PPT: Parsimonious Parser Transfer for Unsupervised Cross-Lingual Adaptation

Kemal Kurniawan[1] **Lea Frermann**[1] **Philip Schulz**[2*] **Trevor Cohn**[1]
[1]School of Computing and Information Systems, University of Melbourne
[2]Amazon Research
kemal.kurniawan@student.unimelb.edu.au
lea.frermann@unimelb.edu.au
phschulz@amazon.com
tcohn@unimelb.edu.au

Abstract

Cross-lingual transfer is a leading technique for parsing low-resource languages in the absence of explicit supervision. Simple 'direct transfer' of a learned model based on a multilingual input encoding has provided a strong benchmark. This paper presents a method for unsupervised cross-lingual transfer that improves over direct transfer systems by using their output as implicit supervision as part of self-training on unlabelled text in the target language. The method assumes minimal resources and provides maximal flexibility by (a) accepting any pre-trained arc-factored dependency parser; (b) assuming no access to source language data; (c) supporting both projective and non-projective parsing; and (d) supporting multi-source transfer. With English as the source language, we show significant improvements over state-of-the-art transfer models on both distant and nearby languages, despite our conceptually simpler approach. We provide analyses of the choice of source languages for multi-source transfer, and the advantage of non-projective parsing. Our code is available online.[1]

1 Introduction

Recent progress in natural language processing (NLP) has been largely driven by increasing amounts and size of labelled datasets. The majority of the world's languages, however, are low-resource, with little to no labelled data available (Joshi et al., 2020). Predicting linguistic labels, such as syntactic dependencies, underlies many downstream NLP applications, and the most effective systems rely on labelled data. Their lack hinders the access to NLP technology in many languages. One solution is cross-lingual model transfer, which adapts models trained on high-resource languages to low-resource ones. This paper presents a flexible framework for cross-lingual transfer of syntactic dependency parsers which can leverage *any* pre-trained arc-factored dependency parser, and assumes no access to labelled target language data.

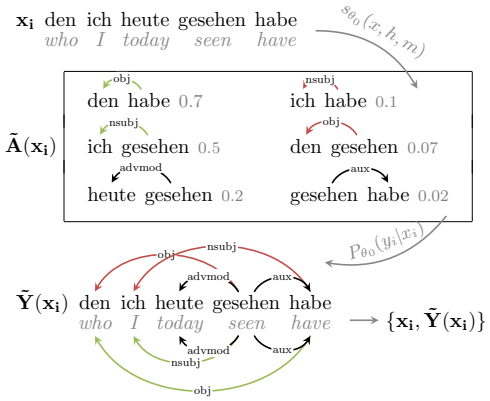

Figure 1: Illustration of our technique. For a target language sentence (x_i), a source parser P_{θ_0} predicts a set of candidate arcs $\tilde{A}(x_i)$ (subset shown in the figure), and parses $\tilde{Y}(x_i)$. The highest scoring parse is shown on the bottom (green), and the true gold parse (unknown to the parser) on top (red). A target language parser P_θ is then fine-tuned on a data set of ambiguously labelled sentences $\{x_i, \tilde{Y}(x_i)\}$.

One straightforward method of cross-lingual parsing is direct transfer. It works by training a parser on the source language labelled data and subsequently using it to parse the target language directly. Direct transfer is attractive as it does not require labelled target language data, rendering the approach fully unsupervised.[2] Recent work has shown that it is possible to outperform direct transfer if unlabelled data, either in the target lan-

*Work done outside Amazon.

[1]https://github.com/kmkurn/ppt-eacl2021

[2]Direct transfer is also called zero-shot transfer or model transfer in the literature.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2907–2918
April 19 - 23, 2021.

guage or a different auxiliary language, is available (He et al., 2019; Meng et al., 2019; Ahmad et al., 2019b). Here, we focus on the former setting and present flexible methods that can adapt a pre-trained parser given unlabelled target data.

Despite their success in outperforming direct transfer by leveraging unlabelled data, current approaches have several drawbacks. First, they are limited to generative and projective parsers. However, discriminative parsers have proven more effective, and non-projectivity is a prevalent phenomenon across the world's languages (de Lhoneux, 2019). Second, prior methods are restricted to single-source transfer, however, transfer from multiple source languages has been shown to lead to superior results (McDonald et al., 2011; Duong et al., 2015a; Rahimi et al., 2019). Third, they assume access to the source language data, which may not be possible because of privacy or legal reasons. In such source-free transfer, only a pre-trained source parser may be provided.

We address the three shortcomings with an alternative method for unsupervised target language adaptation (Section 2). Our method uses high probability edge predictions of the source parser as a supervision signal in a self-training algorithm, thus enabling unsupervised training on the target language data. The method is feasible for discriminative and non-projective parsing, as well as multi-source and source-free transfer. Building on a framework introduced in Täckström et al. (2013), this paper for the first time demonstrates their effectiveness in the context of state-of-the-art neural dependency parsers, and their generalizability across parsing frameworks. Using English as the source language, we evaluate on eight distant and ten nearby languages (He et al., 2019). The single-source transfer variant (Section 2.1) outperforms previous methods by up to 11 % UAS, averaged over nearby languages. Extending the approach to multi-source transfer (Section 2.2) gives further gains of 2 % UAS and closes the performance gap against the state of the art on distant languages. In short, our contributions are:

1. A conceptually simple and highly flexible framework for unsupervised target language adaptation, which supports multi-source and source-free transfer, and can be employed with any pre-trained state-of-the-art arc-factored parser(s);
2. Generalisation of the method of Täckström et al. (2013) to state-of-the-art, non-projective dependency parsing with neural networks;
3. Up to 13 % UAS improvement over state-of-the-art models, considering nearby languages, and roughly equal performance over distant languages; and
4. Analysis of the impact of choice of source languages on multi-source transfer quality.

