Natural Language Processing for Non-standard Languages and Dialects - Challenges and Current Approaches

Barbara Plank, CIS LMU

Over the last decades, Natural Language Processing (NLP) has largely focused on standard languages with many speakers and abundant data. More recently, there is an increasing interest in moving away from the treatment of languages as monoliths, towards recognizing and modelling the linguistic variability inherent in how many of us, including speakers of minority and non-standard languages, actually use language. From an NLP perspective, languages with few speakers and/or no generally accepted standard are interesting in that they require learning from data that are sparse as well as heterogeneous. This is in stark contrast to large language models (LLMs) trained on massive amounts of data, much of which follows some de-facto standard. In this talk, I will reflect on this asymmetry of how to process small languages with large language models, and discuss the current challenges that NLP research is facing. These challenges include three major dimensions: resources (and resource awareness), modeling non-standard data and human-centric design.