

# Fine-Grained Prediction of Syntactic Typology: Discovering Latent Structure with *Supervised* Learning

Dingquan Wang and Jason Eisner

Department of Computer Science, Johns Hopkins University  
{wdd, eisner}@jhu.edu

## Abstract

We show how to predict the basic word-order facts of a novel 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 directionalities of all dependency relations. Such typological properties could be helpful in grammar induction. While such a problem is usually regarded as unsupervised learning, our innovation is to treat it as *supervised* learning, using a large collection of realistic synthetic languages as training data. The supervised learner must identify *surface* features of a language’s POS sequence (hand-engineered or neural features) that correlate with the language’s *deeper* structure (latent trees). In the experiment, we show: 1) Given a small set of real languages, it helps to add many synthetic languages to the training data. 2) Our system is robust even when the POS sequences include noise. 3) Our system on this task outperforms a grammar induction baseline by a large margin.

## 1 Introduction

Descriptive linguists often characterize a human language by its *typological properties*. For instance, English is an SVO-type language because its basic clause order is Subject-Verb-Object (SVO), and also a prepositional-type language because its adpositions normally precede the noun. Identifying basic word order must happen early in the acquisition of syntax, and presumably guides the initial interpretation of sentences and the acquisition of a finer-grained grammar. In this paper, we propose a method for doing this. While we focus on word order, one could try similar methods for other typological classifications (syntactic, morphological, or phonological).

The problem is challenging because the language’s true word order statistics are computed from syntax trees, whereas our method has access only to a POS-tagged corpus. Based on these POS sequences alone, we predict the *directionality* of each type of dependency relation. 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). For example, the `dobj` relation points from a verb to its direct object (if any), so a directionality of 0.9—meaning that 90% of `dobj` dependencies are right-directed—indicates a dominant verb-object order. (See Table 1 for more such examples.) Our system is trained to predict the relative frequency of rightward dependencies for each of 57 dependency types from the Universal Dependencies project (UD). We assume that all languages draw on the same set of POS tags and dependency relations that is proposed by the UD project (see §3), so that our predictor works across languages.

Why do this? Liu (2010) has argued for using these directionality numbers in  $[0, 1]$  as fine-grained and robust *typological descriptors*. We believe that these directionalities could also be used to help define an *initializer, prior, or regularizer* for tasks like grammar induction or syntax-based machine translation. Finally, the vector of directionalities—or the feature vector that our method extracts in order to predict the directionalities—can be regarded as a *language embedding* computed from the POS-tagged corpus. This language embedding may be useful as an input to multilingual NLP systems, such as the cross-linguistic neural dependency parser of Ammar et al. (2016). In fact, some multilingual NLP systems already condition on typological properties looked up in the World Atlas of Language Structures, or WALS (Dryer and Haspelmath, 2013), as

Typology	Example
Verb-Object (English)	<p>She gave me a raise</p>
Object-Verb (Hindi)	<p>She me a raise gave vah mujhe ek uthaane diya</p>
Prepositional (English)	<p>She is in a car</p>
Postpositional (Hindi)	<p>She a car in is vah ek kaar mein hai</p>
Adjective-Noun (English)	<p>This is a red car</p>
Noun-Adjective (French)	<p>This is a car red Ceci est une voiture rouge</p>

Table 1: Three typological properties in the World Atlas of Language Structures (Dryer and Haspelmath, 2013), and how they affect the directionality of Universal Dependencies relations.

we review in §8. However, WALS does not list all properties of all languages, and may be somewhat inconsistent since it collects work by many linguists. Our system provides an automatic alternative as well as a methodology for generalizing to new properties.

More broadly, we are motivated by the challenge of determining the structure of a language from its superficial features. Principles & Parameters theory (Chomsky, 1981; Chomsky and Lasnik, 1993) famously—if controversially—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. Here we investigate whether such a system can be tuned by gradient descent. It is at least plausible that useful superficial features do exist: e.g., if nouns often precede verbs but rarely follow verbs, then the language may be verb-final.

## 2 Approach

We depart from the traditional approach to latent structure discovery, namely unsupervised learning. Unsupervised syntax learners in NLP tend to be rather inaccurate—partly because they are failing to maximize an objective that has many local optima, and partly because that objective does not capture all the factors that linguists consider when assigning

syntactic structure. Our remedy in this paper is a supervised approach. We simply imitate how linguists have analyzed other languages. This supervised objective goes beyond the log-likelihood of a PCFG-like model given the corpus, because linguists do not merely try to predict the surface corpus. Their dependency annotations may reflect a cross-linguistic theory that considers semantic interpretability and equivalence, rare but informative phenomena, consistency across languages, a prior across languages, and linguistic conventions (including the choice of latent labels such as *dobj*). Our learner does not consider these factors explicitly, but we hope it will identify correlates (e.g., using deep learning) that can make similar predictions. Being supervised, our objective should also suffer less from local optima. Indeed, we could even set up our problem with a *convex* objective, such as (kernel) logistic regression, to predict each directionality separately.

Why hasn’t this been done before? Our setting presents unusually sparse data for supervised learning, since *each training example is an entire language*. The world presumably does not offer enough natural languages—particularly with machine-readable corpora—to train a good classifier to detect, say, Object-Verb-Subject (OVS) languages, especially given the class imbalance problem that OVS languages are empirically rare, and the non-IID problem that the available OVS languages may be evolutionarily related.<sup>1</sup> We mitigate this issue by training on the Galactic Dependencies treebanks (Wang and Eisner, 2016), a collection of more than 50,000 human-like synthetic languages. The treebank of each synthetic language is generated by stochastically permuting the subtrees in a given real treebank to match the word order of other real languages. Thus, we have many synthetic languages that are Object-Verb like Hindi but also Noun-Adjective like French. We know the true directionality of each synthetic language and we would like our classifier to predict that directionality, just as it would for a real language. We will show that our system’s accuracy benefits from fleshing out the training set in this way, which can be seen as a form of regularization.

<sup>1</sup>Properties shared within an OVS language family may appear to be consistently predictive of OVS, but are actually confounds that will not generalize to other families in test data.

A possible criticism of our work is that obtaining the input POS sequences requires human annotators, and perhaps these annotators could have answered the typological classification questions as well. Indeed, this criticism also applies to most work on grammar induction. We will show that our system is at least robust to noise in the input POS sequences (§7.4). In the future, we hope to devise similar methods that operate on raw word sequences.

### 3 Data

We use two datasets in our experiment:

**UD: Universal Dependencies version 1.2 (et al., 2015)** A collection of dependency treebanks for 37 languages, annotated in a consistent style with POS tags and dependency relations.

**GD: Galactic Dependencies version 1.0 (Wang and Eisner, 2016)** A collection of projective dependency treebanks for 53,428 synthetic languages, using the same format as UD. 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. Using this “mix-and-match” procedure, the GD collection fills in gaps in the UD collection—which covers only a few possible human languages.

### 4 Task Formulation

We now formalize the setup of the fine-grained typological prediction task. Let  $\mathcal{R}$  be the set of universal relation types, with  $N = |\mathcal{R}|$ . We use  $\overset{r}{\rightarrow}$  to denote a rightward dependency token with label  $r \in \mathcal{R}$ .

**Input** for language  $L$ : A POS-tagged corpus  $\mathbf{u}$ . (“ $\mathbf{u}$ ” stands for “unparsed.”)

**Output** for language  $L$ : Our system predicts  $p(\rightarrow | r, L)$ , the probability that a token in language  $L$  of an  $r$ -labeled dependency will be right-oriented. It predicts this for *each* dependency relation type  $r \in \mathcal{R}$ , such as  $r = \text{dobj}$ . Thus, the output is a vector of predicted probabilities  $\hat{\mathbf{p}} \in [0, 1]^N$ .

