Improving Low-Resource Cross-lingual Parsing with Expected Statistic Regularization

Thomas Effland
Columbia University, USA
teffland@cs.columbia.edu

Michael Collins
Google Research, USA
mjcollins@google.com

Abstract

We present Expected Statistic Regularization (ESR), a novel regularization technique that utilizes low-order multi-task structural statistics to shape model distributions for semi-supervised learning on low-resource datasets. We study ESR in the context of cross-lingual transfer for syntactic analysis (POS tagging and labeled dependency parsing) and present several classes of low-order statistic functions that bear on model behavior. Experimentally, we evaluate the proposed statistics with ESR for unsupervised transfer on 5 diverse target languages and show that all statistics, when estimated accurately, yield improvements to both POS and LAS, with the best statistic improving POS by +7.0 and LAS by +8.5 on average. We also present semi-supervised transfer and learning curve experiments that show ESR provides significant gains over strong cross-lingual-transfer-plus-fine-tuning baselines for modest amounts of label data. These results indicate that ESR is a promising and complementary approach to model-transfer approaches for cross-lingual parsing.[1]

1 Introduction

In recent years, great strides have been made on linguistic analysis for low-resource languages. These gains are largely attributable to transfer approaches from (1) massive pretrained multilingual language model (PLM) encoders (Devlin et al., 2019; Liu et al., 2019b); (2) multi-task training across related syntactic analysis tasks (Kondratyuk and Straka, 2019); and (3) multilingual training on diverse high-resource languages (Wu and Dredze, 2019; Ahmad et al., 2019; Kondratyuk and Straka, 2019). Combined, these approaches have been shown to be particularly effective for cross-lingual syntactic analysis, as shown by UDify (Kondratyuk and Straka, 2019).

However, even with the improvements brought about by these techniques, transferred models still make syntactically implausible predictions on low-resource languages, and these error rates increase dramatically as the target languages become more distant from the source languages (He et al., 2019; Meng et al., 2019). In particular, transferred models often fail to match many low-order statistics concerning the typology of the task structures. We hypothesize that enforcing regularity with respect to estimates of these structural statistics—effectively using them as weak supervision—is complementary to current transfer approaches for low-resource cross-lingual parsing.

To this end, we introduce Expected Statistic Regularization (ESR), a novel differentiable loss that regularizes models on unlabeled target datasets by minimizing deviation of descriptive statistics of model behavior from target values. The class of descriptive statistics usable by ESR are expressive and powerful. For example, they may describe cross-task interactions, encouraging the model to obey structural patterns that are not explicitly tractable in the model factorization. Additionally, the statistics may be derived from constraints dictated by the task formalism itself (such as ruling out invalid substructures) or by numerical parameters that are specific to the target dataset distribution (such as relative substructure frequencies). In the latter case, we also contribute a method for selecting those parameters using small amounts of labeled data, based on the bootstrap (Efron, 1979).

Although ESR is applicable to a variety of problems, we study it using modern cross-lingual syntactic analysis on the Universal Dependencies data, building off of the strong model-transfer framework of UDify (Kondratyuk and Straka, 2019). We show that ESR is complementary to transfer-based approaches for building parsers on low-resource languages. We present several interesting classes of statistics for the tasks and perform

[1] We have published for our implementation and experiments at https://github.com/teffland/expected-statistic-regularization.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 122–138, 2023. https://doi.org/10.1162/tacl_a_00537
Action Editor: Alexander Clark. Submission batch: 4/2022; Revision batch: 7/2022; Published 1/2023.

extensive experiments in both oracle unsupervised and realistic semi-supervised cross-lingual multi-task parsing scenarios, with particularly encouraging results that significantly outperform state-of-the-art approaches for semi-supervised scenarios. We also present ablations that justify key design choices.

2 Expected Statistic Regularization

We consider structured prediction in an abstract setting where we have inputs $x \in \mathcal{X}$, output structures $y \in \mathcal{Y}$, and a conditional model $p_\theta(y|x) \in \mathbb{P}$ with parameters $\theta \in \Theta$, where $\mathbb{P}$ is the distribution space and Θ is the parameter space. In this section we assume that the setting is semi-supervised, with a small labeled dataset $\mathcal{D}_L$ and a large unlabeled dataset $\mathcal{D}_U$; let $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{m}$ be the labeled dataset of size m and similarly define $\mathcal{D}_U = \{x_i\}_{i=m+1}^{m+n}$ as the unlabeled dataset.

Our approach centers around a vectorized statistic function f that maps unlabeled samples and models to real vectors of dimension d_f:

$$f : \mathbb{D} \times \mathbb{P} \to \mathbb{R}^{d_f} \tag{1}$$

where $\mathbb{D}$ is the set of unlabeled datasets of any size, (i.e., $\mathcal{D}_U \in \mathbb{D}$). The purpose of f is to summarize various properties of the model using the sample. For example, if the task is part-of-speech tagging, one possible component of f could be the expected proportion of NOUN tags in the unlabeled data $\mathcal{D}_U$. In addition to f, we assume that we are given vectors of target statistics $t \in \mathbb{R}^d$ and margins of uncertainty $\sigma \in \mathbb{R}^d$ as its supervision signal. We will discuss the details of f, t, and σ shortly but first describe the overall objective.

2.1 Semi-Supervised Objective

Given labeled and unlabeled data $\mathcal{D}_L$ and $\mathcal{D}_U$, we propose the following semi-supervised objective O, which breaks down into a sum of supervised and unsupervised terms L and C:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} O(\theta; \mathcal{D}_L, \mathcal{D}_U) \tag{2}$$

$$O(\theta; \mathcal{D}_L, \mathcal{D}_U) = L(\theta; \mathcal{D}_L) + \alpha C(\theta; \mathcal{D}_U) \tag{3}$$

where $\alpha > 0$ is a balancing coefficient. The supervised objective L can be any suitable supervised loss; here we will use the negative log-likelihood of the data under the model. Our contribution is the unsupervised objective C.

For C, we propose to minimize some distance function ℓ between the target statistics t and the value of the statistics f calculated using unlabeled data and the model p_θ. (ℓ will also take into account the uncertainty margins σ.) A simple objective would be:

$$C(\theta; \mathcal{D}_U) = \ell(t, \sigma, f(\mathcal{D}_U, p_\theta))$$

This is a dataset-level loss penalizing divergences from the target level statistics. The problem with this approach is that this is not amenable to modern hardware constraints requiring SGD. Instead, we propose to optimize this loss in expectation over unlabeled mini-batch samples $\mathcal{D}_U^k$, where k is the mini-batch size and $\mathcal{D}_U^k$ is sampled uniformly with replacement from $\mathcal{D}_U$. Then, C is given by:

$$C(\theta; \mathcal{D}_U) = \mathbb{E}_{\mathcal{D}_U^k}[\ell(t, \sigma, f(\mathcal{D}_U^k, p_\theta))] \tag{4}$$

This objective penalizes the model if the statistic f, when applied to samples of unlabeled data $\mathcal{D}_U^k$, deviates from the targets t and thus pushes the model toward satisfying these target statistics.