2 Supervision via Transfer

In our scenario of unsupervised cross-lingual parsing, we assume the availability of a pre-trained source parser, and unlabelled text in the target language. Thus, we aim to leverage this data such that our cross-lingual transfer parsing method out-performs direct transfer. One straightforward method is self-training where we use the predictions from the source parser as supervision to train the target parser. This method may yield decent performance as direct transfer is fairly good to begin with. However, we may be able to do better if we also consider a set of parse trees that have high probability under the source parser (cf. Fig. 1 for illustration).

If we assume that the source parser can produce a set of possible trees instead, then it is natural to use all of these trees as supervision signal for training. Inspired by Täckström et al. (2013), we formalise the method as follows. Given an unlabelled dataset $\{x_i\}_{i=1}^{n}$, the training loss can be expressed as

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log \sum_{y \in \tilde{Y}(x_i)} P_\theta(y|x_i) \qquad (1)$$

where θ is the target parser parameters and $\tilde{Y}(x_i)$ is the set of trees produced by the source parser. Note that $\tilde{Y}(x_i)$ must be smaller than the set of all trees spanning x (denoted as $\mathcal{Y}(x_i)$) because $\mathcal{L}(\theta) = 0$ otherwise. This training procedure is a form of self-training, and we expect that the target parser can learn the correct tree as it is likely to be included in $\tilde{Y}(x_i)$. Even if this is not the case, as long as the correct arcs occur quite frequently in $\tilde{Y}(x_i)$, we expect the parser to learn a useful signal.

We consider an arc-factored neural dependency parser where the score of a tree is defined as the sum of the scores of its arcs, and the arc scoring function is parameterised by a neural network. The probability of a tree is then proportional to its score.

Formally, this formulation can be expressed as

$$P_\theta(y|x) = \frac{\exp s_\theta(x, y)}{Z(x)} \quad (2)$$

$$s_\theta(x, y) = \sum_{(h,m)\in A(y)} s_\theta(x, h, m) \quad (3)$$

where $Z(x) = \sum_{y\in\mathcal{Y}(x)} \exp s_\theta(x, y)$ is the partition function, $A(y)$ is the set of head-modifier arcs in y, and $s_\theta(x, y)$ and $s_\theta(x, h, m)$ are the tree and arc scoring function respectively.

2.1 Single-Source Transfer

Here, we consider the case where a single pre-trained source parser is provided and describe how the set of trees is constructed. Concretely, for every sentence $x = w_1, w_2, \ldots, w_t$ in the target language data, using the source parser, the set of high probability trees $\tilde{Y}(x)$ is defined as the set of dependency trees that can be assembled from the high probability arcs set $\tilde{A}(x) = \bigcup_{m=1}^{t} \tilde{A}(x, m)$, where $\tilde{A}(x, m)$ is the set of high probability arcs whose dependent is w_m. Thus, $\tilde{Y}(x)$ can be expressed formally as

$$\tilde{Y}(x) = \{y | y \in \mathcal{Y}(x) \wedge A(y) \subseteq \tilde{A}(x)\}. \quad (4)$$

$\tilde{A}(x, m)$ is constructed by adding arcs (h, m) in order of decreasing arc marginal probability until their cumulative probability exceeds a threshold σ (Täckström et al., 2013). The predicted tree from the source parser is also included in $\tilde{Y}(x)$ so the chart is never empty. This prediction is simply the highest scoring tree. This procedure is illustrated in Fig. 1.

Since $\mathcal{Y}(x)$ contains an exponential number of trees, efficient algorithms are required to compute the partition function $Z(x)$, arc marginal probabilities, and the highest scoring tree. First, arc marginal probabilities can be computed efficiently with dynamic programming for projective trees (Paskin, 2001) and Matrix-Tree Theorem for the non-projective counterpart (Koo et al., 2007; McDonald and Satta, 2007; Smith and Smith, 2007). The same algorithms can also be employed to compute $Z(x)$. Next, the highest scoring tree can be obtained efficiently with Eisner's algorithm (Eisner, 1996) or the maximum spanning tree algorithm (McDonald et al., 2005; Chu and Liu, 1965; Edmonds, 1967) for the projective and non-projective cases, respectively.

