SIAM: Système d'Indexation des Articles Médicaux

Jihen Majdoubi*, Mohamed Tmar*, Faiez Gargouri*

*Laboratoire MIRACL,
Institut Supérieur d'Informatique et du Multimédia-Sfax Tunisie, Route de Tunis Km 10 B.P. 242- SFAX 3
majdoubi_jihene@yahoo.fr, mohamed.tmar@isims.rnu.tn, faiez.gargouri@fsegs.rnu.tn

1 Contexte

Dans le domaine médical, les articles scientifiques forment une source d'information très riche pour les chercheurs du domaine. Ainsi, durant ces dernières années la gestion efficace de ce volume impressionnant des articles scientifiques a constitué un défi majeur pour les systèmes documentaires médicaux (Medline[1], CisMef[2],...), et les travaux de recherche sur ce thème. Parmi ces travaux, nous pouvons citer l'indexation sémantique qui permet de décrire le contenu des documents à l'aide d'une Ressource Terminologique ou Ontologique (ontologie, thésaurus,. . .) afin de faciliter l'accès, la recherche et l'exploitation du document Khelif (2006), Ghoula et al. (2008). Dans cet article, nous présentons notre système SIAM (Système d'Indexation des Articles Médicaux), pour indexer des articles médicaux à l'aide du thésaurus Mesh. Cet outil a été testé sur un corpus construit à partir de documents web indexés par le catalogue CisMef.

2 Architecture de notre système d'indexation

Dans Majdoubi et al. (2009), nous avons présenté une approche automatique pour l'indexation des articles médicaux en utilisant le thésaurus Mesh. Cette approche est articulée autour de quatre étapes : nous commençons en premier lieu par le prétraitement qui consiste à à effectuer une série d'analyses linguistiques sur le texte afin de le préparer pour les prochains traitements. Dans une deuxième étape, les concepts Mesh sont extraits à partir de l'article à annoter. Au niveau de la troisième étape ces concepts extraits seront pondérés en utilisant une méthode qui combine le contenu, la structure et la sémantique. Une fois les concepts Mesh sont pondérés, une phase d'optimisation des concepts fréquents est nécessaire afin de générer l'annotation résultat.

[1]http ://medline.cos.com/
[2]CisMef : Catalogue et Index des Sites Médicaux francophone (http ://www.chu-rouen.fr/cismef/)

3 Expérimentation et résultat

Le corpus sur lequel nous avons évalué notre méthode d'indexation est un sous-ensemble d'articles de CISMeF comptant 500 documents annotés manuellement et utilisés pour l'apprentissage et l'évaluation des outils d'extraction d'entités biomédicales. Pour ce faire, nous avons utilisé deux mesures fréquemment utilisées dans l'évaluation des systèmes d'extraction d'informations, à savoir la précision et le rappel.

$$Précision = \frac{Nombre\ de\ concepts\ corrects}{Nombre\ total\ des\ concepts\ extraits}$$

$$Rappel = \frac{Nombre\ de\ concepts\ corrects}{Nombre\ total\ des\ concepts\ corrects\ qui\ auraient\ du\ être\ extraits}$$

Pour chacune de ces mesures, nous avons obtenu les résultats suivants :

Annotés manuellement	Extraits automatiquement	Corrects	Rappel	Précision
236	192	140	0.59	0.72

TAB. 1 – *Précision et rappel du SIAM.*

Références

Ghoula, N., K. Khelif, et R. Kuntz (2008). Vers une fouille sémantique des brevets : application au domaine biomédical. *Extraction et Gestion des Connaissances (EGC'08)*.

Khelif, K. (2006). *Web sémantique et mémoire d'expériences pour l'analyse du transcriptome*. Thèse de doctorat, Université Nice-Sophia Antipolis.

Majdoubi, J., M.Tmar, et F. Gargouri (2009). Using the mesh thesaurus to index a medical article: Combination of content, structure and semantics. *13th International Conference on Knowledge-Based and Intelligent Information Engineering Systems(KES'09), Santiago, Chile*.

Summary

This paper proposes an automatic indexing system which, starting from a medical article and the Mesh thesaurus, generates a describtion of the content's article.
To do this, our system uses firstly the NLP (Natural Language Processing) techniques to extract the indexing terms. Second, it extracts the Mesh concepts from this set of indexing terms. Then, these concepts are weighed based on their frequencies, locations in the article and their semantic relationships according to MeSH. Finally, the structured result annotation is built.