans le texte.

Catégorie 12 : si une partie des mots de descripteur manquant est identifiée dans le texte.

La quatrième catégorie qui est la catégorie 13 est identifiée manuellement.

Toutes les catégories d'erreurs sont détaillées dans la Section 3.2.

Il faut signaler que l'analyse des erreurs (les étapes automatiques et manuelles) a été réalisée par un informaticien (WC), cela en consultants le portail terminologique de santé (PTS) (http ://pts.chu-rouen.fr/) qui comprend toutes les informations détaillées qui sont en relation avec chaque descripteur MeSH (ou d'autres terminologies), notamment sa définition.

Tableau 2
Exemple d'indexation automatique et manuelle d'un titre et sous-titre long d'un document.

Titre et sous-titres long	Titre « Avis relatif à la nécessité d'établir des recommandations particulières sur l'allaitement maternel au vu des bénéfices et des risques d'exposition au chlordécone pour les nourrissons martiniquais et guadeloupéens – Afssa Saisine n° 2007-SA-0350 »
Index automatique	Allaitement maternel ; directives ; exposition maternelle ; képone ; nourrisson ; risque ; risques et bénéfices
Index manuel	Allaitement maternel ; exposition maternelle ; grossesse ; képone ; lait humain ; Guadeloupe ; Martinique ; insecticides ; nourrisson ; nouveau-né

```
Fonction Liste_Termes_Source_Erreurs

Entrée :
Texte   //est le texte à indexer (titre, sous-titres ou résumé)
Sortie :
 Liste_TSE // liste des termes source d'erreur de la ressource
Début
TDNC←Descripteurs_noncorrects(IA, IM)/* la fonction Descripteurs_noncorrects retourne la liste des descripteurs non corrects d'une ressource, cette liste est stockée dans le tableau TDNC  */
T←remplissage(T,Texte) //remplir le tableau T par les mots du texte
g←0
Pour i de 1 à m faire// parcourir TDNC, m est la taille de TDNC
        TSE← «» // initialiser TSE (le  terme source de génération d'un descripteur non correct) à null
        Pour k de 1 à nbrmt  faire // parcourir  T : nbrmt est la taille de T
        texttest←Texte
        texttest ←Supprimer(texttest, T[k])
/* texttest est une variable qui va  recevoir les sous-ensembles de texte après la suppression des mots, utiliser texttest comme une variable intermédiaire permet de garder la valeur initiale de texte*/
        IA←Indexation(texttest,SDfS,SDAS,SDS,SDF)/*faire appel à la fonction d'indexation (figure1), SDfS, SDAS, SDS et SDF sont respectivement les sacs des mots : des descripteurs Français et leur synonymes, descripteurs anglais et leurs synonymes, des descripteurs dé-suffixés et phonémisés*/
        Trouve← false
        j←0
                Tant que (j<nda) et (trouve =false) faire //parcourir IA, nda est la taille de IA
                        Si (TDNC[i] = IA[j])  alors
                        Trouve ←true
                        Fin si
                        j←j+1
                Fin tant que
                Si (trouve =false) alors
                // concaténer TSE à l'espace ceci pour avoir à la fin un terme constitué des mots séparés par un espace
                        TSE ←Concat(«  »,TSE)
                        TSE←Concat(T[k], TSE)
                Fin si
        Fin pour
Liste_TSE [g] ←TSE
g←g+1 ;
Fin pour
retourne(Liste_TSE)
Fin  Liste_Termes_Source_Erreurs
```

Fig. 4. La fonction Liste_Termes_Source_Erreurs.

3. Résultats

3.1. Performance de l'indexation automatique des titres, sous-titres et résumés

Le Tableau 5 présente les résultats de l'évaluation de l'indexation automatique des titres et sous-titres courts, des titres et sous-titres longs, de l'ensemble des titres et sous-titres et des résumés.

Tableau 3
Exemple d'indexation automatique et manuelle d'un titre et sous-titre court d'un document.

Titre	Traitement local de l'otite externe
Index automatique	Thérapeutique, otite externe
Index manuel	Acide acétique, antibactériens glucocorticoïdes, otite externe administration par voie topique ; adulte ; association de médicaments ; résultat thérapeutique ; étude comparative

3.2. Analyse des erreurs

Nous avons classé les erreurs identifiées selon 13 catégories d'erreurs, cela pour les titres et sous-titres ainsi que pour les résumés. Chaque catégorie correspond à une cause de génération d'une erreur. Il faut préciser que nous avons classifié toute erreur identifiée liée ou non l'indexation automatique. On peut distinguer les catégories d'erreurs qui sont à l'origine de la présence de descripteurs non corrects dans l'index automatique, ces catégories sont 1, 2, 3, 4, 5, 6, 7, 8 et 9, et celles qui sont à l'origine de l'absence des descripteurs pertinents dans l'index automatique à savoir les catégories 10, 11, 12 et 13. Nous présentons dans ce qui suit une définition et un exemple de chaque catégorie d'erreur.