The transfer is performed by initialising the target parser with the source parser's parameters and then fine-tuning it with the training loss in Eq. (1) on the target language data. Following previous works (Duong et al., 2015b; He et al., 2019), we also regularise the parameters towards the initial parameters to prevent them from deviating too much since the source parser is already good to begin with. Thus, the final fine-tuning loss becomes

$$\mathcal{L}'(\theta) = \mathcal{L}(\theta) + \lambda||\theta - \theta_0||_2^2 \quad (5)$$

where θ_0 is the initial parameters and λ is a hyperparameter regulating the strength of the L_2 regularisation. This single-source transfer strategy was introduced as ambiguity-aware self-training by Täckström et al. (2013). A difference here is that we regularise the target parser's parameters against the source parser's as the initialiser, and apply the technique to modern lexicalised state-of-the-art parsers. We refer to this transfer strategy as PPT hereinafter.

Note that the whole procedure of PPT can be performed even when the source parser is trained with monolingual embeddings. Specifically, given a source parser trained *only on monolingual embeddings*, one can align pre-trained target language word embeddings to the source embedding space using an offline cross-lingual alignment method (e.g., of Smith et al. (2017)), and use the aligned target embeddings with the source model to compute $\tilde{Y}(x)$. Thus, our method can be used with any pre-trained monolingual neural parser.

2.2 Multi-Source Transfer

We now consider the case where multiple pre-trained source parsers are available. To extend PPT to this multi-source case, we employ the ensemble training method from Täckström et al. (2013), which we now summarise. We define $\tilde{A}(x, m) = \bigcup_k \tilde{A}_k(x, m)$ where $\tilde{A}_k(x, m)$ is the set of high probability arcs obtained with the k-th source parser. The rest of the procedure is exactly the same as PPT. Note that we need to select one source parser as the main source to initialise the target parser's parameters with. Henceforth, we refer to this method as PPTX.

Multiple source parsers may help transfer better because each parser will encode different syntactic biases from the languages they are trained on. Thus, it is more likely for one of those biases to match that of the target language instead of using just a single source parser. However, multi-source transfer may also hurt performance if the languages have very

different syntax, or the source parsers are of poor quality, which can arise from poor quality cross-lingual word embeddings.

3 Experiments

3.1 Setup

We run our experiments on Universal Dependency Treebanks v2.2 (Nivre et al., 2018). We reimplement the self-attention graph-based parser of Ahmad et al. (2019a) that has been used with success for cross-lingual dependency parsing. Averaged over 5 runs, our reimplementation achieves 88.8 % unlabelled attachment score (UAS) on English Web Treebank using the same hyperparameters,[3] slightly below their reported 90.3 % result.[4] We select the run with the highest labelled attachment score (LAS) as the source parser. We obtain cross-lingual word embeddings with the offline transformation of Smith et al. (2017) applied to fastText pre-trained word vectors (Bojanowski et al., 2017). We include the universal POS tags as inputs by concatenating the embeddings with the word embeddings in the input layer. We acknowledge that the inclusion of gold POS tags does not reflect a realistic low-resource setting where gold tags are not available, which we discuss more in Section 3.3. We evaluate on 18 target languages that are divided into two groups, distant and nearby languages, based on their distance from English as defined by He et al. (2019).[5]

During the unsupervised fine-tuning, we compute the training loss over all trees regardless of projectivity (i.e. we use Matrix-Tree Theorem to compute Eq. (1)) and discard sentences longer than 30 tokens to avoid out-of-memory error. Following He et al. (2019), we fine-tune on the target language data for 5 epochs, tune the hyperparameters (learning rate and λ) on Arabic and Spanish using LAS, and use these values[6] for the distant and nearby languages, respectively. We set the threshold $\sigma = 0.95$ for both PPT and PPTX following Täckström et al. (2013). We keep the rest of the hyperparameters (e.g., batch size) equal to those of Ahmad et al. (2019a). For PPTX, unless otherwise stated, we consider a leave-one-out scenario where we use all languages except the target as the source language. We use the same hyperparameters as the English parser to train these non-English source parsers and set the English parser as the main source.

3.2 Comparisons

We compare PPT and PPTX against several recent unsupervised transfer systems. First, HE is a neural lexicalised DMV parser with normalising flow that uses a language modelling objective when fine-tuning on the unlabelled target language data (He et al., 2019). Second, AHMAD is an adversarial training method that attempts to learn language-agnostic representations (Ahmad et al., 2019b). Lastly, MENG is a constrained inference method that derives constraints from the target corpus statistics to aid inference (Meng et al., 2019). We also compare against direct transfer (DT) and self-training (ST) as our baseline systems.[7]

3.3 Results

Table 1 shows the main results. We observe that fine-tuning via self-training already helps DT, and by incorporating multiple high probability trees with PPT, we can push the performance slightly higher on most languages, especially the nearby ones. Although not shown in the table, we also find the PPT has up to 6x lower standard deviation than ST, which makes PPT preferrable to ST. Thus, we exclude ST as a baseline from our subsequent experiments. Our results seem to agree with that of Täckström et al. (2013) and suggest that PPT can also be employed for neural parsers. Therefore, it should be considered for target language adaptation if unlabelled target data is available. Comparing to HE (He et al., 2019), PPT performs worse on distant languages, but better on nearby languages. This finding means that if the target language has a closely related high-resource language, it may be better to transfer from that language as the source and use PPT for adaptation. Against AHMAD (Ahmad et al., 2019b), PPT performs better on 4 out of 6 distant languages. On nearby languages, the average UAS of PPT is higher, and the average LAS is on par. This result shows that leveraging unlabelled data for cross-lingual parsing without access to the source data is feasible. PPT also performs

[3] Reported in Table 4.

