Cross-Lingual Dependency Parsing via Self-Training

Meishan Zhang[1] and **Yue Zhang**[2*]
1. School of New Media and Communication, Tianjin University, China
2. Institute of Advanced Technology, Westlake Institute for Advanced Study
mason.zms@gmail.com,
zhangyue@westlake.edu.cn,

Abstract

Recent advances of multilingual word representations weaken the input divergences across languages, making cross-lingual transfer similar to the monolingual cross-domain and semi-supervised settings. Thus self-training, which is effective for these settings, could be possibly beneficial to cross-lingual as well. This paper presents the first comprehensive study for self-training in cross-lingual dependency parsing. Three instance selection strategies are investigated, where two of which are based on the baseline dependency parsing model, and the third one adopts an auxiliary cross-lingual POS tagging model as evidence. We conduct experiments on the universal dependencies for eleven languages. Results show that self-training can boost the dependency parsing performances on the target languages. In addition, the POS tagger assistant instance selection can achieve further improvements consistently. Detailed analysis is conducted to examine the potentiality of self-training in-depth.

1 Introduction

Cross-lingual dependency parsing has received increasing attention in recent years (Hwa et al., 2005; McDonald et al., 2011; Tiedemann et al., 2014; Guo et al., 2016a; Agić et al., 2016; Schlichtkrull and Søgaard, 2017; Rasooli and Collins, 2017; Rasooli and Collins, 2019; Zhang et al., 2019), which aims to parse target low-resource language with the supervision of resource-rich language. In this paper, we focus on the unsupervised setting (Ma and Xia, 2014; Guo et al., 2015; Rasooli and Collins, 2015; Tiedemann and Agić, 2016; Agić et al., 2016; Schlichtkrull and Søgaard, 2017; Ahmad et al., 2019), where no targeted dependency treebank is given.

Recent advances of multilingual word representations (Smith et al., 2017; Chen and Cardie, 2018; Mulcaire et al., 2019; Pires et al., 2019; Lample and Conneau, 2019; Wang et al., 2019; Wu and Dredze, 2019) has substantially promoted cross-lingual dependency parsing, especially serving as the basic input features for model transfer methods (Guo et al., 2016a; Schuster et al., 2019; Wang et al., 2019). They reduce the input divergences between languages significantly. As a result, the cross-lingual transfer learning setting can be considered highly similar to the monolingual semi-supervised and cross-domain settings. In light of this, the self-training strategy, which is widely adopted for cross-domain parsing (Reichart and Rappoport, 2007; Rush et al., 2012; Yu et al., 2015; Saito et al., 2017; More et al., 2019), can be potentially applicable for cross-lingual dependency parsing as well. However, relatively little work has demonstrated the effects of this potential method.

Instance selection for the next-round training is the key to self-training (Mihalcea, 2004; McClosky et al., 2006a; McClosky et al., 2006b; He and Zhou, 2011; Artetxe et al., 2018), which requires a certain criterion to rank the automatic outputs from the baseline model (Goldwasser et al., 2011; Yu et al., 2015; Zou et al., 2019). Such criteria are typically derived from the baseline model directly, for example, the prediction probability (Zou et al., 2018), and the delta probability between the final output and the second-best candidate output (Yu et al., 2015). Here we hypothesize that we can improve the performance of self-training by an auxiliary task which is highly corrective with the target task. A natural auxiliary task for cross-lingual dependency parsing is universal Part-of-speech (POS) tagging.

*Corresponding author.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 808-820, Hainan, China, October 31 - Novermber 1, 2020.

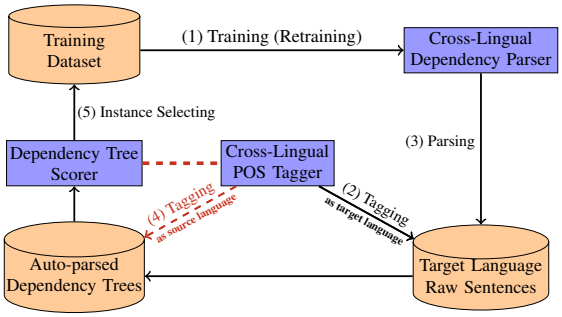


Figure 1: The overall architecture of self-train, where cross-lingual POS tagging is used to assist the instance selection in this work.

POS tags have served as one basic feature for dependency parsing (Zhang and Nivre, 2011; Kiperwasser and Goldberg, 2016; Dozat and Manning, 2016), and universal POS tags have been one important feature source for cross-lingual dependency parsing (McDonald et al., 2011; Petrov et al., 2012). The construction of a POS tagging corpus for a target language has a much lower cost than that of a dependency treebank, leading to the majority work of cross-lingual dependency parsing assuming gold-standard POS tags as inputs (Guo et al., 2016a; Rasooli and Collins, 2015; Tiedemann and Agić, 2016; Rasooli and Collins, 2017). We assume that a POS tag training corpus for the target language is available.

