Improving Unsupervised Multi-Lingual Dependency Parsing via the Dynamic Feature Alignment

Anonymous ACL submission

Abstract

Multi-lingual dependency parsing aims to leverage shared syntactic structures across languages to improve parsing accuracy in low-resource scenarios. However, direct transfer often yields suboptimal performance due to significant linguistic variations across diverse languages. To address this issue, we propose a novel approach for unsupervised multi-lingual dependency parsing via dynamic feature alignment. Specifically, we first construct multilingual aligned dependency treebanks by leveraging the collaborative annotation of multiple Large Language Models (LLMs). Subsequently, we design a dynamic feature alignment network to select beneficial syntactic features and filter out harmful ones automatically. Experiments on multiple benchmark datasets demonstrate that our proposed method significantly outperforms all strong baselines. In-depth comparison experiments confirm that dynamic feature alignment enables the model to adaptively fuse features from multiple high-source languages. Besides, detailed error analysis further validates that our designed feature selection strategy is suitable for dynamic parameter adaptation. Our code and data are available at https://github.com/**.

1 Introduction

Dependency parsing, as a foundational task in Natural Language Processing (NLP), aims to identify grammatical structures of the input text (Gan et al., 2022; Zhang et al., 2022; Aziz et al., 2024), which can help various downstream applications, including machine translation (Chen et al., 2017), information retrieval, and semantic analysis (Zhang et al., 2018). Recently, supervised dependency parsing on high-resource languages has achieved remarkable success. Specifically, the dependency parsing models in English and Chinese have exceeded 97% and 94% labeled attachment scores, respectively (Wang et al., 2021). However, these

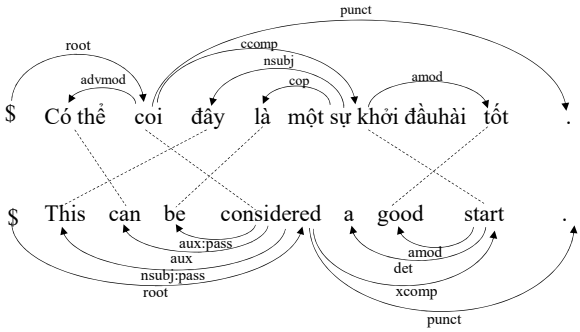


Figure 1: An example of Vietnamese-English aligned dependency trees.

supervised dependency parsing models heavily depend on the scale and quality of training data, which creates a severe bottleneck for the low-resource languages (Joshi et al., 2020). Unsupervised multi-lingual dependency parsing aims to induce syntactic structures directly from raw corpora by leveraging shared linguistic. Generally, existing approaches can be categorized into two primary paradigms: *generative approaches* and *discriminative approaches*. Generative approaches model the joint probability of the sentence and the parse tree (Yang et al., 2020; Yang and Tu, 2021; Lin et al., 2023). Yang et al. (2020) proposed a second-order neural dependency model to capture more complex syntactic structures. Lin et al. (2023) utilize specific prompts to guide LLMs in emitting linearized dependency trees in a zero-shot cross-lingual manner. Discriminative approaches exploit the latent syntactic knowledge within Pre-trained Language Models model the conditional probability given the sentence(Wu et al., 2020; Li and Tu, 2020; Srivastava et al., 2022). Wu et al. (2020) assess the impact of masking one word on another within BERT to derive a dependency tree without any parameter training. Li and Tu (2020) employ a CRF Autoencoder framework to facilitate parser adaptation from source to target languages, utiliz-

ing reconstruction objectives to filter noisy transfers. However, despite these progresses, effectively filtering structural noise caused by linguistic divergence remains a critical challenge, often leading to negative transfer in multilingual scenarios.

As illustrated in Figure 1, while the Vietnamese and the English sentences are perfectly aligned at the sentence level, their internal word-level dependencies are mismatched. Specifically, the syntactic structures and word order between the two languages exhibit significant divergences, a phenomenon rooted in typological variation that challenges standard alignment methods (Ahmad et al., 2019; Ponti et al., 2019). In such scenarios, directly fusing features by adding English syntactic information to the Vietnamese representation at the sentence level is problematic. Because the same position words and their grammatical roles do not line up, the source language features often act as noise rather than helpful guidance, leading to a drop in parsing performance on the target language. Consequently, when the source and target languages diverge syntactically, these models lack the flexibility to filter out irrelevant information, forcing alignment where none exists.

To address these challenges, we propose a novel framework for unsupervised multilingual dependency parsing via Dynamic Feature Alignment. Specifically, we first leverage a collaborative multi-LLM framework to construct aligned multilingual treebanks. Subsequently, we propose a fine-grained word-level alignment mechanism that utilizes cosine similarity to assign higher weights to semantically aligned features, effectively filtering out structural noise caused by misalignment. Furthermore, we introduce a dynamic gating parameter λ to regulate the information flow from different source languages. Optimized via a novel alignment-aware loss, this parameter is adaptively updated during training, enabling the model to automatically select the most beneficial source language representations for the target language. Experiments on four low-resource benchmark datasets demonstrate that our model achieves significant improvements, surpassing the strong shared-private baseline by an average of 1.68/1.72 points in LAS/UAS scores and establishing new state-of-the-art results. Comparative experiments confirm our method effectively filters structural noise while amplifying beneficial cross-lingual signals. Detailed analyses further validate that the word-level vector alignment and the alignment-aware loss, enabling the model to adaptively select optimal source languages and greatly enhance cross-lingual dependency parsing performance in data-scarce scenarios.