[4] UAS and LAS are reported excluding punctuation tokens.

[5] We exclude Japanese and Chinese based on Ahmad et al. (2019a), who reported atypically low performance on these two languages, which they attributed to the low quality of their cross-lingual word embeddings. In subsequent work they excluded these languages (Ahmad et al., 2019b).

[6] Reported in Table 5.

[7] ST requires significantly less memory so we only discard sentences longer than 60 tokens. Complete hyperparameter values are shown in Table 5.

Target	UAS							LAS				
	DT	ST	PPT	PPTX	HE	AHMAD	MENG	DT	ST	PPT	PPTX	AHMAD
fa	37.5	38.0	39.5	53.6	**63.2**	—	—	29.2	30.5	31.6	**44.5**	—
ar†	37.6	39.2	39.5	48.3	**55.4**	39.0	47.3	27.3	30.0	29.9	**38.5**	27.9
id	51.6	49.9	50.3	**71.9**	64.2	51.6	53.1	45.2	44.4	44.7	**59.0**	45.3
ko	35.1	37.1	**37.5**	34.6	37.0	34.2	37.1	16.6	**18.2**	18.0	16.1	16.1
tr	36.9	38.1	**39.2**	38.4	36.1	—	35.2	18.5	19.5	19.0	**20.6**	—
hi	33.7	34.7	34.0	36.4	33.2	37.4	**52.4**	25.4	26.6	26.4	**28.3**	28.0
hr	62.0	63.4	63.8	**71.9**	65.3	63.1	63.7	51.9	54.2	54.2	**61.2**	53.6
he	56.6	59.2	60.5	64.2	**64.8**	57.2	58.8	47.6	50.5	51.1	**53.9**	49.4
average	43.9	45.0	45.5	**52.4**	**52.4**	—	—	32.7	34.2	34.4	**40.3**	—
bg	77.7	80.0	81.2	**81.9**	73.6	79.7	79.7	66.2	68.9	70.0	**70.2**	68.4
it	77.9	79.7	81.4	**83.7**	70.7	80.7	82.0	71.1	74.0	75.5	**77.7**	75.6
pt	74.1	76.3	77.1	**81.0**	66.6	77.1	77.5	65.1	67.6	68.3	**70.6**	67.8
fr	74.8	77.5	78.6	**80.6**	67.7	78.3	79.1	68.1	71.7	72.8	**74.5**	73.3
es†	72.5	74.9	75.2	**78.3**	64.3	74.1	75.8	63.8	66.5	67.0	**69.2**	65.8
no	77.9	80.4	**81.2**	80.0	65.3	81.0	80.4	69.1	71.9	72.7	71.8	**73.1**
da	75.3	76.0	**77.3**	76.6	61.1	76.3	76.6	66.3	67.4	**68.6**	67.9	68.0
sv	78.9	80.5	**82.1**	81.0	64.4	80.4	80.5	71.1	72.7	74.2	72.7	**76.7**
nl	68.0	68.9	69.9	**74.4**	61.7	69.2	67.6	59.5	60.7	61.5	**65.4**	60.5
de	66.8	69.9	69.5	**74.1**	69.5	71.1	70.8	56.4	60.0	59.7	**63.5**	61.8
average	74.4	76.4	77.4	**79.1**	66.5	76.8	77.0	65.7	68.1	69.0	**70.3**	69.1

Table 1: Test UAS and LAS (avg. 5 runs) on distant (top) and nearby (bottom) languages, sorted from most distant (fa) to closest (de) to English. PPTX is trained in a leave-one-out fashion. The numbers for HE, AHMAD, and MENG are obtained from the corresponding papers, direct transfer (DT) and self-training (ST) are based on our own implementation. † indicates languages used for hyper-parameter tuning, and thus have additional supervision through the use of a labelled development set.

better than MENG (Meng et al., 2019) on 4 out of 7 distant languages, and slightly better on average on nearby languages. This finding shows that PPT is competitive to their constrained inference method.

Also reported in Table 1 are the ensemble results for PPTX, which are particularly strong. PPTX outperforms PPT, especially on distant languages with the average UAS and LAS absolute improvements of 7 % and 6 % respectively. This finding suggests that PPTX is indeed an effective method for multi-source transfer of neural dependency parsers. It also gives further evidence that multi-source transfer is better than the single-source counterpart. PPTX also closes the gap against the state-of-the-art adaptation of He et al. (2019) in terms of average UAS on distant languages. This result suggests that PPTX can be an option for languages that do not have a closely related high-resource language to transfer from.

