# DiaDP@XLLM25: Advancing Chinese Dialogue Parsing via Unified Pretrained Language Models and Biaffine Dependency Scoring

**Shuoqiu Duan[a], Xiaoliang Chen[a,b]*, Duoqian Miao[b], Xu Gu[c], Xianyong Li[a], Yajun Du[a]**

[a]School of Computer and Software Engineering, Xihua University, Chengdu 610039, P. R. China

[b]College of Electronic and Information Engineering, Tongji University, Shanghai 201804, P. R. China

[c]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, P. R. China

duanshuoqiu@stu.xhu.edu.cn, dqmiao@tongji.edu.cn
chenxl@mail.xhu.edu.cn, guxu24@mails.ucas.ac.cn
lixy@mail.xhu.edu.cn, duyajun@mail.xhu.edu.cn

## Abstract

Dialogue-level dependency parsing is crucial for understanding complex linguistic structures in conversational data, yet progress has been hindered by limited annotated resources and inadequate modeling of dialogue dynamics. Existing methods often fail to capture both intra- and inter-utterance dependencies effectively, particularly in languages like Chinese with rich contextual interactions. To address these challenges, we propose InterParser, a novel framework that integrates a pretrained language model (PLM), bidirectional GRU (BiGRU), and biaffine scoring for comprehensive dependency parsing. Our model encodes token sequences using a PLM, refines representations via deep BiGRU layers, and employs separate projections for "head" and "dependent" roles to optimize arc and relation prediction. For cross-utterance dependencies, speaker-specific feature projections are introduced to enhance dialogue-aware scoring. Joint training minimizes cross-entropy losses for both intra- and inter-utterance dependencies, ensuring unified optimization. Experiments on a standard Chinese benchmark demonstrate that InterParser significantly outperforms prior methods, achieving state-of-the-art labeled attachment scores (LAS) for both intra- and inter-utterance parsing.

## 1 Introduction

Dialogue-level dependency parsing is crucial for enhancing the capabilities of dialogue understanding systems. This approach seeks to create a unified tree structure that captures both intra-sentence syntactic dependencies and inter-utterance discourse relations. While sentence-level dependency parsing has been extensively researched in languages such as English and Chinese (Xue et al., 2005; Jiang et al., 2018), applying this approach to multi-turn dialogues presents unique challenges. Dialogues inherently involve complex hierarchical interactions: within utterances, there are syntactic dependencies (e.g., subject-verb-object structures), and across utterances, there are discourse dependencies (e.g., question-answer pairs or causal reasoning). In Chinese, a language characterized by flexible word order and context-dependent semantics, performing such hierarchical parsing is particularly challenging.The root nodes of each subtree are often predicates that reflect single semantic events, illustrated in Figure 1.

Recent advancements have started to bridge this gap. Jiang et al. (2023) initiated the development of the Chinese Dialogue-level Dependency Treebank (CDDT), which integrates syntactic dependencies from existing treebanks (Jiang et al., 2018) and discourse relations based on Rhetorical Structure Theory (RST). They proposed rule-based signal detection and pseudo-labeling strategies to address data scarcity in resource-limited scenarios. However, this method's reliance on heuristic transformations, such as mapping syntactic 'root' nodes to discourse dependencies, and its multi-step pipeline process, can lead to error propagation and limited generalization. Subsequently, Zhang et al. (2024) introduced a Large Language Model (LLM)-assisted data augmentation technique, generating synthetic dialogues through various perturbations at the word, syntax, and discourse levels. Although effective, this approach necessitates extensive prompt engineering and can struggle to maintain structural consistency between the generated text and the corresponding dependency labels.

To address these limitations, we introduce an innovative end-to-end neural architecture tailored for Chinese dialogue-level dependency parsing. This approach circumvents the need for intermediate rule-based procedures by integrating the modeling of both intra- and inter-utterance dependencies through a cohesive feature learning mechanism.

