Seq2Seq modelling of complex morphology: A Georgian verb conjugator

Konuralp Kizilay

Universität zu Köln

Learning a Second Language is notoriously difficult. When this language has exceptionally irregular paradigms and little resources, that difficulty is heightened. In my thesis and this talk that presents results from its experimental phase, I am proposing a Deep Learning approach for the low-resource Georgian language, focusing on its complex verb morphology as an approach for developing second language acquisition tools. It has been demonstrated that rule-based systems are brittle and can't handle out-of-vocabulary tokens, fail at more complex forms, explode in size, and do not process minute morphosyntactic or semantic contexts (Faruqui et al. 2016, Feng 2023, Oflazer 1996, Samuelsson & Stajner 2022). Further, conventional machine learning approaches have yielded unsatisfactory results (Elizbarashvili 2024, based on Elizbarashvili et al. 2023: https://github.com/aelizbarashvili/KartuVerbs), with the model not generalising on morphologic features but rather overfitting on lexical IDs. Deep Learning models are universal approximators (Hornk, Stinchcombe & White 1989). Given enough data, a Deep Neural Network is going to be able to represent complex sequential patterns such as conjugation paradigms. By taking this approach, the model generalises on morphologic features. In other words, a learner could input any given verb and yield its paradigm, even if the verb might not have been included in the training data. As an exploratory study of the capabilities of such a model in the context of Georgian morphology, I am working with a classical Encoder-Decoder Architecture as devised by Sutskever et al. 2014 and Cho et al. 2014.