Representation Alignment and Adversarial Networks for Cross-lingual Dependency Parsing

Anonymous ACL submission

Abstract

Thanks to the strong representation capability of pre-trained language models, dependency parsing in rich-resource language has achieved remarkable improvements. However, the parsing accuracy drops sharply when the model is transferred to low-resource language due to distribution shifts. To alleviate this issue, we propose a representation alignment and adversarial model to filter out useful knowledge from rich-resource language and ignore useless ones. Our proposed model consists of two components, i.e., an alignment network in the input layer for selecting useful language-specific representation features and an adversarial network in the encoder layer for augmenting the language-invariant contextualized features. Experiments on the benchmark datasets show that our proposed model outperforms RoBARTa-enhanced strong baseline models by 1.37 LAS and 1.34 UAS. Detailed analysis shows that both alignment and adversarial networks are equally important in alleviating the distribution shifts problem and can benefit from each other. In addition, the comparative experiments demonstrate that both the alignment and adversarial networks can substantially facilitate extracting and utilizing relevant target language features, thereby increasing the adaptation capability of our proposed model.

1 Introduction

Dependency parsing, as an important fundamental task of natural language processing, aims to identify grammatical and syntax relationships between two words in the input sentence via a dependency tree. Figure 1 shows a dependency tree instance, where a dependency from the headword "voi (elephant)" to the modified word "thông minh (intelligent)" with the relation label "amod" means "thông minh (intelligent)" as an adjective modifies "voi (elephant)". Dependency trees are widely applied to various artificial intelligence tasks, such as machine translation (Zhang et al., 2019), grammatical error correction (Zhang et al., 2022), and information extraction (Tian et al., 2022).

In the past decades, pre-trained language model enhanced dependency parsers have achieved outstanding performances in rich-resource languages (Clark et al., 2018; Li et al., 2022; Nishida and Matsumoto, 2022; Mohammadshahi and Henderson, 2021; Yan et al., 2020). Most significantly, Dozat and Manning (2017) propose a BiAffine parser that leverages multi-layer BiLSTMs to encode input sentences and a BiAffine operation to compute scores, thus achieving better performance on various languages. Then, Li et al. (2019) develop a self-attentive BiAffine parser and further improve the model performance with ELMo and BERT representations. However, these model performances drop sharply in low-resource languages due to the lack of annotated data (Wang et al., 2020; Effland and Collins, 2023; Rotman and Reichart, 2019; Vania et al., 2019).

As shown in Figure 1, both sentences from Vietnamese and Chinese have a similar core grammatical structure "subject-predicate-object", but they also have differences in the attributive positions where Vietnamese adopts "post-modifier" while Chinese is the opposite. Hence, how to construct the discrepancy and similarity between different languages becomes the key challenge for cross-lingual dependency parsing (Ahmad et al., 2019; Üstün et al., 2022; Ozaki et al., 2021; Liu et al., 2020; Xu and Koehn, 2021). A series of previous works have explored feature transfer to improve low-resource parsing. Most recently, Al Ghiffari et al. (2023) propose a hierarchical transfer learning (HTL) approach to exploit a source and an intermediate language to improve the parsing accuracy in low-resource languages. Similarly, Choudhary and O'riordan (2023) incorporate linguistic typology knowledge as an auxiliary task, further im-

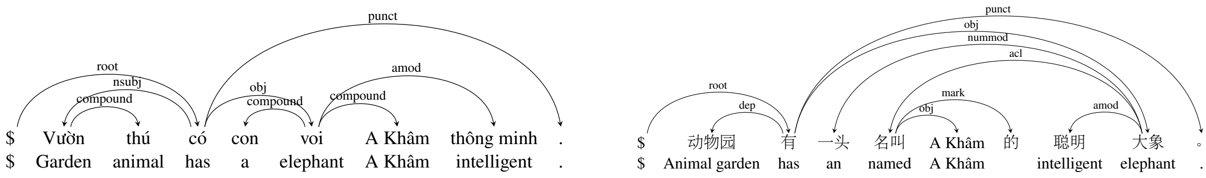

Figure 1: Examples of dependency tree from Universal Dependencies (UD) dataset, where the left sentence is from the low-resource Vietnamese treebanks (VTB) and the right one is from the rich-resource simplified Chinese treebanks (GSDSimp).

proving the low-resource dependency parsing performances. Although transfer learning from rich-resource to low-resource language has shown its promising advantages, how to further emphasize the helpful knowledge and filter out the harmful ones automatically is still an important problem.

To address this issue, we propose a novel representation alignment and adversarial networks for cross-lingual dependency parsing. On the one hand, we propose an alignment network on the input layer to select useful language-specific word information. On the other hand, a language-aware adversarial network is applied on the encoder layer to excavate potential language-invariant knowledge. Experiments on the benchmark dataset show that our proposed model achieves notable performance improvements, leading to new state-of-the-art results. Detailed analysis shows that alignment and adversarial networks are complementary and can benefit from each other. In-depth comparative experiments demonstrate that both alignment and adversarial networks are equally important for filtering out effective knowledge from the source language. In addition, our codes are released at https://github.com/noteljj/align to facilitate future research.

