Empirical Models for an Indic Language Continuum

Niyati Bafna

July 20, 2022

Many Indic languages and dialects of the so-called "Hindi Belt" and surrounding regions in the Indian subcontinent, spoken by more than 100 million people, are severely under-resourced and under-researched in NLP, individually and as a dialect continuum. We first collect monolingual data for 26 Indic languages and dialects, 16 of which were previously zero-resource, and perform exploratory character, lexical and subword cross-lingual alignment experiments for the first time on this linguistic system. We present a novel method for unsupervised cognate/borrowing identification from monolingual corpora designed for low and extremely low resource scenarios, based on combining noisy semantic signals from joint bilingual spaces with orthographic cues modelling sound change; to the best of our knowledge, this is the first work to do so, especially in a (truly) low-resource setup. We create bilingual evaluation lexicons against Hindi for 20 of the languages, and show that our method outperforms both traditional orthography baselines as well as EM-style learnt edit distance matrices, showing that even noisy bilingual embeddings can act as good guides for this task. We release our crawled data in a new collection called "HinDialect"; we also release the evaluation data, code, and results here: `https://github.com/niyatibafna/north-indian-dialect-modelling`