Importantly, the objective in Eq. 4 is more general than typical objectives in that the outer loss function ℓ does not necessarily break down into a sum over individual input examples—the aggregation over examples is done inside f:

$$\ell(t, \sigma, f(\mathcal{D}_U, p_\theta)) \neq \sum_{x \in \mathcal{D}_U} \ell(t, \sigma, f(x, p_\theta)) \tag{5}$$

This generality is useful because components of f may describe statistics that aggregate over inputs, estimating expected quantities concerning sample-level regularities of the structures. In contrast, the right-hand side of Eq. 5 is more stringent, imposing that the statistic be the same for all instances of x. In practice, this loss reduces noise compared to a per-sentence loss, as is shown in Section 5.3.1.

2.2 The Statistic Function f

In principle the vectorized statistic function f could be almost any function of the unlabeled data and model, provided it is possible to obtain its gradients with respect to the model parameters θ, however, in this work we will assume f has the following three-layer structure.

First, let g be another vectorized function of ''sub-statistics'' that may have a different dimensionality than f and takes individual x, y pairs as input:

$$g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d_g} \quad (6)$$

Then let $\bar{g}$ be the expected value of g under the model p_θ summed over the sample $\mathcal{D}_U$:

$$\bar{g} = \sum_{x \in \mathcal{D}_U} \mathbb{E}_{p_\theta(y|x)}[g(x, y)] \quad (7)$$

Given $\bar{g}$, let the f's j'th component be the result of an aggregating function $h_j : \mathbb{R}^{d_g} \rightarrow \mathbb{R}$ on $\bar{g}$:

$$f_j(\mathcal{D}_U, p_\theta) = h_j(\bar{g}) \quad (8)$$

The individual components g_i will mostly be counting functions that tally various substructures in the data. The $\bar{g}_i$'s then are expected substructure counts in the sample, and the h_j's aggregate small subsets of these intermediate counts in different ways to compute various marginal probabilities. Again, in general f does not need to follow this structure and any suitable statistic function can be incorporated into the regularization term proposed in Eq. 4.

In some cases—when the structure of g does not follow the model factorization either additively or multipicatively—computation of the model expectation $\mathbb{E}_{p_\theta(y|x)}[g(x, y)]$ in Eq. 7 is intractable. In these situations, standard Monte Carlo approximation breaks differentiability of the objective with respect to the model parameters θ and cannot be used. To remedy this, we propose to use the ''Stochastic Softmax'' differentiable sampling approximation from Paulus et al. (2020) to allow optimization of these functions. We propose several such statistics in the application (see Section 4.3).

2.3 The Distance Function ℓ

For the distance function ℓ, we propose to use a smoothed hinge loss (Girshick, 2015) that adapts with the margins σ. Letting $\bar{f} = f(\mathcal{D}_U^k, p_\theta)$, the i'th component of ℓ is given by:

$$\ell_i = \begin{cases} \frac{(\bar{f}_i - t_i)^2}{2\sigma_i} & \text{if } |\bar{f}_i - t_i| < \sigma_i \\ |\bar{f}_i - t_i| - \sigma_i & \text{else} \end{cases} \quad (9)$$

The total loss ℓ is then the sum of its components:

$$\ell(t, \sigma, f(\mathcal{D}_U^k, p_\theta)) = \sum_i \ell_i(t_i, \sigma_i, \bar{f}_i) \quad (10)$$

We choose this function because it is robust to outliers, adapts its width to the margin parameter σ_i, and expresses a preference for $f_i = t_i$ (as opposed to max-margin losses). We give an ablation study in Section 5.3.2 justifying its use.

3 Choosing the Targets and Margins

There are several possible approaches to choosing the targets t and margins σ, and in general they can differ based on the individual statistics. For some statistics it may be possible to specify the targets and margins using prior knowledge or formal constraints from the task. In other cases, estimating the targets and margins may be more difficult. Depending on the problem context, one may be able to estimate them from related tasks or domains (such as neighboring languages for cross-lingual parsing). Here, we propose a general method that estimates the statistics using labeled data, and is applicable to semi-supervised scenarios where at least a small amount of labeled data is available.

The ideal targets are the expected statistics under the ''true'' model p^* are: $t^* = \mathbb{E}_{\mathcal{D}_U^k}[f(\mathcal{D}_U^k, p^*)]$, where k is the batch size. We can estimate this expectation using labeled data $\mathcal{D}_L$ and bootstrap sampling (Efron, 1979). Utilizing $\mathcal{D}_L$ as a set of point estimates for p^*, we sample B total minibatches of k labeled examples uniformly with replacement from $\mathcal{D}_L$ and calculate the statistic f for each of these minibatch datasets. We then compute the target statistic as the sample mean:

$$t = \frac{1}{B} \sum_{i=1}^{B} f(\mathcal{D}_L^{(i)}), \quad |\mathcal{D}_L^{(i)}| = k, \ \forall i \quad (11)$$

where we have slightly abused notation by writing $f(\mathcal{D}_L)$ to mean f computed using the inputs $\{x : (x, y) \in \mathcal{D}_L\}$ and the point estimates $p^*(y|x) = 1, \ \forall (x, y) \in \mathcal{D}_L$.

In addition to estimating the target statistics for small batch sizes, the bootstrap gives us a way to estimate the natural variation of the statistics for small sample sizes. To this end, we propose to utilize the standard deviations from the bootstrap samples as our margins of uncertainty σ:

$$\sigma = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} (f(\mathcal{D}_L^{(i)}) - t)^2} \quad (12)$$

This allows our loss function ℓ to adapt to more or less certain statistics. If some statistics are naturally too variable to serve as effective supervision, they will automatically have weak contribution to ℓ and little impact on the model training.

4 Application to Cross-Lingual Parsing

Now that we have described our general approach, in this section we lay out a proposal for applying it to cross-lingual joint POS tagging and dependency parsing. We choose to apply our method to this problem because it is an ideal testbed for controlled experiments in semi-supervised structured prediction. By their nature, the parsing tasks admit many types of interesting statistics that capture cross-task, universal, and language-specific facts about the target test distributions.

We evaluate in two different transfer settings: oracle unsupervised and realistic semi-supervised. In the oracle unsupervised settings, there is no supervised training data available for the target languages (and the L term is dropped from Eq. 3), but we use target values and margins calculated from the held-out supervised data. This setting allows us to understand the impact of our regularizer in isolation without the confounding effects of direct supervision or inaccuracte targets. In the semi-supervised experiments, we vary the amounts of supervised data, and calculate the targets from the small supervised data samples. This is a realistic application of our approach that may be applied to low-resource learning scenarios.

4.1 Problem Setup and Data

We use the Universal Dependencies (Nivre, 2020) v2.8 (UD) corpus as data. In UD, syntactic annotation is formulated as a labeled bilexical dependency tree, connecting words in a sentence, with additional part-of-speech (POS) *tags* annotated for each word. The labeled tree can be broken down into two parts: the *arcs* that connect the *head* words to *child* words, forming a tree, and the dependency *labels* assigned to each of those arcs. Due to the definition of UD syntax, each word is the child of exactly one arc, and so both the attachments and labels can be written as sequences that align with the words in the sentence.

More formally then, for each labeled sentence $x_{1:n}$ of length n, the full structure y is given by the three sequences $y = (t_{1:n}, e_{1:n}, r_{1:n})$, where $t_{1:n}$, $t_i \in \mathcal{T}$ are the POS tags, $e_{1:n}$, $e_i \in \{1, \ldots, n\}$ are the head attachments, and $r_{1:n}$, $r_i \in \mathcal{R}$ are the dependency labels.