Catégorie 1. Regroupe les erreurs liées au mauvais ordre des mots suite à la génération des sous-ensembles de sacs de mots.

Cette erreur se produit à l'étape 12, où fonction d'indexation génère les sous-ensembles de sacs de mots sans prendre compte l'ordre exact des mots dans le texte. Cette tâche peut entraîner la génération des descripteurs non pertinents.

Tableau 4
Exemple d'indexation automatique et manuelle d'un résumé d'un document intitulé « Recommandations de l'Académie Nationale de Médecine concernant la prise en charge extrahospitalière de l'arrêt cardiocirculatoire ».

Résumé	Les arrêts cardiocirculatoires inopinés sont responsables d'environ 50 000 morts subites par an en France. Plus de la moitié d'entre eux sont liés à une fibrillation ventriculaire. Le taux de survie observé à 1 mois est actuellement inférieur à 3. Un appel immédiat aux unités mobiles de secours, des manœuvres simples de réanimation à la portée de tous (massage cardiaque externe en particulier), une défibrillation cardiaque très précoce, devraient pouvoir faire passer ce taux de survie à plus de 3. L'apparition des défibrillateurs externes entièrement automatiques doit permettre leur utilisation par l'ensemble de la population informée
Index automatique	*Acanthosis nigricans* ; France ; mort subite ; massage cardiaque ; mobile ; réanimation ; premiers secours ; unités du SI ; défibrillation ; survie ; automatisme ; population ; sous-population, défibrillateurs
Index manuel	Arrêt cardiaque ; défibrillateurs ; éducation sanitaire ; gestion des soins aux patients ; massage cardiaque ; mort subite cardiaque ; premiers secours ; services des urgences médicales

Exemple : Titre : « Recommandations de la Société française de cardiologie concernant les conditions de compétence, d'activité et d'environnement requises pour l'implantation et la surveillance des stimulateurs cardiaques ».

Le descripteur non correct généré est « surveillance de l'environnement ».

Catégorie 2. Regroupe les erreurs liées à la génération d'un sous-ensemble de sac de mots constitué d'un seul mot.

La fonction d'indexation peut générer à l'étape 12 un sous-ensemble de sac de mots constitué d'un seul mot, ce dernier peut être apparié à un descripteur sans considérer sa liaison avec les mots voisins. On peut alors trouver des descripteurs très généraux et moins précis par rapport au contenu du texte.

Exemple : Le Tableau 3 présente l'indexation automatique et manuelle d'un titre et sous-titre court.

Le descripteur « thérapeutique » est généré vu la présence de son synonyme MeSH « traitement » dans le titre. Ce descripteur est très général par rapport au sens précis du titre.

Catégorie 3. Regroupe les erreurs liées à la présence d'acronymes.

La fonction d'indexation peut ne pas distinguer le bon terme d'un acronyme (plusieurs termes peuvent avoir le même acronyme) (Exemple 1). Elle peut également interpréter un mot comme étant un acronyme (Exemple 2). Cela est du au manque d'une étape de détection d'acronyme.

Exemple 1 : Titre et sous-titre : « Gestion d'une structure anatomie et cytologie pathologique (ACP) – Recommandations et réglementations ».

Le descripteur non correct est « protéine ACP ». Ce descripteur est généré car son synonyme MeSH est « Acyl Carrier Protein ». Le second descripteur non correct est « protocol ACP », ce dernier est généré car son synonyme anglais est « Adriamycine Cyclophosphamide Prednisone ».

Exemple 2 : Titre : « Circulaire interministérielle DGS/DUS/DHOS/DSC/DGAS/2009/358 du 30/11/2009 précisant les actions à mettre en œuvre au niveau local pour prévenir et faire face aux conséquences sanitaires propres à la période hivernale ».

Le descripteur non correct est « dsc ». Ce descripteur est un acronyme du terme « Syndrome de De Sanctis-Cacchione », or le sens de ce terme ne correspond pas au contexte du titre.

Catégorie 4. Regroupe les erreurs liées à l'ambiguïté d'un terme.

Un terme du texte peut être ambigu en ayant plusieurs sens. Dans ce cas, la fonction d'indexation ne peut pas distinguer le bon sens du terme selon le contexte de la phrase à laquelle il appartient. Il est nécessaire donc d'intégrer une étape de désambiguïsation sémantique.

Exemple : Titre : « Tatouages éphémères noirs à base de henné : mise en garde ».

Le descripteur non correct est « population d'origine africaine ». Ce descripteur est généré car son synonyme « noirs » est présent dans le titre. Le mot « noirs » en français possède deux sens : la couleur noir et la population d'origine africaine, dans cet exemple la fonction d'indexation n'a pas distingué le bon sens du mot.

Catégorie 5. Regroupe les erreurs liées aux mots vides.

La table des mots vides doit être mise à jour. L'ajout de quelques mots vides est nécessaire vu que leur présence dans le texte cause des erreurs d'indexation.

Exemple : Titre : « Avis n° 38 du 13 novembre 2006 relatif aux tests génétiques en vue d'établir la filiation après le décès ».

