CCL Frontier Forum 2024

The 23rd Chinese National Conference on Computational Linguistics

Proceedings of the Evaluation Workshop

August 24 - August 28, 2024

Taiyuan, China

©The 23rd Chinese National Conference on Computational Linguistics

Order copies of this and other CCL proceedings from:

Chinese National Conference on Computational Linguistics (CCL)

Courtyard 4, South Fourth Street, Zhongguancun , Haidian District, Beijing 100190, China

Tel: + 010-62562916

Fax: + 010-62661046

cips@iscas.ac.cn

Introduction

Welcome to the Evaluation Workshop of CCL 2024, abbreviated as CCL24-Eval. Since 2017, the yearly CCL conference has included evaluation workshops to facilitate proposal of important new tasks and accompanying evaluation metrics, and release of new datasets in the language information processing community in China. We expect participants to thoroughly utilize state-of-the-art models and propose novel approaches to help us understand the technique boundaries and potential directions for future research.

The major change of CCL24-Eval, compared with previous ones, is that we publish selected overview reports and system reports at ACL/CCL anthology, making it easier for future researchers to read and cite the papers. We also plan to release the talk videos and slides on the CCL 2024 website. All our efforts are aimed at promoting technique accumulation and peer communication. CCL2024-Eval includes 10 tasks, 4 tasks on semantic parsing, 3 tasks on native Chinese learners' test, 2 tasks on multi-modal data processing, and 1 task on ancient Chinese processing. Each task has an organizing committee responsible for defining the task, providing training/dev/test data, collecting submissions, running evaluation, releasing results, and writing an overview report. For each task, the organizing committee builds a separate reviewer pool, probably including the committee themselves, for reviewing system reports of the same task. Each team is encouraged to submit a system report, especially if they have achieved good results or attempted novel approaches. Each system report is reviewed by at least two reviewers, and is required to make corresponding modification according to their reviews. The overview reports are reviewed by at least two members of the CCL24-Eval committee.

Finally, we have accepted 10 overview reports and 36 system reports. Here we sincerely thank all reviewers, organizers, and participants for their hard work.

Hongfei Lin, Bin Li, Hongye Tan

July 2024

Organizers

Evaluation Workshop Chairs

Hongfei Lin	Dalian University of Technology, China
Hongye Tan	Shanxi University, China
Bin Li	Nanjing Normal University, China

Table of Contents

System Report for CCL24-Eval Task 1: Construction of CFSP Model Based on Non-Finetuning Large Language Model

Fugeng Huang[a,1], **Zhongbin Guo**[b,1], **Wenting Li**[c,†], **Haibo Cheng**[d,†]
[a]School of Software and Microelectronics, Peking University, Beijing, 102600, China
[b]School of Computer and Technology, Beijing Institute of Technology, Beijing, 100081, China
[c]School of Information Engineering, Beijing Institute of Graphic Communication, 102600, China
[d]National Engineering Research Center for Software Engineering, Peking University, 100871, China
[a]2301210243@stu.pku.edu.cn
[b]1120220508@bit.edu.cn
[c,d]{wentingli, hbcheng}@pku.edu.cn

Abstract

Chinese Frame Semantic Parsing (CFSP) is an important task in the field of Chinese Natural Language Processing(NLP). Its goal is to extract the frame semantic structure from the sentence and realize the deep understanding of the events or situations involved in the sentence. This paper mainly studies the application of Large Language Model (LLM) for reasoning through Prompt Engineering without fine-tuning the model, and completes three subtasks of Chinese Framework Semantic Parsing tasks: frame identification, argument Identification and role identification. This paper proposes a Retrieval Augmented Generation (RAG) method for target words, and constructs more refined sample Few-Shot method. We achieved the second place on the B rankings in the open track in the "CCL2024-Eval The Second Chinese Frame Semantic Parsing" competition*.

1 Introduction

Chinese Frame Semantic Parsing (CFSP) is a research method based on frame semantics (Charles J. Fillmore, 1982), which is based on the semantic representation and annotation of Chinese FrameNet (CFN) (You et al., 2007; You and Liu, 2005; Li et al., 2024), and achieves the purpose of semantic parsing by extracting the frame semantic structure of sentences (Gildea and Jurafsky, 2002). This method is of great significance for a series of downstream tasks such as reading comprehension (Guo et al., 2020a; Guo et al., 2020b; Wang et al., 2016), text summarization (Guan et al., 2021a; Guan et al., 2021b), and relationship extraction (Zhao et al., 2020).

1.1 Task Definition

Table 1 presents the 'Deciding' framework within the Chinese FrameNet, which illustrates the cognitive process of making decisions among various explicit or potential options. Framework elements refer to the participants in the semantic scene corresponding to the framework. For example, 'Cognizer' in the "Deciding" framework is one of the framework elements.

†Corresponding authors.
*https://tianchi.aliyun.com/competition/entrance/532179
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1-9, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Frame name	Deciding
Frame definition	The cognitive person makes a decision in a variety of explicit or potential choices, which may be an entity or a behavioral process.
Frame element	**Cognizer** The cognitive decides to do something.
	Decision The decision represents the entity or process determined by the cognitive.
	Possibilities The cognizer decides which one to choose from a range of possible options.
	Topic Guided by "about", it indicates the matters involved in the decision, and sometimes the matters involved also indicate the general content of the decision.
	Time Indicates the relative position of the occurrence, progress or end of the action behavior or state in the time dimension, including both time points and time periods.

Table 1: An example of "decision" frame in Chinese FrameNet (CFN)

This evaluation divides Chinese Frame Semantic Parsing into three downstream tasks: Frame Identification (FI) (Su et al., 2021), Argument Identification (AI) and Role Identification (RI). The task of FI is the core task in the research of frame semantics. It requires finding an activated frame for the target word in a given sentence according to its context, which can help the computer identify the key information and semantic framework in the sentence, so as to better understand the meaning of the sentence (Hermann et al., 2014). The main purpose of AI task is to determine the position of argument (i.e. frame element) involved in each target word in the sentence, so as to help the system more accurately identify the semantic role of argument. The RI task aims to determine the semantic role of each argument in its own framework, which plays a vital role in information extraction, relationship extraction and machine translation. The following figure shows the specific work examples of the three downstream tasks of CFSP.

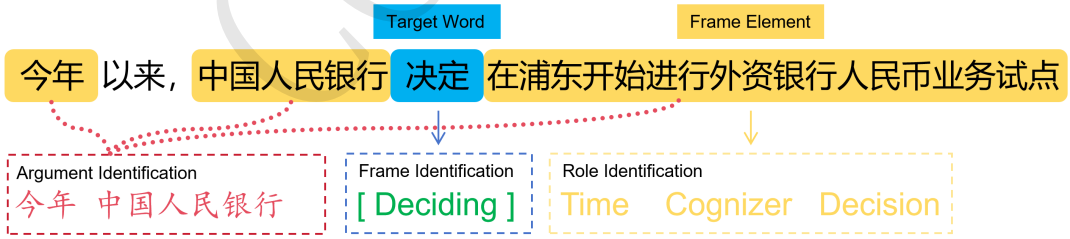


Figure 1: Schematic diagram of Chinese Frame Semantic Parsing task

With the continuous progress of technology, ChatGPT and other Large Language Model (LLM) continue to appear, researchers have also carried out a series of research on the CFSP task based on LLM. The baseline scheme given by the evaluation task shows that under the guidance of Chain of Thinking (CoT), the performance of the three subtasks of CFSP in Zero-Shot and Few-Shot scenarios is not ideal, and LLM cannot understand the text from the perspective of frame, argument and role (Li et al., 2023). This research is mainly committed to using LLM to effectively complete the three downstream tasks under the task of CFSP through the improved scheme.

1.2 Contribution

Our main contributions can be summarized in the following four points:

1. We build a hierarchical index Retrieval Augmented Generation (RAG) system based on target words, use the target word information to filter out some options.
2. We use HanLP to segment sentences, and use the BM25 (Robertson et al., 1994) retrieval algorithm to index, which effectively improves the sample quality of LLM in the Few-Shot scenario.
3. We establish balanced Few-Shot sample categories to ensure that each target word category has a certain amount of data closest to the problem.
4. We change the input from the model to text, and then conduct post-processing to convert it back to the list, so as to avoid wasting attention on the mapping relationship in the learning process of the model.

2 Related Work

Due to the late appearance of LLM, the relatively novel architecture, and the rapid development and updating speed, there are few researches on the application of LLM in the Chinese Frame Semantic Parsing tasks. Yang (Yang et al., 2023) represents the first attempt at leveraging large pre-trained language models (LLM) for SPARQL generation to address Chinese knowledge graph question answering. (Li et al., 2023) tested the effect of ChatGPT on FI task in Zero-Shot and Few-Shot scenarios, and tried to design the Cot to carry out multiple rounds of dialogue, guiding ChatGPT to better complete the tasks of AI and RI, but the final performance was not ideal as well.

3 Model

3.1 Baseline Evaluation

The baseline model given in this task mainly guides LLM by building the Chain-of-Thought (CoT) prompting method (Wei et al., 2023), and tests the results of ChatGPT in the Zero-Shot and Few-Shot scenarios. The specific operation and effect are shown in the following figure:

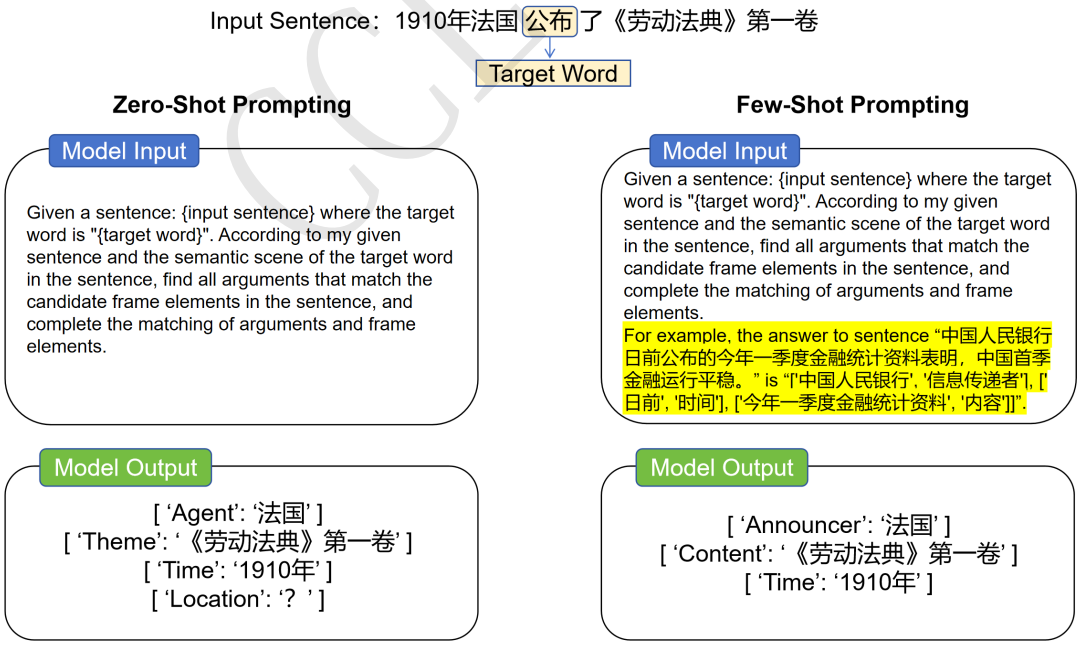


Figure 2: Example of comparison between the effects of Few-Shot and Zero-Shot

https://www.hanlp.com/

It is not difficult to see that compared with Zero-Shot input, The output of the model in Few-Shot scenario is much better, reducing the output of useless information and capturing the corresponding arguments and roles more accurately. The research results of (Li et al., 2023) also show that the accuracy rate of LLM for FI task can only reach 37% in the scenario without samples, while it is significantly increased to 53% in the scenario with samples, which is enough to show the importance of giving samples of corresponding target words for Chinese Frame Semantic Parsing tasks.

3.2 Model Construction

Through the effect analysis of the baseline model, we found that the **target word** is undoubtedly the key to identify the semantic framework, which directly affects the performance of LLM on FI task. Therefore, we build a **hierarchical index RAG system** (Chen et al., 2024) 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, because the Few-Shot scenario greatly improves the performance of LLM, the sample quality provided to LLM is also an important link to determine the performance of the model. Therefore, we first use the HanLP tool to segment the sample sentences to make the sentence structure clearer. Then, in order to make the ability of the model to identify each frame similar, we constructed a **balanced Few-Shot sample category**. For each target word category, we 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, we first analyze the morpheme of the sentence to generate morpheme q_i. In this study, we 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. Its general formula can be written as below:

$$Score(Q,d) = \sum_{i}^{n} IDF(q_i) \cdot R(q_i,d)$$

Where Q is the desired statement, $IDF(q_i)$ indicates the weight of morpheme q_i in the set of attributive sentences, which is called Inverse Document Frequency (IDF). That is, when many sentences contain q_i, The discrimination of q_i is not high, so the importance of using q_i to judge correlation is low, the lower the $IDF(q_i)$. The specific weight calculation method is as follows:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

Where $R(q_i,d)$indicates the correlation score between morpheme q_i and sentence d. In this study, BM25 algorithm in elasticsearch* is applied, and its correlation calculation method is as follows:

$$R(q_i,d) = \frac{f_i \cdot (k_1+1)}{f_i + K}$$

$$K = k_1 \cdot (1 - b + b \cdot \frac{dl}{avgdl})$$

k_1 and b are set to 1.2 and 0.75 by default, f_i represents the frequency of morpheme q_i in Sentence d, $avgdl$represents the average length of all sentences.

To sum up, the correlation score formula of BM25 algorithm in the study can be integrated as follows:

$$Score\,(Q,d) = \sum_{i}^{n} IDF\,(q_i) \cdot \frac{f_i \cdot (k_1+1)}{f_i + k_1 \cdot (1 - b + b \cdot \frac{dl}{avgdl})}$$

*https://www.elastic.co/cn/elasticsearch

On this basis, the framework set corresponding to selected sentences is selected as a candidate frame for LLM to choose and judge. The overall RAG system construction process is as follows:

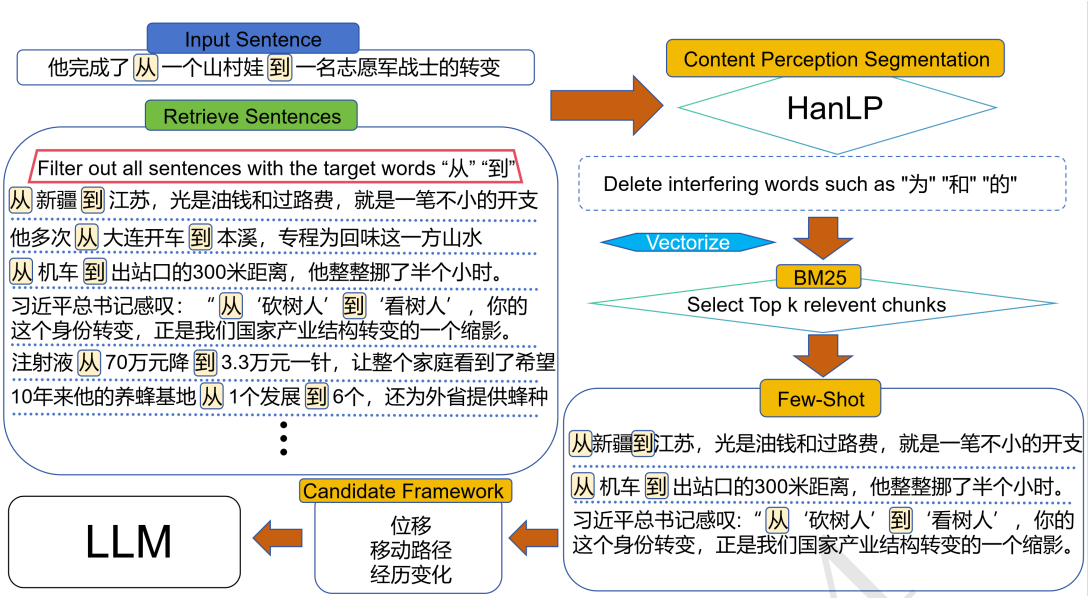


Figure 3: Hierarchical indexing RAG system based on target words

After the above operation, we noticed that about 13% of the data candidate frame information in the test set was empty, and the test found that the probability of the frame identified by LLM belonging to the given semantic frame was less than 4%. We made a special treatment for the target words of this part of Zero-Shot. By giving all the frame options to LLM for judgment, the probability of identifying the frame belonging to the given semantic frame increased to 94.6%.

For AI and FI tasks, we 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, we change the input from the model to text, and then conduct post-processing 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. We also incorporated the Agent features of LLM and limited the specific output format of LLM in prompts. The specific conversion process and prompt example are shown below:

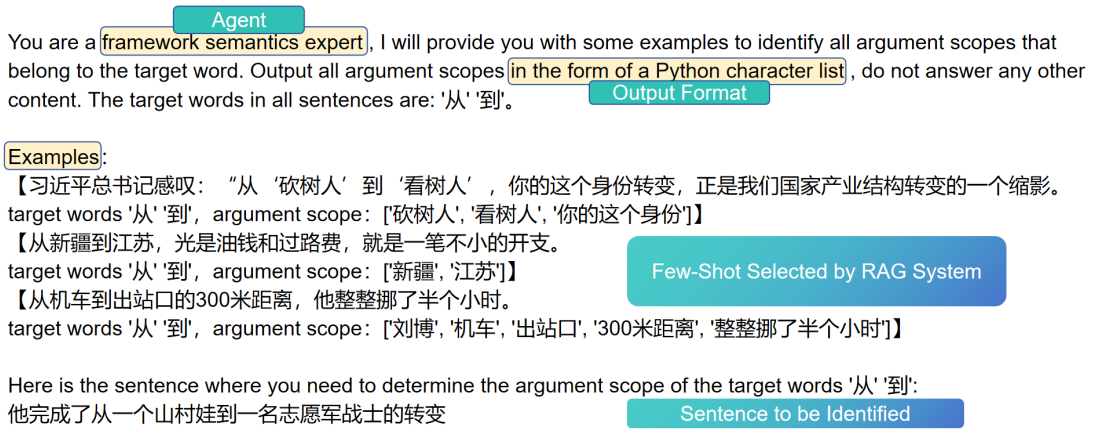


Figure 4: Sample semantic enhancement conversion process example

4 Experiment

4.1 Experimental Setup

The following table shows the official data distribution of this evaluation (Yang et al., 2024). The experimental research in this paper is all based on the ChatGPT (gpt-3.5-turbo) model of Openai[†].

	Training set	Validation set	Test(B) set	Total
Sentences	10700(700)	2300(300)	4600(600)	17600(1600)
Frames	671(32)	354(24)	504(33)	695(86)
Frame Elements	947	649	796	987
Lexical Units	2359	670	572	3132

Table 2: CFN2.1 Dataset distribution

4.2 Evaluation Matrix

1. Frame Identification (FI)

   The only evaluation criterion for FI task is the accuracy:

   $$FI_{acc} = \frac{correct}{total}$$

   Where $correct$ indicates the correct quantity predicted by the model, $total$ is the total number of frames to be identified.

2. Argument Identification (AI)

   AI task adopts precision (P), recall (R) and F1-Score ($F1$) were used as evaluation indexes.

   $$AI_P = \frac{\text{InterSec(gold, pred)}}{\text{Len(pred)}}$$

   $$AI_R = \frac{\text{InterSec(gold, pred)}}{\text{Len(gold)}}$$

   $$AI_{F1} = \frac{2 \times AI_P \times AI_R}{AI_P + AI_R}$$

   Of which, $gold$ and $pred$ represent the real results and the predicted results respectively, $Intersec(*)$ means to calculate the number of tokens shared by both, $Len(*)$ indicates to calculate the number of tokens.

3. Role Identification (RI)

   RI task also use P, R and $F1$ as evaluation indexes.

   $$RI_P = \frac{Count(gold \cap pred)}{Count(pred)}$$

   $$RI_R = \frac{Count(gold \cap pred)}{Count(gold)}$$

   $$RI_{F1} = \frac{2 \times RI_P \times RI_R}{RI_P + RI_R}$$

   Where $gold$ and $pred$ represent the real and predicted results respectively, $Count(*)$ indicates that the number of set elements is calculated.

[†]https://platform.openai.com/docs/models/gpt-3-5-turbo

CCL 2024

The final total evaluation score is the score obtained by weighting the three downstream tasks in the proportion of $(0.3, 0.3, 0.4)$, i.e

$$Score = 0.3 \times FI_{acc} + 0.3 \times AI_{F1} + 0.4 \times RI_{F1}$$

4.3 Evaluation Results and Analysis

The following table lists the scores of each evaluation index of the top three teams of the open track in this evaluation task:

Team ID	Score	AI_R	RI_R	AI_P	RI_P	FI_{acc}	AI_{F1}	RI_{F1}
Tangled	**48.77**	53.84	**38.66**	44.81	**43.97**	**58.62**	48.91	**41.14**
Our Team	40.12	**67.85**	19.52	52.18	14.53	52.54	**58.99**	16.66
UIR-MASTER	21.48	38.87	1.94	**66.64**	2.83	38.90	49.10	2.30

Table 3: CFN data distribution

It can be seen that our team's model has shown excellent results in AI task, and the effect in FI task is similar to that of the first team. Model's performance in RI tasks is poor, and the model needs to be further improved and optimized.

5 Other and Future Work

In terms of model optimization, we analyze and discuss the performance of LLM in Chinese Frame Semantic Parsing tasks. Combined with the relevant research of (Chen et al., 2024), we believe that we can try to use BAAI General Embedding (BGE) (Chen et al., 2024) for semantic embedding in subsequent research, and by using the complementary nature of different ranking algorithms, the Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), which has excellent results in previous studies, can get better search results by mixing their ranking results, so as to obtain better data, and then improve the analytical ability of the model.

In addition, we used the Qwen1.5-7B model[‡] in the local validation phase to evaluate the first 1000 data sets in the test (B) set. Using the above-mentioned statement processing method, we have achieved an accuracy rate of 11.2%, and it is estimated that the accuracy rate of more than 45% can be obtained on the whole test (B) set. It can be seen that after reducing the difficulty of the task, LLM with smaller parameters can also perform well. In the future, we can try to use more open source LLM combined with fine-tuning operations in the corresponding fields to achieve better performance.

Finally, due to the limitation of funds and competition time, we only used the gpt-3.5-turbo model which released a year ago for experiments. We believe that the use of more advanced LLM (such as gpt-4o[§]) will bring considerable performance improvement in the future.

6 Conclusion

In this study, we have developed a hierarchical index Retrieval Augmented Generation (RAG) system based on target words to improve the efficiency and accuracy of Chinese Frame Semantic Parsing (CFSP). By leveraging target word information, our approach filters and indexes documents, which significantly enhances the performance of semantic parsing tasks. The key components of our system include data preparation, text tokenization using the HanLP tokenizer, index creation, data insertion into Elasticsearch (ES) indices, cluster configuration, and index population.

Our hierarchical indexing method ensures efficient and scalable retrieval, which is critical for handling large datasets in CFSP. The utilization of the BM25 retrieval algorithm further optimizes the quality of samples in Few-Shot scenarios, ensuring balanced and relevant data for the model. By refining the input and post-processing stages, we reduce the computational load on the model, allowing it to focus on semantic understanding rather than mapping relationships.

[‡]https://huggingface.co/Qwen/Qwen1.5-7B

[§]https://platform.openai.com/docs/models/gpt-4o

The results of our implementation demonstrate a marked improvement in the retrieval and parsing processes, highlighting the effectiveness of our hierarchical index RAG system. This approach not only enhances the performance of existing models but also provides a scalable solution for future CFSP tasks. As we move forward, exploring the integration of more advanced language models and fine-tuning techniques will be essential to further optimize and refine our system.

Our work underscores the importance of efficient indexing and retrieval methods in natural language processing and sets the stage for future advancements in the field of Chinese Frame Semantic Parsing.

References

Ru Li, Yunxiao Zhao, Zhiqiang Wang, Xuefeng Su, Shaoru Guo, Yong Guan, Xiaoqi Han, Hongyan Zhao 2024. *A Comprehensive Overview of CFN From a Commonsense Perspective*. *Mach. Intell. Res*, 21, 239–256 (2024). https://doi.org/10.1007/s11633-023-1450-8.

Charles J. Fillmore. 1982. *Frame semantics[J]. Linguistics in the Morning Calm*, 1982:111-137.

You L, Liu T, Liu K. 2007. *Chinese FrameNet and OWL Representation[C]. Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007)*, IEEE, 2007: 140-145.

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*.

Daniel Gildea and Daniel Jurafsky. 2002. *Automatic labeling of semantic roles*. *Computational linguistics*, 28(3):245–288.

Shaoru Guo, Ru Li*, Hongye Tan, Xiaoli Li, Yong Guan. 2020. *A Frame-based Sentence Representation for Machine Reading Comprehension[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic (ACL)*, 2020: 891-896.

Shaoru Guo, Yong Guan, Ru Li*, Xiaoli Li, Hongye Tan. 2020. *Incorporating Syntax and Frame Semantics in Neural Network for Machine Reading Comprehension[C]. Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020: 2635-2641.

Yong Guan, Shaoru Guo, Ru Li*, Xiaoli Li, and Hu Zhang. 2021. *Integrating Semantic Scenario and Word Relations for Abstractive Sentence Summarization[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021: 2522-2529.

Yong Guan, Shaoru Guo, Ru Li*, Xiaoli Li, and Hongye Tan. 2021. *Frame Semantic-Enhanced Sentence Modeling for Sentence-level Extractive Text Summarization[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021: 404-4052.

Hongyan Zhao, Ru Li*, Xiaoli Li, Hongye Tan. 2021. *CFSRE: Context-aware based on frame-semantics for distantly supervised relation extraction[J]. Knowledge-Based Systems*, 2020, 210: 106480.

Zhiqiang Wang, Ru Li, Jiye Liang, Xuhua Zhang, Juan Wu, Na Su. 2016. *Jiyu hanyu pianzhang kuangjia yuyi fenxi de yuedu lijie wenda yanjiu*. *Jisuanji xuebao*, 39(4):13.

Juncai Li, Zhichao Yan, Xuefeng Su, Boxiang Ma, Peiyuan Yang1, Ru Li. 2023. *Overview of CCL23-Eval Task 1:Chinese FrameNet Semantic Parsing*. *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 113–123, Harbin, China. Chinese Information Processing Society of China.

K. M. Hermann, D. Das, J. Weston, K. Ganchev. 2014. *Semantic frame identification with distributed word representations[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014: 1448-1458.

Shuangtao Yang, Mao Teng, Xiaozheng Dong, Fu Bo 2023. *Llm-based sparql generation with selected schema from large scale knowledge base[C]. China Conference on Knowledge Graph and Semantic Computing*. Singapore: Springer Nature Singapore, 2023: 304-316.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. *Okapi at TREC-3[J]. Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, USA, November 1994.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, Zheng Liu. 2024. *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. *arXiv preprint*, arXiv: 2402.03216.

G. V. Cormack, C. L. A. Clarke, S. Buettcher. 2009. *Reciprocal rank fusion outperforms condorcet and individual rank learning methods[C]*. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009: 758-759.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. 2023. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. *Advances in neural information processing systems*, 35, 24824-24837.

Xinyue Chen, Pengyu Gao, Jiangjiang Song, Xiaoyang Tan. 2024. *HiQA: A Hierarchical Contextual Augmentation RAG for Massive Documents QA*. *arXiv preprint*, arXiv: 2402.01767.

S. W. Kim, J. M. Gil. 2019. *Research paper classification systems based on TF-IDF and LDA schemes*. *Human-centric Computing and Information Sciences*, 9, 1-21.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z. Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021. *A Knowledge-Guided Framework for Frame Identification*. *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.

Yang Peiyuan and juncai li and Zhichao Yan and Xuefeng Su and Ru Li 2024. *Overview of CCL24-Eval Task1 Chinese Frame Semantic Parsing(CFSP) Evaluation Task*. *Submitted to The 23rd China National Conference on Computational Linguistics (Evaluation Workshop)*, 2024, https://openreview.net/forum?id=ObthsPZyah.

CCL 2024

System Report for CCL24-Eval Task 1: Application of Entity Classification Model Based on Different Position Embedding in Chinese Frame Semantic Parsing

Huirong Zhou, Sujie Tian, Junbo Li, Xiao Yuan
School of Statistics, Beijing Normal University
{202322011204,202322011168,202322011137,202322011194}@mail.bnu.edu.cn

Abstract

This paper addresses three subtasks of Chinese Frame Semantic Parsing based on the BERT and RoBERTa pre-trained models: Frame Identification, Argument Identification, and Role Identification. In the Frame Identification task, we utilize the BERT PLM with Rotary Positional Encoding for the semantic frame classification task. For the Argument Identification task, we employ the RoBERTa PLM with T5 position encoding for extraction tasks. In the Role Identification task, we use the RoBERTa PLM with ALiBi position encoding for the classification task. Ultimately, our approach achieved a score of 71.41 in the closed track of the B leaderboard, securing fourth place and validating the effectiveness of our method.

Keywords: Chinese Frame-Semantic Parsing Relative Position Encoding Multi-Target Words

CCL 2024

1 Introduction

The Chinese FrameNet (CFN) is a Chinese lexical semantic knowledge base for computer use, based on Fillmore's Frame Semantics theory and referencing FrameNet from the University of California, Berkeley, with Chinese real corpus as the basis (Hao et al., 2007). Frame semantic parsing involves a deep understanding of the entities or meanings involved in a sentence, which is of great significance in downstream tasks such as text summarization, relation extraction and reading comprehension.

The basic structure of frame semantic parsing is as follows: as shown in Figure 1, in the example sentence "从海拔2800米的仁青岗村到海拔4600多米的詹娘舍哨所" ("From Renqinggang Village at an altitude of 2800 meters to Zhanniangshe Outpost at an altitude of over 4600 meters"), the phrase "从……到……" ("from …… to ……") can activate a "moving path" frame, where "海拔2800米" ("at an altitude of 2800 meters"), "仁青岗村" ("Renqinggang Village"), "海拔4600多米" ("at an altitude of over 4600 meters"), and "詹娘舍哨所" ("Zhanniangshe Outpost") are the four arguments of the example sentence. The attributes of the arguments are respectively "feature, starting point, feature, end point".

Chinese Frame Semantic Parsing is a semantic parsing task based on Chinese Frame Semantic resources. This task consists of the following three subtasks: frame identification, argument identification, and role identification. The frame identification task involves identifying the corresponding frame category from candidate frame categories based on the target word given in the sentence. The argument identification task determines the boundaries of the arguments in the frame based on the target word in the sentence. The role identification task determines the names of the frame elements corresponding to each argument based on the results of argument identification. Therefore, this paper categorizes the frame identification task and role identification task as classification problems, and the argument identification task as an extraction task.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

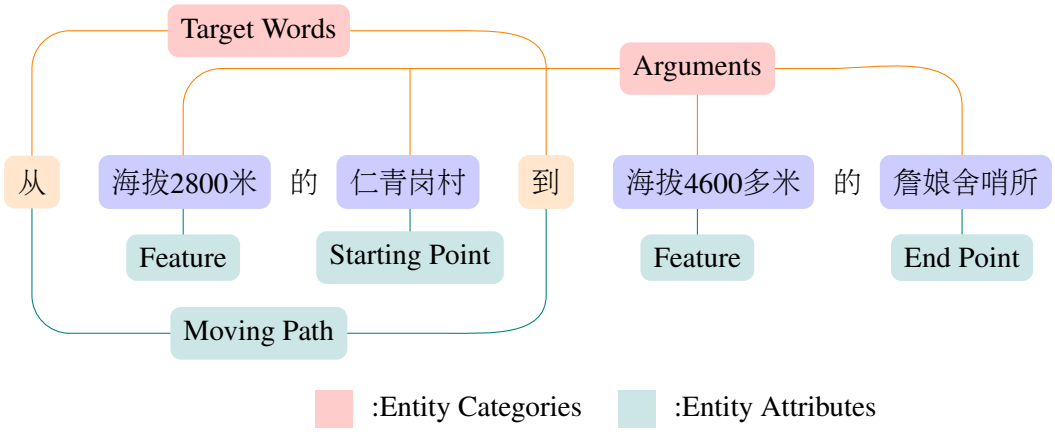


Figure 1: Training Data Examples

2 Related Work

In the task of Chinese Frame Semantic Parsing, traditional machine learning algorithms were initially used, such as maximum entropy (李济洪 et al., 2011), conditional random fields (李济洪, 2010), "OBI data type" annotation, and greedy algorithms (石佼 et al., 2014). However, with the continuous development of deep learning algorithms, Zhao et al. (赵红燕 et al., 2016) used DNN for frame identification, Suhaifeng et al. (Su et al., 2021) utilized frame relationships and definitions for frame identification, Zhou et al. (Zhou et al., 2021) identified frame elements and frames together in end-to-end tasks, and utilized the relationship between target words and frame elements to assist in frame identification during decoding. Wang et al. (王晓晖 et al., 2022) proposed a Chinese Frame Semantic Role Labeling method based on self-attention mechanism to obtain long-distance information from sentences.

The Chinese FrameNet (CFN) contains extensive resources and applications that are significant for natural language understanding tasks such as frame disambiguation and semantic role labeling. CFN mainly focus on common knowledge within contexts(Li et al., 2024), utilizing a combination of top-down, bottom-up, and expert manual curation methods for its creation. (Liu et al., 2023) used different end-to-end frameworks for parsing and employing data augmentation and voting methods to further improve prediction accuracy. (Li et al., 2023) propose an entity classification model based on Rotary Position Embedding (RoPE).(Huang et al., 2023) used a multi-task pipeline strategy and pre-trained language models to address the problem of Chinese Frame Semantic Parsing.

3 Methodology

3.1 Extraction Method for *Span* Type Data

In the task of *Argument Identification*, we need to "extract" entities from sentences. In addition to the token prediction extraction method for "OBI" data, our paper adopts an extraction method for a type of data called "span" data, where predictions are made for the start and end of the arguments. We 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) represents the 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.

As shown in Figure 2, the predicted arguments for the given sentence include "海拔2800米", "仁青岗村", "海拔4600多米",and "詹娘舍哨所". Taking "海拔2800米" as an example, the elements at the starting index (the row corresponding to "海") and the ending index (the column corresponding to "米") in the head-tail matrix should be marked as "1", and the pair of (starting index, ending index) should be treated as an extracted argument span.

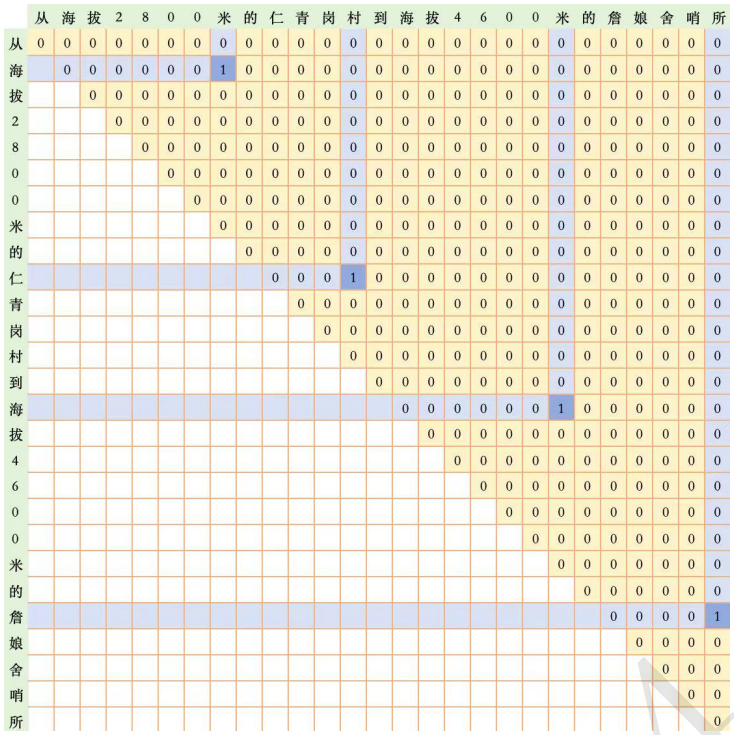


Figure 2: An example of using H-T matrix of "*span*" data

3.2 Method of Frame Identification Based on Multiple Target Words

The main task of *Frame Identification* is to map the semantic structure of a sentence onto a predefined semantic frame. Each frame represents a specific event, action, or situation and defines the roles associated with it, known as *frame elements*. For example, a "purchase" frame may include roles such as buyer, seller, product, and price.

