Cross-lingual Dependency Parsing with Unlabeled Auxiliary Languages

Wasi Uddin Ahmad[1], **Zhisong Zhang**[2], **Xuezhe Ma**[2]
`wasiahmad@cs.ucla.edu,{zhisongz,xuezhem}@cs.cmu.edu`
Kai-Wei Chang[1], **Nanyun Peng**[3]
`kwchang@cs.ucla.edu,npeng@isi.edu`

[1]University of California, Los Angeles, [2]Carnegie Mellon University
[3]University of Southern California

Abstract

Cross-lingual transfer learning has become an important weapon to battle the unavailability of annotated resources for low-resource languages. One of the fundamental techniques to transfer across languages is learning *language-agnostic* representations, in the form of word embeddings or contextual encodings. In this work, we propose to leverage unannotated sentences from auxiliary languages to help learning language-agnostic representations. Specifically, we explore adversarial training for learning contextual encoders that produce invariant representations across languages to facilitate cross-lingual transfer. We conduct experiments on cross-lingual dependency parsing where we train a dependency parser on a source language and transfer it to a wide range of target languages. Experiments on 28 target languages demonstrate that adversarial training significantly improves the overall transfer performances under several different settings. We conduct a careful analysis to evaluate the language-agnostic representations resulted from adversarial training.

1 Introduction

Cross-lingual transfer, where a model learned from one language is transferred to another, has become an important technique to improve the quality and coverage of natural language processing (NLP) tools for languages in the world. This technique has been widely applied in many applications, including part-of-speech (POS) tagging (Kim et al., 2017), dependency parsing (Ma and Xia, 2014), named entity recognition (Xie et al., 2018), entity linking (Sil et al., 2018), coreference resolution (Kundu et al., 2018), and question answering (Joty et al., 2017). Noteworthy improvements are achieved on low resource language applications due to cross-lingual transfer learning.

In this paper, we study cross-lingual transfer for dependency parsing. A dependency parser consists of (1) an encoder that transforms an input text sequence into latent representations and (2) a decoding algorithm that generates the corresponding parse tree. In cross-lingual transfer, most recent approaches assume that the inputs from different languages are aligned into the same embedding space via multilingual word embeddings or multilingual contextualized word vectors, such that the parser trained on a source language can be transferred to target languages. However, when training a parser on the source language, the encoder not only learns to embed a sentence but it also carries language-specific properties, such as word order typology. Therefore, the parser suffers when it is transferred to a language with different language properties. Motivated by this, we study how to train an encoder for generating language-agnostic representations that can be transferred across a wide variety of languages.

We propose to utilize *unlabeled* sentences of one or more auxiliary languages to train an encoder that learns language-agnostic contextual representations of sentences to facilitate cross-lingual transfer. To utilize the unlabeled auxiliary language corpora, we adopt adversarial training (Goodfellow et al., 2014) of the encoder and a classifier that predicts the language identity of an input sentence from its encoded representation produced by the encoder. The adversarial training encourages the encoder to produce language invariant representations such that the language classifier fails to predict the correct language identity. As the encoder is jointly trained with a loss for the primary task on the source language and adversarial loss on all languages, we hypothesize that it will learn to capture task-specific features as well as generic structural patterns applicable to many languages, and thus have better transferrability.

To verify the proposed approach, we conduct experiments on neural dependency parsers trained on English (source language) and directly transfer them to 28 target languages, with or without the assistance of unlabeled data from auxiliary languages. We chose dependency parsing as the primary task since it is one of the core NLP applications and the development of Universal Dependencies (Nivre et al., 2016) provides consistent annotations across languages, allowing us to investigate transfer learning in a wide range of languages. Thorough experiments and analyses are conducted to address the following research questions:

- Does encoder trained with adversarial training generate language-agnostic representations?
- Does language-agnostic representations improve cross-language transfer?

Experimental results show that the proposed approach consistently outperform a strong baseline parser (Ahmad et al., 2019), with a significant margin in two family of languages. In addition, we conduct experiments to consolidate our findings with different types of input representations and encoders. Our experiment code is publicly available to facilitate future research.[1]

2   Training Language-agnostic Encoders

We train the encoder of a dependency parser in an *adversarial* fashion to guide it to avoid capturing language-specific information. In particular, we introduce a language identification task where a classifier predicts the language identity (id) of an input sentence from its encoded representation. Then the encoder is trained such that the classifier fails to predict the language id while the parser decoder predicts the parse tree accurately from the encoded representation. We hypothesize that such an encoder would have better cross-lingual transferability. The overall architecture of our model is illustrated in Figure 1. In the following, we present the details of the model and training method.

