Overview of CCL24-Eval Task1: Chinese Frame Semantic Parsing Evaluation

Peiyuan Yang[1,‡], Juncai Li[1,‡], Zhichao Yan[1,‡], Xuefeng Su[1,3,‡], Ru Li[1,2,⋆,†]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006,China

[3]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology, Jinzhong, Shanxi 030609,China

[‡]{202222407058,202312407010, 202312407023, 201912407008}@email.sxu.edu.cn

[⋆]liru@sxu.edu.cn

Abstract

Chinese Frame-semantic Parsing (CFSP) aims to extract fine-grained frame-semantic structures from texts, which can provide fine-grained semantic information for natural language understanding models to enhance their abilities of semantic representations. Based on the CCL-23 CFSP evaluation task, we introduce construction grammar to expand the targets, as basic units activating frames in texts, from word-style to construction-style, and publish a more challenging CFSP evaluation task in CCL-2024. The evaluation dataset consists of 22,000 annotated examples involving nearly 695 frames. The evaluation task is divided into three subtasks: frame identification, argument identification, and role identification, involving two tracks: close track and open track. The evaluation task has attracted wide attention from both industry and academia, with a total of 1988 participating teams. As for the evaluation results, the team from China University of Petroleum won the first place in the closed track with the final score of 71.34, while the team frome Suzhou University won the first place in the open track with the final socre of 48.77. In this article, we reports the 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 of the CCL-2024 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), its aim is 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 1398 frames, involves 18360 lexical units, and more than 100 thousand annotated sentences.

Currently, in the Chinese FrameNet dataset, the lexical units activating frames are solely single words. However, in certain sentences, individual target words are insufficient to fully illustrate complex semantic scenarios. Take “爱买不买” as an example, it indicates that the speaker doesn't care or is uninterested in whether the counterparty wants to purchase something. Traditional methods analyze this phrase by taking words as units, introducing verbs like “爱” and “买” as target words, activating scenarios of *liking* or *purchasing*, yet falling short of expressing the true meaning of the phrase.

Construction grammar was first proposed by Professor Fillmore in 1988(Fillmore et al., 1988). It advocates that language knowledge consists of fixed, meaningful units, which are referred to as constructions. These units can be as simple as words or phrases, or as complex as sentences or utterances. Thus,

[1]Task website https://tianchi.aliyun.com/competition/entrance/532179/introduction

[†] Corresponding Author

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 32-42, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the phrase “爱买不买” is an entity that expresses semantics, which triggers the $Emotion\_directed$ frame.

To enhance the capability of chinese frame semantic parsing, and achieve a deeper understanding of language, we have expanded our data to include constructions as “target words” for semantic frame parsing. Consequently, we have launched the second 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 is 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 “Change_position_on_a_scale”. This frame represents the semantic scenario conveying the following meaning: “The relative position of an entity on a certain dimension (i.e., a certain attribute) changes, with its attribute value transitioning from the initial value to the final value”.

Frame Element: Frame elements refers 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 “Entity” and “Attribute” in the frame “Change_position_on_a_scale” 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 article, 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.

Frame	Change_position_on_a_scale	
Definition	The relative position of an entity on a certain dimension (i.e., a certain attribute) changes, with its attribute value transitioning from the initial value to the final value.	
Elements	**Name**	**Definition**
	Entity	An entity with a definite quantity on a certain attribute.
	Attribute	Entity's attribute with quantitative variation.
	Initial_value	The starting point of an entity's attribute value variation..
	Final_value	The final quantity reached by the entity.
	Initial_state	The state of the entity before experiencing attribute value changes.
	Final_state	The state of the entity after experiencing attribute value changes.
	Difference	The magnitude of the entity's change in a certain dimension.

Table 1: The “Change_position_on_a_scale” frame and the frame element information it contains.

2.2 Task Description

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

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 don'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:

$$f_t = \underset{f_i \in F, w^t \in S}{\operatorname{argmax}} P(f_i | S, w^t) \tag{3.1}$$

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:

$$r_\tau = \underset{r_i \in R_f, w_t \in S,}{\operatorname{argmax}} P(r_i | S, w_t, f_t, a_\tau) \tag{3.2}$$

Figure 1: Task of Frame Semantic Parsing

3 Evaluation data

The CFN2.1 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 CFN2.0 dataset, CFN2.1 adds two thousand annotated data samples with construction as target words. 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 CFN2.1 dataset is shown in the table2. 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 construction-oriented annotated data.