The target word refers to the word or phrase in the sentence that triggers or activates a specific semantic frame. For instance, in the sentence "小明吃了一个苹果" ("Xiao Ming ate an apple"), the target word "吃" ("ate") activates the semantic frame of "eating". In more complex cases, there may be multiple target words in a given sentence. For example, in the sentence "从早上开始，小明一直在学习，直到傍晚才休息" ("Starting from the morning, Xiao Ming has been studying until taking a break in the evening"), the target phrase "从……到……" ("from……to……") activates the semantic frame of "time span". Unlike the activation of semantic frames with single target words, multiple target words often accompany complex grammatical structures, requiring techniques such as dependency parsing to accurately understand the relationships between sentence structures and components. In our paper, we attempt to use the mean or maximum attention scores of individual target words in the sentence as the predicted scores for the case of multiple target words.

3.3 Rotary Position Embedding(RoPE)

Traditional Transformers adopt either learned or sinusoidal absolute position encoding, lacking relative positional relationship information. However, Rotary Positional Encoding (RoPE) (Su et al., 2024) is an improvement upon the additive positional encoding usually applied in Transformers ($\mathbf{X} + \mathbf{P}$), which provides relative positional information between tokens. RoPE further assists the attention model in memorizing directional (preceding or following) information between tokens by employing multiplicative positional encoding ($\mathbf{X} \otimes \mathbf{P}$).

3.4 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 implement and suitable for fixed-length sequences, it performs poorly when handling sequences of different lengths and capturing relative positional relationships. In contrast, relative positional encoding excels in capturing the relative relationships between elements and achieving translational invariance, making it more suitable for variable-length sequences. Recently the use of relative positional embeddings has become more common. Relative positional embeddings do not use fixed embeddings for each position but generate different learned embeddings based on the offsets between the key and $query$ being compared in the self-attention mechanism.

ALiBi *(Attention with Linear Biases)* positional encoding adds a linearly decreasing penalty proportional to the distance to the dot product of key and $query$ in the $Attention$ model.

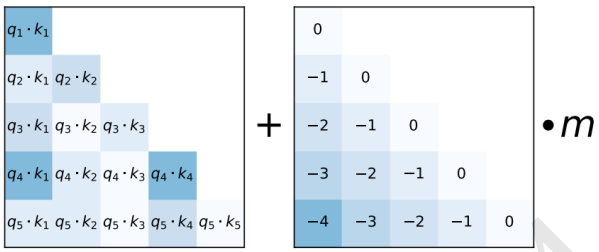


Figure 3: ALiBi

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 $Attention$ (Press et al., 2021).

3.5 T5 Relative Position Encoding

The $T5$ relative position encoding is similar to the $ALiBi$ position encoding, where a penalty term is added to the inner product of the k and q in the corresponding *self-attention score matrix*. The difference is that this penalty term is not the same linearly decreasing penalty as in $ALiBi$. For example, let x_i and x_j represent the tokens at position i and j respectively. According to the original $Transformer$ principle formula (1), the inner product $e_{ij}^{(h)}$ of the k and q in $self-attention$ is computed, and then a penalty term r_{ij} is added according to formula (2) to obtain $\hat{e}_{ij}^{(h)}$, followed by a $softmax$ operation on $\hat{e}_{ij}^{(h)}$, as shown in formula (3). The relative position encoding r_{ij} is a scalar value. The $T5$ (Roberts et al., 2019) uses a partitioning method to map various relative position information into a total of 32 types. For example, the relative position information between nearby tokens is more important and thus needs to be more precise. Conversely, distant relative positions do not need to be as precise, so $T5$ divides these distant positions into different regions, using the same value within the same region, as shown in the mapping relationship in $Table\ 1$.

$$e_{ij}^{(h)} = \frac{\boldsymbol{x_i} W_Q^{(h)} \left(\boldsymbol{x_j} W_K^{(h)}\right)^\top}{\sqrt{d_z/H}}, \tag{1}$$

$$\hat{e}_{ij}^{(h)} = e_{ij}^{(h)} + r_{ij}, \tag{2}$$

$$\alpha_{ij}^{(h)} = \underset{j}{\text{softmax}}\left\{\hat{e}_{ij}^{(h)}\right\}. \tag{3}$$

$i-j$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$r(i-j)$	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9
$i-j$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	$\cdots$
$r(i-j)$	10	10	10	10	10	10	10	11	11	11	11	11	11	11	11	$\cdots$

Table 1: T5 Relative Position Mapping

3.6 Warm-up Strategy

The warm-up learning strategy involves starting with a small learning rate at the beginning of training and then gradually increasing it to the preset learning rate. The purpose of this is to update the model parameters more stably in the early stages of training, preventing model instability or divergence caused by an excessively high initial learning rate. The linear warm-up strategy refers to starting the learning rate from a small value and linearly increasing it to the preset learning rate. The expression for updating the learning rate is as follows:

$$\mathrm{lr}_t = \mathrm{lr}_0 + \frac{(\mathrm{lr}_{\max} - \mathrm{lr}_0) \times t}{T}. \tag{4}$$

In this context, lr_t is the learning rate at the t-th iteration, lr_0 is the initial learning rate (usually set to a very small value), and $\mathrm{lr}_{\max}$ is the preset maximum learning rate (usually the final target learning rate). t is the current iteration number, and T is the total number of warm-up steps. According to the above formula, it is clear that within the first T steps of training, the learning rate increases linearly from the initial value lr_0 to the maximum value $\mathrm{lr}_{\max}$. After exceeding T steps, the learning rate further decreases until convergence.

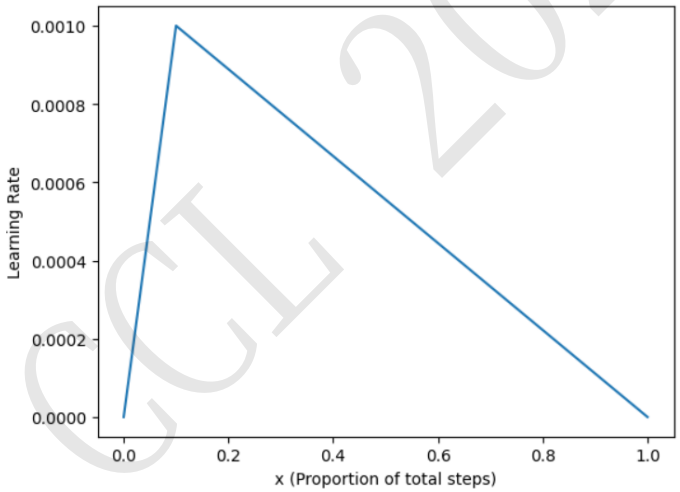


Figure 4: Warm-up Linear Strategy

Figure 4 shows the change in the learning rate during the training process using linear warm-up. In the example shown in the figure, $\mathrm{lr}_{\max} = 0.001$, $\mathrm{lr}_0 = 0$, and T accounts for 0.1 of the total training steps.

4 Model Description

4.1 Task 1: Frame Identification

According to the given sentence $s =< w_1, \cdots, w_i, \cdots, w_j, \cdots, w_n >$ of length n and the context of the target word t, we aim to find the most likely frame f_t to be activated from the frame set $F = \{f_1, f_2, \cdots, f_T\}$, which can be formulated as:

$$f_t = \underset{1 \le j \le T}{\operatorname{argmax}} P\left(f_j \mid s, t\right). \tag{5}$$

CCL 2024

The probability $P(f_j \mid s,t)$ is estimated as follows: In the given sentence s, the target word t has the first and last characters w_{ti} and w_{tj}, respectively. We obtain word vectors $\mathbf{w}_i^f = \mathbf{E}(w_{ti}, s, f)$ and $\mathbf{w}_j^f = \mathbf{E}(w_{tj}, s, f)$ using word embedding methods, and then compute the relative information between the two word vectors $\mathbf{I}\left(\mathbf{w}_i^f, \mathbf{w}_j^f\right)$. For each frame f in the frame set F:

$$P(f \mid s,t) \approx \frac{\exp\left(\mathbf{I}\left(\mathbf{w}_i^f, \mathbf{w}_j^f\right)\right)}{\sum_{\hat{f}\in F} \exp\left(\mathbf{I}\left(\mathbf{w}_i^{\hat{f}}, \mathbf{w}_j^{\hat{f}}\right)\right)}. \tag{6}$$

The architecture of the frame identification task model is illustrated in Figure 5.

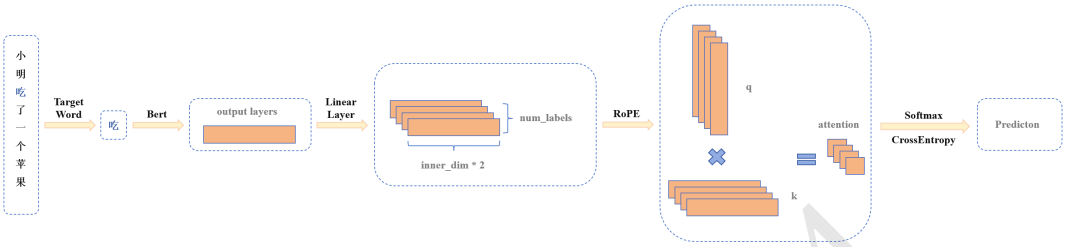


Figure 5: Frame Identification Model

In the model framework shown in Figure 5, for each frame f_t in the frame set F, given the boundaries of the target word w_i, w_j, a corresponding score can be computed as:

$$S_t = S_t\left(w_i^t, w_j^t\right). \tag{7}$$

The activated semantic frame should then be:

$$f = \underset{1\le t\le T}{argmax}\left(softmax\left(\mathbf{S}\left(\mathbf{w}_i, \mathbf{w}_j\right)\right)\right), \tag{8}$$

where:

$$\mathbf{S}\left(\mathbf{w}_i, \mathbf{w}_j\right) = [S_1(w_i^1, w_j^1), S_2(w_i^2, w_j^2), \cdots, S_T(w_i^T, w_j^T)]. \tag{9}$$

The loss function during model training is defined as:

$$CrossEntropyLoss = CrossEntropyLoss(\mathbf{S}\left(\mathbf{w_i}, \mathbf{w_j}\right), real\_frame\_label). \tag{10}$$

4.2 Task 2: Argument Identification

The main purpose of the argument identification task is to determine the positions of each argument (i.e., frame element) involved by each target word in the sentence. That is, given a sentence $s =< w_1, w_2, ..., w_n >$ and a target word t, the task is to find the set of arguments $\hat{T}$ matched with the target word in the set of all consecutive ordered word elements $T = \{t_1, t_2, ..., t_N\}$ within the sentence.

As the training samples provided contain the starting and ending indices of the arguments under the corresponding framework, this problem can be converted into a multi-classification task. Following the approach of $GlobalPointer$ (苏剑林, 2021), this paper adopts the extraction method of predicting the heads and tails of $span$-type data. For a sentence s of length n, there are a total of $n(n+1)/2$ candidate arguments. If the sentence s has k argument indices, then this extraction task is transformed into a multi-label classification problem of selecting k out of $n(n+1)/2$.

For a sentence s of length n, after encoding, it yields a vector sequence $[v_1, v_2, ..., v_n]$. Based on the $Transformer$ transformation, vector sequences $[q_1, q_2, ..., q_n]$ and $[k_1, k_2, ..., k_n]$ are obtained. Then

CCL 2024

$RoPE$ (rotary positional encoding) is introduced, where a transformation matrix R_i satisfies $R_i^T R_j = R_{j-i}$. According to formula (11), rotation is applied to q and k to calculate the attention score $s(i,j)$ between the i-th and j-th word elements, explicitly incorporating relative positional information into the attention score. Meanwhile, for $Task2$, our paper introduces the $T5$ position encoding, adding a penalty term for relative position to the attention score, and obtains $logits$ according to formula (12). Based on the discriminant formula (13), the indices (i,j) where $logits(i,j) = 1$ are obtained, which represent the predicted argument scopes. Finally, our paper employs the loss function used in $GlobalPointer$ as shown in formula (14).

$$s(i,j) = (R_i q_i)^T (R_j k_j) = q_i^T R_i^T R_j k_j = q_i^T R_{j-i} k_j, \tag{11}$$

$$logits(i,j) = s(i,j) - m * T(i,j), \tag{12}$$

$$H\text{-}T_{i,j} = \begin{cases} 1, & \text{if } logits(i,j) \geq 0, \\ 0, & \text{if } logits(i,j) < 0, \end{cases} \tag{13}$$

$$Loss_{AI} = log\left(1 + \sum_{(i,j)\in P} e^{-logits(i,j)}\right) + log\left(1 + \sum_{(i,j)\in Q} e^{-logits(i,j)}\right). \tag{14}$$

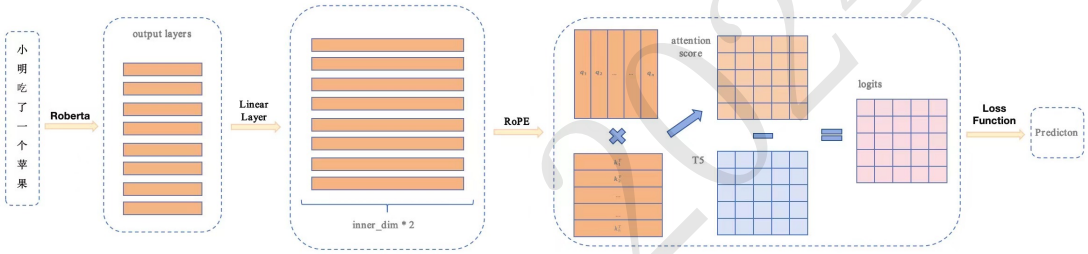


Figure 6: Argument Extraction Model

4.3 Task 3: Role Identification

For the task of the role identification, our paper fully leverages the prediction results from the frame classification module and the argument identification module. Based on the identified argument spans, it further predicts the role classification of each identified argument for a given target word. This task involves finding the corresponding argument c_t from the argument role set $C = \{c_1, c_2, \cdots, c_T\}$, which is essentially similar to $Task1$ as a classification task.

In our paper, we "wrap" the target word with $SpecialToken$ such as $< t >$ and $< /t >$. For different sentences, the number of targets and the relative position between targets and arguments are various, so we have to deal with following scenarios.

The scenarios for when there is one target word are divided into two cases as shown in Figure 7: the target word is either before or after the target argument.

For the case where there are two target words, it is divided into three scenarios as shown in Figure 8: both target words are before the target argument, both target words are after the target argument, and the target argument is between the two target words. At this point, we have implemented the process of incorporating the frame classification module's prediction results as frame features and the argument identification module's prediction results as argument features into the input text. These features are then successfully restored in the final prediction results to obtain the predicted outcome.

The overall model architecture for $Task3$ is shown in Figure 9. Similar to previous tasks, it involves a pre-trained model and a linear layer to produce results through argument frame activation and the RoPE process. However, before applying attention score regularization, the ALiBi method is used to enhance

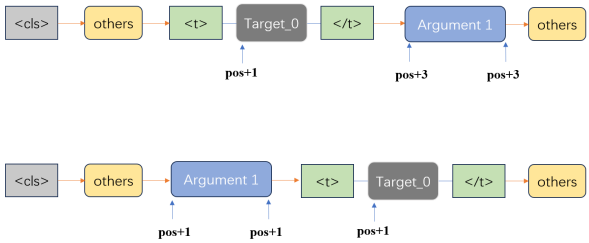


Figure 7: one target word

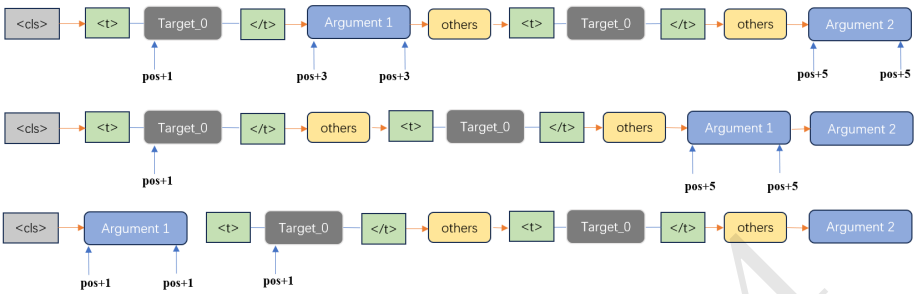


Figure 8: two target words

relative position information. The loss function for $Task3$ has the same form as the loss function in $Task1$, as shown in Equation (10).

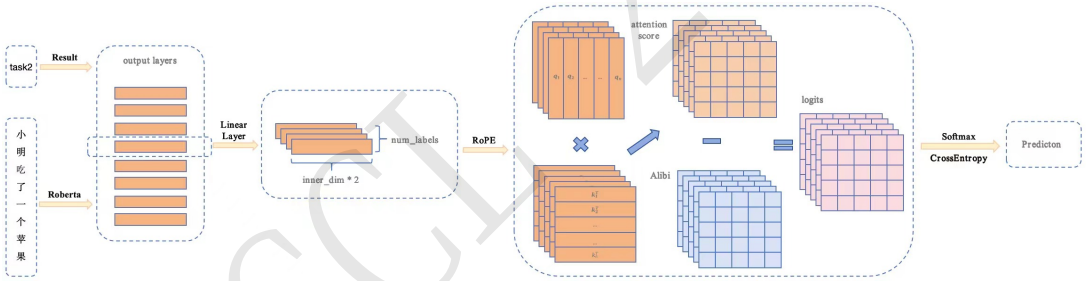


Figure 9: Role Identification Model

5 Experiments

5.1 Selection of Bert Hidden Layers

The BERT model consists of multiple layers of Transformer encoders, where the hidden layers can be divided into different levels, with preceding layers (F,indicating the initial hidden layers) and succeeding layers (L,indicating the final hidden layers) carrying different types of information. Lower-level hidden layers (F) typically capture more local and fundamental lexical-level language features, such as word context and relationships between words, and can be regarded as feature extractors to some extent. As the depth increases, hidden layers (L) gradually capture more abstract and semantically rich information, capable of extracting sentence-level language features.

In the experimental section, we compared the effects of different compositions of BERT output layers (averaging the last four layers $L1 + L2 + L3 + L4$, averaging the first two layers and the last two layers $L1 + L2 + F1 + F2$, and using only the last layer $L1$) on the model's performance on the validation set. The results are shown in Table 2.

CCL 2024

hidden layers	task1 acc	task2 f1	task3 f1
L1	0.732	**0.791**	0.688
L1 + L2 + L3 + L4	**0.738**	0.778	**0.749**
L1 + L2 + F1 + F2	0.721	0.783	0.692

Table 2: Performance for Different Hidden Layer Configurations

It can be observed that, for classification tasks: $task1$ and $task3$, averaging the last four hidden layers as the output layer for BERT yields the best performance. However, for the extraction task: $task2$, directly selecting the last hidden layer as the output layer for BERT achieves the best performance.

5.2 Selection of Linear Warm-up Parameters

By using learning rate warm-up, the initial learning rate during training is kept small, allowing the model to gradually adapt to the training process. The learning rate is then gradually increased to reach a predefined higher value, thereby improving training efficiency and model performance. Through experimentation, we identified the optimal proportion of the linear warm-up phase (in the total training phase) that performs best on the validation set. The results are shown in Table 3.

T proportion	task1 acc	task2 f1	task3 f1
0.02	0.666	0.742	0.701
0.05	0.678	0.788	0.721
0.1	**0.738**	**0.791**	**0.749**
0.12	0.733	0.776	0.747
0.15	0.735	0.751	0.732
0.2	0.729	0.744	0.725

Table 3: Performance for Different Warm-up Percentage

According to the table above, setting the linear warm-up phase in the as first 10% of the total training phase yields the best performance for the model.

5.3 Selection of FGM Adversarial Training Hyper-parameters

The basic idea of FGM is to generate adversarial perturbations by computing gradients of the loss function with respect to the input data based on the original data, and then adding the perturbation to the original data. Typically, the magnitude of the perturbation is controlled by a hyperparameter called ϵ. We found the best ϵ on the validation set through implementation, and the experimental results are shown in Table 4.

ϵ	task1 acc	task2 f1	task3 f1
0.5	0.711	0.758	0.723
0.75	0.733	0.767	0.733
1	**0.738**	**0.791**	**0.749**
1.5	0.698	0.777	0.742
2	0.694	0.759	0.745

Table 4: Performance for Different ϵ

After experimentation, it was found that setting the perturbation magnitude to 1 resulted in the best performance for all three tasks.

5.4 Model Architecture Comparison

For the selection of which type of position encoding to use for each task, this paper conducted ablation experiments, including experiments with the $basemodel$, $basemodel + ALiBi$, $basemodel + ALiBi + Voting$, $basemodel + T5$, and $basemodel + T5 + Voting$.

CCL 2024

First, we compared two base models. We found that BERT performed better on $task1$. In contrast, RoBERTa showed superior performance on $task2$ and $task3$. Additionally, incorporating positional encodings into the RoBERTa model further improved the results.

For $task1$, whether switching to the RoBERTa model or adding various positional encodings, the performance was not as good as the base model. However, for $task2$ and $task3$, simply switching to the RoBERTa model almost always yielded better results than adding various positional encodings to the base model. Therefore, most of our experiments were based on the RoBERTa model for various treatments.By combining RoBERTa with ALiBi, T5, ALiBi+Voting, and T5+Voting, where the first two methods add information from different relative position methods and the last two methods combine the punishment of relative position information with different weights through an ensemble approach, it was found that there was no significant improvement in performance, and even a slight decrease. By further comparing these effects in the T5 model experiments, we had the following conclusions displayed in table 5. It was found that in $task2$, $RoBERTa + T5$ performed the best, with an F1 score reaching a maximum of 0.791. In $task3$, $RoBERTa + ALiBi$ performed the best, reaching 0.749.

Model Architecture	task1 acc	task2 f1	task3 f1
BERT	**0.738**	0.781	0.735
RoBERTa	0.711	0.786	0.738
RoBERTa+ALiBi	0.722	0.785	**0.749**
RoBERTa+ALiBi+Voting	0.715	0.784	0.739
RoBERTa+T5	0.732	**0.791**	0.743
RoBERTa+T5+Voting	0.728	0.787	0.740

Table 5: Performance for Different Model

5.5 Final Model

We determined the best model architectures as well as their hyper-parameters for each sub-task through grid search,the result is shown in the Table 6.

Sub-task	Model Architecture	Hidden Layers	T proportion	ϵ	performance
Frame Identification	BERT	L1+L2+L3+L4	0.1	1	0.738
Argument Identification	RoBERTa + ALiBi	L1	0.1	1	0.791
Role Identification	RoBERTa + T5	L1+L2+L3+L4	0.1	1	0.749

Table 6: Final Model

6 Conclusion

Our article first acknowledges a drawback of traditional Transformer models regarding position encoding, namely the "lack of relative positional information." It proposes addressing this issue by incorporating directional information between tokens into entity classification models through the use of "$T5$" and "$ALiBi$" position encodings. Additionally, it considers scenarios involving multiple target words and enhances the performance of the model in the context of framework identification tasks through aggregation. Through experimentation, comparisons were made regarding various hyperparameters such as the composition of BERT's hidden layers, linear warm-up parameters, adversarial training parameters, and various model architectures. The best combination on the validation set was identified, resulting in achieving the fourth position on the closed-track B leaderboard. Our article also has some shortcomings. For instance, in the framework identification task, we were unable to find downstream model architectures that could effectively leverage RoBERTa. As a result, we had to settle for using BERT only, which led to a performance gap compared to the first-place result. Improving the framework identification model to obtain more accurate and comprehensive framework semantic parsing results is a direction worthy of further research.

CCL 2024

References

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Adam Roberts, Colin Raffel, Katherine Lee, et al. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. Google, Tech. Rep.

Jie Su, Mairin Ahmed, Yutong Lu, et al. 2024. Roformer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Su X, Li R, Li X, et al. 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)*, pp. 5230-5240.

Zhou S, Xia Q, Li Z, et al. 2021. Fast and accurate end-to-end span-based semantic role labeling as word-based graph parsing. arXiv preprint arXiv:2112.02970.

Li R, Zhao Y, Wang Z, et al. 2024. A Comprehensive Overview of CFN From a Commonsense Perspective. *Machine Intelligence Research*, pp. 1-18.

Liu Y, Li Z, Zhang M. 2023. CCL23-Eval 任务 3 系统报告: 苏州大学 CFSP 系统 (System Report for CCL23-Eval Task3: SUDA CFSP System). In: *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pp. 84-93.

Li Z, Guo X, Qiao D, et al. 2023. CCL23-Eval 任务 3 系统报告: 基于旋转式位置编码的实体分类在汉语框架语义解析中的应用 (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)*, pp. 94-104.

Huang S, Shao Q, Li W. 2023. CCL23-Eval 任务 3 系统报告: 基于多任务 pipeline 策略的汉语框架语义解析 (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)*, pp. 105-112.

苏剑林. 2021. 让研究人员绞尽脑汁的Transformer位置编码.

郝晓燕, 刘伟, 李茹, 刘开瑛. 2007. 汉语框架语义知识库及软件描述体系. 中文信息学报, 05: 96-100+138.

李济洪. 2010. 汉语框架语义角色的自动标注技术研究. 博士学位论文. 山西大学.

李济洪, 高亚慧, 王瑞波, 李国臣. 2011. 汉语框架自动识别中的歧义消解. 中文信息学报, 25(03):38-44. 山西大学计算中心, 山西大学数学科学学院, 太原工业学院.

石佼, 李茹, 王智强. 2014. 汉语核心框架语义分析. 中文信息学报, 28(06):48-55.

赵红燕, 李茹, 张晟, 等. 2016. 基于DNN的汉语框架识别研究. 中文信息学报, 30(06):75-83.

王晓晖, 李茹, 王智强, 等. 2022. 基于 Self-Attention 的句法感知汉语框架语义角色标注. 中文信息学报, 36(10):38-44.

System Report for CCL24-Eval Task 1: Leveraging LLMs for Chinese Frame Semantic Parsing

Yahui Liu, Chen Gong, Min Zhang
School of Computer Science and Technology, Soochow University, Suzhou, China
yahuiliu.nlp@foxmail.com
{gongchen18, minzhang}@suda.edu.cn

Abstract

We participate in the open track of the Chinese frame semantic parsing (CFSP) task, i.e., CCL24-Eval Task 1, and our submission ranks first. FSP is an important task in Natural Language Processing, aiming to extract the frame semantic structures from sentences, which can be divided into three subtasks, e.g., Frame Identification (FI), Argument Identification (AI), and Role Identification (RI). In this paper, we use the LLM Gemini 1.0 to evaluate the three subtasks of CFSP, and present the techniques and strategies we employed to enhance subtasks performance. For FI, we leverage mapping and similarity strategies to minimize the candidate frames for each target word, which can reduce the complexity of the LLM in identifying the appropriate frame. For AI and RI subtasks, we utilize the results from small models as auxiliary information and apply data augmentation, self-training, and model ensemble techniques on these small models to further enhance the performance of subtasks.

1 Introduction

Chinese Frame Semantic Parsing (CFSP) is a fine-grained semantic parsing method based on Chinese FrameNet (CFN) (You and Liu, 2005; Li et al., 2023a), which is first proposed during CCL2003-Eval (Li et al., 2023a). It aims to extract frame-semantic information from a sentence (Wang et al., 2020). Previous works have shown that CFSP can help various downstream tasks, 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.

CFN, proposed by Shanxi University, is constructed based on the theory of Frame Semantics, with English FrameNet (Fillmore et al., 2003) as a reference and Chinese realistic corpora as the data foundation (Li et al., 2024). It is a structured representation of knowledge that establishes relationships between vocabulary and concepts through frames. In the CCL2023-Eval CFSP task, part of the CFN data is released for the first time. Compared to the previous evaluation, this evaluation adds information on constructional target words. Constructional target words refer to the words within a construction that are considered central or core. These words play a crucial role in activating the meaning of the entire construction. A construction is a fixed expression unit in a language that has specific form and meaning (Boas, 2021; Willich, 2022). It can be a single word, a phrase, or even a sentence. For example, in the phrase "爱买不买(love to buy or not)", the construction is "爱+V+不+V (love + V + not +V)", where "V" represents a verb. Table 1 provides an example of the frame "量变" (Change position on a scale). The concept expressed by this frame is the relative positional change of a certain attribute of an entity, and its lexemes (also called target word) are "提高(increase)", "从...到... (from...to)" and so on. The word "提高" is a general target word, while "从...到" is a constructional target word. The frame elements are used to capture the semantic information related to the frame in a sentence. Different frame elements are assigned different meanings, for example, "初值(Initial value)" means the starting point of the entity's attribute value change.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 21-31, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Frame Name	量变(Change position on a scale)	
Frame Definition	该框架表示实体在某个维度上（即某属性）的相对位置发生变化，其属性值从初值变至终值 (The framework represents the relative change of an entity along a certain dimension (i.e., attribute), with its attribute value transitioning from an initial value to a final value.)	
Frame Elements	**fe_name**	**fe_def**
	实体(Entity)	在某属性上具有一定量值的事物(Something with a certain quantity value on an attribute)
	属性(Attribute)	实体的有数量变化的属性(The entity's attributes with changing quantities)
	初值（Initial value)	实体的属性值变化的起点(The starting point of the entity's attribute value change)
	终值(Final value)	实体最后达到的量值(The final quantity value reached by the entity)
	初始状态(Initial state)	实体经历属性值变化之前的状态(The state of the entity before undergoing changes in attribute values)
	终状态(Final state)	实体经历属性值的变化之后所达到的状态(The state reached by the entity after undergoing changes in attribute values)
	变幅(Difference)	实体在某维度上变动的幅度(The extent of the entity's variation along a certain dimension)
	值区间(Value range)	属性值的变动范围(The range of variation in attribute values)

Table 1: An example about "量变"(Change position on a scale) frame in Chinese FrameNet.

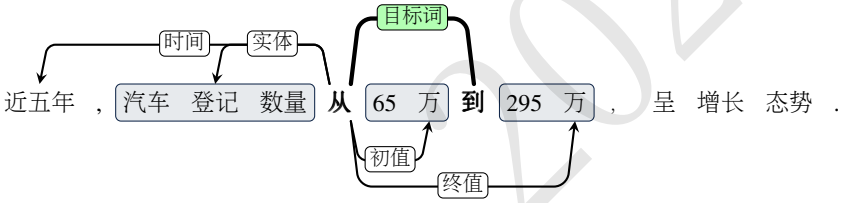


Figure 1: An example of CFSP.

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 constructional target word in the sentence "近五年，汽车登记数量从65万到295万，呈增长态势(Over the past five years, the number of registered cars has increased from 0.65 million to 2.95 million, showing a growth trend)" is "从...到", which corresponds to the frame "量变(Change position on a scale)" in table 1. In this sentence, there are multiple arguments related to this frame, i.e., "近五年", "汽车登记数量", "65万", "295万". "时间", "实体", "初值", and "终值" are their corresponding semantic role labels, respectively.

In this CFSP evaluation task, there are two tracks: open and closed. The open track allows the use of LLM such as ChatGPT for inference. In the open track, we leverage the LLM Gemini (Team et al., 2023) released by Google to evaluate three subtasks of CFSP. For FI subtask, we provide a small number of relevant frames for each target instead of all frames for LLM to select from. For AI and RI subtasks, we enhance the performance of Gemini on CFSP based on the results of frame identification and argument identification from small models, respectively. All of our datasets and codes are available at https://github.com/yahui19960717/CCL2024-CFSP-LLM.git.

2 Related works

CCL2023-Eval (Li et al., 2023a) propose CFSP task based on CFN for the first time. The existing works of CFSP mainly use neural network-based methods (Li et al., 2023b; Huang et al., 2023; Liu et al., 2023; Guan et al., 2023), where Li (2023b) propose a method based on rotational positional encoding (Su et al., 2021) to calculate the attention signal between entities, which has a significant effect on FI subtask. Huang (2023) use a variety of optimization strategies during training to improve the robustness of the model, for example, Exponential Moving Average (EMA) and Fast Gradient Method (Miyato et al., 2016). For FI subtask, Liu (2023) transform CFSP into a word-based graph parsing task, identifying the frame of the target word together with the corresponding arguments in an end-to-end framework. For AI and RI subtasks, they transform CFSP into a tree parsing task to enhance the ability to identify argument boundaries and roles by modeling the internal structure of arguments. Guan (2023) attempt various pre-trained models for different sub-tasks and explore multiple approaches to solving each task from the perspectives of feature engineering, model structure, and other tricks.

In the era of LLM, the performances of LLM on most NLP tasks are still well below the supervised baselines (Sun et al., 2023). Li (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.

For this, we propose different strategies to enhance LLM performance in CFSP task. Given the restriction on the token number in an LLM prompt, for FI subtask, we provide relevant frames in the LLM prompt instead of all frames as candidate frames for each target word. The AI subtask is based on the FI subtask, and RI is a further refinement of the argument. Considering that FI, AI and RI are three interrelated subtasks and the performance of the small model is better than LLM in these subtasks, we utilize the results of small models to parse AI and RI subtasks.

3 Methods

Prompting is a recently-mainstream for LLM and LLM-friendly evaluation method (Zhou et al., 2023). The performance of a task is highly related to the information contained in the prompt. For this, we design appropriate prompts for the three subtasks.

For FI subtask, we need to provide candidate frames for each target word in the prompt. To mitigate the issue of excessively long prompts, we obtain a small number of candidate frames for each word based on the mapping relationships between target words and frames in the given training and dev data. After providing the target word and its triggered frame in the prompt, we can parse the AI task using LLM. Similarly, after providing the target word and its corresponding arguments in the sentence in the prompt, we can parse the RI task. Due to the relatively low performance of the LLM in the AI and FI subtasks, we use small models to obtain FI and AI results, which are then used to assist AI and RI subtasks, respectively. To enhance the performance of the small models, we employ data augmentation techniques (including copying and generating data) for the FI task and self-training methods for the AI task. In both tasks, we adopt model fusion techniques (i.e., multiple model voting) to further improve the model performance.

3.1 Frame Identification

The FI subtask is to identify the frames triggered by the given target word in the sentence from a total of 715 provided frames. If we treat all frames as candidates in the prompt for each target word, the performance of FI subtask will be low (Gemini 1.0 obtains an accuracy of only 15.5% on FI for testA). We suspect that this is probably because it exceeds the maximum number of tokens limit imposed by Gemini in a prompt. The maximum number of tokens allowed in a prompt for Gemini 1.0 is 2048. However, just the length of the Chinese name of all the frames connected by commas is 2940, which is too long for a single prompt.

Stratigies of FI subtask. For this, we reduce the number of candidate frames for each target word. Specifically, we identify all corresponding frames for each target word through the mapping between words and frames in the dataset (there are a total of 4,093 target words with corresponding candidate frames in the testB). For target words without an existing mapping (507 target words in the testB), we compute the cosine similarity between the word vectors of the target words and each frame, selecting the six most similar frames as candidates. We chose six because the average number of frames per target word in the dataset is six. Some frame names are long and not tokenized, such as the frame "事件发生时间变化(Change event time)". In such cases, we use Jieba for tokenization. The word vectors we use are from FastText [0], but some target words or tokenized frame names might not be in FastText. In these instances, we sum the word vectors of each character from the tokenization to obtain the final word vector representation.

FI prompt. After obtaining the candidate frames for each target word, we designed appropriate task formats to accommodate the generative characteristics of LLM. Figure 2 shows an example of a prompt we provided. In the prompt, we present the LLM with a sample, and the model selects the most suitable frame based on the sample, the input text, the target word, and the candidate frames for the target word.

给定句子和其目标词,根据目标词在句子中的语义，从框架集合中选择最符合目标词触发的框架。请以字符串的形式输出。
示例:
输入:
text: "双方还就各自的责任、权利以及资金的运作程序等达成了协议。"
target:"达成"
框架集合："['观点一致', '成就']"
输出:
"观点一致"
给定输入:
text:"很快，他完成了从一个山村娃到一名志愿军战士的转变"
target:"从...到"
框架集合:"['事件时量发生变化', '时量场景', '量变', '经历变化', '移动路径', '位移']"
请输出其合适的目标词。

Figure 2: An example prompt for the FI subtask.