2.1   Architecture

Our model consists of three basic components, (1) a general encoder, (2) a decoder for parsing, and (3) a classifier for language identification. The encoder learns to generate contextualized representations for the input sentence (a word sequence)
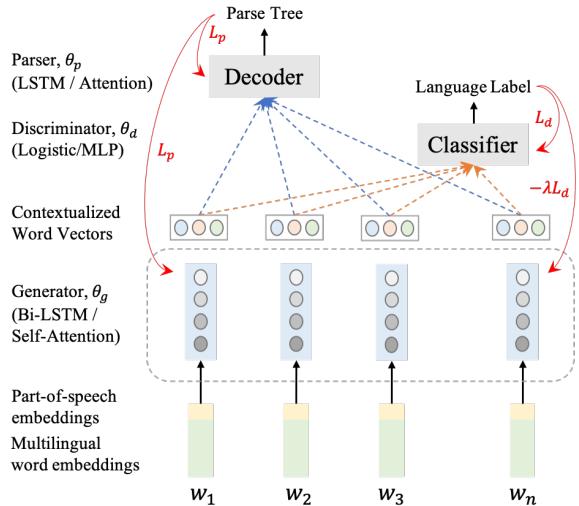

Figure 1: An overview of our experimental model consists of three basic components: (1) Encoder, (2) (Parsing) Decoder, and (3) (Language) Classifier. We also show how parsing and adversarial losses (L_p and L_d) are back propagated for parameter updates.

which are fed to the decoder and the classifier to predict the dependency structure and the language identity (id) of that sentence.

The encoder and the decoder jointly form the parsing model and we consider two alternatives[2] from (Ahmad et al., 2019): "SelfAtt-Graph" and "RNN-Stack". The "SelfAtt-Graph" parser consists of a modified self-attentional encoder (Shaw et al., 2018) and a graph-based deep bi-affine decoder (Dozat and Manning, 2017), while the "RNN-Stack" parser is composed of a Recurrent Neural Network (RNN) based encoder and a stack-pointer decoder (Ma et al., 2018).

We stack a classifier (a linear classifier or a multi-layer Perceptron (MLP)) on top of the encoder to perform the language identification task. The identification task can be framed as either a word- or sentence-level classification task. For the sentence-level classification, we apply average pooling[3] on the contextual word representations generated by the encoder to form a fixed-length representation of the input sequence, which is fed to the classifier. For the word-level classification, we perform language classification for each token individually.

---

[2] Ahmad et al. (2019) studied order-sensitive and order-free models and their performances in cross-lingual transfer. In this work, we adopt two typical ones and study the effects of adversarial training on them.

[3] We also experimented with max-pooling and weighted pooling but average pooling resulted in stable performance.

Algorithm 1 Training procedure.

Parameters to be trained: Encoder (θ_g), Decoder (θ_p), and Classifier (θ_d)

X^a = Annotated source language data

X^b = Unlabeled auxiliary language data

I = Number of warm-up iterations

k = Number of learning steps for the discriminator (D) at each iteration

λ = Coefficient of $\mathcal{L}_d$

α_1, α_1 = learning rate; B = Batch size

Require:

 1: **for** $j = 0, \cdots, I$ **do**
 2:     Update $\theta_g := \theta_g - \alpha_1 \nabla_{\theta_g} \mathcal{L}_p$
 3:     Update $\theta_p := \theta_p - \alpha_1 \nabla_{\theta_p} \mathcal{L}_p$

 4: **for** $j = I, \cdots, num\_iter$ **do**
 5:     **for** k steps **do**
 6:         $(x_a^i)_{i=1}^{B/2} \leftarrow$ Sample a batch from X^a
 7:         $(x_b^i)_{i=1}^{B/2} \leftarrow$ Sample a batch from X^b
 8:         Update $\theta_d := \theta_d - \alpha_2 \nabla_{\theta_d} \mathcal{L}_d$

 9:     Total loss $\mathcal{L} := \mathcal{L}_p - \lambda \mathcal{L}_d$
10:     Update $\theta_g := \theta_g - \alpha_1 \nabla_{\theta_g} \mathcal{L}$
11:     Update $\theta_p := \theta_p - \alpha_1 \nabla_{\theta_p} \mathcal{L}$

In this work, following the terminology in adversarial learning literature, we interchangeably call the encoder as the generator, G and the classifier as the discriminator, D.

2.2 Training

Algorithm 1 describes the training procedure. We have two types of loss functions: $\mathcal{L}_p$ for the parsing task and $\mathcal{L}_d$ for the language identification task. For the former, we update the encoder and the decoder as in the regular training of a parser. For the latter, we adopt adversarial training to update the encoder and the classifier. We present the detailed training schemes in the following.

2.2.1 Parsing

To train the parser, we adopt both cross-entropy objectives for these two types of parsers as in (Dozat and Manning, 2017; Ma et al., 2018). The encoder and the decoder are jointly trained to optimize the probability of the dependency trees (y) given sentences (x):

$$\mathcal{L}_p = -\log p(y|x).$$

The probability of a tree can be further factorized into the products of the probabilities of each token's (m) head decision ($h(m)$) for the graph-based parser, or the probabilities of each transition step decision (t_i) for the transition-based parser:

$$\text{Graph:} \quad \mathcal{L}_p = -\sum_m \log p(h(m)|x, m),$$
$$\text{Transition:} \quad \mathcal{L}_p = -\sum_i \log p(t_i|x, t_{<i}).$$

2.2.2 Language Identification

Our objective is to train the contextual encoder in a dependency parsing model such that it encodes language specific features as little as possible, which may help cross-lingual transfer. To achieve our goal, we utilize adversarial training by employing unlabeled auxiliary language corpora.