2 Related Work

Cross-Lingual Dependency Parsing. Cross-lingual dependency parsing has emerged as a crucial component of natural language processing, with distinct methodologies contributing to its advancement. Among these, three primary categories stand out: *transfer learning*, *multilingual model adaptation*, and *subword representation alignment*. Transfer learning techniques, epitomized by the work of Chen et al. (2019), Liu et al. (2023b) and Niu et al. (2022), leverage resources from rich-resource languages to improve parsing accuracy in low-resource languages, demonstrating the versatility of transferring syntactic knowledge across linguistic boundaries. In multilingual model adaptation, researchers like Pfeiffer et al. (2021). Wang et al. (2020) and Dione (2021) have adapted multilingual BERT models to enhance parsing performance across various languages, illustrating the power of transformer-based methods in handling diverse linguistic environments. Meanwhile, the subword representation alignment approach, as explored by Schuster et al. (2019); Yaari et al. (2022), focuses on the fine-grained alignment of word or subword representations between languages, addressing the challenge of representing low-resource languages in pre-trained models. Collectively, these approaches underscore the dynamism and complexity of cross-lingual dependency parsing, highlighting both its progress and the ongoing challenges of syntactic alignment and resource disparity. This landscape sets the stage for our investigation into the effective transfer of subword representations from Chinese to Vietnamese, a venture that seeks to mitigate the representation gap for low-resource languages and contribute to the evolving narrative of linguistic adaptability in computational models.

Adversarial Learning. Adversarial learning has become increasingly central in NLP, notably for its role in fortifying model robustness and counteracting data biases (Lowd and Meek, 2005), Zalmout and Habash (2019) and Chen et al. (2021) have demonstrated the efficacy of adversarial examples in bolstering the resilience of NLP models to linguistic variations and malicious attacks. Extending this, Lu et al. (2023) and Zou et al. (2021) have successfully integrated adversarial learning into domain adaptation, effectively reducing domain-specific biases. A recent novel approach by Han et al. (2021) and Zhang et al. (2018) involves using adversarial training to mitigate biases in training. Additionally, the advent of adversarial data augmentation, as investigated by Tan et al. (2022), has shown promise in diversifying training datasets, further enhancing model robustness. Despite these advancements, adversarial learning still confronts challenges in balancing model stability and performance, particularly when dealing with highly complex and nuanced linguistic data, underscoring the need for ongoing research and development in this dynamic area of NLP.

Feature Alignment and Transfer. In the field of feature alignment and transfer, existing research can be categorized into *deep learning-based methods*, *instance-based methods*, and *model-based methods*. Deep learning-based methods automatically learn feature mapping relationships between source and target domains through neural networks, such as aligning feature distributions in the space through adversarial training (Riemer et al., 2015), (Kumar et al., 2023) and (Hazem et al., 2022). Instance-based methods select and weight examples from the source domain to have a greater impact in the target domain, like instance selection based on conditional adversarial learning (Basu Roy Chowdhury et al., 2019; Glavaš and Vulić, 2020). Model-based methods focus on how to use the source domain's model to assist learning in the target domain, such as progressive neural networks that learn to transfer knowledge across domains (Chawla and Yang, 2020; Liu et al., 2023a). These methods have their own advantages and can effectively improve the performance of cross-domain learning in different scenarios.

3 Our Approach

Considering not all rich-source language information is equally important for cross-lingual dependency parsing, we propose the alignment and adversarial networks for effective representation selection. Concretely, we first leverage the multi-lingual pre-trained language model XLM-RoBERTa to improve the word representation capability of both source and target languages. Then, a representation alignment network is applied on the input layer to emphasize useful language-specific information and ignore the harmful one. Next, we exploit an adversarial network on the encoder layer to enhance language-invariant representations. Finally, all selected representations are utilized to search for the best dependency tree. Figure 2 illustrates the framework of our proposed model, which is organized into three components, i.e., *Input layer based on the alignment network*, *Encoder layer enhanced with an adversarial network*, *MLP and BiAffine layers*.

3.1 Input Layer Based on Representation Alignment Network

Given an input sentence $w_1, w_2, \ldots, w_n$, the input layer maps them into dense vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. For the source language Chinese, we directly use the normal embeddings as its input vectors. For the target language Vietnamese, we exploit a representation alignment network to select helpful Chinese word information, further enhancing the Vietnamese representation capability.

Input vectors for Chinese. As shown in Equation 1, each Chinese vector $\mathbf{x}_i^{ch}$ is the concatenation of its word representation and corresponding character representation $\mathbf{word}_i^{\text{char}}$, where word representation is the addition of XLM-RoBERTa representation $\mathbf{rep}_i^{\text{XLM-R}}$ and a random initialization word embedding $\mathbf{emb}_i^{\text{word}}$. The character representation $\mathbf{word}_i^{\text{char}}$ is generated by a BiLSTM network, which first encodes the constituent characters of each word w_i^{ch}, and then combines the hidden vectors of two directions (Lample et al., 2016).

$$\mathbf{x}_i^{\text{ch}} = (\mathbf{rep}_i^{\text{XLM-R}} + \mathbf{emb}_i^{\text{word}}) \oplus \mathbf{word}_i^{\text{char}} \quad (1)$$

Input vectors for Vietnamese. Different from Chinese input vectors, Vietnamese input vector $\mathbf{x}_i^{\text{vi}}$ utilizes an additional aligned representation $\mathbf{emb}_i^{\text{vi-FT}}$ to fuse more useful Chinese word information, which is calculated in Equation 2,

$$\begin{aligned}\mathbf{x}_i^{\text{vi}} = &(\mathbf{emb}_i^{\text{vi-FT}} + \mathbf{rep}_i^{\text{XLM-R}} + \mathbf{emb}_i^{\text{word}}) \\ &\oplus \mathbf{word}_i^{\text{char}}\end{aligned} \quad (2)$$

where $\mathbf{emb}_i^{\text{vi-FT}}$ is generated by our alignment network and other representations are obtained similarly to Chinese.

