Orthography Standardization in Arabic Dialects

Abstract

Christian Cayralat[1]

[1] Charles University

Spontaneous orthography in Arabic dialects poses one of the biggest obstacles in the way of Dialectal Arabic NLP applications. As the Arab world enjoys a wide array of these widely spoken and recently written, non-standard, low-resource varieties, this thesis presents a detailed account of this relatively overlooked phenomenon. It sets out to show that continuously creating additional noise-free, manually standardized corpora of Dialectal Arabic does not free us from the shackles of non-standard (spontaneous) orthography. Because real-world data will most often come in a noisy format, it also investigates ways to ease the amount of noise in textual data. As a proof of concept, we restrict ourselves to one of the dialectal varieties, namely, Lebanese Arabic. It also strives to gain a better understanding of the nature of the noise and its distribution. All of this is done by leveraging various spelling correction and morphological tagging neural architectures in a multi-task setting, and by annotating a Lebanese Arabic corpus for spontaneous orthography standardization, and morphological segmentation and tagging, among other features. Additionally, a detailed taxonomy of spelling inconsistencies for Lebanese Arabic is presented and is used to tag the corpus. This constitutes the first attempt in Dialectal Arabic research to try and categorize spontaneous orthography in a detailed manner.