Setup We adopt the basic generative adversarial network (GAN) for the adversarial training. We assume that X^a and X^b be the corpora of the source and auxiliary language sentences, respectively. The discriminator acts as a binary classifier and is adopted to distinguish the source and auxiliary languages. For the training of the discriminator, weights are updated according to the original classification loss:

$$\mathcal{L}_d = \mathbb{E}_{x \sim X^a}[\log D(G(x)] +$$
$$\mathbb{E}_{x \sim X^b}[\log (1 - D(G(x))].$$

For the training of dependency parsing, the generator, G collaborates with the parser but acts as an adversary with respect to the discriminator. Therefore, the generator weights (θ_g) are updated by minimizing the loss function,

$$\mathcal{L} = \mathcal{L}_p - \lambda \mathcal{L}_d,$$

where λ is used to scale the discriminator loss ($\mathcal{L}_d$). In this way, the generator is guided to build language-agnostic representations in order to fool the discriminator while being helpful for the parsing task. Meanwhile, the parser can be guided to rely more on the language-agnostic features.

Alternatives We also consider two alternative techniques for the adversarial training: Gradient Reversal (GR) (Ganin et al., 2016) and Wasserstein GAN (WGAN) (Arjovsky et al., 2017). As opposed to GAN based training, in GR setup, the discriminator acts as a multiclass classifier that predicts language identity of the input sentence, and we use multi-class cross-entropy loss. We also study Wasserstein GAN (WGAN), which is proposed by Arjovsky et al. (2017) to improve the stability of GAN based learning. Its loss function is shown as follows.

$$\mathcal{L}_d = \mathbb{E}_{x \sim X^a}[D(G(x)] - \mathbb{E}_{x \sim X^b}[D(G(x)],$$

Language Families	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austronesian	Indonesian (id)
IE.Baltic	Latvian (lv)
IE.Germanic	Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)
IE.Indic	Hindi (hi)
IE.Latin	Latin (la)
IE.Romance	Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)
IE.Slavic	Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
Korean	Korean (ko)
Uralic	Estonian (et), Finnish (fi)

Table 1: The selected 29 languages for experiments from UD v2.2 (Nivre et al., 2018).

here, the annotations are similar to those in the GAN setting.

3 Experiments and Analysis

In this section, we discuss our experiments and analysis on cross-lingual dependency parsing transfer from a variety of perspectives and show the advantages of adversarial training.

Settings. In our experiments, we study single-source parsing transfer, where a parsing model is trained on one source language and directly applied to the target languages. We conduct experiments on the Universal Dependencies (UD) Treebanks (v2.2) (Nivre et al., 2018) using 29 languages, as shown in Table 1. We use the publicly available implementation[4] of the "SelfAtt-Graph" and "RNN-Stack" parsers.[5] Ahmad et al. (2019) show that the "SelfAtt-Graph" parser captures less language-specific information and performs better than the 'RNN-Stack' parser for distant target languages. Therefore, we use the "SelfAtt-Graph" parser in most of our experiments. Besides, the multilingual variant of BERT (mBERT) (Devlin et al., 2019) has shown to perform well in cross-lingual tasks (Wu and Dredze, 2019) and outperform the models trained on multilingual word embeddings by a large margin. Therefore, we consider conducting experiments with both multilingual word embeddings and mBERT. We use aligned multilingual word embeddings (Smith

et al., 2017; Bojanowski et al., 2017) with 300 dimensionss or contextualized word representations provided by multilingual BERT[6] (Devlin et al., 2019) with 768 dimensions as the word representations. In addition, we use the Gold universal POS tags to form the input representations.[7] We freeze the word representations during training to avoid the risk of disarranging the multilingual representation alignments.

We select six auxiliary languages[8] (French, Portuguese, Spanish, Russian, German, and Latin) for unsupervised language adaptation via adversarial training. We tune the scaling parameter λ in the range of $[0.1, 0.01, 0.001]$ on the source language validation set and report the test performance with the best value. For gradient reversal (GR) and GAN based adversarial objectives, we use Adam (Kingma and Ba, 2015) to optimize the discriminator parameters, and for WGAN, we use RMSProp (Tieleman and Hinton, 2012). The learning rate is set to 0.001 and 0.00005 for Adam and RMSProp, respectively. We train the parsing models for 400 and 500 epochs with multilingual BERT and multilingual word embeddings respectively. We tune the parameter I (as shown in Algorithm 1) in the range of $[50, 100, 150]$.

Language Test. The goal of training the contextual encoder adversarially with unlabeled data from auxiliary languages is to encourage the encoder to capture more language-agnostic representations and less language-dependent features. To test whether the contextual encoders retain language information after adversarial training, we train a multi-layer Perceptron (MLP) with softmax on top of the *fixed* contextual encoders to perform a 7-way classification task.[9] If a contextual encoder performs better in the language test, it indicates that the encoder retains language specific information.