3.2 Argument Identification

The AI subtask refers to identifying the arguments relevant to specific semantic frames triggered by the target word in a sentence. Therefore, it is necessary to first obtain the right frames for target words. However, the FI performance of Gemini is relatively poor, and errors in FI propagate to the AI subtask. We utilize the FI results of the small model as the basis for the AI subtask.

Stratigies of AI subtask. For the small model, we adopted the models proposed by Li (2023b) (A sequence labeling model based on rotation position encoding) and Liu (2023) (A word-based graph parsing model). In the training set, we observe a severe imbalance in the distribution of frames, as illustrated in Figure 3. 76 frames had no corresponding examples, while 14 frames had over 100 corresponding examples. To improve the performance of the small model, we utilize the LLM to generate examples for frames without any examples. For frames with fewer than 15 examples, we augment the existing examples by duplicating them to reach a total of 15 examples. Figure 4 illustrates the prompt we used for sentence generation with the large model. Specifically, we train the models using the original training

[0]https://fasttext.cc/docs/en/crawlvectors.html

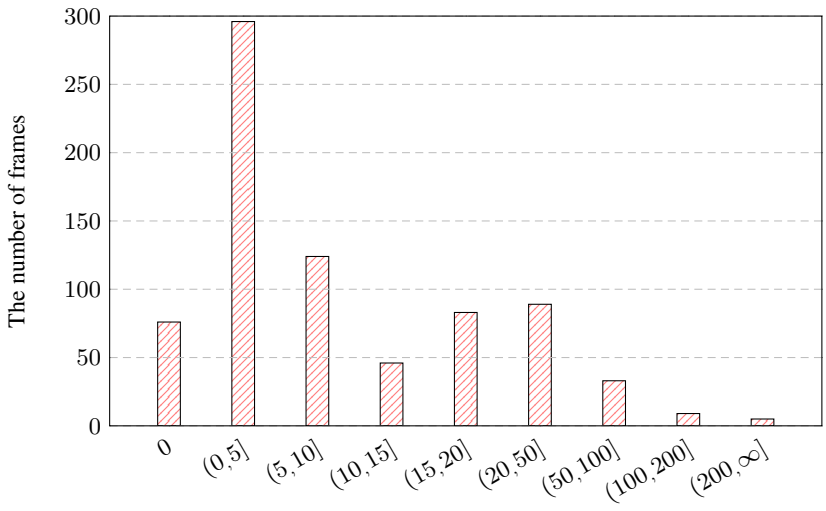


Figure 3: The distribution of the number of example sentences corresponding to each frame.

data	train	dev	testB	frames
CFN	10,700	2,300	4,600	715

Table 2: Data statistic of CFN.

data along with the constructed data. Each model is trained three times, and the final results are obtained through voting.

AI prompt. Finally, the performance of the FI task on the small model is 72%. After obtaining the FI results, we design an appropriate prompt to use Gemini to parse the AI task. Figure 5 provides an example of our prompt.

3.3 Role Identification

Stratigies of RI subtask. The RI subtask aims to assign semantic labels to the arguments identified in the sentence that are related to the frame. Therefore, we first need to identify all the arguments in the sentence. We use the small model to obtain more accurate AI results for RI subtask. Like the AI subtask, we also use the models proposed by Li (2023b) and Liu (2023). To enhance the performance of the small model, we employ the self-training and ensemble techniques. Specifically, we first use a trained model to predict the data from CoNLL09 and CPB1.0. Then, we 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, and the performance of the AI task on the small model is 84%.

RI prompt. After obtaining the AI results, we design an appropriate prompt to use Gemini for parsing the RI task. Figure 6 provides an example of our prompt.

4 Experiments

4.1 Data and Settings.

Table 2 shows the data statistic of CFN. For the AI subtask, we generated 1,509 of data using Gemini 1.0 and constructed 4,688 of data using the copying method. For the RI subtask, we provided a total of 124,270 of data from CTB1.0 and CoNLL09 for self-training.

For the small model, we use the model settings of Li (2023b) and Liu (2023). For the FI subtask, after voting by 6 models, we selected the results with votes greater than or equal to 2 as the FI results. Similarly, we consider results with 2 or more votes as the final AI results. As for the RI subtask, we find that results with 3 or more votes are more reliable.

使用所提供的框架和框架信息，生成与之相关的20个句子，并标识出框架中的目标词(谓词），以json格式输出。
框架信息如下：
框架名称:盗窃
框架定义:以非法占有为目的，秘密窃取数额较大的公私财物或者多次盗窃公私财物的行为。
示例：
输入：
框架名称：到达
框架定义：指转移体朝目的地方向的移动。目的地可直接表达出来，或从上下文中得到理解，动词本身隐含目标之义。
输出2个句子：
{'text': ['运动会', '闭幕', '后', '，', '他们', '将', '在', '北京', '继续', '逗留', '两', '天', '，', '同', '中国', '有关', '方面', '开展', '交流', '活动', '并', '参观', '游览', '，', '于', '１３日', '返回', '日本', '。'], 'target': ['返回'], 'text': ['迎接', '澳门', '回归', '系列', '图书', '出版'], 'target': ['回归']}

Figure 4: An example prompt for the AI subtask.

Organizer	Score
SUDA	48.77
DVTC	40.12
UIR-ISC	21.48

Table 3: Reproduction results of the top three teams.

The LLM settings are configured as follows for all subtasks. The primary model utilized is Gemini 1.0, specified by the parameter model = "gemini-pro". For the generation process, the following configuration is applied: "temperature" is set to 0, "top_p" to 0.95, and "top_k" to 0. Other settings remained at their default values.

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:

$$\text{ACC} = \frac{\text{the right number of frames}}{\text{Total number of frames}} \tag{1}$$

$$P = \frac{(gold \cap pred)}{Count(pred)} \tag{2}$$

$$R = \frac{Count(gold \cap pred)}{Count(gold)} \tag{3}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{4}$$

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.

CCL 2024

作为汉语框架语义学家，你的任务是抽取给定句子中与目标词相关的所有框架论元成分，并根据框架的论元定义为它们分配相应的论元标签。请注意，论元成分是指句子中与目标词直接相关的部分，而论元标签描述了这些成分的具体角色。
输出格式：每个输出元素应该是一个列表，包含两个字符串元素：
一个是论元成分,一个是论元标签，即[[论元成分1，论元标签1],[论元成分2，论元标签2],...]。
只需要以Python数组的形式输出，不要添加其他内容。
文本：长寿区生态环境局局长刘刚介绍，该区乡镇污水收集处理率由原来的不足30%提升到现在的80%。
目标词：“由...到”
框架名称:量变
论元标签:
{‘实体’, ’属性’, ’初值’, ’终值’, ’初状态’, ’变幅’, ’值区间’, ’环境条件’, ’倚变因素’, ’时量’, ’倚变起点’, ’倚变终点’, ’方式’, ’路径’, ’处所’, ’速度’, ’时间’, ’终状态’, ’角色事件时间’, ’动作时量’, ’结果时量’, ’频率’, ’次数’, ’特定次数’, ’角色位置’, ’亚区’, ’接受者’, ’受益人’, ’受损者’, ’伴随实体’, ’身份’, ’形容’, ’事件评价’, ’背景事件’, ’并行事件’, ’相关变量’, ’原因’, ’目的’, ’工具’, ’材料’, ’根据’, ’方法’, ’物量’, ’角色范围’, ’程度’, ’媒介’, ’主题’, ’视角’, ’结果’, ’领域’, ’现象’, ’解释’}
示例：
输入：
文本：[’运动会’, ’闭幕’, ’后’, ’，’, ’他们’, ’将’, ’在’, ’北京’, ’继续’, ’逗留’, ’两’, ’天’, ’，’, ’同’, ’中国’, ’有关’, ’方面’, ’开展’, ’交流’, ’活动’, ’并’, ’参观’, ’游览’, ’，’, ’于’, ’１３日’, ’返回’, ’日本’, ’。’]
目标词：[’返回’]
框架名称:到达
论元标签:
{’终点’, ’转移体’, ’伴随体’, ’形容’, ’目的地状态’, ’方式’, ’方法’, ’移动模式’, ’路径’, ’起点’, ’时间’, ’方向’, ’角色事件时间’, ’动作时量’, ’结果时量’, ’频率’, ’次数’, ’特定次数’, ’处所’, ’角色位置’, ’亚区’, ’接受者’, ’受益人’, ’受损者’, ’伴随实体’, ’身份’, ’事件评价’, ’环境条件’, ’背景事件’, ’并行事件’, ’相关变量’, ’原因’, ’目的’, ’工具’, ’材料’, ’根据’, ’物量’, ’角色范围’, ’程度’, ’媒介’, ’主题’, ’视角’, ’结果’, ’领域’, ’现象’, ’解释’}
输出:
“[[’他们’, ’转移体’], [’于１３日’, ’时间’], [’日本’, ’终点’]]”

Figure 5: An example prompt for the AI subtask.

CCL 2024

作为汉语框架语义学家，你的任务是根据给定的句子的目标词、框架以及目标词在句子中的相关论元成分，从候选的论元标签中为它们分配相应的论元标签。请注意，论元成分是指句子中与目标词直接相关的部分，而论元标签描述了这些成分的具体角色。
输出格式：每个输出元素应该是一个列表，包含两个字符串元素：
一个是论元成分,一个是论元标签，即[[论元成分1，论元标签1],[论元成分2，论元标签2],...]。
只需要以Python数组的形式输出，不要添加其他内容，如果没有相应的标签就返回”。
文本：长寿区生态环境局局长刘刚介绍，该区乡镇污水收集处理率由原来的不足30%提升到现在的80%。
目标词：”由...到”
框架名称:量变
论元标签：
{’实体’, ’属性’, ’初值’, ’终值’, ’初状态’, ’变幅’, ’值区间’, ’环境条件’, ’倚变因素’, ’时量’, ’倚变起点’, ’倚变终点’, ’方式’, ’路径’, ’处所’, ’速度’, ’时间’, ’终状态’, ’角色事件时间’, ’动作时量’, ’结果时量’, ’频率’, ’次数’, ’特定次数’, ’角色位置’, ’亚区’, ’接受者’, ’受益人’, ’受损者’, ’伴随实体’, ’身份’, ’形容’, ’事件评价’, ’背景事件’, ’并行事件’, ’相关变量’, ’原因’, ’目的’, ’工具’, ’材料’, ’根据’, ’方法’, ’物量’, ’角色范围’, ’程度’, ’媒介’, ’主题’, ’视角’, ’结果’, ’领域’, ’现象’, ’解释’} 论元成分:
{’现在的80%’}
示例：
输入：
文本：[’运动会’, ’闭幕’, ’后’, ’，’, ’他们’, ’将’, ’在’, ’北京’, ’继续’, ’逗留’, ’两’, ’天’, ’，’, ’同’, ’中国’, ’有关’, ’方面’, ’开展’, ’交流’, ’活动’, ’并’, ’参观’, ’游览’, ’，’, ’于’, ’１３日’, ’返回’, ’日本’, ’。’]
目标词：[’返回’]
框架名称:到达
论元标签：
{’终点’, ’转移体’, ’伴随体’, ’形容’, ’目的地状态’, ’方式’, ’方法’, ’移动模式’, ’路径’, ’起点’, ’时间’, ’方向’, ’角色事件时间’, ’动作时量’, ’结果时量’, ’频率’, ’次数’, ’特定次数’, ’处所’, ’角色位置’, ’亚区’, ’接受者’, ’受益人’, ’受损者’, ’伴随实体’, ’身份’, ’事件评价’, ’环境条件’, ’背景事件’, ’并行事件’, ’相关变量’, ’原因’, ’目的’, ’工具’, ’材料’, ’根据’, ’物量’, ’角色范围’, ’程度’, ’媒介’, ’主题’, ’视角’, ’结果’, ’领域’, ’现象’, ’解释’}
论元成分：
{’他们’, ’北京’, ’于13日’, ’日本’} 输出:
“[[’他们’, ’转移体’], [’北京’,”], [’于１３日’, ’时间’], [’日本’, ’终点’]]”

Figure 6: An example prompt for the RI subtask.

	FI		AI		RI			总分
	ACC	P	R	F1	P	R	F1	
SUDA (Gemini-w/ small model)	**58.62**	44.84	**53.91**	48.96	**44.09**	**38.75**	**41.24**	**48.77**
SUDA (Gemini-w/o small model)	58.62	**72.39**	40.76	**52.15**	21.49	14.05	16.99	40.03
baseline (ChatGPT)	53.00	60.98	22.52	32.90	6.38	7.59	6.93	28.54

Table 4: Comparison of results between ours and ChatGPT on the three sub-tasks of CFSP.

4.3 Results and Analysis.

Table 3 shows the final reproducibility scores of the top three teams on the testB dataset in the open track. It can be seen that we ranked first, with a lead of 8.77 points over the second place and 27.29 points over the third place.

Previous studies have also evaluated CFSP on LLM in CCL2003-Eval (Li et al., 2023a). They extracted a portion of samples from the test set, constructed different prompt information, and used ChatGPT (gpt-3.5-turbo-16k) to complete the corresponding subtasks. For the FI subtask, they tested the results of ChatGPT in ZeroShot and FewShot scenarios. For the AI and RI subtasks, they guided ChatGPT through multiple rounds of dialogue using a chain-of-thought prompting method to generate more reliable results. As shown in Figure 4, we compared our results with theirs, where their FI results are from the few-shot scenario. It can be seen that our score is 20% higher than theirs, proving the effectiveness of our techniques. However, we can observe that our improvement in the FI task is relatively small, possibly due to the incompleteness of the frame found for each target word using the mapping method. In addition, we find that the improvement in the RI sub-task is significant, indicating that the accuracy of argument identification is crucial for the RI task.

In addition, for AI and RI subtasks, we conducted experiments on Gemini without relying on small models, with results as shown in the major second row. We can find that the performance of RI is lower compared to when small models are used, which is as expected. This is because small models perform well in FI and AI tasks, where FI can provide labels for RI to use, and RI also needs to utilize the results from AI during evaluation. However, it is interesting that in the AI task, the F1 score is higher than when relying on the small model. Intuitively, using small models seems preferable because predicate argument identification occurs after determining frames. We speculate that the reason for the good performance in the AI subtask might be that AI has a weak dependence on results of FI and FI information in the prompt could potentially interfere with AI task generation. Additionally, we can observe that the AI results without using small models excel primarily in accuracy, reaching 72.39%.

5 Conclusion

In this evaluation task, we construct appropriate prompts in Gemini for the three subtasks of CFSP for parsing. For the FI task, we use the mapping between the target words and frames to obtain candidate frames for each target word. For the AI and RI subtasks, we leverage the results from the small model to enhance the prediction performance of the LLM. Ultimately, we achieve first place in the open track of the CFSP.

However, our work still has some notable limitations. For instance, in the FI subtask, the limited candidate frames we provided may constrain the choices of the LLM. On the closed track, the small model has achieved 72.82%, 86.97%, and 60.27% respectively for the three subtasks, indicating ample room for improvement for the large model in CFSP task. In the future, we can enhance the performance of the three subtasks by providing more examples or utilizing information provided by the framework, such as frame definitions and argument definitions. The output of the large model is diverse, and we can further enhance its performance by generating multiple prediction results and using voting mechanisms.

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), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No.23KJB520034), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Hans C Boas. 2021. Construction grammar and frame semantics. In *The Routledge handbook of cognitive linguistics*, pages 43–77.

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. *Advances in neural information processing systems*, 33:1877–1901.

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 medel 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 chiness 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*, 21:239–256.

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 Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

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.

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*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

Alexander Willich. 2022. Introducing construction semantics (cxs): A frame-semantic extension of construction grammar and constructicography. *Linguistics Vanguard*, 8:139–149.

Liping You and Kaiying Liu. 2005. Building Chinese Framenet database. In *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*, 210:106480.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? An evaluation by asking natural language questions. *arXiv preprint arXiv:2311.08287*.

CCL 2024

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

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

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.

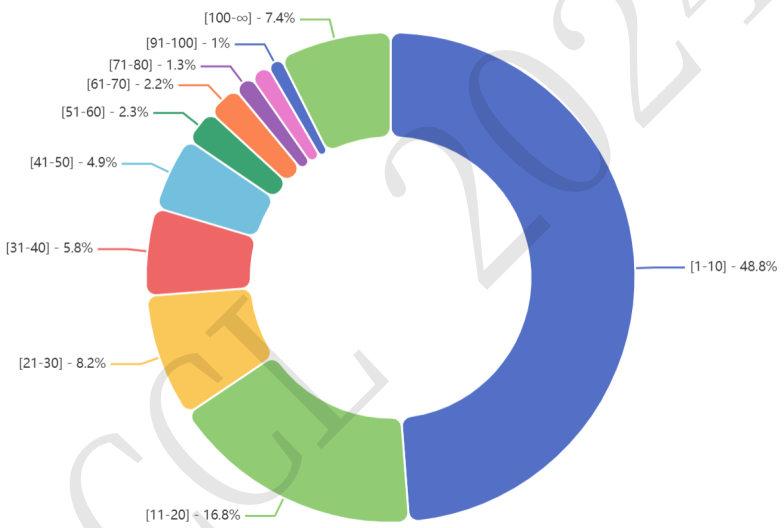


Figure 2: Sentence Range and Proportion in Frame

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

CCL 2024

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 $Attention$.

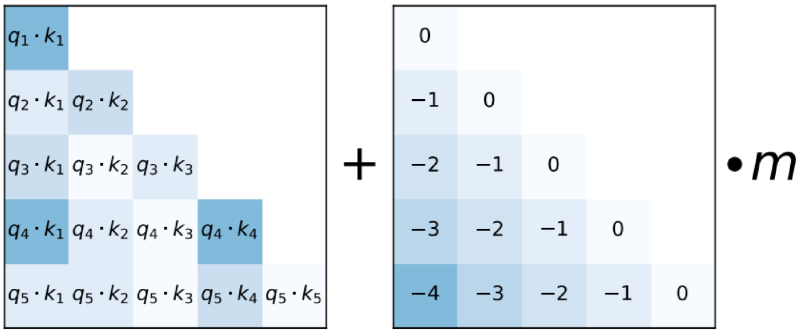


Figure 3: Attention with Linear Biases(ALiBi)

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.

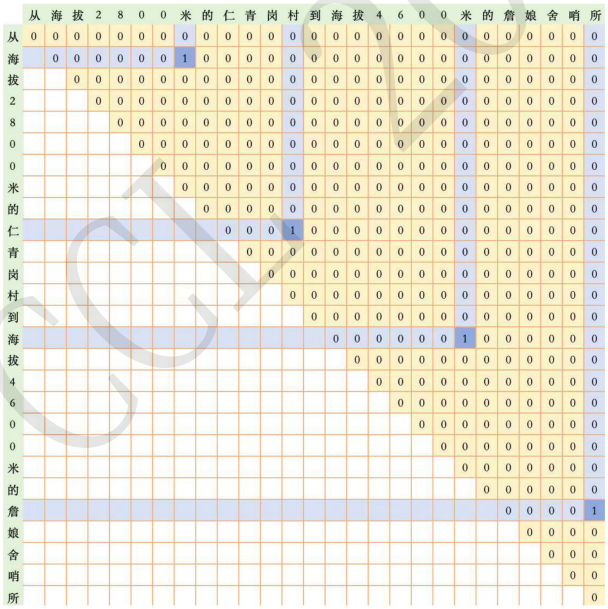


Figure 4: An example of using H-T matrix of “span” data

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

CCL 2024

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.

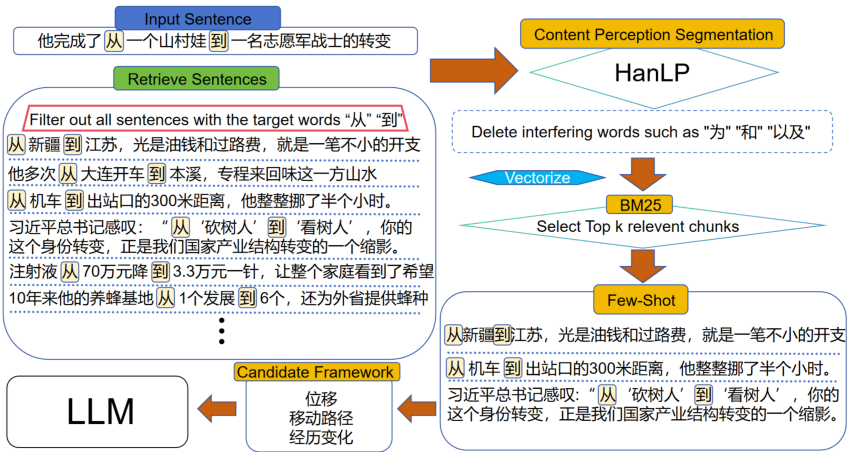


Figure 5: Hierarchical indexing RAG system based on target words

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.

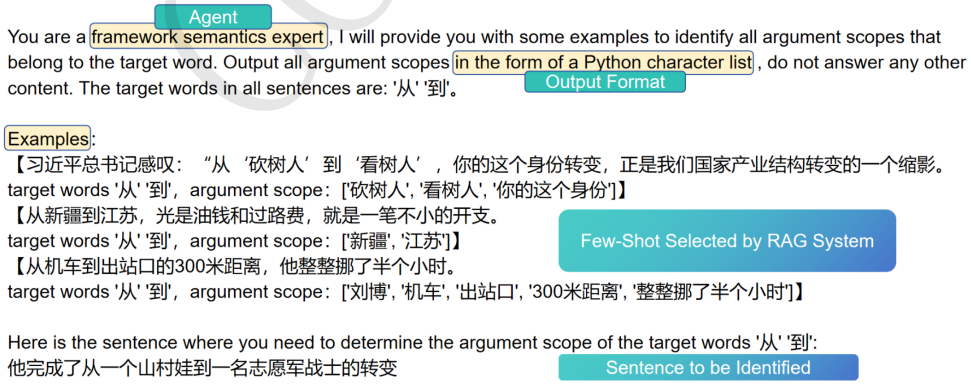


Figure 6: Sample semantic enhancement conversion process example

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.

CCL 2024

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.

CCL 2024

CCL24-Eval任务2系统报告：基于多个大语言模型微调的中文意合图语义解析

李让
北京理工大学计算机科学与技术学院
lirang@bit.edu.cn

摘要

中文意合图对句中成分间的关系进行层次化标注，能有效表示汉语的深层语义结构。传统方法难以对中文意合图中的特殊成分进行特征表示，而近期大语言模型性能的快速提高为复杂自然语言处理任务提供了一种全新思路。在本次任务中，我们尝试使用Prompt-Response方式对大模型进行LoRA微调，让大模型根据输入直接生成格式化的中文意合图三元组序列。我们广泛测试来自不同研发团队、拥有不同参数规模的七个主流大模型，评估基座模型、参数规模、量化训练等因素对微调后模型性能的影响。实验表明，我们的方法展现出远超依存模型的性能，在测试集和盲测集上的F1分别为0.6956和0.7206，获得了本次评测榜一的成绩。

关键词： 中文意合图；语义解析；大模型微调；模型对比

System Report for CCL24-Eval Task 2: Chinese Parataxis Graph Parsing Based on Fine-Tuning Multiple Large Language Models

Rang Li
School of Computer Science and Technology, Beijing Institute of Technology
lirang@bit.edu.cn

Abstract

Chinese Parataxis Graph (CPG) annotates the relationships between sentence components hierarchically, which can effectively represent the deep semantic structure of Chinese language. Traditional methods struggle to extract features from the special components in CPG, while the rapid improvement in the performance of large language models has provided a novel approach for complex natural language processing tasks like this. In this task, we attempt to fine-tune large language models using LoRA method with Prompt-Response, allowing the models to directly generate formatted triple sequences of CPG based on our input. We extensively tested seven popular large language models from different teams with varying parameter sizes, evaluating the influence of base, size, and quantization on the performance of fine-tuned models. Experiments indicate that our method outperformed existing models by a significant margin, with F1 scores of 0.6956 and 0.7206 on the test set and blind test set respectively, achieving the top position in this evaluation.

Keywords: Chinese Parataxis Graph , Semantic Parsing , Large Language Model Fine-Tuning , Model Comparison

第二十三届中国计算语言学大会论文集，第43页-第50页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

语义解析是自然语言处理（NLP）中的一个核心任务，旨在将自然语言转换成一种高度格式化的逻辑结构，可供进一步执行多种下游任务使用。这一过程涉及到识别词汇的语义角色、解析句中的实体关系以及理解语境中的隐含意义等。

由于语法结构的灵活性，中文的语义解析相比英文等语法规则相对严格的语言更加困难。以省略现象为例，中文表达经常涉及到句中某些成分的省略，其中大部分情况下是论元的省略，如谓词所对应的主体或客体。对于针对中文的语义解析方法而言，识别和处理此类特殊现象对于正确理解句子意义十分重要。

本次评测任务的中文意合图（Chinese Parataxis Graph）以事件为中心构建单根有向图，图中的节点对应承载事件、实体、属性的单元，有向边表示单元间的语义关系。中文意合图脱离了传统方法中句法形式的限制，将事件结构相对独立于句法结构进行表示，能较好地表示中文的特殊语义结构 (Guo et al., 2024)。由于中文意合图是一个较新的概念，领域内相关研究较少，本次评测中给出的依存模型也仅取得31.03%的F1分数。因此，我们决定从问题转化出发，分析使用传统方法和大模型微调的可行性。

2 问题转化

在本次任务中，我们需要以三元组的形式表示中文意合图中两个节点和它们之间的关系，这极大地依赖于句中每个词的上下文语义。因此，使用传统方法的一个直接的思路是先使用BERT (Devlin et al., 2018)获得句中每个词结合上下文语义的编码表示，接着对句中任意两个词的组合进行多分类，判断它们之间的关系。然而，考虑到中文意合图结构的特殊性，以下问题不能得到很好的处理。根据体系标签的定义，中文意合图的节点并不只包含句中的词，还可能包含表示句子结构的“ROOT”、表示逻辑关系的“因果关系”以及表示省略的“ls”等特殊成分 (Guo et al., 2024)，它们难以使用BERT进行编码表示，但在多分类时又和句中词语处于同等地位。由于每个句子中存在这些关系的种类和数量的不确定性，简单地添加特殊token对它们进行表示也不能取得很好的效果。

考虑到以上分析，我们决定将目光转向当今发展迅速的大语言模型（LLM）。近年来，大模型性能的快速提高为众多NLP任务的实现提供了一种全新的方式，即构建合适的Prompt进行输入，引导大模型给出问题的答案。对于更为复杂的任务，我们可以采用对大模型进行微调的方式，使用训练集构造合适的Prompt-Response集对大模型进行微调训练，显著提高大模型在解决特定任务方面的能力。在以上分析中，我们已经看到，由于中文意合图体系结构定义的复杂性，传统模型的实现会变得格外繁琐，且种种限制下也很难取得令人满意的分类性能。同时，若只是采用Zero-shot或Few-shot的方式让大模型完成该任务，Prompt中的描述甚至不足以将中文意合图符号的完整定义表达清楚。综合以上原因，我们决定将本次任务转化为大模型微调。

具体而言，我们首先对数据集进行格式化处理构造Prompt和Response，以便大模型更好地理解任务内容。接着，我们广泛采用目前中文性能较好的各种开源与闭源大模型进行微调并进行性能对比，其中开源大模型包括通义千问发布的Qwen-1.5系列 (Bai et al., 2023)中7B、14B、32B、72B四个参数规模的模型以及百川智能发布的Baichuan2-7B (Yang et al., 2023)和零一万物发布的Yi-1.5-9B (Young et al., 2024)，闭源大模型包括百度文心系列的ERNIE4.0-Speed-8K。最后，我们将测试集的Prompt输入大模型，得到相应的推理结果并进行后处理，丢弃无用的脏数据并转化为符合要求的格式。

3 数据预处理

本次任务是对于每一个给定的已分词句子，输出其相应的中文意合图三元组表示。原始数据集中每一个句子及其中包含的所有三元组被以字典的形式给出，若直接输入大模型显然太过繁琐。因此，我们要采用合适的策略在原始数据集之上构造Prompt和Response，以尽可能高的效率和大模型传递信息。同时，考虑到原始训练集为2000条，验证集为1000条，我们对其进行重新配比，取3000条数据中的5%作为验证集，其余均作为训练集。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

3.1 Prompt构造

在原始数据集中，句子作为字典中的一个键值对，键为固定字符串“sent”，值为一个列表，列表中的元素为按顺序排列的分词结果。由于我们是使用Prompt-Response对大模型进行微调，输入的Prompt只需要为分词后的句子即可。因此，我们丢弃原始数据集中的冗余信息，只保留分词信息，以字符‘/’按照原始的分词信息分隔句子，并且句子的末尾也以字符‘/’标记。表1为一个Prompt构造的示例。

原始句子	“sent”: [“你”, “还是”, “自己”,“观察”, “吧”, “。”]
Prompt构造	你/还是/自己/观察/吧/。/

Table 1: Prompt构造示例

3.2 Response构造

在原始数据集中，句子中的每一个三元组关系都被编码成一个字典加入关系列表中，包含词语内容和索引编号的信息。由于采用字典结构，原始的关系编码中存在大量的冗余字符，如重复出现的“word1”“relData”等键名。因此，我们将原始数据集中的关系字典按照三元组的格式构造成Response，省略了大量重复的内容，只结构清晰地保留需要传递的词语内容、索引编号、关系名。构造的三元组中，前两个元素是(词，位置) 的二元组，第三个元素是关系名的字符串。同一句子的多个关系三元组之间不使用任何分隔符。表2为一个Response构造的示例。

原始关系字典	“relData”: [{ “word1”: {“word”: “你”, “idx”: 0},“word2”: { “word”: “Ref”,“idx”: -4},“relVal”: “Entity”},{ “word1”: { “word”: “还是”, “idx”: 1 }, “word2”: {“word”: “观察”,“idx”: 3},“relVal”: “Mod”}]
Response构造	(('你', 0), ('Ref', -4), 'Entity')(('还是', 1), ('观察', 3), 'Mod')

Table 2: Response构造示例

4 模型微调

4.1 微调方式选择

在大模型微调领域，目前常用的方案主要有两种，即全参数微调和LoRA (Hu et al., 2021)微调。全参数微调对整个模型的参数进行更新，需要耗费大量的显存，而LoRA方式利用低秩矩阵的性质，能够在保证性能的前提下大大减少所需更新的参数量，显著降低了对显存的需求。本次任务我们决定采取LoRA方式进行微调。

4.2 损失函数

由于我们将中文意合图三元组的构建任务转化为了大语言模型的序列生成任务，微调过程中的优化目标是最小化负对数似然损失函数。具体到本任务而言，优化目标的形式化描述如下。

设训练集$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$，其中$x^{(i)}$ 是第i个包含分词信息的句子，$y^{(i)}$ 是该句所包含的中文意合图三元组按前文格式拼接成的序列，即

$$y^{(i)} = y_1^{(i)} \oplus y_2^{(i)} \oplus \cdots \oplus y_{t_i}^{(i)} \tag{1}$$

其中，$y_j^{(i)}$ 为第i 个句子中包含的第j 个格式化的中文意合图三元组，t_i 表示第i 个句子中包含的三元组数量。

我们需要通过优化模型参数来最小化生成序列的负对数似然损失，即

$$\min_{\theta} -\sum_{i=1}^{N} \log P(y^{(i)} \mid x^{(i)}; \theta) \tag{2}$$

(c) 2024 中国中文信息学会计算语言学专业委员会

其中，$P(y^{(i)} \mid x^{(i)}; \theta)$表示模型在给定已分词句子$x^{(i)}$的情况下预测出该句的中文意合图三元组序列$y^{(i)}$的概率，该概率由模型参数$\theta$确定。

4.3 环境和参数

对于开源模型微调，我们的硬件平台采用8块4090显卡阵列，根据模型大小分配合适的显卡数量；软件平台采用CentOS7.9操作系统，安装LLaMA-Factory (Zheng et al., 2024)，使用DeepSpeed (Rasley et al., 2020) ZeRO-3平均分配显存。针对参数规模较小的模型，如Qwen-1.5-7B、Baichuan2-7B和Yi-1.5-9B，我们使用两块4090显卡进行LoRA微调；中等参数规模的Qwen-1.5-14B模型使用四块4090显卡进行LoRA微调；大参数规模的Qwen-1.5-32B模型使用八块4090显卡进行LoRA微调。对于超大参数规模的Qwen-1.5-72B模型，八块4090显卡阵列已经不能满足LoRA微调的显存需要，因此我们采取FSDP (Zhao et al., 2023)结合QLoRA (Dettmers et al., 2024)的方式，基于4bit量化进行微调。对于闭源的文心系列ERNIE4.0-Speed-8K模型，我们使用百度智能云的千帆大模型平台创建微调任务。

在超参数的选择上，开源模型根据不同基座模型的官方文档并综合考虑参数规模进行配置。对于ERNIE4.0-Speed-8K闭源模型，我们也采用LoRA方法进行训练，超参数如下表所示。

超参数	数值	说明
迭代轮次	5	过小可能欠拟合，过大可能过拟合，根据训练集大小调整
梯度累计步数	0	累加多次梯度一次性进行更新，取0代表让千帆平台自动计算
学习率	0.0003	过高或过低都会影响微调收敛，使用平台默认值
LoRA所有线性层	True	启用后可以提高模型表达能力，但计算量也会增加
LoRA 策略中的秩	8	过高增加计算复杂度，过低可能限制模型性能
学习率调整计划	linear	训练时学习率的变化方式，千帆平台默认为线性
序列长度	4096	每个输入序列的最大长度，根据数据集输入长度调整
随机种子	42	用于结果复现的随机种子
预热比例	0.1	训练开始时学习率预热阶段所占比例
正则化系数	0.01	用于防止过拟合，但过大可能导致性能下降

Table 3: ERNIE4.0-Speed-8K超参数

5 数据后处理

虽然我们使用标准格式的Prompt-Response对大模型进行训练，但考虑到大模型生成内容具有一定的随机性，大模型生成的内容中可能存在部分不符合格式要求的脏数据。因此，我们在将大模型生成的内容转化为最终预测结果的过程中，需要对相关数据进行后处理操作，具体包括以下措施。需要说明的是，以下特殊情况出现的概率较小，大模型的推理结果总体上是符合格式要求的。

- **忽略无效行**：某些输入句子的内容可能会触发预训练大模型中的判定机制，得到类似“作为一个人工智能语言模型，我还没学习如何回答这个问题”的回复。对于这种情况，我们的后处理程序需要忽略该行内容并返回空的关系三元组列表。
- **确保索引值为数字**：大模型可能在本该是数字的索引值位置返回其他字符，这会导致提供的F1计算代码抛出错误。对于此类三元组，我们认为大模型没有预测出准确的索引值，将问题处索引值置为0。
- **删去格式错误的三元组**：大模型返回的三元组还可能出现一系列的格式错误，例如字符串缺少引号导致出现SyntaxError，缺少元素导致出现IndexError等等不可预知的情况。因此，我们在逐个处理三元组时使用try/except语句，对于一切属于Exception类的错误均跳过。

在完成上述特殊情况的处理后，我们得到了严格符合前文所述Response编码的数据，再转化为题目所需标准的字典格式，并按照所给分词结果加入原始语句，最后写入json文件即可。

6 实验结果

6.1 实验数据

我们首先测试采用普通LoRA方式进行微调的五个开源模型和一个闭源模型，经过量化的Qwen-1.5-72B模型在6.4中进行评估。下表为各开源模型使用LoRA方式的资源消耗和训练推理速度的有关数据。表中'-'表示微调工具不支持反馈该模型的该项数据。

模型	4090显卡数	每秒训练样本数	每秒推理样本数
Baichuan2-7B	2	1.647	-
Qwen-1.5-7B	2	1.452	2.739
Qwen-1.5-14B	4	0.881	1.648
Qwen-1.5-32B	8	0.985	1.937
Yi-1.5-9B	2	1.163	2.334