**Training:** We set the parameters of our system using a collection of training pairs  $(\mathbf{u}, \mathbf{p}^*)$ , each of which corresponds to some UD or GD training language  $L$ . Here  $\mathbf{p}^*$  denotes the true vector of probabilities as empirically estimated from  $L$ ’s treebank.

**Evaluation:** Over pairs  $(\mathbf{u}, \mathbf{p}^*)$  that correspond to held-out *real* languages, we evaluate the expected loss of the predictions  $\hat{\mathbf{p}}$ . We use  $\varepsilon$ -insensitive loss<sup>2</sup> with  $\varepsilon = 0.1$ , so our evaluation metric is

$$\sum_{r \in \mathcal{R}} p^*(r | L) \cdot \text{loss}_\varepsilon(\hat{p}(\rightarrow | r, L), p^*(\rightarrow | r, L)) \quad (1)$$

where

- $\text{loss}_\varepsilon(\hat{p}, p^*) \equiv \max(|\hat{p} - p^*| - \varepsilon, 0)$
- $p^*(\rightarrow | r, L) = \frac{\text{count}_L(\overset{r}{\rightarrow})}{\text{count}_L(r)}$  is the empirical estimate of  $p(\rightarrow | r, L)$ .
- $\hat{p}(\rightarrow | r, L)$  is the system’s prediction of  $p^*$

The aggregate metric (1) is an expected loss that is weighted by  $p^*(r | L) = \frac{\text{count}_L(\overset{r}{\rightarrow})}{\sum_{r' \in \mathcal{R}} \text{count}_L(r')}$ , to emphasize relation types that are more frequent in  $L$ .

Why this loss function? We chose an L1-style loss, rather than L2 loss or log-loss, so that the aggregate metric is not dominated by outliers. We took  $\varepsilon > 0$  in order to forgive small errors: if some predicted directionality is already “in the ballpark,” we prefer to focus on getting other predictions right, rather than fine-tuning this one. Our intuition is that errors  $< \varepsilon$  in  $\hat{\mathbf{p}}$ ’s elements will not greatly harm downstream tasks that analyze individual sentences, and might even be easy to correct by grammar reestimation (e.g., EM) that uses  $\hat{\mathbf{p}}$  as a starting point.

In short, we have the intuition that if our predicted  $\hat{\mathbf{p}}$  achieves small  $\text{loss}_\varepsilon$  on the frequent relation types, then  $\hat{\mathbf{p}}$  will be helpful for downstream tasks, although testing that intuition is beyond the scope of this paper. One could tune  $\varepsilon$  on a downstream task.

### 5 Simple “Expected Count” Baseline

Before launching into our full models, we warm up with a simple baseline heuristic called *expected count* (EC), which is reminiscent of Principles & Parameters. The idea is that if ADJs tend to precede nearby NOUNs in the sentences of language  $L$ , then `amod` probably tends to point leftward in  $L$ . After all, the training languages show that when ADJ and NOUN are nearby, they are usually linked by `amod`.

Fleshing this out, EC estimates directionalities as

$$\hat{p}(\rightarrow | r, L) = \frac{\text{ecount}_L(\overset{r}{\rightarrow})}{\text{ecount}_L(\overset{r}{\rightarrow}) + \text{ecount}_L(\overset{r}{\leftarrow})} \quad (2)$$

<sup>2</sup>Proposed by V. Vapnik for support vector regression.

where we estimate the expected  $\leftarrow^r$  and  $\rightarrow^r$  counts by

$$\text{ecount}_L(\rightarrow^r) = \sum_{u \in \mathbf{u}} \sum_{\substack{1 \leq i < j \leq |u| \\ j-i < w}} p(\rightarrow^r | u_i, u_j) \quad (3)$$

$$\text{ecount}_L(\leftarrow^r) = \sum_{u \in \mathbf{u}} \sum_{\substack{1 \leq i < j \leq |u| \\ j-i < w}} p(\leftarrow^r | u_i, u_j) \quad (4)$$

Here  $u$  ranges over tag sequences (sentences) of  $\mathbf{u}$ , and  $w$  is a window size that characterizes “nearby.”<sup>3</sup>

In other words, we ask: given that  $u_i$  and  $u_j$  are nearby tag tokens in the test corpus  $\mathbf{u}$ , are they likely to be linked? Formulas (3)–(4) count such a pair as a “soft vote” for  $\rightarrow^r$  if such pairs tended to be linked by  $\rightarrow^r$  in the treebanks of the training languages,<sup>4</sup> and a “soft vote” for  $\leftarrow^r$  if they tended to be linked by  $\leftarrow^r$ .

**Training:** For any two tag types  $t, t'$  in the universal POS tagset  $\mathcal{T}$ , we simply use the training treebanks to get empirical estimates of  $p(\cdot | t, t')$ , taking

$$p(\rightarrow^r | t, t') = \frac{\sum_L s_L \cdot \text{count}_L(t \rightarrow^r t')}{\sum_L s_L \cdot \text{count}_L(t, t')} \quad (5)$$

and similarly for  $p(\leftarrow^r | t, t')$ . This can be interpreted as the (unsmoothed) fraction of  $(t, t')$  within a word window where  $t$  is the  $r$ -type parent of  $t'$ , computed by micro-averaging over languages. To get a fair average over languages, equation (5) downweights the languages that have larger treebanks, yielding a *weighted* micro-average in which we define the weight  $s_L = 1 / \sum_{t \in \mathcal{T}, t' \in \mathcal{T}} \text{count}_L(t, t')$ .

As we report later in Table 5, even this simple supervised heuristic performs significantly better than state-of-the-art grammar induction systems. However, it is not a *trained* heuristic: it has no free parameters that we can tune to optimize our evaluation metric. For example, it can pay too much attention to tag pairs that are not discriminative. We therefore proceed to build a trainable, feature-based system.

## 6 Proposed Model Architecture

To train our model, we will try to minimize the evaluation objective (1) averaged over the training lan-

<sup>3</sup>In our experiment, we chose  $w = 8$  by cross-validation over  $w = 2, 4, 8, 16, \infty$ .

<sup>4</sup>Thus, the EC heuristic examines the correlation between relations and tags in the training treebanks. But our methods in the next section will follow the formalization of §4: they do not examine a training treebank beyond its directionality vector  $\mathbf{p}^*$ .

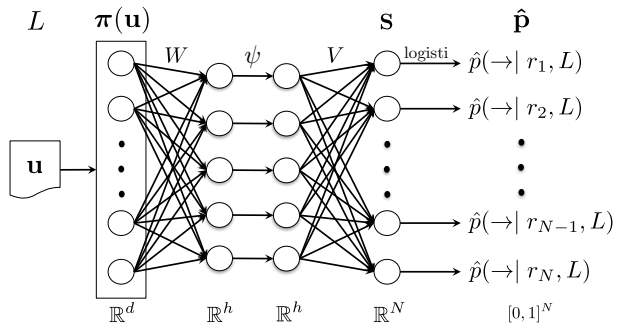


Figure 1: Basic predictive architecture from equations (6)–(7).  $\mathbf{b}_W$  and  $\mathbf{b}_V$  are suppressed for readability.

guages, plus a regularization term given in §6.4.<sup>5</sup>

### 6.1 Directionality predictions from scores

Our predicted directionality for relation  $r$  will be

$$\hat{p}(\rightarrow | r, L) = 1 / (1 + \exp(-\mathbf{s}(\mathbf{u})_r)) \quad (6)$$

$\mathbf{s}(\mathbf{u})$  is a parametric function (see §6.2 below) that maps  $\mathbf{u}$  to a **score** vector in  $\mathbb{R}^N$ . Relation type  $r$  should get positive or negative score according to whether it usually points right or left. The formula above converts each score to a directionality—a probability in  $(0, 1)$ —using a logistic transform.

