UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTE DES SCIENCES DHAR EL MAHRAZ
FES

AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mr **: SARROUTI Mourad**

Soutiendra **: le vendredi 06/07/2018** à **16H** Lieu : **salle réunion de géologie**

Une thèse intitulée:

Contributions to the Improvement of Question Answering Systems in the Biomedical Domain

En vue d'obtenir le Doctorat

FD : Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité : Informatique

	NOM ET PRENOM	GRADE	ETABLISSEMENT
Président	Pr. MEKNASSI Mohammed	PES	Faculté des Sciences Dhar El Mahraz - Fès
Directeur de thèse	Pr. OUATIK EL ALAOUI Said	PES	Faculté des Sciences Dhar El Mahraz - Fès
Rapporteurs	Pr. OUAHBI Brahim	PES	Ecole Nationale Supérieure d'Arts et Métiers - Meknès
	Pr. BEHJA Hicham	PH	Ecole Nationale Supérieure d'Electricité et de Mecanique - Casablanca
	Pr. ZINEDINE Ahmed	PH	Faculté des SciencesDhar El Mahraz - Fès
Membres	Pr. SATORI Khalid	PES	Faculté des SciencesDhar El Mahraz - Fès
	Pr. EN-NAHNAHI Noureddine	PH	Faculté des SciencesDhar El Mahraz - Fès
	Pr. AOURAGH SI Lhoussain	PH	Faculté des SciencesJuridiques, Economiques et Sociales - Salé

Contributions à l'amélioration des systèmes de questions-réponses en domaine biomédical

Résumé:

Ce travail de thèse s'inscrit dans le cadre des systèmes de questions-réponses (SQR)dans le domaine biomédical où plusieurs défis spécifiques à ce domaine sont relevés tels que lelexique et la terminologie, le type des questions posées, et la particularité des documents traités. Nous nous intéressons particulièrement à l'étudeet l'amélioration des méthodes permettant de déterminer les réponses précises à desquestions biomédicales exprimées en langage naturel (en anglais). Les SQR visent à fournir, à partir d'une collection de documents, des réponses succinctes et précises à desquestions en langage naturel. Généralement, ils sont constitués de quatre composantes principales : (1) l'analyse et la classification des questions, (2) la sélection des documents pertinents, (3) la recherche des passages pertinents, et (4) l'extraction des réponses. Dans ce travail dethèse, nous apportons quatre contributions. Dans la première, nous avons proposéune méthode de classification permettant de déterminer le type dela question. Etant basée à la fois sur les patrons lexico-syntaxiques et l'apprentissage automatique, celle-ci est exploitée par le SQR dans la phase d'extraction dela réponse appropriée à une question formulée en langage naturel. Dans le butde déterminer le type sémantique de la réponse attendue (i.e., un ou plusieurs sujets), nous avons proposé une variante de cette méthode qui s'appuie sur d'autrescaractéristiques de type lexical, morpho-syntaxique et sémantique. Le type sémantique d'une question biomédicale peut être pharmacologie, analyse, traitement,etc. Cette information permet de réduire le nombre de documents parcourus lors de la recherche des réponses. La deuxième contribution consiste à suggérer une méthode de recherche des documents pertinents à la question à partir de la base de données MEDLINE.Nous avons également proposé une alternative permettant la recherche des passages(i.e., extraits des documents) pertinents susceptibles de contenirles réponses candidates aux questions biomédicales. Latroisième contribution propose des méthodes d'extraction des réponses appropriéespermettant de générer à la fois les réponses *exactes* et *idéales*. En fin, dans la quatrième contribution, l'ensemble des méthodes proposéessont développées et intégrées au sein d'un système global de questions-réponses,appelé SemBioNLQA. Celui-ci accepte en entrée une variété de questions et retourne desréponses *exactes* et *idéales*. L'ensemble des contributions sont évaluées en utilisantla collection standard de questions fournies par la compagne d'évaluation BioASQ. Les résultats obtenus montrent l'intérêt de notre propos. Deplus, un sous-système de SemBioNLQA a été présentéau challenge BioASQ 2017 et classé parmi les premiers vainqueurs.

Mots clés: Systèmes de questions-réponses, Fouille de texte biomédical, Recherche d'information, Traitement automatique de la langue, Apprentissage automatique, Approche Sémantique.

Abstract:

This thesis work falls within the framework of question answering (QA) in the biomedical domain where several specific challenges are addressed, such as specialized lexicons and terminologies, the types of treated questions, and the characteristics of targeted documents. We are particularly interested in studying and improving methods that aim at finding accurate and short answers to biomedical natural language questions from a large scale of biomedical textual documents in English. QA aims at providing inquirers with direct, short and precise answers to their natural language questions. A typical QA system can be viewed as a pipeline composed of four main components including (1) question analysis and classification, (2) document retrieval, (3) passage retrieval, and (4) answer extraction. We consider that the improvement of such fundamental dimensions of the usefulness of QA has to take into account and solve the problems lying in each of these components. In this Ph.D. thesis, we propose four contributions to improve the performance of QA in the biomedical domain. In our first contribution, we propose a machine learning-based method for question type classification to determine the types of given questions which enable to a biomedical QA system to use the appropriate answer extraction method. We also propose an another machine learning-based method to assign one or more topics (e.g., pharmacological, test, treatment, etc.) to given questions in order to determine the semantic type of the expected answer which is very useful in generating specific answer retrieval strategies. In the second contribution, we first propose a document retrieval method to retrieve a set of relevant documents that are likely to contain the answers to biomedical questions from the MEDLINE database. We then present a passage retrieval method to retrieve a set of relevant passages to questions. In the third contribution, we propose specific answer extraction methods to generate both exact and ideal answers. Finally, in the fourth contribution, we develop a fully automated semantic biomedical QA system called SemBioNLQA which is able to deal with a variety of natural language questions and to generate appropriate answers by providing both exact and ideal answers. SemBioNLQA is derived from our established methods. Our proposals are evaluated on a common experimental design that considers large standard and well-known datasets for biomedical questions and answers provided by the BioASQ challenge. The experimental results support the validity of our contributions. In addition, a subsystem of SemBioNLQA was presented at the 2017 BioASQ challenge and was one of the challenge winners.

Keywords: Question answering, Biomedical text mining, Information retrieval, Natural language processing, Machine learning, Semantic approach.