Table 4: 训练和推理速度报告

我们用测试集1000条语句对以上模型进行评测，结果如下表所示。其中，我们选择表现最好的ERNIE4.0-Speed-8K以及在开源模型中表现最好的Yi-1.5-9B绘制训练损失变化曲线。

模型	Precision	Recall	F1
Baichuan2-7B	0.4739	0.4861	0.4800
Qwen-1.5-7B	0.5294	0.5425	0.5359
Qwen-1.5-14B	0.5083	0.5174	0.5128
Qwen-1.5-32B	0.5226	0.5328	0.5276
Yi-1.5-9B	0.6003	0.6039	0.6021
ERNIE-Speed	0.6778	0.7142	0.6956

Table 5: 各模型的性能指标

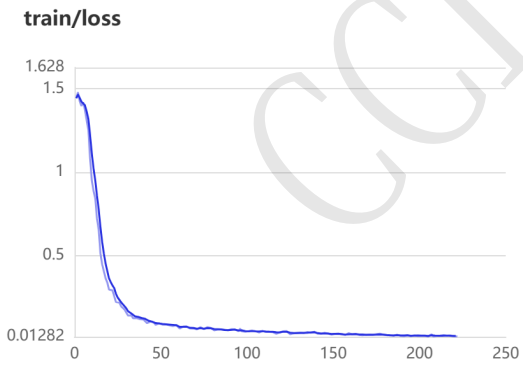


Figure 1: ERNIE-SPEED损失变化

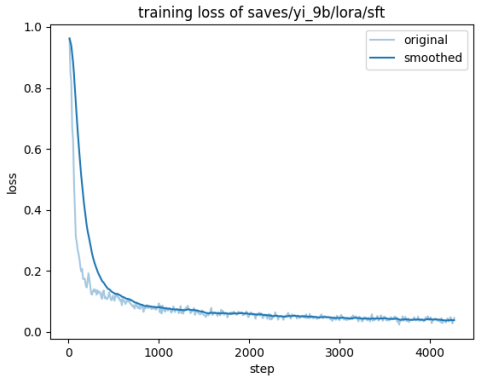


Figure 2: Yi-1.5-9B损失变化

6.2 性能分析

通过分析以上模型的F1分数和损失变化，我们初步得到如下结论。

- **大模型表现普遍优异：**排除受量化影响的72B模型，即使表现最差的Baichuan2-7B模型也达到了0.48的F1分数，表现最好的文心模型更是达到了接近0.7的F1分数，远高于依存模型的表现。在过去的一年中各类大模型的性能提升迅速，在诸多任务上经过微调性能可以达

(c) 2024 中国中文信息学会计算语言学专业委员会

到或超过传统模型，而在一年前大模型的表现通常不及传统模型。使用大模型完成NLP任务逐渐成为一种可靠的选择。

- **文心闭源模型性能突出**：经过我们的测试，一众开源模型微调后的F1分数都大致处于0.4-0.6的区间；而文心闭源模型在测试集的F1分数接近0.7，提交评测后在盲测集的F1分数超过了0.72，与开源模型之间拉开了明显的差距。对比训练集最终损失，开源模型中最优的Yi-1.5-7B模型只能降低到0.077，而文心模型降低到了0.013。

由于ERNIE4.0-Speed-8K闭源模型的详细参数和技术细节均未公布，为了进一步探究各种因素对于模型性能影响的程度，我们采取控制变量的思想对开源模型的实验数据进行对照评估，分别研究参数规模和基座模型系列两个因素的影响。

6.3 对照评估

6.3.1 参数规模

首先，我们探究参数规模对模型性能的影响。为此，我们选择同属Qwen-1.5系列的7B、14B、32B三种参数规模模型进行对比。这三个模型属于同一研发团队的同一系列，能有效排除无关因素的影响。超参数选择上，根据模型大小略有不同，但均保证足够的迭代轮数，并观察到训练结束时损失曲线趋于平稳。下表为各模型在测试集1000条语句上的评估结果。

模型	Qwen-1.5-7B	Qwen-1.5-14B	Qwen-1.5-32B
Precision	0.5294	0.5083	0.5226
Recall	0.5425	0.5174	0.5328
F1	0.5359	0.5128	0.5276

Table 6: 同系列基座不同参数规模对比

我们发现，参数规模对模型性能的影响较小。虽然14B模型和32B模型的参数量分别为7B模型的两倍和四至五倍，但在本任务上的表现却完全处于同一水平，并不符合参数越多性能越好的直观认知。

6.3.2 基座系列

接着，我们探究基座模型系列对模型性能的影响。为此，我们选择Baichuan2-7B、Qwen-1.5-7B和Yi-1.5-9B三个开源模型，它们拥有大致相等的参数规模。同时，我们采用完全相同的超参数和工具进行微调，最大程度上排除了其他因素的影响。其中，迭代轮数均设置为6，训练损失曲线后期均趋于平稳。下表为各模型在测试集1000条语句上的评估结果。

模型	Baichuan2-7B	Qwen-1.5-7B	Yi-1.5-9B
Precision	0.4739	0.5294	0.6003
Recall	0.4861	0.5425	0.6039
F1	0.4800	0.5359	0.6021

Table 7: 不同系列基座相近参数规模对比

我们看到，三个参数规模相近的开源模型采用完全相同的超参数，但F1分数却相差很大，其中最低的Baichuan2-7B模型F1分数仅有0.48，最高的Yi-1.5-9B模型F1分数超过了0.60，这充分说明了基座模型对最终性能的影响。由于来自不同研发团队的基座模型的技术细节和预训练语料均有所不同，模型性能和所擅长领域均存在差异。

综合以上分析，我们认为文心系列ERNIE4.0-Speed-8K模型在本任务上的出色表现主要并不来源于参数规模，而是可能与百度的预训练语料、模型的技术细节以及千帆平台的超参数组合等因素更为相关。

(c) 2024 中国中文信息学会计算语言学专业委员会

6.4 量化损失与错误分析

如前文所述，由于Qwen-1.5-72B模型参数规模很大，八块4090显卡已经不能满足LoRA微调的显存需要，因此我们采取FSDP (Zhao et al., 2023)结合QLoRA (Dettmers et al., 2024)的方式，基于4bit量化进行微调。这种方式使用4bit取代原有浮点数精度对模型的权重进行量化，可以有效降低显存需求，但也会带来模型性能的损失。经过初步测试，4bit量化后的Qwen-1.5-72B模型在测试集前250条语句上的F1值仅为0.2173，远低于上文中所有未经过量化的模型。

为了对模型表现和量化损失有更具体的了解，我们对模型的预测结果进行错误分析。考虑到中文意合图的体系标签定义中除了显式出现的词语，还存在“ROOT”“因果关系”和“ls”等隐式事件词 (Guo et al., 2024)，我们将关系集合分为“只包含显式事件词”和“包含隐式事件词”两类，分别评估模型在这两类关系上的预测表现。下表为Qwen-1.5系列和ERNIE-Speed在测试集前250条语句上错误分析的结果。表中*标识说明微调和推理过程对模型进行了量化处理。

模型	显式F1	隐式F1	总F1
Qwen-1.5-7B	0.5853	0.4575	0.5482
Qwen-1.5-14B	0.5870	0.4232	0.5390
Qwen-1.5-32B	0.5797	0.4466	0.5400
Qwen-1.5-72B*	0.2142	0.2266	0.2173
ERNIE-Speed	0.7663	0.5929	0.7137

Table 8: 部分模型的错误分析

- **隐式错误率更高：** 对于未量化的模型而言，在“只包含显式事件词”的关系上F1均明显高于“包含隐式事件词”的关系，说明在隐式事件词关系上错误率更高。考虑到隐式事件词关系的自由度更高、预测难度更大，该结果符合一般预期。
- **量化损失严重：** 对于经过量化的Qwen-1.5-72B，其显式和隐式两类关系的F1均显著低于同系列未经过量化的模型，说明在本次实验中量化损失是全面的。与Qwen-1.5-32B相比，两类F1下降幅度分别为0.3655和0.2200，在“只包含显式事件词”的关系上性能损失更加严重。

7 结语

在本次任务中，我们分析了使用传统方法完成中文意合图语义解析所遇到的困难，并且阐述了选择大模型微调的原因。通过对数据集和生成内容进行有效预处理和后处理操作，我们测试了七个主流大模型微调后的性能表现，并最终选择文心系列ERNIE4.0-SPEED-8K模型的预测结果进行评测。该模型在测试集和盲测集上分别取得0.6956和0.7206的F1分数，获得本次评测榜一的成绩。我们工作的主要贡献有：

- 将大语言模型微调的方法引入中文意合图语义解析，取得远超依存模型的性能，获得本次评测榜一的成绩，证明了该方法的有效性。
- 广泛测试多个主流大模型，控制变量地评估参数规模和基座系列对模型性能的影响程度；讨论量化损失对模型性能的影响，并进行简要的错误分析。这些工作对后续优化该任务的大模型方案具有一定的方向性指导意义。

下一步，我们将继续探究全参数微调、更多的超参数设置、不同的Prompt-Response构造方式等因素对模型性能的影响。

致谢

本工作受北京理工大学计算机学院辛欣老师所开《知识工程》课程的启发，感谢辛欣老师对相关工作的指导与支持。感谢CCL评测组委会和北京语言大学任务组织者荀恩东老师、饶高琦老师、唐共波老师以及任务联系人郭梦溪、李梦的支持。

(c) 2024 中国中文信息学会计算语言学专业委员会

参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. 2023. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

郭梦溪, 荀恩东, 李梦, 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

郭梦溪, 李梦, 荀恩东, 饶高琦, 于钟洋. 2024. 基于意合图语义理论的结构标注体系与资源建设. 第二十三届中国计算语言学大会.

CCL 2024

System Report for CCL24-Eval Task 2: Chinese Parataxis Graph(CPG) Parsing Based on Large Language Models

YueYi Sun
Beijing Institute of Technology
No.5 Yard, Zhongguancun South Street,
Haidian District, Beijing
1120223534@bit.edu.cn

Yuxuan Wang
Beijing Institute of Technology
No.5 Yard, Zhongguancun South Street,
Haidian District, Beijing
wangyuxuanbilly@bit.edu.cn

Abstract

This paper presents the work submitted for the 23rd China National Conference on Computational Linguistics(Evaluation Workshop)(CCL24-Eval), focusing on the Chinese Parataxis Graph (CPG) Parsing task. CPG represents Chinese natural language hierarchically through relational triplets, providing a consistent representation for linguistic units of varying levels. Our approach has used large-scale language models through full fine-tuning, achieving the result with F1 value at 71.6% in the contest and 74.76% after the contest. Furtehrmore, our team has proposed a combined model that integrates multiple LoRA fine-tuned medium-scale models after the contest. This approach is able to minimize the time and space consumption while keeping the performance of CPG construction task relatively high.

1 Introduction

In recent years, the development of LLM fine-tuning technology has progressed rapidly. LLMs can be fine-tuned to adapt to new tasks and be customized for specific needs, making them easily applicable to downstream vertical domains. Fine-tuning techniques can fully utilize the knowledge in pre-trained models, achieve excellent performance with less training data, significantly improve training efficiency, and avoid retraining the model for each task. This also focuses the model more on the target task, reduces model development costs, and improves model application flexibility and scalability. Fine-tuning for specific tasks can effectively enhance the model's accuracy and generalization ability on those tasks.

Previous research (**?**; **?**) has introduced the definition and rules of CPG. Meanwhile, earlier studies (kommineni2024human, 1972; **?**; **?**; **?**) have experimentally and theoretically demonstrated the completion of corresponding tasks through fine-tuning single large model. Our goal is to find a method to construct CPG through fine-tuning LLMs， which is a combination of the works that are listed. However, there are quite a few significant limitations of previous research. Firstly, the adjustment of large model parameters is insufficient. Secondly, previous tasks are completed only through fine-tuning large models, leading to high space occupancy and time requirements. In our work, we take the task of CPG construction as an example to introduce the results of achieving high performance through fine-tuning large models, reaching an F1 score of 74.76%. Additionally, it proposes a method that achieves similar results through the interaction and integration learning of middle-scale models, significantly reducing

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 51-61, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

time and space consumption during the task process.

According to the Scaling Law (**?**), as the number of model parameters and the size of data increases, the model's performance improves significantly. However, large-scale models require substantial computational resources and storage space. In contrast, middle-scale models require significantly fewer computational resources and storage space and usually have a much faster training speed than large models. They can iterate and adjust quickly, having lower deployment costs, and can more easily adapt to different tasks and environments, achieving good performance in multiple application scenarios with minimal adjustments and fine-tuning.

2 Preliminaries

In this task, we use the Wenxin Qianfan's large-scale model platform of Baidu AI Cloud and the ERNIE Speed closed-source model for fine-tuning. The model version is ERNIE-Speed-8K, released on February 5, 2024. It is a high-performance industry-level knowledge-enhanced large language model developed by Baidu.

This model has the following advantages for this task:

Chinese Context: ERNIE Speed is specially optimized and trained for the Chinese context, allowing it to better understand and process Chinese text.

Context Handling: ERNIE Speed has a high context length processing capability in inference scenarios, enabling it to better handle contextual dependencies in Chinese text.

Lightweight Design: For this task, we introduced the ERNIE-Speed-8K version, which is a lightweight large language model with strong performance and high processing speed.

Efficient Training: Due to its lightweight design, ERNIE Speed has a relatively short training time, improving model iteration efficiency.

During the task, we also tried fine-tuning other large language models provided by the Wenxin Qianfan's large-scale model platform, such as Baidu's lightweight large language model ERNIE Lite and the Qianfan-Chinese-Llama-2-13B-v1 large model, which is based on Llama-2, with an expanded Chinese vocabulary and enhanced pre-training and instruction fine-tuning using large-scale Chinese-English data. Although these models also achieved good performance, the final results were not as good as the ERNIE Speed model. We believe this is mainly due to the parameter scale, so we ultimately chose the ERNIE Speed model.

3 Single-LLM Task

In this section, we mainly present the methods we implemented in the CCL24-Task2 competition.

The core process of fine-tuning large models includes prompt engineering and parameter adjustment. This section will focus on demonstrating the methods of prompt engineering and parameter adjustment.

3.1 Prompt Engineering

During prompt engineering, we tested various prompts, such as removing word segmentation and using the entire paragraph as a prompt. However, we found that this caused significant redundancy in

CCL 2024

the model output and mismatched results with the original word segmentation, making post-processing difficult. Therefore, we explicitly specified word segmentation information. On the other hand, we found that overly detailed and complex prompts may caused the model output to ”repeat” itself, such as including numbering in the prompt and adding prompts about the numbers. Hence, we ultimately adopted a concise prompt.

In our model, the data used to train the large language model consists of a prompt and a response. During the training of the large model, the prompt is the input, and the response is the expected output. In this task, we define the prompt as the sentence being studied and the response as the relationship set in the intention diagram, specifically defined as follows:

Prompt: Composed of the original sentence in segmented form, prefixed with ”segmentation as,” as shown in the example below:

"prompt": "分词为[中国,的,南方,水乡,, ,一旦,进入,“,梅雨,季节,”,, ,阴雨天,往往,会,持续,数,日,。]"

Response：Presented in the form of triplets, showing two words and their relationship, as shown in the example below:

"response": "(中 国,南 方,EntityRel),(南 方,水 乡,EntityRel),(水乡,进 入,A0),(一 旦,进 入,Conj),(进 入,ROOT,CoreWord),(进 入,条 件 关 系,条 件事 件),(梅 雨,季 节,EntityRel),(季 节,进 入,状 态 终 点),(阴 雨 天,持 续,A0),(往往,持 续,Time),(会,持 续,Mod),(持 续,ROOT,CoreWord),(持 续,条 件 关 系,推 论 事件),(数,日,EntityRel),(日,持续,Time)"

Overall：


```
[
    {
        "prompt": "分词为[中国,的,南方,水乡,, ,一旦,进入,“,梅雨,季节,”,, ,阴雨天,往往,会,持续,数,日,。]",
        "response": "(中 国,南 方,EntityRel),(南 方,水乡,EntityRel),(水 乡,进 入,A0),(一 旦,进 入,Conj),(进 入,ROOT,CoreWord),(进入,条件关系,条件事件),(梅雨,季节,EntityRel),(季节,进入,状态终点),(阴雨天,持续,A0),(往 往,持 续,Time),(会,持 续,Mod),(持 续,ROOT,CoreWord),(持 续,条 件 关系,推论事件),(数,日,EntityRel),(日,持续,Time)"
    }
]
```


3.2 Data Preprocessing

We clean the data by removing words corresponding to incomplete or missing examples. Then, we convert each dataset into the prompt format.

3.3 Data post-processing

The output of the large language model is constructed as follows, we first display an example:


```
[
    {
        "index":1,
        "inference_dataset_id":"ds-gf1cr64223rfekth",
        "prompt":"分词为[培养,一条,优秀,的,警犬,是,极,不,容易,的,事,,
,我们,的,警犬,当然,首先,要,为,中国,军队,服务,。,”]",
        "response":"",
        "completion":"(培养,是,A0),(一条,警犬,EntityRel),(优秀,警
犬,EntityRel),(警 犬,培 养,A1),(是,ROOT,CoreWord),(极,容 易,Mod),(不,容
易,Mod),(容 易,事,EntityRel),(事,是,A1),(我 们,警 犬,EntityRel),(警 犬,服
务,A0),(当然,服务,Mod),(首先,服务,Time),(要,服务,Mod),(为,军队,PN),(中国,军
队,EntityRel),(军队,服务,A2),(服务,ROOT,CoreWord)"
    }
]
```


We extract the useful parts, including the prompt section and the completion section, and then we convert them into the final required output format.

3.4 Model Tunning

Considering that with sufficient computing power, full fine-tuning yields the best results among various LLM fine-tuning strategies, but it also requires the most computational resources and storage space. With the cloud computing services provided by the Wenxin Qianfan's large-scale model platform, we have ample computing resources for convenient fine-tuning. Therefore, we ultimately choose the full parameter fine-tuning strategy.

4 Multi-LLM Ensembling and Tuning

In this section, we introduced some smaller-scale LLMs into the task. When using these smaller-scale LLMs separately, they perform not as well as large-scale models. Our approach can be mainly described as two parts: Multi-Model Reinforcement Circum-block and IU-Block. Through these two tools, we came up with a similar result compared with using large-scale LLM.

4.1 Multi-Model Reinforcement Circum-block

We integrate large models with smaller parameter scales through a serial approach, using this method to train the model. We designed a training model named reinforcement circum-block(RC-Block) and the model is illustrated in Figure 1. The main structure of the model consists of two different large language models with small parameter scales(no more than 7B). Our training process is carried out in the following steps:

CCL 2024

We first introduce the notations in this section. Each notation s or t denotes a subset of the test set. The superscripts of each notation represent the step number from which the set is generated, and the subscripts of each notation represents the LLM from which the set is generated.

Step 1: First, we input the dataset into LLM1(i.e. Qwen1.5-7B). The data that generates correct outputs after training is passed to LLM2(i.e. ChatGLM3-6B))for further training, with the results (including prompt and request) denoted as s_2^1. The data that does not generate correct outputs after training is fed back into LLM1 for retraining, with the results denoted as t_1^1. LLM2 is then trained with s_2^1. The results that generate correct outputs after training are recorded as z_2^1, and those that generate incorrect outputs are recorded as t_2^1.

Step 2: z_2^1 is used to train LLM1, generating the correct result set s_2^2 and the incorrect result set t_1^2. Then s_1^2 is used to train LLM2, generating the correct result set z_2^2 and the incorrect result set t_2^2.

Subsequent steps follow this pattern, and the entire model structure is shown in the figure below.

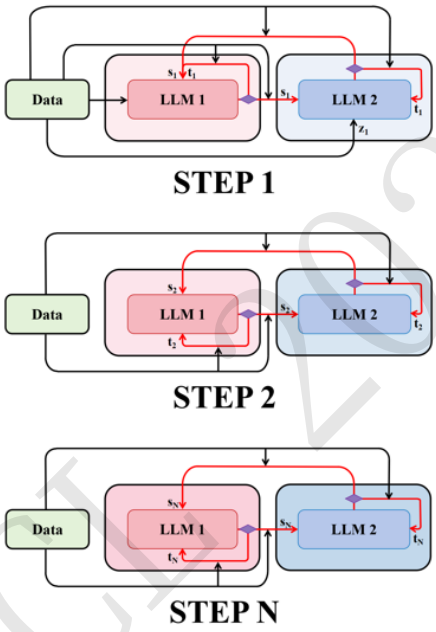


Figure 1: Training Process

4.2 Running and IU-Block

In this section, we mainly show how we design a Result Intersection and Union Block(IU-Block) model. We trained multiple different models and fused the results of these models by taking the intersection ($\cap$) and union ($\cup$).

The purpose of taking the intersection of the output results of two models is that we believe the results predicted to be the same by both models have a higher probability of being correct. At the same time, it excludes redundant incorrect predictions, reducing the total number of predictions and improving the recall rate of the model. The purpose of then taking the union of these intersections is to include as many correct outputs as possible, thereby improving the precision rate of the model.

Our IU-Block model is as follows: First we set the answer generated by the first model(i.e. Qwen1.5-7B) and the second model are A and B, then we may set $A \cap B$ or $A \cup B$ the ultimate answer.

However, the experimental results indicate that none of them are good enough.

This is not difficult to explain, as $A \cap B$ can lead to the model being too "conservative", meaning it tends to make fewer decisions to ensure the accuracy of the predictions made, resulting in higher recall rates. However, it can also lead to lower accuracy due to fewer correct predictions. On the other hand, $A \cup B$ can lead to the model being too "reckless", meaning it tends to make more decisions to ensure that all predictions made cover the majority of correct results, resulting in higher accuracy. However, it can also lead to lower recall rates due to the low proportion of correct predictions made among all predictions made.

So in order to balance the above two situations and fully utilize $A \cap B$ and $A \cup B$, we consider introducing a third large language model (here we introduce Yi1.5-6B, which has been fine tuned on the dataset) for supervision. Firstly, the correct predictions in $A \cap B$ account for a large proportion, so we consider directly incorporating them into the answer. For $A \cup B$, we consider using C (i.e. the predicted result of Yi1.5-6B on the test set) for supervision. If an element in $A \cup B$ appears in C, we believe it is likely to be a correct prediction and incorporate it into the answer. Otherwise, we consider it to be an incorrect prediction. In the end, we obtain the answer:

$$Answer = (A \cap B) \cup ((A \cup B) \cap C)$$

Simplifying the above equation, we have

$$Answer = (A \cap B) \cup (A \cap C) \cup (B \cap C)$$

This equation is not difficult to explain, as the above steps are based on the assumption that the results predicted to be the same by both models have a higher probability of being correct, which means that we take the union of the predicted results of any two models

Similarly, we can introduce a fourth model D (DeepSeek-7B introduced in this article) for supervision and the sets of their output results are denoted as A, B, C, and D, respectively. Then our IU-Block model is as follows: We trained four different large models, and the sets of their output results are denoted as A, B, C, and D, respectively. We perform the following operations on the output results of the four large models.

$$Answer = (A \cap B) \cup (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D) \cup (C \cap D)$$

We take $Answer$ as the final output.

The establishment of the IU-Block model can improve both the recall rate and precision rate of the model, thereby enhancing the overall performance of the model.

5 Experiments

5.1 Experiments on Single-LLM Fine-tuning

5.1.1 Hyperparameter Tuning Experiment

Based on the parameter interface of the ERNIE-Speed-8K model, we made the following adjustments to the model parameters. The experimental data is shown in table 1, table 2 and table 3. below.

CCL 2024

It should be noted that the result of this experiment is tested on the blind test set. The last parameter combination achieved the highest F1 score, indicating the best model performance.

Training Set Size	Validation Set Size	F1/%
3000	1000	40.2
3000	1000	56.28
3000	1000	66.1
3600	400	71.6
4000	0	74.76

Table 1: Hyperparameter Tuning Process Part 1

Learning Rate	Epoch	Learning Rate Adjustment Plan
0.00002	3	linear
0.00002	3	linear
0.00002	6	linear
0.000025	6	cosine
0.000025	6	cosine

Table 2: Hyperparameter Tuning Process Part 2

Number of Cosine Cycle	Regularization Coefficient	Temperature	Diversity
–	0.01	0.95	0.8
–	0.01	0.95	0.8
–	0.01	0.95	0.8
0.5	0.009	0.95	0.8
0.5	0.009	0.6	0.6

Table 3: Continuation of Hyperparameter Tuning Process

5.1.2 Hyperparameter Tuning Process

Utilizing the Wenxin Qianfan's large-scale model platform, we visualized the loss and perplexity (ppl) during the fine-tuning process as shown in Figure 2.

CCL 2024

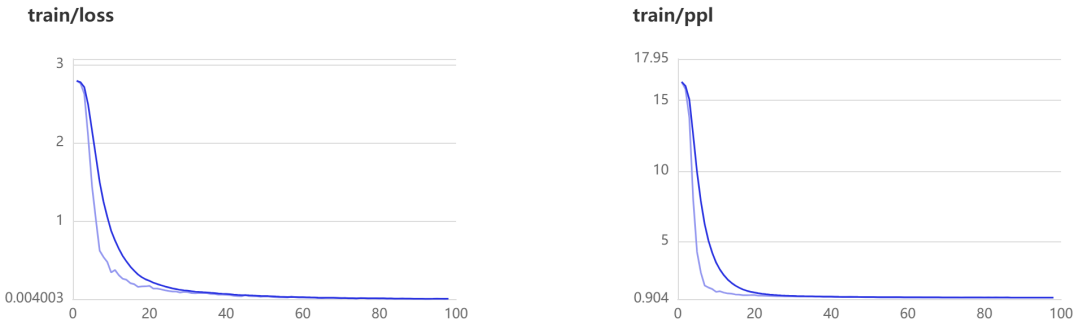


Figure 2: Changes in Loss and Perplexity During Training

5.1.3 Results

Our hyperparameter tuning results are shown in table 4.

Hyperparameter	Value
Number of Iterations	6
Gradient Accumulation Steps	4
Learning Rate	0.000025
End LR for Polynomial Strategy	1e-7
Learning Rate Adjustment Plan	cosine
Sequence Length	4096
Number of Cosine Cycles	0.5
Power for Polynomial Strategy	1
Fake Multi-Turn Probability	0
Checkpoint Saving Interval	256
Random Seed	42
Warmup Proportion	0.1
Regularization Coefficient	0.009
temperature	0.6
top-p	0.6

Table 4: Parameter Description

It should be noted that $F_1 = 74.76\%$ is the result we obtained after the competition. Compared to $F_1 = 71.6\%$, we adjusted the temperature and top-p from both 0.8 to 0.6, and adjusted the gradient accumulation steps from 0 to 4, while incorporating the validation set into the training set.

5.2 Experiments on Multi-LLM Ensembling and Tuning

This section mainly focuses on the performance of large models with smaller parameter scales. In our experiments, we introduced four large models: Qwen1.5-7B, ChatGLM3-6B, Yi1.5-6B, and DeepSeek-7B. First, we present the performance of these four large models after hyperparameter tuning

CCL 2024

and training 6 epochs when independently completing the task, as shown in table 5. The result of the experiment is tested on the test set.

LLM	Precision	Recall	F1
Qwen1.5-7B	0.6101	0.7053	0.6542
ChatGLM3-6B	0.5924	0.6196	0.6057
Yi1.5-6B	0.6337	0.6755	0.6539
DeepSeek-7B	0.6118	0.6757	0.6421

Table 5: Precision, Recall, and F1 Scores of Different LLMs

5.2.1 RC-Block and Cross-Fusion Experiment

In this experiment, we seperately trained the Qwen1.5-7B model, the ChatGLM3-6B model, the Yi1.5-6B model and the DeepSeek-7B model as candidates of the RC-Block, which is shown in figure 1. We recorded their results on the test set as A, B, C and D, respectively.

We compared introducing the models above with or without the method of RC-Block and whether the results are intersected after running the model. The results are shown in the table 6. The result of the experiment is tested on the test set.

RC-Block	Intersect	Precision	Recall	F1
False	False	0.6720	0.7390	0.7039
True	False	0.6694	0.7418	0.7037
False	True	0.6638	0.7395	0.6996
True	True	0.6780	0.7456	0.7102

Table 6: RC-Block Results with Intersection

The experimental results indicate that both the RC-Block and Cross-Fusion methods can effectively improve the performance of large models.

5.2.2 IU-Block Experiment

We completed the training of four models using the RC-Block method and generated outputs A and B for the Qwen1.5-7B and ChatGLM3-6B models using the Cross-Fusion method. The results C and D were obtained by running the test set on the other two models. The results were processed using the IU-Block in different ways, as shown in Table 7. The result of the experiment is tested on the test set.

CCL 2024

Method	Precision	Recall	F1
$A \cap B$	0.4527	0.8513	0.5911
$A \cup B$	0.7257	0.5649	0.6353
ABC+IU-Block	0.6152	0.7941	0.6933
ABD+IU-Block	0.6109	0.7956	0.6911
$ABCD$+IU-Block	0.6780	0.7456	0.7102

Table 7: Results of Different Methods

The experimental results indicate that the IU-Block can effectively improve the performance of large models.

6 Conclusion

This paper proposes a novel method for constructing CPG, significantly improving the accuracy of their construction. By fine-tuning large language models, a model capable of constructing CPG with high accuracy was developed. Considering the substantial time and space requirements of large language models with extensive parameters, this paper explores the use of smaller-scale models for this task. A system integrating multiple small-scale models and a fusion runtime system was constructed. This approach successfully used smaller-scale models to construct CPG, achieving similar results to larger models but with significant savings in both runtime and space utilization.

CCL 2024

Acknowledgements

We would like to express our sincere gratitude to Professor Xin Xin from Beijing Institute of Technology, for his guidance and instruction in the Knowledge Engineering course, which provided the foundation for this project. Additionally, we would like to thank the CPG team at Beijing Language and Culture University, particularly Professor Endong Xun, Gaoqi Rao, Gongbo Tang and graduate studednt Mengxi Guo, for their support and assistance in this project.

References

Kommineni, Vamsi Krishna and König-Ries, Birgitta and Samuel, Sheeba 2024. *From human experts to machines: An LLM supported approach to ontology and knowledge graph construction* arXiv preprint arXiv:2403.08345.

Vizcarra, Julio and Haruta, Shuichiro and Kurokawa, Mori. 2024. 2024 IEEE 18th International Conference on Semantic Computing (ICSC), 231–232. IEEE.

Carta, Salvatore and Giuliani, Alessandro and Piano, Leonardo and Podda, Alessandro Sebastian and Pompianu, Livio and Tiddia, Sandro Gabriele. 2023. terative zero-shot llm prompting for knowledge graph construction. arXiv preprint arXiv:2307.01128.

Yu, Shuang and Huang, Tao and Liu, Mingyi and Wang, Zhongjie. 2023. BEAR: Revolutionizing Service Domain Knowledge Graph Construction with LLM, 339–346. Springer.

郭梦溪 and 荀恩东 and 李梦 and 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

郭梦溪 and 荀恩东 and 李梦 and 饶高琦. 2024. 基于意合图语义理论的结构标注体系与资源建设. 第二十三届中国计算语言学大会.

Aghajanyan, Armen and Yu, Lili and Conneau, Alexis and Hsu, Wei-Ning and Hambardzumyan, Karen and Zhang, Susan and Roller, Stephen and Goyal, Naman and Levy, Omer and Zettlemoyer, Luke. 2023. Scaling Laws for Generative Mixed-Modal Language Models. Proceedings of Machine Learning Research.

CCL 2024

CCL24-Eval 任务2系统报告：基于关系抽取的中文意合图语义解析方法研究

霍虹颖, 黄少平, 刘鹏远
北京语言大学, 信息科学学院, 北京
202111680974@stu.blcu.edu.cn
www17379430207@163.com
liupengyuan@pku.edu.cn

摘要

意合图是以事件为中心的单根有向语义表征图，在语义计算与应用方面具有重要价值。在CCL-2024中文意合图语义解析评测任务中，为克服意合图为单根有向图、意合图包含隐性事件词以及意合图的语义关系类型十分丰富，导致关系类型过多等诸多方面的难点，本文提出一种将该任务转换为关系抽取的方法。该方法首先对标签进行扩充，分为正向标签和反向标签；其次，对输入进行扩充，将隐性事件词添加到输入中，无须额外对隐性事词进行预测；最后，细分为不带隐性事件词和带隐性事件词的关系抽取任务。实验结果表明，本文方法在官方盲测集上的F1值为64.44%，高出基线模型33.41%，证明了本文方法的有效性。

关键词： 关系抽取 ； 语义分析 ； roBERTa ； 意合图

System Report for CCL24-Eval Task 2:A Study on Semantic Parsing Method of Chinese-Parataxis-Graph-Parsing Based on Relational Extraction

Hongying Huo,Shaoping Huang,Pengyuan Liu
School of Information Science, Beijing Language and Culture University, Beijing
202111680974@stu.blcu.edu.cn
www17379430207@163.com
liupengyuan@pku.edu.cn

Abstract

Chinese Parataxis Graph is an event-centered single-root directed semantic representation graph, which is of great value in semantic computation and application. In the task of semantic analysis and evaluation of the Chinese Parataxis Graph in CCL-2024, in order to overcome the difficulties that the Chinese Parataxis Graph is a single-rooted directed graph, contains implicit event words, and its semantic relation types are very rich, resulting in too many relation types. This paper proposes a method to transform this task into relational extraction. In this method, the labels are expanded into forward labels and reverse labels. Secondly, the input is expanded by adding the implicit event words to the input, and there is no need to predict the implicit event words. Finally, it is subdivided into relational extraction tasks without implicit event words and with implicit event words. Experimental results show that the F1 value of the proposed method on the official blind test set is 64.44%, which is 33.41%higher than the baseline model, which proves the effectiveness of the proposed method.

