Invited Talk:
The Way Towards Massively Multilingual Language Models

François Yvon
Sorbonne Université, CNRS, ISIR
yvon@isir.upmc.fr

Abstract

In this talk, I will discuss the training and evaluation of massively multilingual language models, that can handle dozens or even hundreds of languages. After motivating the development of such models, I will draw some lessons learned in the course of developing Glot500, a language model covering 500 languages, and some associated resources such as language identification softwares. I will also focus on the challenges raised by "low resourced" languages, i.e. languages for which (a) the available training data is often incomplete, highly specialised and also possibly very noisy; (b) the evaluation data are non existent, requiring to use innovative evaluation strategy, e.g. based on various cross-lingual alignment tasks.

Keywords: multilingual language model, low-resource languages, language identification

Bio

François Yvon is a senior CNRS researcher at the ISIR laboratory of Sorbonne-Université in Paris, France, working on Machine Translation and Multilingual Language Models. Before this, F. Yvon has been leading activities in Machine Translation at LISN / LIMSI in Orsay for about 15 years, resulting in more than one hundred scientific publications on all aspects related to the development and evaluation of multilingual language processing technologies, from word and sentence alignment to translation modelling and evaluation, including recent work on multi-domain adaptation in Machine Translation and on cross-lingual transfert learning issues. He has acted as coordinator or Principal Investigator in multiple past national and international projects in Machine Translation and has supervised more than 20 PhDs on related topics. Between 2013 and 2020, Dr. Yvon has also been the general director of the LIMSI laboratory in Orsay. He is a board member of the European chapter of the Association for Computational Linguistics, of the MetaNet network, and has recently contributed as an expert on linguistic technologies for the French language to several European projects (European Language Resource Collection, ELE – European Language Equality, ELG – European Language Grid).