When Big Data is Bad (and Why Small Data is Enough)

Charles Yang

The sparsity of linguistic data has been rightly vilified in natural language processing although it has not been fully appreciated by language scientists. In a psychological setting such as language acquisition, Big Data may turn out to be a curse rather than blessing. Linguistic regularities such as rules and constructions can easily be washed out by the large number of lexical items that lie on the long tail of Zipf’s law. Building on a range of examples from child language, I argue that most aspects of the grammar must be acquirable under a modest amount of primary linguistic data, and perhaps only the portion that contain relatively frequent lexical items. These considerations in turn put considerable constraints on both the learning model and the complexity of the grammar model that needs to be assumed. A detailed case study will focus on the double object and to-dative dative construction in English. I show that the sufficient and necessary conditions for their productivity can be inductively learned from a reasonable corpus of child directed English speech, following a general principle of learning (Yang 2016).