3.1 Results and Analysis

Table 2 presents the main transfer results of the "SelfAtt-Graph" parser when training on only English (en, baseline), English with French (en-fr), and English with Russian (en-ru). The re-

---

[4]https://github.com/uclanlp/CrossLingualDepParser
[5]We adopt the same hyper-parameters, experiment settings and evaluation metrics as those in (Ahmad et al., 2019).

[6]https://github.com/huggingface/pytorch-transformers
[7]We concatenate the word and POS representations. In our future work, we will conduct transfer learning for both POS tagging and dependency parsing.
[8]We want to cover languages from different families and with varying distances from the source language (English).
[9]With the source (English) and six auxiliary languages.

Lang	Multilingual Word Embeddings			Multilingual BERT		
	(en)	(en-fr)	(en-ru)	(en)	(en-fr)	(en-ru)
en	**90.23/88.23**	90.01/88.08	89.93/87.93	**93.19/91.21**	92.81/90.97	92.77/90.86
no	80.82/72.94	80.60/72.83	**80.98/73.10**	**85.81/79.03**	85.50/78.64	85.43/78.76
sv	80.33/72.54	79.90/72.16	**80.43/72.68**	85.61/78.34	**85.64/78.58**	85.44/78.33
fr	77.71/72.35	78.49†/73.30†	78.31/73.29	85.22/80.78	84.76/80.26	**85.91†/81.63†**
pt	76.41/67.35	76.88†/67.74	**77.09†/67.81**	82.93/73.33	82.71/73.13	**83.43†/73.88†**
da	**76.58/68.11**	75.99/67.64	76.25/68.03	82.36/73.53	**82.40/73.68**	82.36/**73.86†**
es	73.76/65.46	**74.14**/65.78	74.08/**65.84**	80.81/72.66	81.11/72.80	**81.38†/73.29†**
it	80.89/75.61	**81.33†/76.14†**	80.70/75.57	87.07/82.38	86.90/82.22	**87.41/82.67**
hr	62.21/52.67	**63.38†/53.83†**	63.11†/53.62†	72.96/62.65	73.39†/62.20	**74.20†/63.55†**
ca	73.18/64.53	**73.46†**/64.71	73.40/**64.90†**	80.40/71.42	80.30/71.42	**80.75/71.78**
pl	74.65/62.72	75.65†/63.31†	**75.93/63.60**	81.51/69.25	82.33†/69.91†	**82.48†/70.54†**
uk	59.25/51.92	60.58†/52.72†	**60.81†**/52.66†	69.98/61.52	70.24/61.61	**71.21†/62.84†**
sl	67.51/56.42	68.14/56.52	**68.40/56.87**	75.15/63.12	74.60/62.52	**75.50/63.65†**
nl	68.54/59.99	68.80/60.23	**69.23†/60.51†**	76.76/68.35	**76.94**/68.28	76.89/**68.76†**
bg	79.09/67.61	**80.01†/68.42**	79.72/68.39	86.82/75.47	87.08/75.40	**87.61†/75.94†**
ru	60.91/52.03	61.42†/52.27†	**61.67†/52.41†**	71.92/62.09	72.31/62.15	**72.88†/62.94†**
de	**71.41/61.97**	70.70/61.41	71.05/61.84	78.66/69.81	78.04/69.23	**79.08†/70.26†**
he	55.70/48.08	**57.33†/49.37†**	57.15†/49.36†	64.46/**55.82**	64.97†/55.63	**65.30†**/55.76
cs	63.30/54.14	63.94†/54.63†	**64.37†/55.08†**	73.78/63.52	**74.57†/63.86**	74.56/**64.17†**
ro	65.13/53.98	**65.86/54.76**	65.57/54.42	75.10/62.99	75.85†/63.92†	**76.06†/63.78†**
sk	66.79/58.23	**67.46†/58.77**	67.42†/58.70	76.30/67.38	77.08†/67.57	**77.86†/68.28†**
id	49.85/44.09	**52.05†/45.76†**	51.57/45.31	56.80/50.24	**57.45†**/50.27	57.30†/**50.70†**
lv	70.45/49.47	70.03/49.38	**70.67†/49.61†**	75.63/53.93	75.27/53.78	75.62/**54.29**
fi	66.11/48.73	65.84/48.61	**66.28/48.82**	71.59/**53.81**	71.35/53.63	**71.74**/53.79
et	65.01/44.78	65.31†/45.12†	**65.38†/45.32†**	71.55/50.98	**71.73/51.27**	71.25/51.16
ar	37.63/27.48	38.72†/**28.00†**	**38.98†**/27.89†	49.27/37.62	50.37†/39.37†	**50.95†/39.57†**
la	47.74/34.90	48.80†/35.64†	**49.17†/35.73†**	51.83/38.20	51.48/38.00	**52.20/38.28**
ko	**34.44/16.18**	33.98/15.93	34.23/16.08	38.10/20.62	38.03/20.59	**38.98†/21.54†**
hi	36.34/27.43	36.72/27.40	**37.37†/28.01†**	45.40/35.03	**47.74†/35.90†**	46.10†/34.74
Average	65.92/55.86	66.40†/56.22†	**66.53†/56.32†**	73.34/62.93	73.55/62.99	**73.88†/63.43†**

