From Machine Learning to Machine Training: How linguistically informed models can improve machine learning approaches in comparative linguistics

Johann-Mattis List
Department of Linguistic and Cultural Evolution, MPI-EVA, Leipzig

Machine learning methods, specifically those methods that are based on concepts of artificial intelligence, are currently revolutionizing science. They beat humans at board games like chess and go, they are actively used by scientists for all kinds of tasks where large amounts of data need to be scanned for specific signals, and they have even begin to tackle really hard problems, like protein folding, which were for a long time thought to be unsolvable without human insights. However, the recent success stories of artificial intelligence and machine learning bear a certain danger for scientific research. Given the increased trust in the power of data-driven, theory-blind approaches, in which signal is supposed to be found in the data alone and theory-driven models are treated with suspicion, we risk to overestimate what machines can do, underestimating the importance of scientific modeling. This has led to a situation where the role of expert knowledge, experience in scientific modeling, and data quality in scientific research are systematically downplayed. Starting from the idea that machine learning methods -- no matter how evolved -- need to be applied to high-quality data, carefully curated by experts, I will present a scenario for future machine learning approaches in the field of comparative linguistics, in which human expertise and computational power are reconciled within a new framework.