Modeling language learning from the ground up

Christos Christodoulopoulos*[1], Dan Roth[2], and Cynthia Fisher[3]

[1]Amazon Research Cambridge – United Kingdom
[2]University of Pennsylvania [Philadelphia] – United States
[3]University of Illinois at Urbana-Champaign [Urbana] – United States

Abstract

We present a computational model of the developmental origins of Syntactic Bootstrapping (the proposal that children use syntactic knowledge to infer the meanings of unknown words; Landau & Gleitman, 1985). We explore how the representations needed for Syntactic Bootstrapping might be derived from distributional properties of words, a few known 'seed' nouns, and minimal built-in assumptions about syntax and meaning.
We used a Semantic Role Labeling system, trained on child-directed speech data annotated for semantic roles. We present three experiments in which the model's task is to predict the semantic roles of words (e.g., agent, patient, goal).

Experiment 1 shows that simple syntactic features (the set of nouns in the sentence, or position before or after the verb) are sufficient to learn semantic roles. These features are based on the Structure Mapping hypothesis (Fisher et al., 2010). In training, via a supervised machine-learning paradigm, the model learns to use these simple features to predict semantic roles. Here, we provide veridical semantic-role feedback. We also assume complete knowledge of all nouns and verbs in each utterance ('gold-standard' part-of-speech tags).

Experiment 2 removes this assumption and addresses how the representations needed for Structure Mapping can themselves be bootstrapped. We designed a word category learner based on distributional information. Following Connor, Fisher, & Roth (2012), we used an unsupervised Hidden Markov Model (HMM) and tagged some HMM states as noun states, based on a list of up to 77 'seed' nouns comprising mostly concrete objects, the children's names, and a few pronouns. Any HMM state that contained 4 or more seed nouns was labelled as a noun state. With only 20 seed nouns, noun identification was above 80%. To identify verb states, we compared the method of Connor, Fisher, & Roth (2012) – a heuristic based on the Structure Mapping hypothesis – against a random-guessing baseline using only the exclusionary principle that nouns and function words cannot be verbs. Excluding nouns and function words was enough to successfully identify verbs even when using the guessing baseline (75% precision). Furthermore, we show that these random verb predictions (combined with the noun heuristic) provide useful abstract syntactic features to guide semantic role learning, enabling similar or better results to the ones of Experiment 1. We show that, surprisingly, verbs can be identified without any 'seed' verbs, simply via the prior assumptions that (a) a sentence should contain a verb, (b) and it can't also be a noun or a function word.

Experiment 3 relaxes the assumption of veridical semantic feedback. Using a much weaker

*Speaker

signal, based on the animacy properties of a few common concrete nouns, the model is still able to predict correct semantic roles.

Connor, M., Fisher, C., & Roth, D. (2012). Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision. Cognitive Aspects of Computational Language Acquisition.

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. Wiley Interdisciplinary Reviews: Cognitive Science.

Landau, B., & Gleitman, L. R. (1985). Language and experience: evidence from the blind child. Harvard University Press.

Keywords: syntactic bootstrapping, semantic role labeling, structure mapping