Alignment network. The key to our alignment network is to enhance the Vietnamese word representation capability by emphasizing useful Chinese words and ignoring harmful ones. First, we construct an alignment matrix based on a new high-quality bilingual dictionary to map Vietnamese and Chinese representations into a close space.

Since the bilingual dictionary significantly affects the performance of our alignment matrix, we adopt automatic generation and manual annotation strategy to ensure the quality of the Vietnamese-Chinese dictionary. Concretely, we first download the dump data backup file from Wikipedia[1] and a

[1]https://en.wikipedia.org/

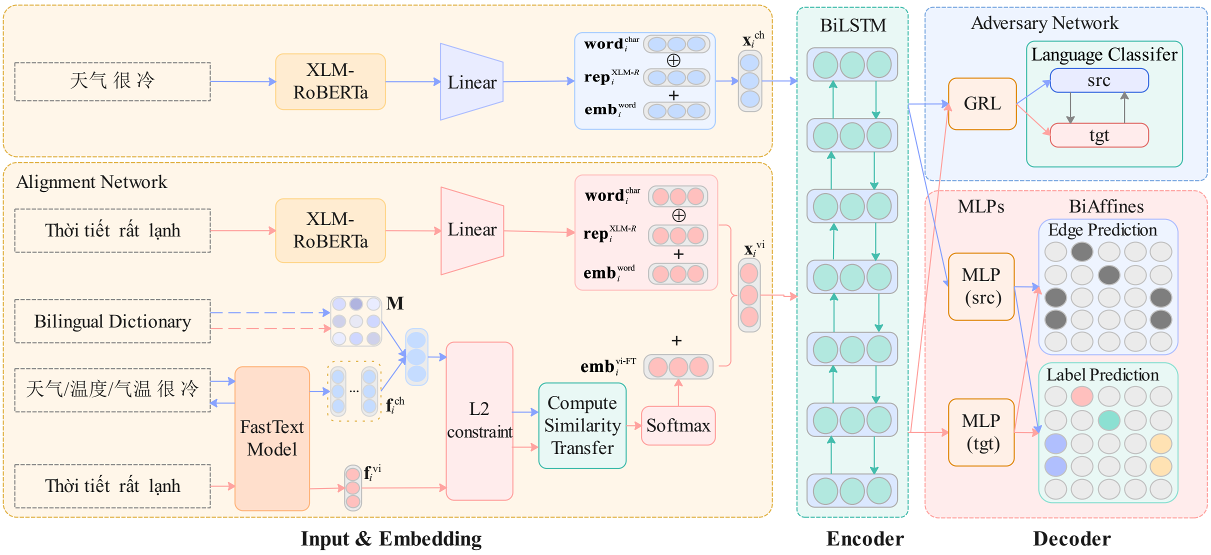

Figure 2: Framework of our proposed model.

simple bilingual dictionary[2]. Second, we use regular expressions to iteratively match and extract the Vietnamese-Chinese alignment titles and subheadings. Third, the alignment word pairs are used to augment the original bilingual dictionary. Finally, the automatic generation dictionary is manually proofread by Vietnamese speakers, thus obtaining a high-quality Vietnamese-Chinese dictionary that contains about $20,000$ word pairs. based on the new dictionary, we use the pre-trained Fasttext models [3] to obtain Vietnamese matrix $V \in \mathcal{R}^{n \times d_1}$ and Chinese matrix $C \in \mathcal{R}^{n \times d_1}$ where n is the number of our dictionary and d_1 denotes the dimension of Fasttext representations. Meanwhile, we exploit an orthogonal similarity transformation to obtain our alignment matrix $M \in \mathcal{R}^{d_1 \times d_1}$ that can be regarded as a linear mapping between Vietnamese and Chinese based on the semantic similarity.