Treebank Leakage The success of our cross-lingual transfer can be attributed in part to treebank leakage, which measures the fraction of dependency trees in the test set that are isomorphic to a tree in the training set (with potentially different words); accordingly these trees are not entirely unseen. Such leakage has been found to be a particularly strong predictor for parsing performance in monolingual parsing (Søgaard, 2020). Fig. 2 shows the relationship between treebank leakage and parsing accuracy, where the leakage is computed between the English training set as source and the target language's test set. Excluding outliers which are Korean and Turkish because of their low parsing accuracy despite the relatively high leakage, we find that there is a fairly strong positive correlation ($r = 0.57$) between the amount of

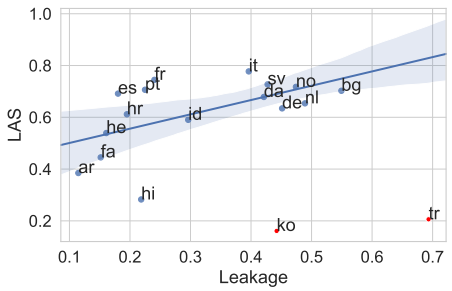

Figure 2: Relationship between treebank leakage and LAS for PPTX. Shaded area shows 95 % confidence interval. Korean and Turkish (in red) are excluded when computing the regression line.

leakage and accuracy. The same trend occurs with DT, ST, and PPT. This finding suggests that cross-lingual parsing is also affected by treebank leakage just like monolingual parsing is, which may present an opportunity to find good sources for transfer.

Use of Gold POS Tags As we explained in Section 3.1, we restrict our experiments to gold POS tags for comparison with prior work. However, the use of gold POS tags does not reflect a realistic low-resource setting where one may have to resort to automatically predicted POS tags. Tiedemann (2015) has shown that cross-lingual delexicalised parsing performance degrades when predicted POS tags are used. The degradation ranges from 2.9 to 8.4 LAS points depending on the target language. Thus, our reported numbers in Table 1 are likely to decrease as well if predicted tags are used, although we expect the decline is not as sharp because our parser is lexicalised.

3.4 Parsimonious Selection of Sources for PPTX

In our main experiment, we use all available languages as source for PPTX in a leave-one-out setting. Such a setting may be justified to cover as many syntactic biases as possible, however, training dozens of parses may be impractical. In this experiment, we consider the case where we can train only a handful of source parsers. We investigate two selections of source languages: (1) a representative selection (PPTX-REPR) which covers as many language families as possible and (2) a pragmatic selection (PPTX-PRAG) containing truly high-resource languages for which quality pre-trained parsers are likely to exist. We restrict the selections to 5 languages each. For PPTX-REPR, we use English, Spanish, Arabic, Indonesian, and Korean as source languages. This selection covers Indo-European (Germanic and Romance), Afro-Asiatic, Austronesian, and Koreanic language families respectively. We use English, Spanish, Arabic, French, and German as source languages for PPTX-PRAG. The five languages are classified as exemplary high-resource languages by Joshi et al. (2020). We exclude a language from the source if it is also the target language, in which case there will be only 4 source languages. Other than that, the setup is the same as that of our main experiment.[8]

We present the result in Fig. 3 where we also include the results for PPT, and PPTX with the leave-one-out setting (PPTX-LOO). We report only LAS since UAS shows a similar trend. We observe that both PPTX-REPR and PPTX-PRAG outperform PPT overall. Furthermore, on nearby languages except Dutch and German, both PPTX-REPR and PPTX-PRAG outperform PPTX-LOO, and PPTX-PRAG does best overall. In contrast, no systematic difference between the three PPTX variants emerges on distant languages. This finding suggests that instead of training dozens of source parsers for PPTX, training just a handful of them is sufficient, and a "pragmatic" selection of a small number of high-resource source languages seems to be an efficient strategy. Since pre-trained parsers for these languages are most likely available, it comes with the additional advantage of alleviating the need to train parsers at all, which makes our method even more practical.

Analysis on Dependency Labels Next, we break down the performance of our methods based on the dependency labels to study their failure and success patterns. Fig. 4 shows the UAS of DT, PPT, and PPTX-PRAG on Indonesian and German for select dependency labels.

Looking at Indonesian, PPT is slightly worse than DT in terms of overall accuracy scores (Table 1), and this is reflected across dependency labels. However, we see in Fig. 4 that PPT outperforms DT on `amod`. In Indonesian, adjectives follow the noun they modify, while in English the opposite is true in general. Thus, unsupervised target language adaptation seems able to address these kinds of discrepancy between the source and target language. We find that PPTX-PRAG outperforms both DT and PPT across dependency labels, especially on `flat` and `compound` labels as shown in Fig. 4. Both labels are related to multi-word expressions (MWEs), so PPTX appears to improve parsing MWEs in Indonesian significantly.

For German we find that both PPT and PPTX-PRAG outperform DT on most dependency labels, with the most notable gain on `nmod`, which appear in diverse, and often non-local relations in both languages many of which do not structurally translate, and fine-tuning improves performance as expected. Also, we see PPTX-PRAG significantly underperforms on `compound` while PPT is better than DT. German compounds are often merged into a single token, and self-training appears to alleviate over-prediction of such relations. The multi-source case may contain too much diffuse

[8] Hyperparameters are tuned; values are shown in Table 5.