### 6.2 Design of the scoring function $\mathbf{s}(\mathbf{u})$

To score all dependency relation types given the corpus  $\mathbf{u}$ , we use a feed-forward neural network with one hidden layer (Figure 1):

$$\mathbf{s}(\mathbf{u}) = V \psi(W \pi(\mathbf{u}) + \mathbf{b}_W) + \mathbf{b}_V \quad (7)$$

$\pi(\mathbf{u})$  extracts a  $d$ -dimensional feature vector from the corpus  $\mathbf{u}$  (see §6.3 below).  $W$  is a  $h \times d$  matrix that maps  $\pi(\mathbf{u})$  into a  $h$ -dimensional space and  $\mathbf{b}_W$  is a  $h$ -dimensional bias vector.  $\psi$  is an element-wise activation function.  $V$  is a  $N \times h$  matrix whose rows can be regarded as learned embeddings of the dependency relation types.  $\mathbf{b}_V$  is a  $N$ -dimensional bias vector that determines the default rightwardness of each relation type. We give details in §7.5.

The hidden layer  $\psi(W \pi(\mathbf{u}) + \mathbf{b}_W)$  can be regarded as a latent representation of the language’s word order properties, from which potentially *correlated* predictions  $\hat{\mathbf{p}}$  are extracted.

<sup>5</sup>We gave all training languages the same weight. In principle, we could have downweighted the synthetic languages as out-of-domain, using cross-validation to tune the weighting.

### 6.3 Design of the featurization function $\pi(\mathbf{u})$

Our current feature vector  $\pi(\mathbf{u})$  considers only the POS tag sequences for the sentences in the unparsed corpus  $\mathbf{u}$ . Each sentence is augmented with a special boundary tag # at the start and end. We explore both hand-engineered features and neural features.

**Hand-engineered features.** Recall that §5 considered which tags appeared near one another in a given order. We now devise a slew of features to measure such co-occurrences in a variety of ways. By training the weights of these many features, our system will discover which ones are actually predictive.

Let  $g(t | j) \in [0, 1]$  be some measure (to be defined shortly) of the **prevalence** of tag  $t$  near token  $j$  of corpus  $\mathbf{u}$ . We can then measure the prevalence of  $t$ , both overall and just near tokens of tag  $s$ :<sup>6</sup>

$$\pi_t = \text{mean}_j g(t | j) \quad (8)$$

$$\pi_{t|s} = \text{mean}_{j: T_j=s} g(t | j) \quad (9)$$

where  $T_j$  denotes the tag of token  $j$ . We now define versions of these quantities for particular prevalence measures  $g$ .

Given  $w > 0$ , let the **right window**  $W_j$  denote the sequence of tags  $T_{j+1}, \dots, T_{j+w}$  (padding this sequence with additional # symbols if it runs past the end of  $j$ 's sentence). We define quantities  $\pi_{t|s}^w$  and  $\pi_t^w$  via (8)–(9), using a version of  $g(t | j)$  that measures the fraction of tags in  $W_j$  that equal  $t$ . Also, for  $b \in \{1, 2\}$ , we define  $\pi_{t|s}^{w,b}$  and  $\pi_t^{w,b}$  using a version of  $g(t | j)$  that is 1 if  $W_j$  contains at least  $b$  tokens of  $t$ , and 0 otherwise.

For each of these quantities, we also define a corresponding **mirror-image** quantity (denoted by negating  $w > 0$ ) by computing the same feature on a reversed version of the corpus.

We also define “truncated” versions of all quantities above, denoted by writing  $\hat{\cdot}$  over the  $w$ . In these, we use a **truncated window**  $\hat{W}_j$ , obtained from  $W_j$  by removing any suffix that starts with #

<sup>6</sup>In practice, we do backoff smoothing of these means. This avoids subsequent division-by-0 errors if tag  $t$  or  $s$  has count 0 in the corpus, and it regularizes  $\pi_{t|s}/\pi_t$  toward 1 if  $t$  or  $s$  is rare. Specifically, we augment the denominators by adding  $\lambda$ , while augmenting the numerator in (8) by adding  $\lambda \cdot \text{mean}_{j,t} g(t | j)$  (unsmoothed) and the numerator in (9) by adding  $\lambda$  times the smoothed  $\pi_t$  from (8).  $\lambda > 0$  is a hyperparameter (see §7.5).

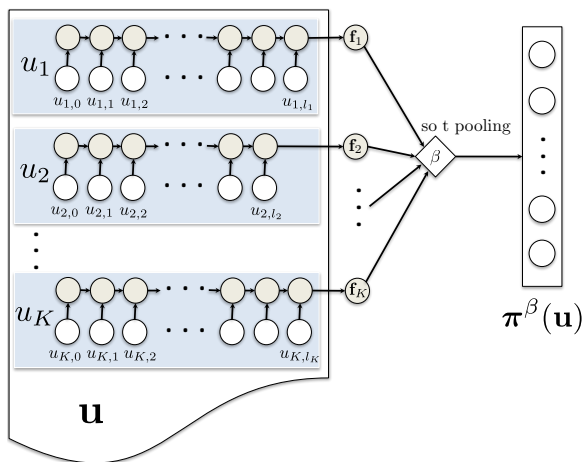


Figure 2: Extracting and pooling the neural features.

or with a copy of tag  $T_j$  (that is,  $s$ ).<sup>7</sup> As an example,  $\pi_{N|V}^{\hat{8},2}$  asks how often a verb is followed by at least 2 nouns, within the next 8 words of the sentence *and before the next verb*. A high value of this is a plausible indicator of a VSO-type or VOS-type language.

We include the following features for each tag pair  $s, t$  and each  $w \in \{1, 3, 8, 100, -1, -3, -8, -100, \hat{1}, \hat{3}, \hat{8}, \hat{100}, -\hat{1}, -\hat{3}, -\hat{8}, -\hat{100}\}$ :<sup>8</sup>

$$\pi_t^w, \pi_{t|s}^w, \pi_{t|s}^w \cdot \pi_s^w, \pi_{t|s}^w / \pi_t^w, \pi_t^w / \pi_{t|s}^w, \pi_{t|s}^w / \pi_{t|s}^{-w}$$

where we define  $x//y = \min(x/y, 1)$  to prevent unbounded feature values, which can result in poor generalization. Notice that for  $w = 1$ ,  $\pi_{t|s}^w$  is bigram conditional probability,  $\pi_{t|s}^w \cdot \pi_s^w$  is bigram joint probability, the log of  $\pi_{t|s}^w / \pi_t^w$  is bigram pointwise mutual information, and  $\pi_{t|s}^w / \pi_{t|s}^{-w}$  measures how much more prevalent  $t$  is to the right of  $s$  than to the left. By also allowing other values of  $w$ , we generalize these features. Finally, our model also uses versions of these features for each  $b \in 1, 2$ .

**Neural features.** As an alternative to the manually designed  $\pi$  function above, we consider a neural approach to detect predictive configurations in the sentences of  $\mathbf{u}$ , potentially including complex long-distance configurations. Linguists working with Principles & Parameters theory have supposed that a single telltale sentence—a **trigger**—may be enough

<sup>7</sup>In the “fraction of tags” features,  $g(t | j)$  is undefined ( $\frac{0}{0}$ ) when  $\hat{W}_j$  is empty. We omit undefined values from the means.

<sup>8</sup>The reason we don’t include  $\pi_{t|s}^{-w} / \pi_{t|s}^w$  is that it is included when computing features for  $-w$ .

to determine a typological parameter, at least given the settings of other parameters (Gibson and Wexler, 1994; Frank and Kapur, 1996).

