6,909 Reasons to Mess Up Your Data

Anders Søgaard
Københavns Universitet
soegaard@hum.ku.dk

ABSTRACT
In computational linguistics we develop tools and on-line services for everything from literature to social media data, but our tools are often optimized to minimize expected error on a single annotated dataset, typically newspaper articles—and evaluated on held-out data sampled from the same dataset. Significance testing across data points randomly sampled from a standard dataset only tells us how likely we are to see better performance on more data points sampled this way, but says nothing about performance on other datasets. This talk discusses how to modify learning algorithms to minimize expected error on future, unseen datasets, with applications to PoS tagging and dependency parsing, including cross-language learning problems. It also discusses the related issue of how to best evaluate NLP tools (intrinsically) taking their possible out-of-domain applications into account.

KEYWORDS: Domain Variation, PoS Tagging, Dependency Parsing, Evaluation.