Based on the above settings, we investigate the capacity of self-training for cross-lingual dependency parsing empirically. Taking the BiAffine parser (Dozat and Manning, 2016) as the major architecture and enriching the model with multilingual BERT word representations (Devlin et al., 2019), we evaluate two widely-adopted instance selection strategies of self-training, and further propose a POS tagging guided criterion, which is illustrated in Figure 1. In particular, a supervised cross-lingual POS tagging model is trained to guide the instance selection in self-training, which uses a language-aware parameter generation network (PGN) (Platanios et al., 2018; Jia et al., 2019) for language switching. Our goal is to choose the target language sentences for which the POS tag outputs change relatively little when they are intentionally marked as source language sentences.

We conduct experiments on the Universal Dependencies (McDonald et al., 2013; Nivre et al., 2016) to study the effectiveness of self-training. English is selected as the source language, and eleven target languages belonging to four different families are investigated. Results show that self-training is an effective way for cross-lingual dependency parsing, boosting the dependency parsing performances of all selected target languages. In addition, POS-guided instance selection achieves further improvements. Finally, we conduct detailed analysis to understand the effectiveness of our self-training methods on four representative languages, one for each language family. All codes and datasets will be released publicly available on `https://github.com/zhangmeishan/selftraining` for research purpose under Apache License 2.0.

2 Related Work

Existing work on cross-lingual dependency parsing can be classified into two categories, namely model transferring and annotation projection, respectively. The first aims to train a dependency parsing model on the source-language treebank (McDonald et al., 2011; Guo et al., 2016a; Guo et al., 2016b), and then use it for target languages directly. Language independent features are exploited in order to minimize the gapping between the source and target languages, including multilingual word clusters (Täckström et al., 2012), word embeddings (Guo et al., 2015; Duong et al., 2015b; Duong et al., 2015a; Zhang and Barzilay, 2015; Guo et al., 2016b; Ammar et al., 2016; Wick et al., 2016; de Lhoneux et al., 2018), universal POS tags (McDonald et al., 2011; McDonald et al., 2013) and multilingual contextualized word representations (Wang et al., 2019; Wu and Dredze, 2019). In this work, we build our baselines with multilingual BERT, which has demonstrated state-of-the-art effort for cross-lingual model transferring (Wang et al., 2019).

Annotation projection aims to construct an automatic target-language dependency treebank by projecting

source language dependencies into target language sentences (Hwa et al., 2005; Ganchev et al., 2009). It relies on a parallel corpus, where source dependencies can be obtained by a source parser (Ma and Xia, 2014; Rasooli and Collins, 2015; Xiao and Guo, 2015; Agić et al., 2016; Schlichtkrull and Søgaard, 2017) or manually annotated (Tiedemann et al., 2014; Tiedemann, 2015; Tiedemann and Agić, 2016). Word-level alignment between sentence pairs has been used to project source dependencies into the target sentences. Our self-training strategy is similar in constructing automatic training datasets for target languages, while the key idea is significantly different.

Self-training has been shown effective for a number of NLP tasks (Mihalcea, 2004; McClosky et al., 2006a; Sagae, 2010; Goldwasser et al., 2011; He and Zhou, 2011; Artetxe et al., 2018). For dependency parsing, Rush et al. (2012) show that it fails to improve the performance under a supervised setting. Several studies have demonstrated its effectiveness on neural dependency parsing under the fully supervised multilingual setting (Rybak and Wróblewska, 2018). Lightly supervised learning and cross-domain adaption are more successful settings for self-training (McClosky et al., 2006b; Reichart and Rappoport, 2007; Rush et al., 2012; Yu et al., 2015; More et al., 2019). Our work applies self-training in the unsupervised cross-lingual setting. There is only one work of a similar setting. Rasooli and Collins (2017) add a number of auto-parsed outputs to enlarge the training dataset as an auxiliary technique. Their auto labeling is limited to the small-scale raw corpora with gold-standard POS tags, obtaining much smaller improvements than our work. To our knowledge, we are the first work to study self-training systematically.

3 Models

In this section, we describe the dependency parsing and POS tagging models, and the key details which would be used in the self-training.

3.1 Dependency Parsing

We use the BiAffine dependency parsing model (Dozat and Manning, 2016) as the baseline parser, adapting it for cross-lingual parsing with multilingual BERT inputs (Devlin et al., 2019).

Input. An input sentence $w_1 \cdots w_n$ is fed directly into a pretrained multilingual BERT module. BERT would split each word into pieces. We adopt averaged pooling to obtain word-level representations from the piece-level outputs. The top-k layer outputs of the BERT are used, which are combined by a parameterized scalar vector into a single representation layer.[1] Finally, we obtain word-level representations $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$ by this process.