Le descripteur non correct est « vision oculaire ». Ce descripteur est généré vu que son synonyme MeSH (« vue ») présent dans le titre. Dans ce cas il est nécessaire d'ajouter le terme « en vue de » à la liste des mots vides puisqu'il est fréquent dans les ressources, il n'est pas utile pour l'indexation et comprend un mot qui génère des erreurs.

Catégorie 6. Regroupe les erreurs liées à la détection de la langue du texte.

Les deux vocabulaires français et anglais (actuellement CISMeF n'indexe que des documents en santé en français et en

Tableau 5
Performance de l'indexation automatique des titres, sous-titres et des résumés.

	Précision	Rappel	F-mesure	Nombre de descripteurs extraits manuellement	Nombre de De descripteurs extraits automatiquement	Nombre de descripteurs corrects
Titres et sous-titres courts	0,56	0,21	0,30	641	240	135
Titres et sous-titres longs	0,39	0,27	0,30	2642	1845	720
Total des titres et sous-titres	0,41	0,26	0,31	3283	2085	855
Résumés	0,23	0,61	0,33	1295	3434	790

anglais) ont un ensemble de mots en commun et qui n'ont pas nécessairement le même sens or la fonction d'indexation ne distingue pas si le terme qui va être apparié aux descripteurs MeSH est un terme français ou un terme anglais. Par conséquent, un mot en anglais peut être considéré comme un mot en français (et inversement).

Exemple : Titre : « Instruction DGAS/3B n° 2008-167 du 20 mai 2008 relative aux groupes d'entraide mutuelle pour personnes handicapées psychiques ».

Le descripteur non correct est « films et vidéos pédagogiques ». Ce descripteur est généré vu que son synonyme MeSH en anglais est « instruction » (étape 7 de l'indexation). La fonction d'indexation n'a pas détecté que la langue du titre est le français et elle a interprété « instruction » comme un mot en anglais.

Catégorie 7. Regroupe les erreurs liées à l'étape de dé-suffixation.

Cette erreur est due à l'étape 8 de la fonction d'indexation automatique. Lors de cette phase, les mots constituants les sacs de mots du texte sont réduits à leur racine. L'algorithme d'indexation cherche ensuite les sacs de mots des descripteurs qui ont la même racine que les sacs de mots du texte. Cependant, les mots qui ont la même racine n'ont pas tous nécessairement le même sens, ce qui peut expliquer la présence des descripteurs non pertinents dans l'index automatique.

Exemple : Titre et sous-titre : « Recommandations de dépistage et de diagnostic de l'autisme et des autres troubles envahissants du développement (TED) – recommandations pour les équipes spécialisées dans les troubles du développement ».

Le descripteur incorrect est « équipement et fournitures ».

La fonction d'indexation a retourné ce descripteur car il a la même racine que le terme « équipe ».

Catégorie 8. Regroupe les erreurs liées à l'ordre des étapes de la fonction d'indexation.

L'exécution de quelques étapes de la fonction d'indexation avant d'autres génère des erreurs, ce qui nécessite le changement de leur ordre. Comme la tâche de désaccentuation qui est réalisée avant la suppression des mots vides or il y a des mots vides accentués, ce qui empêche leur suppression comme : « à » et « après ».

La tâche de suppression des signes de ponctuation tels que les apostrophes « ' » est réalisée avant la suppression des mots vides ce qui empêche la suppression des mots vides comme « l'un » « j' » qui sont fréquents dans les titres et sous-titres.

L'étape de segmentation en mots est réalisée avant l'étape de suppression des mots vides ce qui empêche la suppression des mots vides composés comme : « en particulier », « tandis que », etc.

Catégorie 9. Regroupe les erreurs liées à la suppression du tiret « - » lors de l'étape 2 de l'indexation.

Lors de l'analyse de la fonction d'indexation nous avons remarqué l'importance de la présence de tiret entre deux mots. En effet, dans le cas ou un mot vide et un mot non vide sont séparés par un tiret, ces deux mots constituent un seul terme non vide. La suppression de mot vide dans ce cas entraîne le changement du sens du terme. Ainsi, le tiret doit être supprimé de la liste des séparateurs.

Exemple : Le Tableau 4 présente l'indexation automatique et manuelle d'un résumé.

Le descripteur « sous-population » est un descripteur non correct. Il est généré suite à la suppression du mot vide « sous », le mauvais descripteur devient alors « population » ce qui correspond à un mot du texte.

Catégorie 10. Regroupe les erreurs liées à l'étape 1 de l'évaluation de la fonction d'indexation (Section 2.4).

Cette catégorie d'erreurs est liée à une étape de l'évaluation qui consiste à comparer l'indexation manuelle de l'ensemble du document à l'indexation automatique des titres et sous-titres et l'indexation automatique des résumés. En effet, le contenu d'un document peut comprendre des descripteurs pertinents qui n'existent pas dans le titre et sous-titres ou dans le résumé. Par conséquent, ces descripteurs ne vont pas être générés par la fonction d'indexation. Si l'indexation manuelle des titres et sous-titres seuls et l'indexation manuelle des résumés seuls étaient disponibles, cette catégorie d'erreurs ne serait pas rencontrée.

