Overview of CCL25-Eval Task2: Chinese Frame Semantic Parsing Evaluation

Hao Xu[1,‡], Juncai Li[1,‡], Zhichao Yan[1,‡], Haikun Liu[1,‡]
Xuefeng Su[1,3,‡], Jiayang Zhang[1,‡], Ru Li[1,2,⋆,†]

[1]School of Computer and Information Technology, Shanxi University

[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education

[3]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology

[‡]{202322407052,202312407010,202312407023,202322407023,201912407008,202322407063}@email.sxu.edu.cn

[⋆]liru@sxu.edu.cn

Abstract

Chinese Frame Semantic Parsing (CFSP) aims to extract fine-grained frame semantic structures from text, providing rich semantic information to enhance the capabilities of natural language understanding models in semantic representation and downstream applications. Building on the CCL-2024 CFSP evaluation task and motivated by the prevalent phenomenon of semantic roles nesting in sentences, we update the nested role annotation data by simultaneously labeling all nested semantic roles. Based on this enhancement, we publish a more challenging CFSP evaluation task for CCL-2025. The evaluation dataset consists of 22,000 annotated examples involving 703 frames, including nested annotations covering 101 semantic roles. The evaluation task, divided into three subtasks: frame identification, argument identification, and role identification, has attracted wide attention from both industry and academia, with a total of 156 teams participating. As for the evaluation results, Yongqing Huang from Guangdong Province won first place with a final score of 70.76. In this paper, we report key information about the evaluation task, including key concepts, evaluation dataset, top-3 results and corresponding methods. More information about this task can be found on the website for the CCL-2025 CFSP evaluation task [1].

1 Introduction

Frame Semantic Parsing (FSP) is a fine-grained semantic analysis task based on frame semantics (Kate et al., 2005), and it aims to extract frame semantic structures from sentences, thereby achieving in-depth understanding of events or situations within the sentence. FSP plays a pivotal role in downstream tasks such as reading comprehension (Guo et al., 2020b; Guo et al., 2020a), text summarization (Guan et al., 2021a; Guan et al., 2021b), and relation extraction (Zhao et al., 2020).

Chinese FrameNet (CFN) (Li et al., 2024; You and Liu, 2005) is a semantic knowledge base for the Chinese language, constructed on the theoretical basis of Frame Semantics and developed from Chinese corpus materials, referring to the FrameNet (FN) of the University of California, Berkeley. It comprises a frame library, a sentence corpus, and a lexical unit library. Currently, it contains 1,400 frames, involves 18,896 lexical units, and over 100 thousand annotated sentences.

In the existing Chinese FrameNet dataset, semantic roles with larger argument spans are prioritized, while finer-grained roles are often overlooked, resulting in the loss of some semantic information. For instance, in the sentence “我的眼睛什么也看不见了,” the phrase “我的眼睛” serves as the Body Part role within the Perception Active frame, while the word “我”, which simultaneously functions as the Perceiver Agentive role, is often ignored.

With the rapid advancement of large language models(LLMs), considerable progress has been achieved in coarse-grained semantic analysis tasks, suggesting that models are capable of

[1]Task website https://tianchi.aliyun.com/competition/entrance/532338

[†] Corresponding Author

Proceedings of the 24th China National Conference on Computational Linguistics, pages 60-69, Jinan, China, August 11-14, 2025.

basic semantic understanding. Nevertheless, their performance in fine-grained and complex semantic scenarios remains suboptimal, indicating significant potential for further improvement. To enhance the capabilities of Chinese frame semantic parsing and facilitate a deeper understanding of language, we expand our dataset to include nested roles annotation by simultaneously labeling all nested semantic roles. Specifically, we annotate 690 sentences with nested structures, involving a total of 101 semantic roles. Consequently, we launch the third Chinese semantic frame parsing evaluation.