2 Related Work

2.1 Unsupervised Dependency Parsing.

Unsupervised dependency parsing aims to induce syntactic structures directly from raw text, serving as a critical foundation for low-resource scenarios (Klein and Manning, 2004; Han et al., 2020; Jiang et al., 2016). Osa et al. (2023) proposed a lightweight joint framework that simultaneously learns part-of-speech tagging and dependency parsing directly from raw text, challenging the necessity of pipelined supervision. Furthermore, the development of Large Language Models (LLMs) has significantly propelled the progress of this field (Zhang et al., 2024; Chen et al., 2024). For instance, Chen et al. (2024) proposed integrating Conditional Mutual Information with LLMs to interpret bi-lexical dependencies, effectively imposing grammatical constraints without gold annotations. Similarly, Zhang et al. (2024) demonstrated that LLMs can serve as robust data augmentors, generating high-quality pseudo-treebanks that boost parsing performance in data-scarce settings. In parallel to LLM-driven approaches, Choenni et al. (2023) explored structural adaptation by designing language-specific subnetworks, refining the transfer of universal parsers to specific target languages.

2.2 Multi-lingual Dependency Parsing

Multi-lingual Dependency Parsing aims to transfer syntactic knowledge from resource-rich languages to low-resource ones (Ruder et al., 2019; Pikuliak et al., 2021; Wu and Dredze, 2019). Most existing approaches rely on multi-lingual pre-trained language models to align different languages within a shared representation space, thereby facilitating cross-lingual transfer (**?**Liu et al., 2024, 2025b). Notably, Devlin et al. (2019) introduced mBERT, pre-trained on Wikipedia data from 104 languages, enabling effective zero-shot transfer. Conneau et al. (2020) proposed XLM-R, trained on 2.5TB of CommonCrawl data, producing more robust multi-lingual representations. To fully capitalize on these shared representation spaces, various advanced strategies have been proposed to optimize cross-lingual knowledge transfer (Kondratyuk and Straka, 2019; Wang et al., 2019;

Li et al., 2024; Liu et al., 2025a). Specifically, Kondratyuk and Straka (2019) concatenate treebanks from 75 languages and fine-tuned the UDify parser to capture universal syntactic patterns across languages. Wang et al. (2019) propose Cross-Lingual BERT Transformation, which maps source embeddings to target languages, reducing distributional shifts and outperforming static embeddings in zero-shot tasks.

2.3 Cross-Lingual Feature Alignment

Cross-lingual feature alignment aims to bridge the gap between resource-rich and low-resource languages by mitigating distributional shifts, thereby enabling effective generalization across unseen languages. Existing research can be categorized into shared representation spaces, contrastive learning, and language-specific adaptation. Shared representation spaces project multiple languages into a common semantic embedding layer to establish universal linguistic features without requiring direct parallel data (Conneau et al., 2020; Xue et al., 2021; Feng et al., 2022). Contrastive learning optimizes feature consistency by maximizing the similarity between semantically related cross-lingual examples while minimizing the distance between unrelated ones (Chen et al., 2020; Chi et al., 2021; Wei et al., 2021). Language-specific adaptation tailors cross-lingual models to accommodate unique linguistic variations using techniques such as adapters or meta-learning to enhance performance in low-resource settings (Peters et al., 2019; Pfeiffer et al., 2020; Hu et al., 2022). Despite these advancements, existing methods predominantly operate at a static, coarse-grained level, which often indiscriminately aligns noisy features alongside beneficial ones. To address this, we propose Dynamic Feature Alignment, a novel framework that leverages word-level cosine similarity to filter structural noise and employs an adaptive gating mechanism to selectively regulate cross-lingual information flow.

3 Our Approach

Multi-lingual dependency parsing aims to leverage linguistic knowledge from multiple source languages to enhance the parsing performance of a target language. However, the direct incorporation of multiple languages often introduces harmful interference due to significant differences in lexical and syntactic distributions across languages. To address this issue, we propose an unsupervised multilingual dependency parsing via dynamic feature alignment. By designing a dynamic alignment network, our method effectively extracts beneficial linguistic features while filtering out irrelevant information. Figure 2 illustrates the overall architecture of our proposed framework, which is structured into three distinct stages: *pseudo treebank generation*, *basic model parser pre-training*, and *aligned model parser fine-tuning*. Initially, the *pseudo treebank generation* stage leverages a multi-LLM collaborative framework to construct high-quality aligned multilingual treebanks. Subsequently, *basic model parser pre-training* utilizes these generated treebanks to establish a robust shared semantic space. Finally, *aligned model parser fine-tuning* facilitates fine-grained syntactic knowledge transfer by adaptively weighting contributions from different source languages. Both stages consist of Input & Embedding, Encoder, and Decoder layers.

3.1 Pseudo Treebanks Generation.

To acquire a high-quality alignment high-quality dependency parsing treebank, we first employ Chain-of-Thought (CoT) prompting on two distinct base Large Language Models (LLMs) to generate initial dependency parsing trees. Subsequently, we design a specialized prompt to guide an Expert LLM in synthesizing these initial trees. By analyzing structural discrepancies and integrating the complementary strengths of the base models, the Expert LLM produces a refined, high-quality dependency parsing treebank.