Encoder. The BiAffine dependency parsing simply adopts a three-layer BiLSTM as encoder, which can be formalized as:

$$\boldsymbol{h}_1^l \cdots \boldsymbol{h}_n^l = \text{BiLSTM}(\boldsymbol{h}_1^{l-1} \cdots \boldsymbol{h}_n^{l-1}), \tag{1}$$

where $l = \{1, 2, 3\}$, $\boldsymbol{h}_1^0 \cdots \boldsymbol{h}_n^0 = \boldsymbol{x}_1 \cdots \boldsymbol{x}_n$, and $\boldsymbol{h}_1^3 \cdots \boldsymbol{h}_n^3$ is our desired outputs.

Decoder. The BiAffine operation is used to calculate head and dependency label scores for each sentential word. Take head prediction as an example. First, two MLP layers are used to obtain the features for a word as head ($\boldsymbol{h}_1^{\text{head}} \cdots \boldsymbol{h}_n^{\text{head}}$) and child ($\boldsymbol{h}_1^{\text{child}} \cdots \boldsymbol{h}_n^{\text{child}}$), respectively. Then for each word w_i, we find its head word by calculating:

$$\boldsymbol{s}(w_i^{\curvearrowleft} w_j) = \text{BiAffine}(\boldsymbol{h}_i^{\text{child}}, \boldsymbol{h}_j^{\text{head}}), \tag{2}$$

where $j \in [1, n]/\{i\}$, and the highest-scored j is selected as the head for word w_i. For dependency relation prediction, we simply extend the scale $\boldsymbol{s}(w_i^{\curvearrowleft} w_j)$ into a vector $\boldsymbol{s}^{\text{rel}}(w_i^{\curvearrowleft} w_j)$, whose dim size equals the relation size. After the head word j is specified, we obtain the dependency relation label by the highest-scored index.

Dependency Probability. The probability for each dependency arc will be used as the confidence score in self-training. For each sentential word w_i, the probability of a given head j is calculated by:

$$p(w_i^{\curvearrowleft} w_j) = \frac{\exp(\boldsymbol{s}(w_i^{\curvearrowleft} w_j))}{\sum_{k \in [1,n]/\{i\}} \exp(\boldsymbol{s}(w_i^{\curvearrowleft} w_k))}. \tag{3}$$

[1] In this work, we set $k = 6$ and freeze BERT parameters according to the preliminary experiments.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 808-820, Hainan, China, October 31 - Novermber 1, 2020.

The probability is computed in terms of words since the BiAffine decoder classifies heads at the word level. The conditional dependency relation probability $p(r_i|w_i, h_i)$ is computed similarly by softmax over $\boldsymbol{s}^{\text{rel}}(w_i \curvearrowright w_j)$. The reader is referred to as Dozat and Manning (2016) for more details.

3.2 POS Tagging

POS Tagging is exploited for two purposes related to self-training. On the one hand, we produce automatic POS tag inputs for automatic dependency parsing, as it is impractical to assume a very large corpus with gold-standard POS tags. On the other hand, we use the tagging model to rank auto-parsed dependency trees for instance selection. Here, we introduce the POS tagging model in detail, which is adapted from a typical BiLSTM POS tagger (Huang et al., 2015; Plank et al., 2016).

Input. Given a sentence $w_1 \cdots w_n$, we obtain $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$ by going through a multilingual BERT module, which is exactly the same as that of the dependency parsing model. The details can be found in the input part of Section 3.1 directly.

Encoder. For the encoder, we exploit PGN-BiLSTM (Jia et al., 2019) instead of a standard BiLSTM, taking the language ID as input to choose parameters for the BiLSTM module, which enables the model better capture the language differences.

For convenience, we formalize the standard BiLSTM by:

$$\boldsymbol{h}_1 \cdots \boldsymbol{h}_n = \text{BiLSTM}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n, \boldsymbol{V}), \tag{4}$$

where $\boldsymbol{V}$ denotes the flattened equivalent of all the BiLSTM parameters $\{\boldsymbol{W}_1 \cdots \boldsymbol{W}_K\}$. $\boldsymbol{V}$ can be implemented by $\boldsymbol{V} = \text{Vec}(\boldsymbol{W}_1) \oplus \cdots \oplus \text{Vec}(\boldsymbol{W}_K)$, where $\text{Vec}(\cdot)$ indicates vectorizing to reshape tensors into vectors, and $\oplus$ denotes concatenation.

In PGN-BiLSTM, we produce $\boldsymbol{V}$ dynamically according to the input language ID. Formally, the PGN-BiLSTM can be formalized as:

$$\begin{aligned} \boldsymbol{h}_1 \cdots \boldsymbol{h}_n &= \text{PGN-BiLSTM}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n, \boldsymbol{e}_{\text{lg}}) \\ &= \text{BiLSTM}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n, \boldsymbol{V}_{\text{lg}}), \\ &= \text{BiLSTM}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n, \boldsymbol{W}_{\text{pgn}}\boldsymbol{e}_{\text{lg}}), \end{aligned} \tag{5}$$

where $\boldsymbol{e}_{\text{lg}}$ is the embedding of the input language ID, and $\boldsymbol{W}_{\text{pgn}}$ is a meta model parameter of PGN-BiLSTM. In this way, we obtain different encoder parameters when the input language ID changes.