2 Relevant Concepts and Task Description

2.1 Relevant Concepts

Frame semantics is an important branch of cognitive linguistics, which was first proposed and advocated by Fillmore. Frame semantics introduces the cognitive structure of the concept of "frame" into semantics, providing a cognitive-level explanation for understanding word meanings, sentence meanings, and discourse meanings. It has unique advantages in implementing cognitive understanding of language in computers. The Chinese FrameNet is a Chinese frame semantic knowledge base built on the theoretical foundation of frame semantics. There are several important concepts in the Chinese FrameNet.

Frame: A frame is a schematic cognitive scene activated by words in the user's brain, which is the background and motivation for understanding and using language. Table 1 demonstrates the basic information about frame "量变". This frame represents the semantic scenario conveying the following meaning: "实体在某个维度上（即某属性）的相对位置发生变化，其属性值从初值变至终值".

Frame Element: Frame elements refer to the participants in the semantic scenario corresponding to the frame, which is also called semantic roles in frame semantic parsing task. For example, the "实体" and "属性" in the frame "量变" are two frame elements of this frame. The frame elements greatly enrich the semantic information of the frame.

Lexical Unit: The lexical unit refers to a word that can activate a certain frame in the CFN frame library. Each lexical unit can typically activate one or more frames, but in a specific sentence, each lexical unit can only belong to a specific frame. In the example shown in this paper, in addition to the construction "从 A 到 B", the "量变" frame includes lexical unit such as "增加" and "上升".

Target Word: A word or construction in the sentence to be annotated that can activate the frame, usually a lexical unit or construction from the CFN library. In the example sentence in Figure 1, "从 A 到 B" is the target word that activates the frame.

框架名称	量变	
框架定义	实体在某个维度上（即某属性）的相对位置发生变化，其属性值从初值变至终值。	
框架元素	框架元素名称	框架元素定义
	实体	在某属性上具有一定量值的事物。
	属性	实体的有数量变化的属性。
	初值	实体的属性值变化的起点。
	终值	实体最后达到的量值。
	初状态	实体经历属性值的变化之前的状态。
	终状态	实体经历属性值的变化之后所达到的状态。
	变幅	实体在某维度上变动的幅度。

Table 1: The "量变" frame and its frame element information.

2.2 Task Description

The task of CFSP is divided into three sub-tasks: Frame Identification (FI), Argument Identification (AI), and Role Identification (RI).

Proceedings of the 24th China National Conference on Computational Linguistics, pages 60-69, Jinan, China, August 11-14, 2025.

Frame Identification: Frame Identification is the task of selecting the most suitable semantic frame from multiple candidate frames for a given target word that can activate a frame, based on the context. As shown in the part of Frame Identification in the Figure 1, the target word can activate frames like "量变" and "到达". But the "量变" frame can be finally determined based on the context. The formal definition of this task is as follows: Given a sentence S that contains the target word, denoted as $S = (w_1, w_2, ..., w_n)$, w_i stands for the ith word in the sentence, where $1 \leq i \leq n$. The target word to be identified is denoted as $w^t = \{w_1^t, w_2^t, \cdots, w_m^t\}$, $w_j^t \in S, m \leq n$.The word in w^t doesn't have to be consecutive. The task is to select an appropriate frame f_t from a given frame library $F = \{f_1, f_2, ...f_n\}$ based on the semantic context, which is expressed as:

$$f_t = \underset{f_i \in F, w^t \in S}{\text{argmax}} P(f_i | S, w^t) \tag{2.1}$$

Argument Identification: Argument identification is a subtask that identifies the starting and ending positions of an argument in a sentence. That is, given a sentence and a target word, it automatically identifies the boundaries of the semantic roles governed by the target word under the condition that the target word is known. In the Figure 1, the target word "从 A 到 B" governs arguments including "新注册登记新能源汽车", "数量", "65 万辆", and "295 万辆", while "新能源汽车" is an incorrect argument. The formal definition of argument identification is as follows: for a given sentence $S = (w_1, w_2, ..., w_n)$ and its target word $w_t \in S$, the objective of this task is to find the boundary range i_τ^s and i_τ^e for an argument $a_\tau \in \{a_1, a_2, ..., a_k\}$ such that $a_\tau = w_{i_\tau^s}, ...w_{i_\tau^e}$.

