Perturbation learning for general-purpose text validation

Anonymous NAACL submission

Abstract

Language learners and generative models alike are often in need of text validation: checking how natural a certain sentence sounds within a given language or style. In this paper, we propose an approach to training a statistical validation model on a text corpus with no supervision. This is achieved by applying random perturbations to sentences from the corpus and training a recurrent neural network to discriminate between the original sentences and the perturbed ones. Choosing the right perturbation model, however, is far from trivial: the resulting validation model has to generalize beyond the specific perturbation we introduced and be able to recognize previously unseen kinds of deviations from the norm it learned from the corpus. We develop several perturbation models, demonstrate and compare their generalization ability.