3.2 Input & Embedding

Given the aligned sequence $w_1, w_2, \ldots, w_n$ from any dependency parsing tree of these dependency parsing treebanks, the input layer converts them into high-dimensional vectors. As illustrated in equation 1, each word vector $\mathbf{X}_*^i$ comprises a composite word representation and a corresponding character representation $\mathbf{Char}_*^i$. The word representation is formed by the summation of the XLM-RoBERTa output $\mathbf{Rep}_*^i$ and a randomly initialized embedding $\mathbf{Emb}_*^i$. The character representation $\mathbf{Char}_*^i$ is generated by a Char-BiLSTM network, which utilizes a one-layer BiLSTM to encode the constituent characters of each word w_i and merges the hidden states from two directions. Consequently, we obtain the unified input vectors

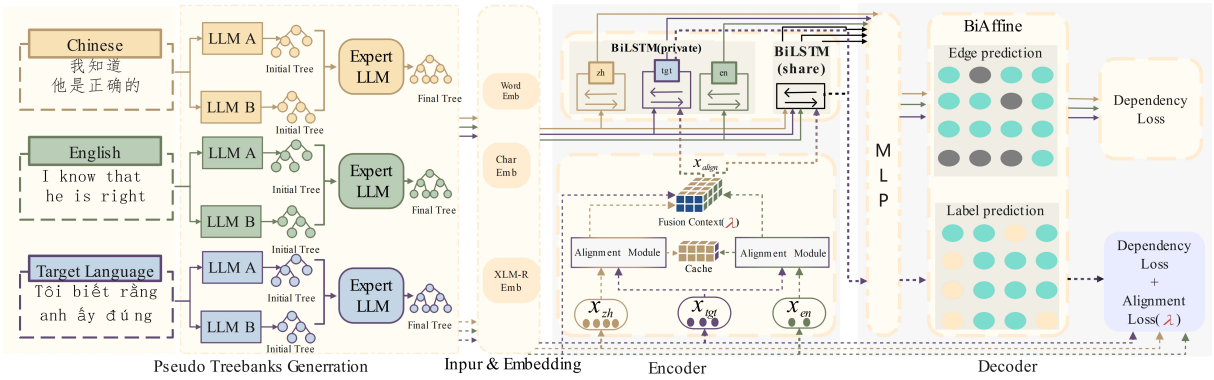


Figure 2: The overall architecture of our method. The solid lines indicate pre-training, and the dashed lines are alignment training.

$\mathbf{X}_{\text{zh}}$, $\mathbf{X}_{\text{en}}$, and $\mathbf{X}_{\text{tar}}$ for subsequent layers:

$$\mathbf{X}^{\text{i}}_{*} = (\mathbf{Rep}^{\text{i}}_{*} + \mathbf{Emb}^{\text{i}}_{*}) \oplus \mathbf{Char}^{\text{i}}_{*} \tag{1}$$

3.3 Encoder Layer with Dynamic Word-Level Vector Alignment (λ)

Throughout both training stages, we employ a Shared-Private BiLSTM structure to capture common and specific features. In the second stage, we further augment this process by introducing a word-level dynamic vector alignment mechanism, allowing the model to bridge the cross-lingual gap more effectively during alignment training.

Word-Level Dynamic Vector Alignment. This module operates primarily during the second stage (alignment fine-tuning) to integrate syntactic knowledge from high-resource languages into the target language at a fine-grained word level. Initially, during the Chinese (*zh*) and English (*en*) training phases, the respective word vectors are cached. Subsequently, in the target (*tar*) phase, these cached vectors are loaded to calculate similarity matrices. Specifically, for each i-th word w_i^{tar} in the target sentence, we compute its cosine similarity against every j-th word w_j^{zh} in the aligned Chinese sentence and every k-th word w_k^{en} in the English sentence, as defined in equation 2.

$$\begin{aligned} \text{sim}_{\text{tz}}(i,j) &= \frac{\mathbf{tw}_i^{\top}\mathbf{zw}_j}{\|\mathbf{tw}_i\|\|\mathbf{zw}_j\|} \\ \text{sim}_{\text{te}}(i,k) &= \frac{\mathbf{tw}_i^{\top}\mathbf{ew}_k}{\|\mathbf{tw}_i\|\|\mathbf{ew}_k\|} \end{aligned} \tag{2}$$

From this, we construct the Target-Chinese similarity matrix $\mathbf{CS}_{\text{tz}} \in \mathbb{R}^{n\times m}$ and the Target-English similarity matrix $\mathbf{CS}_{\text{te}} \in \mathbb{R}^{n\times h}$, where n is the number of words in the target language sentence, and m and h are the word counts in the corresponding Chinese and English sentences. For each target language word tw_i, we select the Top-k most similar words in the matrices $\mathbf{CS}_{\text{tz}}[i,:]$ and $\mathbf{CS}_{\text{te}}[i,:]$, denoted as equation 3

$$\begin{aligned} I_{\text{zh}}(i) &= \text{TopK}(\mathbf{CS}_{\text{tz}}[i,:],k) \\ I_{\text{en}}(i) &= \text{TopK}(\mathbf{CS}_{\text{te}}[i,:],k) \end{aligned} \tag{3}$$

Next, we apply Softmax normalization to these Top-k similarity values to generate normalized attention weights as shown in equation 4.