4.2 The Model and Training

We now turn to the parsing model that is used as the basis for our approach. Though the general ideas of our approach are adaptable to other models, we choose to use the UDify architecture because it is one of the state-of-the-art multilingual parsers for UD.

4.2.1 The UDify Model

The UDify model is based on trends in state-of-the-art parsing, combining a multilingual pretrained transformer language model encoder (mBERT) with a deep biaffine arc-factored parsing decoder, following Dozat and Manning (2017). These encodings are additionally used to predict POS tags with a separate decoder. The full details are given in Kondratyuk and Straka (2019), but here it suffices to characterize the parser by its top-level probabilistic factorization:

$$\begin{aligned} &p(t_{1:n}, e_{1:n}, r_{1:n} | x_{1:n}) \\ &\quad = p(e_{1:n}|x_{1:n}) p(t_{1:n}|x_{1:n}) p(r_{1:n}|e_{1:n}, x_{1:n}) \quad (13) \\ &\quad = p(e_{1:n}|x_{1:n}) \prod_{i=1}^{n} p(t_i|x_{1:n}) p(r_i|e_i, x_{1:n}) \quad (14) \end{aligned}$$

This model is scant on explicit joint factors, following recent trends in structured prediction that forgo higher-arity factors, instead opting for shared underlying contextual representations produced by a mBERT that implicitly contain information about the sentence and structure as a whole. This factorization will prove useful in Section 4.3 where it will allow us to compute many of the supervision statistics under the model exactly.

4.2.2 Training

The UDify approach to training is simple: It begins with a multilingual PLM, mBERT, then fine-tunes the parsing architecture on the concatenation of the source languages. With vanilla UDify, transfer to target languages is zero-shot.

Our approach begins with these two training steps from UDify, then adds a third: adapting to the target language using the target statistics and possibly small amounts of supervised data (Eq. 3).

4.3 Typological Statistics as Supervision

We now discuss a series of statistics that we will use as weak supervision. Most of the proposed statistics describe various probabilities for different (but related) grammatical substructures and can ultimately be broken down into ratios of ‘‘count’’ functions (sums of indicators), which tally various types of events in the data. We propose statistics that cover surface level (POS-only), single-arc, two-arc, and single-head substructures, as well as conditional variants. Due to space constraints, we omit their mathematical descriptions.

Surface Level: One simple set of descriptive statistics are the unigram and bigram distributions over POS tags. POS unigrams can capture some basic relative frequencies, such as our expectation that nouns and verbs are common to all languages. POS bigrams will allow us to capture simple word-order preferences.

Single-Arc: This next set of statistical families all capture information about various choices in single-arc substructures. A single arc substructure carries up to 5 pieces of information: the arc’s direction, label, and distance, as well as the tags for the head and child words. Various subsets of these capture differing forms of regularity, such as ‘‘the probability of seeing tag t_h head an arc with label r in direction d’’.

Universally Impossible Arcs: In addition to many single-arc variants, we also consider the specific subset of (head tag, label, child tag) single-arc triples that are never seen in the any UD data. These combinations correspond to the impossible arrangements that do not ‘‘type-check’’ within the UD formalism and are interesting in that they could in principle be specified by a linguist without any labeled data whatsoever. As such, they represent a particularly attractive use-case of our approach, where a domain expert could rule out all invalid substructures dictated from the task formalism without the model having to learn it implicitly from the training data. With complex structures, this can be a large proportion of the possibilities: in UD we can rule out 93.2% (9,966/10,693) of the combinations.

Two-Arc: We also consider substructures spanning two connected arcs in the tree. They may be useful because they cover many important typological phenomena, such as subject-object-verb ordering. They also have been known to be strong features in higher-order parsing models, such as the parser of Carreras (2007), but are also known to be intractable in non-projective parsers (McDonald and Pereira, 2006).

Following McDonald and Pereira (2006), we distinguish between two different patterns of neighboring arcs: *siblings* and *grandchildren*. Sibling arc pairs consist of two arcs that share a single head word, while grandchild arc pairs share an intermediate word that is the child of one arc and the head of another.

Head-Valency: One interesting statistic that does not fall into the other categories is the valency of a particular head tag. This corresponds to the count of outgoing arcs headed by some tag. We convert this into a probability by using a binning function that allows us to quantify the ‘‘probability that some tag heads between a and b children’’. Like the two-arc statistics, expected valency statistics are intractable under the model and we must approximate their computation.

Conditional Variants: Further, each of these statistics can be described in conditional terms, as opposed to their full joint realizations. To do this, we simply divide the joint counts by the counts of the conditioned-upon sub-events. Conditional variants may be useful because they do not express preferences for probabilities of the sub-events on the right side of the conditioning bar, which may be hard to estimate.

Average Entropy: In addition to the above proposed relative frequency statistics, we also include average per-token, per-edge, and MST tree entropies as additional regularization statistics that are always used. Though we do not show it here, each of these functions may be formulated as a statistic within our approach. The inclusion of these statistics amounts to a form of Entropy Regularization (Grandvalet and Bengio, 2004) that keep the models from optimizing the other ESR constraints with degenerate constant predictions (Mann and McCallum, 2010).

5 Oracle Unsupervised Experiments

We begin with oracle unsupervised transfer experiments that evaluate the potential of many

Language	Code	Treebank	Family	Train Sents	UDPRE LAS
Arabic	ar	PADT	Semitic	6.1k	80.5
Basque	eu	BDT	Basque	5.4k	77.0
Chinese	zh	GSD	Sino-Tibetan	4.0k	62.3
English	en	EWT	IE, Germanic	12.5k	88.1
Finnish	fi	TDT	Uralic	12.2k	84.4
Hebrew	he	HTB	Semitic	5.2k	80.5
Hindi	hi	HDTB	IE, Indic	13.3k	87.0
Italian	it	ISDT	IE, Romance	13.1k	91.8
Japanese	ja	GSD	Japanese	7.1k	73.6
Korean	ko	GSD	Korean	4.4k	79.0
Russian	ru	SynTagRus	IE, Slavic	15.0k*	89.1
Swedish	sv	Talbanken	IE, Germanic	4.3k	85.7
Turkish	tr	IMST	Turkic	3.7k	61.7
German	de	HDT	IE, Germanic	153.0k	82.7
Indonesian	id	GSD	Austronesian	4.5k	50.4
Maltese	mt	MUDT	Semitic	1.1k	20.9
Persian	fa	PerDT	IE, Iranian	26.2k	57.0
Vietnamese	vi	VTB	Austro-Asiatic	1.4k	48.1

Table 1: *Training and Evaluation Treebank Details*. The final column shows UDPRE test set performance after UDify training (evaluation treebank performance is zero-shot). ($*$): downsampled to the same 15k sentences as Üstün et al. (2020) to reduce training time and balance the data.

types of statistics and some ablations. In this setting, we do not assume any labeled data in the target language, but do assume accurate target statistics and margins, calculated from held-out training data using the method of Section 3. This allows us to study the potential of our proposed ESR regularization term C on its own and without the confounds of supervised data or inaccurate targets.