Keywords: Relational Extraction , Semantic analysis , roBERTa , Chinese Parataxis Graph

(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

意合图是围绕事件的语义表征，事件词是事件的核心表达，事件词关联起事件内的各个成分，在事件间建立关系时，以事件词代表整个事件。意合图由事件结构和实体结构构成，均有内外结构之分，内部结构是构成要素间的关系，外部结构是整体间的关系。在CCL2024-CPG数据集中，定义了包含论元结构、关系事件、特殊标签等7个标签层级。例如，在句子“他哭肿了眼睛”中所包含的语义标签有：(他,哭,A0), (眼睛,肿,A0), (他,眼睛,EntityRel), (了,哭,Time), (了,肿,Time), (哭,因果关系,原因事件), (肿,因果关系,结果事件), (哭,ROOT,CoreWord)。将其看作关系抽取任务，则前两个元素相当于关系抽取中需要识别出来的实体，第三个元素则为这两个实体之间的关系。比如“他”和“哭”之间存在的语义关系为A0，A0代表着论元结构中的主体关系，即实体“他”和“哭”存在着动作发起的主体关系。

意合图理论的提出为解决自然语言处理中的语义难题提供了一种全新的思路，然而这一新兴领域，目前尚无成熟的研究方法和标注策略，本文的工作主要对意合图语义框架的识别提出了一种研究思路，我们通过关系抽取的方法，对语义标签识别并分类标注。

意合图语义分析与普通的关系抽取任务有所不同。意合图为单根有向图，意味着意合图的两个实体的先后位置固定，不可任意交换，实体在三元组中的第一个位置和第二个位置的标签含义并不相同。其次，意合图包含隐性事件词，例如当两个实体是并列关系，与其他事件词构成语义关系时，以“And”标签来指代这两个并列实体这一整体，“And”这样的标签就是隐性事件词，它并未在原始分词后的句子中出现。除此之外，意合图整体体系丰富，语义关系标签众多，这也为关系抽取带来了一定难度。

关系抽取是信息抽取的一项重要子任务，旨在从非结构化文本中识别并提取实体之间的语义关系，主要包括实体识别和关系分类。实体识别需要识别文本中的命名实体，如人名、地名等，关系分类则是判断实体之间是否存在某种特定的关系，并确定关系的类型，在涉及到更复杂的句子结构时，还需要识别事件及其参与者。通过我们的分析发现，意合图语义解析可以认为是一种关系抽取任务，即识别出句子中存在有向关系的实体并进行标签分类。

我们充分利用基于roBERTa的关系抽取方法，来从意合图中解析出事件与实体的语义关系。主要使用的是chinese-roBERTa-wwm-ext-large模型，利用roBERTa模型对文本进行编码，识别出文本中的实体，采用roBERTa模型对任意两个实体之间的关系进行识别和分类。与传统的基于特征工程的方法相比，我们的方法不需要手工设计特征，而是直接从原始文本中学习特征表示，具有更好的泛化能力和适应性。

总的来说，本文采用的方法是将意合图语义框架识别转换成关系抽取任务，针对意合图的单根有向性、包含隐性事件词、语义标签丰富的问题对数据进行了特殊处理，在评估中，我们的模型在盲测集上的F1得分为64.44%，在基线依存模型的基础上提升了33.41%。在chinese-roBERTa-wwm-ext-large模型的基础上，我们也尝试使用大模型进行标签标注，我们考虑到通过输入大模型能够理解的语义知识并提供示例的方法，来让大模型学习不同的语义标签并对句子进行标注。整合两种方法的结果最终得分为64.42%。

2 相关工作

在关系抽取领域的早期研究中，学者们主要探索了基于规则和基于监督学习的方法。基于规则的方法主要根据设定的规则以及固定模式来识别实体之间的关系，这种方法通常受限于规则的覆盖范围和适用性。

关系抽取的过程需要依赖大量标注的训练样本，随着深度学习技术的发展，传统的关系抽取方法逐渐被基于深度神经网络的关系抽取方法所取代。这样的方法可以改善传统方法中的人工选择部分，有效改善特征在抽取过程中的误差累积。监督学习的关系抽取模型以Zeng et al. (2014)提出的卷积神经网络CNN和Lin et al. (2018)提出的循环神经网络RNN为两大代表，Gormley et al. (2015) 提出了一种基于因子的组合嵌入模型(FCM),该模型用依存树和命名实体从单词嵌入构建句子级和子结构嵌入。dos Santos et al. (2015a) 提出了排序CNN（CR-CNN），通过一个新的成对排名损失函数，减少人工类的影响，该方法比使用CNN+softMax classifier的模型更有效。Shen and Huang (2016) 将CNN编码器与句子表示结合使用，该句子表示通过关注目标实体和句子中的单词之间的注意力对单词进行加权，以提高性能。Wang et

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

al. (2016) 为了捕获异构上下文中的模式以对关系进行分类，提出了一种具有两层注意力级别的卷积神经网络体系结构。Lee et al. (2019) 开发了一个端到端的循环神经网络模型，该模型结合了实体感知的注意力机制和潜在的实体类型以进行关系分类。然而训练数据的质量将大幅影响这两种模型的性能，即模型依赖大量高质量的人工标注数据进行模型训练。之后Mintz et al. (2009) 等人提出了远程监督的关系抽取方法，该方法通过将数据自动对齐远程知识库来解决大量无标签数据的自动标注问题，在一定程度上能减少对人工的依赖。利用少量标注的样本进行学习也是一种方案。少样本学习属于监督学习的范畴，通过给出每个类的少量标注信息的实例，使得程序在处理该任务时能实现性能的提升，在训练过程中不改变已有训练好的模型，借助每类少数几个标注样本就可以快速学习并完成新类的分类。

这些方法利用深度神经网络模型来自动学习文本中的特征表示，并在端到端的方式下进行关系抽取。尤其是Transformer模型的出现，特别是其变种roBERTa的问世，为关系抽取任务带来了新的突破，roBERTa模型通过预训练大规模语料库，能够捕捉文本中丰富的语义信息，使得关系抽取模型能够更好地理解文本语境，从而提高了关系抽取的性能。Wu and He (2019) 提出了一个模型，既利用预训练的bert语言模型，又结合来自目标实体的信息来完成关系分类。基于R-BERT-CNN模型的实体关系抽取论文将实体级信息融入预训练模型获取目标实体的语义；采用CNN提取句子级的语义信息；连接句子向量、标签向量和目标实体向量，获得全局信息；通过softmax分类器抽取实体关系。

相较于GPT-3这样的超大型模型，roBERTa的参数规模较小，因此在训练和推理时需要的计算资源更少。roBERTa模型通过优化模型结构和训练方法，提高了计算效率，在训练过程中使用了动态掩码的技术，可以更好地利用计算资源，从而加快速度，所以在同样的数据集和硬件条件下，相较于诸如GPT-3这样的模型，训练roBERTa模型所需的时间更短。roBERTa在BERT的基础上进行了一系列训练技巧和超参数的调整，以提升了模型在文本分类、命名实体识别等自然语言处理任务上的性能，这也就意味着在相同的关系抽取任务和数据集上，使用roBERTa模型可以获得更好地表现而不需要使用更大的模型。

意合图首次参与评测活动，前人的研究主要是在理论与资源构建方面的工作 (郭梦溪 et al., 2024a)。对于意合图语义分析，目前没有太多可参考的方法。但是意合图属于语义分析的一种 (郭梦溪 et al., 2024b)，所以可以借鉴其他语义分析的计算方法，比如抽象语义表示、语义依存等。这些方法在处理复杂的语义关系和结构方面已经取得了显著成果，可以为意合图的进一步研究提供有价值的参考。本文采用的基线模型即为依存模型。

3 基于关系抽取的中文意合图语义解析方法

通过分析CCL2024-CPG数据集，我们已经发现每条数据的输入与输出与传统的关系抽取任务十分类似，抽取结果都为三元组$\{word_1, word_2, Relval\}$。但相比之下，意合图语义框架的难点在于：首先，中文意合图是单根有向图，即$\{word_1, word_2, Relval\}$的顺序不能随意更换。其次，该任务中需要先识别“实体对”再对关系进行分类，同时，该任务需要识别的“实体对”不仅会出现在输入中，由于意合图包括事件外结构，所以还有隐性事件词需要额外预测。最后，该数据集的关系标签分布也不平衡，因为需要考虑“实体对”的顺序，关系标签会加倍，从而标签不平衡的现象更加明显。

为了解决上述问题，我们分别提出了以下解决方法：

1. **针对$word_1, word_2$的顺序问题**：我们将分类标签加倍，$word_1, word_2$在输入中的位置为从前到后，则为正常标签，反之，$word_1, word_2$的位置在输入中为从后到前则为反标签，记作“标签_reverse”。

2. **针对隐性事件词的预测**：我们对输入进行扩充，转化为传统的关系抽取任务，即我们整理了所有可能出现的隐性事件词，将其全部以特定的格式“<隐性事件词>”添加到输入末尾。这时，隐性事件词也出现在输入中，不需要进行额外的预测，从而转化成关系抽取任务。图 3是一个例子，我们将所有隐性事件词加到原始句之后，同时对标签中与前后顺序相反的标签添加“_reverse”。

3. **针对标签不平衡的问题**：考虑到标签加倍后，会影响模型对其他标签的预测性能，我们将任务划分成了两个子任务，即不包含隐性事件词的关系抽取和包含隐性事件词的关系抽取。

(c) 2024 中国中文信息学会计算语言学专业委员会

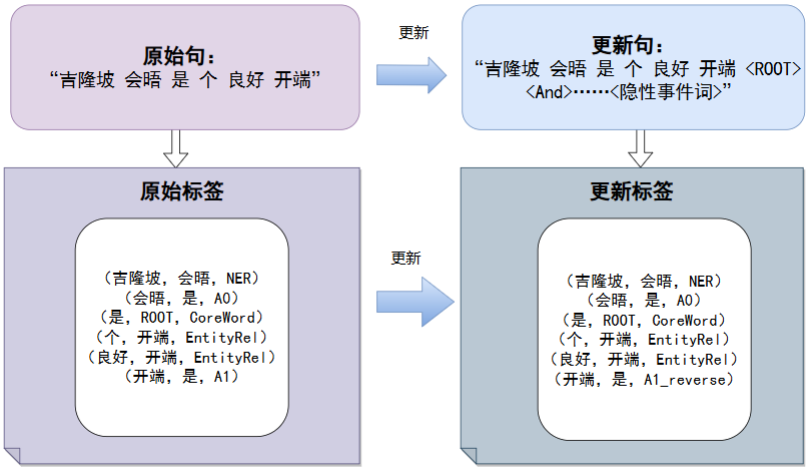


图 1: 展示原始句子标签和更新的句子标签

为了解决顺序问题和隐性事件词的出现，我们对数据进行预处理，以预处理后实体对是否包含有隐性事件词的标签为依据，划分为两个子任务。我们对这两个子任务分别使用对称的关系识别模型和关系分类模型（关系识别模型和关系分类模型均使用chinese-roBERTa-wwm-ext-large模型），从而得到两个子任务在关系抽取方法下的结果，通过对两个结果的整合，从而来得到最终的意合图语义分析结果。图2是该方法的整体流程图。

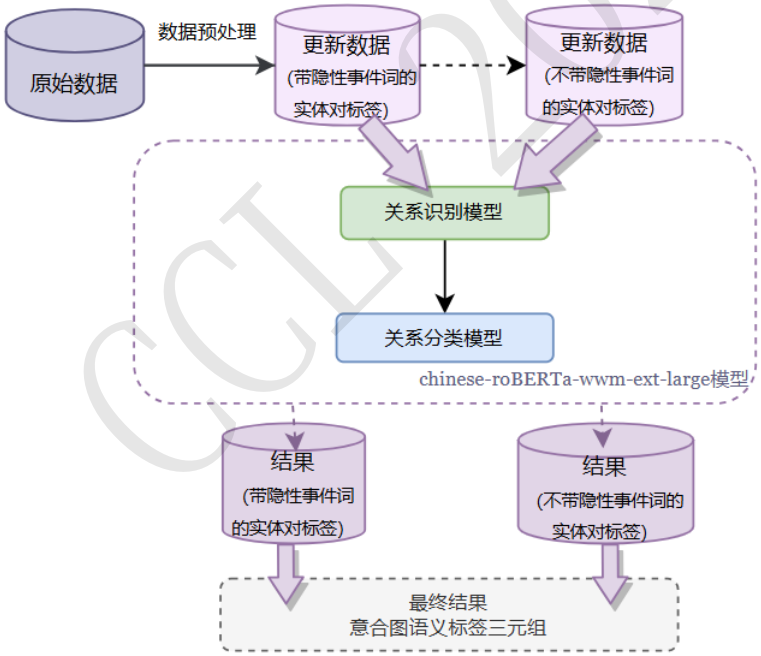


图 2: 整体流程图

首先，对于关系识别模型，我们将原数据集中的不带隐性事件词的实体对和带隐性事件词的实体对分别整理成两个数据集，再把两个数据集分别按7:3的比例划分成训练集和验证集。我们先对预处理后的数据进行关系识别，每条数据格式如表1所示，包含实体1、实体2、标签和句子。标签有类“yes”和“no”。“yes”代表两个实体之间有关系，“no”代表两个实体之间没有关系。

其次，对于关系分类模型，由于不带隐性事件词的实体对和带隐性事件词的实体对的关系

实体1	实体2	标签	句子
今天	发挥	yes	但今天中国队发挥较好，因此获得了胜利。
和	两端	no	佩雷斯说：“中国和以色列地处亚洲的两端，相隔万水千山。

表 1: 关系识别数据集样例

类型有重叠，但大部分不一样，所以我们分别整理了所有可能出现的关系类型，加倍后，前者共64种，后者共94种。同时，所有可能出现的隐性事件词数量为13种，这些隐性事件词以“<隐性事件词>”的格式扩充到输入中。具体见表2。最后，将所有含答案数据按照不带隐性事件词的关系类型和带隐性事件词的关系类型分别整理成两个数据集，同样按7:3的比例划分成训练集和验证集。每条数据的格式与表1相同，标签部分为预处理后的更改标签。

chinese-roBERTa-wwm-ext-large模型的架构包括三个主要部分：roBERTa编码层、实体特征处理层和分类层。

3.1 roBERTa编码层

首先，将输入句子中的token通过roBERTa模型编码，得到每个token的隐藏状态和句子级别的表示（即[CLS] token的表示）。

假设输入句子表示为$X = \{x_1, x_2, \ldots, x_n\}$，其中$n$ 为句子的长度。通过roBERTa模型编码后，得到每个token的隐藏状态$H = \{h_1, h_2, \ldots, h_n\}$ 和句子级别的表示$h_{[\text{CLS}]}$：

$$H, h_{[\text{CLS}]} = \text{roBERTa}(X) \tag{1}$$

其中，$H \in \mathbb{R}^{n \times d}$，$h_{[\text{CLS}]} \in \mathbb{R}^d$，$d$ 为隐藏状态的维度。

3.2 实体特征处理层

对于关系识别模型中存在关系的两个实体，通过对实体的所有token的隐藏状态取平均值来表示实体的特征。具体来说，假设实体e_1 的token索引范围为$[i, j]$，则实体e_1 的特征表示为：

$$e_1 = \frac{1}{j - i + 1} \sum_{k=i}^{j} h_k \tag{2}$$

为了简化计算，我们定义实体掩码$e_1^{\text{mask}} \in \{0, 1\}^n$，当且仅当$k \in [i, j]$ 时，$e_1^{\text{mask}}[k] = 1$。则上述平均操作可以表示为：

$$e_1 = \frac{e_1^{\text{mask}} \cdot H}{\sum_{k=1}^{n} e_1^{\text{mask}}[k]} \tag{3}$$

同理，实体e_2 的特征表示为：

$$e_2 = \frac{e_2^{\text{mask}} \cdot H}{\sum_{k=1}^{n} e_2^{\text{mask}}[k]} \tag{4}$$

为了增强实体特征的表示能力，将其通过一个线性层和非线性激活函数处理：

$$e_1' = \tanh(W_e e_1 + b_e) \tag{5}$$

$$e_2' = \tanh(W_e e_2 + b_e) \tag{6}$$

其中，$W_e \in \mathbb{R}^{d \times d}$，$b_e \in \mathbb{R}^d$。

CCL 2024

3.3 分类层

将句子的[CLS]表示与两个实体的特征拼接起来，形成最终的特征向量：

$$h_{\text{concat}} = [h_{[\text{CLS}]}, e'_1, e'_2] \tag{7}$$

其中，$h_{\text{concat}} \in \mathbb{R}^{3d}$。

为了提高模型的泛化能力，对拼接后的特征向量进行层归一化和dropout处理：

$$h_{\text{concat\_norm}} = \text{LayerNorm}(h_{\text{concat}}) \tag{8}$$

$$h_{\text{concat\_drop}} = \text{Dropout}(h_{\text{concat\_norm}}) \tag{9}$$

最后，通过一个线性层将特征向量映射到关系类别空间：

$$\text{logits} = W_{\text{out}} h_{\text{concat\_drop}} + b_{\text{out}} \tag{10}$$

其中，$W_{\text{out}} \in \mathbb{R}^{k \times 3d}$，$b_{\text{out}} \in \mathbb{R}^{k}$，$k$ 为关系类别的数量。

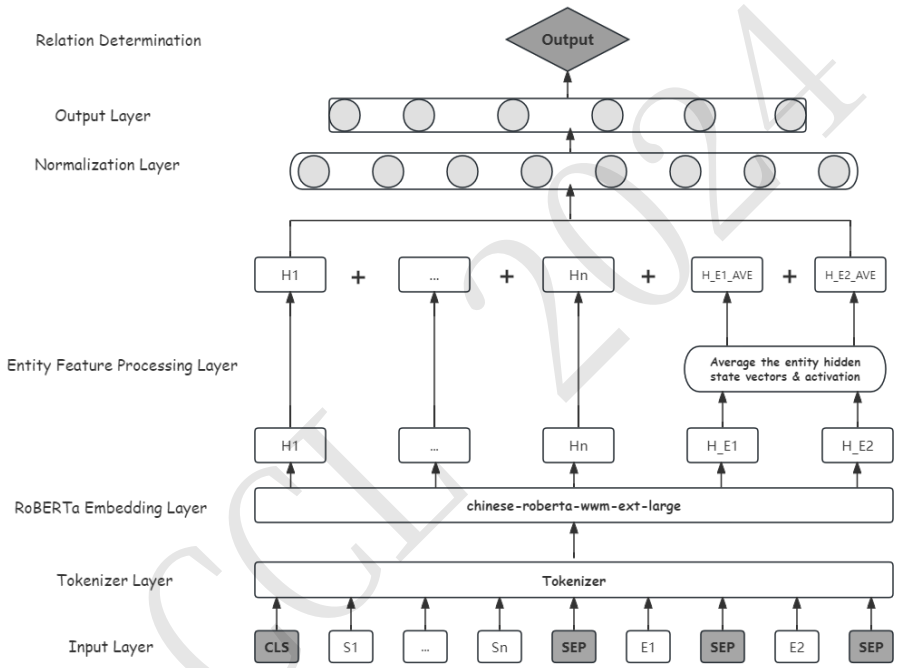


图 3: 模型框架图

通过以上三个主要部分的协同工作，chinese-roBERTa-wwm-ext-large模型能够有效地进行关系抽取任务，从而识别出文本中不同实体之间的关系。

4 实验

4.1 实验数据与评价方法

本文实验数据来自CCL2024-CPG数据集，包含来自国际中文教育阅读语料和中文宾州树库的新闻语料。每条数据包含分词后的句子、意合图语义标签、$word_1, word_2$的idx，即特定序号等标注信息。本文使用数据集中的训练集和验证集共3000句语料来训练模型。本次评测采用F1值作为模型表现的评价标准，其计算方式如下：

$$Precision = \frac{Matching\ Tuples}{Generated\ Tuples} \tag{11}$$

(c) 2024 中国中文信息学会计算语言学专业委员会

$$Recall = \frac{Matching\ Tuples}{Gold\ Tuples} \tag{12}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

其中，Generated Tuples 为模型预测的三元组集合数，Gold Tuples 为测试集/盲测集的三元组集合数，Matching Tuples 为模型预测的三元组集合与测试集/盲测集的三元组集合间的最大匹配个数。

4.2 实验参数设置

本文采用哈工大版本的chinese-roBERTa-wwm-ext-large模型，实验的相关参数设置如下表所示。

参数名	参数值
attention_probs_dropout_prob	0.1
bos_token_id	0
directionality	bidi
eos_token_id	2
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	1024
initializer_range	0.02
intermediate_size	4096
layer_norm_eps	1e-12
max_position_embeddings	512
model_type	bert
num_attention_heads	16
num_hidden_layers	24
output_past	true
pad_token_id	0
pooler_fc_size	768
pooler_num_attention_heads	12
pooler_num_fc_layers	3
pooler_size_per_head	128
pooler_type	first_token_transform
type_vocab_size	2
vocab_size	21128

表 2: 实验参数设置表

4.3 实验结果

经过10轮的迭代训练后，关系识别模型在不含隐性事件词的任务中F1值最佳时为0.9653，关系识别模型在含隐性事件词的任务中F1最高值为0.9840。尽管F1值很高，但实际在盲测集进行关系识别时，是对所有实体进行一一枚举，让模型判断是否有关系，输出“yes”或者“no”，如果直接利用训练好的模型进行推理，包含隐性事件词的数据集预测的实体对会过多。因此我们对预测结果向量设置了一个最小阈值，该向量为二维向量，分别表示两个标签的置信度，由于该二维向量未经过Softmax以及归一化成概率，故将其称为logits，只有“yes”标签对应的logit大于2时，预测结果为“yes”的实体对才被模型输出，否则无效，其中最小阈值是一个超参数，经过多轮实验对比，最小阈值为2时，实验结果最优。

关系分类仍然利用基于roBERTa的关系抽取模型，为了缓解标签过多，造成对模型性能的影响，我们训练了两个关系分类模型，相对而言，由于不带隐性事件词的实体对数量多于带

(c) 2024 中国中文信息学会计算语言学专业委员会

隐性事件词的实体对数量，且后者中个别标签数量太少，所以从Marco F1值来说，不带隐性事件词的实体的关系分类模型整体性能较好，但从Weighted F1来看，带隐性事件词的关系分类模型性能好于不带隐性词的关系分类模型性能，前者为F1为0.9524，后者为0.9287。最后，对两个子任务的结果进行整合，对更新后的标签进行还原，并增加位置属性idx，得到最终结果，按照评价方法计算在盲测集上得到F1得分为64.44%。F1值在整合前高是由于在关系识别阶段，“yes”和“no”标签分布不均衡，大部分分类结果都是“no”，所以F1值会呈现过高的情况，在整合结果后，所有分类结果为“no”的情况已被筛出，所评价的均为分类结果为“yes”的数据，故F1值会下降。

4.3.1 扩展研究

我们整理了chinese-roBERTa-wwm-ext-large模型的结果，发现对于数据集中数量少的语义标签，抽取效果并不理想。我们通过对验证集结果的整理分析，筛选出所有未被chinese-roBERTa-wwm-ext-large模型抽取出来的语义标签，基于大模型强大的语义理解和生成能力，考虑使用大模型来完成这一部分语义标签的标注任务。通过对意合图体系及标签定义的学习，以及对数据集的观察，我们发现不同的语义标签在意合图语义理论的定义（如，关系事件：关系事件由关系论元和关系事件词构成，关系论元即具有关系的实体或事件，表示实体间关系或事件间关系的抽象关系词，被视为关系事件的核心表达，作为关系事件的事件词，如“因果关系”“领属关系”等，属于关系型隐性事件词）之外，还可以与语法知识、上下文语境等来通俗地描述。由此，我们考虑到通过输入大模型能够理解的语义知识并提供示例的方法，来让大模型学习不同的语义标签并对句子进行标注。对于这些标签，我们使用大模型包括GPT-4、ChatGLM、deepseek等，查找相关定义以及在训练集中找到对应例子，基于语法、上下文语境等的理解同时使用思维链、上下文学习等方法来调用大模型API得到三元组，如：

prompt=

“‘对于句子“中国社会科学院的经济学者朱运法、张延群说：“住房分配货币化已是大势所趋。”，有一个三元组(And,说,A0)，把“中国社会科学院的经济学者朱运法、张延群”意思是这两个人作为一个整体，标注为And，这个整体和“说”是一种A0的关系，相当于说A0是一个主语与谓语之间的关系。

对于句子“中国古代在造纸的技术、设备、加工等方面为世界各国提供了一整套先进的工艺体系。”，三元组(And,提供,范围),将“一整套先进的工艺体系”And标签，在该句子中与“提供”构成“范围”标签。

当然，不是所有的句子都会出现这样的关系，不用标注其他关系，也不需要标注And的所指内容。

请你一步一步思考，识别以下分词了之后的句子中*And和A0*或者*And和A1*或者*And和A2*或者*And和范围*的关系，并按照这样的格式输出：

sent=[”分词后的句子”] results=[(”And”,”句子里的词”,”A0/A1/A2/范围”)]”’

4.3.2 与基准系统相比

本文提出的基于关系抽取的意合图语义分析方法得分为64.44%，与基线依存模型相比提升了33.41%，与基线大模型GPT-4交互式相比提升了27.15%。在拓展研究中让大模型对在chinese-roBERTa-wwm-ext-large模型中F1得分低的标签如PN、CompPN、Conj、FW、TM、PF、SF等进行重新整理。整合大模型和小模型的结果，最终F1为64.42%。相较单独使用chinese-roBERTa-wwm-ext-large模型得分有所下降。

4.3.3 错误分析

对实验结果进一步分析发现，本文提出的方法仍然存在不足之处。一方面，本文的标签翻倍虽然在一定程度上解决了意合图的有向问题，但是这会导致标签数量过多，从而影响模型性能。其次，对于标签位置idx的标注，本文在数据整合阶段单独标注，这会导致在某些出现多次的隐性事件词标签的位置idx标注只能有一个值，而这将会降低模型的F1值。另一方面，在大模型的应用上，本文的方法较为浅显，并没有很好地利用到数据集进行微调，所采用的提示工程方法也较为单一。

除此之外，可以观察到在扩展研究中使用大模型对chinese-roBERTa-wwm-ext-large模型未能抽取出来的标签重新单独抽取，得到的F1值相较不使用大模型并没有提高，对此，我们对大

(c) 2024 中国中文信息学会计算语言学专业委员会

实验系统	**F1得分**（%）
基线依存模型	31.03
基线GPT4交互式	37.29
chinese-roBERTa-wwm-ext-large模型	**64.44**
混合模型（GPT-4、ChatGLM、deepseek、chinese-roBERTa-wwm-ext-large模型）	64.42

表 3: 实验结果

模型标注的数据进行了错例分析。我们发现:

1. **大模型生成不可控**：即使在提示语中约束的情况下，除了对特定标签的标注，大模型还会自己生成一些不在范围内的标签，甚至还会生成二元组、四元组等，大模型的生成不可控，需要对格式、标签种类进行筛选。

2. **抽取标签数量过多**：我们观察到，这些特定标签本身出现概率较小，而大模型几乎会在每一个句子都抽取特定标签，其中大部分句子并不含有该标签，因此会导致F1值下降。

3. **缺乏上下文示例**：由于输入输出长度限制，大模型无法学习每一个包含标签的句子，所以缺乏上下文学习的示例，这也导致大模型的抽取效果不佳。

5 结论

本文采用的基于关系抽取的中文意合图语义解析方法,通过给分类标签加上顺序来解决意合图中的有向问题；我们将所有可能出现的隐性事件词添加到输入末尾从而也将隐性事件词的预测转换成关系抽取任务。为了保持标签的平和，我们划分成两个子任务，即不包含隐性事件词的关系抽取和包含隐性事件词的关系抽取，同时训练两个子任务模型并对于这一部分未能标注出的数据采用大模型进行尝试。实验结果证明，本文提出的关系抽取方法不仅可以考虑到$word_1, word_2$的顺序，还能通过扩充输入的方式统一非隐性事件词之间、隐性事件词之间、隐性事件词和非隐性事件词之间的关系抽取任务，在基线依存模型的基础上F1提升了33.41%，可见我们的方法有效且具有创新性。在下一步的研究工作中，可以考虑通过微调大模型来提升大模型的效果，也可以将关系抽取的方法运用到篇章的语义解析任务上。

参考文献

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765, Berlin, Germany, August. Association for Computational Linguistics.

Cícero dos Santos, Bing Xiang, and Bowen Zhou. 2015a. Classifying relations by ranking with convolutional neural networks. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634, Beijing, China, July. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015b. Classifying relations by ranking with convolutional neural networks.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784, Lisbon, Portugal, September. Association for Computational Linguistics.

Joohong Lee, Sangwoo Seo, and Yong Suk Choi. 2019. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing.

(c) 2024 中国中文信息学会计算语言学专业委员会

Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In Alberto Lavelli, Anne-Lyse Minard, and Fabio Rinaldi, editors, *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 165–176, Brussels, Belgium, October. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In Phil Blunsom, Shay Cohen, Paramveer Dhillon, and Percy Liang, editors, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado, June. Association for Computational Linguistics.

Yatian Shen and Xuanjing Huang. 2016. Attention-based convolutional neural network for semantic relation extraction. In Yuji Matsumoto and Rashmi Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany, August. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification.

Mo Yu, Matthew R. Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1374–1379, Denver, Colorado, May–June. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In Junichi Tsujii and Jan Hajic, editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

郭梦溪, 李梦, 荀恩东, 饶高琦, and 于钟洋. 2024a. 基于意合图语义理论的结构标注体系与资源建设. 第二十三届中国计算语言学大会.

郭梦溪, 荀恩东, 李梦, and 饶高琦. 2024b. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

CCL 2024

CCL24-Eval任务2系统报告：基于样本设计工程和大模型微调的中文意合图语义解析*

司函[1]，罗智勇[1†]
[1]北京语言大学 信息科学学院
202221198690@stu.blcu.edu.cn，luo_zy@blcu.edu.cn

摘要

本文介绍了我们在第二十三届中国计算语言学大会中文意合图语义解析评测中提交的参赛系统。中文意合图（Chinese Parataxis Graph，CPG）是以事件为中心的语义表征图，可以对不同层级的语言单元作一贯式表示，是一种通用性与扩展性兼具的语义表征方法。鉴于大语言模型在语义解析任务中的优越性能，我们对Llama3-Chinese-8B-Instruct模型进行了LoRA微调，使其能够生成结构化的意合图表征三元组，并采用了样本设计工程（Sample Design Engineering，SDE）技巧进行微调样本的设计。此外，我们还对不同标签进行了分类微调，探究大模型在不同语义标签预测能力上的差异。最终，我们的参赛系统在任务发布的评测集上F1值达到0.6461，在本次评测任务中获得了第三名的成绩。

关键词： 意合图；语义解析；大语言模型；样本设计工程

System Report for CCL24-Eval Task 2: The Chinese Parataxis Graph Parsing Based on Sample Design Engineering and Fine-Tuning Large Language Model

Han Si[1]，Zhiyong Luo[1*]
[1]School of Information Science, Beijing Language and Culture University
202221198690@stu.blcu.edu.cn，luo_zy@blcu.edu.cn

Abstract

This paper introduces the system we submitted in the shared task of Chinese Parataxis Graph (CPG) Parsing at the Twenty-three Chinese National Conference on Computational Linguistics. The Chinese Parataxis Graph is an event-centered semantic representation graph that provides a consistent representation of language units at different levels, offering a versatile and extensible method of semantic representation. Considering the superior performance of large language models in semantic parsing tasks, we fine-tuned the Llama3-Chinese-8B-Instruct model using LoRA to enable it to generate structured parataxis graph representation triples. We also employed the Sample Design Engineering (SDE) technique for the design of fine-tuning samples. Furthermore, we conducted classification fine-tuning for different labels to explore the model's performance in predicting various semantic labels. Ultimately, our system achieved a F1 score of 0.6461 on the evaluation set provided by the task organizers, securing the third place in this evaluation task.

* 基金项目：国家自然科学基金（62076037）
† 通讯作者

第二十三届中国计算语言学大会论文集，第72页-第79页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

Keywords: Chinese Parataxis Graph , Semantic Parsing , Large Language Model , Sample Design Engineering

1 引言

语义解析是指将自然语言文本转化为结构化的语义表示，使得计算机能够理解和执行人类的指令。语义解析是自然语言处理领域亟待突破的瓶颈，精准把握自然语言语义需要准确且完备的语义表示方法。英文语义表示的研究发展较早，最具典型的就是抽象语义表示(Abstract Meaning Representation，AMR) (Banarescu, 2013)，它是句子级语义表示方法的一种，它将句子中的事件、状态、属性等内容抽象为语义概念，通过图的方式表示不同概念之间的语义关系，图中的边则表示不同概念节点之间存在的语义关系。随着AMR逐渐受到大家的关注，其他非英语语言的AMR语料库也得到了丰富，Li et al. (2016)将AMR推广到中文，称为中文抽象语义表示(Chinese Abstract Meaning Representation, CAMR)。近些年，荀恩东(2023)提出意合图理论，它是以事件为中心的语义表征图，为单根有向图，图中的节点对应承载事件、实体、属性的单元，边为有向边，表示单元间的语义关系，意合图力求能够对句子、段落、篇章等不同层级的语言单元作一贯式表示(郭梦溪et al., 2024)。

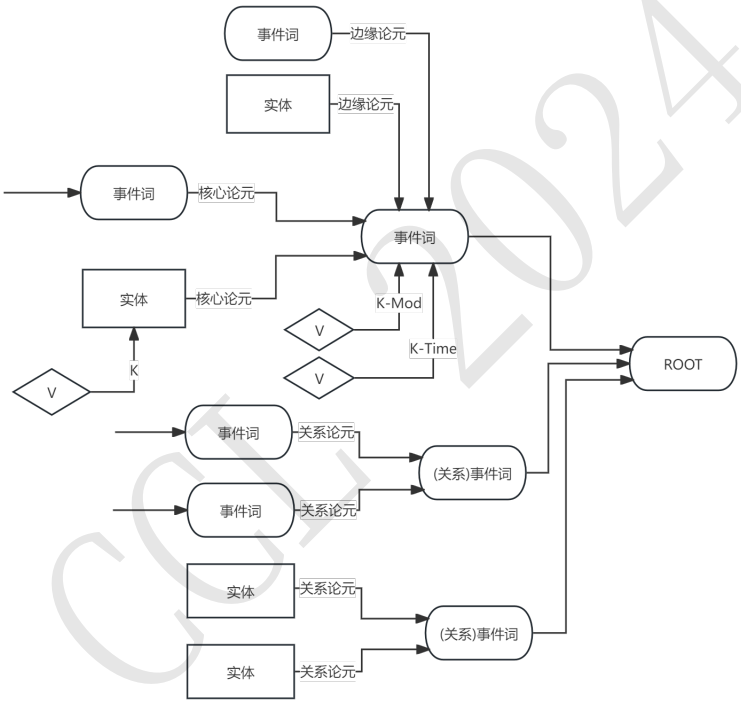


图 1: 意合图抽象表示

语义表示方法的发展也推动了语义解析技术的进步。目前，针对AMR语义解析的主流方法主要分为三类：基于图的方法(Graph-based)(Flanigan et al., 2014; Zhang et al., 2019)、基于转移的方法(Transition-based) (Wang et al., 2015; Ballesteros and Al-Onaizan, 2017; Astudillo et al., 2020)和基于序列到序列的方法(Seq2Seq-based)(Barzdins and Gosko, 2016; Peng et al., 2017; Noord and Bos, 2017; Konstas et al., 2017; Xu et al., 2020)。随着大模型时代的到来，一些研究将大模型技术引入语义解析任务中。Yang et al.(2023)探索了大型语言模型是否能够借鉴Seq2Seq建模方式在复杂结构化预测任务上得以运用。高文炀et al.(2023)选择微调Baichuan-7B模型来以端到端的形式从文本直接生成序列化的CAMR。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

受此启发，我们选择了Llama3的中文微调模型Llama3-Chinese-8B-Instruct作为基座模型，在此基础上对其进行了LoRA (Hu et al., 2021)微调来进行中文意合图的语义解析。由于微调样本的不同设计会显著影响大模型微调后的效果，我们进行了样本设计工程(Sample Design Engineering, SDE)(Guo et al., 2024)设计。此外，在实验过程中，经过统计发现，大模型对不同语义标签解析难度不同，为此，我们设置了不同的模型训练策略，并对解析结果进行了组合分析。

2 方法

我们首先对中文意合图数据集进行了分析统计，根据统计规律以及任务性质进行了SDE设计。在设计好的样本集上微调Llama3-Chinese-8B-Instruct模型生成结构化的意合图表征三元组。为了提高大模型输出结果的质量和可靠性，我们设计了一些规则进行后处理。

2.1 SDE设计

根据简单的任务分析与统计，我们发现中文意合图数据集只有3000条可用的训练数据，共出现了63个语义标签，且除了要挖掘句子中词语之间的语义关系外，还需要对隐式事件词进行补充，所以我们认为中文意合图语义解析任务是一个较为复杂的下游任务。而细致地考虑大模型微调样本的设计，可以使用更少的样本训练出在下游任务上表现更好的模型(Guo et al., 2024)。因此，我们对评测提供的中文意合图数据集进行了简单分析，使用了SDE技巧来完成微调样本的设计。

图 2: 输入样本格式

(c) 2024 中国中文信息学会计算语言学专业委员会

最终，我们的大模型输入样本如图2所示。整个输入样本主要包含了四部分内容，分别为：上下文、指令、输入数据和输出指示。

- 上下文：对意合图的概念做了简单解释，帮助大模型更好的理解中文意合图的表示方法；
- 指令：概述模型需要执行意合图语义解析任务；
- 输入数据：告知模型需要进行语义解析的句子，句子的输入形式为json格式，“sent”字段中为按照特定方式分好词的句子；
- 输出指示：详细说明模型输出的格式，模型最后需要输出结构化的json格式，并向模型解释每个字段的意义和值域选项。

在输入样本格式的设计过程中，我们采用了Guo et al.(2024)工作中经过实验验证的SDE设计技巧。首先，我们将上下文、指令以及输出指示放置于输入的任务文本之前，这样更有助于提升模型的任务理解能力；其次，因为输出设计越格式化，格式输出错误的几率就越低，并且评测任务最终要求提交的是json格式，所以我们选择最结构化的输出格式json；最后，我们按照在训练集中出现的频次，从大到小排列语义标签，以提高模型对出现次数多的标签的关注程度。