Given a Vietnamese sentence, we first utilize Fasttext models to obtain word segmentation sequences. Then, for each Vietnamese word, we select multiple corresponding Chinese words based on our dictionary. Next, all selected words are dotted with an alignment matrix M, and L2 constraint is applied on them to yield stable and aligned word representations $\hat{\mathbf{f}}_i$. The formula for this operation is as follows,

$$\hat{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{f}_i^2 + \varepsilon}} \tag{3}$$

where $\mathbf{f}_i$ represents the i-th word vector from the FastText model, ε is a very small positive number used to prevent division by zero. Considering each Vietnamese word may align with several Chinese words, we employ the cosine function to compute semantic similarity as alignment weights. The formulas are shown as follows,

$$\begin{aligned} S_{i,j}^{ch,vi} &= \frac{(\hat{\mathbf{f}}_i^{\,ch})^T \hat{\mathbf{f}}_j^{\,vi}}{\parallel \hat{\mathbf{f}}_i^{\,ch} \parallel \parallel \hat{\mathbf{f}}_j^{\,vi} \parallel} \\ \mathbf{w}_{i,j}^{ch,vi} &= \exp(\mathbf{S}_{i,j}^{ch,vi}/\tau) \end{aligned} \tag{4}$$

where $\mathbf{S}_{i,j}^{ch,vi}$ denotes the similarity score between the Chinese word i and the Vietnamese word j. τ denotes the temperature coefficient. $\mathbf{w}_{i,j}^{ch,vi}$ is the corresponding weight. Finally, We construct the final alignment Vietnamese representation $\mathbf{emb}_i^{\text{vi-FT}}$ using constrained word vectors and alignment weights to emphasize useful words and ignore harmful ones. The formula is as follows,

$$\mathbf{emb}_i^{\text{vi-FT}} = \frac{\sum_{ch \in \mathcal{J}_{vi}} \mathbf{w}_{i,j}^{ch,vi} \cdot \hat{\mathbf{f}}_i^{\,ch}}{\sum_{ch \in \mathcal{J}_{vi}} \mathbf{w}_{i,j}^{ch,vi}} \tag{5}$$

where $\mathcal{J}_{vi}$ represents a collection of Chinese words that exhibit the highest degree of similarity to a Vietnamese word.

[2]https://github.com/CPJKU/wechsel/tree/main/dicts/data

[3]https://fasttext.cc/docs/en/crawl-vectors.html/

3.2 Encoder Layer Enhanced with Adversarial Network

Different from the traditional BiLSTM encoder, we employ an adversarial network above the encoder to ensure it imply more potential language-invariant knowledge.

BiLSTM encoder. Following Dozat and Manning (2017), we also adopt a three-layer BiLSTM network as the encoder to generate original contextualized vectors. Since BiLSTM is able to encode the words in a sentence from two directions, each word can obtain contextualized information $\mathbf{h}_i$.

$$\mathbf{h}_i = \mathbf{BiLSTM}(\mathbf{x}_i, \theta_{\text{BiLSTM}}) \tag{6}$$

where θ_{BiLSTM} is the BiLSTM parameters.

Adversarial network. The adversarial network mainly contains three components, i.e., the shared BiLSTM encoder, the Gradient Reversal Layer (GRL), and a language classifier. First, Sentence from Chinese or Vietnamese are fed into the shared BiLSTM layer to obtain contextualized word representations $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$. Then, they pass the GRL which inverts the gradient during backpropagation, thus fostering BiLSTM to learn more shared features between Vietnamese and Chinese. The forward and backward propagation equations for GRL are as follows,

$$\begin{aligned} \mathrm{GRL}_\gamma(\mathbf{h}_i) &= \mathbf{h}_i \\ \frac{d\mathrm{GRL}_\gamma(\mathbf{h}_i)}{d(\mathbf{h}_i)} &= -\gamma \mathbf{I} \end{aligned} \tag{7}$$

where γ is a hyperparameter to balance the impact of adversarial learning and dependency parsing on the shared BiLSTM. Then, we use a multilayer perceptron (MLP) to compute the language distribution scores and a softmax function to obtain the language distribution probabilities. The formula is as follows,

$$\mathbf{re}_i = \mathrm{softmax}\left(\mathrm{MLP}\left(\mathbf{h}_i\right)\right) \tag{8}$$

Finally, we employ a standard cross-entropy loss to optimize all parameters of the adversarial network,

$$\mathcal{L}^{adv} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(\tilde{\mathbf{re}}_{i,j})\log\left((\mathbf{re}_{i,j})\right) \tag{9}$$

where m is the number of languages, n is the word number of input sentence, and $\tilde{\mathbf{re}}_{i,j}$ represents the gold-standard language distribution vector, where only one element is 1 corresponding to the language index where the sentence comes from.

3.3 MLP and BiAffine Layer

The MLP layer employs the enhanced contextualized vector $\mathbf{h}_i$ as its input and reduce the dimension of $\mathbf{h}_i$, extracting its head representation $\mathbf{r}_i^{\mathrm{h}}$ and modifier representation $\mathbf{r}_i^{\mathrm{d}}$ for each word w_i.

$$\begin{aligned} \mathbf{r}_i^{\mathrm{h}} &= \mathrm{MLP}_h\left(\mathbf{h}_i\right) \\ \mathbf{r}_i^{\mathrm{d}} &= \mathrm{MLP}_d\left(\mathbf{h}_i\right) \end{aligned} \tag{10}$$

where $\mathrm{MLP}_{\mathrm{h}}(*)$ and $\mathrm{MLP}_{\mathrm{d}}(*)$ have a single hidden layer with the ReLU activation function. Then, a BiAffine computes $\mathrm{score}(i \leftarrow j)$ between the current word w_i and the other word w_j. Simultaneously, $\mathrm{score}(i \xleftarrow{l} j)$ is calculated by another separated BiAffine layer as equation 11