* Corresponding author



*Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 266–273
August 1, 2025 

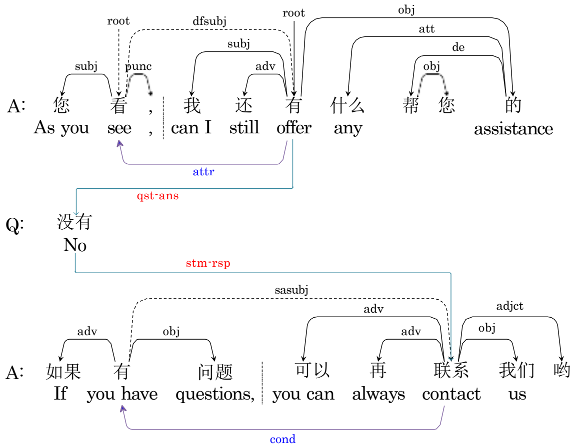


Figure 1: Example of Dialogue-Level Dependency Parsing: Vertical dashed lines indicate EDU boundaries, with arcs above words representing intra-EDU dependencies and arcs below or crossing utterances indicating inter-EDU dependencies.

Our framework specifically targets three pivotal challenges:

**Hierarchical Integration of Linguistic Features:** Effective dialogue parsing necessitates the concurrent modeling of various linguistic dimensions, including character-level, word-level, and utterance-level representations. This is especially crucial in languages like Chinese, characterized by intricate morphological structures and word agglutination. Our model strategically fuses these multifaceted features to optimize the representation of both syntactic and discursive elements.

**Speaker-Aware Interaction Modeling:** In multi-party dialogues, comprehending the roles and interactions between participants is essential. Our method incorporates explicit modeling of speaker roles to capture dependencies that are unique to different interlocutors, such as those between a customer and a service agent. This aspect is often underrepresented in conventional dependency parsing methodologies.

**Enhancing Low-Resource Robustness:** The scarcity of annotated dialogue data presents a significant challenge in training reliable models, particularly in low-resource settings. Our model addresses this issue by leveraging syntactic priors from existing treebanks, while meticulously preventing overfitting to sparse discourse patterns.

Our contributions include the following key innovations:

(1) Dynamic Subword Weighting: Our model incorporates a trainable attention mechanism that adaptively aggregates subword embeddings to construct word-level representations. This approach surpasses traditional static averaging, effectively capturing nuanced semantic variations.

(2) Gated Multi-Level Fusion: We employ a hierarchical encoding structure that seamlessly integrates character, word, and speaker features through sigmoid-gated interactions. This mechanism enhances the model's contextual awareness across various linguistic granularities.

(3) Unified Biaffine Decoding: Our model employs dual biaffine attention mechanisms to concurrently capture syntactic and discourse dependencies. This design enables the model to effectively specialize in both local syntactic and global discourse dependency patterns.

(4) Curriculum Joint Training: We implement a phased optimization strategy that progressively shifts the training focus from syntax, utilizing treebanks, to discourse dependencies, leveraging dialogue data. This approach ensures stable knowledge transfer and enhances the model's generalization capabilities.

Our model, evaluated on the CDDT benchmark, achieves state-of-the-art performance in Chinese dialogue-level dependency parsing. It effectively captures both syntactic and discourse dependencies, surpassing existing heuristic-based and multi-step pipeline methods.

## 2 Method

To address the speed and performance inefficiencies of traditional sentence parsing models, especially when dealing with an increasing number of words and dependency parsing tags, we employ a hierarchical decoding strategy for inner-EDU and inter-EDU dependencies. Additionally, we utilize the Chinese-electra-180g-base-discriminator for pre-training our large model and incorporate a state-of-the-art biaffine parser (Dozat and Manning, 2017) to enhance parsing efficiency. Our modelling framework is shown in Figure 2

First, we are provided with an input dialogue text, represented as a sequence of $n$ words $\mathbf{x} = [w_1, w_2, \ldots, w_n]$, and its corresponding EDU-level sequence $\mathbf{E} = [E_1, E_2, \ldots, E_m]$, where $m$ denotes the number of EDUs. Each EDU $E_k$ ($k \in [1, m]$)