2.2 大模型微调

Llama3-8B是一个基于仅解码器Transformer架构的多语言模型，拥有近80亿参数，在超过15万亿个标记(tokens)的公开数据上预训练，支持8192的上下文长度。通过引入分组查询注意力机制(Group Query Attention, GQA)(Ainslie et al., 2023)、扩大模型规模、更新分词器、增加词表大小和使用更为庞大的训练数据集，Llama3-8B展现出了强大的语言理解和生成能力。Llama3-Chinese-8B-Instruct1[1]是以Llama3-8B为基座模型的中文指令微调版本。我们使用LoRA(Hu et al., 2021)对Llama3-Chinese-8B-Instruct模型进行了监督微调，微调使用的参数设置如表1所示。

参数	参数值	参数含义
截断长度	4096	输入序列分词后的最大长度
学习率	5e-5	AdmaW优化器的初始学习率
训练轮数	6.0	需要执行训练总轮数
最大样本数	3000	每个数据集最多使用的样本数
计算类型	Fp16	训练使用的混合精度类型
批处理大小	1	批处理的样本数量

表 1: 微调参数设置

在微调过程中，我们以最小化交叉熵损失函数为优化目标：

$$J(\theta) = -\sum_{z_i \in V} z_i log(P(\hat{z}_i)) \tag{1}$$

$$P(\hat{z}_i) = \frac{e^{\hat{z}_i}}{\sum_{z_i \in V} e^{\hat{z}_i}} \tag{2}$$

其中，V是词汇表的大小，z_i为真实分布中第i个词的值（对于one-hot编码，目标词的z_i为1，其余为0），$\hat{z}_i$为模型预测该词的概率。

[1]https://www.modelscope.cn/FlagAlpha/Llama3-Chinese-8B-Instruct.git

(c) 2024 中国中文信息学会计算语言学专业委员会

模型	F1值
chatglm3-6b[1]	0.6572
Qwen1.5-7B[2]	0.6595
Qwen2-7B[3]	**0.6968**
Chinese-Falcon-7B[4]	0.4189
Llama3-Chinese-8B-Instruct	0.6251
Yi-1.5-9B[5]	0.6739
依存模型	0.3103
GPT4交互式	0.3729

表 2: 微调后不同基座模型以及基线的预测结果

2.3 后处理

大语言模型是概率模型，只是基于模型输入预测输出，因此大模型的输出不可避免地会产生一定的幻觉。而且，由于模型输出上下文长度有限，会导致大模型生成的json格式不完整。为了缓解这些问题对最终结果的影响，我们采取了以下三种后处理方式：

- json格式补全：在生成过程中，由于上下文长度的限制，模型生成的结果可能会被提前截断。对于这些无法直接进行json解析的预测结果，我们采用正则表达式匹配找到最后一个完整的三元组，舍弃剩余部分，并补全json格式，以尽可能多地保留模型生成的结果。

- 节点编号对齐：由于大模型对数字不敏感，在预测节点编号时可能会产生错误。为解决这一问题，我们采用规则的方法进行了节点编号的对齐处理。具体而言，如果节点内容词是隐式事件词，我们按照编号的对应关系直接对齐；如果节点内容词出现在输入句子中，则需考虑两种情况：1）若内容词在输入句子中仅出现一次，则将节点编号与内容词在输入句子中的索引对齐；2）若内容词在输入句子中出现多次，则将内容词定位到与输入句子中最接近预测编号的位置，节点编号与该处的索引对齐。

- 节点内容词修正：在模型生成的结果中，可能会出现生成的节点内容词既不是隐式事件词，又不在输入句子中的情况。针对这一情况，我们首先判断生成的节点编号是否合法，即该编号是否大于0且小于输入句子列表的长度。若编号合法，我们将生成的内容词替换为在输入句子中以该节点编号为索引的词；若编号不合法，则直接删除该节点的三元组。

3 实验

3.1 全语义标签微调

按照图2中的样本格式作为模型输入，在表1所示的参数设置下对Llama3-Chinese-8B-Instruct进行了监督微调。模型使用评测任务中的所有有标签中文意合图数据集作为训练集，对全部63个语义标签进行预测。为了评估其性能，我们在相同实验条件下微调了一系列同等规模的开源模型，实验结果如表2所示。由实验结果可知，微调后的Llama3-Chinese-8B-Instruct模型在该任务评测集上的F1值明显优于依存模型和GPT4交互式基线模型，且展现出了比Chinese-Falcon-7B模型更好的性能。然而，与其他中文模型相比，其性能仍存在不足。由此可见，在解决中文意合图语义解析任务方面，相较于使用中文语料指令微调的英文大模型，中文大模型展现出更高的适应性和优越性。

[1]https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
[2]https://www.modelscope.cn/qwen/Qwen1.5-7B.git
[3]https://www.modelscope.cn/qwen/Qwen2-7B.git
[4]https://huggingface.co/Linly-AI/Chinese-Falcon-7B
[5]https://www.modelscope.cn/01ai/Yi-1.5-9B.git

(c) 2024 中国中文信息学会计算语言学专业委员会

此外，为了评估大模型对不同语义标签的预测能力，我们在微调后的Llama3-Chinese-8B-Instruct模型上对3000条训练集数据的进行了推理预测。定义语义标签l的错误率err_l，该错误率衡量的是语义标签l的预测准确性，具体公式如下：

$$err_l = \frac{\sum_{i=1}^{3000} |\{(word1, word2, relVal) \in P_i \cap T_i : relVal = l\}|}{\sum_{i=1}^{3000} |\{(word1, word2, relVal) \in T_i : relVal = l\}|} \tag{3}$$

其中，P_i表示第i个样例的预测结果中含有语义标签l的三元组集合，T_i表示第i个样例的真实结果中含有语义标签l的三元组集合。

各语义标签的错误率如图3所示。根据实验结果，我们发现，在全语义标签微调下，大模型对不同语义标签的预测能力存在差异。因此，基于错误率是否小于0.5，我们将语义标签分为易预测和难预测两类，并进行了进一步实验。

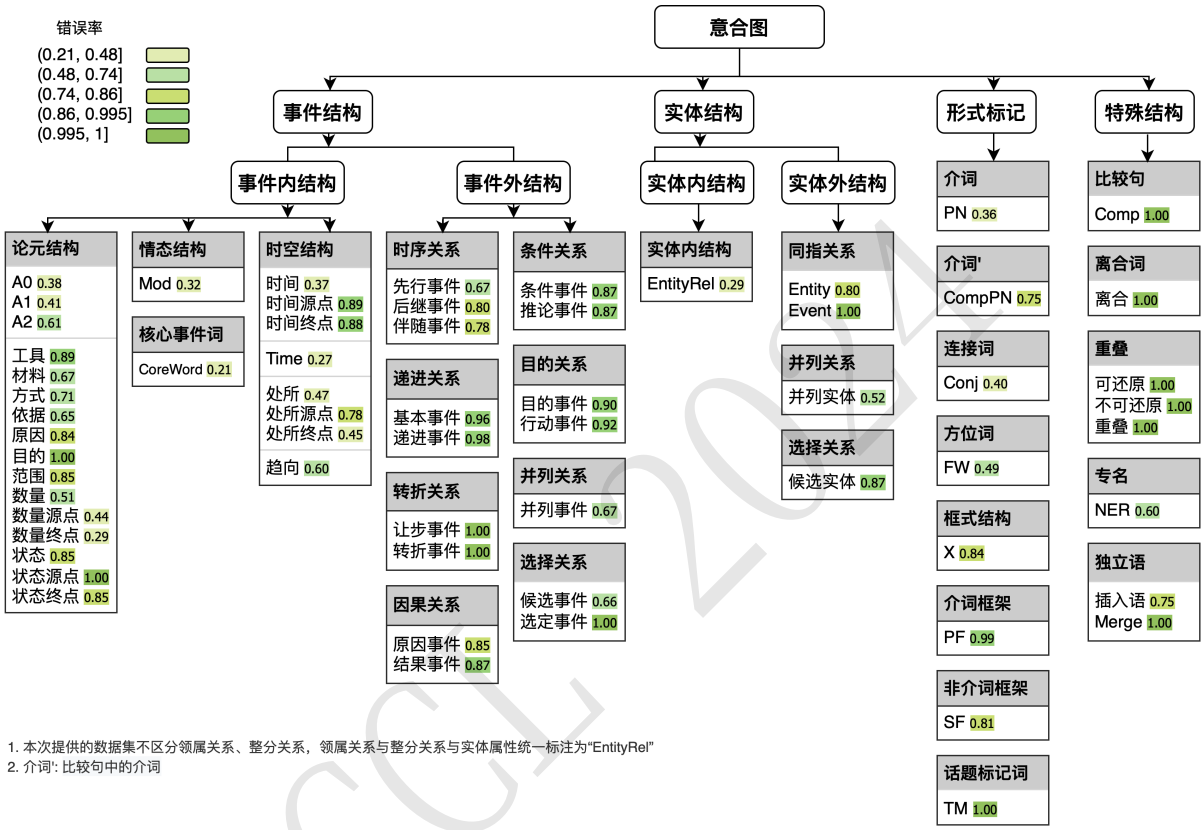


图 3: 全语义标签微调模型在训练集上各语义标签的错误率

3.2 语义标签分类微调

在本节中，我们针对易预测(l_easy)与难预测(l_hard)两种语义标签类别，将整个数据集划分为简单集与困难集两部分，简单数据集中仅包含标签类别为" l_easy"的三元组，而困难数据集则仅包含标签类别为"l_hard"的三元组。我们分别在这两部分数据集进行了Llama3-Chinese-8B-Instruct模型的监督微调。为了得到更优的语义解析结果，我们对不同微调模型的预测结果按照语义标签类别进行了组合，实验结果如表3所示。

由实验结果可知，难预测的语义标签预测结果的F1值非常低，进一步探究其原因，我们发现一些语义标签（如"Merge" "状态源点" "可还原" "目的" "离合" "重叠" "不宜还原" "TM" "选定事件"）在训练集中只有不到25条数据，少者甚至只有1条。这些语义标签由于训练数据的缺乏，导致训练不充分，从而对模型的理解和预测造成困难。此外，一些语义标签的训练数据并不少（如"行动事件" "目的事件" "结果事件"等），但其预测准确率依然很低。通过对这些语义标签的更进一步分析，我们发现，标注为这些语义标签的两个词中往往有一个词为"隐性事件词"。如在下面例句中，"出去"和"目的关系"标注为"行动事件"，其中"目的关系"则是"隐性事

模型	语义标签	F1值
Total	L_total	0.625
	L_easy	0.606
	L_hard	0.109
Easy	L_easy	0.629
Hard	L_hard	0.074
Easy	L_easy	0.625
Hard	L_hard	
Total	L_hard	**0.646**
Easy	L_easy	
Total	L_easy	0.604
Hard	L_hard	

表 3: 不同微调模型的预测结果。其中，Total为全语义标签微调的模型，Easy为在简单数据集上训练的模型，Hard为在困难数据集上训练的模型；L_total为全部语义标签的集合，L_easy为属于"*l_easy*"类别的语义标签集合，L_hard为属于"*l_hard*"类别的语义标签集合。

件词"，若要成功预测生成该类语义标签下的三元组，则须先将"隐性事件词"识别并抽取出来。因此，我们认为这些语义标签对大模型而言较难学习和掌握。

例句：

['老王','退休','以后','觉得','生活','没有','意思',',',',','所以','妻子','让','他','出去','找','点儿','事','干','。']

{'word1': 'word': '出去', 'idx': 12,'word2': 'word': '目的关系', 'idx': -10,'relVal': '行动事件'}

我们还发现，使用简单数据集进行微调有助于提高易预测语义标签的预测性能，然而，使用困难数据集微调反而降低了难预测语义标签预测的F1值，我们推测，这可能是因为模型在预测其他语义标签时学到了意合图语义解析的一般规律，这对预测难预测语义标签是有益的。

4 结语

在本次评测任务中，我们利用SDE进行了输入样本格式设计，并在此基础上对开源大模型进行了微调，以解决中文意合图语义解析任务。实验结果表明，使用中文语料指令微调的英文大模型解决该任务的能力逊于中文大模型，并且大模型对不同语义标签的预测能力存在显著差异。最终，我们的参赛系统在评测集上F1值达到了0.6461，在该任务中取得了第三名的成绩。然而，该任务仍存在一些需要进一步研究的问题。未来，我们将更加深入地分析大模型在不同标签上预测能力不同的原因，并将继续探索如何提高大模型在难预测语义标签上的预测性能。

参考文献

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. *In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with Chinese AMRs. *In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

(c) 2024 中国中文信息学会计算语言学专业委员会

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* , volume 1, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR Parsing as Sequence-to-Graph Transduction. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. Boosting Transition-based AMR Parsing with Refined Actions and Auxiliary Analyzers. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

Miguel Ballesteros, and Yaser Al-Onaizan. 2017. AMR Parsing using Stack-LSTMs. *ArXiv, abs/1707.07755*.

Ramo n Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodget, and Radu Florian. 2020. Transition-based Parsing with Stack-Transformers. *ArXiv, abs/2010.10669*.

Guntis Barzdins, and Didzis Gosko. 2016. RIGA at SemEval-2016 Task 8:Impact of smatch extensions and character-level neural translation on AMR parsing accuracy. *In Proceedings of International Workshop on Semantic Evaluations (SemEval)*, pages 1143-1147, San Diego, USA. Association for Computer Linguistics.

Xiaochang Peng, Chuan Wang, Daniel Gildea, and Nianwen Xue. 2017. Addressing the data sparsity issue in neural AMR parsing. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 366–375, Valencia, Spain. Association for Computational Linguistics.

Rik van Noord, and Johan Bos. 2017. Neural Semantic Parsing by Character-based Translation: Experiments with Abstract Meaning Representations. *ArXiv, abs/1705.09980*.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 146-157.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2020. Improving AMR parsing with sequence-to-sequence pre-training. *In Proceedings of the EMNLP*, pages 2501-2511.

Yifei Yang, Ziming Cheng , and Hai Zhao. 2023. CCL23-Eval任务2系统报告：基于大型语语言模型的中文抽象语义表示解析.

高文炀, 白雪峰, and 张岳. 2023. CCL23-Eval任务2系统报告：WestlakeNLP，基于生成式大语言模型的中文抽象语义表示解析.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv, abs/2106.09685*.

Biyang Guo, He Wang, Wenyilin Xiao, Hong Chen, Zhuxin Lee, Songqiao Han, and Hailiang Huang. 2024. Sample Design Engineering: An Empirical Study of What Makes Good Downstream Fine-Tuning Samples for LLMs. *ArXiv, abs/2404.13033*.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *ArXiv, abs/2305.13245*.

郭梦溪，荀恩东，李梦，and 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会

CCL 2024

CCL24-Eval任务2总结报告：中文意合图语义解析评测

郭梦溪[1]，李梦[1]，靳泽莹[1]，吴晓靖[1]，饶高琦[2]，唐共波[1]，荀恩东[3*]
[1]北京语言大学 信息科学学院
[2]北京语言大学 国际中文教育研究院
[3]北京语言大学 语言资源高精尖中心
guo_mengxi@foxmail.com

摘要

中文意合图是近年提出的中文语义表示方法。本次评测是首次基于意合图理论的语义分析评测，旨在探索面向意合图理论的语义计算方法，评估机器的语义分析能力。本次评测共有14支队伍报名，最终有7支队伍提交结果，其中有5支队伍提交技术报告与模型，均成功复现。在评测截止时间内，表现最好的队伍使用大语言模型LoRA微调方法获得了F1值为72.06%的成绩。在最终提交技术报告的5支队伍中，有4支队伍使用了大语言模型微调方法，在一定程度上表明了目前技术发展的趋势。

关键词： 意合图；语义分析；通用语义框架；评测任务

Overview of CCL24-Eval Task2: Chinese Parataxis Graph Parsing Evaluation

Mengxi Guo[1],Meng Li[1], Zeying Jin[1],Xiaojing Wu[1],
Gaoqi Rao[2],Gongbo Tang[1],Endong Xun[3*]
[1]School of Information Science, Beijing Language and Culture University
[2]Research Institute of International Chinese Language Education,
Beijing Language and Culture University
[3]Beijing Advanced innovation Center for language Resources,
Beijing Language and Culture University
guo_mengxi@foxmail.com

Abstract

The Chinese Parataxis Graph is a recently proposed method for Chinese semantic representation. This evaluation is the first semantic analysis evaluation based on the theory of the Chinese Parataxis Graph. It aims to explore semantic computation methods oriented towards the Chinese Parataxis Graph theory and to evaluate the semantic analysis capabilities of machines.A total of 14 teams registered for this evaluation, with 7 teams submitting results. Among them, 5 teams submitted technical reports and models, all successfully replicated the results. By the evaluation deadline, the best-performing team, which used the LoRA fine-tuning method with a large language model, achieved an F1 score of 72.06%. Among the 5 teams that submitted technical reports, 4 teams used the fine-tuning method with large language models, indicating a current trend in technological development to some extent.

Keywords: Chinese Parataxis Graph , Semantic parsing , Evaluation , Universal semantic framework

* 通讯作者

(c) 2024 中国中文信息学会计算语言学专业委员会

1 评测背景

随着自然语言处理技术的不断发展，语义分析作为语言理解的重要组成部分，受到了广泛的关注和研究。意合图是荀恩东近年来提出的一种以事件为中心的语义表示方法，采用单根有向图的形式承载事件、实体、属性及其相互关系。过往的语义表示多是在不同层级单位进行相应的表示，如词、句子、段落、篇章等不同层次的语言单元都有不同的语义表示方法。而意合图力求通过统一的表示框架，对不同层级的语言单元进行一致表示。

意合图理论经历了早期理论架构 (荀恩东, 2023)，以及基于早期理论架构的工程实践 (王诚文, 2021; 王贵荣, 2023)。通过工程实践我们优化并完善了的意合图理论架构，构建了完整的意合图通用语义体系 (郭梦溪等, 2024)，并基于意合图理论与通用语义体系构建了一批意合图语义标注资源 (郭梦溪等, 2024)。为了探索意合图的最佳计算方法，并评估当前机器的语义分析能力，我们组织了本次中文意合图语义解析评测。

2 评测任务

2.1 相关概念

意合图将通过语言所表征的事件定义为两种，一种是现实世界或可能世界中的事物的动作行为或关系描述，另一种是现实世界或可能世界中的事物间关系，相对应地，意合图将第一种事件称为意合图所表征的一般事件，第二种事件称为意合图所表征的关系事件。意合图将事件词作为事件的核心表达，其中一般事件的事件词常为在句中出现的连续或非连续谓词性语言单元，如汉语的离合词即为非连续事件词；如果事件词在句中省略情况或汉语特殊的名词谓语句，则对事件词进行补全；关系事件的事件词为抽象出的关系概念词，如“因果关系”“同指关系”等。

意合图在符合人类对语言认知的基础上，充分考虑落地应用的可操作性，使其尽可能地层次化，以便于后续语义分析路径的设计，实现通用性与扩展性兼具的语义表征方案。按照层次可将意合图分为事件结构与实体结构两大部分。事件结构分为事件内事件内结构与事件外结构，事件内结构可进一步分为以事件词为核心的论元结构、情态结构、时空结构，事件外结构为多个事件构成的关系事件结构；实体结构分为实体内结构与实体外结构，实体内结构即实体属性与属性值结构，实体外结构即多个实体构成的实体关系事件结构。本次评测所发布的标注语料的语义标签体系如表1所示，关系论元与关系事件具有对应性，单独于表2展示。其抽象表示如图1所示。

层级	包含的语义标签类
论元结构 （一般论元）	A0, A1, A2, 工具, 材料, 方式, 依据, 原因, 目的, 范围, 数量, 数量源点, 数量终点, 状态, 状态源点, 状态终点 （关系事件论元见表2）
情态结构	Mod（此次评测不对情态内部细分）
时空结构	时间, 时间源点, 时间终点,Time（此次评测不对时态时制等时间信息细分）； 处所, 处所源点, 处所终点, 趋向
实体属性	EntityRel（此次评测不细分实体属性）
特殊标签	Comp，NER，插入语，Merge，离合，重叠，不宜还原，宜还原
形式标记	PN，CompPN，Conj，FW，TM，PF，SF, X

Table 1: 数据集语义标签（部分）

关系事件（词）	对应的关系论元
时序关系	先行事件，后继事件，伴随事件
递进关系	基本事件，递进事件

Table 2: 关系事件及关系论元

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

Table 2 – 续

关系事件（词）	对应的关系论元
转折关系	让步事件，转折事件
因果关系	原因事件，结果事件
条件关系	条件事件，关系事件
目的关系	目的事件，行动事件
并列关系（And）	并列事件，并列实体
选择关系（Or）	候选事件，选定事件，候选实体
同指关系（Ref）	Entity，Event
领属关系	（此次评测不区分领属关系）
整分关系	（此次评测不区分整分关系）

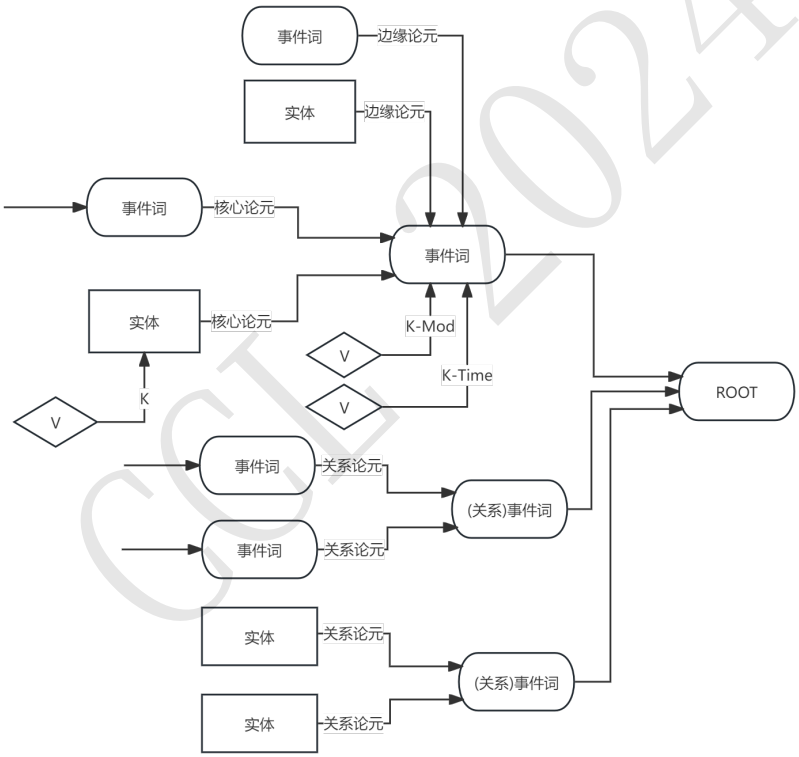


Figure 1: 意合图的抽象表示

2.2 任务要求

本次中文意合图语义解析评测任务仅要求生成句子级意合图框架即可，即输入单元为分词后的句子，输出为意合图框架结构，无需生成细化实体结构、情态结构、时空结构等的内部语义分类，仅判断是否属于该结构成分即可，所提供的语料也为粗粒度标签。此外，本次评测任务的参赛队伍可自行借助形式标记辅助个别语义的识别，但最终不要求对形式标记进行识别，不在结果中进行计算。

(c) 2024 中国中文信息学会计算语言学专业委员会

3 评测数据

3.1 数据分布

本次评测任务的数据集语料抽取自BCC（Beijing Language and Culture University Corpus Center）(荀恩东等, 2016)国际中文教育阅读语料库与中文宾州树库8.0的新闻语料，共计5000条标注语料。数据分布如表3所示。

	国际中文教育阅读语料	新闻语料	总计
训练集	701	1299	2000
验证集	351	649	1000
测试集	351	649	1000
盲测集	351	649	1000
总计	1754	3246	5000

Table 3: 数据分布

需要说明的是，本次评测任务允许参赛队伍根据需求对除盲测集外的数据集分布进行重分配。盲测集仅提供给参赛队伍分词后的输入，由参赛队伍进行结果推理，将预测结果返回。最终根据盲测集结果进行排名。

3.2 数据标注

本次评测的数据集由8位具有语言学背景的研究生在标注规范的指导下完成标注。每次任务先由随机两人进行独立标注，然后双方再根据管理者返回的不一致标注结果进行讨论，确定唯一标注结果。无法达成一致的情况，由管理者介入进行确认。最后管理者对标注结果进行全检，再次修正错误。因此，每条语料经过多次确认，以保障标注数据的质量。且数据集中的每条语料的标注结果均经验证，能够生成完整意合图，即标注结果中不存在游离成分或违反意合图原则的情况。

3.3 数据格式

除盲测集外，其他发布给参赛队伍的数据集文件均为UTF8编码，Json格式，包括句子的分词信息和标注三元组。如样例（图2）所示，“sent”值域是句子分词结果，“relData”值域是该句子的所有标注信息，其中“word1”和“word2”值域均包含节点内容word和节点编号idx，当word为句中词汇时，idx为该词在句中的编号（从0开始），当word为隐式事件词或实体省略标签时，idx的对应关系如表4所示。需要注意的是，当句中存在不止一个某一种隐式事件时，word对应的内容将在字符串后面增加数字，并向上叠加，idx也将在-13的基础上减1。例如，句中存在三个因果关系，则其对应的word分别为：因果关系、因果关系1、因果关系2，idx分别为：-8、-14、-15。


```
[
  {"sent":["我","对不起","大家",",","我","没有","完成","任务","。"],
"relData":[
{"word1":{"word":"我","idx":0},"word2":{"word":"对不起","idx":1},"relVal":"A0"},
{"word1":{"word":"对不起","idx":1},"word2":{"word":"ROOT","idx":-1},"relVal":"CoreWord"},
{"word1":{"word":"对不起","idx":1},"word2":{"word":"因果关系","idx":-8},"relVal":"结果事件"},
{"word1":{"word":"大家","idx":2},"word2":{"word":"对不起","idx":1},"relVal":"A1"},
{"word1":{"word":"我","idx":4},"word2":{"word":"完成","idx":6},"relVal":"A0"},
{"word1":{"word":"没有","idx":5},"word2":{"word":"完成","idx":6},"relVal":"m否定"},
{"word1":{"word":"完成","idx":6},"word2":{"word":"ROOT","idx":-1},"relVal":"CoreWord"},
{"word1":{"word":"完成","idx":6},"word2":{"word":"因果关系","idx":-8},"relVal":"原因事件"},
{"word1":{"word":"任务","idx":7},"word2":{"word":"完成","idx":6},"relVal":"A1"},]}
]
```


Figure 2: 数据格式

(c) 2024 中国中文信息学会计算语言学专业委员会

隐式事件词	ROOT	And	Or	Ref	时序关系	递进关系	转折关系	因果关系	条件关系	目的关系	重叠	Is	QS
idx	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13

Table 4: 隐式事件词索引

4 评价指标

本次评测采用F1值作为模型表现的评价标准与排名的主要依据，计算方式如下：

$$P = \frac{\text{count(Matching Tuples)}}{\text{count(Generated Tuples)}} \tag{1}$$

$$R = \frac{\text{count(Matching Tuples)}}{\text{count(Gold Tuples)}} \tag{2}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

其中“Generated Tuples”为模型预测的三元组集合数，“Gold Tuples”为测试集/盲测集的三元组集合数，“Matching Tuples”为模型预测的三元组集合与测试集/盲测集的三元组集合间的最大匹配个数。

5 评测结果

5.1 提交结果

本次评测共有14支队伍报名参赛，包括北京大学、北京理工大学、北京语言大学、郑州大学、湘潭大学、南京理工大学、沈阳航空航天大学等高校，中科院计算所等研究所，中国太平洋保险集团、中电信人工智能科技有限公司、北京恺望数据科技有限公司等企业。最终有7支队伍提交了结果，其中有5支队伍提交技术报告，并成功复现其模型。由于本次评测中大多参赛队伍使用了大模型，无法保证每次复现结果完全一致，因此我们接受复现结果与提交结果存在小幅度差异，如复现没有问题，最终排名仍以参赛队伍提交的结果为准。在评测结果提交的规定时间内所提交的盲测集结果如表5所示，来自北京理工大学的队伍1所提交的结果F1值最高，为72.06%。

队伍编号	队伍单位	**Precision(%)**	**Recall(%)**	**F1-score(%)**
1	北京理工大学	70.11	**74.10**	**72.06**
2	北京理工大学	**70.28**	73.18	71.70
3	北京语言大学	62.65	66.70	64.61
4	北京语言大学	64.22	64.66	64.44
5	中国太平洋保险公司	57.80	61.95	59.80
6	湘潭大学	57.89	58.53	58.21
7	沈阳航空航天大学	55.89	57.51	56.69

Table 5: 盲测集结果排名

5.2 方法概述

5.2.1 基于大语言模型的微调方法

随着大语言模型的不断发展，其参数量不断增大，模型的泛化能力不断增强，微调方法便捷。提交技术报告的5支队伍中，有4支队伍使用了大语言模型微调方法。所使用的基座模型涉及闭源模型百度文心系列的ERNIE4.0-Speed-8K，开源模型Llama3-Chinese-8B-Instruct、Chinese-Llama2-7B-Chat，且均使用了LoRA微调方法。

来自北京理工大学的队伍1和队伍2均使用了百度文心系列的ERNIE4.0-Speed-8K闭源模型作为基座模型在百度智能云的千帆大模型平台进行微调。二者的微调方法不同，队伍1使用

(c) 2024 中国中文信息学会计算语言学专业委员会

了LoRA微调方法，队伍2使用了全参数微调。在Prompt-Response的设计上，二者的Prompt设计类似，输入均为分词后的句子，无意合图概念性描述。Response的设计上，队伍1保留了索引编号，而队伍2未直接输出索引编号。在训练集条数上，队伍1自行分配了2850条语料，队伍2分配了3600句语料。在超参数设置上，两个队伍在迭代轮次、学习率、学习率调整计划等设置上存在差异。最终两个队伍所提交的盲测集预测结果的F1值较为接近，分别为72.06%、71.70%。在结果提交截止后，队伍2又将训练语料条数增加至4000句，在参数上进行少量调整，使得模型盲测集预测结果的F1值提升至74.76%（该结果经过复现核对），并且又进一步探讨了使用小参数规模的模型达到比赛中所用的闭源模型在这一任务上的表现的可行性。该队伍集成了包括Qwen1.5-7B，ChatGLM3-6B，Yi1.5-6B，DeepSeek-7B在内的国内开源模型的轻量级版本，设计了一种循环增强微调训练模块，并采用一种级联监督的方式融合模型的输出结果。这种方法成功地实现结合小规模参数的开源模型达到与百度文心系列的ERNIE-Speed-8K模型相近的效果。

来自北京语言大学的队伍3使用了开源模型Llama3-Chinese-8B-Instruct在本地进行LoRA微调。该参赛队伍对数据集分布进行了统计，根据统计结果进行了样本设计工程（Sample Design Engineering，SDE），将上下文、指令及输出指示放置于输入的任务文本之前以提升模型的任务理解能力，并且根据意合图标签的出现频次，由高到低排序以提高模型对出现次数多的标签的关注程度。该队伍通过实验发现大模型对不同语义标签解析难度不同，并设计了不同的模型训练策略，对解析结果进行了组合分析。最终该队伍所提交的盲测集预测结果的F1值为64.61%。

来自湘潭大学的队伍6使用Chinese-Llama2-7B-Chat在阿里云人工智能平台PAI进行LoRA微调。最终所提交的盲测集预测结果的F1值为58.21%。

5.2.2 基于roBERTa的关系抽取方法

来自北京语言大学的队伍4将意合图语义解析任务转换为传统的关系抽取任务。而意合图与传统关系抽取任务相比，其三元组内词的顺序不可变，且存在句外词，即意合图的隐式事件词、根节点“ROOT”、实体省略标签“QS”等。对此，该队伍将三元组内不符合原句语序的“实体对”的关系改为“关系标签_reverse”，以解决词对顺序不可变问题；将句外词添加在原句末尾作为输入，以此解决句外词问题。通过上述处理，将意合图语义解析任务转变为了关系抽取任务。但该处理方式也使得原本就不平衡的标签分布加重，因此该参赛队伍将任务划分为两个子任务，即不包含隐式事件词的关系抽取和包含隐式事件词的关系抽取。该参赛队伍将任务分为关系识别与关系分为两部分，均使用了哈工大版本的chinese-roBERTa-wwm-extlarge模型。最终该参赛队伍取得了F1值为64.44%的成绩。

5.3 其他分析

本次评测中有些参赛队伍不仅完成了评测，还在过程中进行了对比实验、结果分析等，对意合图的进一步研究提供了参考。

来自北京理工大学的队伍1对六个开源大模型进行了测试，Yi-1.5-9B在测试集上的F1值为60.21%，其余五个在测试集上的F1值均不足60%，而ERNIE-Speed在同一份测试集上的F1值达69.56%。为进一步探究各种因素对于模型性能影响的程度，该队伍对参数规模和基座模型系列两个因素的影响进行探究。实验结果表明，同属Qwen-1.5系列的四种参数规模的模型在该任务上的表现基本一致，并没有随着参数量的增加而在该任务上表现更优越。其次，参赛队伍选用Baichuan2-7B、Qwen-1.5-7B和Yi-1.5-9B三个参数规模类似的不同系列的模型，采用完全相同的超参数和工具进行微调。实验结果表明，三者表现差异较大，说明了基座模型对于该任务的影响较大。通过该实验，参赛队伍得出其所使用的文心系列ERNIE4.0-Speed-8K模型在该任务中的出色表现与百度的预训练语料、模型的技术细节以及千帆平台的微调实现等因素更为相关。

来自北京语言大学的队伍3对模型在训练集上各语义标签的错误率进行排序，其中错误率最低的语义标签为“CoreWord”，即核心事件词的判断；错误率最高的语义标签为“选定事件”。对易预测错误的语义标签进行分析，发现一些语义标签错误率高可能是由于训练集中该类数据过少，也存在一些语义标签在训练集中数据不少，但大模型仍难掌握。此外，该队伍还在相同实验条件下微调了六个同等规模的开源大模型，尤其关注了中文大模型与中文语料增量预训练的英文大模型。实验结果显示，在该任务下中文大模型展现出更高的适应性与优越性。

(c) 2024 中国中文信息学会计算语言学专业委员会

6 总结

本次评测是意合图正式提出后首次参与评测活动，共吸引了来自高校、研究院以及企业的14支队伍报名，最终有7支队伍提交了结果，其中来自北京理工大学的队伍以F1值达72.06%的成绩取得了本次评测的第一名。在提交技术报告的5支队伍中，有4支队伍使用了大模型微调的方法。该情况充分展现了大模型微调已成为当前技术研究与应用中的主流选择，更多团队倾向于采用更为高效的微调方法来解决各种任务。

参赛队伍的实验结果为我们的研究提供了宝贵的参考经验。基于本次评测所反映出的情况，我们将在未来的研究中，探索基于更优秀的基座模型和更精细的微调工程，获得更优异的分析效果。为典型的语义关系挖掘提供更多表达方式作为备选资源，使得能够更大程度地提高标签精度，为更精准的语义分析提供支持。在资源建设方面，评测中所呈现的方法将助力我们构建更高质量的意合图语义资源，更高效地推动意合图在各类应用场景中的实践。我们将不断优化和改进意合图的构建和应用方法，期待在语义分析领域取得更大的突破，推动语义分析的发展。

致谢

本研究得到国家自然科学基金“中文意合图的表征与生成方法研究”（62076038）与中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）的支持。本次评测所使用的数据集由北京语言大学李梦、何晴、胡星雨、王静怡、吴晓靖、张可芯、周书帆、朱奕瑾（按姓氏排）八位研究生完成标注，感谢各位标注员所作出的贡献。数据集标注所使用的在线标注平台最初由张梦圆构建，于钟洋、宋玉良完成了新功能的实现，感谢三位研究生对数据集构建所作出的贡献。

参考文献

Liyang Pang, Chengwen Wang, Guirong Wang, Gaoqi Rao, and Endong Xun. 2021. *Prepositional Frame Extraction and Semantic Classification Based on Chinese ChunkBank.* CLSW, Nanjing.

Shufan Zhou , Chengwen Wang, Endong Xun. 2023. *Recognition of Disyllabic Intransitive Verbs and Study on Disyllabic Intransitive Verbs Taking Objects Based on Structure Retrieval.* Springer Nature Switzerland, 2023: 265–282.