$$\begin{aligned} \mathrm{score}(i \leftarrow j) &= \begin{bmatrix} \mathbf{r}_i^{\mathrm{d}} \\ 1 \end{bmatrix}^{\mathrm{T}} \mathbf{U}_1 \mathbf{r}_j^{\mathrm{h}} \\ \mathrm{score}(i \xleftarrow{l} j) &= \mathbf{r}_j^{\mathrm{h}} \mathbf{U}_2 \mathbf{r}_i^{\mathrm{d}} + (\mathbf{r}_j^{\mathrm{h}} \oplus \mathbf{r}_i^{\mathrm{d}}) \mathbf{U}_3 + b \end{aligned} \tag{11}$$

where $\mathbf{U}_1$ $\mathbf{U}_2$, $\mathbf{U}_3$, and b are parameters. l denotes the relation label. After obtaining the scores of dependency arcs and dependency labels, we use the typical Maximum Spanning Tree (MST) algorithm to find the highest-score tree as our final parsing result. Finally, for each position i, if the gold-standard head of word w_i is word w_j and its corresponding gold relation label is l, the parsing loss is computed as follows,

$$\begin{aligned} \mathcal{L}^{\mathrm{par}} = &- \log \frac{e^{\mathrm{score}(i \leftarrow j)}}{\sum\limits_{0 \le k \le n, k \ne i} e^{\mathrm{score}(i \leftarrow k)}} \\ &- \log \frac{e^{score(i \xleftarrow{l} j)}}{\sum_{l' \in \mathcal{L}} e^{score(i \xleftarrow{l'} j)}} \end{aligned} \tag{12}$$

where $\mathrm{score}(i \leftarrow k)$ denotes the dependency arc score from head w_i to modefier W_k. $\mathcal{L}$ refers to the collection of all dependency labels l'.

3.4 Cyclic Cross-lingual Training

In this work, we propose a cyclic training strategy to mitigate data imbalance between source and target languages, as outlined in Algorithm 1. Considering the data scale of the source language is much larger than the target one, we divide the first n_1 mini-batches of the source language as s^f and the last as s^l where n_1 is the mini-batch number of the target language. During training, we take turns

Algorithm 1: Cyclic Training Procedure

Input: Source language data S, target language data T
Hyper-parameters: Loss weight α, training iterations k
1: Initialize $iter = 0$
2: **Repeat**
3: Sample mini-batch x alternately from S or T
4: **if** $x \in S^f$:
5: Update parameter by minimizing $\mathcal{L}^{par} + \alpha\mathcal{L}^{adv}$
6: **elif** $x \in S^l$:
7: Update parameter by minimizing $\mathcal{L}^{par}$
8: **else** $x \in T$:
9: Compute $\mathbf{emb}_i^{vi-FT}$ = alignment(θs)
11: Update parameters by minimizing $\mathcal{L}^{par} + \alpha\mathcal{L}^{adv}$
12: $iter += 1$
13: **until** $iter = k$ or convergence

Table 1: Cyclic Cross-lingual Training Procedure.

to sample mini-batch x of source and target languages. If x comes from the first part of the source language S^f, we update parsing and adversarial parameters by minimizing parsing and adversarial losses. While x belongs to S^l, we only update the parser parameters θ_1 by minimizing the parsing loss. If x comes from the target language T, we compute an alignment representation $\mathbf{emb}_i^{vi-FT}$ via an alignment network. and update all parameters by minimizing parsing and adversarial losses. Finally, we iteratively train all the data until it converges or stops prematurely.

Dataset	Train	Dev	Test
GSDSimp	3,997	500	500
VTB	1,400	800	800

Table 2: Dataset statistics in sentence number.

4 Experiments

4.1 Settings

Datasets. To compare with previous work fairly, we use the shared multi-language Universal Dependencies (UD) 2.12 treebank as our benchmark datasets [4]. Concretely, we choose Chinese as our source language and Vietnamese as our target language. The detailed illustrations of our datasets are shown in Table 2.

Evaluation. Following Hajic et al. (2009), we employ the Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) as our evaluation indicators. Each model is trained for at most $1,000$ iterations, and the performance is evaluated on the dev data after each iteration for model selection. We stop the training if the peak performance does not increase in 100 consecutive iterations.

Hyper-parameter choices. We mostly maintain the hyper-parameter settings of Li et al. (2019), such as MLP and BiAffine dimensions, dropout ratios, and so on. The adversary loss weight α, neighbor, and temperature, which are set as 1, 10, and 0.1 respectively. The character embeddings are initialized randomly with a dimension of 100.

Baseline. To validate the advantages and effectiveness of our proposed model, we choose the following approaches as our strong baselines.

- **Pre-training method.** BiAffine parser is first proposed by Dozat and Manning (2017), and then is widely used on various dependency parsing tasks. Different from the original BiAffine parser, we first exploit the Vietnamese pre-trained language model XLM-RoBERTa-base [5] to enhance the parsing performance. Then, we pre-train the enhanced BiAffine parser exclusively on the Vietnamese Universal Dependencies (UD) dataset, which is used as our strong baseline model.

- **Fine-tuning method.** Shi et al. (2022) propose to fine-tune the basic model twice and achieve selective differential privacy for large language models. In this work, we also utilize the idea of fine-tuning method to improve the adaptation capability of the enhanced BiAffine parser in Vietnamese. We first use the Chinese dataset for initial training, and then fine-tune the pre-trained model with the Vietnamese dataset, thus transferring the syntactic knowledge contained in the Chinese treebank to Vietnamese.