We map each corpus sentence  $u_i$  to a finite-dimensional real vector  $\mathbf{f}_i$  by using a gated recurrent unit (GRU) network (Cho et al., 2014), a type of recurrent neural network that is a simplified variant of an LSTM network (Hochreiter and Schmidhuber, 1997). The GRU reads the sequence of one-hot embeddings of the tags in  $u_i$  (including the boundary symbols #). We omit the part of the GRU that computes an output sequence, simply taking  $\mathbf{f}_i$  to be the final hidden state vector. The parameters are trained jointly with the rest of our typology prediction system, so the training procedure attempts to discover predictively useful configurations.

The various elements of  $\mathbf{f}_i$  attempt to detect various interesting configurations in sentence  $u_i$ . Some might be triggers (which call for max-pooling over sentences); others might provide softer evidence (which calls for mean-pooling). For generality, therefore, we define our feature vector  $\boldsymbol{\pi}(\mathbf{u})$  by **soft-pooling** of the sentence vectors  $\mathbf{f}_i$  (Figure 2). The tanh gate in the GRU implies  $f_{ik} \in (-1, 1)$  and we transform this to the positive quantity  $f'_{ik} = \frac{f_{ik}+1}{2} \in (0, 1)$ . Given an “inverse temperature”  $\beta$ , define<sup>9</sup>

$$\pi_k^\beta = \left( \text{mean}_i (f'_{ik})^\beta \right)^{1/\beta} \quad (10)$$

This  $\pi_k^\beta$  is a pooled version of  $f'_{ik}$ , ranging from max-pooling as  $\beta \rightarrow -\infty$  (i.e., does  $f'_{ik}$  fire strongly on any sentence  $i$ ?) to min-pooling as  $\beta \rightarrow \infty$ . It passes through arithmetic mean at  $\beta = 1$  (i.e., how strongly does  $f'_{ik}$  fire on the average sentence  $i$ ?), geometric mean as  $\beta \rightarrow 0$  (this may be regarded as an arithmetic mean in log space), and harmonic mean at  $\beta = -1$  (an arithmetic mean in reciprocal space).

Our final  $\boldsymbol{\pi}$  is a concatenation of the  $\pi^\beta$  vectors for  $\beta \in \{-4, -2, -1, 0, 1, 2, 4\}$ . We chose these  $\beta$  values experimentally, using cross-validation.

**Combined model.** We also consider a model

$$\mathbf{s}(\mathbf{u}) = \alpha \mathbf{s}_H(\mathbf{u}) + (1 - \alpha) \mathbf{s}_N(\mathbf{u}) \quad (11)$$

where  $\mathbf{s}_H(\mathbf{u})$  is the score assigned by the hand-feature system,  $\mathbf{s}_N(\mathbf{u})$  is the score assigned by the

<sup>9</sup>For efficiency, we restrict the mean to  $i \leq 1e4$  (the first 10,000 sentences).

neural-feature system, and  $\alpha \in [0, 1]$  is a hyperparameter to balance the two.  $\mathbf{s}_H(\mathbf{u})$  and  $\mathbf{s}_N(\mathbf{u})$  were trained separately. At test time, we use (11) to combine them linearly before the logistic transform (6). This yields a weighted-product-of-experts model.

#### 6.4 Training procedure

**Length thresholding.** By default, our feature vector  $\boldsymbol{\pi}(\mathbf{u})$  is extracted from those sentences in  $\mathbf{u}$  with length  $\leq 40$  tokens. In §7.3, however, we try concatenating this feature vector with one that is extracted in the same way from just sentences with length  $\leq 10$ . The intuition (Spitkovsky et al., 2010) is that the basic word order of the language can be most easily discerned from short, simple sentences.

**Initialization.** We initialize the model of (6)–(7) so that the estimated directionality  $\hat{p}(\rightarrow | r, L)$ , regardless of  $L$ , is initially a weighted mean of  $r$ ’s directionalities in the training languages, namely

$$\bar{p}_r \equiv \sum_L w_L(r) p^*(\rightarrow | r, L) \quad (12)$$

$$\text{where } w_L(r) \equiv \frac{p^*(r|L)}{\sum_{L'} p^*(r|L')} \quad (13)$$

This is done by setting  $V = 0$  and the bias  $(\mathbf{b}_V)_r = \log \frac{\bar{p}_r}{1-\bar{p}_r}$ , clipped to the range  $[-10, 10]$ . As a result, we make sensible initial predictions even for rare relations  $r$ , which allows us to converge reasonably quickly even though we do not update the parameters for rare relations as often.

We initialize the recurrent connections in the GRU to random orthogonal matrices. All other weight matrices in Figure 1 and the GRU use “Xavier initialization” (Glorot and Bengio, 2010). All other bias weight vectors are initialized to 0.

**Regularization.** We add an L2 regularizer to the objective. When training the neural network, we use dropout as well. All hyperparameters (regularization coefficient, dropout rate, etc.) are tuned via cross-validation; see §7.5.

**Optimization.** We use different algorithms in different feature settings. With scoring functions that use only hand features, we adjust the feature weights by stochastic gradient descent (SGD). With scoring functions that include neural features, we use RMSProp (Tieleman and Hinton, 2012).

Train		Test
cs, es, fr, hi,	en, nl, da, fi,	la, hr, ga, he, hu,
de, it, la_itt,	got, grc, et,	fa, ta, cu, el, ro,
no, ar, pt	la_proiel,	sl, ja_ktc, sv,
	grc_proiel, bg	fi_ftb, id, eu, pl

Table 2: Data split of the 37 real languages, adapted from Wang and Eisner (2016). (Our “Train,” on which we do 5-fold cross-validation, contains both their “Train” and “Dev” languages.)

## 7 Experiments

### 7.1 Data splits

We hold out 17 UD languages for testing (Table 2). For training, we use the remaining 20 UD languages and tune the hyperparameters with 5-fold cross-validation. That is, for each fold, we train the system on 16 real languages and evaluate on the remaining 4. When augmenting the 16 real languages with GD languages, we include only GD languages that are generated by “mixing-and-matching” those 16 languages, which means that we add  $16 \times 17 \times 17 = 4624$  synthetic languages.<sup>10</sup>

Each GD treebank  $\mathbf{u}$  provides a standard split into train/dev/test portions. In this paper, we primarily restrict ourselves to the train portions (saving the gold trees from the dev and test portions to tune and evaluate some future grammar induction system that consults our typological predictions). For example, we write  $\mathbf{u}_{\text{train}}$  for the POS-tagged sentences in the “train” portion, and  $\mathbf{p}_{\text{train}}^*$  for the empirical probabilities derived from their gold trees. We always train the model to predict  $\mathbf{p}_{\text{train}}^*$  from  $\mathbf{u}_{\text{train}}$  on each *training language*. To evaluate on a *held-out language* during cross-validation, we can measure how well the model predicts  $\mathbf{p}_{\text{train}}^*$  given  $\mathbf{u}_{\text{train}}$ .<sup>11</sup> For our fi-

<sup>10</sup>Why  $16 \times 17 \times 17$ ? As Wang and Eisner (2016, §5) explain, each GD treebank is obtained from the UD treebank of some **substrate** language  $S$  by stochastically permuting the dependents of verbs and nouns to respect typical orders in the **superstrate** languages  $R_V$  and  $R_N$  respectively. There are 16 choices for  $S$ . There are 17 choices for  $R_V$  (respectively  $R_N$ ), including  $R_V = S$  (“self-permutation”) and  $R_V = \emptyset$  (“no permutation”).

