DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **1er octobre 2021**

Nom de famille et prénom de l'auteur : **Monsieur SAGE Clément**

Titre de la thèse : *« Apprentissage profond pour l'extraction de l'information des documents commerciaux »*

Résumé

En raison de la quantité massive et croissante de documents reçus chaque jour et du nombre d'étapes pour les traiter, les plus grandes entreprises se sont tournées vers des logiciels d'automatisation des processus documentaires afin d'atteindre de faibles coûts de traitement. Une étape cruciale d'un tel logiciel est l'extraction de l'information des documents, en particulier la récupération des champs qui apparaissent régulièrement dans les documents entrants. Pour faire face à la variabilité de la structure de l'information contenue dans ces documents, les systèmes industriels et académiques sont progressivement passés de méthodes basées sur des règles à des modèles d'apprentissage profond pour effectuer la tâche d'extraction. L'objectif de cette thèse est d'apporter des méthodes pour apprendre à extraire l'information des documents commerciaux.

Dans la première partie de ce manuscrit, nous adoptons l'approche d'étiquetage de séquence en entraînant des réseaux de neurones profonds à classer le type d'information porté par chaque *token* des documents. Lorsque les étiquettes des *tokens* utilisées pour l'apprentissage sont parfaites, nous montrons que ces classificateurs de *tokens* peuvent extraire des champs tabulaires complexes de documents dont l'émetteur et la mise en page étaient inconnues au moment de l'apprentissage du modèle. Cependant, lorsque la supervision au niveau du *token* doit être déduite de la vérité terrain de haut niveau naturellement produite par la tâche d'extraction, nous démontrons que les classificateurs de *tokens* extraient l'information de documents du monde réel avec une précision nettement inférieure en raison du bruit introduit dans les étiquettes.

Dans la deuxième partie de cette thèse, nous explorons des méthodes qui apprennent à extraire de l'information directement à partir de la vérité terrain de haut niveau à notre disposition, évitant ainsi une supervision au niveau des *tokens* coûteuse. Nous adaptons un modèle séquence à séquence basé sur un mécanisme d'attention afin de copier les *tokens* du document portant de l'information pertinente et de générer les balises XML structurant le schéma d'extraction en sortie. Contrairement aux travaux antérieurs en extraction d'information de bout en bout, notre approche permet de retrouver n'importe quel schéma d'information, quelle que soit sa structure. En comparant ses performances d'extraction avec les

classificateurs de *tokens* précédemment étudiés, nous montrons que les méthodes de bout en bout sont compétitives avec les approches d'étiquetage de séquence et peuvent largement les surpasser lorsque les étiquettes des *tokens* ne sont pas immédiatement accessibles.

Enfin, dans une troisième partie, nous confirmons qu'utiliser des modèles pré-entraînés pour extraire de l'information réduit considérablement les besoins en documents annotés. Nous exploitons un modèle de langage existant basé sur l'architecture Transformer qui a été pré-entraîné sur une large collection de documents commerciaux. Lorsqu'il est adapté à une tâche d'extraction d'information via l'approche d'étiquetage de séquence, le modèle de langage nécessite très peu de documents d'entraînement pour atteindre des performances d'extraction proches du maximum. Cela souligne que les modèles pré-entraînés sont significativement plus efficients en matière de données que les modèles apprenant la tâche d'extraction à partir de zéro. Nous révélons également de précieuses capacités de transfert de connaissances pour ce modèle de langage puisque les performances sont améliorées en apprenant au préalable à extraire de l'information sur un autre jeu de données, même si ses champs ciblés diffèrent de la tâche initiale.