Well-Behaved Nonparametric Statistical Learning for Parsing and Translation

Khalil Sima'an
University of Amsterdam
k.simaan@uva.nl

Abstract

The lion's share of the success in automatic parsing and translation of text over the last decade is attributed to statistical learning from data. Interestingly, both in parsing and in translation the successful models employ a large, redundant set of fragments as model parameters. Full sentences together with their parse trees or translations, and all their connected subgraphs are considered fragments. These models, including Data-Oriented Parsing (DOP) and Phrase-Based Statistical Machine Translation (PBSMT), are nonparametric models as they grow to fit any data they are trained on. The Maximum-Likelihood Estimator (MLE) is inapplicable for such models as it overfits. The statistical commonly used estimator for DOP and PBSMT is known to behave badly, i.e., it is inconsistent – does not converge to the relative frequency as data grows large. For good performance this estimator often demands tweaking and adjustments (by intuition or on development data). In this talk I will discuss this estimation problem and show how it relates to statistical density estimation, including smoothing by Leave-One-Out, K-Nearest Neighbor and Parzen Window. I will also explain how good performance is being achieved in practice with a badly behaved estimator. Consequently I will apply this knowledge to formulate the bias that underlies a family of well-behaved (consistent) estimators known to provide good performance.

Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning, page 65,
Tilburg, May 2009.