5.1 Experimental Setup

Next we describe setup details for the experiments. These settings additionally apply to the rest of the experiments unless otherwise stated.

5.1.1 Datasets

In all experiments, the models are first initialized from mBERT, then trained using the UDify code (Kondratyuk and Straka, 2019) on 13 diverse treebanks, following Kulmizev et al. (2019); Üstün et al. (2020). This model, further referred to as **UDPRE**, is used as the foundation for all approaches.

As discussed in Kulmizev et al. (2019), these 13 training treebanks were selected to give a diverse sample of languages, taking into account factors such as language families, scripts, morphological complexity, and annotation quality.

We evaluate all proposed methods on 5 held-out languages, similarly selected for a diversity in language typologies, but with the additional factor of tranfser performance of the **UDPRE** baseline.[2]

A summary table of these training and evaluation treebanks is given in Table 1.

5.1.2 Approaches

We compare our approach to two strong baselines in all experiments, based on recent advances in the literature for cross-lingual parsing. These baselines are implemented in our code so that we may fairly compare them in all of our experiments.

- **UDPRE**: The first baseline is the UDify (Kondratyuk and Straka, 2019) model-transfer approach. Multilingual model-transfer alone

[2] While we would like to evaluate on as many UD treebanks as possible, budgetary constraints required that we restrict the number of test languages when experimenting with settings that combinatorially vary in other dimensions. We do however experiment with more languages in Section 6.2.

is currently one of the state-of-the-art approaches to cross-lingual parsing and is a strong baseline in its own right.

- **UDpre-PPT:** We also apply the Parsimonious Parser Transfer (PPT) approach from Kurniawan et al. (2021). PPT is a nuanced self-training approach, extending Täckström et al. (2013), that encourages the model to concentrate its mass on its most likely predicted parses for the target treebank. We use their loss implementation, but apply it to our UDpre base model (instead of their weaker base model) for a fair comparison, so this approach combines UDify with PPT.
- **UDpre-ESR:** Our proposed approach, Expected Statistic Regularization (ESR), applied to UDpre as an unsupervised-only objective. In individual experiments we will specify the statistics used for regularization.

5.1.3 Training and Evaluation Details

For metrics, we report accuracy for POS tagging, coarse-grained labeled attachment score (LAS) for dependency trees, and their average as a single summary score. The metrics are computed using the official CoNLL-18 evaluation script.[3] For all scenarios, we use early-stopping for model selection, measuring the POS-LAS average on the specified development sets.

We tune learning rates and α for each proposed loss variant at the beginning of the first experiment with a low-budget grid search, using the settings that achieve best validation metric on average across the 5 language validation sets for all remaining experiments with that variant. We find generally that a base learning rate of 2×10^{-5} and $\alpha = 0.01$ worked well for all variants of our method. We train all models using AdamW (Loshchilov and Hutter, 2019) on a slanted triangular learning rate schedule (Devlin et al., 2019) with 500 warmup steps. Also, since the datasets vary in size, we normalize the training schedule to 25 epochs at 1000 steps per epoch. We use a batch size of 8 sentences for training and estimating statistic targets. When bootstrapping estimates for t and σ we use $B = 1000$ samples.

5.2 Assessing the Proposed Statistics

In this experiment we evaluate 32 types of statistics from Section 4.3 for transfer of the UDpre model (pretrained on 13 languages) to the target languages. The purpose of this experiment is to get a sense of the effectiveness of each statistic for improving model-based cross-lingual transfer.[4] To prevent overfitting to the test sets for later experiments, all metrics for this experiment are calculated on the development sets.

Results: The results of the experiment are presented in Table 2, ranked from best to worst. Due to space constraints, we only show the top 10 statistics in addition to the Universal-Arc statistic. Generally we find that all of the 32 proposed statistics improve upon the UDpre and UDpre-PPT models on average, with many exhibiting large boosts. The best performing statistic concerns (Child Tag, Label, Direction) substructures, yielding an average improvement of +7.0 POS and +8.5 LAS, an average relative error rate reduction of 23.5%. Many other statistics are not far behind, and overall statistics that bear on the child tag and dependency label had the highest impact. This indicates that, with accurate target estimates, the proposed statistics are highly complementary to multilingual parser pretraining (UDpre) and substantially improve transfer quality in the unsupervised setting. By comparison, the PPT approach provides marginal gains to UDpre of only +1.4 average POS and +1.5 average LAS.

Another interesting result is that several of the intractable two-arc statistics were among the best statistics overall, indicating that the use of the differentiable SST approximation does not preclude the applicability of intractable statistics. For example the directed grandchild statistic of cooccurrences of incoming and outgoing edges for certain tags was the second highest performing, with an average improvement of +7.0 POS accuracy and +8.5 LAS (21.3% average error rate reduction).

Results for the conditional variants (not shown) were less positive. Generally, conditional variants were worse than their full joint counterparts (e.g., ‘‘Child | Label’’ and ‘‘Label | Child’’ are worse than ‘‘Child, Label’’), performing worse in 15/16 cases. This makes sense, as we are using

[3] https://universaldependencies.org/conll18/evaluation.html.

[4] While it would also be possible to try out different combinations of the various statistics, due to cost considerations we leave these experiments to future work.

	POS					LAS					
Statistic	de	id	fa	vi	mt	de	id	fa	vi	mt	avg
UDpre	89.3	80.3	83.0	64.7	41.4	82.7	50.4	57.0	48.1	20.9	61.8
UDpre-PPT	+0.4	+5.6	−1.5	−0.1	+3.1	+0.2	+8.1	−5.5	−0.3	+4.6	+1.5
†**Child, Label**	+3.3	+5.7	+8.0	+4.0	+14.0	+3.5	+10.1	+18.4	+0.5	+10.2	+7.8
*†Child, Label, Grand-label	+1.5	+2.8	+5.3	+5.9	+15.7	+2.5	+9.2	+16.2	+3.2	+12.5	+7.5
†Head, Child, Label	+2.5	+4.9	+7.2	+4.1	+14.2	+2.8	+9.9	+17.2	+1.3	+10.2	+7.4
†Head, Label	+1.7	+3.2	+5.2	+6.3	+14.2	+2.7	+9.0	+16.4	+3.8	+11.0	+7.3
†Head, Label | Child	+3.0	+5.2	+5.8	+5.7	+10.9	+2.8	+9.7	+15.3	+2.1	+5.6	+6.6
†Label	−0.2	+4.4	+5.1	+0.0	+11.3	+2.9	+8.4	+17.1	+4.0	+9.5	+6.2
†Label, Distance	−0.2	+3.7	+3.9	−0.1	+11.7	+2.8	+8.9	+16.3	+4.1	+9.4	+6.0
*†Head, Sibling Children Tags	+2.2	+5.1	+5.6	+7.8	+14.0	+1.6	+2.7	+11.3	−3.0	+11.1	+5.8
†Head, Child	+2.3	+5.3	+5.9	+3.7	+14.0	+1.9	+3.3	+12.2	−0.7	+10.1	+5.8
†Label | Child	+1.2	+4.3	+4.8	−0.5	+10.0	+3.3	+9.5	+16.6	+2.0	+5.6	+5.7
Universal Arc	+2.4	+4.7	+1.4	+1.4	+8.7	+2.1	+8.1	+4.0	−3.1	+3.9	+3.4

Table 2: *Unsupervised Oracle Statistic Variant Results. (Top):* Baseline methods that do not use ESR. *(Bottom):* Various statistics used by ESR as unsupervised loss on top of UDpre. Scores are measured on target treebank development sets. Bold names mark statistics used in later experiments. $(*)$: All statistics with $*$ are intractable and utilize the SST relaxation of Paulus et al. (2020). (†): All statistics with † also include directional information.

accurate statistics and full joints are strictly more expressive.

