

System Report for CCL25-Eval Task 2: Enhanced Chinese Frame Semantic Parsing with Pre-trained Model and Linguistic Features

Yahui Liu, Ziheng Qiao, Chen Gong, Min Zhang

School of Computer Science and Technology, Soochow University, Suzhou, China

{yhliunlp, zhqiao}@stu.suda.edu.cn

{gongchen18, minzhang}@suda.edu.cn

Abstract

This paper presents our system submitted to the Chinese Frame Semantic Parsing evaluation task at the 24th China National Conference on Computational Linguistics (CCL2025). For the three subtasks of Frame Identification (FI), Argument Identification (AI), and Role Identification (RI), we utilized a larger Chinese pre-trained model, as the foundation and adopted specific optimization strategies for FI and RI subtasks. Specifically, we incorporated word segmentation structure information and updatable pre-trained target word embeddings in the FI subtask, and explored the use of Focal Loss combined with target word embeddings and word segmentation structure information in the RI subtask. Furthermore, a voting mechanism was employed in both the FI and RI subtasks to enhance performance. Our system ultimately achieved first place on the TestA and second place on the TestB.

Keywords: Chinese Frame Semantic Parsing, Linguistic Features, Pre-trained Model

1 Introduction

As a critical task in fine-grained semantic understanding, Chinese Frame Semantic Parsing (CFSP) leverages the rich lexical-semantic knowledge encoded in Chinese FrameNet (CFN) (You and Liu, 2005; Li et al., 2023a). This task is first introduced as a evaluation task during the CCL2023-Eval (Li et al., 2023a). CFSP aims to extract frame-semantic information from a sentence (Wang et al., 2020). Its effectiveness in enhancing a range of downstream applications, including text summarization (Guan et al., 2021a; Guan et al., 2021b), reading comprehension (Guo et al., 2020a; Guo et al., 2020b), relation extraction (Zhao et al., 2020), etc.

Chinese FrameNet (CFN), a structured knowledge representation developed by Shanxi University, grounds itself in the principles of Frame Semantics, drawing inspiration from English FrameNet (Fillmore et al., 2003) and utilizing authentic Chinese corpora as its foundational data source (Li et al., 2024). By establishing explicit connections between vocabulary and underlying concepts through the notion of “frames”, CFN provides a rich semantic resource. Notably, a portion of the CFN dataset was made publicly available for the first time during the CCL2023-Eval Chinese Frame Semantic Parsing (CFSP) task. Building on this, the subsequent CCL2024-Eval CFSP introduced additional information pertaining to constructional target words (Peiyuan et al., 2024). This evaluation focuses towards the models’ capacity to analyze sentences exhibiting semantic nesting phenomena.

The pervasive phenomenon of semantic role nesting in sentences, such as in the sentence “我的眼睛什么也看不见了。” (My eyes can’t see anything anymore.), where “我的眼睛” (my eyes) functions as the “Body_part” role within the “自由感知(Perception_active)” frame, while simultaneously “我” (I) acts as the “自主感知者” (Perceiver) role, has led to a prioritization of annotating semantic roles with broader argument spans in previous evaluations, neglecting finer-grained roles and consequently resulting in the omission of some semantic information. This evaluation places a greater emphasis on the model’s ability to analyze sentences exhibiting semantic nesting phenomena and aims to improve existing analysis tools

Frame Name	肢体移动(Body movement)	
Frame Definition	施动者使用其身体部位进行移动或行动的词汇。 (An agent performs movement or action using parts of their body.)	
Frame Elements	fe_name	fe_def
	施动者(Agent)	施动者使用其具体身体部位来执行某行为。(An agent uses specific body parts to carry out an action.)
	身体部位(Body part)	将行为中的身体部位由动词的意义表达出来, 并不能单独进行表达。 (The body part involved in the action is expressed through the meaning of the verb and cannot be expressed independently.)
	程度(Degree)	表示事件发生的程度。(Denotes the extent of the event.)
	方向(Direction)	肢体移动时的方向。(The direction of the body movement.)
	时间(Time)	表示身体运动发生的时间。(The time when the body movement takes place.)
	起点(Source)	身体部位开始移动的地方。(Indicates the source location where the movement of the body part is initiated.)

Table 1: An example about “肢体移动” (Body movement) frame in Chinese FrameNet.

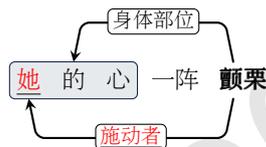


Figure 1: An example of CFSP.

that are prone to overlooking partial role information when faced with linguistic phenomena such as the nesting and fusion of frame elements.