Role Identification: The task of role identification is the final step in CFSP task. This task aims to determine the corresponding frame element for each argument in the sentence, that is, the semantic role of each argument within its corresponding frame. For example, in the Figure 1, the semantic role of "新注册登记新能源汽车" is "实体". The formal definition of this task is as follows: given a sentence $S = (w_1, w_2, ..., w_n)$, the target word $w_t \in S$ in the sentence, and the frame f activated by the target word, for the argument $a_\tau = w_{i_\tau^s}, ...w_{i_\tau^e}$ with known boundary range, the aim of the task is to identify the correct semantic roles (frame element) r_τ, where $a_\tau \in \{a_1, a_2, ..., a_k\}$, $r_\tau \in R_f$, R_f contains all the frame elements in the frame f. The task definition is denoted as:

$$r_\tau = \underset{r_i \in R_f, w_t \in S,}{\text{argmax}} P(r_i | S, w_t, f_t, a_\tau) \tag{2.2}$$

Figure 1: Task of Frame Semantic Parsing

3 Evaluation Data

The CFN2.2 dataset, which has recently been made publicly, originates from the Chinese Information Processing Team at Shanxi University and their Chinese FrameNet (CFN) initiative. The CFN dataset has been continuously developed since 2004 and now comprises a large-scale dataset with over 100,000 annotated sample sentences.

Compared to the CFN2.1 dataset, CFN2.2 includes an additional 690 annotated samples featuring nested semantic role labels, covering 101 types of semantic roles. The dataset consists

of two sections, frame information and annotated sentences. The corpus is drawn from over 1,100 press releases covering a wide variety of fields. The annotated content includes framings activated by target words as well as the semantic roles dominated by these target words. Each annotated sentence has gone through a double-blind annotation process, dual review, and expert clarification to ensure the quality of the annotated data.

The scale of the CFN2.2 dataset is shown in the Table 2. It's worth noting that in the counting process, for the same sentence, if its target words are different, it will be considered as a different sentence for counting purposes. The number in the brackets denotes the volume of nested semantic roles annotated data.

Dataset Division	Train	Dev	Test_A	Test_B	ALL
Sentences	10700(423)	2300(52)	4400(81)	4600(134)	22000(690)
Frames	638(234)	368(36)	570(25)	507(38)	703(268)
Frame Elements	682(90)	388(15)	571(17)	534(24)	779(101)
Lexical Units	2302	678	1697	1271	3124

Table 2: Statistics of CFN2.2 Dataset

In the task of frame semantic parsing, different frames often contain different semantic information, and the combination of their frame elements is also complex and diverse. These characteristics pose higher requirements for frame semantic analysis models. In addition, in the correspondence between frames and example sentences, a large number of frames only have a few example sentences. As shown in the Figure 2, more than half of the frames only have less than 20 example sentences. In contrast, the frame with the most example sentences has 910 sentences. Although this presents a long tail distribution phenomenon, it conforms to the real rules when humans describe in natural language, which adds to the complexity of the data.