Decoder. Finally, the decoder consists of a single MLP layer:

$$\boldsymbol{o}_1 \cdots \boldsymbol{o}_n = \text{MLP}(\boldsymbol{h}_1 \cdots \boldsymbol{h}_n), \tag{6}$$

which is used to score all POS candidates directly for each word. The highest-scored tag index of each $\boldsymbol{o}_i$ is the final POS predictions.[2]

POS Probability. We also need to calculate POS probabilities for self-training. This is conducted straightforwardly by softmax since word-level prediction is used in our POS tagging model:

$$p(t|w_i, \text{lg}) = \frac{\exp(\boldsymbol{o}_{i,t})}{\sum \exp(\boldsymbol{o}_{i,*})}, \tag{7}$$

where t is the desired tag for word w_i.

4 Self-Training

The self-training framework for cross-lingual dependency parsing is as follows. First, a cross-lingual dependency parser (Section 3.1) trained on a source language corpus is used to parse the raw corpus of a target language. In particular, POS tags of the raw corpus are produced by a supervised cross-lingual POS tagger (Section 3.2). Next, we select a number of auto-parsed dependency trees from the outputs, and use them as the extra corpus to enhance the dependency parser. Instance selection is a key factor to the performance of self-training. We investigate two instance selection strategies based on the baseline dependency parser, and further suggest another alternative by using the cross-lingual POS tagger.

[2]We do not exploit CRF as its final impact on self-training is marginal while introduces addition calculation cost.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 808-820, Hainan, China, October 31 - Novermber 1, 2020.

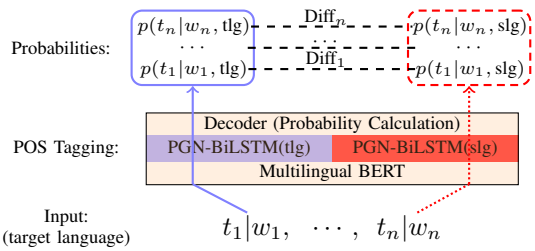


Figure 2: Illustration of the POS tagging guided instance selection, where the inner structures of the POS tagging model is described in Section 3.2, tlg and slg denote the target and source languages, respectively.

4.1 Strategies based on Dependency Parsing

Prediction Probability. The prediction probability is a widely-adopted strategy for instance selection in self-training (Yu et al., 2015; Zou et al., 2019), where auto-parsed dependency trees are ranked according to their tree probabilities, and the top probability trees are used for next-round training. Given a sentence $w_1 \cdots w_n$, assuming the output heads by our dependency parsing model are $h_1 \cdots h_n$, we calculate the score of the output dependency tree by the following formula:

$$s_{\text{prob}} = \prod_{i=1}^{n} p(w_i \curvearrowright w_{h_i}), \tag{8}$$

where $p(w_i \curvearrowright w_{h_i})$ is defined by Formula 3, which can be regarded as the confidence value of the current dependency arc.[3] We refer to this strategy as `prob` for simplicity.

Delta Probability. The second strategy is to use the delta value of the probabilities between the output head and the second-best head for each sentential word (Mejer and Crammer, 2012; Yu et al., 2015), where auto-parsed trees with larger delta values are selected for self-training.[4] For the sentence $w_1 \cdots w_n$, where the output heads and the second-best heads are $h_1 \cdots h_n$ and $h'_1 \cdots h'_n$, respectively, the selection score is defined by:

$$s_{\text{delta}} = \prod_{i=1}^{n} \left(p(w_i \curvearrowright w_{h_i}) - p(w_i \curvearrowright w_{h'_i})\right). \tag{9}$$

Note that there are cases where the final output head is not the highest-probability head because of the tree constraints, which are excluded directly. We use `delta` to denote this method for short.

4.2 POS Tagging Enhanced Criterion

Ranking the output sentences from the cross-lingual dependency parsing model itself may be biased, as it captures little knowledge on the differences between the source and target languages. Instead, the cross-lingual POS tagging model can offer such information, since it learns a universal model from the gold-standard training corpora of both the source and target languages. In addition, POS tagging is closely related to dependency parsing because they are both syntax-oriented, but POS tagging is much more light-weighted than dependency parsing, which makes our method more feasible in practice. Our goal is to select a target sentence which behaves highly similar across languages. We use these sentences to bridge the syntactic knowledge from the source into the target. Figure 2 illustrates the idea of the confidence computation strategy in detail.

Formally, given a target language sentence $w_1 \cdots w_n$, we first go through POS tagging as introduced in Section 3.2, feeding the target language ID into the PGN-BiLSTM encoder and computing the POS tagging probabilities of the best predictions $t_1 \cdots t_n$ at the word level by Equation 7. Then we compute another set of POS tagging probabilities by using the source language ID instead, feeding it into the PGN-BiLSTM encoder and computing the POS tagging probabilities of $t_1 \cdots t_n$. The process can be

[3] We do not use the relation probability for simplicity and meanwhile more importantly because it brings little influence.