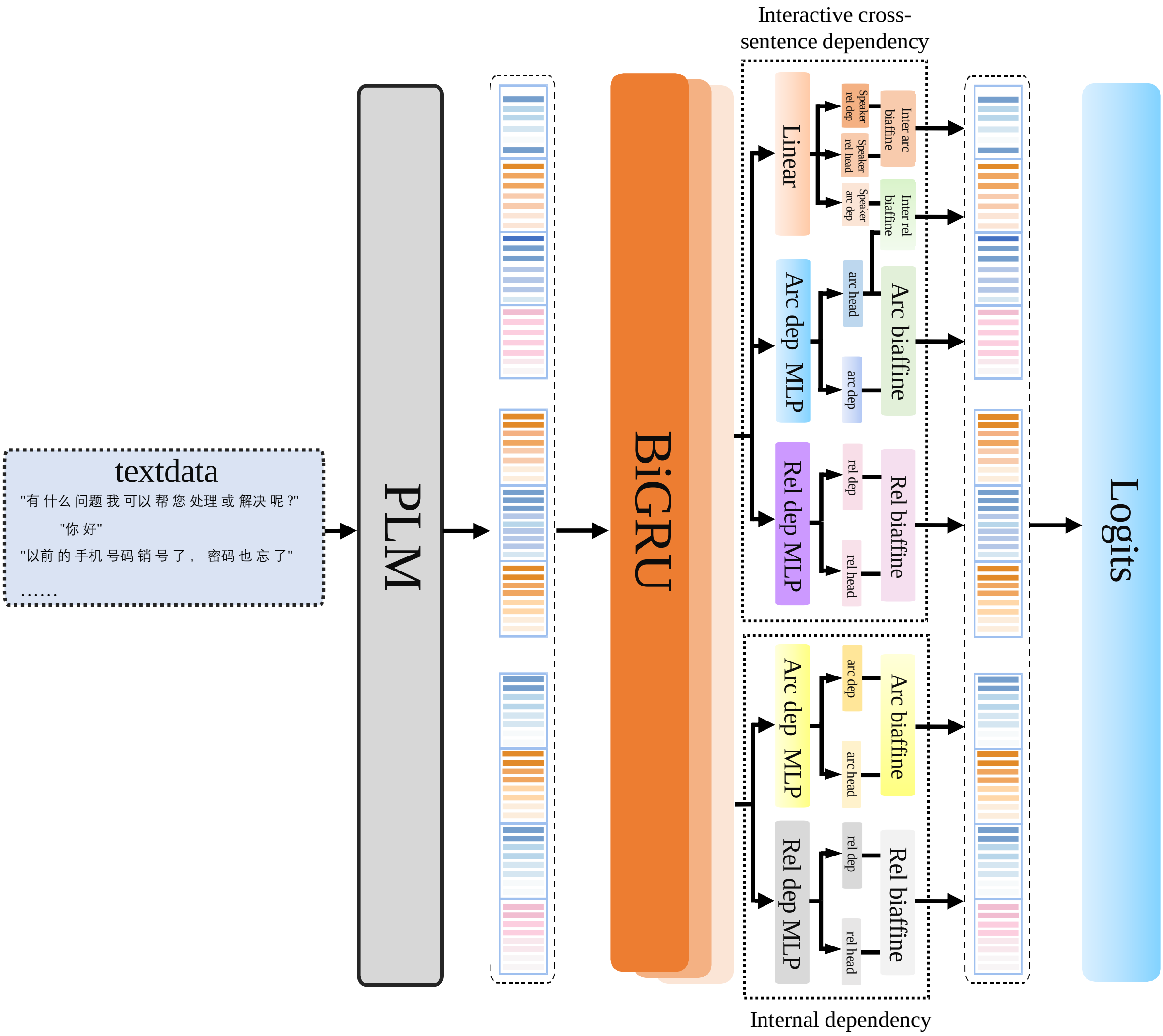


Figure 2: Our model diagram, arc denotes an arc of dependence, rel denotes the type of relationship, inter arc/rel biaffine: handling interaction dependencies across sentences, arc/rel biaffine :dealing with internal dependencies.

encompasses a subsequence of words, represented as $[w_{k,1}, \ldots, w_{k,s_k}]$, where $s_k$ is the number of words within the $k$-th EDU. We proceed to illustrate the baseline parser using an encoding-decoding framework.