郭梦溪, 荀恩东, 李梦, 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

郭梦溪, 李梦, 荀恩东, 饶高琦, 于钟洋. 2024. 基于意合图语义理论的结构标注体系与资源建设. 第二十三届中国计算语言学大会.

邵田, 翟世权, 饶高琦, 荀恩东. 2023. 基于结构树库的状位动词语义分类及搭配库构建. 中文信息学报,37(06):44-51+66.

田思雨, 邵田, 荀恩东, 饶高琦. 2023. 基于结构树库的补语位形容词语义分析及搭配构建. 第二十二届中国计算语言学大会论文集,第420页-第432页,哈尔滨.

王诚文, 钱青青, 荀恩东, 邢丹, 李梦, 饶高琦. 2020. 三元搭配视角下的汉语动词语义角色知识库构建. 中文信息学报,34(09):19-27.

王诚文. 2021. 面向意合图的汉语动词论知识构建研究. 北京语言大学博士论文.

王贵荣. 2023. 意合图事件结构标注及分析研究. 北京语言大学博士论文.

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

荀恩东. 2023. 自然语言结构计算：*BCC*语料库. 人民邮电出版社, 北京.

荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制. 语料库语言学.

(c) 2024 中国中文信息学会计算语言学专业委员会

CCL24-Eval任务3系统报告：基于参数高效微调与半监督学习的空间语义理解

李晨阳[1,2]，张龙[1,2]，郑秋生[1,2]
[1]中原工学院 前沿信息技术研究院，河南 郑州 450007
[2]河南省网络舆情监测与智能分析重点实验室，河南 郑州 450007
2312826399@qq.com

摘要

本文介绍了我们在第二十三届中文计算语言大会的第四届中文空间语义理解评测任务中提交的参赛模型。该任务旨在测试机器的中文语义理解水平。现有研究显示，机器的中文语义理解水平与人类平均水平相比仍有较大差距。近年来，生成式大规模语言模型在自然语言处理任务中展现了出色的生成和泛化能力。在本次评测中，我们采用了对Qwen1.5-7b模型进行高效微调的方法，以端到端的形式实现空间语义的推理过程，并结合prompt优化和半监督学习提升推理表现。实验结果表明，我们的模型在该任务中取得了领先的效果。

关键词： 大语言模型 ； 高效微调 ； 半监督学习 ； prompt

System Report for CCL24-Eval Task 3: Spatial Semantic Understanding Based on Parameter-efficient Fine-tuning and Semi-supervised Learning

Chenyang Li[1,2], Long Zhang[1,2],Qiusheng Zheng[1,2]
[1]Frontier Information Technology Research Institute,
Zhongyuan University of Technology, Zhengzhou 450007 China
[2]Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou China
2312826399@qq.com

Abstract

This paper introduces the models we submitted for the Fourth Chinese Spatial Semantic Understanding Evaluation Task at the 23rd Chinese Computational Linguistics Conference. This task aims to test the level of machine understanding of Chinese semantics. Existing research indicates that the level of machine understanding of Chinese semantics still lags significantly behind the average human level. In recent years, generative large-scale language models have demonstrated excellent generative and generalization capabilities in natural language processing tasks. For this evaluation, we employed an efficient fine-tuning method on the Qwen1.5-7b model to achieve the spatial semantic reasoning process in an end-to-end manner. We also combined prompt optimization and semi-supervised learning to enhance reasoning performance. Experimental results show that our model achieved leading performance in this task.

Keywords: Large Language Model , Efficient Fine-Tuning , Semi-Supervised Learning , prompt

1 引言

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

空间表达是自然语言中常见的现象，用来描绘物体之间的空间位置关系。空间范畴是人类认知总重要的基础范畴，大量空间信息存在于自然语言文本中。在通往人工智能的道路上，空间语义理解是不可绕开的关键步骤。著名认知语言学家Jackendoff(2004)在其概念语义学理论中指出，空间结构是语言系统的四种基本结构之一。理解文本中的空间表达语义，除了语言知识外，还需借助空间认知能力来构建空间场景，并依据世界知识进行有关空间方位信息的推理。通常认为，对文本中空间信息的理解，不仅需要掌握句段中词汇、句法语义知识，还需要具备一定的常识或背景知识，甚至是超出语言范畴的空间想象等认知能力，以此来构建空间场景。

在这样的背景下，第二十三届中国计算语言学大会发布了SpaCE2024技术评测。与之前相比，SpaCE2024不再划分子任务，而是以选择题的形式考察五个层次的空间语义理解能力：空间信息实体识别，要求从四个选项中选出文本信息的参照物。空间信息角色识别。要求从四个选项中选出文本信息的语义角色，或者选出与语义角色相对应的空间表达形式。空间信息异常识别，要求从四个选项中选出文本空间信息异常的语言表达。空间方位信息推理。要求基于文本给出的推理条件进行空间方位推理，从四个选项中选出推理结果。空间异形同义识别。要求从四个选项中选出能使两个文本异形同义或异义的空间义词语。

近期，随着预训练语言模型的参数量和数据量不断增大，大规模语言模型在人工智能领域取得了革命性进展，使得一些过去被视为仅限于人类能力范围的自然语言处理任务变得可行。相较于传统预训练模型，大模型能够记忆更广泛的世界知识，关注到更多的信息，并展现出传统模型不具备的涌现能力。通过指令微调和代码预训练等技术，大模型具备了理解人类指令、泛化到新任务、代码理解与生成、利用思维链推理以及处理长距离依赖等多项能力。基于此，我们采用了阿里发布的中文大模型Qwen1.5-7b进行端到端微调，以实现空间语义的推理过程。并结合prompt优化和由半监督学习产生的伪标签来提升推理表现。

2 相关工作

为了评测机器的空间语义理解能力，自然语言处理领域的评测任务主要分为以下三类：1.空间信息标注任务，要求机器根据给定的语义角色标注文本中的空间实体和空间关系，形式上与语义角色标注任务和时间抽取任务相似，代表性工作有SpRL任务(Kolomiyets et al., 2013)和SpaceEval任务(Pustejovsky et al., 2015)。2.空间关系推理任务，要求机器根据文本中已有的空间信息回答涉及空间关系推理的问题，代表性工作又bAbI任务集(Weston et al., 2015)中的位置推理任务和路径推理任务，以及SpartQA任务(Mirzaee et al., 2021)。3.空间语义异常判断任务，要求机器判断文本是否存在空间信息异常以及异常的归隐类型，詹卫东等(2022)首次提出该任务，认为如果机器能够识别错误的空间信息并进行正确的归因，就说明机器具有一定的空间语义理解能力。

空间关系提取任务可以分为传统的机器学习方法和神经网络方法。前者高度依赖于手动特征或显式句法结构。Nichols等(2015)提出了一种基于筛选的模型，它使用多层来提取空间元素，然后引入分类器来分类空间关系。D'Souza等(2015)提出了一种基于筛选的模型，通过贪心的特征选择技术来生成各种手动特征。还有研究人员(Salaberri et al., 2015)引入外部知识作为空间信息的补充，在此过程中，提供了许多空间元素的信息。Kim等(2016) 提出了一种韩语空间关系提取模型，使用依赖关系来找到适合角色的元素。

随着神经网络的广泛应用，Ramrakhiyani等(2019)提出了一种通过依存句法生成候选关系，并使用BiLSTM 模型对候选关系进行分类的方法。Shin等(2020)提出了一种使用CRF 进行空间关系分类的模型，Wu等(2019)则提出了基于AR-BERT 的方法来提取空间关系。此外，一些研究关注于多模态空间关系提取。例如，Dan等(2020)提出了一种空间BERT，它同时从文本以及包含实体的图片中来预测实体之间的空间关系。

3 实现方法

如图1所示，我们先对文本内容进行格式化处理，然后对主流大模型采用qlora方法(Dettmers et al., 2024)进行端到端微调，实现语义空间的推理过程。接着，选出效果较好的模型作为基线模型，最后采用prompt优化和半监督学习提升推理表现。

3.1 数据预处理

为了使得本评测任务的数据集能直接应用于大模型微调，在数据预处理阶段，我们编写

(c) 2024 中国中文信息学会计算语言学专业委员会

图 1: 模型的数据预处理、预测以及后处理过程

了代码对原始数据集进行了处理。如图2所示，首先识别原数据的qid字段，并将识别后的能力代号、题目类别加入到文本序列中，各项信息用“#”进行分隔，正确答案也作为一个序列用“#”进行分隔。我们还尝试了除此之外的其他prompt模板，以优化模型输入格式。


```
{
  "qid": "4-dev-m-1079",
  "text": "赵云、曹操、关羽、孙坚四人来到火锅店吃火锅，选了四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人。面对面而坐。已知：赵云在关羽右边坐，曹操在孙坚同侧左边。",
  "question": "关羽和()不是并排坐的。",
  "option": {
    "A": "孙坚",
    "B": "赵云",
    "C": "曹操",
    "D": "以上选项都不是"
  },
  "answer": [
    "A",
    "C"
  ]
}
```

```
{
  "messages": [
    {
      "role": "user",
      "content": "赵云、曹操、关羽、孙坚四人来到火锅店吃火锅，选了四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人。面对面而坐。已知：赵云在关羽右边坐，曹操在孙坚同侧左边。#空间推理#多选题#关羽和()不是并排坐的。#A:孙坚 B:赵云 C:曹操 D:以上选项都不是"
    },
    {
      "role": "assistant",
      "content": "A#C"
    }
  ]
}
```


图 2: 左为原格式数据，右为拼接后格式化数据

3.2 大模型微调

Qwen是一个全能的语言模型系列(Bai et al., 2023)，其使用了高达3万亿个token的数据进行预训练，涵盖多个类型、领域和任务，不仅包括基本的语言能力，还包括算术、编码和逻辑推理等高级技能。同时使用了复杂的流程进行数据清晰和质量控制。

在微调阶段，我们采用了qlora方法，这是一种高效的微调方法，可减少内存使用量。构造的输入数据包括任务指令和利用prompt拼接后的原句，标签则是对应的选项答案。我们把该任务视为序列生成任务，而不是直接生成符合本任务的数组形式，这样也是为了避免模型生成不符合格式的答案。我们发现，参数秩设置为64和96时效果最佳。如图3所示，针对于输入的问题，语言模型会将输入问题的上文作为参考进行信息的搜索，并以此来进行空间位置的推理，从而得出正确答案。

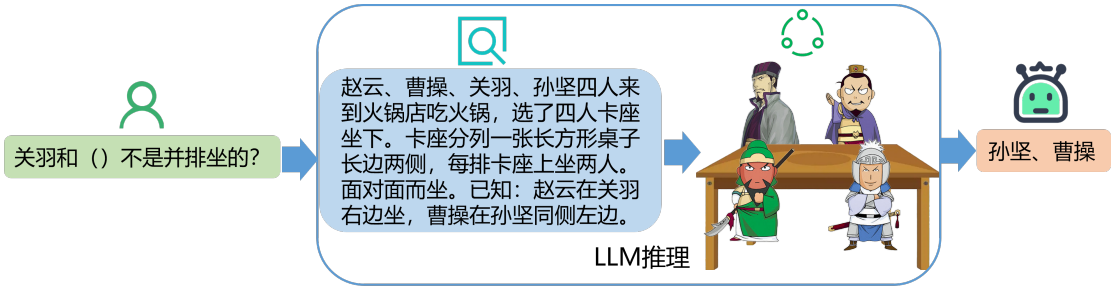


图 3: 大模型语义空间的推理过程

(c) 2024 中国中文信息学会计算语言学专业委员会

3.3 数据分析

为了进行后续分数的提升，我们对5种层次题型类别的数量和每个类别的准确率进行了统计，如图4所示，我们发现不论在训练集还是验证集中，都存在数据分布不平均，部分标签对应的数量较少的问题(Li et al., 2023)，如在训练集中，空间方位信息推理类型数量达到了1210条，其对应的准确率也较高。而空间异形同义识别类型的训练集只有5条，其对应的准确率也较低。为了提升该类的数据量，我们尝试采用增加数据集的方式来使得模型在训练时学习到更多的相关特征，我们首先采用同义替换、随机词插入等方法对该类的数据进行了数据增强，在模型上进行微调后，验证集分数有所提升，但在测试集上效果却较差，出现了过拟合的现象，我们猜测可能是数据增强引入了过多错误的噪声，进而破坏了数据原始的分布。于是我们采用了同样能增加训练集数量的伪标签方法。

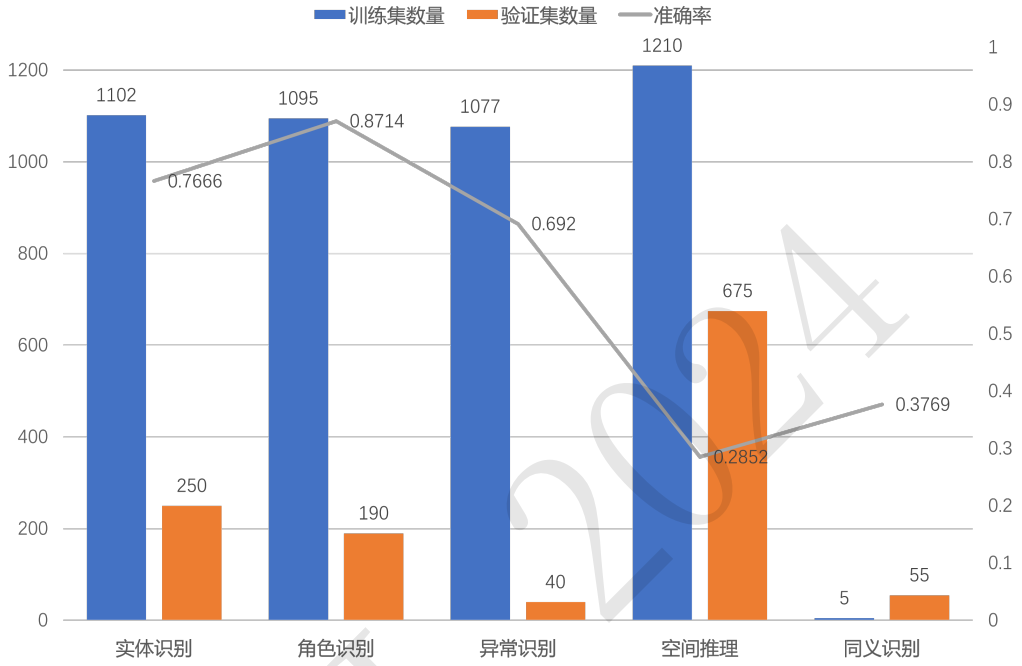


图 4: 大模型语义空间的推理过程

3.4 基于半监督学习的伪标签生成

伪标签方法来自于半监督学习，其核心思想是借助无标签的数据来提升有监督过程中的模型性能(Rizve et al., 2021; Berry et al., 2019)。由半监督学习生成伪标签的过程如图5所示，其主要是将模型对无标签的测试数据的预测结果加入到训练集中，从而增大数据量以提升模型效果。该方法适用于模型精度比较高的情况。由于我们未采用传统分类模型进行训练，从而无法得到预测类别的概率值。我们采用多个预测结果较高的文件，并取预测标签相同的部分作为伪标签加入到训练集中重新训练。经过统计，每次得到的伪标签数量都接近3k条，这表明我们的数据集在原本数量的基础上又增加了3k条数据。经过多轮的伪标签训练后，筛选出的伪标签会越来越接近，最终模型达到了拟合的状态，此时在进行后续的伪标签训练已经无法进一步再提升测试集的准确率。

3.5 模型融合

关于模型融合，周志华(2021)教授在其《机器学习》一书中提到：模型融合要好而不同，即模型差异性越大，融合效果越好。我们从两个方面来增加差异化，一是使用不同的多个模型，Qwen和ChatGLM4，二是重新划分训练集和验证集来改变模型输入。在使用这两个模型进行预测后，选择出每个层次类别中较好的部分进行结果的融合。

CCL 2024

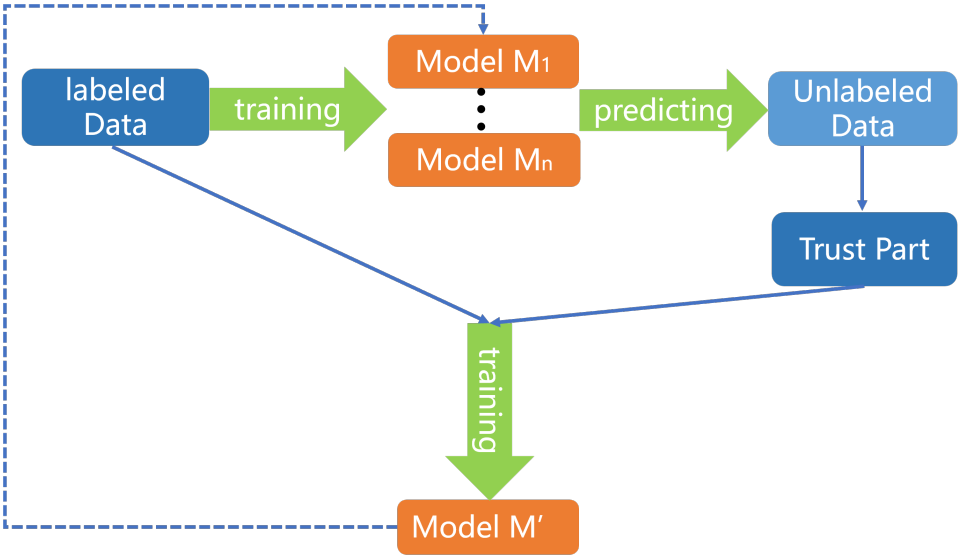


图 5: 基于半监督学习的伪标签生成

3.6 prompt模板构造

如图6所示，我们构造了两种prompt模板，并用不同颜色区分原数据的信息。尽管第二种模板比第一种更加流畅，但在实验中发现，在使用模型微调情况下，这两种prompt模板对结果的影响基本可以忽略不计，然而，由于第二种模板增加了一些额外的文本信息，导致其最大长度增加，从而增加了训练时的额外开销。相反，在调用大模型接口时，prompt的质量对结果影响非常显著。

他扶着移动的扶手，四下张望。右边有一个迷宫，里面满是叽叽喳喳叫嚷着的鹦鹉，左边是一家日杂品店，里面到处闪着铬的光芒。#实体识别#单选题#()右边有一个迷宫。#A:他 B:扶手 C:铬 D:鹦鹉"

请阅读以下文本并回答问题：他扶着移动的扶手，四下张望。右边有一个迷宫，里面满是叽叽喳喳叫嚷着的鹦鹉，左边是一家日杂品店，里面到处闪着铬的光芒。题型：实体识别单选题。问题：()右边有一个迷宫。括号中应该填入什么？选项：A.他 B.扶手 C.铬 D.鹦鹉 请选择正确答案。

图 6: 左为prompt模板1，右为prompt模板2

4 实验

4.1 实验结果

如表4.1所示，我们列出了采用ChatGLM3、Qwen-7b、Qwen1.5-7b和调用ChatGLM4接口的测试集结果。在选择基线模型的过程中，我们发现Qwen1.5-7b模型要明显优于其他模型。但为了更好的使用半监督学习来生成高质量的伪标签，我们同样也使用了除Qwen1.5以外的模型进行训练，以融合出更高质量的伪标签。随着Qwen模型逐轮加入伪标签数据后，模型预测的准确率不断提升。同时我们发现，在使用微调时，prompt模板的优劣对结果不会有太明显的影响。而在调用大模型接口时，prompt模板的质量对结果影响非常明显。最终，通过多轮伪标签的应用，我们模型预测准确率达到了0.5516的分数，再通过多个模型结果的融合，最终达到了0.566的分数。如表4.1所示，实验结果证明，我们的模型取得了较为先进的效果，并在本任务中取得第4名的成绩。

CCL 2024

模型	方法	准确率	实体识别	角色识别	异常识别	空间推理	同义识别
Baseline	微调	0.4792	0.7509	0.8818	0.6860	0.2196	0.4200
ChatGLM3-6b [0]	微调	0.4370	0.6720	0.7727	0.6320	0.2348	0.3185
Qwen-7b[1]	微调	0.5035	0.7667	0.8714	0.6920	0.2853	0.3769
	+1轮伪标签	0.5344	0.8175	0.9039	0.7580	0.3157	0.3630
	+2轮伪标签	0.5423	0.8316	0.9000	0.7920	0.3122	0.3954
	+3轮伪标签	0.5434	0.8421	0.9051	0.7980	0.3078	0.3970
Qwen1.5-7b[2]	微调	0.5100	0.7795	0.8810	0.6900	0.2966	0.4277
	+1轮伪标签	0.5465	0.8351	0.9103	0.7740	0.3181	0.4046
	+2轮伪标签	0.5485	0.8403	0.8961	0.7960	0.3196	0.4092
	+3轮伪标签	0.5516	0.8526	0.9078	0.7860	0.3210	0.4092
ChatGLM-4[3]	api+模板1	0.3870	0.5543	0.7987	0.4580	0.1838	0.3353
	api+模板2	0.4832	0.6070	0.9064	0.7080	0.2309	0.4923
-	结果融合	0.5660	0.8526	0.9142	0.7980	0.3210	0.4923

表 1: 方法结果

排名	队伍	准确率	实体识别	角色识别	异常识别	空间推理	同义识别
1	TeleAI	0.6024	0.8947	0.9364	0.8480	0.3471	0.5631
2	zyy	0.5969	0.8491	0.9143	0.8100	0.3716	0.5131
3	涛涛不绝	0.5949	0.7719	0.9429	0.7800	0.3711	0.5877
4	Prompt	0.5660	0.8526	0.9142	0.7980	0.3210	0.4923
5	猪门永存	0.5647	0.7368	0.9286	0.7620	0.3240	0.5862
6	龙年旺旺空间站	0.5620	0.7965	0.9429	0.7420	0.3064	0.5692
7	panda	0.5448	0.7509	0.9117	0.7540	0.3044	0.5231
8	一个短篇	0.5355	0.7333	0.9013	0.7960	0.2858	0.5123
9	欣崽全球研究生后援会	0.5199	0.8053	0.9000	0.7020	0.34076	0.2415
10	CPIC1	0.4865	0.7667	0.8610	0.6220	0.2603	0.4031
11	少小离家	0.4724	0.6719	0.8182	0.7000	0.2735	0.3369
12	Azur Promilia	0.3364	0.4947	0.6727	0.2160	0.2172	0.2662
-	Baseline	0.4792	0.7509	0.8818	0.6860	0.2196	0.4200

表 2: 测试集排行榜[4]

4.2 结果分析

在使用大模型之前，我们也尝试使用了类似Bert的传统分类模型，然而，传统模型在测试集上的表现要远低于基线模型的结果。这一结果表明基座模型对于推理效果有着显著的影响，而大模型在预训练获得的世界知识和涌现能力对空间语义理解能力任务有着重要帮助。我们在对大模型微调时也面临着一个普遍问题，即幻觉现象(Huang et al., 2023)。当模型生成的文本不遵循原文或者不符合事实时，我们就认为模型出现了幻觉，尽管在我们训练集的选项中只有4个选项可选，但模型在结果预测时仍会产生4个选项以外的答案，为此，我们暂时只采用正则表达式来过滤掉这些无效答案。

[0]https://github.com/THUDM/ChatGLM3
[1]https://github.com/QwenLM/Qwen
[2]https://github.com/QwenLM/Qwen1.5
[3]https://chatglm.cn
[4]https://2030nlp.github.io/SpaCE2024/leaderboard.html

(c) 2024 中国中文信息学会计算语言学专业委员会

通过大量实验发现，数据增强方法在大多数任务中，尤其是小样本任务，通常会有不同程度的提升效果。然而，在本任务中，由于数据量规模并不小，采用类似随机词插入和随机词删除等通过添加噪声来实现数据增强的方法可能改变了原本的数据分布，反而没有显著提升效果。相反，使用半监督学习生成的伪标签可以增加数据集的规模，提升模型预测的准确率，并增加模型的泛化性。在使用多轮伪标签方法后，后续筛选得出的伪标签几乎不会有变化，导致模型的性能不再有提升，这时可以采用模型融合技术，取差异较大的多个模型，分别学习不同的输入，使得多个模型之间学到的知识特征尽量不同，这样使得多个模型可以更好的融合，提升性能。

5 总结

在本任务中，除了提供的训练集和测试集以外，任务组织者还额外提供了空间语义词表信息，但我们在仅使用了训练集和验证集的情况下就达到了超过基线模型的效果，后续并采用prompt优化和半监督学习的方式来进一步提升推理的表现。为了进一步优化结果，未来我们将针对使用训练集存在的过拟合问题，考虑在划分训练集和验证集时进行数据均衡。使用模型融合时，采用五折交叉验证法来训练多个模型，然后在对多个预测结果取平均，以取得更好的预测效果。

参考文献

Ray Jackendoff. Précis of foundations of language: Brain, meaning, grammar, evolution. *Behavioral and Brain Sciences*, 26(6):651–65; discussion 666–707, 2004.

Eric Nichols and Fadi Botros. Sprl-cww: Spatial relation classification with independent multi-class models. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 895–901, 2015.

Jennifer D'Souza and Vincent Ng. Sieve-based spatial relation extraction with expanding parse trees. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 758–768, 2015.

Haritz Salaberri, Olatz Arregi, and Beñat Zapirain. Ixagroupehuspaceeval:(x-space) a wordnet-based approach towards the automatic recognition of spatial information following the iso-space annotation scheme. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 856–861, 2015.

Bogyum Kim and Jae Sung Lee. Extracting spatial entities and relations in korean text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2389–2396, 2016.

Nitin Ramrakhiyani, Girish Palshikar, and Vasudeva Varma. A simple neural approach to spatial role labelling. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*, pages 102–108. Springer, 2019.

Hyeong Jin Shin, Jeong Yeon Park, Dae Bum Yuk, and Jae Sung Lee. Bert-based spatial information extraction. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 10–17, 2020.

Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364, 2019.

Soham Dan, Hangfeng He, and Dan Roth. Understanding spatial relations through multiple modalities. *arXiv preprint arXiv:2007.09551*, 2020.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

(c) 2024 中国中文信息学会计算语言学专业委员会

Chenyang Li, Long Zhang, Qiusheng Zheng, Zhongjie Zhao, and Ziwei Chen. User preference prediction for online dialogue systems based on pre-trained large model. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 349–357. Springer, 2023.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens, and Steven Bethard. Semeval-2013 task 3: Spatial role labeling. In *Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 255–262, 2013.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL, 2015.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*, 2021.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 空间语义理解能力评测任务设计的新思路—space2021数据集的研制. 语言文字应用, pages 99–110, 2022.

Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.

CCL 2024

CCL24-Eval任务3系统报告: 基于大型语言模型的中文空间语义评测

霍世图[1], 王钰君[1], 吴童杰[1]
[1]北京师范大学　国际中文教育学院　北京　100875
{20222109017, 202221090022, 202221090014}@mail.bnu.edu.cn

摘要

本研究的任务旨在让大模型进行实体识别、角色识别、异常识别、信息推理、同义识别任务，综合评估大模型的空间语义理解能力。其中，我们使用普通提示词、工作流提示词和思维链三种提示词策略来探讨大模型的空间语义理解能力，最后发现ERNIE-4在1-shot的普通提示词上表现最佳。最终，我们的方法排名第六，总体准确率得分为56.20%。

关键词： 空间语义；大语言模型；提示词工程

System Report for CCL24-Eval Task 3: Evaluation of Chinese Spatial Semantics Based on Generative Language Models

Shitu Huo[1], Yujun Wang[1], Tongjie Wu[1]
[1]Beijing Normal University, School of International Chinese Education, Beijing 100895
{20222109017, 202221090022, 202221090014}@mail.bnu.edu.cn

Abstract

The task of this paper aims to comprehensively assess large models' spatial semantic understanding capabilities through entity recognition, role recognition, anomaly detection, information inference, and synonym recognition tasks. We explored the spatial semantic understanding capabilities of large models using three prompt strategies: general prompts, workflow prompts, and chain-of-thought prompts. In the end, we found that ERNIE-4 performed best with 1-shot general prompts. Our system ranked sixth overall, with an accuracy score of 56.20%.

关键词： Spatial Semantics；Large Language Model；Prompt Engineering

1 引言

在自然语言处理领域，大型语言模型取得了显著进展。现有的大模型主要基于注意力机制的Transformer (Vaswani et al., 2017)，利用缩放定律（Scaling Laws）大幅提升模型性能 (Kaplan et al., 2020)，从而使得BERT (Devlin et al., 2018)和GPT-3 (Floridi and Chiriatti, 2020)等模型能够捕捉复杂的语言结构和语境关系，甚至可以批量改写和生成逼真的文本，从而端对端地完成机器翻译、语义分析等自然语言处理任务 (Brown et al., 2020; Zhao et al., 2023)。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

大语言模型具有较好的语义理解能力，与人类的语义判断具有极高的相关性，并展现出作为理论语言学新兴研究工具的潜力 (Tjuatja et al., 2023)。在早期，大语言模型常被视为一只“随机鹦鹉”，其学习机制是单纯统计某个词语出现的频率 (Bender et al., 2021)。然而后续的研究表明，大模型在没有遵循语言学理论的情况下，能够有效整合语义和语法的多重信息 (Piantadosi, 2023)，区分同一个词在不同语境中的多种用法 (Petersen and Potts, 2023)，甚至可以有效表征时间和空间 (Gurnee and Tegmark, 2023)。

空间语义评测是评估自然语言处理系统对空间表达理解能力的重要手段。传统的空间语义评测方法主要依赖人工标注，但这种方法标注成本高且可扩展性差。如今，基于生成式语言模型开展评测任务逐渐成为趋势。基于上述背景，本研究拟探究以下问题：1）大模型对空间语义的理解程度如何？2）在理解空间语义的具体任务上，大模型各有哪些优劣？

本研究基于第四届中文空间语义理解评测任务（SpaCE2024），首先介绍空间语义评测的背景和相关研究，然后通过实验分析不同模型的空间语义理解能力，最后对实验结果进行讨论和分析，以更好地了解大模型在空间语义理解方面的能力边界。

2 相关研究

这一部分首先介绍与空间语义有关的语言学研究，然后总结自然语言处理领域关于空间语义评测的研究概况。

2.1 语言学视角下的空间语义研究

认知语言学基于人们对世界的经验和对世界进行感知和概念化的方法 (张敏, 1998)，借助隐喻等方式将抽象概念具象化。空间关系在认知语言学中占据了重要地位，是人类最早习得的能力之一 (Akhundov, 1986; Clark, 1973; 张敏, 1998; 赵艳芳, 2001)。Lakoff and Turner (1989)指出，作为一种意象图式隐喻（image schema metaphor），空间隐喻将具体的空间概念投射到抽象的语言结构中，这样的投射可以传递空间关系及其内在逻辑。这是空间语义形成的认知基础。[0]

具体来说，人在感知外部世界时总是对空间敏感，包括运动、方向、地点等信息。Langacker (1982)高度重视空间语义对语言形成的作用，提出空间语法（Space Grammar）。Jackendoff (1983)的主题关系假设（Thematic Relations Hypothesis, TRH）认为人类语言概念结构中的事件（event）和状态（state）都是通过空间概念化组织起来的，所有语义场内的关系都类似空间组织关系。Lakoff (1987)的形式空间化假设（SFH）与之类似，并且更进一步从空间语义的视角讨论了一些基本句型的形成。Johnson (1987)概括出了和空间语义相关的27个最重要的意象图式，认为这是人类空间范畴推理的基础。Pütz and Dirven (1996)也指出地点位置对人类概念的形成有基础性作用。很多和空间范畴相关的概念也已经成为认知语言学中的重要工具和分析手段，如Talmy (1983)的图形-背景理论、Langacker (1987)的“基体-侧面”“射体-界标”关系等。

在这些基本概念和技术手段的指导下，国内外都出现了一批有关空间语义的微观研究。国外研究中，Hawkins (1984)基于空间认知对英语介词进行了全面研究。Vandeloise (1994)系统考察了法语空间介词的句法结构和语义表征。Herskovits (1986)对英语的空间表达式进行了跨学科的调查。Svorou (1993)则从认知普遍性的角度对空间介词进行了跨语言比较研究。国内研究中，廖秋忠 (1986)引进参照点的概念研究方位词，突破了传统语法对方位词静态研究的局限。刘宁生 (1994)讨论了汉语如何选择空间方位的参照物、目的物和方位词，从而表达物体的空间关系。齐沪扬 (1998)建立了现代汉语空间系统的理论框架，极大地开拓了汉语空间语义研究的视野。

汉语方位词“上”“下”始终是微观研究的重点。崔希亮 (2000)对“在X上”进行解析，显示了其空间语义及其心理延伸。蓝纯 (2003)比较了汉语的“上”“下”和英语的up和down，指出两种语言中相似方位词存在不同空间语义。白丽芳 (2006)更加深入地考察了汉语的“上”和“下”，指出两者在同一语言系统中也有不对称，“上”比“下”更基本、语义更丰富。徐丹 (2008)考察了汉语时空表达的语言特点，认为和其他语言相比，汉语特有的采用“上”“下”这样的纵向结构表示时间，体现了汉人独特的认知观念。

[0] “空间语义”和“空间范畴”是一对内涵和外延都类似的概念，本体研究中一般不作区分。如无特殊必要，下文也不区分两者。

(c) 2024 中国中文信息学会计算语言学专业委员会

除了关于方位词的静态研究，有关位移动词和位移事件的动态研究也是国内空间语义研究的热门话题。陆俭明 (2002)最先明确界定位移动词，认为这类动词“含有向着说话者或离开说话者位移的语义特征”，这个概念蕴含了很强的空间语义观。此后的研究中，动词的位移性逐渐作为动词的一种语义特征使用，和语义特征分析法或构式研究结合在一起，如张国宪 (2006)对表状态“在+处所+V状”和“V状+在+处所”两类构式的研究，雍茜 (2013)对三类“在+L”构式的研究，曾传禄 (2014)对“V起来”“V得（不）过来/过去”等结构的语义分化研究等。

近来对空间语义的研究，则主要集中于汉（语）方（言）比较、汉（语）外（语）比较、空间语义的认知和心理现实性等问题，如贾红霞 (2009)、尹蔚彬 (2014)、李云兵 (2016; 2020)、祝克懿 (2018)等。可以说，对空间语义的研究在国内已经步入了一个新的阶段，跨学科和实证研究不断涌现，对包括汉语在内的人类语言空间语义的认识也在不断提高。

2.2 自然语言处理领域的空间语义评测研究

自然语言处理领域的空间语义评测研究主要关注如何从自然语言中提取和理解与物理空间相关的信息。深度学习方法出现之前，自然语言处理中的空间语义任务大致经历了以下阶段：阶段一主要关注空间语义网络的层级和关系定义 (Tappan, 2004)，通过明确空间实体之间的层级关系和语义连接，初步奠定了空间信息处理的基础；阶段二则侧重特定空间语义任务，如空间实体识别 (Kordjamshidi et al., 2011)、空间关系判定等，使用机器学习方法在特定数据集上进行的半监督或无监督的训练。

然而，此前的研究大多采用非语言的形式化方法，没有充分考虑人类在自然语言中表达空间关系的方式，因此并不能有效理解自然语言中的抽象空间概念 (Stock, 1998; Renz and Nebel, 2007; Bateman et al., 2007)。例如，人类在交流空间信息时，自然语言片段中存在着不确定性和模糊性，通过构建与空间相关的知识分类，如拓扑关系和度量关系 (Tappan, 2004)等类别的方式发挥的效果有限。

为了解决自然语言中空间表述的模糊性问题，尤其是空间介词的语义识别，Kordjamshidi 等 (2011)提出了空间角色标注（Space Role Labelling，SpRL）任务，Roberts (2012)使用联合方法来识别和分类空间角色，首先从训练和测试数据中使用CRF模型提取特征，捕捉词语之间的依赖关系，并使用最大熵和朴素贝叶斯分类器来消除介词含义的歧义。同时，利用SemEval-2007 (Litkowski and Hargraves, 2007)的介词项目（TPP）的注释数据来学习介词的空间意义，然后在识别轨迹和地标角色的空间角色标签器中使用介词消歧的结果，从而实现介词的空间或非空间意义的二元分类。该方法能够同时考虑空间关系中的所有元素（如轨迹物、地标和指示物），允许使用基于整个关系的特征集 (Roberts and Harabagiu, 2012)，实现了空间关系介词消歧。