Exemple : Le Tableau 2 présente l'indexation automatique et manuelle d'un titre et sous-titre long. On peut remarquer que le descripteur « insecticides » est manquant, vu qu'il est présent que dans le contenu du texte (n'est pas présent dans le titre et sous-titre).

Catégorie 11. Regroupe les erreurs liées aux descripteurs implicites.

Les descripteurs implicites sont les descripteurs (et leurs synonymes) qui ne sont pas présents (même un seul mot du descripteur) dans le texte, mais jugés pertinents pour l'indexation de document par les indexeurs, cela en se basant sur le sens du texte. Bien évidemment, la fonction d'indexation ne peut pas générer ces descripteurs.

Exemple : Titre et sous-titre « Bon usage des opioïdes forts dans le traitement des douleurs chroniques non cancéreuses – Mise au point ».

Le descripteur extrait manuellement est « soins palliatifs », ce descripteur n'existe ni dans le titre et sous-titre ni dans le contenu. L'indexeur a déduit que ce descripteur est pertinent d'après le sens du texte.

Catégorie 12. Regroupe les erreurs liées aux descripteurs partiellement explicites dans le texte.

Les descripteurs partiellement explicites sont des descripteurs pertinents pour l'indexation de documents (présents dans l'IM) et n'ayant qu'un sous-ensemble de leurs mots dans le texte ce qui aide les indexeurs à trouver ces descripteurs. Cependant, la fonction d'indexation ne peut pas les générer même si la partie explicite des descripteurs existe dans les titres et sous-titres ou dans les résumés.

Exemple : Titre : « Évaluation médicale de l'aptitude à conduire ».

L'IM de ce titre comprend le descripteur « examen du permis de conduire automobile », ce descripteur n'existe pas tel qu'il est dans le texte ce qui n'a pas permis à la fonction d'indexation de l'extraire. Cependant, des sous-ensembles de ce descripteur sont présents dans le texte, ces descripteurs sont « permis de conduire » et « conduite automobile ».

Tableau 6
Fréquences absolues des catégories d'erreurs par titre et sous-titre et par résumé.

Catégories d'erreurs	Fréquence absolue par titre et sous-titre court	Fréquence absolue par titre et sous-titre long	Fréquence absolue par titre et sous-titre (court ou long)	Fréquence absolue par résumé
1	0,0270	0,094	0,08	0,2
2	0,638	1,737	1,5	5,71
3	0,027	0,119	0,1	0,4
4	0,037	0,104	0,09	0,94
5	0,027	0,056	0,05	1,4
6	0,018	0,099	0,082	0,6
7	0,175	0,617	0,522	3,64
8	0,009	0,017	0,016	0,25
9	0,009	0,022	0,02	0,08
10	2,7	2,57	2,6	0,91
11	0,92	1,007	0,99	0,71
12	0,98	1,22	1,17	0,80
13	0,074	0,094	0,09	0,1

Catégorie 13. Regroupe les erreurs liées à la variation morphologique de type dérivation.

Dans la langue française, il existe des mots d'une même famille mais qui n'ont pas la même racine, c'est la variation morphologique de type dérivation [3]. Cette variation permet d'ajouter les suffixes et les préfixes autour d'une racine pour obtenir par exemple la forme adjectivale d'un mot (exemples : enfant/infantile et cœur/cardiaque). Si un descripteur est une variante morphologique de type dérivation d'un terme du texte, ce descripteur ne va pas être extrait par la fonction d'indexation.

Exemple 2 : Titre : Botulisme infantile (Le).

Le descripteur « enfant » est un descripteur manquant vu que la fonction d'indexation n'a pas reconnu que « infantile » et « enfant » sont de la même famille.

4. Discussion

Les résultats présentés dans le Tableau 5 montrent une précision acceptable ($p = 0{,}56$) pour les titres et sous-titres courts et une précision faible ($p = 0{,}39$) pour les titres et sous-titres longs. Cette baisse de la valeur de précision est due à l'augmentation du nombre de mots par titre et sous-titre (6,62 mots par titre et sous-titres court vs 20,11 mots par titre et sous-titres long, d'après le Tableau 1), ce qui génère plus d'erreurs d'indexation liées à la présence des descripteurs non corrects dans l'index automatique (0,97 erreur par titre et sous-titres court vs 2,67 erreurs par titre et sous-titres long d'après le Tableau 6 et la Fig. 5). Le nombre élevé des titres et sous-titres longs dans le corpus de test, qui dépasse le triple de nombre des titres et sous-titres courts (108 titres et sous-titres courts vs 392 titres et sous-titres longs), explique donc la précision faible totale de 0,41 des titres et sous-titres.