### 2.1 Hierarchical Encoding

Our encoding pipeline consists of three sequential transformations to derive parsing-oriented representations:

(1) **Contextual Embedding**: The input sequence $\mathbf{x} = [w_1, w_2, \ldots, w_n]$ is processed by a pre-trained language model (e.g., BERT or ELECTRA) to obtain contextualized token embeddings:

$$\begin{aligned} e_{1:n} &\rightarrow e_1, e_2, \ldots, e_n \\ &= \mathrm{PLM}\left(w_1, w_2, \ldots, w_n\right) \end{aligned} \tag{1}$$

(2) **Sequential Abstraction**: A $L$-layer bidirectional GRU is employed to capture position-aware linguistic patterns from the contextualized embeddings:

$$\begin{aligned} h_{1:n} &\rightarrow h_1, h_2, \ldots, h_n \\ &= \mathrm{BiGRU}^{\times L}\left(e_1, e_2, \ldots, e_n\right) \end{aligned} \tag{2}$$

(3) **Dependency-Specific Projection**: Parallel $K$-layer MLPs are used to transform the abstracted features into dependency-parsing oriented representations. These MLPs provide

distinct perspectives for dependency analysis:

$$\begin{cases} \boldsymbol{z}_{1:n}^{d} \rightarrow \boldsymbol{z}_{1}^{d}, \boldsymbol{z}_{2}^{d}, \ldots, \boldsymbol{z}_{n}^{d} \\ \qquad = \mathrm{MLP}^{\times K}\left(\boldsymbol{h}_{1}, \boldsymbol{h}_{2}, \ldots, \boldsymbol{h}_{n}\right) \\ \boldsymbol{z}_{1:n}^{h} \rightarrow \boldsymbol{z}_{1}^{h}, \boldsymbol{z}_{2}^{h}, \ldots, \boldsymbol{z}_{n}^{h} \\ \qquad = \mathrm{MLP}^{\times K}\left(\boldsymbol{h}_{1}, \boldsymbol{h}_{2}, \ldots, \boldsymbol{h}_{n}\right) \end{cases} \tag{3}$$

Here, $\boldsymbol{z}^{d} = [\boldsymbol{z}_{1}^{d}, \ldots, \boldsymbol{z}_{n}^{d}]$ represents the dependency-centric feature matrix, providing insights into the relationships between words as dependents. Conversely, $\boldsymbol{z}^{h} = [\boldsymbol{z}_{1}^{h}, \ldots, \boldsymbol{z}_{n}^{h}]$ denotes the head-centric feature matrix, focusing on the roles of words as heads in dependency structures. These matrices offer orthogonal perspectives for subsequent dependency analysis, enhancing the model's parsing capabilities.

### 2.2 Decoding

The decoding of the dialogue-level dependency tree is executed in two phases. Initially, we conduct inner-EDU dependency parsing. For each $\mathrm{E}_k = [w_{k,1}, \ldots, w_{k,s_k}]$, we derive their corresponding dependency-aware and head-aware representations $z_{k,1:sk}^{d} = [z_{k,1}^{d}, \ldots, z_{k,s_k}^{d}]$ and $z_{k,1:s_k}^{h} = [z_{k,1}^{h}, \ldots, z_{k,s_k}^{h}]$ through direct indexing. Subsequently, we compute the candidate head scores for each word $w_{k,j}$ using the biaffine operation:

$$\boldsymbol{o}_{k,j}^{\mathrm{IN}} = \boldsymbol{z}_{k,1:k,s_k}^{h} \boldsymbol{U}^{\mathrm{IN}} \boldsymbol{z}_{k,j}^{d} + \boldsymbol{z}_{k,1:k,s_k}^{h} \boldsymbol{u}^{\mathrm{IN}} \tag{4}$$

$$\boldsymbol{o}_{k,j}^{\mathrm{IN,ARC}} = \sum_{l} \boldsymbol{o}_{k,j}^{\mathrm{IN}}[\cdot][l] \tag{5}$$

In these equations, $\boldsymbol{U}^{\mathrm{IN}}$ and $\boldsymbol{u}^{\mathrm{IN}}$ are trainable parameters. The candidate heads for each word $w_{k,j}$ are confined to within its EDU, and only syntactic relation labels are considered at this stage. The tensor $\boldsymbol{o}_{k,j}^{\mathrm{IN}}$ encompasses scores for both head selection and label classification: its slice $\boldsymbol{o}_{k,j}^{\mathrm{IN}}[i]$ represents a vector of relation scores for head candidate $i$. During inference, we first apply the minimum spanning tree algorithm to the arc scores $\boldsymbol{o}_{k,j}^{\mathrm{IN,ARC}}$ to retrieve a well-formed dependency tree, and then assign each predicted arc the relation label with the highest score. For cross-utterance relation modeling, we augment the biaffine mechanism with discourse-specific adaptations. Two essential feature sequences are extracted from EDU root nodes: $z_{r_1:r_m}^{d} = z_{1,r_1}^{d}, \ldots, z_{m,r_m}^{d}$ and $z_{r_1:r_m}^{h} = z_{1,r_1}^{h}, \ldots, z_{m,r_m}^{h}$, where $r_*$ denotes the root word index of each EDU. The discourse dependency scores are calculated as follows:

$$\boldsymbol{o}_{k}^{\mathrm{IT}} = \boldsymbol{z}_{r_1:r_m}^{h} \boldsymbol{U}^{\mathrm{IT}} \boldsymbol{z}_{r_k}^{d} + \boldsymbol{z}_{r_1:r_m}^{h} \boldsymbol{u}^{\mathrm{IT}} \tag{6}$$

$$o_{k}^{\mathrm{IT,ARC}} = \sum_{l} o_{k}^{\mathrm{IT}}[\cdot][l] \tag{7}$$

Here, $\boldsymbol{U}^{\mathrm{IT}}$ and $\boldsymbol{u}^{\mathrm{IT}}$ are learnable parameters. The tensor $\boldsymbol{o}_{k}^{\mathrm{IT}}$ forms a 2D structure capturing head candidates and relation labels, while $\boldsymbol{o}_{k}^{\mathrm{IT,ARC}}$ determines the dependency tree structure.

### 2.3 Curriculum Optimization

The joint loss function is modified to incorporate phased weighting, as expressed below:

$$\mathcal{L} = \alpha_t \mathcal{L}syn + (1 - \alpha_t)\mathcal{L}disc \tag{8}$$

In this equation, $\alpha_t = \max(0.5, 1 - \frac{t}{T})$ serves as a dynamic weight that evolves over $T$ epochs, implementing a curriculum learning strategy. This approach ensures a balanced focus on syntactic and discourse-level losses throughout the training process.

### 2.4 Training

We optimize a standard cross-entropy objective, which consists of separate terms for dependency arc prediction and relation classification. Let $\boldsymbol{o}_{*}^{*}$ represent either the inner-EDU scores $\boldsymbol{o}_{k,j}^{\mathrm{IN}}$ or the inter-EDU scores $\boldsymbol{o}_{k}^{\mathrm{IT}}$. We apply softmax over the arc logits $\boldsymbol{o}_{*}^{*,\mathrm{ARC}}$ and over the label logits $\boldsymbol{o}_{*}^{*}[\hat{\boldsymbol{y}}_h]$ with the $\hat{\boldsymbol{y}}_h$ representing the ground-truth head assignments to obtain probability distributions over all candidate heads and syntactic/discourse labels, respectively. The overall loss is the sum of the negative log-likelihoods of the correct heads and labels. This training procedure follows the biaffine parser framework of Dozat and Manning (2017).

We train our baseline parser by dividing its supervision into two complementary subtasks: inner-EDU (syntax) parsing and inter-EDU (dialogue) parsing.

- **Inner-EDU** parsing is fully supervised. We utilize a large-scale syntactic treebank in conjunction with the 50 gold-standard dialogue instances provided by Jiang et al. (2023). This combination offers dense, in-domain dependency annotations, ensuring the reliable convergence of this component.

| Statistic | Train | Test |
|---|---|---|
| # dialogue | 50 | 800 |
| avg.# turns | 23 | 25 |
| avg.# words | 194 | 212 |
| # inner | 9129 | 159803 |
| # inter | 1671 | 29200 |

Table 1: The Statistics of CDDT. "#" and "avg.#" Indicate "Count" and "Average Count"

- **Inter-EDU** parsing faces the challenge of annotation sparsity. To address this issue, we adopt the approach of Jiang et al. (2023), leveraging their rule-based silver dialogue corpus in addition to the same 50 gold-standard instances. This strategy merges pseudo-labeled and gold supervision, facilitating the training of the cross-utterance dependency component.