[4] This is a simplified version of Yu et al. (2015).

regarded as by intentionally treating the target language sentence as a source language sentence. Finally, we obtain the confidence value for each sentence by:

$$\begin{aligned} \text{Diff}_i &= \|p(t_i|w_i, \text{tlg}) - p(t_i|w_i, \text{slg})\|, \\ s_{\text{pos}} &= \prod_{i=1}^{n}(1 - \text{Diff}_i), \end{aligned} \tag{10}$$

where the first equation indicates the language gaps, and the sentences with smaller gaps are chosen for self-training. We use `pos` to denote it for short.

4.3 Confidence-Aware Training of Dependency Parsing

Although with relatively high quality, the selected auto-parsed trees can nevertheless include noise. In order to address the influence of the noise, we introduce the confidence-aware training for the cross-lingual dependency parsing. The idea is inspired by Li et al. (2014), who solve parse ambiguities for monolingual self-training.

The standard training objective of the dependency parsing model mentioned in Section 3.1 is a cross-entropy loss over the dependency trees in the training corpus. Given a sentence $w_1 \cdots w_n$ and the corresponding dependency structure $(h_1, r_1) \cdots (h_n, r_n)$, where h and r indicate the head and dependency relation, respectively, the loss function is defined as follows:

$$\mathcal{L} = -\frac{\sum \log p(h_i, r_i|w_i)}{n}, \tag{11}$$

where $p(h_i, r_i|w_i) = p(w_i^{\frown} w_{h_i})p(r_i|w_i, h_i)$.

We use the word-level confidence values to regularize the loss function, which is defined by:

$$\mathcal{L}_{\text{conf}} = -\frac{\sum \tilde{p}(w_i^{\frown} w_{h_i}) \log p(h_i, r_i|w_i)}{n}, \tag{12}$$

where $\tilde{p}(w_i^{\frown} w_{h_i})$ is the confidence, defined by the dependency probability obtained from the original baseline dependency parsing model.

In particular, when the training corpus of the source and target languages is mixed to train a target language parser, we adopt a hyper-parameter α as the word-level confidence to rescale all the source language dependencies.

5 Experiments

5.1 Data and Settings

We conduct experiments on the Google Universal Dependency Treebanks (v2.2) (McDonald et al., 2013; Nivre et al., 2016) to verify the effectiveness of our models.[5] We adopt English as the source language. and choose eleven target languages, including German (de), Dutch (nl) and Swedish (sv) of the IE.Germanic family,[6] Spanish (es), French (fr) and Portuguese (pt) of the IE.Romance family, Polish (pl), Slovak (sk) and Slovenian (sl) of the IE.Romance family, and Estonian (et) and Finnish (fi) of the Uralic family. For each language, we use the same treebank type as Wang et al. (2019).[7]

We collect 500,000 raw sentences for each target language, respectively. The raw sentences are all selected from the Europarl v8 parallel corpus, which are download from the OPUS website directly. These sentences are already tokenized by the OPUS. We exclude the sentences shorter than 5 words or longer than 100 words, and then randomly sample 500,000 from the remaining.

For dependency parsing, we train models on the source English dataset and the auto-parsed dependency trees produced by self-training. During evaluation, gold POS tags are used as inputs on the test datsets for

[5] http://hdl.handle.net/11234/1-2837

[6] English also belongs to this family.

[7] The data statistics are omitted due to the space limitation.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 808-820, Hainan, China, October 31 - Novermber 1, 2020.

Model.	IE.Germanic			IE.Romance			IE.Slavic			Uralic		AVG
	de	nl	sv	es	fr	pt	pl	sk	sl	et	fi	
Source Only												
baseline	75.31	75.22	81.35	78.35	81.51	78.84	79.80	72.08	72.22	69.30	72.15	76.01
Target Only												
prob	76.44‡	76.55‡	82.40‡	77.20↓	81.86	78.45↓	80.13	72.76	73.34‡	71.05	72.59	76.62
delta	76.85‡	76.29‡	82.67‡	77.31↓	82.11	78.07↓	80.22	72.60	73.48‡	71.24‡	72.51	76.67
pos	**77.52‡**	**76.74‡**	**83.20‡**	**78.54**	**82.35‡**	**79.17**	**80.85‡**	**73.32‡**	**73.71‡**	**71.64‡**	**73.51‡**	**77.32**
Δ(pos)	+2.21	+1.52	+1.85	+0.19	+0.84	+0.33	+1.05	+1.24	+1.49	+2.34	+1.36	+1.31
Standard Self-Training (Source + Target)												
prob	78.00‡	76.68‡	83.06‡	78.41	82.37‡	79.05	80.38	73.13‡	74.04‡	71.39‡	73.13‡	77.24
delta	77.85‡	76.54‡	83.23‡	78.53	82.14	79.52‡	80.17	73.37‡	74.09‡	71.50‡	73.22‡	77.29
pos	**78.45‡**	**77.22‡**	**83.58‡**	**79.42‡**	**82.80‡**	**80.01‡**	**80.70‡**	**73.74‡**	**74.21‡**	**72.04‡**	**73.94‡**	**77.83**
Δ(pos)	+3.14	+2.00	+2.23	+1.07	+1.29	+1.17	+0.90	+1.66	+1.99	+2.74	+1.79	+1.82