On peut également remarquer que, d'après le Tableau 5, les valeurs de rappel sont faibles indépendamment de la longueur des titres et sous-titres. Cela est dû essentiellement à la nature des titres et des sous-titres qui ne contiennent pas tout l'ensemble des mots clés du document. En effet, l'analyse des erreurs présentées dans le Tableau 7 et la Fig. 6 indique que 53,6 % des descripteurs extraits manuellement et ne sont pas présents dans les index automatique des titres (courts ou longs) sont des descripteurs qui existent dans le contenu de document (catégorie d'erreurs n° 10).

Concernant les résumés, la valeur du rappel dépasse de 0,35 celle des titres et sous-titres, Cette augmentation est normale, étant donné que le résumé comprend plus de mots clés et reflète plus le contenu de document.

En revanche, la précision de l'indexation des résumés est très faible ($p = 0{,}23$) cela est expliqué par le nombre élevé de descripteurs non corrects présents dans l'index automatique. En effet, d'après le Tableau 6 et la Fig. 5, on trouve 13,22 erreurs par résumé liées à la présence des descripteurs non corrects parmi 17,17 descripteurs extraits automatiquement

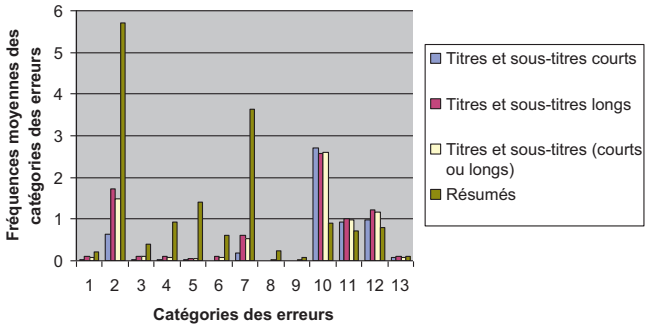

Fig. 5. Fréquences absolues des catégories d'erreurs par titre et sous-titre et par résumé.

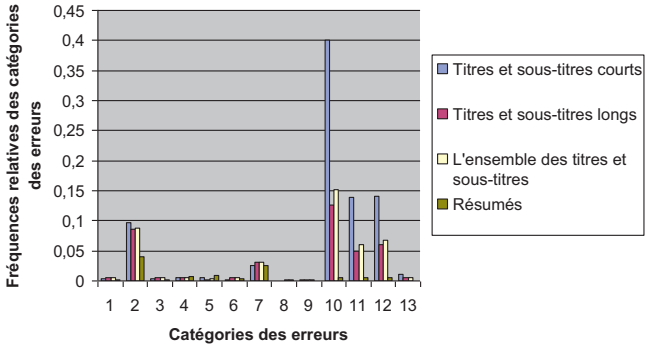

Fig. 6. Fréquences relatives des catégories d'erreurs dans les titres et sous-titres et dans les résumés.

Tableau 7
Fréquences relatives des catégories d'erreurs dans les titres et sous-titres et dans les résumés.

Catégorie d'erreurs	Fréquence relative dans les titres et sous-titres courts	Fréquence relative dans les titres et sous-titres longs	Fréquence relative dans l'ensemble des titres et sous-titres	Fréquence relative dans les résumés
1	0,0040	0,0046	0,0046	0,0013
2	0,0963	0,0863	0,0872	0,0396
3	0,0040	0,0059	0,0058	0,0027
4	0,0055	0,0051	0,0052	0,0065
5	0,0047	0,0027	0,0029	0,0097
6	0,0027	0,0049	0,0047	0,0041
7	0,0264	0,0305	0,0303	0,0252
8	0,0001	0,0008	0,00093	0,0017
9	0,0001	0,001	0,0011	0,0005
10	0,4	0,127	0,1511	0,0063
11	0,139	0,05	0,0601	0,0049
12	0,14	0,061	0,068	0,0055
13	0,011	0,0046	0,0052	0,0006

par résumé. Cela est expliqué par le nombre de mots élevé présent dans les résumés (144 mots par résumé) dont la plupart ne sont pas des mots clés. Or, la fonction d'indexation prend en considération tous les mots sans exception qui ont un descripteur MeSH correspondant, ce qui augmente le nombre de descripteurs non corrects. Par exemple, dans le Tableau 4, on remarque que les descripteurs « automatisme », « population » et « sous-population » sont des descripteurs non corrects qui existent dans l'index automatique et n'appartiennent pas à l'IM. L'adjectif « automatique » est à l'origine de l'apparition de descripteur « automatisme » dans l'index automatique puisqu'ils ont la même racine « automat ». Le nom « population » est à l'origine de deux descripteurs « population » et « sous-population » (« sous » est un mot vide).

Nous accordons de l'importance à la fois à la précision et au rappel. En effet, les patients, les professionnels de santé ou les étudiants en médecine cherchent à trouver une information médicale non seulement précise, ce qui nécessite l'amélioration de la précision, mais aussi complète, ce qui nécessite l'amélioration du rappel.

Nous analysons dans la suite la fréquence absolue et la fréquence relative des catégories d'erreurs dans les titres et sous-titres et dans les résumés.