## 3 Experiment

### 3.1 Dataset

We employ the publicly accessible Chinese Dialogue-Level Dependency Treebank (CDDT), introduced by Jiang et al. (2023). This dataset serves as the sole benchmark for Chinese dialogue-level dependency parsing. The statistics of this dataset are detailed in Table 1.

### 3.2 Settings

**Evaluation Methodology.** We assess model performance using the conventional dependency parsing metrics, Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS), with punctuation tokens explicitly excluded from the calculations. To facilitate detailed diagnostic analysis, we separate the evaluation into two distinct components: intra-EDU dependencies (relationships within Elementary Discourse Units) and inter-EDU dependencies (syntactic links across units). Notably, the inter-EDU evaluation focuses on the lexical dependency level, rather than abstract EDU representations. This necessitates the accurate identification of EDU head tokens as a fundamental step for valid cross-unit dependency assessment.

In scenarios where resources are limited and development sets are not accessible, we use the checkpoint from the final training iteration for model validation. To ensure reproducibility, all implementations were carried out on a consistent computational platform equipped with an NVIDIA RTX3090 GPU, which has 24GB of VRAM.

**Hyper-parameters.** Our PLM is a Chinese variant of ELECTRA, as implemented by Cui et al. (2020). We utilize the base scale discriminator[1] for fine-tuning purposes. The hidden size of both our Parser and MLM components is set to 200, with a dropout rate of 0.1. For model training, we employ the AdamW optimizer, initializing the learning rate of the PLM at 2e-5 and that of the subsequent modules at 1e-4. A linear warmup is applied for the first 10% of the training steps. The weight decay is set to 0.1, and to prevent gradient explosion, we implement gradient clipping with a maximum value of 2.0. The training batch size is configured to 64, and the total number of epochs is 25.

### 3.3 Results

| Training Data | Few-shot | | | |
|---|---|---|---|---|
| | Inner-EDU | | Inter-EDU | |
| | UAS | LAS | UAS | LAS |
| Jiang et al. (2023) | 91.74 | 88.2 | 71.09 | 55.73 |
| baseline(Zhang et al., 2024) | 91.66 | 89.12 | 71.59 | 56.32 |
| GPT-3.5-Turbo-0613 | | | | |
| +wrd | 92.37 | 90.01 | 73.06 | 58.50 |
| +syn | 92.13 | 89.94 | 73.22 | 59.33 |
| +dis | 92.35 | 90.11 | 73.57 | 59.68 |
| +wrd & syn | 92.38 | 90.16 | 73.52 | 59.47 |
| +wrd & dis | 92.19 | 90.04 | 73.84 | 59.81 |
| +syn & dis | 92.23 | 90.18 | **73.88** | 59.94 |
| +wrd & syn & dis | 92.46 | 90.35 | 73.81 | **60.17** |
| Llama2-7B | | | | |
| +wrd | 91.91 | 89.73 | 72.33 | 57.63 |
| +syn | 91.65 | 89.51 | 72.31 | 58.28 |
| +dis | 91.90 | 89.85 | 72.76 | 58.45 |
| +wrd & syn | 91.87 | 89.81 | 72.56 | 58.38 |
| +wrd & dis | 91.82 | 89.63 | 73.13 | 58.75 |
| +syn & dis | 91.76 | 89.91 | 72.92 | 58.79 |
| +wrd & syn & dis | 91.97 | 89.89 | 72.95 | 59.01 |
| Qwen-7B | | | | |
| +wrd | 92.03 | 89.88 | 72.68 | 57.94 |
| +syn | 91.94 | 89.69 | 72.80 | 58.46 |
| +dis | 92.01 | 89.97 | 73.19 | 58.85 |
| +wrd & syn | 91.84 | 89.97 | 73.05 | 58.74 |
| +wrd & dis | 91.87 | 89.76 | 73.47 | 59.05 |
| +syn & dis | 92.07 | 89.99 | 73.42 | 59.14 |
| +wrd & syn & dis | 91.96 | 89.85 | 73.52 | 59.31 |
| Ours | **92.56** | **90.66** | 72.81 | 59.92 |

Table 2: The test results under the few-shot settings. "wrd", "syn", "dis" denote the "word-level", "syntax-level", and "discourse-level", respectively.