<sup>11</sup>In actuality, we measured how well it predicts  $\mathbf{p}_{\text{dev}}^*$  given  $\mathbf{u}_{\text{dev}}$ . That was a slightly less sensible choice. It may have harmed our choice of hyperparameters, since dev is smaller than train and therefore  $\mathbf{p}_{\text{dev}}^*$  tends to have greater sampling error. Another concern is that our typology system, having been specifically tuned to predict  $\mathbf{p}_{\text{dev}}^*$ , might provide an unrealistically accurate estimate of  $\mathbf{p}_{\text{dev}}^*$  to some future grammar induction system that is being cross-validated against the same dev set, harming that system’s choice of hyperparameters as well.

Architecture		$\varepsilon$ -insensitive loss	
Scoring	Depth	UD	+GD
EC	-	0.104	0.099
$s_H$	0	0.057	<b>0.037*</b>
$s_H$	1	<b>0.050</b>	<b>0.036*</b>
$s_H$	3	0.060	0.048
$s_N$	1	<b>0.062</b>	0.048
$\alpha s_H + (1 - \alpha) s_N$	1	<b>0.050</b>	<b>0.032*</b>

Table 3: Average expected loss over 20 UD languages, computed by 5-fold cross-validation. The first column indicates whether we score using hand-engineered features ( $s_H$ ), neural features ( $s_N$ ), or a combination (see end of §6.3). As a baseline, the first line evaluates the EC (expected count) heuristic from §5. Within each *column*, we boldface the best (smallest) result as well as all results that are not significantly worse (paired permutation test by language,  $p < 0.05$ ). A starred result is significantly better than the other model in the same *row*.

nal test, we evaluate on the 17 test languages using a model trained on all training languages (20 treebanks for UD, plus  $20 \times 21 \times 21 = 8840$  when adding GD) with the chosen hyperparameters. To evaluate on a *test language*, we again measure how well the model predicts  $\mathbf{p}_{\text{train}}^*$  from  $\mathbf{u}_{\text{train}}$ .

### 7.2 Comparison of architectures

Table 3 shows the cross-validation losses (equation (1)) that are achieved by different scoring architectures. We compare the results when the model is trained on real languages (the “UD” column) versus on real languages plus synthetic languages (the “+GD” column).

The  $s_H$  models here use a subset of the hand-engineered features, selected by the experiments in §7.3 below and corresponding to Table 4 line 8.

Although Figure 1 and equation (7) presented an “depth-1” scoring network with one hidden layer, Table 3 also evaluates “depth- $d$ ” architectures with  $d$  hidden layers. The depth-0 architecture simply predicts each directionality separately using logistic regression (although our training objective is not the usual convex log-likelihood objective).

Some architectures are better than others. We note that the hand-engineered features outperform the neural features—though not significantly, since they make complementary errors—and that combining them is best. However, the biggest benefit comes from augmenting the training data with GD languages; this consistently helps more than changing the architecture.

ID	Features	Length	Loss (+GD)
0	$\emptyset$	—	0.076
1	conditional	40	0.058
2	joint	40	0.057
3	PMI	40	0.039
4	asymmetry	40	0.041
5	rows 3+4	40	0.038
6	row 5+b	40	<b>0.037</b>
7	row 5+t	40	<b>0.037</b>
8*	row 5+b+t	40	<b>0.036</b>
9	row 8	10	0.043
10	row 8	10+40	<b>0.036</b>

Table 4: Cross-validation losses with different subsets of hand-engineered features from §6.3. “ $\emptyset$ ” is a baseline with no features (bias feature only), so it makes the same prediction for all languages. “conditional” =  $\pi_{t|s}^w$  features, “joint” =  $\pi_{t|s}^w \cdot \pi_s^w$  features, “PMI” =  $\pi_{t|s}^w / \pi_t^w$  and  $\pi_t^w / \pi_{t|s}^w$  features, “asymmetry” =  $\pi_{t|s}^w / \pi_{t|s}^{-w}$  features, “b” are the features superscripted by *b*, and “t” are the features with truncated window. “+” means concatenation. The “Length” field refers to length thresholding (see §6.4). The system in the starred row is the one that we selected for row 2 of Table 3.

### 7.3 Contribution of different feature classes

To understand the contribution of different hand-engineered features, we performed forward selection tests on the depth-1 system, including only some of the features. In all cases, we trained in the “+GD” condition. The results are shown in Table 4. Any class of features is substantially better than baseline, but we observe that most of the total benefit can be obtained with just PMI or asymmetry features. Those features indicate, for example, whether a verb tends to attract nouns to its right or left. We did not see a gain from length thresholding.

### 7.4 Robustness to noisy input

We also tested our directionality prediction system on noisy input (without retraining it on noisy input). Specifically, we tested the depth-1  $s_H$  system. This time, when evaluating on the dev split of a held-out language, we provided a noisy version of that input corpus that had been retagged by an automatic POS tagger (Nguyen et al., 2014), which was trained on just 100 gold-tagged sentences from the train split of that language. The average tagging accuracy over the 20 languages was only 77.26%. Nonetheless, the “UD”-trained and “+GD”-trained systems got respective losses of 0.052 and 0.041—nearly as good as in Table 3, which used gold POS tags.

	MS13	N10	EC	$\emptyset$	UD	+GD
loss	0.166	0.139	0.098	0.083	0.080	<b>0.039</b>

Table 5: Cross-validation average expected loss of the two grammar induction methods, MS13 (Mareček and Straka, 2013) and N10 (Naseem et al., 2010), compared to the EC heuristic of §5 and our architecture of §6 (the version from the last line of Table 3). In these experiments, the dependency relation types are ordered POS pairs. N10 harnesses prior linguistic knowledge, but its improvement upon MS13 is not statistically significant. Both grammar induction systems are *significantly* worse than the rest of the systems, including even our two baseline systems, namely EC (the “expected count” heuristic from §5) and  $\emptyset$  (the no-feature baseline system from Table 4 line 0). Like N10, these baselines make use of some cross-linguistic knowledge, which they extract in different ways from the training treebanks. Among our own 4 systems, EC is significantly worse than all others, and +GD is significantly better than all others. (Note: When training the *baselines*, we found that including the +GD languages—a bias-variance tradeoff—harmed EC but helped  $\emptyset$ . The table reports the better result in each case.)

### 7.5 Hyperparameter settings

For each result in Tables 3–4, the hyperparameters were chosen by grid search on the cross-validation objective (and the table reports the best result). For the remaining experiments, we select the depth-1 combined models (11) for both “UD” and “+GD,” as they are the best models according to Table 3.

The hyperparameters for the selected models are as follows: When training with “UD,” we took  $\alpha = 1$  (which ignores  $s_N$ ), with hidden layer size  $h = 256$ ,  $\psi = \text{sigmoid}$ ,  $\text{L2\_coeff} = 0$  (no L2 regularization), and  $\text{dropout} = 0.2$ . When training with “+GD,” we took  $\alpha = 0.7$ , with different hyperparameters for the two interpolated models:  $s_H$  uses  $h = 128$ ,  $\psi = \text{sigmoid}$ ,  $\text{L2\_coeff} = 0$ , and  $\text{dropout} = 0.4$ , while  $s_N$  uses  $h = 128$ ,  $\text{emb\_size} = 128$ ,  $\text{rnn\_size} = 32$ ,  $\psi = \text{relu}$ ,  $\text{L2\_coeff} = 0$ , and  $\text{dropout} = 0.2$ . For both “UD” and “+GD,” we use  $\lambda = 1$  for the smoothing in footnote 6.

### 7.6 Comparison with grammar induction

Grammar induction is an alternative way to predict word order typology. Given a corpus of a language, we can first use grammar induction to parse it into dependency trees, and then estimate the directionality of each dependency relation type based on these (approximate) trees.