- **Adversarial learning method.** Li et al. (2021) apply the adversarial network on the BiAffine parser, thus achieving impressive results on cross-domain dependency parsing. In this work, we attempt to apply an adversarial network on BiAffine parser to capture more similarities between Chinese and Vietnamese.

4.2 Main Results

Table 3 displays the final results on our test data and gives a detailed comparison with previous works. First, we find that our model outperforms the "Adversary" model, demonstrating that our alignment

[4] https://universaldependencies.org/

[5] https://huggingface.co/xlm-roberta-base

Model	LAS	UAS
Results of previous works		
UDPipe (2019)	62.56	70.38
UDify(2019)	66.00	74.11
UDPipe2.0+WCBF(2019)	65.41	72.94
TOWER (2021a)	63.50	72.40
Pre-training	67.61	75.47
Fine-tuning	68.09	75.93
Adversary	68.47	76.39
Our model	**68.98**	**76.81**

Table 3: Main results on the Vietnamese UD test dataset.

network can emphasize useful language-specific representations from the source language and ignore the harmful ones, thus further improving the cross-lingual dependency parsing accuracy. Second, compared with the "Fine-tuning" model, the " Adversary" model achieves better performance, revealing that an adversarial network can extract potential language-invariant knowledge to construct the in-depth relationship between source and target languages. Finally, we can see that our proposed model outperforms all strong baselines, indicating that our proposed representation alignment and adversarial networks are extremely useful for cross-language dependency parsing.

We also compare with previous works in the top block. Kondratyuk and Straka (2019) first propose the UDpipe model, which integrates a tokenizer, morphological analyzer, POS tagger, lemmatizer, and dependency parser into a single model for comprehensive natural language processing. Then, they propose a UDify framework based on a multilingual BERT self-attention model with tagging and parser joint training, which fine-tunes a multilingual pre-trained model with 104 languages to improve parsing accuracy. Straka et al. (2019) enhance the UDPipe model by incorporating various embeddings, including BERT and Flair. Lastly, Glavaš and Vulić (2021b) propose a TOWER model, which uses hierarchical language clustering to improve the low-resource dependency parsing performance. Compared with these works, we find that our model can achieve the best performance with only a single target language, highlighting the efficiency and powerful parsing capabilities of our proposed model.

4.3 Ablation Study

Results of ablation studies are shown in Table 4. First, we find that removing either the adversarial network or the representation alignment network can lead to a decrement in parsing performance. This outcome suggests that each module plays a crucial role in mitigating the potential conflicts arising from direct language transfer. Second, removing adversarial and alignment modules simultaneously leads to a significant decline in dependency parsing accuracy, revealing that the two modules are complementarity and benefit from each other. Most notably, the performance deteriorates to its lowest when the source language is excluded altogether, affirming that the source language encompasses valuable information beneficial for the target language. This observation not only emphasizes the importance of preserving source language features but also reinforces the necessity of their strategic filtration.

Model	LAS	UAS
Our model	**68.98**	**76.81**
w/o Adv	68.71	76.53
w/o Ali	68.47	76.39
w/o Adv & Ali	68.09	75.93
w/o Adv & Ali & Ch	67.61	75.47

Table 4: Ablation study on reducing the component of our model on test data, where "w/o Adv", "w/o Ali", and "w/o Ch" mean removing the adversarial network, representation alignment network or the Chinese UD training dataset.

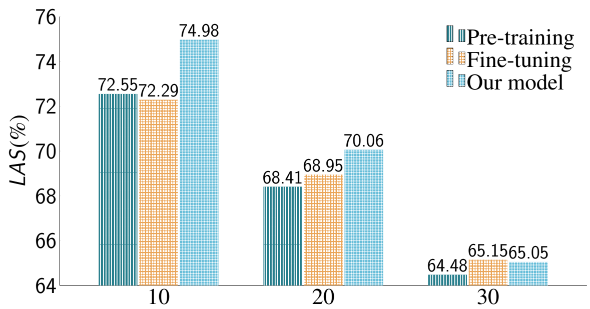

Figure 3: LAS regarding diverse sentence lengths.

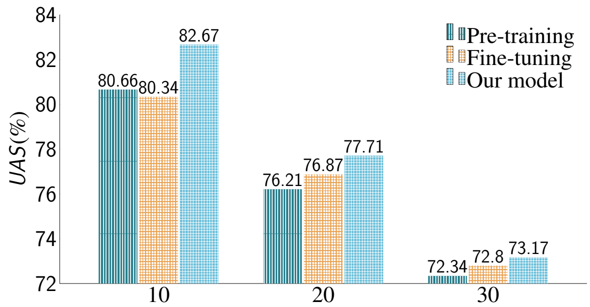

Figure 4: UAS regarding diverse sentence lengths.