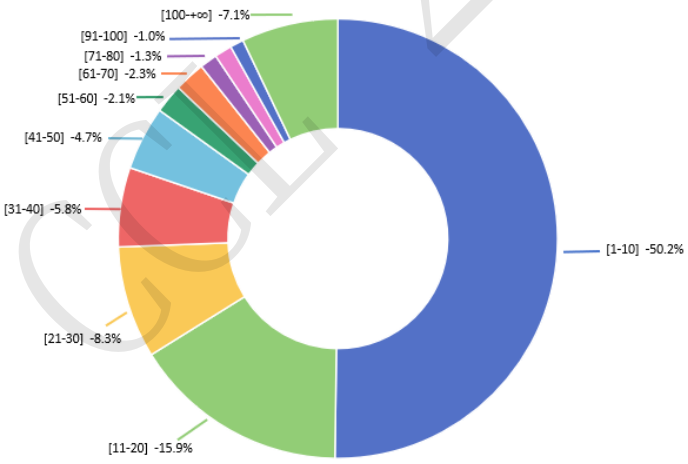

Figure 2: Sentence Range and Proportion in Frame

4 Evaluation Metrics

For the three subtasks in the Chinese Frame Semantic Parsing, the evaluation metrics of this evaluation mainly include the accuracy of frame identification(Acc), the F1-score of argument identification, and the F1-score of role identification. Finally, the scores of the three subtasks are weighted and summed to obtain the final evaluation score.

Frame Identification: The accuracy of frame identification is scored by calculating the ratio of the number of example sentences correctly identified by the model to the total number

of example sentences. The specific calculation formula is:

$$\text{task1\_acc} = \text{correct}/\text{total} \tag{4.1}$$

where correct is the number of predictions made correctly by the model, and total is the total data volume.

Argument Identification: The evaluation method for this task is to calculate the F1 value between the argument range recognized by the model and the actual argument range of the data. The specific calculation formula is:

$$\begin{aligned} \text{task2\_precision} &= \frac{\text{InterSec(gold,pred)}}{\text{Len(pred)}} \\ \text{task2\_recall} &= \frac{\text{InterSec(gold,pred)}}{\text{Len(pred)}} \\ \text{task2\_f1} &= \frac{2*\text{task2\_precision}*\text{task2\_recall}}{\text{task2\_precision}+\text{task2\_recall}} \end{aligned} \tag{4.2}$$

where gold and pred represent the actual result and the predicted result respectively. InterSec(∗) represents calculating the number of tokens shared by both, and Len(∗) represents calculating the number of tokens.

Role Identification: This task strictly judges the boundaries and roles of each argument, also using F1 as an evaluation indicator:

$$\begin{aligned} \text{task3\_precision} &= \frac{\text{Count(gold,pred)}}{\text{Count(pred)}} \\ \text{task3\_recall} &= \frac{\text{Count(gold,pred)}}{\text{Count(gold)}} \\ \text{task3\_f1} &= \frac{2*\text{task3\_precision}*\text{task3\_recall}}{\text{task3\_precision}+\text{task3\_recall}} \end{aligned} \tag{4.3}$$

where gold and pred represent the actual and predicted semantic role sets respectively, and $Count(*)$ represents the number of elements in the set.

Final score: Considering the difficulty of the three sub-tasks, the final score of this evaluation is the weighted sum of the scores of three subtasks, and the specific calculation method is:

$$\text{final\_score} = 0.3 * \text{task1\_acc} + 0.3 * \text{task2\_f1} + 0.4 * \text{task3\_f1} \tag{4.4}$$

5 Submit Results

During the evaluation period, a total of 156 teams registered for the competition, and 16 of them made it into the rematch of the B-rank track. In the end, we chose to reproduce the results of a total of 3 teams.

Rank	Institution	Number	Task1	Task2			Task3			Final
			Acc	P	R	F1	P	R	F1	
1	Individual	Team.1	**71.80**	**85.84**	86.07	85.96	56.97	**60.31**	**58.59**	**70.76**
2	SUDA	Team.2	71.90	87.86	83.30	85.52	57.04	58.21	57.62	70.27
3	Lianyungang Daily	Team.3	71.08	87.53	83.05	85.23	**58.42**	57.15	57.78	70.01
4	Baseline	Team.4	70.43	87.59	78.05	82.54	54.70	53.75	54.22	67.58

Table 3: B-Rank Reproduction Results of Participating Teams(%)

The table lists the scores of 3 participating teams and the baseline model in detail (the scores are based on the reproduction results), and the ranks are based on the final scores. For tasks 2 and 3, the table lists the precision, recall rate and F1 value of each team. In the following text, we will refer to the team numbers in the table to represent different teams for ease of subsequent expression.