Dataset Division	Train	Dev	Test_A	Test_B	ALL
Sentences	10700(700)	2300(300)	4400(400)	4600(600)	22000(2000)
Frames	671(32)	354(24)	432(32)	504(33)	695(86)
Frame Elements	947	649	711	796	987
Lexical Units	2359	670	931	572	3132

Table 2: Statistics of CFN2.1 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 figure2, more than half of the frames only have less than 20 example sentences. In contrast, the frame with the most example sentences has 904 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.

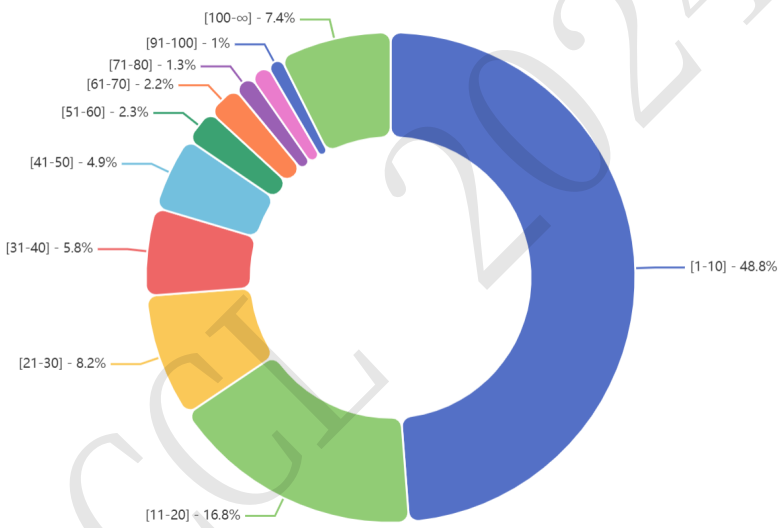

Figure 2: Sentence Range and Proportion in Frame

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:

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

where gold and pred represent the actual result and the predicted result respectively. $\text{InterSec}(*)$ represents calculating the number of tokens shared by both, and $\text{Len}(*)$ 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:

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

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 1988 teams registered for the competition, and 29 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 10 teams from both tracks.

Track	Rank	Institution	Number	task1	task2			task3			final
				Acc	P	R	F1	P	R	F1	
Closed	1	Individual	Team.1	**72.49**	90.19	83.14	86.53	60.20	58.01	59.09	**71.34**
	2	BNU	Team.2	72.42	89.08	83.50	86.20	59.34	**58.97**	59.15	71.25
	3	SQU	Team.3	71.13	**90.46**	**83.75**	**86.97**	**60.22**	58.57	**59.38**	71.18
Open	1	SUDA	Team.4	**58.62**	44.83	53.91	48.95	**44.08**	**38.74**	**41.24**	**48.77**
	2	PKU	Team.5	52.54	52.17	**67.84**	**58.99**	14.52	19.52	16.65	40.12
	3	UIR	Team.6	23.06	**67.66**	35.62	46.67	1.75	1.17	1.40	21.48

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

The table lists the scores of 6 participating teams 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 accuracy, 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.

In the closed track, even though each team proposed a variety of methods to improve performance, the scores of all teams eventually fluctuated around 71.2. 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 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 involving a large number of

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.

The experimental results on the open track show that the performance of LLM does not stand out in the frame identification task. In this task, $Team.4$ has relatively favorable results. They used word vector cosine similarity to pre-select part of the frames instead of choosing from all the frames. We suspect this phenomenon occurs because the large language models can't distinguish subtle differences among a large number of frames. Meanwhile, $Team.4$ also achieved the best result in Task 3, suggesting that the accuracy of frame identification has a significant impact on role identification.