Table 2: Cross-lingual transfer performances (UAS%/LAS%, excluding punctuation) of the SelfAtt-Graph parser (Ahmad et al., 2019) on the test sets. In column 1, languages are sorted by the word-ordering distance to English. (en-fr) and (en-ru) denotes the source-auxiliary language pairs. '†' indicates that the adversarially trained model results are statistically significantly better (by permutation test, $p < 0.05$) than the model trained only on the source language (en). Results show that the utilization of unlabeled auxiliary language corpora improves cross-lingual transfer performance significantly.

sults demonstrate that the adversarial training with the auxiliary language identification task benefits cross-lingual transfer with a small performance drop on the source language. When multi-lingual embedding is employed, the performance significantly improves, in terms of UAS of 0.48 and 0.61 over the 29 languages when French and Russian are used as the auxiliary language, respectively. When richer multilingual representation technique like mBERT is employed, adversarial training can still improve cross-lingual transfer performances (0.21 and 0.54 UAS over the 29 languages by using French and Russian, respectively).

Next, we apply adversarial training on the "RNN-Stack" parser and show the results in Table 3. Similar to the "SelfAtt-Graph" parser, the "RNN-Stack" parser resulted in significant improvements in cross-lingual transfer from unsu-

pervised language adaptation. We discuss our detailed experimental analysis in the following.

3.1.1 Impact of Adversarial Training

To understand the impact of different adversarial training types and objectives, we apply adversarial training on both word- and sentence-level with gradient reversal (GR), GAN, and WGAN objectives. We provide the average cross-lingual transfer performances in Table 4 for different adversarial training setups. Among the adversarial training objectives, we observe that in most cases, the GAN objective results in better performances than the GR and WGAN objectives. Our finding is in contrast to Adel et al. (2018) where GR was reported to be the better objective. To further investigate, we perform the language test on the encoders trained via these two objectives. We find that the GR-based trained encoders

perform consistently better than the GAN based ones on the language identification task, showing that via GAN-based training, the encoders become more language-agnostic. In a comparison between GAN and WGAN, we notice that GAN-based training consistently performs better.

Comparing word- and sentence-level adversarial training, we observe that predicting language identity at the word-level is slightly more useful for the "SelfAtt-Graph" model, while the sentence-level adversarial training results in better performances for the "RNN-Stack" model. There is no clear dominant strategy.

In addition, we study the effect of using a linear classifier or a multi-layer Perceptron (MLP) as the discriminator and find that the interaction between the encoder and the linear classifier resulted in improvements.[10]

3.1.2 Adversarial v.s. Multi-task Training

In section 3.1.1, we study the effect of learning language-agnostic representation by using auxiliary language with adversarial training. An alternative way to leverage auxiliary language corpora is by encoding language-specific information in the representation via multi-task learning. In the multi-task learning (MTL) setup, the model observes the same amount of data (both labeled and unlabeled) as the adversarially trained (AT) model. The only difference between the MTL and AT models is that in the MTL models, the contextual encoders are encouraged to capture language-dependent features while in the AT models, they are trained to encode language-agnostic features.

The experiment results using multi-task learning in comparison with the adversarial training are presented in Table 5. Interestingly, although the MTL objective sounds contradiction to adversarial learning, it has a positive effect on the cross-lingual parsing, as the representations are learned with certain additional information from new (unlabeled) data. Using MTL, we sometimes observe improvements over the baseline parser, as indicated with the † sign, while the AT models consistently perform better than both the baseline and the MTL model (as shown in Columns 2–5 in Table 5). The comparisons on parsing performances do not reveal whether the contextual encoders learn to

