Nom et Prénom : ALAMI HAMZA
Date de soutenance : 10/10/2021
Directeur de Thèse : OUATIK EL ALAOUI SAID

Sujet de Thèse :

Apprentissage profond pour l'amélioration des systèmes question réponses en langue Arabe

Résumé :

Construire des Systèmes Questions-Réponses (SQR), permettant de répondre de manière précise aux questions exprimées en langage naturel, est l'une des tâches les plus difficiles en domaines de l'intelligence artificielle et du Traitement Automatique du Langage Naturel (TLN). Cette thèse aborde les systèmes questions-réponses à domaine ouvert en langue Arabe en proposant des méthodes permettant de retourner des réponses exactes aux questions de l'utilisateur. La langue Arabe présente plusieurs défis pour les applications du TLN, et notamment, pour les SQR. En effet, plusieurs difficultés liées à la morphologie complexe de cette langue, telles que sa nature dérivative et flexionnelle, la présence de signes diacritiques, l'absence de majuscules et le manque de ressources arabes, rendent l'analyse et la fouille de textes une tâche extrêmement ardue. Dans cette thèse, nous nous sommes basés sur des modèles de réseaux de neurones profonds, puisqu'ils se sont avérés plus efficaces pour l'apprentissage des propriétés linguistiques riches et l'amélioration des performances de diverses tâches du Traitement Automatique du Langage Naturel. Pour répondre au problème de la classification des questions, nous proposons deux principales contributions : (i) Une nouvelle taxonomie inspirée des études linguistiques Arabes pour construire un classificateur de questions en s'appuyant sur les algorithmes d'apprentissages automatique et profond. Nous adoptons une représentation continue et distribuée des mots afin de capturer les relations sémantiques et syntaxiques entre les mots. (ii) Nous intégrons la représentation contextuelle des mots pour construire des classificateurs de questions Arabes en se basant à la fois sur des méthodes d'apprentissage profond et des techniques ensemblistes. L'objectif est de pallier la limitation de la représentation statique des mots qui ne tient pas compte du contexte des mots dans le texte. Dans la troisième contribution relative à la détection des questions dupliquées, nous proposons une nouvelle méthode qui comprend un module de classification des questions Arabes capable de filtrer les différentes questions en fonction de leurs classes ou catégories. De plus, un détecteur neuronal de questions en double est construit à l'aide d'un mécanisme d'attention neuronale et d'une représentation contextuelle. La détection des questions en double permet de réduire le temps de réponse et le coût global de calcul de la réponse aux questions en domaine ouvert. Dans la quatrième et dernière contribution, nous suggérons un système Questions-Réponses Arabe en domaine ouvert nommé DAQAS. Le système construit intègre trois composants principaux : (1) Un module de détection des questions en double permettant d'identifier et de renvoyer la réponse aux questions auxquelles la réponse est déjà connue. (2) Un module de récupération des passages pertinents qui consiste à appliquer l'expansion des requêtes et en utilisant un processus de recherche d'information (3) Un module de lecture qui extrait la réponse exacte et précise à partir des questions et des passages pertinents récupérés à l'étape précédente. Diverses expérimentations ont été menées pour montrer l'efficacité de l'ensemble des méthodes proposées. Les résultats obtenus montrent l'intérêt de nos propos notamment pour les futurs systèmes de réponse aux questions à domaine ouvert en langue Arabe.

Absract :

Building computer systems that have the ability to answer natural language questions is one of the most challenging tasks in Artificial Intelligence (AI) and Natural Language Processing (NLP). This thesis tackles the problem of Arabic Open-Domain Question Answering Systems by designing and building systems that search for exacting answers to human questions. We focus on the Arabic language, where several specific challenges are addressed. For instance, the difficulties related to the complex morphology of this language, such as its derivational and inflectional nature, the presence of diacritical marks, the absence of capital letters, and the lack of Arabic resources. In addition, we aim at adopting deep neural network models as they have proven to be superior, by a large margin, to traditional sparse and hand-designed feature-based approaches, in learning rich linguistic phenomena and improving performance of various Natural Language Processing (NLP) tasks. To tackle the problem of questions classification, we propose two main contributions: (i) a new

taxonomy inspired from Arabic linguistic studies to build shallow learning and deep learning question classifier. We adopt continuous and distributed word representation which catch semantic and syntactic relations between words. (ii) We overcome the limitation of static word representation, which does not consider the context, by integrating contextual word representation to build Arabic question classifiers. The classifiers are based on deep learning methods and ensemble techniques. In the third contribution related to detecting Arabic duplicate questions, we propose a new method that comprises a question classification module able to filter different questions according to question class labels. Also, a neural duplicate question detector is built using a neural attention mechanism and contextual representation. Detecting duplicate questions reduces the response time and diminishes the computation cost of the overall Open-Domain Question Answering. In the four and last contribution, we design and build an Arabic Open-Domain Question Answering system, namely DAQAS. The system integrates three main components: Duplicate Question Detector module able to identify and return the answer to previously answered questions; Retriever module which aims at retrieving relevant passages by expanding queries with neural language generation and using an Information Retrieval system; Reader module that takes questions and its relevant passages retrieved to extract the exact and precise answer. Various experiments have been conducted to demonstrate the effectiveness of these contributions. We believe that they hold great promise for future Arabic Open-Domain Question Answering systems.