$$\begin{aligned} \text{wei}_{\text{zh}}[j] &= \frac{\exp(\mathbf{CS}_{\text{tz}}[i, I_{\text{zh}}(i)[j]])}{\sum_{j'=1}^{k}\exp(\mathbf{CS}_{\text{tz}}[i, I_{\text{zh}}(i)[j']])} \\ & \quad j = 1,\ldots,k \\ \text{wei}_{\text{en}}[k] &= \frac{\exp(\mathbf{CS}_{\text{te}}[i, I_{\text{en}}(i)[k]])}{\sum_{k'=1}^{k}\exp(\mathbf{CS}_{\text{te}}[i, I_{\text{en}}(i)[k']])} \\ & \quad k = 1,\ldots,k \end{aligned} \tag{4}$$

We perform a weighted aggregation of the retrieved features. This process generates the context vectors $\mathbf{C}_i^{\text{zh}}$ and $\mathbf{C}_i^{\text{en}}$ by selectively amplifying the signals of highly aligned words while suppressing less relevant ones. To further enhance computational efficiency and facilitate reuse, these vectors are stored in a cache, as formulated in equation 5:

$$\begin{aligned} \mathbf{C}_i^{\text{zh}} &= \sum_{j=1}^{k}\text{wei}_{\text{zh}}[j]\cdot\mathbf{zw}_{I_{\text{zh}}(i)[j]} \\ \mathbf{C}_i^{\text{en}} &= \sum_{m=1}^{k}\text{wei}_{\text{en}}[m]\cdot\mathbf{ew}_{I_{\text{en}}(i)[m]} \end{aligned} \tag{5}$$

Subsequently, we employ a dynamic gating parameter λ to perform a weighted fusion yielding the final fused representation fusion_i as 6:

$$\text{fusion}_i = \lambda \cdot \mathbf{C}_i^{\text{zh}} + (1-\lambda) \cdot \mathbf{C}_i^{\text{en}} \quad (6)$$

Finally, the fused vector is added to the original target language word vector to obtain the aligned enhanced representation:

$$x_{\text{align},i} = \mathbf{tw_i} + \text{fusion}_i \quad (7)$$

This process is executed in parallel for all target language words $i = 1, \ldots, n$, and the final output is the enhanced word vector sequence $X_{\text{align}} \in \mathbb{R}^{B \times L \times D}$.

Shared & Private Encoder. Unlike traditional BiLSTM-based encoders, we utilize a hybrid encoding architecture that employs both a shared BiLSTM and multiple language-specific private BiLSTMs to extract common and specific features. Exhibit in equation 8, the private BiLSTM captures unique language-specific patterns, denoted as $\mathbf{private}_i$, while the shared BiLSTM extracts universal linguistic characteristics $\mathbf{shared}_i$. Finally, the language-specific and common features are fused via addition to obtain the final representation $\mathbf{H}_i$:

$$\begin{aligned} \mathbf{private}_i &= \text{BiLSTM}_{\text{pri}}(\mathbf{x}_i; \theta_{\text{pri}}) \\ \mathbf{shared}_i &= \text{BiLSTM}_{\text{sha}}(\mathbf{x}_i; \theta_{\text{sha}}) \\ \mathbf{H}_i &= \mathbf{private}_i + \mathbf{shared}_i \end{aligned} \quad (8)$$

where θ_{pri} and θ_{sha} represent the learnable parameters of the private and shared BiLSTM modules, respectively.

3.4 Decoder with Vector Space Alignment Loss (λ)

The decoder predicts dependency arcs and relation labels using a Deep Biaffine architecture. In the first stage, the model is optimized via a conventional dependency parsing loss. In the second stage, we incorporate a novel alignment-aware loss that explicitly enforces semantic consistency between the target language and its high-resource counterparts, guided by the dynamic gating parameter λ.

MLP and Biaffine Layer. The decoder employs an MLP layer to project the enhanced contextualized vector $\mathbf{H}_i$ into lower-dimensional spaces, extracting the head representation $\mathbf{r}_i^{\text{h}}$ and the modifier representation $\mathbf{r}_i^{\text{d}}$ for each word w_i:

$$\begin{aligned} \mathbf{r}_i^{\text{h}} &= \text{MLP}_h\left(\mathbf{H}_i\right) \\ \mathbf{r}_i^{\text{d}} &= \text{MLP}_d\left(\mathbf{H}_i\right) \end{aligned} \quad (9)$$

where MLP_h and MLP_d share a single hidden layer with ReLU activation. Subsequently, a BiAffine layer computes the arc score $\text{score}(i \leftarrow j)$ between word w_i and w_j. Simultaneously, a separate BiAffine layer calculates the label score $\text{score}(i \overset{l}{\leftarrow} j)$:

$$\begin{aligned} \text{score}(i \leftarrow j) &= \left[\mathbf{r}_i^{\text{d}}; 1\right]^\top \mathbf{U}_1 \mathbf{r}_j^{\text{h}} \\ \text{score}(i \overset{l}{\leftarrow} j) &= \mathbf{r}_j^{\text{h}} \mathbf{U}_2 \mathbf{r}_i^{\text{d}} + (\mathbf{r}_j^{\text{h}} \oplus \mathbf{r}_i^{\text{d}}) \mathbf{U}_3 + b \end{aligned} \quad (10)$$

where $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ and b are learnable parameters. At inference, the Maximum Spanning Tree algorithm is applied to identify the highest-scoring tree as the final parsing result.