However, what are the dependency relation types? Current grammar induction systems produce unlabeled dependency edges. Rather than try to obtain



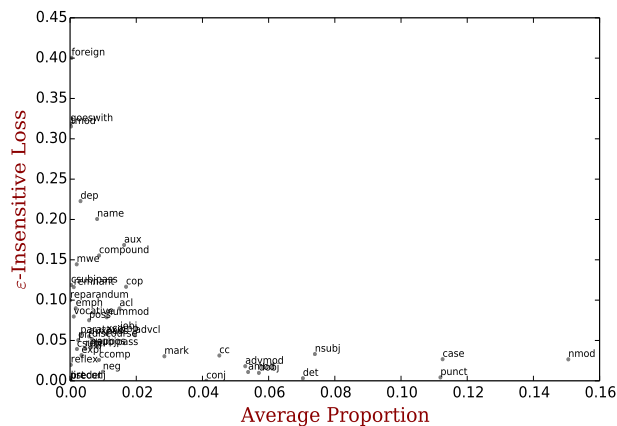


Figure 3: Cross-validation loss broken down by relation. We plot each relation  $r$  with  $x$  coordinate = the proportion of  $r$  in the average training corpus =  $\text{mean}_{L \in \text{Train}} p_{\text{train}}^*(r | L) \in [0, 1]$ , and with  $y$  coordinate = the weighted average  $\sum_{L \in \text{Heldout}} w_L(r) \text{loss}_\varepsilon(\hat{p}_{\text{dev}}(\rightarrow|r, L), p_{\text{dev}}^*(\rightarrow|r, L))$  (see (13)).

a UD label like  $r = \text{amod}$  for each edge, we label the edge deterministically with a POS pair such as  $r = (\text{parent} = \text{NOUN}, \text{child} = \text{ADJ})$ . Thus, we will attempt to predict the directionality of each POS-pair relation type. For comparison, we retrain our supervised system to do the same thing.

For the grammar induction system, we try the implementation of DMV with stop-probability estimation by Mareček and Straka (2013), which is a common baseline for grammar induction (Le and Zuidema, 2015) because it is language-independent, reasonably accurate, fast, and convenient to use. We also try the grammar induction system of Naseem et al. (2010), which is the state-of-the-art system on UD (Noji et al., 2016). Naseem et al. (2010)’s method, like ours, has prior knowledge of what typical human languages look like.

Table 5 shows the results. Compared to Mareček and Straka (2013), Naseem et al. (2010) gets only a small (insignificant) improvement—whereas our “UD” system halves the loss, and the “+GD” system halves it again. Even our baseline systems are significantly more accurate than the grammar induction systems, showing the effectiveness of casting the problem as supervised prediction.

## 7.7 Fine-grained analysis

Beyond reporting the aggregate cross-validation loss over the 20 training languages, we break down the cross-validation predictions by relation type. Figure 3 shows that the *frequent* relations are all quite

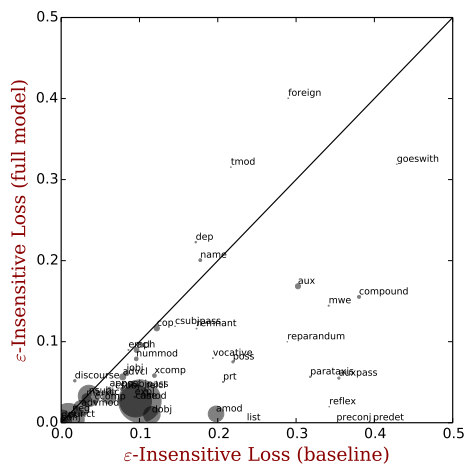


Figure 4: The  $y$  coordinate is the average loss of our model (Table 4 line 8), just as in Figure 3, whereas the  $x$  coordinate is the average loss of a simple baseline model  $\emptyset$  that ignores the input corpus (Table 4 line 0). Relations whose directionality varies more by language have higher baseline loss. Relations that beat the baseline fall below the diagonal line. The marker size for each relation is proportional to the  $x$ -axis in Figure 3.

predictable. Figure 4 shows that our success is not just because the task is easy—on relations whose directionality varies by language, so that a baseline method does poorly, our system usually does well.

To show that our system is behaving well across languages and not just on average, we zoom in on 5 relation types that are particularly common or of particular interest to linguistic typologists. These 5 relations together account for 46% of all relation tokens in the average language:  $\text{nmod} = \text{noun-nominal modifier order}$ ,  $\text{nsubj} = \text{subject-verb order}$  (feature 82A in the World Atlas of Language Structures),  $\text{dobj} = \text{object-verb order}$  (83A),  $\text{amod} = \text{adjective-noun order}$  (87A), and  $\text{case} = \text{placement of both adpositions and case markers}$  (85A).

As shown in Figure 5, most points in the first five plots fall in or quite near the desired region. We are pleased to see that the predictions are robust when the training data is unbalanced. For example, the  $\text{case}$  relation points leftward in most real languages, yet our system can still predict the right directionality of “hi”, “et” and “fi.” The credit goes to the diversity of our training set, which contains various synthetic  $\text{case-right}$  languages: the system fails on these three languages if we train on real languages only. That said, apparently our training set is still not diverse enough to do well on the outlier “ar” (Arabic); see Figure 4 in Wang and Eisner (2016).

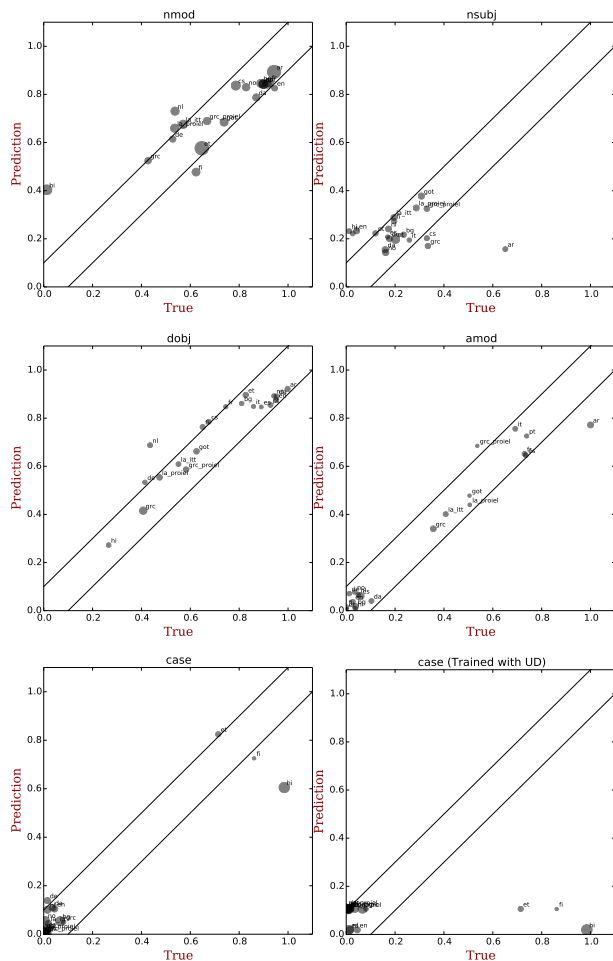


Figure 5: Scatterplots of predicted vs. true directionalities (by cross-validation). In the plot for relation type  $r$ , each language appears as a marker at  $(p^*, \hat{p})$  (see §4), with the marker size proportional to  $w_L(r)$  (see (13)). Points that fall between the solid lines ( $|\hat{p} - p^*| \leq \varepsilon$ ) are considered “correct,” by the definition of  $\varepsilon$ -insensitive loss. The last plot (bottom right) shows worse predictions for `case` when the model is trained on UD only.

