Online Franco-African Seminar in Digital Sciences - LIRIMA
Wednesday November 16 2022, 4:00 pm

Paulin MELATAGIA YONTA, University of Yaoundé I, Cameroon

Machine learning to support the use of low resourced languages in digital technologies

Abstract
Machine learning (ML) algorithms for natural language and speech processing learn on huge quantities of data that are necessary to reach state-of-the-art performances and allow industrial exploitation of tools in which they are implemented. This is the principal limitation of using the existing algorithms for low resourced languages which constitutes a very large percentage of languages used across the world. In Cameroon for example there are more than 200 low resourced languages. It's important to tackle the specific aspects of such languages that affect the development of reliable systems. The first limitation is the lack of annotated corpora, this is why we are working to develop systems that learn from limited training data (text and speech). Since structural, social and cultural aspects (tone, agglutination, dialects for example) related to the targeted languages bring additional challenges, we have to deal with the most efficient representation (using among other things multilingual embedding) of the data for ML processing, automatic annotation (using self-supervised learning for example) and extraction of semantic information (NER, WSD, ...) for automatic transcription and translation applications.

Biography:
Since 2008, Paulin Melatagia is a lecturer and researcher at the Department of Computer Science of the Faculty of Sciences of University of Yaounde I in Cameroon where he obtained his Ph.D. in 2012. He is the team leader, since 2017, of the Idasco (data science and complex systems modeling) research team. His research are focused on natural language processing and speech processing with an emphasis for low resourced languages. The main issues of his work are representation learning, automation of labeling and semantic information extraction.

Paulin Melatagia is the scientific secretary of the CRI (Conférence de Recherche en Informatique) since 2015. He is a member of the UMMISCO unit council (http://www.ummisco.fr/) since 2021 and during CARI'2022 he was appointed as member of the first executive committee of ASDS (The African Society in Digital Science).

References :

- 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