Vector Space Alignment Loss (λ). In the first training stage, the model is optimized using the classic dependency parsing loss $\mathcal{L}_{\text{par}}$:

$$\begin{aligned} \mathcal{L}_{\text{par}} = &- \log \frac{e^{\text{score}(i \leftarrow j)}}{\sum_{0 \le k \le n, k \ne i} e^{\text{score}(i \leftarrow k)}} \\ &- \log \frac{e^{\text{score}(i \overset{l}{\leftarrow} j)}}{\sum_{l' \in \mathcal{L}} e^{\text{score}(i \overset{l'}{\leftarrow} j)}} \end{aligned} \quad (11)$$

where j and l denote the gold-standard head and relation label for word w_i, respectively.

In the second stage, we introduce a cross-lingual alignment loss $\mathcal{L}_{\text{align}}$ to refine semantic consistency. This loss constrains the distance between the target language vectors X_{tar} and the cached high-resource vectors (X_{zh}, X_{en}) using the Frobenius norm:

$$\begin{aligned} \mathcal{L}_{\text{align}} = &\lambda \cdot \|X_{\text{zh}} - X_{\text{tar}}\|_F^2 \\ &+ (1-\lambda) \cdot \|X_{\text{en}} - X_{\text{tar}}\|_F^2 \end{aligned} \quad (12)$$

The dynamic gating parameter λ is adaptively learned via a Sigmoid function:

$$\lambda = 0.1 + 0.8 \cdot \sigma(\theta_\lambda), \quad \sigma(x) = \frac{1}{1+e^{-x}} \quad (13)$$

This λ is shared with the *Dynamic Vector Alignment* module and is jointly optimized via backpropagation, allowing the model to adaptively determine the semantic affinity between the target and source languages. The total objective function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{par}} + \mathcal{L}_{\text{align}} \quad (14)$$

The model is saved once $\mathcal{L}_{\text{total}}$ reaches convergence.

Dataset	Train	Dev	Test	All
Universal dependency Treebanks				
Chinese(GSDSimp)	-	500	500	1,000
English (EWT)	-	2,001	2,077	4,078
Vietnamese (VTB)	-	1,123	800	1,923
Tamil (TTB)	-	80	120	180
Telugu (MTG)	-	131	146	277
Maltese (MUDT)	-	433	518	951
FLORES-200 parallel sentences				
Chinese	2,000	-	-	2,000
English	2,000	-	-	2,000
Vietnamese	2,000	-	-	2,000
Tamil	2,000	-	-	2,000
Telugu	2,000	-	-	2,000
Maltese	2,000	-	-	2,000

Table 1: Dataset statistics in sentence number.

4 Experiments

4.1 Experimental Setups

Datasets. 1) Datasets for parser training. We collected high-quality parallel sentences from FLORES-200 [1] as the source of our six-language (Chinese, English, Vietnamese, Tamil, Telugu, and Maltese) synthetic treebanks, During training, Chinese and English are leveraged as high-resource source languages to enhance the syntactic parsing performance of the other low-resource languages."

2) Datasets for method validation. We collected validation and test sets for four languages to optimize model performance, i.e., Vietnamese (vi), Tamil (ta), Telugu (te), and Maltese (mt), all of which are derived from the Universal Dependencies (UD) v2.13 treebanks [2]. Detailed dataset statistics are presented in Table 1.

Evaluation. We utilize Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) as evaluation metrics (Liu et al., 2025b). All models are trained for no more than 500 iterations, and their performances are evaluated on the development dataset after each iteration to guide the model selection. Model training is stopped if the peak performance does not increase for 50 consecutive iterations.

Baselines. We reproduce the following baseline models for our experiments. ***Full Shared Model (FulSha).*** We employ fully shared encoder parameters across three dependency graph formalisms to capture cross-formalism commonalities, thereby enhancing heterogeneous dependency parsing (Peng et al., 2017). ***Language Embedding Model (LanEmb).*** Leveraging the insight that injecting domain embeddings as auxiliary inputs improves cross-domain parsing by informing the model of domain-specific characteristics and introducing 8-dimensional language embeddings to explicitly encode language identity, guiding the model in distinguishing between different language structures (Li et al., 2019). ***Shared and Private Model (ShaPri).*** Adopting the shared-private framework (Liu et al., 2017), we decompose feature extraction into two distinct streams. We use a single shared XLM-RoBERTa to capture universal linguistic patterns across all languages, while assigning a specific private XLM-RoBERTa for each language to extract its unique, language-specific characteristics. ***w/ roberta.*** For all typical models above, we use the XLM-RoBERTa-base [3] pre-training model to extract the corresponding feature representations of the input words and add them to the random word embeddings of the above models to enhance the contextual representation of the words.

4.2 Main Results

Table 2 presents our main results and a comparison with prior works. First, we find that the parsing accuracies of all baselines improve significantly after integrating our method, illustrating that our proposed mechanism universally enables more effective cross-lingual knowledge transfer. Second, we observe that *Our LanEmb* slightly underperforms *Our FulSha*. This indicates that injecting language embedding provides insufficient discriminative power for the model to capture complex linguistic structures, thereby limiting the efficacy of dynamic fusion. In contrast, *Our ShaPri* achieves the most significant improvement and the best overall performance. This demonstrates that explicitly disentangling language-invariant and language-specific features creates a more distinct representation space, allowing our dynamic gating mechanism to more precisely calibrate the balance between cross-lingual transfer and target-specific adaptation.