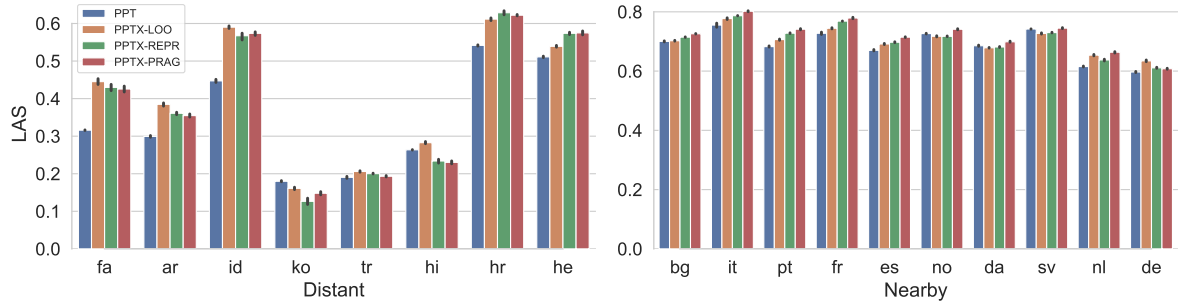

Figure 3: Comparison of selection of source languages for PPTX on distant and nearby languages, sorted from most distant (fa) to closest (de) to English. PPTX-LOO is trained in a leave-one-out fashion. PPTX-REPR uses the representative source language set, while PPTX-PRAG is adapted from five high-resource languages. A source language is excluded from the source if it is also the target language.

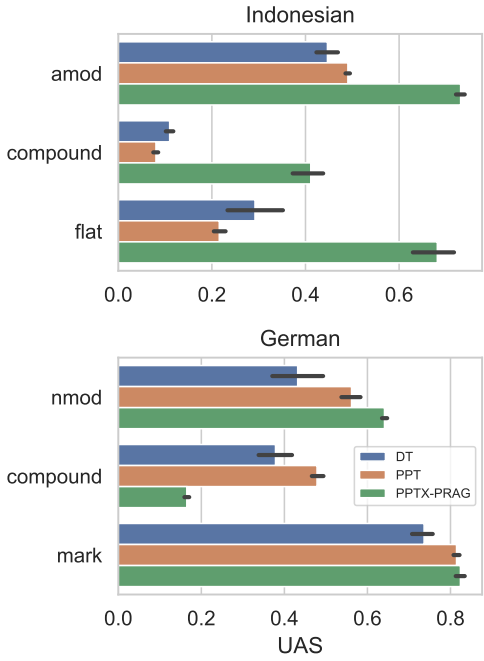

Figure 4: Comparison of direct transfer (DT), PPT, and PPTX-PRAG on select dependency labels of Indonesian (top) and German (bottom).

signal on `compound` and thus the performance is worse than that of DT. We find that PPT and PPTX improves over DT on `mark`, likely because markers are often used in places where German deviates from English by becoming verb-final (e.g., subordinate clauses). Both PPT and PPTX-PRAG seem able to learn this characteristic as shown by their performance improvements. This analysis suggests that the benefits of self-training depend on the syntactic properties of the target language.

Model	Target				AVG
	id	hr	fr	nl	
Non-projective					
DT	45.2	51.9	68.1	59.5	56.2
PPT	44.7	54.2	72.8	61.5	58.3
PPTX-PRAG	57.4	62.2	77.9	66.4	66.0
Projective					
DT	45.7	52.1	68.4	59.6	56.4
PPT	45.0	54.0	72.3	61.7	58.3
PPTX-PRAG	57.5	61.1	78.1	67.7	66.1

Table 2: Comparison of projective and non-projective direct transfer (DT), PPT, and PPTX-PRAG. Scores are LAS, averaged over 5 runs.

3.5 Effect of Projectivity

In this experiment, we study the effect of projectivity on the performance of our methods. We emulate a projective parser by restricting the trees in $\tilde{Y}(x)$ to be projective. In other words, the sum in Eq. (1) is performed only over projective trees. At test time, we search for the highest scoring projective tree. We compare DT, PPT, and PPTX-PRAG, and report LAS on Indonesian (id) and Croatian (hr) as distant languages, and on French (fr) and Dutch (nl) as nearby languages. The trend for UAS and on the other languages is similar. We use the dynamic programming implementation provided by `torch-struct` for the projective case (Rush, 2020). We find that it consumes more memory than our Matrix-Tree Theorem implementation, so we set the length cutoff to 20 tokens.[9]

Table 2 shows result of our experiment, which suggests that there is no significant performance difference between the projective and non-projective

[9]Hyperparameters are tuned; values are shown in Table 5.

Model	Target	
	ar	es
DT	28.1	64.1
PPT	30.8	67.3
PPTX^{EN5}	30.9	66.3
PPTX-PRAG^{S}	36.5	70.3
PPTX-PRAG	36.5	71.9

Table 3: Comparison of LAS on Arabic and Spanish on the development set, averaged over 5 runs. PPTX^{EN5} is PPTX with 5 English parsers as source, each trained on 1/5 size of the English corpus. PPTX-PRAG^{S} is PPTX with the pragmatic selection of source languages (PPTX-PRAG) but each source parser is trained on the same amount of data as PPTX^{EN5}.

variant of our methods. This result suggests that our methods generalise well to both projective and non-projective parsing. That said, we recommend the non-projective variant as it allows better parsing of languages that are predominantly non-projective. Also, we find that it runs roughly 2x faster than the projective variant in practice.