## 7.8 Binary classification accuracy

Besides  $\varepsilon$ -insensitive loss, we also measured how the systems perform on the coarser task of binary classification of relation direction. We say that relation  $r$  is dominantly “rightward” in language  $L$  iff  $p^*(\rightarrow | r, L) > 0.5$ . We say that a system predicts “rightward” according to whether  $\hat{p}(\rightarrow | r, L) > 0.5$ .

We evaluate whether this binary prediction is correct for each of the 20 most frequent relations  $r$ , for each held-out language  $L$ , using 5-fold cross-validation over the 20 training languages  $L$  as in the previous experiment. Tables 6 and 7 respectively summarize these results by relation (equal average

Relation	Rate	EC	$\emptyset$	UD	+GD
nmod	0.15	0.85	0.9	0.9	0.9
punct	0.11	0.85	0.85	0.85	0.85
case	0.11	0.75	0.85	0.85	1
nsubj	0.08	0.95	0.95	0.95	0.95
det	0.07	0.8	0.9	0.9	0.9
dobj	0.06	0.6	0.75	0.75	0.85
amod	0.05	0.6	0.6	0.75	0.9
advmod	0.05	0.9	0.85	0.85	0.85
cc	0.04	0.95	0.95	0.95	0.95
conj	0.04	1	1	1	1
mark	0.03	0.95	0.95	0.95	0.95
advcl	0.02	0.85	0.85	0.85	0.8
cop	0.02	0.75	0.75	0.65	0.75
aux	0.02	0.9	0.6	0.75	0.65
iobj	0.02	0.45	0.55	0.5	0.6
acl	0.01	0.45	0.85	0.85	0.8
nummod	0.01	0.9	0.9	0.9	0.9
xcomp	0.01	0.95	0.95	0.95	1
neg	0.01	1	1	1	1
ccomp	0.01	0.75	0.95	0.95	0.95
Avg.	-	0.81	0.8475	0.855	0.8775

Table 6: Accuracy on the simpler task of binary classification of relation directionality. The most common relations are shown first: the “Rate” column gives the average rate of the relation across the 20 training languages (like the  $x$  coordinate in Fig. 3).

over languages) and by language (equal average over relations). Keep in mind that these systems had not been specifically trained to place relations on the correct side of the artificial 0.5 boundary.

Binary classification is an easier task. It is easy because, as the  $\emptyset$  column in Table 6 indicates, most relations have a clear directionality preference shared by most of the UD languages. As a result, the better models with more features have less opportunity to help. Nonetheless, they do perform better, and the EC heuristic continues to perform worse.

In particular, EC fails significantly on `dobj` and `iobj`. This is because `nsubj`, `dobj`, and `iobj` often have different directionalities (e.g., in SVO languages), but the EC heuristic will tend to predict the same direction for all of them, according to whether NOUNs tend to precede nearby VERBs.

## 7.9 Final evaluation on test data

All previous experiments were conducted by cross-validation on the 20 training languages. We now train the system on all 20, and report results on the 17 blind test languages (Table 8). In our evaluation metric (1),  $\mathcal{R}$  includes all 57 relation types that appear in training data, plus a special UNK type for

target	EC	∅	UD	+GD
ar	0.8	0.8	0.75	0.85
bg	0.85	0.95	0.95	0.95
cs	0.9	1	1	0.95
da	0.8	0.95	0.95	0.95
de	0.9	0.9	0.9	0.95
en	0.9	1	1	1
es	0.9	0.9	0.95	0.95
et	0.8	0.8	0.8	0.8
fi	0.75	0.85	0.85	0.85
fr	0.9	0.9	0.9	0.95
got	0.75	0.8	0.85	0.8
grc	0.6	0.7	0.7	0.75
grc_proiel	0.8	0.8	0.85	0.9
hi	0.6	0.45	0.45	0.7
it	0.9	0.9	0.9	0.95
la_itt	0.7	0.85	0.8	0.85
la_proiel	0.7	0.7	0.75	0.7
nl	0.95	0.85	0.85	0.85
no	0.9	1	1	0.95
pt	0.8	0.85	0.9	0.9
Avg.	0.81	0.8475	0.855	0.8775

Table 7: Accuracy on the simpler task of binary classification of relation directionality for each training language. A detailed comparison shows that EC is *significantly* worse than UD and +GD, and that ∅ is *significantly* worse than +GD (paired permutation test over the 20 languages,  $p < 0.05$ ). The improvement from UD to +GD is *insignificant*, which suggests that this is an easier task where weak models might suffice.

relations that appear only in test data. The results range from good to excellent, with synthetic data providing consistent and often large improvements.

These results could potentially be boosted in the future by using an even larger and more diverse training set. In principle, when evaluating on any one of our 37 real languages, one could train a system on *all* of the other 36 (plus the galactic languages derived from them), not just 20. Moreover, the Universal Dependencies collection has continued to grow beyond the 37 languages used here (§3). Finally, our current setup extracts only one training example from each (real or synthetic) language. One could easily generate a variant of this example each time the language is visited during stochastic optimization, by bootstrap-resampling its training corpus (to add “natural” variation) or subsampling it (to train the predictor to work on smaller corpora). In the case of a synthetic language, one could also generate a corpus of new trees each time the language is visited (by re-running the stochastic permutation procedure, instead of reusing the particular permutation released by the Galactic Dependencies project).

Test			Train		
target	UD	+GD	target	UD	+GD
cu	0.024	0.024	ar	0.116	0.057
el	0.056	0.011	bg	0.037	0.015
eu	0.250	0.072	cs	0.025	0.014
fa	0.220	0.134	da	0.024	0.017
fi_ftb	0.073	0.029	de	0.046	0.025
ga	0.181	0.154	en	0.025	0.036
he	0.079	0.033	es	0.012	0.007
hr	0.062	0.011	et	0.055	0.014
hu	0.119	0.102	fi	0.069	0.070
id	0.099	0.076	fr	0.024	0.018
ja_ktc	0.247	0.078	got	0.008	0.026
la	0.036	0.004	grc	0.026	0.007
pl	0.056	0.023	grc_proiel	0.004	0.017
ro	0.029	0.009	hi	0.363	0.191
sl	0.015	0.031	it	0.011	0.008
sv	0.012	0.008	la_itt	0.033	0.023
ta	0.238	0.053	la_proiel	0.018	0.021
			nl	0.069	0.066
			no	0.008	0.010
			pt	0.038	0.004
Test Avg.	0.106	0.050*	All Avg.	0.076	0.040*

Table 8: Our final comparison on the 17 test languages appears at left. We ask whether the average expected loss on these 17 real target languages is reduced by augmenting the training pool of 20 UD languages with +20\*21\*21 GD languages. For completeness, we extend the table with the cross-validation results on the training pool. The “Avg.” lines report the average of 17 test or 37 training+testing languages. We mark both “+GD” averages with “\*” as they are significantly better than their “UD” counterparts (paired permutation test by language,  $p < 0.05$ ).

## 8 Related Work

Typological properties can usefully boost the performance of cross-linguistic systems (Bender, 2009; O’Horan et al., 2016). These systems mainly aim to annotate low-resource languages with help from models trained on similar high-resource languages. Naseem et al. (2012) introduce a “selective sharing” technique for generative parsing, in which a Subject-Verb language will use parameters shared with other Subject-Verb languages. Täckström et al. (2013) and Zhang and Barzilay (2015) extend this idea to discriminative parsing and gain further improvements by conjoining regular parsing features with typological features. The cross-linguistic neural parser of Ammar et al. (2016) conditions on typological features by supplying a “language embedding” as input. Zhang et al. (2012) use typological properties to convert language-specific POS tags to UD POS tags, based on their ordering in a corpus.