In addition, we compare our approach with several previous works, i.e., *ESR* (Effland and Collins, 2023) and *Subnet* (Choenni et al., 2023). *ESR* introduces regularization constraints and *Subnet* employs subnet selection for transfer. Our model sets a new state-of-the-art among unsupervised meth-

[1] https://github.com/facebookresearch/flores/tree/main/flores200

[2] https://universaldependencies.org/

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

Model	Train Mode	Vietnamese		Tamil		Telugu		Maltese		Avg.	
		LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
Results of previous works											
Subnet(2023)	Unsupervised	44.62	-	51.48	-	-	-	30.05	-	42.05	-
ESR(2023)	Unsupervised	52.2	-	-	-	-	-	33.40	-	42.80	-
ESR(2023)	Semi-Supervised	60.80	-	66.40	-	80.10	-	74.20	-	70.38	-
Compare with traditional models											
FulSha	Unsupervised	59.78	75.78	58.12	72.56	71.15	89.32	63.88	75.55	63.23	78.30
LamEmb	Unsupervised	59.61	75.59	57.82	72.65	71.71	89.32	63.84	75.31	63.25	78.22
ShaPri	Unsupervised	60.67	76.33	59.08	73.10	72.68	89.46	66.25	76.90	64.67	78.95
Our FulSha	Unsupervised	61.33	77.58	60.96	75.12	72.41	89.88	66.95	77.43	65.41	80.00
Our LamEmb	Unsupervised	61.17	77.58	59.73	75.12	73.37	89.88	65.54	77.43	64.95	80.00
Our ShaPri	Unsupervised	**61.92**	**77.67**	**62.18**	**76.50**	**73.49**	**90.11**	**67.79**	**78.41**	**66.35**	**80.67**

Table 2: Main results of four languages on the test dataset.

Model	Tamil	
	LAS	UAS
Ours	**62.18**	**76.50**
w/o alignment loss	61.56	75.86
w/o all	60.96	75.12
w/o word-level vector alignment	59.76	74.33

Table 3: Ablation study on the Tamil dev-dataset.

ods and achieves performance superior to even certain semi-supervised approaches. These results highlight the potential of leveraging dynamic representation fusion to achieve high-quality parsing in low-resource languages.

4.3 Ablation Study

Table 3 presents the ablation study results on the Tamil development set. First, the removal of the alignment loss leads to a noticeable performance drop. This confirms that our alignment-aware loss, guided by the dynamic parameter λ, effectively allows the parser to adaptively calibrate semantic weights. Furthermore, removing the word-level vector alignment causes the performance to decline even further, falling below the baseline. This significant decrease demonstrates that direct sentence-level fusion may introduce noise or harmful interference, thereby highlighting the necessity of our fine-grained word-level alignment approach for capturing precise cross-lingual features.

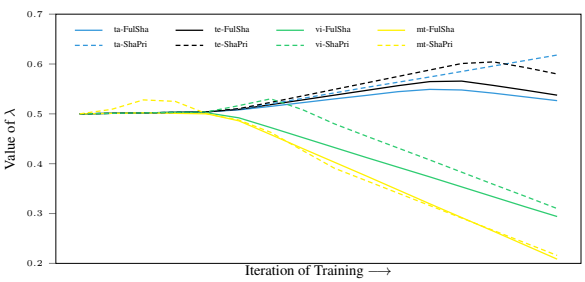


Figure 3: Evolution of the gating parameter λ

4.4 Compare Study

Figure 3 illustrates the evolutionary lines of the gating parameter λ. Overall, λ undergoes continuous adaptation across all languages. For Vietnamese and Maltese, the parameter exhibits a consistent downward trend, indicating a progressive shift towards English-derived representation. In contrast, Tamil and Telugu display sustained trends, reflecting a continued reliance on Chinese-derived representations. This highlights the parsers' capacity to dynamically recalibrate their reliance on cross-lingual knowledge to optimize parsing performance. Ultimately, our mechanism tailors the fusion process to each target, dynamically balancing the integration of external knowledge with intrinsic linguistic representations.

4.5 Error Analysis

Word-level alignment analysis. Figure 4 illustrates the cosine similarity scores between word-level embeddings of Tamil and its corresponding Chinese (left) and English (right). Notably, structural differences across these languages cause the highest cosine similarity score for a specific Tamil word to correspond to Chinese and English tokens at different positions. This misalignment indicates that sentence-level mixing introduces significant semantic noise, potentially degrading model performance.

Optimal topk analysis. Figure 5 presents the experimental results under two distinct thresholding strategies. The left illustrates the number of selected tokens (k), while the right depicts the effect of varying the cosine similarity (CS) threshold.

From an overall perspective, the parsing performance exhibits a consistent upward trend as the number of tokens allocated for fusion increases. Optimal performance is achieved when all tokens

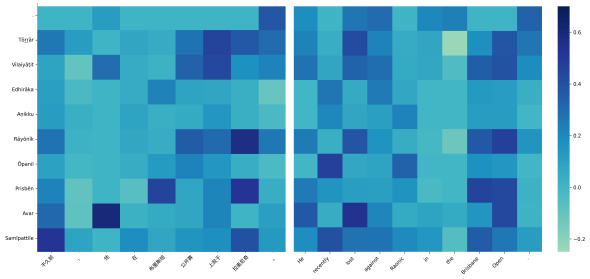


Figure 4: Heatmap of cosine similarity scores between word-level embeddings of Tamil and its corresponding Chinese (left) and English (right) parallel sentences.

