AVIS DE SOUTENANCE D'UNE THESE DE DOCTORAT

Le Doyen de la Faculté des Sciences a le plaisir d'informer le public qu'une soutenance de thèse de Doctorat en

«**Mathématiques informatique et applications**»

aura lieu le 13/07/2024 à 10H30 à l'ENSA, Kénitra

La Thèse sera présentée par Mme LAHBARI IMANE

Sous le thème :

Contributions aux Systèmes Questions-Réponses en langue Arabe par l'apprentissage profond et la sémantique implicite

Devant le jury composé de :

Nom et Prénom	Titre	Etablissement
LAMRINI MOHAMED	Président	Faculté des Sciences, Dhar El Mahraz, Fès
BEHJA HICHAM	Rapporteur	Ecole Nationale Supérieure d'Electricité et de Mécanique, Casablanca
AOURAGH SI LHOUSSAIN	Rapporteur	Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Rabat
OUMSIS MOHAMMED	Rapporteur	Ecole Supérieure de Technologie, Salé
ALAOUI ZIDANI KHALID	Examinateur	Faculté des Sciences, Dhar El Mahraz, Fès
CHOUGDALI KHALID	Examinateur	ENSA, Kénitra
EL HAMI NORELISLAM	Examinateur	ENSA, Kénitra
OUATIK EL ALAOUI SAID	Directeur de thèse	ENSA, Kénitra

UNIVERSITE IBN TOFAIL
PÔLE DES ETUDES DOCTORALES
KENITRA

ⵜⴰⴳⵍⴷⵉⵜ ⵏ ⵍⵎⵖⵔⵉⴱ
ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵓⵙⵙⵍⵎⴷ ⴰⵏⴰⴼⵍⵍⴰ
ⴷ ⵓⵔⵣⵣⵓ ⴰⵎⴰⵙⵙⴰⵏ ⴷ ⵓⵙⵏⵓⴼⵍ

المملكة المغربية
وزارة التعليم العالي
و البحث العلمي و الإبتكار

جامعة ابن طفيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵉⴱⵏ ⵟⵓⴼⴰⵢⵍ
Ibn Tofaïl University

Nom et Prénom : LAHBARI IMANE
Date de soutenance : 13/07/2024
Directeur de Thèse : OUATIK EL ALAOUI SAID

Sujet de thèse :

Contributions aux Systèmes Questions-Réponses en langue Arabe par l'apprentissage profond et la sémantique implicite

Résumé:

L'objectif des systèmes de questions-réponse (SQR) est de répondre précisément aux questions exprimées en langage naturel. De manière générale, les SQR se composent de trois modules principaux : le traitement des questions, l'extraction de passages et l'extraction de réponses. Dans ce travail de thèse, nous examinons chaque partie pour améliorer les techniques utilisées et obtenir des résultats supérieurs. L'état actuel de la technique suggère diverses approches pour la classification des questions ; après avoir mené une analyse comparative, nous avons constaté qu'une approche hybride combinant des techniques basées sur des règles et des techniques d'apprentissage automatique était le moyen le plus efficace de classer les questions arabes. Nous avons ensuite reformulé et élargi cette approche en incorporant le marquage POS et l'arabe WordNet.
Dans la deuxième partie, nous extrayons des documents à l'aide de l'API Google et d'une base de données Wikipédia. Nous appliquons ensuite diverses incorporations de mots, telles que Word2vec, GloVe et FastText, pour extraire les passages pertinents. En utilisant des représentations d'incorporation de phrases comme AraBERT, Elmo et FastText, nous améliorons le module d'extraction. En classant les passages et en extrayant les plus pertinents, la réponse finale est extraite en affinant les paramètres AraBERT.
Parallèlement, nous avons participé à deux éditions de SemEval, une série internationale d'ateliers de recherche sur le traitement automatique du langage naturel (TALN) dont le but est de repousser les limites de l'analyse sémantique. Notre stratégie pour traiter les questions-réponses communautaires lors de ce concours est basée sur l'utilisation à la fois des textes arabes originaux et de leurs traductions en anglais, sur lesquelles des techniques d'apprentissage automatique supervisé sont appliquées.

Abstract:

The goal of Question Answering Systems (QASs) is to precisely respond to inquiries expressed in natural language. Generally speaking, QASs consist of three primary parts: question processing, passage extraction, and answer extraction. In this thesis, we examine each part to enhance the used techniques and get superior outcomes. The current state of the art suggests various approaches for question classification; after conducting a comparative analysis, we found that a hybrid approach that combined rule-based and machine-learning techniques was the most effective way to classify Arabic questions. We then reformulated and expanded this approach by incorporating POS tagging and Arabic WordNet.

In the second part, we extract documents using the Google API and a database of Wikipedia. We then apply various word embeddings, such as Word2vec, GloVe, and FastText, to extract pertinent passages. By utilizing sentence embedding representations like AraBERT, Elmo, and FastText, we improve the extraction module. By ranking passages and extracting the most pertinent ones, the final response is extracted by fine-tuning the AraBERT parameters.

Simultaneously, we took part in two editions of SemEval, an international series of workshops for natural language processing (NLP) research whose goal is to push the boundaries of semantic analysis. Our strategy for dealing with the Community Question Answering at this competition is based on utilizing both the original Arabic texts and their English translations, over which supervised machine learning techniques are applied.