Lang	(en)	(en-fr)	(en-ru)
en	89.65/87.43	**89.88/87.66**	89.67/87.56
no	80.20/72.11	80.42/72.49	**80.73†/72.65†**
sv	81.02/72.95	81.14/**73.44†**	**81.20**/73.37
fr	77.42/72.27	77.45/72.72	**77.78/73.10**
pt	75.94/67.40	76.09/67.47	**76.39†/67.85†**
da	76.87/68.06	77.43†/68.62†	**77.92†/69.24†**
es	73.92/65.95	74.32†/66.35†	**74.83†/66.83†**
it	80.09/75.36	80.98†/76.00†	**81.04†/76.06†**
hr	59.53/49.19	60.00†/50.02†	**60.16†/50.16†**
ca	73.62/64.97	73.73/65.11	**74.18†/65.59†**
pl	71.48/57.43	72.48†/**59.19†**	**72.55†**/58.38†
uk	57.23/49.67	58.38†/**51.04†**	**58.57†**/50.88†
sl	65.48/53.40	66.11†/**54.21†**	**66.23†**/54.09†
nl	67.13/59.15	67.57/59.71†	**67.76†/59.96†**
bg	77.28/65.77	77.79†/**66.66†**	**78.02†**/66.53†
ru	58.70/49.34	59.77†/**50.77†**	**59.98†**/50.51†
de	69.71/58.51	70.03/**59.45†**	**70.05**/59.38†
he	52.97/45.73	53.63†/46.49†	**54.72†/47.34†**
cs	60.99/51.63	61.60†/52.41†	**61.81†/52.45†**
ro	62.01/51.03	62.49/51.30	**63.22†/51.91†**
sk	64.44/56.01	65.03†/56.65†	**65.36†/56.67†**
id	45.08/40.00	45.46/40.61†	**46.82†/41.63†**
lv	70.22/48.46	**71.08†/49.10†**	70.76/48.86
fi	65.39/47.78	**65.59/48.31†**	65.42/47.84
et	64.73/43.84	65.01/**44.27**	**65.04**/44.16
ar	30.98/23.83	31.91†/**24.72†**	**32.83†/25.34†**
la	**45.28**/33.08	44.94/32.94	45.12/**33.11**
ko	**33.50/14.36**	32.87/14.10	32.60/14.11
hi	27.63/19.16	**27.66/19.22**	26.72/18.96
Average	64.09/53.93	64.51†/54.52†	**64.74†/54.64†**

Table 3: Cross-lingual transfer results (UAS%/LAS%, excluding punctuation) of the RNN-Stack parser on the test sets. '†' indicates that the adversarially trained model results are statistically significantly better (by permutation test, $p < 0.05$) than the model trained only on the source language (en).

encode language-agnostic or dependent features.

Therefore, we perform language test with the MTL and AT (GAN based) encoders, and the results are shown in Table 5, Columns 6–7. The results indicate that the MTL encoders consistently perform better than the AT encoders, which verifies our hypothesis that adversarial training motivates the contextual encoders to encode language-agnostic features.

3.1.3 Impact of Auxiliary Languages

To analyze the effects of the auxiliary languages in cross-language transfer via adversarial training, we perform experiments by pairing up[11] the source language (English) with six different lan-

---

[10]This is a known issue in GAN training as the discriminator becomes too strong, it fails to provide useful signals to the generator. In our case, MLP as the discriminator predicts the language labels with higher accuracy and thus fails.

[11]We also conduct experiments on multiple languages as the auxiliary language. For GAN and WGAN-based training, we concatenate the corpora of multiple languages and treat them as one auxiliary language. In these set of experiments, we do not observe any apparent improvements.

AT	SelfAtt-Graph				RNN-Stack			
	en-fr		en-ru		en-fr		en-ru	
	word	sent	word	sent	word	sent	word	sent
GR	66.19	66.21	66.38	66.38	**64.51**	**64.51**	64.52	64.52
GAN	**66.40**	66.29	**66.53**	66.41	64.40	**64.51**	64.63	**64.74**
WGAN	66.24	66.18	66.40	66.27	64.29	64.34	64.57	64.57

Table 4: Average cross-lingual transfer performances (UAS%, excluding punctuation) on the test sets using different adversarial training objective and setting. Multilingual word embeddings are used for these experiments.

Lang (Src. + Aux.)	Auxiliary Language Perf.		Average Cross-lingual Perf.		Lang. Test Perf.	
	AT	MTL	AT	MTL	AT	MTL
en + fr	**78.49/73.30**[†]	78.26/72.98[†]	**66.40/56.22**	66.18/56.04	**62.25**	59.94
en + pt	**76.53/67.45**[†]	75.88/66.75	**66.40/56.22**	66.27/56.08	60.17	**72.02**
en + es	73.66/65.48	**74.04/65.83**[†]	**66.38/56.24**	66.22/56.12	56.78	**74.52**
en + ru	**61.67/52.41**[†]	61.08/52.04	**66.53/56.32**	66.35/56.20	37.34	**60.56**
en + de	**71.65/62.11**[†]	71.17/61.88	**66.41/56.13**	66.18/56.12	61.22	**72.08**
en + la	**49.22/35.94**[†]	48.04/35.09[†]	**66.45/56.20**	66.17/56.05	50.04	**64.91**

Table 5: Comparison between adversarial training (AT) and multi-task learning (MTL) of the contextual encoders. Columns 2–5 demonstrate the parsing performances (UAS%/LAS%, excluding punctuation) on the auxiliary languages and average of the 29 languages. Columns 6–7 present accuracy (%) of the language label prediction test. '†' indicates that the performance is higher than the baseline performance (shown in the 2nd column of Table 2).