6 Method overview

After analyzing the technical reports submitted by 6 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. In the closed track, $Team.1$ adopts the Token-Aware Virtual Adversarial Training (TA-VAT) method to improve the performance of the model. $Team.2$ proposed the Extraction Method for Span Type Data for the Argument Identification task, which achieved good results. $Team.3$ achieved pretty results by using a large language model and data augmentation techniques. In the open track, $Team.4$ reduce the number of candidate frames for each target words by using word vector cosine similarity. $Team.5$ build a hierarchical index RAG system based on target words. These methods effectively improved the performances of large models in the task of chinese frame semantic parsing.

6.1 Closed Track

Data Augmentation.

$Team.3$ using large language models like ChatGPT for automated data augmentation, diverse and coherent text variants are created which enriches the diversity of the training data. This significantly reduces the data preparation time, rapidly generates a large quantity of high-quality samples, accelerates the model training, and enhances model performance and robustness.

Post Hoc Exponential Moving Average.

$Team.1$ addressed the excessive influence of initialization on the final EMA model in traditional EMA methods by adopting a Post Hoc EMA method. This method uses a dynamically changing decay factor, defined as:

$$\beta(t) = (1 - 1/t)^{1+\gamma} \tag{6.1}$$

This is divided into two parts: saving EMA model copies for different γ and recovering any γ EMA model after training. After the training process ends, any γ EMA model can be restored through the saved EMA model copies. This method allows flexibility in adjusting the smoothness of the model after training, avoiding retraining, and significantly enhancing model training efficiency and outcomes.

ALiBi Relative Position Encoding.

Since the self-attention mechanism in Transformer is independent of the text order, it is usually necessary to provide explicit positional signals to the Transformer. The original Transformer uses sinusoidal or learned positional embeddings. Although absolute positional encoding is simple to imple- ment and suitable for fixed-length sequences, it performs poorly when handling sequences of different lengths and capturing relative positional relationships. Thus $Team.2$ use ALiBi (Attention with Linear Biases) (Press et al., 2021) positional encoding, which adds a linearly decreasing penalty propor- tional to the distance to the dot product of key and query in the Attention model,this encoding approach has achieved good results.

As shown in Figure 3, the left diagram is similar to the traditional $Transformer$, where the initial attention score is obtained through the dot product of key and $query$. The right diagram shows a relative distance matrix, where the elements of the matrix are the differences between the indices i and j of q_i

and k_j. The third term m is a fixed slope parameter, which depends on the number of $heads$ in the $Attention$.

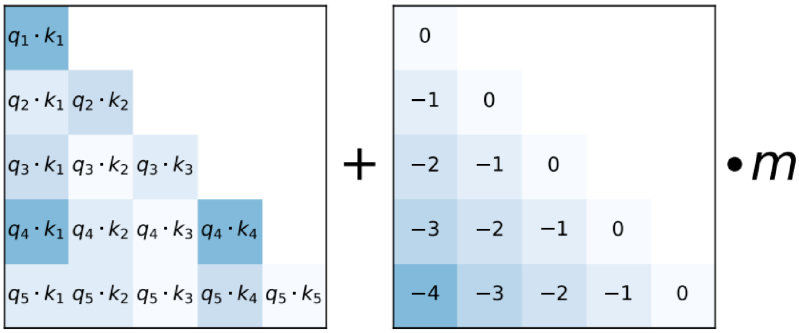

Figure 3: Attention with Linear Biases(ALiBi)

Extraction Method for Span Type Data.

Considering the characteristics of the Argument Identification task, $Team.2$ adopts an extraction method for a type of data called “span” data, where predictions are made for the start and end of the arguments. They treat each given sentence S as a “span” type data and label it with a “head-tail matrix”. The head-tail matrix is an upper triangular matrix, and can be used as follows: the row number (vertical coordinate) represents the starting index of the predicted argument, while the column number (horizontal coordinate) representsthe ending index of the predicted argument. “1” is marked at the start and end indices of the predicted argument, while “0” is marked in all other positions of the matrix.