In the few-shot learning scenario, our model is trained on a dataset comprising 50 human-annotated instances supplemented with silver-standard corpus data. We conduct a systematic evaluation that compares four configurations: (1) base-

[1] huggingface.co/hfl/chinese-electra-180g-base-discriminator

line methods, (2) individual augmentation strategies, (3) pairwise combinations, and (4) the full integration of all three data augmentation techniques alongside our proposed model. This comprehensive evaluation framework allows for a rigorous assessment of the potential capabilities of our model. As depicted in Table 2, our approach achieves statistically significant improvements over the baseline methods across all evaluation metrics. The experimental results yield two key insights:

- For *Inner-EDU* evaluation, our model surpasses all baseline approaches and exhibits superior performance compared to three large-scale reference models.
- In the *Inter-EDU* assessment, the proposed method remains competitive with the current state-of-the-art large models.

Specifically concerning attachment scores, our model achieves the following improvements:

- **UAS Improvement**: An increase of 0.9% for Inner-EDU and 1.22% for Inter-EDU compared to the baselines.
- **LAS Enhancement**: Absolute gains of 1.54% for Inner-EDU and 3.60% for Inter-EDU.

### 3.4 Ablation Study

Our ablation study systematically investigates the consequences of removing the meticulously optimized Bidirectional Gated Recurrent Unit (BiGRU) from our model architecture. As illustrated in Table 3, the removal of this architectural component led to a notable decline in performance across all evaluation metrics. This empirical evidence underscores the critical role of our carefully designed BiGRU layer in the model's operation, especially in terms of capturing sequential dependencies and contextual patterns.

| Model | Few-shot | | | |
|---|---|---|---|---|
| | Inner-EDU | | Inter-EDU | |
| | UAS | LAS | UAS | LAS |
| Ours | 92.56 | 90.66 | 72.81 | 59.92 |
| w/o BiGRU | 90.64 | 88.03 | 69.13 | 53.19 |

Table 3: The results of the ablation experiments.

## 4 Related Work

**Dependency Parsing.** Several Chinese dependency parsing paradigms and corresponding treebanks have been developed (Xue et al., 2005; Che et al., 2012; McDonald et al., 2013; Qiu et al., 2014). These efforts primarily concentrate on sentence-level dependency parsing, with document-level parsing being significantly less explored. Li et al. (2014) applied a dependency parsing approach to discourse parsing, although their EDU-wise method overlooks the parsing within EDUs. Recent advancements by Jiang et al. (2023) have initiated Chinese dialogue-level dependency parsing, establishing a unified schema that encompasses both inner-EDU syntactic dependencies and inter-EDU discourse dependencies. Building upon this, Zhang et al. (2024) have further refined the framework by incorporating LLM-assisted data augmentation, tackling the challenges of low-resource settings through hierarchical transformations at the word, syntax, and discourse levels.

Meanwhile, cross-lingual transfer methods have emerged as complementary approaches. Guo et al. (2022) proposed a curriculum-style fine-grained adaptation technique for unsupervised cross-lingual dependency transfer, demonstrating that syntactic knowledge can be effectively transferred across languages through progressive difficulty scheduling and parameter generation networks. This approach achieves state-of-the-art performance on Universal Dependencies treebanks by combining curriculum learning with self-training strategies.

**Dialogue Parsing.** Discourse structures in dialogue can be represented by various theories, including RST (Mann and Thompson, 1987, 1988), SDRT (Asher and Lascarides, 2003), and PTDB (Prasad et al., 2008). While datasets such as STAC (Afantenos et al., 2015) and Molweni (Li et al., 2020) concentrate on English multi-party dialogues, Jiang et al. (2023) introduced the first Chinese dialogue-level dependency treebank (CDDT), which merges RST-inspired discourse relations with syntactic dependencies. This work bridges the gap between EDU-based discourse parsing and word-wise dependency structures. Expanding on this, Zhang et al. (2024) developed a three-level augmentation strategy using Large Language Models (LLMs) to create varied pseudo-instances while maintaining discourse hierarchies, leading to significant improvements in handling inter-EDU dependencies.