Aux. lang	Avg. Dist. to other lang	multilingual Word Emb.	multilingual BERT
pt	0.144	66.40/56.22	73.47/63.11
ru	0.146	**66.53/56.32**	73.88/63.43
de	0.151	66.41/56.13	**73.92/63.56**
es	0.151	66.38/56.24	71.71/62.49
fr	0.160	66.40/56.22	73.55/62.99
la	0.242	66.45/56.20	73.69/63.29

Table 6: Average cross-lingual transfer performances (UAS%/LAS%, w/o punctuation) on the test sets using SelfAtt-Graph parser when different languages play the role of the auxiliary language in adversarial training.

guages (spanning Germanic, Romance, Slavic, and Latin language families) as the auxiliary language. The average cross-lingual transfer performances are presented in Table 6 and the results suggest that Russian (ru) and German (de) are better candidates for auxiliary languages.

We then dive deeper into the effects of auxiliary languages trying to understand whether auxiliary languages particularly benefit target languages that are closer to them[12] or from the same family. Intuitively, we would assume when the auxiliary language has a smaller average distance to all the target languages, the cross-lingual transfer performance would be better. However, from the results in Table 6, we do not see such a pattern. For example, Portuguese (pt) has the smallest average distance to other languages among the aux-

---

[12]The language distances are computed based on word order characteristics as suggested in Ahmad et al. (2019).

iliary languages we tested, but it is not among the better auxiliary languages.

We further zoom in the cross-lingual transfer improvements for each language families as shown in Table 7. We hypothesis that the auxiliary languages to be more helpful for the target languages in the same family. The experimental results moderately correlate with our expectation. Specifically, the Germanic family benefits the most from employing German (de) as the auxiliary language; similarly Slavic family with Russian (ru) as the auxiliary language (although German as the auxiliary language brings similar improvements). The Romance family is an exception because it benefits the least from using French (fr) as the auxiliary language. This may due to the fact that French is too closed to English, thus is less suitable to be used as an auxiliary language.

4 Related Work

Unsupervised Cross-lingual Parsing. Unsupervised cross-lingual transfer for dependency parsing has been studied over the past few years (Agić et al., 2014; Ma and Xia, 2014; Xiao and Guo, 2014; Tiedemann, 2015; Guo et al., 2015; Aufrant et al., 2015; Rasooli and Collins, 2015; Duong et al., 2015; Schlichtkrull and Søgaard, 2017; Ahmad et al., 2019; Rasooli and Collins, 2019; He et al., 2019). Here, "unsupervised transfer" refers to the setting where a parsing model trained only on the source language is directly

Lang	(en,ru) - en	(en,fr) - en	(en,de) - en	(en,la) - en
	IE.Slavic Family			
hr	1.24/0.90	0.43/-0.45	1.52/1.02	0.06/-0.13
sl	0.35/0.53	-0.55/-0.60	-0.04/0.14	-0.17/-0.50
uk	1.23/1.32	0.26/0.09	1.54/1.33	-0.29/-0.09
pl	0.97/1.29	0.82/0.66	0.82/0.98	1.03/0.98
bg	0.79/0.47	0.26/-0.07	0.49/0.41	0.01/0.04
ru	0.96/0.85	0.39/0.06	1.07/1.11	0.20/0.34
cs	0.78/0.65	0.79/0.34	0.91/0.81	-0.08/0.05
sk	1.56/0.90	0.78/0.19	1.88/1.04	0.56/0.66
Avg.	**0.98/0.86**	**0.4/0.03**	**1.02/0.86**	**0.17/0.17**
	IE.Romance Family			
pt	0.50/0.55	-0.22/-0.20	0.54/0.80	0.49/0.60
fr	0.69/0.85	-0.46/-0.52	0.49/0.16	0.95/0.86
es	0.57/0.63	0.30/0.14	0.45/0.39	0.44/0.51
it	0.34/0.29	-0.17/-0.16	-0.22/-0.17	0.26/0.18
ca	0.35/0.36	-0.10/0.00	0.64/0.70	0.10/0.28
ro	0.96/0.79	0.75/0.93	1.32/1.32	1.62/1.73
Avg.	**0.57/0.58**	**0.02/0.03**	**0.54/0.53**	**0.64/0.69**
	IE.Germanic Family			
en	-0.42/-0.35	-0.38/-0.24	-0.35/-0.25	-0.15/-0.20
no	-0.38/-0.27	-0.31/-0.39	-0.41/-0.15	-0.22/-0.24
sv	-0.17/-0.01	0.03/0.24	-0.12/0.35	-0.02/0.18
da	0.00/0.33	0.04/0.15	-0.15/0.08	-0.46/-0.25
nl	0.13/0.41	0.18/-0.07	0.95/0.89	0.57/0.42
de	0.42/0.45	-0.62/-0.58	1.41/1.40	0.25/0.43
Avg.	**-0.07/0.09**	**-0.18/-0.15**	**0.22/0.39**	**0.00/0.06**

Table 7: Average cross-lingual performance difference between the SelfAtt-Graph parser trained on the source (en) and an auxiliary (x) language and the SelfAtt-Graph parser trained only on English (en) language (UAS%/LAS%, excluding punctuation). We use multilingual BERT in this set of experiments.

transferred to the target languages. In this work, we relax the setting by allowing unlabeled data from one or more auxiliary (helper) languages other than the source language. This setting has been explored in a few prior works. Cohen et al. (2011) learn a generative target language parser with unannotated target data as a linear interpolation of the source language parsers. Täckström et al. (2013) adopt unlabeled target language data and a learning method that can incorporate diverse knowledge sources through ambiguous labeling for transfer parsing. In comparison, we leverage unlabeled auxiliary language data to learn language-agnostic contextual representations to improve cross-lingual transfer.