As before, this evaluation divides CFSP task into three subtasks: Frame Identification, Argument Identification and Role Identification. **Frame Identification (FI)** aims to find the corresponding frame for the target word in all frames according to its meaning in the sentence. Frame elements, also referred to as arguments, can be a single word or a span in the sentence. The goal of **Argument Identification (AI)** is to locate all arguments within the sentence and determine their boundaries. The purpose of **Role Identification (RI)** is to assign appropriate semantic role labels to the identified arguments. Taking Figure 1 as an example, the target word in the sentence “她的心一阵颤栗” (Her heart trembled for a moment) is “颤栗” (trembled), which corresponds to the frame “肢体移动” (Body movement) in Table 1. In this sentence, there are multiple arguments related to this frame, i.e., “她” (her) and “她的心” (her heart). “施动者” (agent) and “身体部位” (Body part) are their corresponding semantic role labels, respectively. It is noted that the two arguments, “她” (her) and “她的心” (her heart), exhibit a nested structure.

To enhance performance across the various subtasks of this evaluation, we implemented distinct strategies, with a particular emphasis on improving the outcomes for FI and RI subtasks. Across all three tasks, we upgraded the base architecture to a larger Chinese BERT model. Specifically, for FI subtask, we incorporated word segmentation structure information and pre-trained embeddings for target words (fasttext), allowing the target word embeddings to be updated during training. For RI subtask, we replaced the standard Entropy Loss with Focal Loss and also integrated pre-trained target word embeddings and word segmentation structure information. Finally, we employed a voting technique to refine the predictions for both FI and RI. All of our datasets and codes are available at <https://github.com/yahui19960717/CCL-2025-CFN.git>.

2 Related Work

The Chinese Frame Semantic Parsing (CFSP) task, a novel challenge based on CFN, was first presented at CCL2023-Eval (Li et al., 2023a). Current CFSP research largely employs neural network methods (Li et al., 2023b; Huang et al., 2023; Liu et al., 2023; Guan et al., 2023). Li et al. (2023b) achieved significant gains in Frame Identification (FI) by using rotational positional encoding (Su et al., 2021) to improve entity attention. Huang et al. (2023) enhanced model robustness through training strategies like Exponential Moving Average (EMA) and the Fast Gradient Method (Miyato et al., 2017). For FI, Liu (2023) framed CFSP as a word-based graph parsing task for end-to-end frame and argument identification. For Argument Identification (AI) and Role Identification (RI), they utilized tree parsing to better model argument structure for boundary and role recognition. Guan (2023) explored various pre-trained models and a range of techniques in feature engineering, model design, and other optimizations for each subtask.

In the era of LLM, the performances of LLM on most NLP tasks are still well below the supervised baselines (Sun et al., 2023; Yan et al., 2024). Li et al. (2023a) test the ability of the LLM on different subtasks of CFSP using ChatGPT (gpt-3.5-turbo-16k) (Brown et al., 2020). They construct different prompts for the three subtasks on part of the test set and guide the model to generate more reliable results by designing the chain-of-thought. Finally, they find that the performance of ChatGPT on the three sub-tasks of CFSP is not ideal. In CCL2024-Eval, Yang et al. (2024) introduce construction grammar to expand the targets, as basic units activating frames in texts, from word-style to construction-style. This evaluation includes two tracks: close track and open track. Open track allows for inference using large language models such as ChatGPT. Liu et al. (2024) won first place by employing a method that combines small and large models, but the performance of LLM does not stand out in the CFSP task.

In this evaluation, considering the poor performance of large models on this task, we opted to leverage linguistic features and pre-trained models with smaller models to enhance the performance of each subtask.

3 Methods

Our work builds upon the baseline model proposed by Li et al. (2023b). Their baseline utilizes an entity classification method based on Rotational Positional Encoding (RoPE) (Su et al., 2022) to calculate attention scores between entities for classification. Next, we will introduce the techniques and strategies used during the competition.

3.1 Pre-trained Model

Initially, for all three subtasks, we replaced the original chinese-roberta-wwm-ext model¹ with the larger pre-trained model.² Our primary rationale for this substitution was that larger pre-trained models typically possess greater model capacity and learning ability, enabling them to capture more intricate language patterns and knowledge. Theoretically, this should contribute to better generalization performance on downstream tasks. Our experimental results on TestA validated this hypothesis, demonstrating performance improvements across all three tasks with this substitution.