As shown in Table 3, although each team proposed a variety of methods to improve performance, the scores of all teams eventually fluctuated around 70.8. This reflects the difficulty for

models like BERT to fully represent all fine-grained semantic information under the constraint of parameter scale. In the future, we are considering introducing larger models or attempting methods such as knowledge distillation. Moreover, many teams did not handle annotation data with constructions as target words and nested semantic roles in a special way. We believe this also to be one of the reasons why it's hard to further improve the final results.

At the same time, we noticed that many methods did not perform as expected on Task 3, while they achieved better results on Task 2. We believe this is related to Task 3, which involves some relatively more fine-grained semantic roles. Clearly, the model can effectively identify the related arguments of the target words in the sentence. However, the current methods cannot effectively identify the semantic roles of each argument in the scene triggered by the target word.

6 Method Overview

After analyzing the technical reports submitted by 3 participating teams and reproducing their model results, we have compiled the main methods used by the teams, in order to analyze the advantages each team has on different tasks.

Team.1 employs effective optimization strategies to enhance model performance. *Team*.2 proposes a method with a pre-trained model and linguistic features, which achieves good results. *Team*.3 achieves impressive results through a unified prompt-based LLM framework with iterative refinements. These methods effectively improved the performance of models in the task of chinese frame semantic parsing.

6.1 Combining Multiple Optimization Strategies

Team.1 formulates frame identification and role identification as classification tasks, and argument identification as an extraction task. To mitigate model instability and improve robustness and generalization, *Team*.1 applies a range of optimization techniques, including exponential moving average, grouped learning rates, and efficient rotational positional encoding.

Post Hoc EMA

Exponential Moving Average (EMA) is a technique that assigns greater weight to recent data, making parameter updates dependent on historical values over time. To mitigate the excessive influence of initialization on the final EMA model in traditional approaches, *Team*.1 adopted the Post Hoc EMA method. This method introduces a dynamically changing decay factor, defined as:

$$x_t = \left(1 - \frac{1}{t}\right)^{1+\gamma} \tag{6.1}$$

where γ is a hyperparameter that controls the rate of decay. The algorithm consists of two main parts: saving EMA model copies for different γ and recovering any γ EMA model after training. This approach enables flexible post-training adjustment of model smoothness without requiring retraining, thereby improving training efficiency and enhancing final model performance.

Token-Aware Virtual Adversarial Training

Team.1 utilizes the Token-Aware Virtual Adversarial Training (TA-VAT) method to improve model performance. TA-VAT consists of two main components: token-level perturbation initialization and token-level perturbation constraint.

The initialization step involves creating a global perturbation matrix for the vocabulary. During each iteration of virtual adversarial training, the accumulated perturbations are used to initialize the corresponding token perturbations, effectively reducing noise caused by random initialization.

In the constraint step, perturbations are updated via gradient ascent and confined within a small normalized sphere to ensure minimal magnitude. Unlike the traditional VAT method, which applies uniform normalization constraints across the entire sequence, TA-VAT introduces a token-level constraint mechanism. This mechanism allows tokens with larger gradients to have larger perturbation boundaries, while tokens with smaller gradients are more strictly constrained. The method contributes to enhancing the robustness of neural networks and yields promising results. The detailed algorithm procedure is as follows.