D'après les Tableaux 6 et 7 et les Fig. 5 et 6, nous avons constaté que les catégories d'erreurs de 1 à 9, qui sont liées à la présence des descripteurs non corrects dans l'index automatique, ont une fréquence absolue beaucoup plus importante dans les résumés que dans les titres et sous-titres, mais une fréquence relative à peu près égale dans les titres et sous-titres et dans les résumés. De plus, les catégories d'erreurs 10, 11 et 12, qui sont liées à l'absence des descripteurs pertinents dans l'index automatique, ont une fréquence absolue et une fréquence relative plus importantes dans les titres et sous-titres que dans les résumés. On peut déduire alors d'après cette analyse que la perte de précision sur les résumés est uniquement due au fait que les résumés sont plus longs. Par conséquent, si on corrige, même partiellement, ces erreurs, l'indexation par les résumés pourraient potentiellement devenir plus intéressante que celle par les titres et sous-titres.

Les résultats trouvés nous permettent de déduire que la fonction d'indexation est actuellement applicable que pour les phrases simples et courtes comme celles des titres et sous-titres courts. L'utilisation de la fonction d'indexation, initialement destinée pour faire de l'extension des requêtes des utilisateurs, pour indexer les ressources de CISMeF a, sans doute, diminué ces performances. D'abord les requêtes des utilisateurs sont des phrases simples et courtes or les titres peuvent être courts ou longs et les résumés sont longs. De plus, les requêtes des utilisateurs peuvent comprendre des fautes d'orthographes, ce qui explique l'utilité de l'étape de phonémisation pour l'extension des requêtes. En revanche, les fautes d'orthographe sont presque absentes dans les ressources médicales puisqu'elles sont collectées à partir des sites de qualité tels que les sites institutionnels. L'étape de phonémisation est ainsi inutile pour l'indexation automatique des ressources et peut aussi générer des erreurs d'indexation.

Il faut signaler que l'équipe CISMeF a réalisé en 2007 une évaluation de l'indexation automatique [26] sur les documents de type enseignement. Par rapport à l'étude antérieure, et pour mieux évaluer l'indexation automatique, nous avons augmenté le volume de corpus de test et nous avons choisi un autre type de documents. La fonction d'indexation depuis 2007 n'a pas subi de changement (ajout ou suppression d'étapes) sauf des corrections purement techniques à savoir le règlement de quelques anomalies, l'exclusion de quelques termes et l'optimisation des performances. La comparaison des résultats de l'évaluation réalisée en 2007 et l'évaluation actuelle (Tableau 8) montre une baisse dans la précision (0,54 en 2007 vs 0,41 en 2012) et une augmentation de la valeur de rappel (0,16 en 2007 vs 0,26 en 2012). On peut déduire alors que le changement de type de ressources a un impact important sur la performance de la fonction d'indexation actuelle. En effet, documents de type recommandation sont caractérisés, par rapport aux documents de type enseignement, par des titres longs et courts (le nombre des titres long est supérieur au nombre des titres courts) et la présence des codes dans les titres qui peuvent générer des erreurs de catégorie 3, comme l'expression « DGS/DUS/DHOS/DSC/DGAS/2009/358 du 30/11/2009 » présente dans le titre de l'exemple 2 de catégorie 3 d'erreurs

Tableau 8
Tableau comparatif de [26] et l'évaluation de la fonction d'indexation actuelle de CISMeF.

	Étude de Névéol et al. [26]	Étude de la fonction d'indexation actuelle de CISMeF
Taille de corpus de test	99	500
Type de ressource	Enseignement	Recommandation
Taille des titres	Court	Long et court + sous-titres
Référence	Indexation manuelle de l'ensemble de document	Indexation manuelle de l'ensemble de document
Précision	0,54	0,41
Rappel	0,16	0,26
F-mesure	0,24	0,31

qui a généré le mauvais descripteur « dsc » considéré à tort comme un acronyme. D'après le Tableau 7, cette catégorie d'erreur (catégorie 3) est la troisième la plus fréquente parmi les catégories d'erreurs liées à la présence des descripteurs non corrects dans les titres et sous-titres longs.

L'évaluation de la fonction d'indexation en utilisant les documents de type recommandation qui sont actuellement indexés manuellement, a permis d'identifier les sources et la fréquence des erreurs. Ainsi, proposer des solutions même pour une partie des erreurs permettra sans doute d'améliorer l'indexation automatique de ce type de ressource.

Nous pouvons également comparer notre évaluation à celles d'autres systèmes d'indexation des documents en français par le thésaurus MeSH.