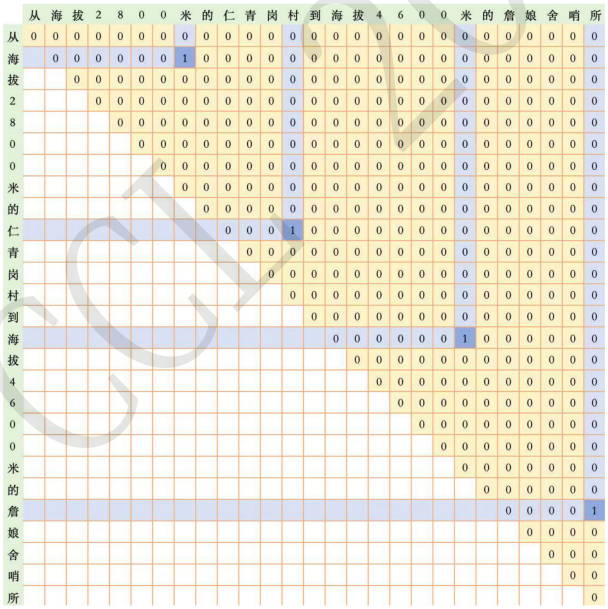

Figure 4: An example of using H-T matrix of “span” data

Token-Aware Virtual Adversarial Training.

$Team.1$ adopts the Token-Aware Virtual Adversarial Training (TA-VAT) method to improve the performance of the model. The TA-VAT method mainly includes two steps: initialization of word-level perturbation and constraint of word-level perturbation. The initialization of word-level perturbation establishes a global perturbation vocabulary, and in each virtual adversarial training process, the accumulated perturbation is used to initialize the corresponding word perturbation, avoiding the noise brought by random initialization. The constraint of word-level perturbation uses gradients to update the pertur-

bations after initialization, and constrains these perturbations within a small normalized sphere to keep them minimal. While the traditional VAT method (Miyato et al., 2018) applies normalizing spheres to the entire sequence, TA-VAT proposes a word-level constraint method, where words with larger gradients are allowed a larger perturbation boundary, and words with smaller gradients are subject to smaller constraints. These two methods effectively increase the robustness of neural networks and achieve good results. he algorithm procedure is shown 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

6.2 Open Track

Frame Identification.

When $Team.4$ is building a frame identification task prompt, they find it exceeds the maximum number of tokens limit imposed by Gemini in a prompt. For this, they reduce the number of candidate frames for each target words. Specifically, they identify all corresponding frames for each target word through the mapping between words and frames in the dataset. For target words without an existing mapping they compute the cosine similarity between the word vectors of the target words and each frame, selecting the six most similar frames as candidates. In addition, some frame names are long and not tokenized, they use Jieba for tokenization. The word vectors they use are from FastText 0, but some target words or tokenized frame names might not be in FastText. In these instances, they sum the word vectors of each character from the tokenization to obtain the final word vector representation.

$Team.5$ build a hierarchical index RAG system based on target words, which uses keyword information to filter out a certain amount of options, reduces the length of tokens, and avoids the decline of LLM reasoning ability caused by long tokens. At the same time,they use the HanLP tool to segment the sample sentences to make the sentence structure clearer.Then, they constructed a balanced Few-Shot sample category. For each target word category, they matched the nearest pieces of data as a Few-Shot, ensuring that each category of the target word had the same number of data as samples, and at the same time using BM25 (Robertson et al., 1994) to ensure that the selected data were the closest to the problem. As for the specific principle of BM25 retrieval algorithm, they first analyze the morpheme of the sentence to generate morpheme q_i. They directly regard the process of word segmentation through hanlp as morpheme analysis, and each word segmentation is regarded as morpheme q_i. Then, for each search

statement d, the correlation score of each morpheme q_i and d is calculated. Finally, the correlation score of q_i relative to d is weighted and summed to obtain the correlation score of the sentence and d.