4.4 Error Analysis

Sentence length. Figure 3 and Figure 4 present the LAS and UAS scores regarding diverse sentence lengths. First, it is clear that all models perform better with shorter sentences. For sentences under 10 words, the LAS and UAS scores hover around 73 and 82, respectively. However, there is a noticeable drop of over 9 points in scores for sentences approximately 30 words in length, indicating that the parsing difficulty is sharply improved with the increase in sentence length. Then, we can see that the "Pre-training" model records the lowest scores across all length categories. Notably, incorporating the Chinese corpus enhances its performance across most lengths, except for the 10-word category, The reason may be that pronounced structural disparities between short Chinese and Vietnamese sentences. Finally, our model significantly mitigates the performance decline observed with the "Fine-tuning" model, achieving substantial improvements across all sentence lengths.

DEP	Precision (%)		
	Pre-training	Fine-tuning	Our
amod	67.45	63.78	**67.97**
cc	87.34	86.74	**88.64**
ccomp	54.33	54.64	**56.45**
compound	73.03	73.47	**74.75**
conj	63.69	64.50	**66.60**
cop	81.35	81.94	**82.05**
discourse	44.12	**53.57**	52.78
mark	73.00	73.33	**73.58**
nmod	70.84	71.99	**73.12**
nsubj	83.42	83.47	**83.85**
obj	79.86	81.17	**81.67**
root	79.64	79.71	**80.14**

Table 5: Precisions of dependency labels on different models.

Dependency labels. Table 5 presents the precisions of main dependency labels on different models. These models include the Chinese training dataset to analyze inter-language connections. First, the "Pre-training" model registers the lowest scores across all dependency labels. Then, the " Fine-tuning" model achieves better performance on most dependency labels. The reason may be that the dependency trees in the target language contain abundant language-specific syntax information. Finally, our proposed model consistently obtains the highest scores on almost all labels, further proving the effectiveness of our proposed model.

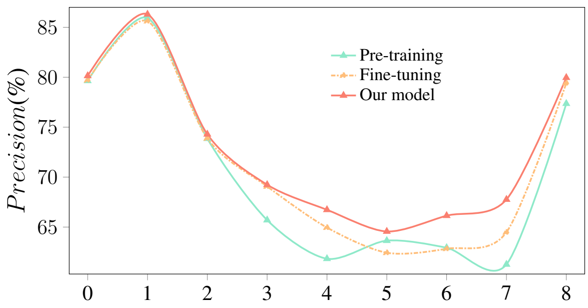

Figure 5: Precision of diverse models regarding different binned head absolute distances with punctuation.

Absolute distance. Figure 5 shows the effects of absolute distances from the head word to the modifier word on dev data. First, the "Pre-training" model achieves the lowest performance at most absolute distances, revealing that not all knowledge of source language is equally important to improve cross-lingual dependency performance. Second, compared with the "Pre-training" model, the "Fine-tuning" model achieves better performance at distances above 6, demonstrating that target language data can facilitate our model to capture the long dependency relationship. Finally, our model substantially enhances performances on all absolute distances, highlighting the importance of filtering source language information.

5 Conclusion

We propose a feature selection approach to emphasize useful representative features and ignore the useless ones, thus improving the performance of cross-lingual dependency parsing. our model not only exploits a representation alignment network that selectively filters advantageous source language representations at the input layer but also utilizes an adversarial network to strengthen context-invariant features within the encoding layer. Experiments on a benchmark dataset illustrate that our proposed model significantly outperforms several strong baseline models. Detailed comparative experiments show that both the alignment and adversarial networks can substantially facilitate extracting and utilizing relevant target language features, thereby increasing the adaptation capability of our model. Furthermore, in-depth analysis reveals that our model achieves notable improvements in parsing long-distance dependencies and exhibits robustness capabilities, confirming its comprehensive applicative value in cross-lingual settings.

Limitations

Our proposed representation alignment and adversarial networks require a high-quality bilingual dictionary to facilitate language associations through matrix alignment. Hence, when there exists a bilingual dictionary, our method can be easily adapted to other cross-lingual dependency parsing tasks. Meanwhile, our constructed Vietnamese-Chinese bilingual dictionary will be released to facilitate future researches.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of CoNLL*, pages 372–382.

Fadli Aulawi Al Ghiffari, Ika Alfina, and Kurniawati Azizah. 2023. Cross-lingual transfer learning for javanese dependency parsing. In *Proceedings of IJCNLP-AACL*, pages 1–9.

Somnath Basu Roy Chowdhury, Annervaz M, and Ambedkar Dukkipati. 2019. Instance-based inductive deep transfer learning by cross-dataset querying with locality sensitive hashing. In *Proceedings of DeepLo*, pages 183–191.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of EMNLP*, pages 2340–2354.

Guandan Chen, Kai Fan, Kaibo Zhang, Boxing Chen, and Zhongqiang Huang. 2021. Manifold adversarial augmentation for neural machine translation. In *Findings of ACL-IJCNLP*, pages 3184–3189.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of ACL*, pages 3098–3112.

Chinmay Choudhary and Colm O'riordan. 2023. Multi-lingual end-to-end dependency parsing with linguistic typology knowledge. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 12–21.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of EMNLP*, pages 1914–1925.

Cheikh M. Bamba Dione. 2021. Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba. In *Proceedings of IWPT*, pages 84–92.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR*.

Thomas Effland and Michael Collins. 2023. Improving low-resource cross-lingual parsing with expected statistic regularization. *TACL*, pages 122–138.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL*, pages 7548–7555.

