Proceedings of the Society for Computation in Linguistics

Volume 1 Article 39

2018

Predicting Fine-Grained Syntactic Typology from Surface Features

Dingquan Wang
Johns Hopkins University, wddabc@gmail.com

Jason Eisner
Johns Hopkins University, jason@cs.jhu.edu

Recommended Citation

Wang, Dingquan and Eisner, Jason (2018) "Predicting Fine-Grained Syntactic Typology from Surface Features," *Proceedings of the Society for Computation in Linguistics*: Vol. 1 , Article 39.
DOI: https://doi.org/10.7275/R5F769RV
Available at: https://scholarworks.umass.edu/scil/vol1/iss1/39

Predicting Fine-Grained Syntactic Typology from Surface Features

Abstract of Wang and Eisner (2017), previously published in TACL journal & presented at ACL 2017

Dingquan Wang and Jason Eisner
Department of Computer Science, Johns Hopkins University
{wdd,eisner}@jhu.edu

Motivation

We are motivated by the longstanding challenge of determining the structure of a language from its superficial features. Principles & Parameters theory (Chomsky, 1981) hypothesized that human babies are born with an evolutionarily tuned system that is specifically adapted to natural language, which can predict typological properties ("parameters") by spotting telltale configurations in purely linguistic input (Gibson and Wexler, 1994). Here we investigate whether such configurations even exist, by asking an artificial system to find them.

Synopsis

We show how to predict the basic word-order facts[1] in a treebank of an unspecified and possibly unknown language, given only a corpus of part-of-speech (POS) sequences. We predict how often direct objects follow their verbs, how often adjectives follow their nouns, and in general the directionality of each of the 57 dependency relations recognized by the Universal Dependencies (UD) project (Nivre et al., 2016).

We define the directionality to be a real number in $[0, 1]$: the fraction of tokens of this relation that are "right-directed," in the sense that the child (modifier) falls to the right of its parent (head). Results for the nominal subject, direct object, and case/adposition relations are shown below on 20 treebanks spanning 18 different languages (e.g., nl=Nederlands=Dutch). The points corresponding to a given treebank in a language were jointly predicted without any observation of the true trees or directionalities in that treebank.

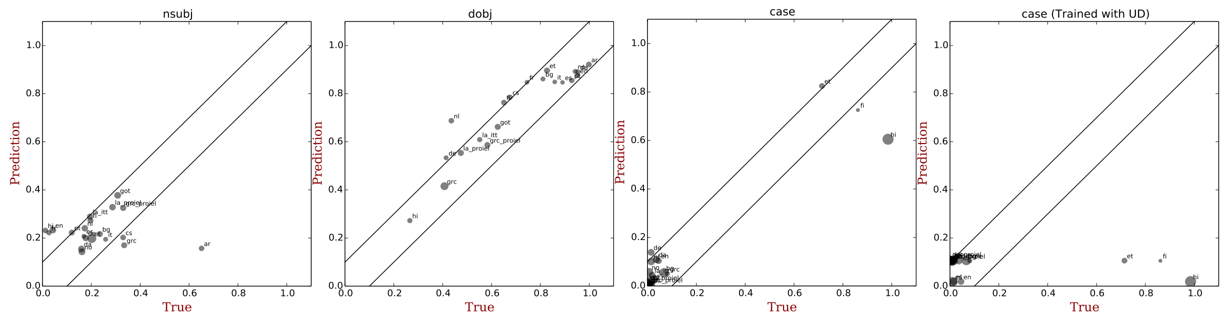


Figure 1: Scatterplots of predicted (y-axis) vs. true (x-axis) directionalities for each language. The first 3 plots show predictions by our full system. The 4th shows how performance degrades without the use of synthetic data to illustrate the surface word order of postpositional languages (see below). These plots are typical; error is also low according to aggregate performance metrics.

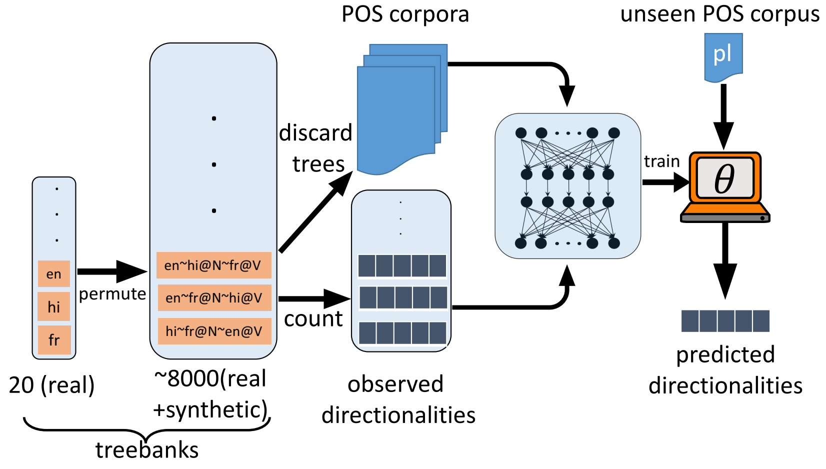
[1] Some theories may posit deeper parameters ("head-final") that underlie and correlate these basic word-order facts.

Methods

The best version of our system considers over 40,000 numerical measurements from the collection of POS sequences. E.g., among all cases where an adjective and a noun appear within the same 4-word window, we measure the fraction in which the adjective follows the noun, which would tend to correlate with the directionality of the adjectival modifier relation. The system predicts directionalities using a neural network with these numerical features as input. Training the neural network determines which of the features or co-occurrences of features are actually predictive.

While determining the syntactic structure of a language is usually regarded as an *unsupervised* learning problem, our innovation is to treat it as a *supervised* one, with supervision provided by analyses of *other* languages.

Standard unsupervised learning methods such as Expectation Maximization (Dempster et al., 1977) try to explain the observed data using latent variables such as syntax trees. A metric such as log-likelihood is used to measure the quality of the explanation. Unfortunately, over the past 4 decades, this strategy has not led to much success on the problem of grammar induction. First, finding an explanation that scores highly on log-likelihood is computationally intractable in theory and also difficult in practice: known methods get badly stuck in poor local maxima. Second, locally improving the log-likelihood metric often *degrades* the agreement with the syntactic structures assigned by linguists. The reason is that log-likelihood can be improved by using the latent variables in linguitically unconventional ways, perhaps using them to capture non-syntactic regularities (e.g., register or topic) or to capture syntactic regularities in non-canonical ways.



Indeed, we determined empirically that the state-of-the-art grammar induction systems fail dramatically at typology prediction. We do far better with a supervised approach, where we train a system to predict how *linguists* will annotate a language. We show our system the directionalities chosen by linguists in the 20 treebanks of 18 human languages. However, this is a tiny amount of training data, given the number of free parameters in our system. So (crucially) we augment it by also training the system on about 8000 synthetic languages that are, roughly speaking, in the space of possible human languages. The treebank of each synthetic language is generated from the UD treebank of some real language by stochastically permuting the dependents of all nouns and/or verbs to match the dependent orders of other real UD languages. These synthetic languages were developed in a previous paper (Wang and Eisner, 2016), where we evaluated their naturalness and the ability of their treebanks to flesh out gaps in the original set of 20.