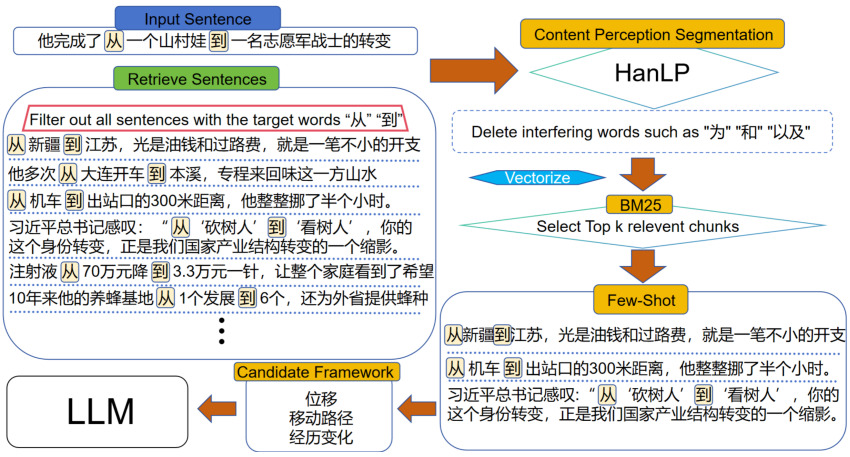

Figure 5: Hierarchical indexing RAG system based on target words

Argument Identification.

Because the FI performance of Gemini is relatively poor, and errors in FI propagate to the AI subtask, to avoid error propagation $Team.4$ utilizes the FI results of the small model as the basis for the AI subtask. During the training process, they found that the frame distribution is severely imbalanced, with many frames only having one or two examples. In order to solve the data imbalance issue and improve the performance of the small model, they use the LLM to generate examples for frames without any examples. For frames with fewer than 15 examples, they augment the existing examples by duplicating them until each frame has a total of 15 examples.

$Team.5$ note that LLM is often not good at regular mapping, but is sensitive to semantics. It can effectively improve the performance of the model by handing over the mechanical work to pre-processing and post-processing. Therefore, based on the same construction of RAG system and high-quality Few-Shot samples, they change the input from the model to text, and then conduct postprocessing to convert it back to the list, in order to reduce the computational load on the model, rather than waste its attention on the mapping relationship.They also incorporated the Agent features of LLM and limited the specific output format of LLM in prompts.

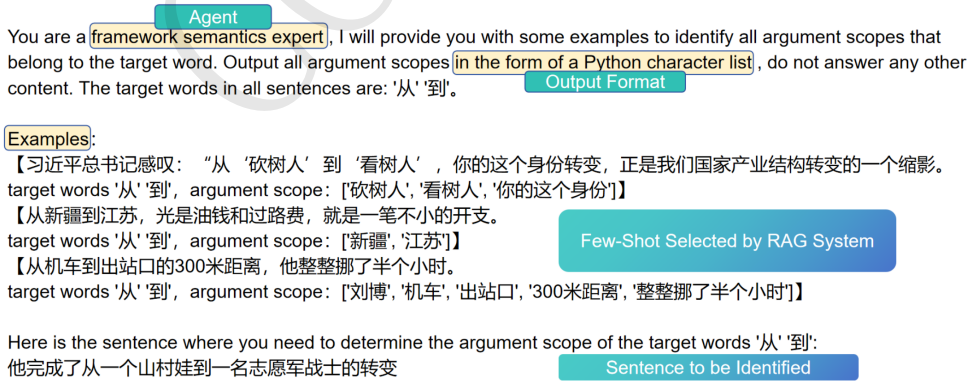

Figure 6: Sample semantic enhancement conversion process example

Role Identification prompt.

$Team.4$ use the small model to obtain more accurate AI results for RI subtask.To enhance the performance of the small model, they employ the self-training and ensemble technique. Specifically, they first use a trained model to predict the data from CoNLL09 and CPB1.0. Then, they used the predicted

data to train a new model. This model is then fine-tuned on the training data, the final argument result is selected by voting.

7 Summary

This evaluation task, based on previous tasks, introduces construction grammar, and increases the data with a construction as the target word. It focuses on sentences in Chinese where some common semantic cores are expressed through a specific structure in the sentence. This enhances the capability of the frame semantic analysis and further enables a deeper understanding of language.

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 vocabulary. 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, adding more data with constructions as target words, 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 Frame Net.

Acknowledgements

Thanks to the support of the key project of the National Natural Science Foundation of China (No. 61936012).

Thanks for the support of the CCL Evaluation Committee.

References

Charles J Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538.

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.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Ofir Press, Noah Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

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.