Table 1: Final UAS results, where the $\Delta(\cdot)$ rows show the improvements over the corresponding baseline without self-training, the negative results are marked with ↓, the results marked with ‡ denote that the p-value is less than 0.001 compared with the baseline by using the pairwise t-test.

all target languages, following the majority of the previous studies. We adopt the unlabeled attachment score (UAS) as the major evaluation metric (excluding the punctuations).[8]

For POS tagging, we train models on the combined dataset of the source English training corpus and the test corpus of each target language. Since gold-standard POS tags are already given as inputs for dependency parsing, it is fair and reasonable to adopt this setting. The POS tagging model is also used to tag raw corpus of the self-training for each language, which is a pre-requisite step for dependency parsing since no POS tag exists in the collected large-scale raw corpus.

There are several hyper-parameters in the neural dependency parsing and POS tagging models. We set them empirically according to previous work. For the input multilingual BERT, we exploit the BERT-Base Multilingual Cased version, where the output dimension size is 768.[9] The POS tag embedding size of the dependency parsing model is 100. The language embedding size of the POS tagging model is 4. The hidden sizes of various BiLSTMs for both parsing and tagging are all 400, and the hidden sizes of the two MLP layers in the dependency parsing model are both 600.

For training, we exploit batch learning with a batch size of 200 and Adam with a learning ratio of 0.002 to optimize the model parameters. Dropout is adopted by a rate of 0.33 for all neural modules except BERT. Since we assume only a test (no development) dataset for the target language, we stop the training after 8,000 iterations. We train each model five times and report the averaged results.

5.2 Results

First, our baseline dependency parsing model achieves a UAS of 96.75 and an LAS of 95.14 on the benchmark English Penn Treebank dataset (Stanford Dependencies v3.5.0) by using the base version of the English BERT, and a UAS of 93.38 and an LAS of 91.34 on the UDT dataset,[10] achieving state-of-the-art dependency parsing performance (Kondratyuk and Straka, 2019). However, when multilingual BERT is exploited, the performance shows a significant decrease, resulting in a UAS of 91.54 and an LAS of 89.30 on the UDT dataset. The observation indicates that monolingual training with language-specific BERT might be better than multilingual BERT.

The final result on the test datasets with self-training is shown in Table 1. 50,000 target language dependency trees are selected for training.[11] First, we focus on the models trained on the selected

[8] LAS is not given for the target languages to save space.

[9] https://github.com/google-research/bert

[10] The scores change very little by fine tuning the BERT.

[11] 50,000 is the closest setting to the best-performance models considering all settings and languages.

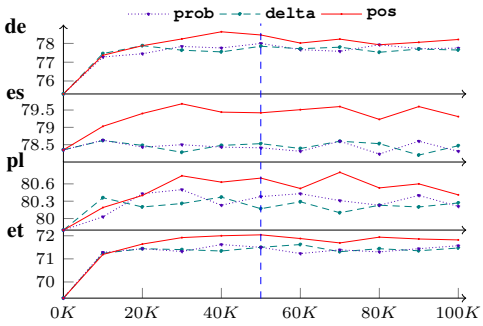


Figure 3: Impact of the selected sentence number.

automatic target dependency trees only, which indicates the effectiveness of the transferred knowledge by the target raw corpus. We list the performances in four groups according to the language family. In this setting, the strategy `prob` and `delta` can bring better performances on the majority languages, except on the language Spanish (es) and Portuguese (pt), which may be due to their differences with the English language making the transferring difficult.

Our final POS guided strategy `pos` can give consistently improved performances on all languages compared with the baseline, demonstrating that it is more effective than the `prob` and `delta` strategies. Although the improvements on all languages are better, the `pos` strategy also shows large variances among the eleven languages, which is similar to that of the `prob` and `delta` strategies. For the language Spanish (es) and Portuguese (pt), the improvements by using `pos` are also much smaller than the other languages. The observation indicates that the individual difference between the source and the target languages is a key factor for the effectiveness of knowledge transferring.

Further, we examine the standard setting of the self-training, merging the selected auto-parsed target dependency trees into the source English trees, and training target language dependency parsing models on both the source and target corpora. We set $\alpha = 0.4$ to reweigh the source English corpus. As shown in Table 1, there are great improvements compared with those of using only the target trees in the majority of cases. After the combination, all three instance selection strategies can obtain large gains. For the strategy `prob` and `delta`, marginal improvements can be obtained for the language Spanish (es) and Portuguese (pt) as well. Thus, self-training can bring improved performances for all the selected languages by using any of the three instance selection strategies, demonstrating the effectiveness of self-training. Overall, we obtain an averaged UAS improvement of $\frac{1.23+1.28+1.82}{3} = 1.44$ considering all selected eleven languages and all instance selection strategies.