This experiment gives a broad but shallow view into the effectiveness of the various proposed statistics. In the rest of the experiments, we evaluate the following two variants in more depth:

1. **ESR-CLD**, which supervises target proportions for (Child Tag, Label, Direction) triples. This is the ‘‘Child, Label’’ row in Table 2.
2. **ESR-UniArc**, which supervises the 9,966 universally impossible (Head Tag, Child Tag, Label) arcs that do not require labeled data to estimate. All of these combinations have targets values of $t = 0$ and margins $\sigma = 0$. This is the ‘‘Universal Arc’’ row in Table 2.

We choose these two because ESR-CLD is the best performing statistic overall and ESR-UniArc is unique in that it does not require labeled data to estimate; we do not evaluate others because of cost considerations.

5.3 Ablation Studies

Next, we perform two ablation experiments to evaluate key design choices of the proposed approach. First, we evaluate the use of batch-level aggregation in the statistics before the loss, versus the more standard approach of loss-per-sentence. In the second, we evaluate the proposed form of ℓ.

We compare the two aggregation variants using the CLD (Child Tag, Label, Direction) statistic (ESR-CLD). We report test set results averaged over all 5 languages. We use the same hyperparameters selected in Section 5.2.

Aggregation Variant	POS avg	LAS avg	avg
Loss per sentence	77.1	58.5	67.8
Loss per batch (ESR)	**79.9**	**60.4**	**70.1**

Table 3: *Loss Aggregation Ablation Results.* Loss per batch outperforms loss per sentence for both POS and LAS on average.

5.3.1 Batch-Level Loss Ablation

In this ablation, we evaluate a key feature of our proposal—the aggregation of the statistic over the batch before loss computation Eq. 5 versus the more standard approach, which is to apply the loss per-sentence. The former, ‘‘Loss per batch’’, has the form: $\ell(t, \sigma, f(\mathcal{D}_U, p_\theta))$ while the latter, ‘‘Loss per sentence’’, has the form:

$$\sum_{x \in \mathcal{D}_U} \ell(t, \sigma, f(x, p_\theta)).$$

The significance of this difference is that ‘‘Loss per batch’’ allows for the variation in individual sentences to somewhat average out and hence is less noisy, while ‘‘Loss per sentence’’ requires that each sentence individually satisfy the targets.

Results: The results are presented in Table 3. From the table we can see that ‘‘Loss per batch’’ has an average POS of 79.9 and average LAS of

ℓ Variant	POS avg	LAS avg	avg
L2 ($\sigma = 0$)	78.0	58.2	68.1
L1 ($\sigma = 0$)	78.5	60.3	69.5
Hard L1 (max-margin)	78.4	59.9	69.2
Smooth L1 (ESR)	**79.9**	**60.4**	**70.1**

Table 4: *Loss Function Ablation Results.* The Smooth L1 loss outperforms the other simpler loss variants for both POS and LAS, averaged over 5 languages.

60.4, compared to ''Loss per sentence'' with average POS of 77.1 and LAS of 58.5, which amount to +2.8 POS and +1.9 LAS improvements. This indicates that applying the loss at the batch level confers an advantage over applying per sentence.

5.3.2 Smooth Hinge-Loss Ablation

Next, we evaluate the efficacy of the proposed smoothed hinge-loss distance function ℓ. We compare to using just L1 or L2 uninterpolated and with no margin parameters ($\sigma = 0$). We also compare to the ''Hard L1'', which is the max-margin hinge $\ell(t, \sigma, x) = \max\{0, |t - x| - \sigma\}$. We use the same experimental setup as the previous ablation.

Results: The results are presented in Table 4. From the table we can see that the Smooth L1 loss outperforms the other variants.

6 Realistic Semi-Supervised Experiments

The previous experiments considered an unsupervised transfer scenario without labeled data. In these next experiments we turn to a realistic semi-supervised application of our approach where we have access to limited labeled data for the target treebank.

6.1 Learning Curves

In this experiment we present learning curves for the approaches, varying the amount of labeled data $|\mathcal{D}_L^{\text{train}}| \in \{50, 100, 500, 1000\}$. To make experiments realistic, we calculate the target statistics t and margins σ from the small subsampled labeled training datasets using Eqs. 11 and 12.

We study two distinct settings. First, we study the multi-source domain-adaptation transfer setting, UDPRE. Second, we study our approach in a more standard semi-supervised scenario where we cannot utilize intermediate on-task pretraining and domain-adaption, instead learning on the target dataset starting ''from scratch'' with the pretrained PLM (MBERT).

We use the same baselines as before, but augment each with a supervised fine-tuning loss on the supervised data in addition to any unsupervised losses. We refer to these models as **UDPRE-FT, UDPRE-FT-PPT,** and **UDPRE-FT-ESR**. That is, models with **FT** in the name have some supervised fine-tuning in the target language.

In these experiments, we subsample labeled training data 3 times for each setting. We report averages over all 5 languages, 3 supervised subsample runs each, for a total of 15 runs per method and dataset size. We also use subsampled development sets so that model selection is more realistic.[5] For development sets we subsample the data to a size of $|\mathcal{D}_L^{\text{dev}}| = \min(100, |\mathcal{D}_L^{\text{train}}|)$, which reflects a 50/50 train/dev split until $|\mathcal{D}_L| \geq 200$, at which point we maximize training data and only hold out 100 sentences for validation.

We use the same hyperparameters as before, except we use 40 epochs with 200 steps per epoch as the training schedule, mixing supervised and unsupervised data at a rate of 1:4.

6.1.1 UDPRE Transfer

In this experiment, we evaluate in the multlingual transfer scenario by initializing from UDPRE. In addition to the two chosen realistic ESR variants, we also experiment with an ''oracle'' version of ESR-CLD, called ESR-CLD*, that uses target statistics estimated from the full training data. This allows us to see if small-sample estimates cause a degradation in performance compared to accurate large-sample estimates.

Results: Learning curves for the different approaches, averaged over all 3 runs for all 5 languages (15 total), are given in Figure 1. From the figure we can discern several encouraging results.

ESR-CLD and ESR-UNIARC add significant benefit to fine-tuning for small data. Both variants significantly outperform the baselines at 50 and 100 labeled examples. For example, relative to UDPRE-FT, the ESR-CLD model yielded gains of +2 POS, +3.6 LAS at 50 examples and +1.8

[5] As is argued by Oliver et al. (2018), using a realistically sized development set is overlooked in much of the semi-supervised literature, leading to inappropriately strong model selection and overly optimistic results.

