Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Nesrine BANNOUR : bannour.nesrine@gmail.com

Titre : Extraction d'informations à partir des dossiers patients informatisés : études en temporalité, confidentialité et impact environnemental

Mots-clés : extraction d'informations, représentation temporelle, traitement automatique des langues cliniques, confidentialité, réseaux de neurones, empreinte carbone.

Title: *Information Extraction from Electronic Health Records: Studies on Temporal Ordering, Privacy and Environmental Impact*

Keywords: *information extraction, temporal representation, clinical natural language processing, confidentiality, neural networks, carbon footprint.*

Thèse de doctorat en informatique, laboratoire interdisciplinaire des sciences du numérique, LISN, UMR 9015, école doctorale sciences et technologies de l'information et de la communication, STIC, Université Paris-Saclay, sous la direction de Mme Aurélie Névéol (DR, CNRS, laboratoire interdisciplinaire des sciences du numérique, LISN), M. Xavier Tannier (Pr, Sorbonne Université, laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS) et M. Bastien Rance (MC, praticien hospitalier, Université Paris-Cité, hôpital européen Georges Pompidou, AP-HP, centre de recherche des Cordeliers). Thèse soutenue le 30/11/2023.

Jury : Mme Aurélie Névéol (DR, CNRS, laboratoire interdisciplinaire des sciences du numérique, LISN, codirectrice), M. Xavier Tannier (Pr, Sorbonne Université, laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, codirecteur), M. Bastien Rance (MC, praticien hospitalier, Université Paris-Cité, hôpital européen Georges Pompidou, AP-HP, centre de recherche des Cordeliers, codirecteur), Mme Fatiha Saïs (Pr, Université Paris-Saclay, présidente), M. Maxime Amblard (Pr, Université de Lorraine, rapporteur), M. Timothy Miller (*associate*

professor, Harvard University, Boston Children's Hospital, Boston, États-Unis, rapporteur), Mme Fleur Mougin (Pr, Université de Bordeaux, examinatrice).

Résumé : *L'extraction automatique des informations contenues dans les dossiers patients informatisés (DPI) est cruciale pour améliorer la recherche clinique. Or, la plupart des informations sont sous forme de texte non structuré. La complexité et le caractère confidentiel du texte clinique présentent des défis supplémentaires. Par conséquent, le partage de données est difficile dans la pratique et est strictement encadré par des réglementations. Les modèles neuronaux offrent de bons résultats pour l'extraction d'informations. Mais ils nécessitent de grandes quantités de données annotées, qui sont souvent limitées, en particulier pour les langues autres que l'anglais. Ainsi, la performance n'est pas encore adaptée à des applications pratiques. Outre les enjeux de confidentialité, les modèles d'apprentissage profond ont un important impact environnemental. Dans cette thèse, nous proposons des méthodes et des ressources pour la reconnaissance d'entités nommées (REN) et l'extraction de relations temporelles dans des textes cliniques en français. Plus précisément, nous proposons une architecture de modèles préservant la confidentialité des données par mimétisme permettant un transfert de connaissances d'un modèle enseignant entraîné sur un corpus privé à un modèle élève. Ce modèle élève pourrait être partagé sans révéler les données sensibles ou le modèle privé construit avec ces données. Notre stratégie offre un bon compromis entre la performance et la préservation de la confidentialité. Ensuite, nous introduisons une nouvelle représentation des relations temporelles, indépendante des événements et de la tâche d'extraction, qui permet d'identifier des portions de textes homogènes du point de vue temporel et de caractériser la relation entre chaque portion du texte et la date de création du document. Cela rend l'annotation et l'extraction des relations temporelles plus faciles et reproductibles à travers différents types d'événements, vu qu'aucune définition ni extraction préalables des événements ne sont requises. Enfin, nous effectuons une analyse comparative des outils existants de mesure d'empreinte carbone des modèles de TAL. Nous adoptons un des outils étudiés pour calculer l'empreinte carbone de nos modèles, en considérant que c'est une première étape vers une prise de conscience et vers un contrôle de leur impact environnemental. En résumé, nous générons des modèles de REN partageables préservant la confidentialité que les cliniciens peuvent utiliser efficacement.*

Nous démontrons également que l'extraction de relations temporelles peut être abordée indépendamment du domaine d'application et que de bons résultats peuvent être obtenus en utilisant des données d'oncologie du monde réel.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04347666

Gaël LEJEUNE : gael.lejeune@sorbonne-universite.fr

Titre : De la variation linguistique et de son influence sur l'application de méthodes de Traitement Automatique des Langues

Mots-clés : tokenisation, n-grammes de caractères, sous-mots, genre textuel, collecte de corpus, nettoyage de pages Web, reconnaissance optique de caractères, reconnaissance d'entités nommées, données bruitées, variation linguistique.

Title: *On Linguistic Variation and Its Impact on the Application of Natural Language Processing Methods*

Keywords: *tokenization, character n-grams, subwords, text genre, corpus collection, Web scraping, optical character recognition, named entity recognition, noisy data, linguistic variation.*

Habilitation à diriger des recherches en informatique, sociologie et informatique pour les sciences humaines, STIH, Sorbonne Université, sous la direction de Mme Virginie Julliard (Pr, Sorbonne Université). Habilitation soutenue le 18/12/2023.

Jury : Mme Virginie Julliard (Pr, Sorbonne Université, directrice), M. Franck Neveu (Pr, Sorbonne Université, président), Mme Cécile Fabre (Pr, Université de Toulouse, rapporteuse), M. Éric Gaussier (Pr, Université Grenoble Alpes, rapporteur), M. Laurent Romary (DR, Inria Paris, rapporteur), M. François Rioult (MC, HDR, Université de Caen, examinateur).

Résumé : *Cette habilitation à diriger les recherches traite de la variation des données textuelles et de son influence sur l'application de méthodes de traitement automatique des langues (TAL). Différents types de variation sont examinés : variation de la langue, variation de la qualité des données, variation de l'homogénéité des corpus et variation du genre textuel.*

Nous posons, d'une part, la question des observables du TAL. Il s'agit d'interroger la pertinence du paradigme, majoritaire dans le domaine, consistant à envisager les documents avant tout à travers des représentations en mots, très sensibles aux variations de toutes sortes, au détriment par exemple d'approches en chaînes de caractères plus robustes.

D'autre part, nous interrogeons les observatoires du TAL en proposant des pistes pour exploiter les genres textuels des documents et pour tirer des corpus desquels ils sont tirés des propriétés utiles au traitement automatique, à rebours d'une approche où

les documents sont simplement des séquences de mots ou de sous-mots. Nous montrons notamment comment la structure des documents et le genre textuel peuvent être exploités pour concevoir des modèles de TAL.

URL où le mémoire peut être téléchargé :
https://hal.science/tel-04360967