How Many Languages Can a Language Model Model? (invited talk)

Robert Östling
Department of Modern Languages
University of Helsinki
robert.ostling@helsinki.fi

Abstract

One of the purposes of the VarDial workshop series is to encourage research into NLP methods that treat human languages as a continuum, by designing models that exploit the similarities between languages and variants. In my work, I am using a continuous vector representation of languages that allows modeling and exploring the language continuum in a very direct way. The basic tool for this is a character-based recurrent neural network language model conditioned on language vectors whose values are learned during training. By feeding the model Bible translations in a thousand languages, not only does the learned vector space capture language similarity, but by interpolating between the learned vectors it is possible to generate text in unattested intermediate forms between the training languages.

Biography

Robert Östling is working on ways to use parallel corpora in computational linguistics, including machine translation, cross-language learning and language typology.

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
page 62, Osaka, Japan, December 12 2016.