جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THESE DE DOCTORAT

Madame Rajae MOUMEN

soutiendra publiquement sa thèse de Doctorat en Informatique

Le Samedi 13 Novembre 2021 à 11h au Grand amphi à l'ENSIAS

Intitulé de la thèse

Traitement automatique de la langue arabe à l'aide des réseaux de neurones profonds

UNIVERSITE MOHAMMED V - RABAT ENSIAS

Devant le Jury composé de :

Président :

Pr. Abdellatif EL AFIA, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Raddouane CHIHEB, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Mohamed BAHAJ, PES, FST, Université Hassan 1er de Settat

Pr. Mustapha MACHKOUR, PH, Faculté des Sciences, Université Ibn Zohr d'Agadir

Pr. Mohamed LAZAAR, PH, ENSIAS, Université Mohammed V de Rabat

Examinateur :

Pr. Fatima OUZAYD, PH, ENSIAS, Université Mohammed V de Rabat

Invité :

Taoufiq ZARRA, Docteur-Ingénieur, Responsable Data et Innovation, Société Nationale de Radiodiffusion et de Télévision (SNRT)

Traitement automatique de la langue arabe à l'aide des réseaux de neurones profonds

Résumé : Durant la dernière décennie, l'avancée technologique en termes de capacité de stockage et de calcul, ainsi que la puissance croissante des algorithmes d'apprentissage automatique a permis d'orienter l'intérêt vers des données jusque-là peu exploitées. Des mégadonnées sont traitées automatiquement presque en temps réel, et de l'information précieuse en est extraite. Dans cette thèse, l'intérêt est porté sur la langue arabe et ses deux registres : l'arabe moderne et le dialectal marocain. La langue arabe est une langue avec peu de ressources qui ne bénéficie que de peu d'engouement en comparaison avec les langues occidentales en particulier l'anglais. On explore, en premier lieu, le domaine de la détection et reconnaissance de texte sous-domaine de la vision par ordinateur (Computer vision). Un système de détection automatique de texte de scène en langue arabe a été mis en place, basé sur le réseau convolutif VGG-16 ; Le Framework proposé fractionne la tâche en deux étapes : un premier réseau qui localise les instances textuelles et affecte de manière régressive une estimation d'échelle, et un deuxième réseau qui détecte le texte de manière précise au niveau du pixel. Devant le manque de bases de données annotées pour la détection de texte en langue arabe, et afin d'entrainer et tester ce Framework, un travail de construction de base de données pour la détection et la reconnaissance de texte arabe a été entrepris, associé à des techniques d'augmentation de data pour obtenir une base de données synthétiques de 1M d'images.

En deuxième lieu, nous investiguons le domaine de la diacritisation de la langue arabe sous-domaine du traitement automatique des langues naturelles (Natural language processing) ; dans ce cadre, deux contributions ont été présentées en ce qui concerne le traitement automatique du dialectal marocain, qui consistent en la création d'un corpus annoté en diacritisation, ainsi qu'un modèle de diacritisation basé sur le réseau de neurones récurent LSTM. Quant à l'arabe moderne, notre contribution consiste à explorer les performances du réseau de neurones GRU (Gated Reccurent Unit) et évaluer ses performances dans la diacritisation.

Mots-clés : GRU, LSTM, Traitement automatique des langues (TALN), Dialectal marocain, Arabe moderne, diacritisation, détection de texte, reconnaissance de texte, réseaux de neurones convolutifs.