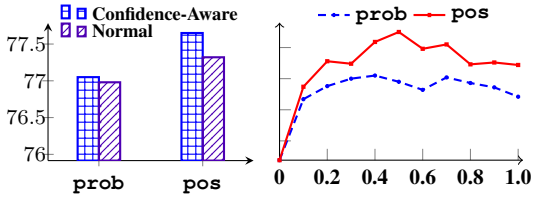
We new look at the performances of self-training with the `pos` instance selection strategy in detail, which is used as our final model. As shown in Table 1, this model achieves the best performances on all languages. The final model can obtain an averaged increase of 1.82 UAS points over all the eleven languages, better than the other two strategies which are 1.23 and 1.28, respectively. In particular, the languages of the IE.Germanic family benefit the most from self-training, leading to an averaged improvement of $\frac{3.14+2.00+2.33}{3} = 2.46$ UAS points, which may be due to the same language family as the source English language. Similarly, the large variations (i.e., the best is 3.14, while the worst is 0.90) of the gains by our final model further demonstrate that the individual difference between the source and the target languages has a strong influence on the effectiveness of self-training.

5.3 Analysis

We choose four languages German (de), Spanish (es), Polish (pl) and Estonian (et) for further analysis, where one language is selected for each family.

Influence of the selected number. First, we examine the performance variations by the selected target dependency tree numbers. Figure 3 shows the tendency, where the start position with zero target tree is our baseline. When the number is surrounding 50,000, the UAS scores remain stable for all languages and instance selection strategies. The `pos` strategy gives more sustainable growth compared with the

Proceedings of the 19th China National Conference on Computational Linguistics, pages 808-820, Hainan, China, October 31 - November 1, 2020.



(a) target reweighing (b) source reweighing

Figure 4: Impact of confidence-aware training.

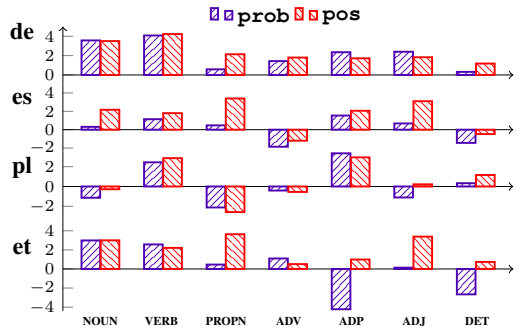


Figure 5: UAS variations with respect to POS tags.

`prob` and `delta` strategies, where the latter two show decreases when the number reaches 20,000. The observation again indicates that `pos` is more effective for instance selection. In addition, we find that `prob` and `delta` are highly similar. Averaged 90% of the selected sentences are identical by the two strategies, while the percentages are lower than 30% when compared to the `pos` strategy, respectively. Thus we exclude the `delta` strategy for the remaining analysis.

Impact of Confidence-Aware Training. Next, we test the effectiveness of confidence-Aware training. Our preliminary experimental results show that their influences are similar across all the four languages. Thus we average their performance to offer overall tendencies of the `prob` and `pos` instance selection strategies. Figure 4 shows the comparison results. For reweighing via the target dependency confidences, the `prob` strategy gains relatively little improvements compared with `pos`, which may be due to repeated information exploited. For source dependency reweighing, the performances remain stable in $[0.4, 0.7]$ for both strategies, resulting in increased UAS values by approximatively 0.3 compared with $\alpha = 1.0$. The observation demonstrates that confidence-aware training can give better performances for self-training.

Performances by POS tags. Further, we analyze the profit distributions of self-training with respect to different POS tags. The delta UAS values by different POS tags (only list seven popular tags) are shown in Figure 5. We see that self-training can not consistently improve the performances over all POS tags, especially for the languages which belong to a different family. By the fine-grained investigation, we can see further that the syntax characteristic of the target language is critical for self-training. The results further indicate that the individual difference between the source and the target languages is important, as mentioned in Section 5.2, as it may determine which kinds of syntax can be accurately captured by self-training. Given a target language, the highly-different syntax attributes might be difficult to learn, as self-training transfers syntax knowledge in a purely unsupervised way. For the language German (de), self-training can obtain better performance on all the seven popular POS tags, while for the other distant language to the English, there exist no consistent findings in more details despite the fact that we can obtain the overall improvements.