Système d'indexation automatique de documents médicaux (SIAM) [28] a été évalué en utilisant un corpus composé d'un sous-ensemble d'articles de CISMeF de 500 documents. Les résultats sont de 0,59 pour le rappel et de 0,71 pour la précision. Dans [8], Névéol a comparé MeSHMap [29] et NomIndex [14], cela en utilisant un corpus en français « misc » et les ressources en français du corpus « ENFR » [30]. Pour un rang égal à 10 (ce rang représente le nombre de mots clés classés selon un score calculé), les valeurs de rappel et de précision sont respectivement $R = 0{,}22$ et $p = 0{,}12$ pour NomIndex, $R = 0{,}18$ et $p = 0{,}11$ pour MeSHMap. Névéol et al. [31] ont comparé Medical Text Indexer (MTI) [2] et MeSH Automatic Indexer for French (MAIF) [8] en utilisant un corpus parallèle en français composé de ressources sélectionnées aléatoirement à partir des ressources de l'institut canadien de santé dans CISMeF. Le rappel et la précision de MAIF pour un rang égal à 10, sont respectivement de $R = 0{,}39$ et $p = 0{,}27$ et celles de MTI sont $R = 0{,}53$ et $p = 0{,}25$. MTI a été évalué aussi par Trieschnig et al. [32] sur 1000 documents choisis aléatoirement à partir de MEDLINE. La précision trouvée au rang 10 est $p = 0{,}32$. Harrathi et al. [18] ont testé la méthode d'indexation proposée sur le corpus Cross Language Evaluation Forum (CLEF) médical 2007. La valeur de précision moyenne trouvée est $p = 0{,}24$ à une valeur de rappel égale à $R = 0{,}40$. Jonquet et al. ont utilisé trois ressources de données biomédicales pour l'évaluation du système d'indexation qu'ils ont proposé dans [24], ces ressources sont ClinicalTrials.gov (des descriptions d'essais cliniques), Gene Expression Omnibus (des données d'expression génétique), ARRS GOLDMiner (des légendes et descriptions d'images radiologiques). Les valeurs de précision sont respectivement : $p = 0{,}87$, $p = 0{,}88$ et $p = 0{,}73$. Zhou, et al. [22] ont exploité le corpus GENIA 3.02. Les valeurs de précision et de rappel sont $p = 0{,}54$ et $R = 0{,}57$.

Cette comparaison entre les méthodes d'indexation automatique reste approximative, car les corpus de test sont différents et indexés manuellement par des indexeurs différents. En effet, on peut voir l'exemple de l'évaluation de MTI sur deux corpus différents (CISMeF et MEDLINE), où deux valeurs de précision différentes sont trouvées au rang 10. On ne peut pas affirmer alors que les méthodes d'indexation proposées [22,24,28] sont meilleures que celle de CISMeF puisque le corpus de test de [22] et [24] est différent de celui utilisé dans cette étude. Le corpus de test dans [28] est un sous-ensemble de ressources de CISMeF, sauf que l'indexation automatique est réalisé sur l'ensemble du document et comparée aussi à l'indexation manuelle de l'ensemble du document, ce qui est différent de l'étude réalisée ici.

En termes de perspectives, nous envisageons d'abord d'améliorer la fonction d'indexation actuelle de CISMeF en proposant des solutions qui visent à minimiser les erreurs générées. Nous testerons ensuite quelques méthodes existantes pour les comparer à la fonction d'indexation de CISMeF améliorée dans les mêmes conditions en les appliquant sur le même corpus de test.

5. Conclusion

L'objectif de ce travail est de déterminer la performance actuelle de la fonction d'indexation automatique utilisée par CISMeF et d'identifier les causes des erreurs générées. Nous avons commencé d'abord par présenter les étapes du processus d'indexation, qui sont basées sur l'algorithme du « sac de mots » complété par d'autres étapes supplémentaires. Nous avons ensuite comparé l'indexation automatique à l'indexation manuelle, ce qui a permis de mettre en évidence les erreurs d'indexation automatique. En effet, nous avons analysé les erreurs cela en les classant selon des catégories qui correspondent aux causes de leurs générations et en déterminant leurs fréquences moyennes par titre et sous-titres et par résumé ainsi que leurs fréquences relatives.

Nous considérons les résultats de ce travail très utiles. En effet, ils ont permis de conclure que la fonction actuelle d'indexation automatique utilisée dans CISMeF est destinée nécessairement à indexer que des phrases simples et courtes ce qui ne correspond pas à la nature des titres et sous-titres des documents de type recommandation.

Cette évaluation nous a également permis de mettre en évidence l'importance des résumés pour l'indexation des doc-

uments vu qu'ils comprennent les informations très utiles qui décrivent le texte.

Nous envisageons à la suite de cette étude de proposer des solutions possibles pour corriger chaque catégorie d'erreur, le but étant d'améliorer les résultats de l'indexation des titres, des sous-titres et des résumés. Cela entraînera une recherche d'information médicale plus fiable dans le moteur de recherche de CISMeF et permettra aussi d'indexer automatiquement un plus grand nombre de documents en santé.

Références

[1] Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B. CISMeF: a structured health resource guide. Methods Inf Med 2000;39(1):30–5.

[2] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Med Health Info 2004;11(Pt 1):268–72.

[3] Soualmia LF. Étude et évaluation d'approches multiples d'expansion de requêtes pour une recherche d'information intelligente : application au domaine de la santé sur l'Internet. Thèse de l'INSA de Rouen, 2004.

[4] Mörchen F, Dejori M, Fradkin D, Etienne J, Wachmann B, Bundschus M. Anticipating annotations and emerging trends in biomedical literature. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2008:954–62.