**Weakly Supervised Learning.** Predicting unseen dependency labels in low-resource settings presents significant challenges (Norouzi et al., 2013). Jiang et al. (2023) tackle this issue by employing signal-based dependency transformation and pseudo-labeled data filtering, utilizing syntactic treebanks and masked language modeling to infer inter-EDU relations. Zhang et al. (2024) build upon this approach by leveraging LLMs' generative abilities for extensive data augmentation, developing prompt-based mechanisms to maintain structural consistency throughout transformations. Their method integrates characterization, chain-of-thought prompting, and constrained generation, illustrating that LLMs can effectively distill syntactic and discourse knowledge without direct supervision. This contrasts with conventional self-training (Scudder, 1965) and co-training (Blum and Mitchell, 1998) methodologies, providing a model-centric solution for low-resource dependency parsing.

**Universal Structured NLP and Demonstration Systems.** Recent efforts have been made to unify structured NLP (XNLP) tasks under a general framework. Fei et al. (2023) proposed XNLP, an interactive demonstration system built upon large language models (LLMs), aiming to model a wide variety of XNLP tasks, such as syntactic parsing, information extraction, semantic role labeling, and sentiment analysis, in a unified manner. By reducing task outputs to span extraction and relation prediction, the system achieves high generalizability and supports zero-shot and weakly supervised learning without task-specific fine-tuning. Furthermore, it offers multi-turn user interaction, structured visualization via brat, and interpretable prediction rationales. These features highlight the potential of LLM-based architectures in managing structurally diverse tasks with minimal supervision, aligning well with the goals of low-resource dependency parsing and dialogue-level analysis. XNLP thus provides both a practical tool and a methodological reference for universal structured prediction under weak supervision.

## 5 Conclusion

In this study, we introduce InterParser, an innovative end-to-end framework designed for Chinese dialogue-level dependency parsing. Our model integrates a pretrained language model, hierarchical BiGRU encoding, and speaker-aware biaffine scoring mechanisms, effectively merging intra-EDU syntactic dependencies with inter-EDU discourse relations. Experimental outcomes showcase notable enhancements over existing techniques, reaching state-of-the-art performance with 92.56% UAS for inner-EDU parsing and 90.66% LAS for inner-EDU parsing in few-shot scenarios. The ablation study further confirms the essential contribution of the BiGRU layer in capturing sequential linguistic structures.

## Limitations

Despite our advancements in low-resource dialogue parsing, certain limitations persist. The dependence on pseudo-labeled data for inter-EDU dependencies may introduce annotation noise. Additionally, our current speaker modeling, which focuses on role disparities, overlooks dynamic interaction nuances. Future research directions include extending cross-lingual adaptation to other resource-scarce languages, incorporating pragmatic elements for comprehensive dialogue comprehension, and developing unified parsing-generation frameworks to more effectively bridge the gap between syntactic and discourse hierarchies. These developments will be instrumental in constructing more robust and interpretable dialogue systems.

## Acknowledgement

This work is supported by the National Key R&D Program of China (Grant nos. 2023YFB3308601, 2022YFB3104700), the National Natural Science Foundation (Grant nos. 62402395,62376198), and the Science and Technology Program of Sichuan Province (Grant no. 2023YFS0424).

## References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Conference on Empirical Methods on Natural Language Processing (EMNLP 2015)*, pages pp–928.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2012. Chinese dependency treebank 1.0 ldc2012t05. *Philadelphia: Linguistic Data Consortium*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pretrained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing.

Hao Fei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Xnlp: an interactive demonstration system for universal structured nlp. *arXiv preprint arXiv:2308.01846*.

Peiming Guo, Shen Huang, Peijie Jiang, Yueheng Sun, Meishan Zhang, and Min Zhang. 2022. Curriculum-style fine-grained adaption for unsupervised cross-lingual dependency transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:322–332.

Gongyao Jiang, Shuang Liu, Meishan Zhang, and Min Zhang. 2023. A pilot study on dialogue-level dependency parsing for chinese. *arXiv preprint arXiv:2305.12441*.

Xinzhou Jiang, Zhenghua Li, Bo Zhang, Min Zhang, Sheng Li, and Luo Si. 2018. Supervised treebank conversion: Data and approaches. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2706–2716.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, Bonnie L Webber, et al. 2008. The penn discourse treebank 2.0. In *LREC*.

Likun Qiu, Yue Zhang, Peng Jin, and Houfeng Wang. 2014. Multi-view chinese treebanking. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 257–268.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. Llm–assisted data augmentation for chinese dialogue–level dependency parsing. *Computational Linguistics*, pages 1–24.