Avis de Soutenance

Monsieur IBRAHIM SOULEIMAN MAHAMOUD

Spécialité : Informatique et Applications

Soutiendra publiquement ses travaux de thèse intitulés

« Apprentissage automatique, et sans modèle a priori, des liens sémantico-structurels entre les champs dans un document. »

dirigés par Monsieur Jean Marc Ogier et Monsieur Mickael COUSTATY

le **jeudi 07 novembre 2024**
à

14h00

Lieu : La Rochelle Université
Pôle Communication, Multimédia et Réseaux
Amphithéâtre Michel Crépeau
44 Av. Albert Einstein 17000 LA ROCHELLE

Composition du jury proposé

M. Mickaël COUSTATY	La Rochelle Université -
Mme Véronique EGLIN	INSA de Lyon -
M. Camille KURTZ	Université Paris Cité -
Mme Aurélie LEMAITRE	Université Rennes 2 -
M. Jean-Marc OGIER	La Rochelle Université -

Résumé :
À l'ère où l'utilisation quotidienne de documents numériques est devenue de plus en plus répandue, le traitement automatique des documents est devenu essentiel. Parmi les processus automatiques applicables à un document, on trouve la classification et l'extraction d'informations (soit à travers des classes prédéfinies, soit en les extrayant via des questions). Avant de concevoir des processus de traitement automatique, il est crucial de comprendre les contraintes auxquelles ces processus sont confrontés. Parmi ces contraintes, on trouve la grande diversité des documents, qui entraîne des déséquilibres et inclut des documents multilingues. De plus, ces solutions doivent être rapides et nécessiter peu de paramètres. Premièrement, nous avons présenté le corpus CHIC, un corpus d'extraction d'informations à travers des questions sur des documents multilingues respectant les contraintes industrielles. Ensuite, nous avons proposé un ensemble de méthodes basées sur des mécanismes d'attention pour relever plusieurs défis. Nous avons d'abord présenté un mécanisme d'auto-attention visant à réduire l'inter-similarité entre les documents. Ensuite, nous avons proposé des mécanismes d'attention pour comprendre les choix du modèle conduisant à des prédictions. Enfin, nous avons présenté des méthodes pré-entraînées à partir de corpus provenant de certaines tâches pertinentes dans notre contexte. Nous avons proposé plusieurs stratégies pour pré-entraîner ces méthodes de manière non-supervisée. Toutes ces contributions ont abouti à des résultats encourageants qui pourraient être utiles dans un contexte industriel.