3.6 Disentangling the Effect of Ensembling and Larger Data Size

The effectiveness of PPTX can be attributed to at least three factors: (1) the effect of ensembling source parsers (*ensembling*), (2) the effect of larger data size used for training the source parsers (*data*), and (3) the diversity of syntactic biases from multiple source languages (*multilinguality*). In this experiment, we investigate to what extent each of those factors contributes to the overall performance. To this end, we design two additional comparisons: PPTX^{EN5} and PPTX-PRAG^{S}.

PPTX^{EN5} is PPTX with only English source parsers, where each parser is trained on 1/5 of the English training set. That is, we randomly split the English training set into five equal-sized parts, and train a separate parser on each. These parsers then serve as the source parsers for PPTX^{EN5}. Thus, PPTX^{EN5} has the benefit of ensembling but not data and multilinguality compared with PPT.

PPTX-PRAG^{S} is PPTX whose source language selection is the same as PPTX-PRAG, but each source parser is trained on the training data whose size is roughly the same as that of the training data of PPTX^{EN5} source parsers. In other words, the training data size is roughly equal to 1/5 of the English training set. To obtain this data, we randomly sub-sample the training data of each source language to the appropriate number of sentences. Therefore, PPTX-PRAG^{S} has the benefit of ensembling and multilinguality but not data.

Table 3 reports their LAS on the development set of Arabic and Spanish, averaged over five runs. We also include the results of PPTX-PRAG that enjoys all three benefits. We observe that PPT and PPTX^{EN5} perform similarly on Arabic, and PPTX^{EN5} has a slightly lower performance on Spanish. This result suggests a negligable effect of ensembling on performance. On the other hand, PPTX-PRAG^{S} outperforms PPTX^{EN5} remarkably, with approximately 6 % and 4 % LAS improvement on Arabic and Spanish respectively, showing that multilinguality has a much larger effect on performance than ensembling. Lastly, we see that PPTX-PRAG performs similarly to PPTX-PRAG^{S} on Arabic, and about 1.6 % better on Spanish. This result demonstrates that data size has an effect, albeit a smaller one compared to multilinguality. To conclude, the effectiveness of PPTX can be attributed to the diversity contributed through multiple languages, and not to ensembling or larger source data sets.

4 Related Work

Cross-lingual dependency parsing has been extensively studied in NLP. The approaches can be grouped into two main categories. On the one hand, there are approaches that operate on the data level. Examples of this category include annotation projection, which aims to project dependency trees from a source language to a target language (Hwa et al., 2005; Li et al., 2014; Lacroix et al., 2016; Zhang et al., 2019); and source treebank reordering, which manipulates the source language treebank to obtain another treebank whose statistics approximately match those of the target language (Wang and Eisner, 2018; Rasooli and Collins, 2019). Both methods have no restriction on the type of parsers as they are only concerned with the data. Transferring from multiple source languages with annotation projection is also feasible (Agić et al., 2016).

Despite their effectiveness, these data-level methods may require access to the source language data, hence are unusable when it is inaccessible due to privacy or legal reasons. In such source-free transfer, only a model pre-trained on the source language data is available. By leveraging parallel data, annotation projection is indeed feasible without ac-

cess to the source language data. That said, parallel data is limited for low-resource languages or may have a poor domain match. Additionally, these methods involve training the parser from scratch for every new target language, which may be prohibitive.

On the other hand, there are methods that operate on the model level. A typical approach is direct transfer (aka., zero-shot transfer) which trains a parser on source language data, and then directly uses it to parse a target language. This approach is enabled by the shared input representation between the source and target language such as POS tags (Zeman and Resnik, 2008) or cross-lingual embeddings (Guo et al., 2015; Ahmad et al., 2019a). Direct transfer supports source-free transfer and only requires training a parser once on the source language data. In other words, direct transfer is unsupervised as far as target language resources.

Previous work has shown that unsupervised target language adaptation outperforms direct transfer. Recent work by He et al. (2019) used a neural lexicalised dependency model with valence (DMV) (Klein and Manning, 2004) as the source parser and fine-tuned it in an unsupervised manner on the unlabelled target language data. This adaptation method allows for source-free transfer and performs especially well on distant target languages. A different approach is proposed by Meng et al. (2019), who gathered target language corpus statistics to derive constraints to guide inference using the source parser. Thus, this technique also allows for source-free transfer. A different method is proposed by Ahmad et al. (2019b) who explored the use of unlabelled data from an auxiliary language, which can be different from the target language. They employed adversarial training to learn language-agnostic representations. Unlike the others, this method can be extended to support multi-source transfer. An older method is introduced by Täckström et al. (2013), who leveraged ambiguity-aware training to achieve unsupervised target language adaptation. Their method is usable for both source-free and multi-source transfer. However, to the best of our knowledge, its use for neural dependency parsing has not been investigated. Our work extends theirs by employing it for the said purpose.