Algorithm 1 Token-Aware Virtual Adversarial Training

Require: Training sample $S = \{(X = [w_0, \cdots, w_i, \cdots], y)\}$, perturbation boundary ϵ, initialization boundary σ, adversarial steps K, adversarial step size α, model parameters θ

$\mathbf{V} \in \mathbb{R}^{N \times D} \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$ // Initialize perturbation vocabulary
for epoch $= 1, \cdots$ **do**
  for batch $B \subset S$ **do**
    $\delta_0 \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$, $\eta_i^0 \leftarrow \mathbf{V}[w_i]$, $g_0 \leftarrow 0$ // Initialize perturbation and gradient
    for $t = 1, \cdots, K$ **do**
      $g_t \leftarrow g_{t-1} + \frac{1}{K} \mathbb{E}_{(X,y) \in B} [\nabla_\theta L(f_\theta(X + \delta_{t-1} + \eta_{t-1}), y)]$ // Accumulate gradient
      Update word-level perturbation η:
      $g_\eta^i \leftarrow \nabla_{\eta^i} L(f_\theta((X + \delta_{t-1} + \eta_{t-1}), y)$
      $\eta_i^t \leftarrow n_i \cdot \frac{\eta_i^{t-1} + \alpha \cdot g_\eta^i / \|g_\eta^i\|_F}{\|g_\eta^i\|_F}$
      $\eta^t \leftarrow \Pi_{\|\eta\|_F \leq \epsilon}(\eta^t)$
      Update instance-level perturbation δ:
      $g_\delta \leftarrow \nabla_\delta L(f_\theta((X + \delta_{t-1} + \eta_{t-1}), y)$
      $\delta^t \leftarrow \Pi_{\|\delta\|_F \leq \epsilon}(\delta_{t-1} + \alpha \cdot g_\delta / \|g_\delta\|_F)$
    end for
    $\mathbf{V}[w_i] \leftarrow \eta_i^K$ // Update perturbation vocabulary
    $\theta \leftarrow \theta - g_K$ // Update model parameters
  end for
end for

Efficient Rotational Position Encoding

Due to the strong correlation between Rotational Position Encoding (RoPE) parameters and the number of entity classes, introducing new entity types leads to significant parameter growth. In tasks such as frame and role identification, which involve numerous classes, this results in substantial redundancy. To mitigate this, *Team*.1 adopts the Efficient GlobalPointer method, which integrates rotation matrices with Fourier transforms to reduce the parameter count. This approach scales the number of position encoding parameters linearly with the embedding dimension, thereby effectively reducing model complexity. Additionally, the sparsity of the rotation matrix and the locality of the Fourier basis enable the model to capture positional relationships more efficiently.

6.2 Leveraging Stronger Pre-trained Language Models and Linguistic Features

Based on the hypothesis that larger pre-trained models possess greater capacity to learn complex language patterns and knowledge, thereby enabling better generalization to downstream tasks, *Team*.2 utilized a larger pre-trained model, resulting in performance improvements across all three subtasks.

In addition, *Team*.2 adopted specific optimization strategies for the FI and RI subtasks. Since the task data already provides explicit target word information—including the start and end boundary indices—*Team*.2 first obtained updatable target word embeddings by combining pre-trained word embeddings with the output of the BERT model. These embeddings were

then refined during training. Subsequently, the entity classification method with Rotational Positional Encoding (RoPE) was applied to compute attention scores and determine the most likely frame for each target word.

Furthermore, considering that structural information from Chinese word segmentation might benefit the tasks, *Team.*2 incorporated segmentation structure into the model. They 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.

Notably, to mitigate the potential bias in model training caused by data distribution imbalance, *Team.*2 replaced the standard Cross-Entropy Loss with Focal Loss, which dynamically adjusts the weight of each sample to focus learning on harder examples. The formulation of Focal Loss is shown in Equation 6.2:

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{6.2}$$

where p_t denotes the probability assigned by the model to the correct class.

Finally, a voting mechanism was employed across the three subtasks to further enhance performance. As a result, *Team.*2 achieved first place on TestA and second place on TestB.