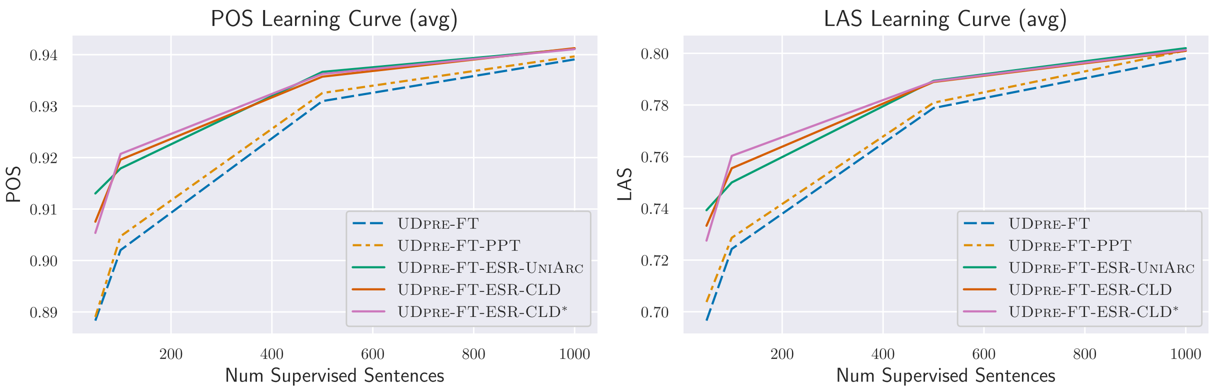

Figure 1: *Multi-Source* UDPRE *Transfer Learning Curves*. Baseline approaches are dotted, while ESR variants are solid. All curves show the average of 15 runs across 5 different languages with 3 randomly sampled labeled datasets per language. The plots indicate a significant advantage of ESR over the baselines in low-data regions.

POS, +3.2 LAS at 100 labeled examples. At 500 and 1000 examples, however, we begin to see diminishing benefits to ESR on top of fine-tuning.

ESR-UNIARC is much more effective in conjunction with fine-tuning. Compared to the unsupervised experiment in Section 5.2 where it ranked 25/32, the ESR-UNIARC statistic is much more competitive with the more detailed ESR-CLD statistics. One potential explanation is that without labeled data (as in Section 5.2) the ESR-UNIARC statistic is under-specified (the 727 allowed arcs are all free to take any value), whereas the inclusion of some labeled data in this experiment fills this gap by implicitly indicating target proportions for the allowed arcs. This suggests that an approach which combines UniArc constraints with elements of self-training (like PPT) that supervise the ‘‘free’’ non-zero combinations could potentially be a useful approach to zero-shot transfer. However, we leave this to future work.

Small-data estimates for ESR-CLD are as good as accurate estimates. Comparing ESR-CLD to the unrealistic ESR-CLD*, we find no significant difference between the two, indicating that, at least for the CLD statistic, using target estimates from small samples is as good as large-sample estimates. This may be due in part to the margin estimates σ, which are wider for the small samples and somewhat mitigate their inaccuracies.

PPT adds little benefit to fine-tuning. Relative to UDPRE-FT, the UDPRE-FT-PPT baseline does not yield much gain, with a maximum average improvement of +0.3 POS and +0.7 LAS over all dataset sizes. This indicates that fine-tuning and PPT-style self-training may be redundant.

6.1.2 MBERT Transfer

In this experiment, we consider a counterfactual setting: What if the UD data was not a massively multilingual dataset where we can utilize multilingual model-transfer, and instead was an isolated dataset with no related data to transfer from? This situation reflects the more standard semi-supervised learning setting, where we are given a new task, some labeled and unlabeled data, and must build a model ‘‘from scratch’’ on that data.

For this experiment, we repeat the learning curve setting from Section 6.1.1, but initialize our model directly with **MBERT**, skipping the intermediate **UDPRE** training.

Results: Learning curves for the different approaches, averaged over all 3 runs for all 5 languages (15 total), are given in Figure 2. The results from this experiment are encouraging; ESR has even greater benefits when fine-tuning directly from MBERT than the previous experiment, indicating that our general approach may be even more useful outside of domain-adaptation conditions.

6.2 Low-Resource Transfer

In previous experiments, we limited the number of evaluation treebanks to 5 to allow for variation in other dimensions (i.e., constraint types, loss types, differing amounts of labeled data). In this experiment, we expand the number of treebanks and evaluate transfer performance in a low-resource setting with only $|\mathcal{D}_L^{\text{train}}| = 50$ labeled sentences in the target treebank, comparing UDPRE, UDPRE-FT, and UDPRE-FT-ESR-CLD. As before, we subsample 3 small datasets per treebank and calculate

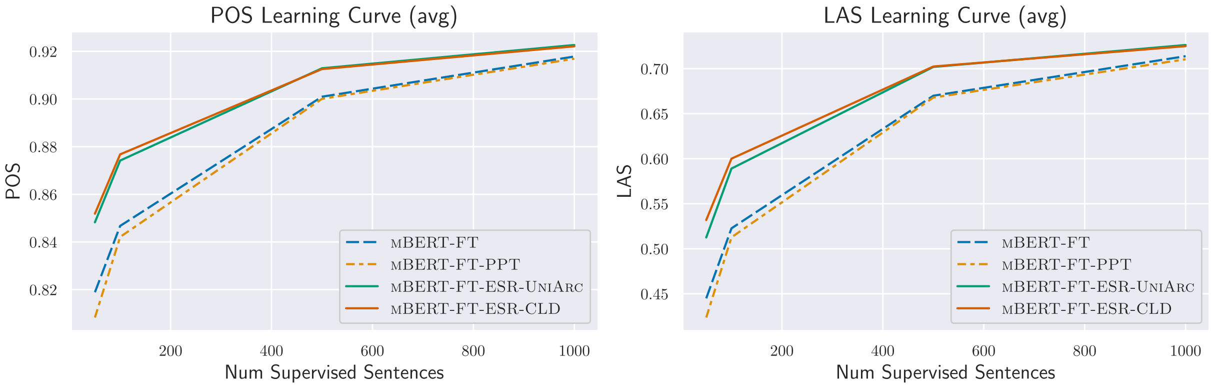

Figure 2: *"From Scratch"* mBERT *Transfer Learning Curves.* Baseline approaches are dotted, while ESR variants are solid. All curves show the average of 15 runs across 5 different languages with 3 randomly sampled labeled datasets per language. The plots indicate a significant advantage of ESR over the baselines in all regions.

the target statistics t and margins σ from these to make transfer results realistic.

We select evaluation treebanks according to the following criteria. For each unique language in UD v2.8 that is not one of the 13 training languages, we select the largest treebank, and keep it if has at least 250 train sentences and a development set, so that we can get reasonable variability in the subsamples. This process yields 44 diverse evaluation treebanks.

Results: The results of this experiment are given in Table 5. From the table we can see the our approach ESR (UDpre-FT-ESR-CLD) outperformed supervised fine-tuning (UDpre-FT) in many cases, often by a large margin. On average, UDpre-FT-ESR-CLD outperformed UDpre-FT by +2.6 POS and +2.3 LAS across the 44 languages. Further, UDpre-FT-ESR-CLD outperformed zero shot transfer, UDpre, by $+10.0$ POS and $+14.7$ LAS on average.