[5] Sohn S, Kim W, Comeau DC, Wilbur WJ, Research Paper:. Optimal training sets for Bayesian prediction of MeSH assignment. J Am Med Inform Assoc 2008;15(4):546–53.

[6] Huang M, Névéol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. J Am Med Inform Assoc 2011;18(5):660–7.

[7] Vasuki V, Cohen T. Reflective random indexing for semiautomatic indexing of the biomedical literature. J Biomed Inform 2010;43(5):694–700.

[8] Névéol A. Automatisation des tâches documentaires dans un catalogue de santé en ligne. Thèse de l'INSA de Rouen, France 2005.

[9] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the American Medical Informatics Association Symposium 2001:17–21.

[10] Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinform J 2006;22(6):658–64.

[11] Singhal A. Modern information retrieval: a brief overview. IEEE Data Eng Bull 2001;24(4):35–43.

[12] Hliaoutakis A, Zervanou K, Petrakis EGM. The AMTEx approach in the medical document indexing and retrieval application. Data Knowledge Eng 2009;68(3):380–92.

[13] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-Value/NC-value method. Int J Digit Libr 2000;3(2):117–32.

[14] Pouliquen B, Delamane D, Lebeux P. Indexation des textes médicaux par extraction de concepts et ses utilisations. 6e Journées internationales d'analyse statistique des données textuelles, JADT 2002:617–28.

[15] Lenoir P, Michel JR, Frangeul C, et Chales G. Réalisation, développement et maintenance de la base de données ADM. Med Inform 1981;6:51–6.

[16] Majdoubi J, Tmar M, Gargouri F. Using the MeSH thesaurus to index a medical article: combination of content, structure and semantics. International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES (1) 2009:277–84.

[17] Dinh D, Tamine L. Biomedical concept extraction based on combining the content-based and word order similarities. Symposium On Applied Computing, SAC 2011:1159–63.

[18] Harrathi F, Roussey C, Maisonnasse L, Calabretto S. Vers une approche statistique pour l'indexation sémantique des documents multilingues. Congrès INFORSID INFormatique des ORganisation et Systèmes d'Information 2010:127–41.

[19] Delbecque T, Zweigenbaum P. Indexation UMLS en français : une expérience. In: Beuscart R, Brunetaud JM, editors. Actes des Journées francophones d'informatique médicale. 2005. p. 1–8.

[20] Volk M, Ripplinger B, Vintar S, Buitelaar P, Raileanu D, Sacaleanu B. Semantic annotation for concept-based cross language medical information retrieval. Int J Med Inform 2002;67(1–3):97–112.

[21] Wermter J, Hahn U. Paradigmatic modifiability statistics for the extraction of complex multi-word terms. Human Language Technology Conference on Empirical Methods in Natural Language Processing, HLT 2005: 843–50.

[22] Zhou X, Zhang X, Hu X. MaxMatcher: biological concept extraction using approximate dictionary lookup. PRICAI Pacific Rim International Conferences on Artificial Intelligence 2006:145–149.

[23] Ruch P, Gobeill J, Lovis C, Geissbühler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak 2008;8(Suppl. 1):S6.

[24] Jonquet C, Coulet A, Shah NH, Musen MA. Indexation et intégration de ressources textuelles à l'aide d'ontologies : application audomaine biomédical. 21e Journées Francophones d'Ingénierie des Connaissances, IC 2010:271–82.

[25] Dai M, Shah NH, Xuan W, Musen MA, Watson SJ, Athey BD, et al. An efficient solution for mapping free text to ontology terms. AMIA Symposium on Translational Bioinformatics 2008; San Francisco, CA, USA.

[26] Névéol A, Pereira S, Kerdelhué G, Dahamna B, Joubert M, Darmoni SJ. Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a French online catalogue. Stud Health Technol Inform 2007;129(Pt 1):407–11.

[27] Manning CD, Schütze H. Fondations of statistical natural language processing. Cambridge, MA: MIT Press; 1999, p. 534–36.

[28] Majdoubi J, Tmar M, Gargouri F. SIAM : système d'indexation des articles médicaux. Conférence Internationale francophone sur l'extraction et la gestion des connaissances, EGC 2010:697–8.

[29] Ruch P, H. Baud R, Geissbühler A. Learning-free text categorization. Conference on Artificial Intelligence in Medicine, AIME LNAI 2003;2780:199–208.

[30] Névéol A, Rogozan A, Darmoni SJ. Indexation automatique de ressources de santé à l'aide de paires de descripteurs MeSH. Actes de Traitement Automatique des Langues Naturelles, TALN 2005;1: 475–80.

[31] Névéol A, Mork JG, Aronson AR, Darmoni SJ. Evaluation of French and English MeSH indexing systems with a parallel corpus. American Medical Informatics Association Annual Symposium Proceedings, AMIA 2005:565–9.

[32] Trieschnigg D, Pezik P, Lee V, et al. MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics 2009;25(11):1412–8.