6.3 Full-Scale Fine-Tuning of Large Language Models

Team.3 proposes a unified approach to Chinese Frame Semantic Parsing by employing a full-scale fine-tuned Qwen3 model, addressing Frame Identification, Argument Identification, and Role Identification within a single pipeline. Initially, an instructive prompt was used to guide the Qwen3 model to perform all three sub-tasks simultaneously, as illustrated in Figure 3.

```
你是汉语框架语义解析大师。
子任务1：框架识别(Frame Identification)，识别句子中给定目标词激活的框架。
子任务2：论元范围识别(Argument Identification)，识别句子中给定目标词所支配论元的边界范围。
子任务3：论元角色识别(Role Identification)，预测子任务2所识别论元的语义角色标签。
原句：餐饮业是天津市在海外投资的重点之一。
分词后的句子：餐饮业/n 是/v 天津市/ns 在/p 海外/nl 投资/v 的/u 重点/n 之一/r 。
目标词1：是，目标词词性：v。
请帮我预测给定目标词激活的框架、给定目标词所支配论元的边界范围和所识别论元的语义角色标签。
```

Figure 3: Prompt template for Chinese Frame Semantic Parsing

To address suboptimal performance in Argument Identification and Frame Identification, the team introduced targeted optimization strategies: Direct Argument Output and Majority Voting, respectively. Specifically, for Argument Identification, the output format was modified to return argument spans as text (e.g., “餐饮业”) instead of numerical indices (e.g., [0, 2]). For Frame Identification, a majority-voting mechanism was implemented, whereby the model generated 16 candidate frame predictions per sentence, and the most frequently predicted frame was selected. These enhancements significantly improved performance across the sub-tasks, highlighting the potential of large language models (LLMs) in complex semantic parsing tasks.

Furthermore, Team.3 explored the integration of Chain-of-Thought (CoT) reasoning by prompting the Qwen3 model to generate brief justifications for its predictions across all three sub-tasks. However, this approach yielded subpar results, potentially due to a misalignment with the model’ s fine-tuning objectives. This indicates that future research could focus on

deeper exploration of CoT integration, particularly through the use of higher-quality reasoning models or more effective prompt engineering strategies.

7 Summary

This evaluation task, based on previous tasks, expands the dataset with more fine-grained annotation data for nested semantic roles. It focuses on Chinese sentences in which common semantic cores are conveyed through specific sentence structures. This enhances frame semantic analysis and further facilitates deeper language understanding.

This evaluation is of great significance for fine-grained semantic analysis, and it has also attracted a large number of teams from the academic and industrial sectors to register for the competition. Due to the high difficulty of the evaluation task, fine-grained semantics, and the target word is no longer a single word. Small models lack semantic understanding when facing a large number of frames, and are unable to cope with a large number of role types in role tagging. Large models lack frame semantic knowledge and cannot distinguish between subtle semantic differences among a large number of frames. They also struggle to correctly identify argument roles in sentences. This reflects that there are still tremendous development prospects for this task.

In general, this evaluation targets the deficiencies of existing models in fine-grained semantic analysis, using the Chinese frame semantic parsing task to assess the model's scenario depiction capabilities. Future evaluations could consider expanding the data coverage fields, covering more semantic scenarios, and evaluating the model's understanding of fine-grained semantic scenarios in a more comprehensive way, further promoting the development of the Chinese FrameNet.

Acknowledgements

We would like to thank the CCL Evaluation Committee for their support and Beijing PARATERA Tech Corp.,Ltd. for sponsoring this evaluation. This work was supported by the National Natural Science Foundation of China (Nos.61936012, 62376144), as well as the Natural Language Processing Innovation Team Project (Sanjin Talent) of Shanxi Province.

References

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.

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.

Rohit J. Kate, Yuk Wah Wong, and Raymond J. Mooney. 2005. Learning to transform natural to formal languages. In *AAAI Conference on Artificial Intelligence*.

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*, pages 1–18.

Liping You and Kaiying Liu. 2005. Building chinese framenet database. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*.

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*, 210:106480.