Moving from engineering to science, lin-

guists seek *typological universals* of human language (Greenberg, 1963; Croft, 2002; Song, 2014; Hawkins, 2014), e.g., “languages with dominant Verb-Subject-Object order are always prepositional.” Dryer and Haspelmath (2013) characterize 2679 world languages with 192 typological properties. Their WALS database can supply features to NLP systems (see previous paragraph) or gold standard labels for typological classifiers. Daumé III and Campbell (2007) take WALS as input and propose a Bayesian approach to discover new universals. Georgi et al. (2010) impute missing properties of a language, not by using universals, but by backing off to the language’s typological cluster. Murawaki (2015) use WALS to help recover the evolutionary tree of human languages; Daumé III (2009) considers the geographic distribution of WALS properties.

Attempts at automatic typological classification are relatively recent. Lewis and Xia (2008) predict typological properties from induced trees, but guess those trees from aligned bitexts, not by monolingual grammar induction as in §7.6. Liu (2010) and Futrell et al. (2015) show that the directionality of (gold) dependencies is indicative of “basic” word order and freeness of word order. Those papers predict typological properties from trees that are automatically (noisily) annotated or manually (expensively) annotated. An alternative is to predict the typology directly from raw or POS-tagged text, as we do. Saha Roy et al. (2014) first explored this idea, building a system that correctly predicts adposition typology on 19/23 languages with only word co-occurrence statistics. Zhang et al. (2016) evaluate semi-supervised POS tagging by asking whether the induced tag sequences can predict typological properties. Their prediction approach is supervised like ours, although developed separately and trained on different data. They more simply predict 6 binary-valued WALS properties, using 6 independent binary classifiers based on POS bigram and trigrams.

Our task is rather close to grammar induction, which likewise predicts a set of real numbers giving the relative probabilities of competing syntactic configurations. Most previous work on grammar induction begins with maximum likelihood estimation of some generative model—such as a PCFG (Lari and Young, 1990; Carroll and Charniak, 1992) or dependency grammar (Klein and Manning, 2004)—

though it may add linguistically-informed inductive bias (Ganchev et al., 2010; Naseem et al., 2010). Most such methods use local search and must wrestle with local optima (Spitkovsky et al., 2013). Fine-grained typological classification might supplement this approach, by cutting through the initial combinatorial challenge of establishing the basic word-order properties of the language. In this paper we only quantify the directionality of each relation *type*, ignoring how *tokens* of these relations interact locally to give coherent parse trees. Grammar induction methods like EM could naturally consider those local interactions for a more refined analysis, when guided by our predicted global directionalities.

## 9 Conclusions and Future Work

We introduced a typological classification task, which attempts to extract quantitative knowledge about a language’s syntactic structure from its surface forms (POS tag sequences). We applied supervised learning to this apparently unsupervised problem. As far as we know, we are the first to utilize synthetic languages to train a learner for real languages: this move yielded substantial benefits.<sup>12</sup>

Figure 5 shows that we rank held-out languages rather accurately along a spectrum of directionality, for several common dependency relations. Table 8 shows that if we jointly predict the directionalities of *all* the relations in a new language, most of those numbers will be quite close to the truth (low aggregate error, weighted by relation frequency). This holds promise for aiding grammar induction.

Our trained model is robust when applied to noisy POS tag sequences. In the future, however, we would like to make similar predictions from raw word sequences. That will require features that abstract away from the language-specific vocabulary. Although recurrent neural networks in the present paper did not show a clear advantage over hand-engineered features, they might be useful when used with word embeddings.

Finally, we are interested in downstream uses. Several NLP tasks have benefited from typological features (§8). By using end-to-end training, our methods could be tuned to extract features (existing or novel) that are particularly useful for some task.

<sup>12</sup>Although Wang and Eisner (2016) review uses of synthetic training data elsewhere in machine learning.

**Acknowledgements** This work was funded by the U.S. National Science Foundation under Grant No. 1423276. We are grateful to the state of Maryland for providing indispensable computing resources via the Maryland Advanced Research Computing Center (MARCC). We thank the Argo lab members for useful discussions. Finally, we thank TACL action editor Mark Steedman and the anonymous reviewers for high-quality suggestions, including the EC baseline and the binary classification evaluation.

## References

- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. Many languages, one parser. *Transactions of the Association of Computational Linguistics*, 4:431–444, 2016.
- Emily M. Bender. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, 2009.
- Glenn Carroll and Eugene Charniak. Two experiments on learning probabilistic dependency grammars from corpora. In *Working Notes of the AAAI Workshop on Statistically-Based NLP Techniques*, pages 1–13, 1992.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.
- Noam Chomsky. *Lectures on Government and Binding: The Pisa Lectures*. Holland: Foris Publications, 1981.
- Noam Chomsky and Howard Lasnik. The theory of principles and parameters. In *Syntax: An International Handbook of Contemporary Research*, Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld, and Theo Vennemann, editors. Berlin: de Gruyter, 1993.
- William Croft. *Typology and Universals*. Cambridge University Press, 2002.
- Hal Daumé III. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601, 2009.
- Hal Daumé III and Lyle Campbell. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, 2007.
- Matthew S. Dryer and Martin Haspelmath, editors. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. <http://wals.info/>.
- Joakim Nivre, et al. Universal Dependencies 1.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague. Data available at <http://universaldependencies.org>, 2015.
- Robert Frank and Shyam Kapur. On the use of triggers in parameter setting. *Linguistic Inquiry*, 27: 623–660, 1996.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics*, pages 91–100, 2015.
- Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010.
- Ryan Georgi, Fei Xia, and William Lewis. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393, 2010.
- Edward Gibson and Kenneth Wexler. Triggers. *Linguistic Inquiry*, 25(3):407–454, 1994.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010.
- Joseph H. Greenberg. Some universals of grammar with particular reference to the order of mean-

- ingful elements. In *Universals of Language*, Joseph H. Greenberg, editor, pages 73–113. MIT Press, 1963.
- John A. Hawkins. *Word Order Universals*. Elsevier, 2014.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.
- Dan Klein and Christopher Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 478–485, 2004.
- Karim Lari and Steve J. Young. The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, 4(1):35–56, 1990.
- Phong Le and Willem Zuidema. Unsupervised dependency parsing: Let’s use supervised parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 651–661, 2015.
- William D. Lewis and Fei Xia. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.
- Haitao Liu. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578, 2010.
- David Mareček and Milan Straka. Stop-probability estimates computed on a large corpus improve unsupervised dependency parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 281–290, 2013.
- Yugo Murawaki. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334, 2015.
- Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, 2010.
- Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, 2012.
- Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20, 2014.
- Hiroshi Noji, Yusuke Miyao, and Mark Johnson. Using left-corner parsing to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43, 2016.
- Helen O’Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. Survey on the use of typological information in natural language processing. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, 2016.
- Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly, and Monojit Choudhury. Automatic discovery of adposition typology. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1037–1046, 2014.
- Jae Jung Song. *Linguistic Typology: Morphology and Syntax*. Routledge, 2014.
- Valentin I. Spitzkovsky, Hiyam Alshawi, and Daniel Jurafsky. From baby steps to leapfrog: How “less is more” in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759, 2010.

- Valentin I. Spitzkovsky, Hiyan Alshawi, and Daniel Jurafsky. Breaking out of local optima with count transforms and model recombination: A study in grammar induction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1983–1995, 2013.
- Oscar Täckström, Ryan McDonald, and Joakim Nivre. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, 2013.
- Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: *Neural Networks for Machine Learning*, 2012.
- Dingquan Wang and Jason Eisner. The Galactic Dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association of Computational Linguistics*, 4: 491–505, 2016. Data available at <https://github.com/gdtreebank/gdtreebank>.
- Yuan Zhang and Regina Barzilay. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867, 2015.
- Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. Learning to map into a universal POS tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1368–1378, 2012.
- Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. Ten pairs to tag—multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, 2016.