3.2 Frame Identification

The FI subtask is to identify the frames triggered by the given target word in the sentence from a total of 713 provided frames. For FI subtask, the task data already provides explicit target word information, including the start and end boundary indices of the target word. We adopted the RoPE-based entity classification method to calculate the scores for the corresponding frames of each target word. To more fully leverage the target word information, we obtained the embedding representation of the target word text using a pre-trained word embedding³, and then concatenated this embedding with the output of the BERT model. This combined representation was learned and adjusted during the model training process.

¹<https://huggingface.co/hfl/chinese-bert-wwm-ext/tree/main>

²<https://huggingface.co/hfl/chinese-roberta-wwm-ext-large/tree/main>

³<https://fasttext.cc/docs/en/pretrained-vectors.html>

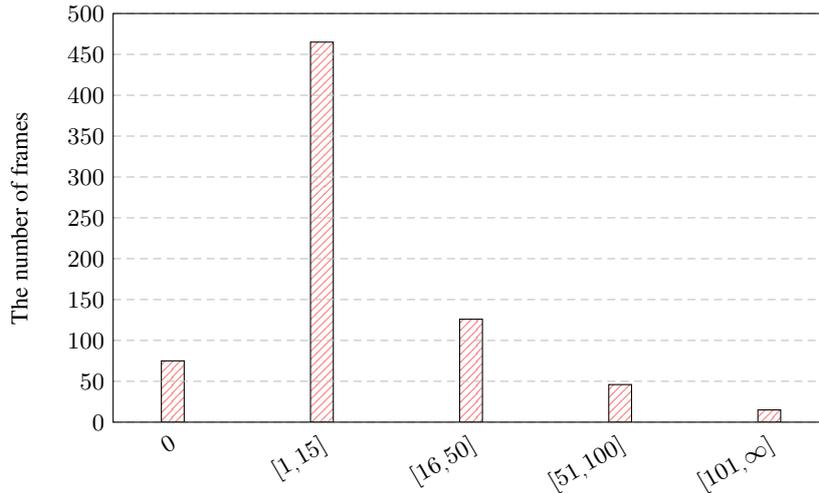


Figure 2: The distribution of the number of example sentences corresponding to each frame in Train set.

In addition, considering that the structural information from Chinese word segmentation might be beneficial for the tasks, we also incorporated the word segmentation structure information in the sentence into our model. We employed the BMES (Beginning, Middle, End, Single) tagging scheme to indicate whether each character in the sentence belongs to the beginning, inside, or outside of a word. As shown in Figure 1, its word segmentation structure information can be expressed as “BEBEBEBEBES BEBEBMEBEBEBMMESBEBES”.

Furthermore, we observed a significant class imbalance issue, exhibiting a long-tailed distribution where a few frame categories accounted for the majority of training samples, while a large number of frame categories had only a small number of instances. As shown in Figure 2, the distribution of the number of sentences corresponding to each frame in the training set is displayed. It can be observed that the largest number of frames have between 1 and 15 sentences, while frames with other sentence counts are relatively few. To address the potential bias in model training caused by this data distribution, we experimented with replacing the standard Cross-Entropy Loss with Focal Loss (Lin et al., 2017). Focal Loss aims to make the model focus more on hard-to-classify samples by dynamically adjusting the weights of samples. The formula of focal loss is shown as Equation 1. However, the experimental results on the TestA unexpectedly showed that applying Focal Loss did not lead to performance improvements for the Frame Identification task. We hypothesize that this might be because there are a large number of frame labels present in TestA that are completely absent from the training set (there are 75 frames that have no corresponding sentences in the training set). Even though Focal Loss can effectively mitigate the class imbalance problem for known categories in the training set, the model lacks the fundamental knowledge to learn relevant features for categories not seen during training, which may explain why using Focal Loss did not yield the expected benefits on TestA.

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t means the probability of the correct class predicted by the model.

Voting is a common strategy used in evaluation competitions to enhance model performance. During the evaluation of each model, we enabled Dropout and ran the model three times using different random seeds: 1, 77, and 777, to obtain three distinct sets of prediction results. For FI subtask, we directly selected the frame that the highest number of votes across these three prediction sets as the final prediction.

3.3 Argument Identification

The AI subtask aims to identify the boundaries of text segments in a sentence that function as arguments of a frame. We primarily relied on the strong contextual understanding capabilities of the pre-trained language model. Although we experimented with incorporating word segmentation structure information

data	train	dev	testA	testB	frames
CFN	10,700	2,300	4,400	4,600	713

Table 2: Data statistic of CFN.

