Which Algorithms Can and Can't Learn Identity Effects in Phonological Grammars

Paul Tupper
Simon Fraser University

Abstract: Suppose you are told that in an alien language the strings AA, MM, DD, RR are all valid words, whereas BF, QG, CE, TM are not. You are then asked if you think EE is a valid word. Most people identify EE as a valid word; they are sensitive to the fact that all the valid words consist of two identical letters, whereas the invalid words do not. This is known as an identity effect, and has been observed in artificial language learning experiments and in a diverse range of natural languages. I give a formal proof that many popular learning framework, including methods for training neural network of arbitrary number of layers, cannot learn such identity effects. The proof exploits symmetries in the architectures and their training regimes to show that such learners cannot perform with human-like behaviour on these grammatical judgment tasks.