Performances by sentence lengths. Finally, we compare the performances in terms of sentence length. Figure 6 shows the results, where the sentence length is categorized into six bins. Overall, self-training

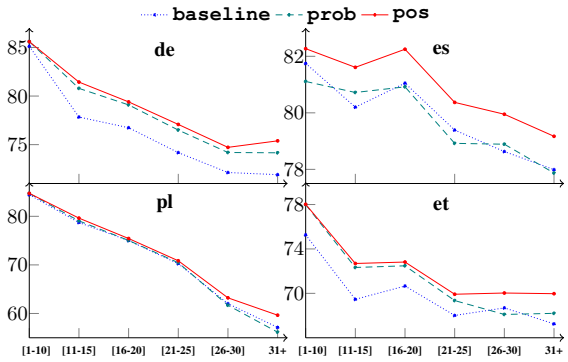


Figure 6: Performances by the sentence length.

brings consistently better performances over all sentence lengths on the four languages, which demonstrates the effectiveness further. We can see that the UAS decreases as a whole as the sentence length grows, which is reasonable since long sentences are difficult to parse (e.g., the head selection range is much larger). By examining the performance differences of the `prob` and `pos` in-depth, we find that `pos` gives larger improvements on longer sentences, which is possibly due to that `prob` tends to select shorter sentences (i.e., averaged 11.4 words compared with 15.2 words by `pos` when 50,000 sentences are selected).

6 Conclusions

We investigated self-training for unsupervised cross-lingual dependency parsing. A baseline dependency parser with multilingual BERT representations is trained and used to parse sentences of a target language and a set of the resulting dependency trees are selected to help training a target language dependency parser. We studied three different instance selection strategies, including two criteria by using the baseline dependency parser, and one criterion guided by a multilingual POS tagger. Results showed that self-training is effective in general for cross-lingual parsing. With the POS-assistant strategy, our final model brings the largest improvements, demonstrating the effectiveness of the method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC No. 61976180 and 61602160), the Westlake University and Bright Dream Joint Institute for Intelligent Robotics.

References

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

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

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *TACL*, 4:431–444.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th ACL*, pages 789–798.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the EMNLP*, pages 261–270.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the EMNLP*, pages 4992–4997.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the CONLL*, 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 ACL*, pages 845–850.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of ACL-IJCNLP*, pages 369–377.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *Proceedings of the 49th ACL*, pages 1486–1495.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL-IJCNLP*, pages 1234–1244.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016a. A distributed representation-based framework for cross-lingual transfer parsing. *JAIR*, 55:995–1023.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016b. A representation learning framework for multi-source transfer parsing. In *AAAI*.

Yulan He and Deyu Zhou. 2011. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606–616.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

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.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th ACL*, pages 2464–2474.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *TACL*, 4:313–327.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the EMNLP-IJCNLP*, pages 2779–2795.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zhenghua Li, Min Zhang, and Wenliang Chen. 2014. Ambiguity-aware ensemble training for semi-supervised dependency parsing. In *Proceedings of the 52nd ACL*, pages 457–467.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of ACL*, volume 1, pages 1337–1348.

David McClosky, Eugene Charniak, and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of the NAACL*, pages 152–159.

David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of the ACL*, pages 337–344.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st ACL*, volume 2, pages 92–97.

Avihai Mejer and Koby Crammer. 2012. Are you sure? confidence in prediction of dependency tree edges. In *Proceedings of the NAACL*, pages 573–576.

Rada Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the CoNLL*, pages 33–40.

Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from modern Hebrew. *TACL*.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of the NAACL*, pages 3912–3918.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the LREC*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th ACL*, pages 4996–5001.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th ACL*, volume 2, pages 412–418.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the EMNLP*, pages 425–435.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of EMNLP*, pages 328–338.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *TACL*, 5:279–293.

Mohammad Sadegh Rasooli and Michael Collins. 2019. Low-resource syntactic transfer with unsupervised source reordering. *arXiv preprint arXiv:1903.05683*.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th ACL*, pages 616–623.

Alexander M Rush, Roi Reichart, Michael Collins, and Amir Globerson. 2012. Improved parsing and pos tagging using inter-sentence consistency constraints. In *Proceedings of the EMNLP*, pages 1434–1444.

Piotr Rybak and Alina Wróblewska. 2018. Semi-supervised neural system for tagging, parsing and lematization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 45–54.

Kenji Sagae. 2010. Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th ICML*, pages 2988–2997.

Michael Schlichtkrull and Anders Søgaard. 2017. Cross-lingual dependency parsing with late decoding for truly low-resource languages. In *Proceedings of the EACL*, volume 1, pages 220–229.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the NAACL*, pages 1599–1613.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the NAACL*, pages 477–487.

Jörg Tiedemann and Zeljko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *JAIR*, 55:209–248.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the EMNLP*, pages 130–140.

Jörg Tiedemann. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th NODALIDA*, number 109, pages 191–199.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of EMNLP*, pages 5725–5731.

Michael Wick, Pallika Kanani, and Adam Pocock. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. In *AAAI*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Min Xiao and Yuhong Guo. 2015. Annotation projection-based representation learning for cross-lingual dependency parsing. In *Proceedings of the CONLL*, pages 73–82.

Juntao Yu, Mohab Elkaref, and Bernd Bohnet. 2015. Domain adaptation for dependency parsing via self-training. In *Proceedings of the IWPT*, pages 1–10.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of EMNLP*, pages 1857–1867.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th ACL*, pages 188–193.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of the EMNLP-IJCNLP*, pages 997–1006.

Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of ECCV*, pages 289–305.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *Proceedings of CVPR*, pages 5982–5991.