Team	FI_acc	AI_f1	RI_f1	task_score
Team1	71.8032	85.9555	58.5900	70.7636
Team2	71.8960	85.5192	57.6181	70.2718
Team3	71.0838	85.2344	57.7761	70.0059

Table 3: Reproduction results of the top three teams.

into the argument identification model, we did not observe significant performance improvements in our experiments, and therefore this feature was not used in the final model. In addition, voting strategies also failed to improve the performance of the model.

3.4 Role Identification

The RI task aims to determine the semantic role of identified arguments within a specific frame. Similar to the FI subtask, we first utilized the boundary information of the target word and concatenated the pre-trained target word embeddings obtained via fasttext with the output of the BERT model. A key improvement was the replacement of the cross-entropy loss function used in the baseline with Focal Loss, with the goal of better handling potential class imbalance issues. We also experimented with incorporating word segmentation structure information. Finally, we applied the same voting technique used in the FI subtask to ensemble the results for the RI task. Specifically, we retained the argument spans that appeared in at least 2 out of the different model prediction results.

4 Experiments

4.1 Data and Settings

The data statistics of CFN are presented in Table 2. We set the Focal Loss parameter γ to 1 tuned on TestA with other model parameters following Li et al. (2023b).

4.2 Evaluation Metrics

We use the evaluation metrics provided by the official. The FI subtask employs accuracy as the metric, while the AI and RI subtasks adopt precision, recall, and F1 scores as measures. The formulas are as follows:

$$ACC = \frac{\text{the right number of frames}}{\text{Total number of frames}} \quad (2)$$

$$P = \frac{Count(\text{gold} \cap \text{pred})}{Count(\text{pred})} \quad (3)$$

$$R = \frac{Count(\text{gold} \cap \text{pred})}{Count(\text{gold})} \quad (4)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (5)$$

where gold and pred denote the correct and predicted results, respectively. For the AI subtask, $Count(*)$ represents the number of tokens in results, whereas it signifies the number of elements in the Argument Role Identification task.

Model	FI.acc	AI.f1	RI.f1
Full Model	71.8960	85.5192	57.6181
- w/o LM	71.5943	82.9840	56.5342
- w/o WS	71.6175	-	57.1250
- w/o TWE	71.7336	-	57.0354
- w/o Voting	72.1977	-	57.4830
- w/o FL	-	-	57.9126

Table 4: Ablation study on three subtasks.

4.3 Results and Analysis.

As shown in Table 3, we compared the performance of our submitted system with the top three teams on the leaderboard. Our system (Team2) achieved an accuracy of 71.8960 on FI subtask, an F1 score of 85.5192 on AI, and an F1 score of 57.6181 on RI, resulting in an overall score of 70.2718, ranking second.

Compared to Team1, our system exhibited performance gaps in both the AI and RI subtasks, which directly contributed to our lower overall score. The higher F1 score achieved by Team1 in AI subtask likely stems from their more effective techniques for identifying argument boundaries. Notably, our system demonstrated a slightly superior performance in the FI subtask, potentially validating the effectiveness of our combined strategies, including the larger pre-trained model, word segmentation structure information, target word embeddings, and the voting mechanism, for frame recognition. While a performance gap exists in AI subtask compared to the top team, the difference is relatively small. Considering that we primarily relied on the representational power of a larger pre-trained model for this subtask, this tentatively suggests the potential of larger models in capturing argument boundary information. However, we observed a more significant performance decline in the RI subtask, which is likely attributed to the performance limitations in our AI subtask as argument role labeling typically relies on accurate argument boundary identification results. Therefore, improving the performance of AI subtask will be crucial for future enhancements in RI subtask.

Ablation Study After the competition, we conducted ablation studies on all three tasks by systematically removing individual technical components. As shown in Table 4, the components include LM (RoBERTa large model), WS (word segmentation structure information), TWE (target word embedding), and FL (focal loss), where “-” indicates that the full model does not incorporate the corresponding technique. As we can see, removing the large pre-trained language model leads to a notable drop in performance across all three tasks, especially in AI (-2.53 F1) and RI (-1.08 F1), confirming that deep contextual representations are critical for semantic understanding.