Goran Glavaš and Ivan Vulić. 2021a. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of ACL-IJCNLP*, pages 4878–4888.

Goran Glavaš and Ivan Vulić. 2021b. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of ACL-IJCNLP*, pages 4878–4888.

Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pages 1–18.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of EACL*, pages 2760–2765.

Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2022. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of LREC*, pages 648–662.

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

Shanu Kumar, Soujanya Abbaraju, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2023. DiTTO: A feature representation imitation approach for improving cross-lingual transfer. In *Proceedings of EACL*, pages 385–406.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Ying Li, Shuaike Li, and Min Zhang. 2022. Semi-supervised domain adaptation for dependency parsing with dynamic matching network. In *Proceedings of ACL*, pages 1035–1045.

Ying Li, Zhenghua Li, Min Zhang, Rui Wang, Sheng Li, and Luo Si. 2019. Self-attentive biaffine dependency parsing. In *Proceedings of IJCAI*, pages 5067–5073.

Ying Li, Meishan Zhang, Zhenghua Li, Min Zhang, Zhefeng Wang, Baoxing Huai, and Nicholas Jing Yuan. 2021. APGN: Adversarial and parameter generation networks for multi-source cross-domain dependency parsing. In *Findings of EMNLP*, pages 1724–1733.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Xunliang Cai, Dongyan Zhao, Ran Wang, and Rui Yan. 2023a. Retrieval-based knowledge transfer: An effective approach for extreme large language model compression. In *Findings of EMNLP*, pages 8643–8657.

Lu Liu, Yi Zhou, Jianhan Xu, Xiaoqing Zheng, Kai-Wei Chang, and Xuan-Jing Huang. 2020. Cross-lingual dependency parsing by pos-guided word reordering. In *Findings of EMNLP*, pages 2938–2948.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei, and Hinrich Schuetze. 2023b. Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. In *Findings of EMNLP*, pages 8376–8401.

Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of ACM SIGKDD*, pages 641–647.

Menglong Lu, Zhen Huang, Yunxiang Zhao, Zhiliang Tian, Yang Liu, and Dongsheng Li. 2023. DaMSTF: Domain adversarial learning enhanced meta self-training for domain adaptation. In *Proceedings of ACL*, pages 1650–1668.

Alireza Mohammadshahi and James Henderson. 2021. Recursive non-autoregressive graph-to-graph transformer for dependency parsing with iterative refinement. *TACL*, 9:120–138.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *TACL*, 10:127–144.

Tong Niu, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. OneAligner: Zero-shot cross-lingual transfer with one rich-resource language pair for low-resource sentence retrieval. In *Findings of ACL*, pages 2869–2882.

Hiroaki Ozaki, Gaku Morio, Terufumi Morishita, and Toshinori Miyoshi. 2021. Project-then-transfer: Effective two-stage cross-lingual transfer for semantic dependency parsing. In *Proceedings of ACL*, pages 2586–2594.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of EMNLP*, pages 10186–10203.

Matthew Riemer, Sophia Krasikov, and Harini Srinivasan. 2015. A deep learning and knowledge transfer based architecture for social media user characteristic determination. In *Proceedings of SocialNLP*, pages 39–47.

Guy Rotman and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *TACL*, 7:695–713.

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 NAACL*, pages 1599–1613.

Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. 2022. Just fine-tune twice: Selective differential privacy for large language models. In *Proceedings of EMNLP*, pages 6327–6340.

Milan Straka, Jana Straková, and Jan Hajic. 2019. Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing. *CoRR*, abs/1908.07448.

Weiting Tan, Shuoyang Ding, Huda Khayrallah, and Philipp Koehn. 2022. Doubly-trained adversarial data augmentation for neural machine translation. In *Proceedings of AMTA*, pages 157–174.

Yuanhe Tian, Yan Song, and Fei Xia. 2022. Improving relation extraction through syntax-induced pre-training with dependency masking. In *Findings of ACL*, pages 1875–1886, Dublin, Ireland.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022. Udapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of EMNLP-IJCNLP*, pages 1105–1116.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of EMNLP*, pages 2649–2656.

Haoran Xu and Philipp Koehn. 2021. Zero-shot cross-lingual dependency parsing through contextual embedding transformation. In *Proceedings of Adapt-NLP*.

Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases, and Andrei Barbu. 2022. The aligned multimodal movie treebank: An audio, video, dependency-parse treebank. In *Proceedings of EMNLP*, pages 9531–9539.

Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. A graph-based model for joint Chinese word segmentation and dependency parsing. *TACL*, 8:78–92.

Nasser Zalmout and Nizar Habash. 2019. Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling. In *Proceedings of ACL*, pages 1775–1786.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of AAAI/ACM AIES*, pages 335–340.

Meishan Zhang, Zhenghua Li, Guohong Fu, and Min Zhang. 2019. Syntax-enhanced neural machine translation with syntax-aware word representations. In *Proceedings of NAACL-HLT*, pages 1151–1161.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022. SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of EMNLP*, pages 2518–2531.

Han Zou, Jianfei Yang, and Xiaojian Wu. 2021. Unsupervised energy-based adversarial domain adaptation for cross-domain text classification. In *Findings of ACL-IJCNLP*, pages 1208–1218.