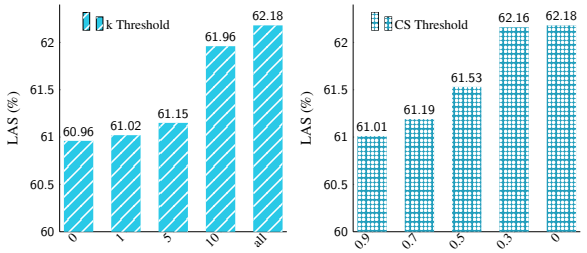


Figure 5: Performance comparison across different word-level fusion thresholds.

are integrated into the fusion process. Moreover, under highly restrictive constraints, i.e., $k = 1$ and $CS = 0.9$, the performance gains are negligible. This indicates that the single token is insufficient to capture the complex semantic information. Ultimately, these results demonstrate that full weighted fusion is superior to restrictive filtering for mitigating cross-lingual discrepancies.

Weight parameter analysis. Figure 6 illustrates the efficacy and necessity of parameter learning for the fusion weight λ. First, it is observed that integrating information from high-resource languages with varying fixed weights consistently yields performance improvements over the baseline, further validating the effectiveness of word-level vector fusion. Moreover, our proposed method achieves the optimal result among all configurations. This demonstrates that dynamic parameter learning outperforms manual tuning, as it helps the model to adaptively determine the optimal balance.

Manual assessment. Table 4 shows the manual assessment scores for dependency labels in Tamil. Our ShaPri model consistently improves the prediction precision across all evaluated labels. Notably, labels such as *advmod*, *det*, and *root* exhibit the most significant gains, indicating that our approach effectively captures local functional and modifier information while gaining a better grasp of the core skeletal structure of the sentence.

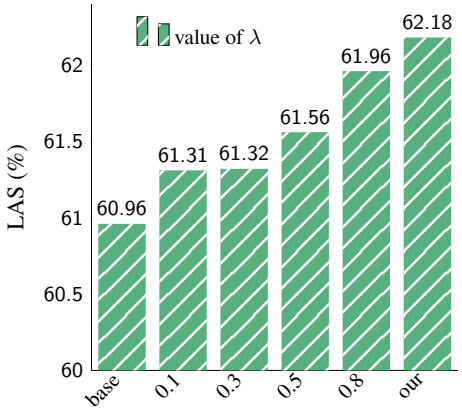


Figure 6: Performance comparison across different weight parameters.

Dep_labels	Accuracy (%)		
	ShaPri	Our ShaPri	Improvement
advmod	63.16	72.73	**+9.57**
det	67.74	75.86	**+8.12**
root	82.44	87.90	**+5.46**
nsubj	72.79	76.92	**+4.13**
advcl	61.25	64.20	**+2.95**
case	48.10	50.70	**+2.60**
conj	74.19	76.88	**+2.69**
ccomp	53.57	55.45	**+1.88**
mark	73.17	74.68	**+1.51**
punct	97.91	98.88	**+0.97**
amod	18.50	19.32	**+0.82**
obj	71.43	72.23	**+0.80**
nmod	80.00	80.61	**+0.61**
nummod	94.64	94.84	**+0.20**
aux	83.77	83.95	**+0.18**
obl	59.11	59.20	**+0.09**

Table 4: Manual assessment of dependency label accuracy in Tamil.

5 Conclusion

We propose a novel unsupervised multi-lingual dependency parsing approach that integrates a collaborative multiple large language models architecture with a dynamic feature alignment network. Benchmark experiments demonstrate that our approach consistently improves multi-lingual dependency parsing performance, leading to state-of -the-art results. In-depth comparisons confirm that our approach adaptively selects optimal high-resource languages. Furthermore, manual evaluations confirm that our method effectively improves the accuracy across various dependency labels. Detailed analysis indicates that dynamic parameter learning and comprehensive word-level fusion are essential for effectively mitigating cross-lingual knowledge.

Limitations

Although we have validated the quality of our synthetic data through three typical cross-lingual models, model confidence evaluation, and manual assessment, additional validation strategies may be necessary. These will be explored in future work.

References

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

Makera Moayad Aziz, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. 2024. Corenlp dependency parsing and pattern identification for enhanced opinion mining in aspect-based sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, 36(4):102035.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of ACL*, pages 1936–1945.

Junjie Chen, Xiangheng He, and Yusuke Miyao. 2024. Language model based unsupervised dependency parsing with conditional mutual information and grammatical constraints. In *Proceedings of NAACL*, pages 6355–6366.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of NAACL*, pages 3576–3588.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, 49(3):613–641.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Ding, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

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

Thomas Effland and Michael Collins. 2023. Improving low-resource cross-lingual parsing with expected statistic regularization. *TACL*, 11:122–138.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embeddings. In *Proceedings of ACL*, pages 873–891.

Leilei Gan, Yuxian Meng, Kun Kuang, Xiaofei Sun, Chun Fan, Fei Wu, and Jiwei Li. 2022. Dependency parsing as mrc-based span-span prediction. In *Proceedings of ACL*, pages 2427–2437.

Wenjuan Han, Yong Jiang, Hwee Tou Ng, and Kewei Tu. 2020. A survey of unsupervised dependency parsing. In *Proceedings of COLING*, pages 2522–2533.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the ICLR*.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *Proceedings of EMNLP*, pages 763–771.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL*, pages 478–485.

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

Ying Li, Jianjian Liu, Zhengtao Yu, Shengxiang Gao, Yuxin Huang, and Cunli Mao. 2024. Representation alignment and adversarial networks for cross-lingual dependency parsing. In *Findings of the EMNLP*, pages 7687–7697.

