Séminaire en ligne Franco-Africain en Sciences du Numérique - LIRIMA
Mercredi 16 novembre 2022, 16:00 (Paris time)

Paulin MELATAGIA YONTA, Université de Yaoundé I, Cameroun

L'apprentissage automatique au service de l'utilisation des langues peu dotées dans les technologies numériques

Résumé
Les algorithmes d'apprentissage automatique (ML) pour le traitement du langage naturel et de la parole apprennent sur d'énormes quantités de données qui sont nécessaires pour atteindre les performances de l'état de l'art et permettre l'exploitation industrielle de des outils dans lesquels ils sont mis en œuvre. C'est la principale limite de l'utilisation des algorithmes existants pour les langues faiblement dotées qui constituent un pourcentage important des langues utilisées dans le monde. Au Cameroun par exemple, il y a plus de 200 langues locales, toutes faiblement dotées. Il est important de s'intéresser aux aspects spécifiques de ces langues qui affectent le développement de systèmes fiables. Le premier handicap est le manque de corpus annotés, c'est pourquoi nous travaillons à développer des modèles d'apprentissage à partir de données d'entraînement limitées (texte et parole). Sachant que les caractéristiques structurelles, sociales et culturelles (tonalité, agglutination, dialectes par exemple) liées aux langues cibles apportent des défis supplémentaires, nous devons mettre en œuvre des techniques de représentation des données efficaces (en utilisant notamment les plongements multilingues) pour l'apprentissage, l'annotation automatique (en utilisant par exemple l'apprentissage auto-supervisé) et l'extraction d'informations sémantiques (NER, WSD, ...) pour les applications de transcription et de traduction automatique.

Biographie :
Paulin Melatagia est enseignant-chercheur, depuis 2008, au Département d'Informatique de la Faculté des Sciences de l'Université de Yaoundé I au Cameroun où il a obtenu son Doctorat/Ph.D. en 2012. Il est responsable, depuis 2017, de l'équipe de recherche Idasco (sciences de données et modélisation des systèmes complexes) du Laboratoire d'Informatique et Applications. Ses travaux de recherche portent depuis 2014 sur le traitement automatique du langage naturel et le traitement de la parole en particulier pour les langues peu dotées. Les questions principales de ses travaux sont l'apprentissage des représentations, l'automatisation de l'étiquetage et l'extraction d'information sémantique.

Paulin Melatagia est secrétaire scientifique du CRI (Conférence de Recherche en Informatique) depuis 2015, membre du conseil d'unité de UMMISCO (http://www.ummisco.fr/) depuis 2021 et lors du CARI'2022 il a été fait membre du premier comité exécutif de l'ASDS (The African Society in Digital Science).

Références :

- Paulin Melatagia Yonta, Michael Franklin Mbouopda, Named Entity Recognition in Low-resource Languages using Cross-lingual distributional word representation. ARIMA J. 33
- Mathieu Leonel Mba, Christian Gamom Ngounou Ewo, Julien Denoulet, Paulin Melatagia Yonta and Bertrand Granado, An efficient FPGA overlay for MPI-2 RMA parallel applications, 2022 20th IEEE Interregional NEWCAS Conference (NEWCAS), 2022, pp. 412-416
- Michael Franklin Mbouopda, Paulin Melatagia Yonta, Guy Stephane B. Fedim Lombo, Neurals Networks for Projecting Named Entities from English to Ewondo, International Conference on Learning Representations 2020
- Dimedrick Vanil Feudjieu, Paulin Melatagia Yonta, Résolution d'anaphores nominales avec les séparateurs à vastes marges sur les arbres syntaxiques, CARI 2020
- Michael Franklin Mbouopda, Paulin Melatagia Yonta. A Word Representation to Improve Named Entity Recognition in Low-resource Languages. SNAMS 2019: 333-337