Multilingual Representation Learning. The basic of the unsupervised cross-lingual parsing is that we can align the representations of different languages into the same space, at least at the word level. The recent development of bilingual or multilingual word embeddings provide us with such shared representations. We refer the readers

to the surveys of Ruder et al. (2017) and Glavaš et al. (2019) for details. The main idea is that we can train a model on top of the source language embeddings which are aligned to the same space as the target language embeddings and thus all the model parameters can be directly shared across languages. During transfer to a target language, we simply replace the source language embeddings with the target language embeddings. This idea is further extended to learn multilingual contextualized word representations, for example, multilingual BERT (Devlin et al., 2019), have been shown very effective for many cross-lingual transfer tasks (Wu and Dredze, 2019). In this work, we show that further improvements can be achieved by adapting the contextual encoders via unlabeled auxiliary languages even when the encoders are trained on top of multilingual BERT.

Adversarial Training. The concept of adversarial training via Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Szegedy et al., 2014; Goodfellow et al., 2015) was initially introduced in computer vision for image classification and received enormous success in improving model's robustness on input images with perturbations. Later many variants of GANs (Arjovsky et al., 2017; Gulrajani et al., 2017) were proposed to improve its' training stability. In NLP, adversarial training was first utilized for domain adaptation (Ganin et al., 2016). Since then adversarial training has started to receive an increasing interest in the NLP community and applied to many NLP applications including part-of-speech (POS) tagging (Gui et al., 2017; Yasunaga et al., 2018), dependency parsing (Sato et al., 2017), relation extraction (Wu et al., 2017), text classification (Miyato et al., 2017; Liu et al., 2017; Chen and Cardie, 2018), dialogue generation (Li et al., 2017).

In the context of cross-lingual NLP tasks, many recent works adopted adversarial training, such as in sequence tagging (Adel et al., 2018), text classification (Xu and Yang, 2017; Chen et al., 2018), word embedding induction (Zhang et al., 2017; Lample et al., 2018), relation classification (Zou et al., 2018), opinion mining (Wang and Pan, 2018), and question-question similarity reranking (Joty et al., 2017). However, existing approaches only consider using the *target* language as the auxiliary language. It is unclear whether the language invariant representations learned by previously proposed methods can perform well on a

wide variety of *unseen* languages. To the best of our knowledge, we are the first to study the effects of language-agnostic representations on a broad spectrum of languages.

5 Conclusion

In this paper, we study learning language invariant contextual encoders for cross-lingual transfer. Specifically, we leverage unlabeled sentences from auxiliary languages and adversarial training to induce language-agnostic encoders to improve the performances of the cross-lingual dependency parsing. Experiments and analysis using English as the source language and six foreign languages as the auxiliary languages not only show improvements on cross-lingual dependency parsing, but also demonstrates that contextual encoders successfully learns not to capture language-dependent features through adversarial training. In the future, we plan to investigate the effectiveness of adversarial training for multi-source transfer to parsing and other cross-lingual NLP applications.

References

Heike Adel, Anton Bryl, David Weiss, and Aliaksei Severyn. 2018. Adversarial neural networks for cross-lingual sequence tagging. *arXiv preprint arXiv:1808.04736*.

Željko Agić, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*.

Wasi Uddin Ahmad, Zhisong Zhang, Zuezhe 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR.

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2015. Zero-resource dependency parsing: Boosting delexicalized cross-lingual transfer with linguistic knowledge. In *COLING 2016, the 26th International Conference on Computational Linguistics*, pages 119–130. The COLING 2016 Organizing Committee.

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.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1226–1240.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 50–61, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. *Internation Conference on Learning Representations*.

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 710–721.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Internation Conference on Learning Representations*.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. 2017. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777.

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)*, volume 1, pages 1234–1244.

Junxian He, Zhisong Zhang, Taylor Berg-Kiripatrick, 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.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 395–400. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, Hervé Jégou, et al. 2018. Word translation without parallel data. In *Internation Conference on Learning Representations*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Internation Conference on Learning Representations*.

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

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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338.

Mohammad Sadegh Rasooli and Michael Collins. 2019. Low-resource syntactic transfer with unsupervised source reordering. *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Sebastian Ruder, Anders Søgaard, and Ivan Vulic. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 71–79.

Michael Schlichtkrull and Anders Søgaard. 2017. Cross-lingual dependency parsing with late decoding for truly low-resource languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 220–229. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Internation Conference on Learning Representations*.

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. Association for Computational Linguistics.

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.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Wenya Wang and Sinno Jialin Pan. 2018. Transition-based adversarial network for cross-lingual aspect extraction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4475–4481. International Joint Conferences on Artificial Intelligence Organization.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448.