The methods of both He et al. (2019) and Ahmad et al. (2019b) have several limitations. The method of He et al. (2019) requires the parser to be generative and projective. Their generative parser is quite impoverished with an accuracy that is 21 points lower than a state-of-the-art discriminative arc-factored parser on English. Thus, their choice of generative parser may constrain its potential performance. Furthermore, their method performs substantially worse than direct transfer on nearby target languages. Because of the availability of resources such as Universal Dependency Treebanks (Nivre et al., 2018), it is likely that a target language has a closely related high-resource language which can serve as the source language. Therefore, performing well on nearby languages is more desirable pragmatically. On top of that, it is unclear how to employ this method for multi-source transfer. The adversarial training method of Ahmad et al. (2019b) does not suffer from the aforementioned limitations but is unusable for source-free transfer. That is, it assumes access to the source language data, which may not always be feasible due to privacy or legal reasons.

5 Conclusions

This paper presents a set of effective, flexible, and conceptually simple methods for unsupervised cross-lingual dependency parsing, which can leverage the power of state-of-the-art pre-trained neural network parsers. Our methods improve over direct transfer and strong recent unsupervised transfer models, by using source parser uncertainty for implicit supervision, leveraging only unlabelled data in the target language. Our experiments show that the methods are effective for both single-source and multi-source transfer, free from the limitations of recent transfer models, and perform well for non-projective parsing. Our analysis shows that the effectiveness of the multi-source transfer method is attributable to its ability to leverage diverse syntactic signals from source parsers from different languages. Our findings motivate future research into advanced methods for generating informative sets of candidate trees given one or more source parsers.

Acknowledgments

We thank the anonymous reviewers for the useful feedback. A graduate research scholarship is provided by Melbourne School of Engineering to Kemal Kurniawan.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019a. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019b. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Miryam de Lhoneux. 2019. *Linguistically Informed Neural Dependency Parsing for Typologically Diverse Languages*. Ph.d. thesis, Uppsala University.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on 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.

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.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 478–485.

Terry Koo, Amir Globerson, Xavier Carreras, and Michael Collins. 2007. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150.

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1058–1063.

Zhenghua Li, Min Zhang, and Wenliang Chen. 2014. Soft cross-lingual syntax projection for dependency parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 783–793.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530.

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.

Ryan McDonald and Giorgio Satta. 2007. On the complexity of non-projective data-driven dependency parsing. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 121–132.

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.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2018. Universal dependencies 2.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Mark A Paskin. 2001. *Cubic-time parsing and learning algorithms for grammatical bigram models*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164.

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.

Alexander Rush. 2020. Torch-struct: Deep structured prediction library. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*, pages 335–342.

David A. Smith and Noah A. Smith. 2007. Probabilistic models of nonprojective dependency trees. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 132–140.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*.

Anders Søgaard. 2020. Some languages seem easier to parse because their treebanks leak. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2765–2770.

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.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Dingquan Wang and Jason Eisner. 2018. Synthetic data made to order: The case of parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1337.

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*.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed treebank. 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 996–1005.

A Hyperparameter values

Here we report the hyperparameter values for experiments presented in the paper. Table 4 shows the hyperparameter values of our English source parser explained in Section 3.1. Table 5 reports the tuned hyperparameter values for our experiments shown in Table 1, Fig. 3, and Table 2.

Hyperparameter	Value
Sentence length cutoff	100
Word embedding size	300
POS tag embedding size	50
Number of attention heads	10
Number of Transformer layers	6
Feedforward layer hidden size	512
Attention key vector size	64
Attention value vector size	64
Dropout	0.2
Dependency arc vector size	512
Dependency label vector size	128
Batch size	80
Learning rate	10^{-4}
Early stopping patience	50

Table 4: Hyperparameter values of the source parser.

Hyperparameter	Value	
	Nearby	Distant
ST		
Sentence length cutoff	60	60
Learning rate	5.6×10^{-4}	3.7×10^{-4}
L2 coefficient (λ)	3×10^{-4}	2.8×10^{-4}
PPT		
Learning rate	3.8×10^{-5}	2×10^{-5}
L2 coefficient (λ)	0.01	0.39
PPTX/PPTX-LOO		
Learning rate	2.1×10^{-5}	5.9×10^{-5}
L2 coefficient (λ)	0.079	1.2×10^{-4}
PPTX-REPR		
Learning rate	1.7×10^{-5}	9.7×10^{-5}
L2 coefficient (λ)	4×10^{-4}	0.084
PPTX-PRAG		
Learning rate	4.4×10^{-5}	8.5×10^{-5}
L2 coefficient (λ)	2.7×10^{-4}	2.8×10^{-5}
Projective PPT		
Sentence length cutoff	20	20
Learning rate	10^{-4}	10^{-4}
L2 coefficient (λ)	7.9×10^{-4}	7.9×10^{-4}
Projective PPTX-PRAG		
Sentence length cutoff	20	20
Learning rate	9.4×10^{-5}	9.4×10^{-5}
L2 coefficient (λ)	2.4×10^{-4}	2.4×10^{-4}

Table 5: Hyperparameter values of ST, PPT, PPTX, PPTX-REPR, PPTX-PRAG, projective PPT, and projective PPTX-PRAG. Sentence length cutoff for PPT, PPTX, PPTX-REPR, and PPTX-PRAG is 30, as explained in Section 3.1.