Interestingly, we found that there were several cases of large performance gains while there were no cases of large performance declines. For example, ESR improved LAS scores by $+17.3$ for Wolof, $+16.8$ for Maltese, and $+12.5$ for Scottish Gaelic, and 9/44 languages saw LAS improvements $\geq +5.0$, while the largest decline was only -2.5. Additionally, ESR improved POS scores by $+20.9$ for Naija, $+11.2$ for Welsh, and 9/44 languages saw POS improvements $\geq +5.0$.

The cases of performance decline for LAS merit further analysis. Of the 20 languages with negative Δ LAS, 18 of these are modern languages spoken in continental Europe (mostly Slavic and Romance), while only 5 of the 24 languages with positive Δ LAS meet this criteria. We hypothesize that this tendency is be due to the training data used for pretraining mBERT, which was heavily skewed towards this category (Devlin et al., 2019). This suggests that ESR is particularly helpful in cases of transfer to domains that are underrepresented in pretraining.

7 Related Work

Related work generally falls into two categories: weak supervision and cross-lingual transfer.

Weak Supervision: Supervising models with signals weaker than fully labeled data has and continues to be a popular topic of interest. Current trends in weak supervision focus on generating instance-level supervision, using weak information such as: relations between multiple tasks (Greenberg et al., 2018; Ratner et al., 2018; Ben Noach and Goldberg, 2019); labeled features (Druck et al., 2008; Ratner et al., 2016; Karamanolakis et al., 2019a); coarse-grained labels (Angelidis and Lapata, 2018; Karamanolakis et al., 2019b); dictionaries and distant supervision (Bellare and McCallum, 2007; Carlson et al., 2009; Liu et al., 2019a; Üstün et al., 2020); or some combination thereof (Ratner et al., 2016; Karamanolakis et al., 2019a).

In contrast, our work is more closely related to older work on population-level supervision. These techniques include Constraint-Driven Learning (CODL) (Cha), posterior regularization (PR) (Ganchev et al., 2010), the measurements framework of Liang et al. (2009), and the generalized expectation criteria (GEC) (Druck et al., 2008, 2009; Mann and McCallum, 2010).

		POS				LAS			
Treebank	Family	UDPRE	FT	ESR	Δ	UDPRE	FT	ESR	Δ
Wolof-WTB	Northern Atlantic	40.6	79.5	**85.4**	+5.9	12.7	55.9	**73.3**	+17.3
Maltese-MUDT	Semitic	35.1	82.6	**91.8**	+9.2	16.0	57.5	**74.2**	+16.8
Scottish_Gaelic-ARCOSG	Celtic	45.7	66.0	**75.9**	+9.9	24.4	56.4	**68.9**	+12.5
Faroese-FarPaHC	Germanic	74.7	86.2	**87.2**	+1.1	43.0	71.4	**80.7**	+9.3
Gothic-PROIEL	Germanic	30.1	67.6	**71.7**	+4.1	12.6	45.8	**54.6**	+8.8
Welsh-CCG	Celtic	71.9	74.7	**85.8**	+11.2	54.8	69.4	**77.6**	+8.1
Western_Armenian-ArmTDP	Armenian	80.6	84.9	**87.1**	+2.2	60.4	67.0	**72.7**	+5.7
Telugu-MTG	Dravidian	82.0	**81.6**	**81.6**	0.0	70.9	74.6	**80.1**	+5.5
Vietnamese-VTB	Viet-Muong	67.0	85.6	**88.5**	+2.9	46.3	55.3	**60.8**	+5.5
Turkish_German-SAGT	Code Switch	76.8	84.4	**85.8**	+1.4	48.0	58.0	**62.1**	+4.1
Afrikaans-AfriBooms	Germanic	90.7	88.0	**91.3**	+3.3	62.0	79.4	**83.4**	+3.9
Hungarian-Szeged	Ugric	87.9	79.9	**89.7**	+9.7	74.0	77.8	**81.7**	+3.9
Galician-CTG	Romance	91.8	89.0	**91.2**	+2.2	60.5	74.3	**77.8**	+3.6
Marathi-UFAL	Marathi	71.4	81.1	**82.**3	+1.1	44.9	59.5	**62.5**	+3.0
Naija-NSC	Creole	46.5	68.0	**88.9**	+20.9	27.9	71.1	**73.4**	+2.3
Greek-GDT	Greek	87.1	**92.8**	92.5	−0.3	78.7	86.3	**88.0**	+1.8
Tamil-TTB	Dravidian	72.3	72.4	**79.6**	+7.2	46.7	64.9	**66.4**	+1.5
Indonesian-GSD	Austronesian	82.3	89.8	**90.2**	+0.5	58.3	72.9	**74.3**	+1.4
Uyghur-UDT	Turkic	23.7	59.8	**65.5**	+5.6	14.0	38.0	**39.2**	+1.3
Old_French-SRCMF	Romance	65.3	74.2	**76.2**	+2.0	44.0	56.7	**57.8**	+1.2
Old_Church_Slavonic-PROIEL	Slavic	37.3	54.7	**61.0**	+6.3	19.2	39.0	**40.1**	+1.1
Portuguese-GSD	Romance	92.1	89.6	**92.8**	+3.3	74.4	84.1	**84.5**	+0.4
Danish-DDT	Germanic	92.0	**92.7**	92.1	−0.6	71.0	75.5	**75.7**	+0.2
Armenian-ArmTDP	Armenian	84.7	**88.1**	88.0	−0.1	64.1	69.0	**69.2**	+0.1
Spanish-AnCora	Romance	94.5	95.2	**95.4**	+0.2	77.8	**83.0**	82.9	−0.1
Catalan-AnCora	Romance	92.9	94.4	**94.6**	+0.3	75.8	**82.5**	82.4	−0.1
Serbian-SET	Slavic	91.2	90.7	**93.1**	+2.4	81.6	**86.5**	86.4	−0.1
Slovak-SNK	Slavic	91.5	91.5	**92.0**	+0.5	81.6	**84.0**	83.9	−0.1
Romanian-Nonstandard	Romance	79.2	83.3	**85.0**	+1.7	54.5	**63.6**	63.4	−0.2
Polish-PDB	Slavic	89.7	90.4	**90.9**	+0.5	76.0	**79.7**	79.4	−0.3
German-HDT	Germanic	89.6	**94.4**	94.2	−0.2	83.0	**88.2**	87.7	−0.5
Lithuanian-ALKSNIS	Baltic	87.0	**87.4**	**87.4**	0.0	65.4	**69.2**	68.6	−0.6
Latin-ITTB	Italic	73.8	80.9	**81.7**	+0.8	51.7	**64.3**	63.7	−0.6
Bulgarian-BTB	Slavic	91.9	**94.7**	94.6	−0.1	78.0	**84.4**	83.7	−0.7
Czech-PDT	Slavic	90.6	92.1	**92.7**	+0.6	78.1	**81.9**	81.1	−0.8
Persian-PerDT	Iranian	79.1	**91.0**	90.8	−0.2	48.4	**74.6**	73.7	−0.9
Slovenian-SSJ	Slavic	89.2	90.9	**91.2**	+0.3	79.6	**84.5**	83.5	−0.9
Croatian-SET	Slavic	91.4	91.7	**92.1**	+0.4	80.0	**84.1**	83.1	−1.0
Urdu-UDTB	Indic	86.9	**90.0**	88.2	−1.8	68.7	**75.7**	74.4	−1.3
Ukrainian-IU	Slavic	91.5	92.0	**92.4**	+0.3	79.6	**81.2**	80.0	−1.3
Dutch-Alpino	Germanic	90.0	**90.6**	**90.6**	0.0	78.9	**81.6**	80.3	−1.3
Norwegian-Bokmaal	Germanic	91.7	91.8	**92.1**	+0.3	80.8	**82.5**	81.0	−1.5
Belarusian-HSE	Slavic	91.5	91.6	**91.9**	+0.3	78.9	**79.8**	78.1	−1.8
Estonian-EDT	Finnic	89.1	**89.6**	89.2	−0.4	70.4	**71.4**	68.9	−2.5
Average		77.3	84.7	**87.3**	+2.6	59.0	71.4	**73.7**	+2.3

