Data Quality for NLP: Challenges and (Possible) Solutions for Multilingualism and Low Resource Languages

A. Seza Doğruöz
Ghent University

Journées Scientifiques du GDR TAL
Centre des Colloques du Campus Condorcet, Paris

Lundi 1er décembre 2025 - 14:15-15h00

Abstract

With the recent advancements in LLMs, data has become more important than ever. However, progress in NLP does not only rely on model architecture but also on the quality of data (both in training and evaluation).

Scarce data sources (e.g., low resource languages) often rely on convenience sampling with limited representation of the linguistic variation and sociocultural context (especially for low resource languages). Annotation is also challenged by limited availability of trained annotators and inconsistent guidelines leading to low inter-annotator agreement and unstable labels.

These upstream issues propagate into benchmarks, which frequently lack coverage and robust evaluation resulting in unreliable model rankings and performance estimates.

Focusing on low resource languages and multilingualism in NLP, we will explore the current data collection and annotation practices (e.g., convenience sampling, issues about representativeness and bias), challenges about reliability and construct validity for benchmarks and evaluations as well as possible solutions to tackle these challenges.