AVIS DE SOUTENANCE DE THESE

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

Mr **: EL ADLOUNI Yassine**

Soutiendra **: le 28/12/2020** à **15h**

Lieu **: Centre de Visio Conférence**

Une thèse intitulée

Contributions to the improvement of arabic information retrieval :Application to arabic biomedical community question-answering

En vue d'obtenir le Doctorat

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

Spécialité: Informatique

Devant le jury composé comme suit :

	Nom et prénom	Grade	Etablissement
Président	Pr. Mohamed LAMRINI	PES	Faculté des Sciences Dhar El Mahraz - Fès
Directeur de thèse	Pr. Noureddine EN NAHNAHI	PH	Faculté des Sciences Dhar El Mahraz - Fès
Co-directeur	Pr. Said OUATIK EL ALAOUI	PES	ENSA-Université Ibn Tofail de Kénitra
Rapporteurs	Pr. Ouafae AMMOR	PES	Faculté des Sciences et Techniques -Fès
	Pr. Brahim OUHBI	PES	Ecole Nationale Supérieure d'Arts et Métiers - Meknès
	Pr. Ayoub AIT LAHCEN	PH	ENSA-Université Ibn Tofail de Kénitra
Membres	Pr. Hamid TAIRI	PES	Faculté des Sciences Dhar El Mahraz - Fès
	Pr. Jaouad BOUMHIDI	PES	Faculté des Sciences Dhar El Mahraz - Fès
Invité	Pr. Mohammed MEKNASSI	PES	Faculté des Sciences Dhar El Mahraz - Fès

Résumé :

Les systèmes de question réponse communautaires ont suscité beaucoup d'intérêt durant les dernières années vu leur capacité de se positionner comme alternative aux systèmes de recherche d'information basés sur les mots clés et les systèmes de question réponse.

Hormis leur but de suggérer une réponse à la question posée par l'utilisateur, ils permettent d'épargner le temps aux utilisateurs tout en optimisant la performance du système à travers l'élimination des doublons.

Cette thèse a pour objectif l'amélioration de la représentation, la restitution ainsi que la compréhension du texte en utilisant la modélisation des thèmes, les plongements distributionnels des mots et les bases de connaissances. Les systèmes de question réponses préalables ont été généralement basés sur l'analyse lexicale et l'analyse syntaxique du texte, ce qui engendrait de faibles performances dues à leur incapacité de détecter les reformulations entre autres. Cette thèse a pour but d'aller au-delà de ces représentations en exploitant la dimension sémantique du texte à travers les modèles d'apprentissage pour la classification et le classement, les réseaux de neurones et les bases de connaissances.

Cette thèse commence par explorer la performance des modèles classiques pour la recherche d'information, présenter une approche à base de la modélisation des sujets et démontrer par la suite comment l'approche par pairs est plus performante que celle par points quand il s'agit de l'apprentissage du classement.

Ensuite, la thèse présente l'utilité d'un ensemble de plongements de mots générés à partir d'une collection d'articles du domaine biomédical afin de générer des représentations vectorielles denses des mots et qui sont par la suite injectées dans une pile de réseaux de neurones convolutifs. On démontre également l'utilité du contexte pour ces représentations en les ajustant pour notre plateforme.

Cette thèse se termine par une présentation de l'approche à base des arbres à travers l'utilisation des fonctions noyaux d'arbres ou les réseaux de neurones récursifs avec comme entrée les arbres binaires de constituants.

Mots clés :

Systèmes Réponse aux questions de la communauté, Recherche d'information, Traitement du langage naturel arabe, Modélisation de sujets, Apprentissage en profondeur, Intégration contextualisée, Approche arborescente.

CONTRIBUTIONS TO THE IMPROVEMENT OF ARABIC INFORMATION RETRIEVAL : APPLICATION TO ARABIC BIOMEDICAL COMMUNITY QUESTION-ANSWERING

Abstract :

Community Question Answering has attracted much attention in recent years due to their ability to stand in between the less verbose search engines offering a keyword-based lookup and the more verbose question answering systems. Besides their obvious purpose of suggesting potential answers to a user's question, they enable users to save considerable time by avoiding duplicate questions.

Previous community question answering systems were mostly based on lexical and syntactic approaches leading to weak performances as they are unable to identify paraphrasing.

This dissertation goes beyond these approaches seeking to improve text comprehension and retrieval by leveraging the semantic dimension of text using learning to rank, neural and knowledge-based models. To this end we propose a bunch novel methods including topic modeling, distributional embeddings, and knowledge bases.

This thesis research starts by exploring the performance of classic approaches to information retrieval based on machine learning, presenting a topic modeling approach to improve text representation and depicting our learning to rank models by showing how pairwise models are more powerful than their counterpart pointwise approaches.

Then this dissertation presents our distributional embeddings trained on a biomedical corpus and used to power a novel neural architecture combining lexical features with a stack of convolutional networks. Afterward, we show how contextualized embeddings perform better than context-free embeddings by fine-tuning them on our task.

This thesis research concludes with a tree-based approach either using kernel trees combined with a machine learning algorithm or using binary constituency trees as inputs to a recursive neural network.

Key Words :

Community Question Answering, Information retrieval, Arabic Natural Language processing, Topic Modeling, Deep Learning, Conextualized Embeddings, Tree-Based Approach