Table 5: *Low-Resource Semi-Supervised Transfer Results.* Transfer results for 44 unseen test languages using 50 labeled sentences in the target language, averaged over 3 subsampled datasets. ''FT'' refers to the UDPRE-FT fine-tuning baseline, ''ESR'' refers to our UDPRE-ESR-CLD approach, and Δ refers to the absolute difference of ESR minus FT. Best performing methods are bolded. Results are ordered from best to worst Δ LAS.

Our work can be seen as an extension of GEC to more expressive expectations and to modern mini-batch SGD training. There are a two more recent works that touch on these ideas, but both have significant downsides compared to our approach. Meng et al. (2019) use a PR approach inspired by Ganchev and Das (2013) for cross-lingual parsing, but must use very simple constraints and require a slow inference procedure that can only be used at test time. Ben Noach and Goldberg (2019) utilize GEC with minibatch training, but focus on using related tasks for computing simpler constraints and do not adapt their targets to small batch sizes.

Cross-Lingual Transfer: Earlier trends in cross-lingual transfer for parsing used delexicalization (Zeman and Resnik, 2008; McDonald et al., 2011; Täckström et al., 2013) and then aligned multilingual word vector-based approaches (Guo et al., 2015; Ammar et al., 2016; Rasooli and Collins, 2017; Ahmad et al., 2019). With the rapid rise of language-model pretraining (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019b), recent research has focused on multilingual PLMs and multi-task fine-tuning to achieve generalization in transfer. Wu and Dredze (2019) showed that a multilingual PLM afforded surprisingly effective cross-lingual transfer using only English as the fine-tuning language. Kondratyuk and Straka (2019) extended this approach by fine-tuning a PLM on the concatenation of all treebanks. Tran and Bisazza (2019), however, show that transfer to distant languages benefit less.

Other recent successes have been found with linguistic side-information (Meng et al., 2019; Üstün et al., 2020), careful methodology for source-treebank selection (Tiedemann and Agic, 2016; Tran and Bisazza, 2019; Lin et al., 2019; Glavaš and Vulić, 2021), self-training (Kurniawan et al., 2021), and paired bilingual text for annotation projection (Rasooli and Tetreault, 2015; Rasooli and Collins, 2019; Liu et al., 2020; Shi et al., 2022).

8 Conclusion

We have presented Expected Statistic Regularization, a general approach to weak supervision for structured prediction, and studied it in the context of modern cross-lingual multi-task syntactic parsing. We evaluated a wide range of expressive structural statistics in idealized and realistic transfer scenarios and have shown that the proposed approach is effective and complementary to the state-of-the-art model-transfer approaches.

Acknowledgments

We would like to thank Chris Kedzie, Giannis Karamanolakis, and the reviewers for helpful conversations and feedback.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1035

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444. https://doi.org/10.1162/tacl_a_00109

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31. https://doi.org/10.1162/tacl_a_00002

Kedar Bellare and Andrew McCallum. 2007. Learning extractors from unlabeled text using relevant databases. In *Sixth international workshop on information integration on the web (AAAI)*.

Matan Ben Noach and Yoav Goldberg. 2019. Transfer learning between related tasks using expected label proportions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 31–42, Hong Kong, China. Association for Computational Linguistics.

Andrew Carlson, Scott Gaffney, and Flavian Vasile. 2009. Learning a named entity tagger from gazetteers with the partial perceptron. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 7–13.

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 957–961, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 595–602, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1390334.1390436

Gregory Druck, Gideon Mann, and Andrew McCallum. 2009. Semi-supervised learning of dependency parsers using generalized expectation criteria. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 360–368, Suntec, Singapore. Association for Computational Linguistics.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26. https://doi.org/10.1214/aos/1176344552

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1996–2006, Seattle, Washington, USA. Association for Computational Linguistics.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.431

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. 2018. Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2824–2829, Brussels, Belgium. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1119

Junxian He, Zhisong Zhang, Taylor Berg-Kirkpatrick, and Graham Neubig. 2019. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3223, Florence, Italy. Association for Computational Linguistics.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019a. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4611–4621, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1468

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019b. Weakly supervised attention networks for fine-grained opinion mining and public health. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1279

Artur Kulmizev, Miryam de Lhoneux, Johannes Gontrum, Elena Fano, and Joakim Nivre. 2019. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2755–2768, Hong Kong, China. Association for Computational Linguistics.

Kemal Kurniawan, Lea Frermann, Philip Schulz, and Trevor Cohn. 2021. PPT: Parsimonious parser transfer for unsupervised cross-lingual adaptation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2907–2918, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.254

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 641–648, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1553374.1553457

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Lu Liu, Yi Zhou, Jianhan Xu, Xiaoqing Zheng, Kai-Wei Chang, and Xuanjing Huang. 2020. Cross-lingual dependency parsing by POS-guided word reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2938–2948, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.265

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019a. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692v1*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Gideon S. Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–88, Trento, Italy. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tao Meng, Nanyun Peng, and Kai-Wei Chang. 2019. Target language-aware constrained inference for cross-lingual dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1117–1128, Hong Kong, China. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D19-1103`

Joakim Nivre. 2020. Multilingual dependency parsing from universal dependencies to sesame street. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 11–29. Springer.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*, pages 3239–3250.

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J. Maddison. 2020. Gradient estimation with stochastic softmax tricks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual.*

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N18-1202`

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293.

Mohammad Sadegh Rasooli and Michael Collins. 2019. Low-resource syntactic transfer with unsupervised source reordering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3845–3856, Minneapolis, Minnesota. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N19-1385`

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *ArXiv*, abs/1503.06733.

Alexander Ratner, Braden Hancock, Jared Dunnmon, Roger E. Goldman, and Christopher Ré. 2018. Snorkel metal: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning, DEEM@SIGMOD 2018, Houston, TX, USA, June 15, 2018*, pages 3:1–3:4. ACM. `https://doi.org/10.1145/3209889.3209898`

Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain*, pages 3567–3575.

Freda Shi, Kevin Gimpel, and Karen Livescu. 2022. Substructure distribution projection for zero-shot cross-lingual dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6563, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.452`

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. 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, Atlanta, Georgia. Association for Computational Linguistics.

Jörg Tiedemann and Zeljko Agic. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, 55:209–248. `https://doi.org/10.1613/jair.4785`

Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 281–288, Hong Kong, China. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, pages 2302–2315. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.180`

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.