Big Data and Little Learners

John C. Trueswell (trueswel@psych.upenn.edu)
(Symposium Organizer) Department of Psychology
University of Pennsylvania, USA

Linda B. Smith (smith4@indiana.edu)
Department of Psychological and Brain Sciences
Indiana University, USA

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, USA

Charles Yang (charles.yang@ling.upenn.edu)
Department of Linguistics and Computer Science
University of Pennsylvania, USA

Keywords: data science; deep learning; poverty of the stimulus; indirect negative evidence; language development

Introduction

Recent advances in the data sciences, particularly within the area of language technology, have been impressive and non-incremental. For example, within the domain of language translation, the application of deep Long Short Term Memory (LSTM) neural networks to large bodies of text have resulted in a 60% reduction in translation errors from traditional methods, significantly closing the gap between machine and human performance (Wu et al., 2016). Similarly impressive advances have been observed in, e.g., speech recognition (Hinton et al., 2012), syntactic parsing (Dyer et al., 2015) and automatic content extraction (Berant et al., 2015).

Clearly, excitement is justified as a new era of linguistic technology is emerging. But should this excitement lead to a fundamental rethinking of our theories of child language and cognition? Doesn't the "poverty of the stimulus" still pose a problem for human language learners? What role do hierarchical linguistic formalisms play within statistical theories of language learning and use? This symposium brings together leading figures in cognitive science who offer different informed perspectives on these matters.

The data and the learner from a developmental perspective

Linda Smith (Indiana University)

The world offers data to learning systems that is massive in total scale and that comes in many forms. However, the relevant data for any learning system are only those that actually engage the learning mechanisms of that system. For living and breathing learners, this engagement begins with their sensory systems. Sensory systems are on bodies that move through the world – constrained by the physics of space and time – and thus the sampled data are constrained and ordered by space in time. Human infants learn their first words during a period in which their bodies (and brains) change dramatically and systematically and do so in ways that put those sensory systems in different parts of the data space at different points in development. The data for learning – and the learning tasks to be solved – are systematically ordered by development itself. This talk will present evidence from a large corpus of head-camera data recorded in infants' homes (over 500 million frames extracted at 1 Hz) that illustrate how human development (and the reality of bodies learning in space and time) fundamentally changes the questions to be asked and the computational answers to how language is learned.

Existence proofs and computational mechanisms

Charles Yang (University of Pennsylvania)

Mathematicians have always drawn a useful distinction between existence and constructive results. An analogy can be made in the study of language acquisition, especially in the age of Big Data and Big Machines. While distributional regularities can be captured by idealized statistical models, it is a different matter whether, and how, such regularities are exploited by computational mechanisms available to human children.

Consider the use of indirect negative evidence in language acquisition. A specific example concerns the predicative and attributive use of the so-called a-adjectives in English: "the cat is asleep/away" vs. "*the asleep/away cat". Indirect negative evidence can be formulated in certain probabilistic models of inference: the conspicuous absence of forms such as "the asleep cat" reduces the learner's confidence in the hypothesis that permits such expressions. While these models are presented as existence proofs without commitments to psychological mechanisms, they are still unlikely to succeed when evaluated against realistic statistical distribution of adjectives in a large corpus of child-directed English speech.

The alternative approach is to avoid the use of indirect negative evidence, and to develop a transparently mechanistic models that can be readily tested. I review the Tolerance Principle, a parameter-free model of inductive generalization, and its application to morphological and syntactic acquisition, including artificial language learning (joint work with Kathryn Schuler and Elissa Newport). Furthermore, the Tolerance Principle suggests that language acquisition may succeed *only* with small data, the kind similar to the small vocabularies of young children. This supports the view that cognitive and maturational constraints support rather than hinder language development.

On what you can't learn from (merely) all the data in the world, and what else is needed

Josh Tenenbaum (University of Edinburgh)

Recent successes with recurrent neural networks and other big-data techniques in AI applications raise the question of whether similar approaches might explain human language acquisition. How far can the data of language take us alone, with little other structure? I will first describe some experiments testing RNN models developed by Google that can perform some truly impressive feats in language technology, yet at the same time fail a number of basic tests of understanding syntax and semantics that cognitive scientists have long been interested in, as well as some new benchmarks that we have come up with. They often fail for interesting reasons, based on the differences between their linear (sequential) processing architecture and the hierarchical structure of thought, their emphasis on character-level modeling as opposed to words and phrases, and their lack of interfaces to core cognition outside language. Their successes and failures illustrate how both advocates and critics of early statistical language learning were correct — Chomsky and Gleitman and Pinker were right after all, but Elman and Hinton were also right. They were just right about different things, and we can learn much by re-interpreting early debates.

As a way forward, I argue for combining smart statistics with more structured, hierarchical representations, interfacing to a cognitively grounded semantics. I report some promising results, although we are far from being able to implement this at the scale Google requires. I will also sketch ideas for how RNNs can make these more structured approaches work better, with the hope of integrating these often-opposing traditions to best make progress.

References

Berant, J., Alon, N., Dagan, I, & Goldberger, J. (2015). Efficient global learning and entailment. *Computational Linguistics, 42,* 221-263.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. (2015).Transition-based dependency parsing with stack long short-term memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics,* 334-343.

Hinton, G. Deng, L., Yu, D., Dahl, G et al., (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, 29,* 82-97.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv E-prints,* 2arXiv:1609.08144.