Enjeux et opportunités de la fouille textes pour stimuler la recherche pluridisciplinaire

Mathieu Roche*

* CIRAD, Montpellier

Résumé

Les travaux de sciences des données textuelles ont connu un formidable essor en ouvrant des perspectives nombreuses depuis l'avènement des modèles de langues et des grands modèles de langues (LLM - Large Language Model). Dans ce contexte, les travaux pluridisciplinaires intégrant des ressources textuelles hétérogènes offrent de nouvelles perspectives.
Dans un premier temps, cette présentation dresse un panorama d'approches de fouille de textes intégrées dans différents projets appliqués à l'agriculture et à la santé dans une perspective One Health (une seule santé). Les recherches pluridisciplinaires peuvent se nourrir mutuellement et conduire à la co-construction de démarches génériques.
Dans un deuxième temps, nous montrerons de quelles manières (i) les travaux disciplinaires peuvent alimenter les travaux pluridisciplinaires et (ii) comment les problématiques thématiques peuvent engendrer de nouveaux verrous scientifiques pour la fouille de textes et les recherches du monde académique.
Enfin, cette présentation discutera la manière dont les nouvelles problématiques méthodologiques et disciplinaires liées aux LLM et à leur usage ouvrent de nouveaux défis pluridisciplinaires en particulier dans les pays du Sud : biais issus des modèles, traitement de langues peu dotées, intégration de connaissances (syntaxiques et sémantiques) dans les modèles, explicabilité, frugalité, etc.

Biographie

Mathieu Roche est chercheur au CIRAD – UMR TETIS (Montpellier, France) depuis 2013. Entre 2005 et 2013, il a été Maître de Conférences à l'Université Montpellier 2, France. Mathieu Roche a obtenu un doctorat en informatique à l'Université Paris 11 (Orsay) en 2004. Il a soutenu son HDR (Habilitation à Diriger des Recherches) en 2011 et a dirigé plusieurs projets académiques et industriels en fouille de textes. Mathieu Roche a été responsable de différentes équipes de recherche de Traitement Automatique du Langage Naturel et de Science des Données (équipes TEXTE (UMR LIRMM), SISO et MISCA (UMR TETIS)). Il est actuellement impliqué dans plusieurs comités éditoriaux de revues internationales (ARIMA, Scientific Data, IDA) et contribue à 2 projets de recherche européens (H2020 MOOD 2020-24 et Horizon CEA-First 2023-27) portant sur la mise en œuvre de méthodes de fouille de textes pour la

surveillance épidémiologique et la sécurité alimentaire. Il a publié plus de 200 articles depuis 2013 (dont 3 best-papers) et a dirigé/encadré 21 doctorants.