Zhao Li and Kewei Tu. 2020. Unsupervised cross-lingual adaptation of dependency parsers using CRF autoencoders. In *Findings of EMNLP*, pages 3025–3035.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of ACL*, pages 2386–2395.

Boda Lin, Xinyi Zhou, Binghao Tang, Xiaocheng Gong, and Si Li. 2023. Chatgpt is a potential zero-shot dependency parser. In *Findings of EMNLP*, pages 10799–10813.

Chaoqun Liu, Yuxin Lai, Zihan Yu, Edoardo Maria Ponti, Trevor Cohn, and Lucia Specia. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

Jianjian Liu, Ying Li, Zhengtao Yu, Shun Su, Shengxiang Gao, and Yuxin Huang. 2025a. Memory-enhanced large language model for cross-lingual dependency parsing via deep hierarchical syntax understanding. In *Findings of EMNLP*, pages 1910–1923.

Jianjian Liu, Zhengtao Yu, Ying Li, Yuxin Huang, and Shengxiang Gao. 2025b. Dynamic syntactic feature filtering and injecting networks for cross-lingual dependency parsing. In *Proceedings of the AAAI*, volume 39, pages 24614–24622.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of ACL*.

Yusuke Osa, Daisuke Kawahara, and Sadao Kurohashi. 2023. Multilingual parsing from raw text with a lightweight joint part-of-speech tagger and dependency parser. In *Findings of the EACL*, pages 621–638.

Hao Peng, Sam Thomson, and Noah A Smith. 2017. Deep multitask learning for semantic dependency parsing. In *Proceedings of ACL*, pages 2037–2048.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of RepL4NLP*, pages 7–14.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the EMNLP*, pages 7654–7673.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of NAACL*, pages 15–18.

Aarohi Srivastava, Subendhu Rongali, Andrew Drozdov, and Andrew McCallum. 2022. Unsupervised parsing with S-DIORA: Single tree encoding for deep inner attention. In *Proceedings of EMNLP*, page System Demonstrations.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated concatenation of embeddings for structured prediction. In *Proceedings of ACL-IJCNLP*, pages 2643–2660.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of EMNLP-IJCNLP*, pages 5721–5727.

Xiangpeng Wei, Rongxiang Huang, Yu Wang, Xinyu Dai, and Jiajun Chen. 2021. On learning universal representations across languages. In *Proceedings of ICLR*.

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

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for forest-based saliency. In *Proceedings of ACL*, pages 415–425.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*, pages 483–498.

Songlin Yang, Yong Jiang, Wenjuan Han, and Kewei Tu. 2020. Second-order unsupervised neural dependency parsing. In *Proceedings of COLING*, pages 3436–3447.

Songlin Yang and Kewei Tu. 2021. Neural bi-lexical prediction for unsupervised dependency parsing. In *Proceedings of ACL*, pages 2661–2667.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. LLM-assisted data augmentation for chinese dialogue-level dependency parsing. *Computational Linguistics*, 50(3):867–891.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of COLING*, pages 4212–4227.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees for relation extraction. In *Proceedings of EMNLP*, pages 2205–2215.

A Prompt Design for Initial Tree Generation

Table 5 illustrates the prompt template used for generating initial dependency trees. This prompt employs a Chain-of-Thought mechanism that guides the Large Language Model to sequentially identify the global sentence structure.

Prompt Template for Initial Generation

[Role] You are a linguistic expert specializing in dependency syntactic parsing.

[Task] Utilizing your linguistic knowledge, perform a rigorous dependency parsing on the input sentence.

[CoT]
1. **Global Analysis:** Identify the core clause structure, determining the main predicate.
2. **Local Analysis:** Examine the contextual information of each remaining node to determine its dependency relations.
3. **Synthesis:** Integrate all structures to construct the complete dependency parsing tree.

Example

[Input] What if Google morphed into GoogleOS?

[Output]

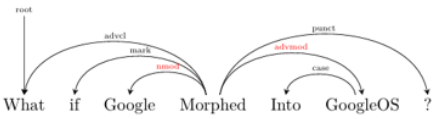


Table 5: The prompt template designed for initial dependency tree generation.

B Prompt Design for Final Tree Generation

Table 6 illustrates the prompt template for the final tree generation stage. This prompt guides an Expert LLM to compare the initial dependency trees, resolve structural discrepancies, and synthesize a high-quality dependency tree.

Prompt Template for Expert Synthesis

[Role] You are a senior linguistic expert and adjudicator in dependency syntactic parsing.

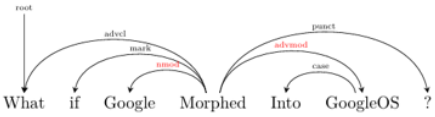
[Task] Analyze the structural discrepancies between these initial trees and integrate their complementary strengths to synthesize a single, high-quality dependency tree.

[CoT]
1. **Comparative Analysis:** Compare the two initial trees to identify structural discrepancies.
2. **Conflict Resolution:** For each discrepancy, analyze the linguistic context to determine which initial tree offers the correct syntactic relation, or propose a better one if both are incorrect.
3. **Optimal Synthesis:** Merge the verified edges into a final, refined dependency tree.

Example

[Sentence] What if Google morphed into GoogleOS?

[Initial Tree 1]



[Initial Tree 2]

[Output]

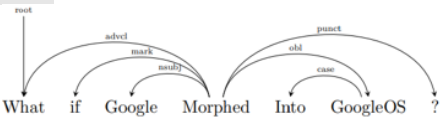


Table 6: The prompt template designed for the final tree generation.