Abstract:

The orthography of many resource-scarce languages includes diacritically marked characters. Falling outside the scope of the standard Latin encoding, these characters are often represented in digital language resources as their unmarked equivalents. This renders corpus compilation more difficult, as these languages typically do not have the benefit of large electronic dictionaries to perform diacritic restoration. This paper describes experiments with a machine learning approach that is able to automatically restore diacritics on the basis of local graphemic context. We apply the method to the African languages of Cilubà, Gĩkũyũ, Kĩkamba, Maa, Sesotho sa Leboa, Tshivenda and Yoruba and contrast it with experiments on Czech, Dutch, French, German and Romanian, as well as Vietnamese and Chinese Pinyin.