Removing the structure information of word segmentation causes a slight decline in FI and RI performance, suggesting that the information contributes modest semantic cues, although it is less critical than LM. Similarly, the impact of removing TWE is comparable to WS, indicating that both tasks are indeed sensitive to target word information, even when the pre-trained model is capable of capturing contextual semantics, the explicit representation of the target word can still provide additional signals.

Interestingly, removing the voting strategy leads to a small increase in FI accuracy, suggesting that the voting mechanism may be unstable. We hypothesize that this may be due to the insufficient number of prediction outputs used for aggregation.

In addition, removing focal loss slightly improves RI F1, which is contrary to its effect on TestA. This may be due to differences in the degree of label imbalance between TestA and TestB, with TestB possibly being more balanced.

5 Conclusion

This paper presented our system for the Chinese Frame Semantic Parsing task at CCL2025. We leveraged a larger pre-trained model and introduced specific optimizations for FI and RI subtasks, including linguistic features and voting. Our system achieved first place on TestA and second place on TestB. While demonstrating competitive performance, particularly achieving a slight lead in FI, our analysis revealed performance gaps in AI and RI subtasks compared to the top-ranked Team1. Future work will focus on enhancing our AI capabilities, which we believe will subsequently improve RI performance, to further advance our overall CFSP system.

Acknowledgements

We thank the reviewers for their constructive feedback which helped to improve the paper. This work was supported by National Natural Science Foundation of China (Grant No.62306202) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, pages 1877–1901.
- Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to framenet. *International journal of lexicography*, 16(3):235–250.
- Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hongye Tan. 2021a. Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4045–4052.
- Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hu Zhang. 2021b. Integrating semantic scenario and word relations for abstractive sentence summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2522–2529.
- Yingxuan Guan, Xunyuan Liu, Lu Zhang, Zexian Xie, and Binyang Li. 2023. System report for CCL23-eval task 3: UIR-ISC pre-trained language model for Chinese frame semantic parsing. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 124–138.
- Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2020a. Incorporating syntax and frame semantics in neural network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2635–2641.
- Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020b. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896.
- Shutan Huang, Qiuyan Shao, and Wei Li. 2023. System report for CCL23-eval task 3: Chinese frame semantic parsing based on multi task pipeline strategy. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 105–112.
- Juncai Li, Zhichao Yan, Xuefeng Su, Boxiang Ma, Peiyuan Yang1, and Ru Li. 2023a. Overview of CCL23-eval task 1: Chinese FrameNet semantic parsing. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 113–123.
- Zuoheng Li, Xuanzhi Guo, Dengjian Qiao, and Fan Wu. 2023b. System report for CCL23-eval task 3: Application of entity classification model based on rotary position embedding in chinese frame semantic parsing. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 94–104.
- Ru Li, Yunxiao Zhao, Zhiqiang Wang, Xuefeng Su, Shaoru Guo, Yong Guan, Xiaoqi Han, and Hongyan Zhao. 2024. A comprehensive overview of cfn from a commonsense perspective. *Machine Intelligence Research*, 21(3):239–256.

- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.
- Yahui Liu, Zhenghua Li, and Min Zhang. 2023. System report for CCL23-eval task3: SUDA CFSP system. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 84–93.
- Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of the 5th International Conference on Learning Representations*, pages 1–11.
- Yang Peiyuan, Li Juncai, Yan Zhichao, Su Xuefeng, and Ru Li. 2024. Chinese frame semantic parsing evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 32–42.
- Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021. A knowledge-guided framework for frame identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5230–5240.
- Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.
- Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, et al. 2023. Pushing the limits of ChatGPT on NLP tasks. *arXiv preprint arXiv:2306.09719*.
- Xiaohui Wang, Ru Li, Zhiqiang Wang, Qinghua Chai, and Xiaoqi Han. 2020. Syntax-aware Chinese frame semantic role labeling based on self-attention. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 616–623.
- Liu Yahui, Gong Chen, and Zhang Min. 2024. Leveraging LLMs for Chinese frame semantic parsing. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 21–31.
- Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Ru Li, and Jeff Z Pan. 2024. Atomic fact decomposition helps attributed question answering. *arXiv preprint arXiv:2410.16708*.
- Liping You and Kaiying Liu. 2005. Building Chinese Framenet database. In *Proceedings of the 2005 international conference on natural language processing and knowledge engineering*, pages 301–306.
- Hongyan Zhao, Ru Li, Xiaoli Li, and Hongye Tan. 2020. CFSRE: Context-aware based on frame-semantics for distantly supervised relation extraction. *Knowledge-Based Systems*, 210(2020):106480.