SpaceEval 2013 (Kolomiyets et al., 2013)扩展了SpRL 任务，在静态空间关系识别之外引入了运动关系（Movelink）和运动标签，用于注释运动动词或名词性运动事件及其类别并从空间语义的角度来分类事件。SpaceEval 2015 (Pustejovsky et al., 2015)则通过设定空间元素识别和分类任务、运动信号识别、运动关系识别等子任务，全面评估系统在识别和分类各种空间概念及其关系中的表现。

在中文方面，SpaCE空间语义测评借鉴了上述成果，构建了一系列高质量评测数据集，为机器的空间语义理解提出了更高要求 (詹卫东et al., 2022; 岳朋雪et al., 2023)。任务设置上，要求模型不仅能在富含空间信息的语料中执行识别和分类任务，还要进行方位推理和异形同义识别等多层次的空间语义理解。此外，在大语言模型掌握世界知识并“涌现”出空间语义识别和规划能力后，针对空间语义理解任务的设定必然更加复杂且全面。

3 数据集

本研究的数据集涵盖五大任务类别和两种选择题形式，共有九个小类题目，包括报刊、文学作品、中小学课本等一般领域与交通事故、体育动作、地理百科等专业领域，旨在考察模型在实体识别、角色识别、异常判断、方位推理和语义识别五个维度的空间语义理解能力。表1分别展示了训练集、验证集和测试集的数据情况，每一条数据都包含了题目编号、文本、选项和答案，评测采用选择题形式，题目选项设置为4个。在数据分布上，空间方位信息推理题目最多，空间异形同义识别题目最少，题型以单选题为主。总的来看，数据集的分布差异呈现了题目的多样性和复杂性，给本次评测也带来了一定的挑战。

(c) 2024 中国中文信息学会计算语言学专业委员会

序号	任务类别	任务要求	题型	训练集	验证集	测试集
1	空间信息实体识别	选出文本空间信息的参照物	单选题	937	226	489
			多选题	161	24	81
2	空间信息角色识别	选出文本空间信息的语义角色，或者选出与语义角色相对应的空间表达形式	单选题	1074	186	746
			多选题	19	4	24
3	空间信息异常识别	从四个选项中选出文本空间信息异常的语言表达	单选题	1077	40	500
4	空间方位信息推理	基于文本给出的推理条件进行空间方位推理，从四个选项中选出推理结果	单选题	909	468	1509
			多选题	301	207	531
5	空间异形同义识别	从四个选项中选出能使两个文本异形同义或异义的空间义词语	单选题	4	44	517
			多选题	1	11	133
总计				**4483**	**1210**	**4530**

Table 1: SPaCE 2024数据集概况

4 实验过程

4.1 模型一览

如表2所示，本研究选取了来自OpenAI、智谱华章、阿里巴巴、百度和深度求索的六个具有代表性的模型，涵盖了不同模型架构和规模。这些模型的参数规模从720亿到2360亿不等，支持的上下文长度从3.2万到12.8万不等，均是在2024年本研究开展期间的最新或较新版本[1]。

模型	版本日期	开发者	模型大小	上下文	词表大小	是否开源	调用方式
GPT-4 Turbo	04-09	OpenAI	未披露	12.8万	10万	否	API
GPT-4o	05-13	OpenAI	未披露	12.8万	20万	否	API
GLM-4	未披露	智谱华章	未披露	12.8万	未披露	否	API
ERNIE-4	03-29	百度	未披露	8千	未披露	否	API
Qwen1.5-72B-chat	未披露	阿里巴巴	720亿	3.2万	15万	是	API
Deepseek-V2-chat	未披露	深度求索	2360亿	3.2万	10万	是	API

Table 2: 模型一览

4.2 提示词工程

本研究的提示词均采用Markdown格式的结构化格式，主要包含提示词策略、提示样本构建两个部分。

在提示词策略上，分别采用普通提示（Vanilla Prompt）、工作流（Workflow）、思维链（Chain of Thought, CoT）三种方式构建提示词。在提示样本构建上，本研究的普通提示词和工作流提示词都采用0-shot、1-shot、3-shot，思维链采用1-shot。对于思维链提示词，我们参考了Wei (2022)的提示词，将其改为“想法”和“答案”两部分，让输出更为结构化，从而方便思维链和答案的提取。

关于样本的选取，训练集每条数据都有一个文本（C）、一个问题（Q）、四个选项（O）和一个答案（A）。对于每条数据，我们将其组织为一个样本$S_i = \{C_i, Q_i, O_i\}$。随后，使用Sentence-BERT (Reimers and Gurevych, 2019)将这些样本转换为向量。接下来，针对每个任务类别，我们计算了所有样本向量的平均值，作为该类别的簇心。最后，通过计算每个样本向

[1]本研究开展日期为2024年5月1日至5月17日。

(c) 2024 中国中文信息学会计算语言学专业委员会

量与簇心的语义相似度，分别找出距离簇心最近的1个和3个样本，作为1-shot和3-shot的训练数据。

在思维链中，样本示例需要有思考过程，因此我们用GPT-4撰写了样本的思维链过程。此外，由于异形同义识别任务的训练集只有1道多选题，我们人工将异形同义识别任务的其中2道单选题改编为多选题，以确保能够构建3-shot。

普通提示词示例

#Goal: 从四个选项中选出文本中的空间信息参照物。注意，只需回答option的一个key，不需要回答value，不需要解释。
*Text:** <text>
*Question:** <question>
*Option:** <option>
*Answer:**

工作流提示词示例

#Role: 你是一位擅长空间信息实体识别的专家。
#Goal: 从四个选项中选出文本中的空间信息参照物。注意，只需回答option的一个key，不需要回答value，不需要解释。
#Workflow: 1.阅读text：细致阅读提供的text，特别关注其中的空间信息描述。2.分析option：查看所有option，识别哪些可能是text中的空间参照物。3.选择正确option：对比text与option，选择最匹配的空间信息参照物。
*Text:** <text>
*Question:** <question>
*Option:** <option>
*Answer:**

思维链提示词示例

#Goal: 从四个选项中选出文本中的空间信息参照物。注意，只需回答option的一个key，不需要回答value，写出Thought和Answer。
*Text:** <text>
*Question:** <question>
*Option:** <option>
*Thought:** <thought>
*Answer:**

4.3 实验设置

针对不同的**提示词输出内容**，我们采用了不同的答案提取方法以优化提取过程。在普通提示和工作流提示中，提示词要求模型直接输出选项，选项之间用英文逗号“,”隔开，以便后续转换为列表格式进行评估。然而，模型有时输出答案后可能继续输出其他内容。对此，我们首先将其转换为列表，接着遍历每个元素，提取每个元素中的首字符。在思维链方法中，提示词要求模型先输出思路，再输出答案。我们使用正则表达式来提取答案，所使用的正则表达式为`<**Answer:**\n(.+?)(\n\n|$)>`。由于不同模型的指令遵循能力存在差异，我们还会自动检查每个答案是否都为A、B、C、D 四个选项之一，如不符合，还需人工检查。

关于**模型输出结果及其评测**，本研究将`temperature` 设为0.1，以确保模型输出结果的稳定性。评测指标采用准确率（Accuracy），即模型答对的题目数量占所有题目的百分比。模型答对为1分，其他情况为0分。其他情况包括：模型认为选项都不符合要求，模型拒绝回答问题，或在多选题中未能全部答对。

(c) 2024 中国中文信息学会计算语言学专业委员会

5 结果

5.1 模型总体表现

表 3和表 4分别是模型验证集、测试集的总体表现，满分为100。ERNIE-4借助1个样本的普通提示得到53.88%，为本研究评测的最高分，而GLM-4在1个样本的工作流提示词得到了第二高分53.14%。最终测试集使用ERNIE-4和GLM-4进行预测，ERNIE-4达到了最高的准确率，为56.20%。

根据所有模型的表现情况，我们总结归纳以下结论：（1）大模型基座能力具有举足轻重的作用。大模型的表现并不是一成不变，但比如ERNIE-4、GLM-4等模型拥有较强的中文语义理解基座能力，能够很好地适应多种有挑战性的任务。（2）提示词的数量对模型结果有重要影响。单样本可以显著提升模型的空间语义理解能力，但相较于0-shot，1/3-shot可以显著提升模型的空间语义理解能力，但从1到3-shot，准确率升降不定（7例上升，5例下降）。（3）提示词策略不一定越复杂越好，简单的提示词策略可能也有出色的效果。思维链可以帮助模型更好地理解语义空间，但在此次空间语义测评中表现并不突出。

模型	普通			工作流			思维链
	0	1	3	0	1	3	1
ERNIE-4	50.25	**53.88**	52.73	**52.23**	52.73	**52.81**	**51.06**
GLM-4	**51.24**	52.01	52.23	50.49	**53.14**	50.41	50.82
GPT-4o	48.92	51.16	**52.89**	48.35	50.99	51.73	50.91
GPT-4 Turbo	48.18	50.99	51.54	47.43	51.49	47.77	50.74
Deepseek-V2-chat	48.84	49.83	49.98	46.69	49.42	49.83	46.78
Qwen1.5-72B-chat	44.71	46.61	46.45	42.81	45.70	45.04	45.45

Table 3: 模型在验证集的总体表现

模型	样本数量	提示词	测试集准确率
ERNIE-4	1	普通	**56.20**
GLM-4	1	工作流	54.52

Table 4: 模型在测试集的最终表现

5.2 模型具体表现

大模型使用某个提示词策略达到最佳结果，但不等于在该提示词策略下各个方面都表现优秀。基于此，本研究进一步探究五个模型的具体表现，涵盖实体识别、角色识别、异常识别、空间推理、同义词识别、单选题和多选题7个维度。为了展示模型的实际表现和最大潜力，表 5列出了每个模型在验证集中的实际最佳性能和潜在最佳性能。其中，实际最佳性能是指某一维度在最优提示词下的得分，潜在最佳性能是该维度在不同提示词中的最高得分。

整体上看，所有模型在角色识别任务的表现最优，在空间推理任务的表现相对最差，单选题得分普遍高于多选题。在实际最佳性能方面，ERNIE-4表现最好，在4个任务和单选题表现优异，GLM-4次之，在角色识别任务和多选题的得分最高。在潜在最佳性能方面，ERNIE-4和GLM-4依然保持稳定，模型可以在实体识别和角色识别这两个任务发挥更大潜力。此外，Deepseek-V2-chat在同义识别任务的潜在表现尤其值得关注，可以达到与ERNIE-4、GLM-4比肩的水平。

下列以ERNIE-4的实际最佳性能的结果为例，探究其在验证集不同任务类别的表现，更为细致地了解模型空间语义理解的特点。

5.2.1 在实体识别题目的表现

实体识别类题目考察模型能否识别空间方位词和语境中已经出现过实体的同指关系。由于这类空间方位词和实体的关系在语境中是固定的，模型比较容易从语境中学到这个知识，其对实体的空间语义关系理解比较准确（如1题和4题）。

(c) 2024 中国中文信息学会计算语言学专业委员会

		实体识别	角色识别	异常识别	空间推理	同义识别	单选题	多选题
ERNIE-4	实际最佳	**79.20**	95.26	**87.50**	29.92	**65.45**	**61.20**	25.20
	潜在最佳	**80.40**	**96.84**	**87.50**	29.92	**65.45**	**61.20**	25.20
GLM-4	实际最佳	78.40	**95.79**	85.00	29.33	60.00	58.30	**32.93**
	潜在最佳	78.40	**96.84**	85.00	29.33	63.63	59.64	**32.93**
GPT-4o	实际最佳	76.40	93.68	80.00	**30.52**	60.00	58.09	32.52
	潜在最佳	76.40	95.79	80.00	**30.52**	65.45	59.34	33.74
GPT-4 Turbo	实际最佳	76.80	95.26	72.50	28.59	54.54	59.54	20.73
	潜在最佳	76.80	95.26	80.00	29.48	61.82	59.92	23.17
Deepseek-V2-chat	实际最佳	74.40	95.26	77.50	26.22	52.73	56.33	24.80
	潜在最佳	74.40	96.84	82.50	29.04	**65.45**	56.74	29.67
Qwen1.5-72B-chat	实际最佳	71.60	91.05	67.50	23.11	52.73	55.50	11.79
	潜在最佳	72.40	93.68	67.50	24.74	54.54	55.50	16.67

Table 5: 模型在验证集的实际最佳性能和潜在最佳性能

1题：周游口袋里只有五元钱……所以蹬三轮车的上来拉生意时，他理都不理他们，而是从西装口袋里掏出个玩具手机，这个玩具手机像真的一样，里面装上一节五号电池，悄悄按上一个键，手机的铃声就会响起来。（题目：___里面装上一节五号电池）

4题：回家以后，她给丈夫算了一笔账：我每天上下班路程要花3个小时，工作8小时，中午吃饭1小时，总共在外边花12小时……（题目：总共在___外边花12小时）

值得一提的是，当需要判断的实体在语境中和其他实体有领属或广义领属关系，尤其是其他实体没有在语境中出现时，ERNIE-4对实体选择的错误率较高。58题没有出现实体“屋子”，67题存在干扰项“大厅”（实际应该选“楼上”，两者有广义领属关系），模型对两句话的判断都出现了问题。

58题：爸爸把我从床头打到床角，从床上打到床下，外面的雨声混合着我的哭声……（题目：___外面的雨声混合着我的哭声）

67题：楼上只有南面的大厅有灯亮。灯亮里有块白长布，写着点什么——林乃久知道写的是什么。其余的三面黑洞洞的……（题目：___三面黑洞洞的）

总的来说，ERNIE-4可以比较准确地判断语境中出现过的单独和方位词关联的实体，但对语境中没有出现的、以及语境中还存在和该实体有领属关系的干扰项时，模型的判断能力比较弱。

5.2.2 在角色识别题目的表现

角色识别类题目考察模型能否识别存在空间交互关系的两个实体。两个实体的空间交互关系有多种来源，包括领属关系带来的空间关系（如251题）、事件带来的空间关系（如第258题）、相对位置带来的空间关系（如259题）等。ERNIE-4对这些关系的认识都非常准确。

251题：时间过去近两个月，木沙江·努尔墩仍清楚地记得……在人工湖边的冰窟中，拉齐尼用一只手臂搂住孩子，另一只手努力托举着孩子……（题目：___的手努力托举着孩子）

258题：……正在站上值班的牛红生例行巡检，走到龙王路段时，发现一辆轿车从百米外的公路上猛然栽进路边的排水渠……牛红生只能在水中摸索，摸到车门后用力拽开，把人拉出水面……（题目：牛红生把___拉出了水面）

259题：文本同258题（题目：轿车栽进去时的初始位置是___）

当题目直接询问实体（包括抽象实体）的复杂空间关系（包括隐域空间关系）时，模型的识别能力下降，出现了较多错误。这类空间关系要么是隐涵的，要么是“元语言”意义上的，不容易从语境中直接得到。这里说的“元语言”，指题目选项中的抽象空间关系表达，如398题的四个选项为“路径”“方向”“起点”和“外部位置”，是抽象的描述方位的语言，和具体题干无关，这和“名词”“动词”“主语”“宾语”等“描述语言”的“元语言”不同。感谢审稿专家指出这一点。

398题：首先是水的气味，宽广的昌江流经鄱阳城奔向鄱阳湖，在城外留下韭菜湖、青山湖、土湖、东湖、球场湖五片湖……（题目：“鄱阳城”属于“昌江”流动时的___信息）

425题：几天以后李光头回来了，他在上海买了一辆红色的桑塔纳轿车，他有专车了……李光头从车里出来时，身穿一身黑色的意大利阿玛尼西装，那身破烂衣服扔在上海的垃圾桶里了……（题目：“破烂衣服扔在上海的垃圾桶里”发生在___）

(c) 2024 中国中文信息学会计算语言学专业委员会

总的来说，ERNIE-4判断两个具体实体的具体空间关系是最准确的，判断抽象实体的具体空间关系次之，判断抽象实体的抽象空间关系是最弱的。这和人类一般的认知能力相似，越是具体的对象和关系就越容易认知和识别。

5.2.3 在异常识别题目的表现

异常识别类题目考察模型能否识别存在异常或错误空间交互关系的若干实体。异常的空间关系要么是不合常理的（如441题），要么是自相矛盾的（如442题）。

441题：小红在下，我在上，走到四楼的东侧.....（题目：异常的空间方位信息是___，要求识别出“小红在下，我在上”）

442题：灵车缓缓地前进，牵动着千万人的心.....人们多么希望车子能停下来，希望时间能停下来！可是灵车渐渐地靠近了，最后消失在苍茫的夜色中了.....（题目：异常的空间方位信息是___，要求识别出“灵车渐渐地靠近并消失在夜色中”）

ERNIE-4的异常空间识别能力较好，但如果异常空间关系复杂，ERNIE-4就不容易将其识别出来。例如在478题中，由东向西行驶的车子左转弯是向南而不是向北。这种空间关系并不直观，模型在识别上也出现了问题。对这种异常空间关系的识别通常需要另外的百科知识或一般推理能力。

478题：经审理查明.....小型客车沿本区亭林镇红梓路由东向西行驶至车亭公路路口时，遇绿灯向北实施左转弯途中.....（题目：异常的空间方位信息是___，要求识别出“小型客车由东向西行驶至车亭公路路口向北左转弯”）

总的来说，ERNIE-4基本可以准确识别相对直观的异常空间关系，但在需要调用推理能力或百科知识判断空间关系是否正常时，模型的表现并不是很好。

5.2.4 在空间推理题目的表现

空间推理类题目，考察模型能否通过简单的推理方式得到正确的实体间空间关系。这类问题中，模型只能在语境里得到条件句命题的前件，后件需要根据实际问题的需要自行推理。

481题：贺知章、李白、陈子昂、骆宾王、王维、孟浩然六个人在海边沙滩上围成一圈坐着，大家都面朝中心的篝火。六人的位置恰好形成一个正六边形。任意相邻两个人之间的距离相等，大约为一米。已知：陈子昂在骆宾王左边数起第1个位置，孟浩然在陈子昂逆时针方向的第5个位置，王维在孟浩然顺时针方向的第1个位置（题目：孟浩然在___的斜对面）

尽管该推理问题并不是很复杂，但ERNIE-4体现出的空间推理能力不是很强。这种推理能力不仅需要正确得到语境中的信息，同时还要调用必要的百科知识（如481题中对“正六边形”的理解等），然后根据这些信息展开推理。空间推理可以理解为复杂的角色识别和实体识别问题，这意味着模型在处理连续实体识别问题上还存在一些问题，无法正确判断这种需要推理和叠加的复杂空间关系。

5.2.5 在同义识别题目的表现

同义识别类题目考察模型对不同空间方位词表达的具体空间关系的认识是否准确。汉语存在一批空间方位词，它们单独使用的语义不同，但和某些空间实体结合时可以表达相同的空间方向。例如在1157题，“回来”和“回去”是两个不同的词，甚至有时被认为是一对反义词，但此时发生替换后空间场景没有明显改变。

1157题：傍晚的时候，宋钢将他带回去的钱用一张旧报纸仔细包好了，放在了枕头下面.....（题目：“回去”替换为___形成的新句可以与原句表达相同的空间场景，要求用“回来”替换“回去”）

由于这类空间方位词单独使用的语义不同，如果要正确替换，模型必须将方位词和与之关联的实体联系在一起替换，这考察了模型对“实体+方位词”序列语义的认识能力，而不是仅仅从语境中找到方位词关联的实体，后者是实体识别的任务。从同义识别题目和实体识别题目的对比表现上看，ERNIE-4可以比较好地找到方位词和方位词关联的实体，但是在其语义的理解上表现得不够出色。

5.2.6 在不同任务类型的任务表现

综合上文分析，虽然测试题目在不同类型空间关系的考察上有所差异，但从ERNIE-4在现有题目的平均表现上还是可以观察到其空间语义理解的特点，具体如表 6。

6 结语

在CCL2024年中文空间语义评测中，我们观察到了不同大模型的空间语义理解能力。本队

(c) 2024 中国中文信息学会计算语言学专业委员会

任务类别	模型表现	影响因素
角色识别	好	具体实体的具体空间关系不受外界影响，但不容易判断抽象实体（元语言对象）的抽象空间关系
实体识别	较好	表示静态的空间关系，基本不受外界影响，但出现与目标实体有领属或广义领属关系的其他实体时容易受干扰
异常识别	较好	简单异常空间关系容易识别，但在需要百科知识或推理能力的问题上易受干扰
同义识别	较差	表示空间关系的联系，受“实体+方位词”语义的影响
空间推理	差	参与空间主体较多，而且需要百科知识和推理能力，易受干扰项影响

Table 6: 模型表现及其影响因素

伍提交的系统在封闭赛道中排名第6，测试集上的准确率得分为56.20。综合来看，为了提升模型的空间语义理解能力，可以深入优化模型的提示词处理机制，或者设计更具结构化和明确性的提示词。

此外，进一步提升大模型的基座能力，也是未来模型发展的重要方向。这包括对模型架构的优化，如通过更多的数据和更先进的训练算法提升模型的表现；同时，结合外部知识库和信息源，使模型能够在更广泛的知识背景下进行推理和生成。通过结合知识增强方法和交互式决策策略，可以显著提高模型的实际应用效果。

参考文献

Murad D. Akhundov. 1986. *Conceptions of Space and Time*. The MIT Press, Cambridge, MA.

John Bateman, Thora Tenbrink and Scott Farrar. 2007. The role of conceptual and linguistic ontologies in interpreting spatial discourse. *Discourse Processes*, 44(3): 175-212.

Emily M. Bender, Timnit Gebrum, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610-623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry and Amanda Askell et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877-1901.

Herbert H. Clark. 1973. *Space, Time, Semantics and the Child: Cognitive Development and the Acquisition of Language*. Academic Press, New York.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, pages 4171-4186.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681-694.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Eric Hawkins. 1984. *Awareness of Language: An Introduction*. Cambridge University Press, Cambridge.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, Cambridge.

Ray S. Jackendoff. 1983. *Semantics and Cognition*. The MIT Press, Cambridge, MA.

Mark Johnson. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason*. The University of Chicago Press, Chicago.

(c) 2024 中国中文信息学会计算语言学专业委员会

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Jeffrey and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255-262.

Parisa Kordjamshidi, Martijn Van Otterlo and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3): 1-36.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, Chicago.

George Lakoff and Mark Turner. 1989. *More Than Cool Reason: A Field Guide to Poetic Metaphor*. The University of Chicago Press, Chicago.

Ronald W. Langacker. 1982. Space grammar, analysability, and the English passive. *Language*, 22–80. JSTOR.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press, Stanford.

Kenneth C. Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24-29.

Erika Petersen and Christopher Potts. 2023. Lexical Semantics with Large Language Models: A Case Study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490-511.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Pustejovsky, J., Kordjamshidi, P., Moens, M. F., et al. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884-894.

Martin Pütz and René Dirven. 1996. *The Construal of Space in Language and Thought*. de Gruyter Mouton, Berlin.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jochen Renz and Bernhard Nebel. 2007. Qualitative spatial reasoning using constraint calculi. In *Handbook of Spatial Logics*, pages 161-215. Dordrecht: Springer Netherlands.

Kirk Roberts and Sanda Harabagiu. 2012. UTD-SpRL: A joint approach to spatial role labeling. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 419-424.

Oliviero Stock. 1998. *Spatial and Temporal Reasoning*. Springer Science & Business Media.

Soteria Svorou. 1993. *The Grammar of Space*. John Benjamins, Amsterdam.

Leonard Talmy. 1983. How language structures space. In H. L. Pick and L. P. Acredolo, editors, *Spatial Orientation*, pages 225-282. Springer, Boston, MA.

Daniel Allen Tappan. 2004. Knowledge-based spatial reasoning for automated scene generation from text descriptions. New Mexico State University.

Lindia Tjuatja, Emmy Liu, Lori Levin and Graham Neubig. 2023. Syntax and Semantics Meet in the "Middle": Probing the Syntax-Semantics Interface of LMs Through Agentivity. *arXiv preprint arXiv:2305.18185*.

(c) 2024 中国中文信息学会计算语言学专业委员会

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Llion, Aiden N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998-6008.

Claude Vandeloise. 1994. Methodology and analyses of the preposition in. *Cognitive Linguistics*, 2: 157-184.

Wei J, Wang X, Schuurmans D, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824-24837.

Zhao Wayne Xin, Zhou Kun, Li Junyi, Tang Tianyi, Wang Xiaolei, Hou Yupeng, Min Yingqian, Zhang Beichen, Zhang Junjie and Dong Zican et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

白丽芳. 2006. “名词+上/下”语义结构的对称与不对称. 语言教学与研究, 4: 58-65.

崔希亮. 2000. 空间方位关系及其泛化形式的认知解释. 载《语法研究和探索（十）》, pages 85-97. 北京: 商务印书馆.

贾红霞. 2009. 普通话儿童空间范畴表达发展的个案研究. 博士学位论文. 中国社会科学院研究生院.

蓝纯. 2003. 从认知角度看汉语和英语的空间隐喻. 外语教学与研究出版社, 北京.

李云兵. 2016. 论苗语空间范畴的认知. 民族语文, 3: 8-34.

李云兵. 2020. 朱坝羌语静态空间范畴的表征与认知. 民族语文, 5: 3-20.

廖秋忠. 1986. 现代汉语篇章中的空间和时间的参考点. 载《廖秋忠文集》, pages 3-15. 北京: 北京语言学院出版社.

刘宁生. 1994. 汉语怎样表达物体的空间关系. 中国语文, 3: 169-179.

陆俭明. 2002. 动词后趋向补语和宾语的位置问题. 世界汉语教学, 1: 5-17+114.

齐沪扬. 1998. 现代汉语空间问题研究. 学林出版社, 上海.

徐丹. 2008. 从认知角度看汉语的两对空间词. 中国语文, 6: 504-510+575.

尹蔚彬. 2014. 拉坞戎语的空间范畴. 语言科学, 3: 268-280.

雍茜. 2013. “在+L”类构式与动词的语义整合. 硕士学位论文. 上海师范大学.

岳朋雪, 王诚文, 孙春晖, 詹卫东and 穗志方. 2023. 中文空间语义理解评测数据集质量评估研究. 语言文字应用, (01): 101-113. DOI: 10.16499/j.cnki.1003-5397.2023.01.006.

曾传禄. 2014. 现代汉语位移空间的认知研究. 商务印书馆, 北京.

张国宪. 2006. 现代汉语形容词功能与认知研究. 商务印书馆, 北京.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐and 秦梓巍. 2022. 空间语义理解能力评测任务设计的新思路—SpaCE2021数据集的研制. 语言文字应用, (02): 99-110. DOI: 10.16499/j.cnki.1003-5397.2022.02.005.

张敏. 1998. 认知语言学与汉语名词短语. 中国社会科学出版社, 北京.

赵艳芳. 2001. 认知语言学概论. 上海外语教育出版社, 上海.

祝克懿. 2018. 心理空间范畴与语言生成机制.

CCL 2024

(c) 2024 中国中文信息学会计算语言学专业委员会

CCL24-Eval 任务3系统报告：基于上下文学习与思维链策略的中文空间语义理解

王士权，付薇薇，方瑞玉，李孟祥，何忠江，李永翔，宋双永
中国电信人工智能研究院
{wangsq23, fuweiwei, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

摘要

本技术报告详细介绍了我们团队参加第四届中文空间语义理解评测（SpaCE2024）的方法和成果。SpaCE2024旨在全面测试机器对中文空间语义的理解能力，包括空间信息实体识别、空间信息实体识别、空间信息异常识别、空间方位信息推理和空间异形同义识别五个不同的任务。我们团队采用精心设计的prompt并结合微调的方式激发大语言模型的空间语义理解能力，构建了一个高效的空间语义理解系统。在最终的评估中，我们在空间信息实体识别题目中准确率为0.8947，在空间信息实体识别题目中准确率为0.9364，在空间信息异常识别题目中准确率为0.8480，在空间方位信息推理题目中准确率为0.3471，在空间异形同义识别题目中准确率为0.5631，测试集综合准确率为0.6024，排名第一。

关键词： 空间语义理解 ；大语言模型

System Report for CCL24-Eval Task 3: Chinese Spatial Semantic Understanding Based on In-Context Learning and Chain of Thought Strategy

Shiquan Wang, Weiwei Fu, Ruiyu Fang, Mengxiang Li, Zhongjiang He, Yongxiang Li , Shuangyong Song
Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd
{wangsq23, fuweiwei, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

Abstract

This technical report provides a detailed introduction to the methods and achievements of our team in the Fourth Chinese Spatial Semantic Understanding Evaluation (SpaCE2024). The SpaCE2024 aims to comprehensively test a machine's ability to understand Chinese spatial semantics across five different tasks: spatial information entity recognition, spatial information entity disambiguation, spatial information anomaly detection, spatial orientation reasoning, and spatial heteronym synonym recognition.Our team employed meticulously designed prompts combined with fine-tuning to enhance the spatial semantic understanding capabilities of large language models, thereby constructing an efficient spatial semantic understanding system. In the final evaluation, our system achieved an accuracy of 0.8947 in spatial information entity recognition, 0.9364 in spatial information entity disambiguation, 0.8480 in spatial information anomaly detection, 0.3471 in spatial orientation reasoning, and 0.5631 in spatial heteronym synonym recognition. The overall accuracy on the test set was 0.6024, earning us a first-place ranking.

Keywords: Spatial semantic understanding , Large Language Model

第二十三届中国计算语言学大会论文集，第106页-第112页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

空间语义理解是自然语言处理（NLP）领域中一个极具挑战性的任务，它要求机器不仅要理解语言的表面形式，还要能够构建和推理出语言背后的空间关系和场景。掌握中文空间语义理解，对于提升机器翻译的准确性、增强人机交互的自然性、提高自动摘要和信息检索的相关性，以及推动智能教育、自动驾驶、室内导航等应用领域的发展具有深远的影响。此外，由于中文在表达空间关系时具有独特的语言现象和语法结构，中文空间语义理解的研究还有助于丰富跨语言的NLP理论和技术，促进人工智能对人类语言和认知过程的深入理解。

第四届中文空间语义理解评测（SpaCE2024）更加注重针对大语言模型的空间语义理解能力的测试，宗旨是在一个测试数据集上考察机器中文空间语义理解的综合能力。为此，SpaCE2024 将不再划分子任务，而是以选择题的形式考察五个层次的空间语义理解能力，分别是空间信息实体识别、空间信息角色识别、空间信息异常识别、空间方位信息推理、空间异形同义识别。

在本次评测任务中，我们首先使用精心设计的提示（prompt）获取模型在选择答案时的推理过程。随后，我们结合上下文学习与思维链（chain of thought）的方法，构建推理提示以提高模型在空间语义理解题目上的表现。其次，我们基于上下文学习的方法对模型进行微调，以进一步增强其在该任务中的性能。最后，通过投票策略整合多种模型的输出，获取最优结果。我们的系统在最终的线上评测的空间信息实体识别题目中准确率为0.8947，在空间信息实体识别题目中准确率为0.9364，在空间信息异常识别题目中准确率为0.8480，在空间方位信息推理题目中准确率为0.3471，在空间异形同义识别题目中准确率为0.5631，测试集综合准确率为0.6024，综合排名第一。

2 相关工作

空间语义理解作为自然语言处理领域的核心任务之一，对于实现类人智能具有重要意义。人类通过空间范畴来组织和理解周围的世界，而在自然语言文本中，空间信息的表达和理解尤为关键。随着自然语言处理技术的不断进步，对机器的空间语义理解能力进行评测变得尤为重要。为了更好的评判模型过对于空间语义的理解能力，Kordjamshidi等(Kordjamshidi et al., 2012)在SemEval-2012举行了基于静态空间关系的空间语义角色标注任务，该任务涉及从自然语言中提取空间语义的主要组成部分：轨迹、地标和空间指示器。除了这些主要组成部分外，还针对它们之间的联系以及空间关系的一般类型，包括区域、方向和距离进行了标注。在SemEval-2013中，Kordjamshidi等(Kolomiyets et al., 2013)对Spatial Role Labeling任务进行了完善。Pustejovsky等人(Pustejovsky et al., 2015)在SemEval-2015提出在SpRL任务中使用ISO-Space标注体系，进一步完善空间语义角色标注任务。

SpaCE系列赛事在探索中文空间语义理解评测上做出了持久的贡献，SpaCE2021(詹卫东et al., 2022)分为三个中文空间语义正误判断、中文空间语义异常归因合理性判断与中文空间语义判断与归因联合任务三个子任务，分别考察模型能否正确区分正常与错误的空间语义表达，模型能否解释空间语义表达错误的原因，模型处理上述两个任务的综合能力。SpaCE2022(Liming et al., 2023)在SpaCE2021的基础上扩大了数据规模，拓宽了语料类型，改进了任务形式，提高了标注质量，共分为中文空间语义正误判断、中文空间语义异常归因与异常文本识别与中文空间实体识别与空间方位关系标注三个子任务。SpaCE2023(Xiao et al., 2023)在SpaCE2022的基础上删除了空间语义正误判断任务，以及异常归因任务；保留了异常文本识别任务，以及空间语义角色标注任务，并改进了数据格式，更新了测试集，提高了标注质量；新增了生成类任务，要求机器判断两段在语言表达形式上有差异的文本是否可以描述相同的空间场景，并说明判断理由。

SpaCE2024将之前的测试题形式改为选择题，并通过设置一定比例的重复题目来测试模型的稳定性，同时提高专业领域的语料占比。SpaCE2024更加注重针对大语言模型(Wang et al., 2024)的空间语义理解能力的测试，以选择题的形式考察以下五个层次的空间语义理解能力：（1）空间信息实体识别。要求从四个选项中选出文本空间信息的参照物。（2）空间信息角色识别。要求从四个选项中选出文本空间信息的语义角色，或者选出与语义角色相对应的空间表达形式。（3）空间信息异常识别。要求从四个选项中选出文本空间信息异常的语言表达。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

（4）空间方位信息推理。要求基于文本给出的推理条件进行空间方位推理，从四个选项中选出推理结果。（5）空间异形同义识别。要求从四个选项中选出能使两个文本异形同义或异义的空间义词语。

3 方法

在本次评测任务中，我们首先使用精心设计的提示获取模型在选择答案时的推理过程。随后，我们结合上下文学习（In-context Learning）与思维链（Chain of Thought）的方法，构建推理提示以提高模型在空间语义理解题目上的表现。其次，我们基于上下文学习的方法对模型进行微调，以进一步增强其在该任务中的性能。最后，通过投票策略整合多个模型的输出，获取最优结果。

3.1 基于上下文学习与思维链策略的模型推理

上下文学习，是大型语言模型中一种重要的能力。ICL允许模型利用输入中的上下文信息来指导其输出，从而在没有额外训练数据的情况下提高模型的泛化能力和灵活性。在大模型中，ICL的作用尤为显著，因为这些模型通常预训练在大规模的数据集上，能够捕捉到丰富的语言模式和世界知识。当给定一个包含示例和问题的上下文时，大型模型能够快速适应新的任务或领域，展现出对上下文的敏感性和适应性。这种能力使得大模型在处理诸如阅读理解、问题解答和代码生成等任务时表现出色，因为它们可以直接从上下文中提取相关信息来生成答案或解决方案。

思维链是一种新兴的策略，用于增强大型语言模型的推理能力。CoT的核心思想是模拟人类解决问题时的思维过程，即将复杂问题分解为一系列简单的推理步骤，并清晰地展示这些步骤。在大模型中，CoT的作用体现在能够引导模型逐步构建答案，而不是仅仅依赖于模式识别或记忆匹配。这种方法特别适用于那些需要多步骤逻辑推理的任务，如数学问题解决、科学推理或编程问题调试。通过明确展示思维过程，CoT不仅提高了模型输出的可解释性，还有助于提升模型在面对新颖或复杂问题时的表现。此外，CoT还可以作为一种正则化手段，帮助模型避免在复杂任务中陷入局部最优解，而是鼓励其探索更广泛的解决方案空间。

输入：
下面是一个关于实体识别的单项选择题，请根据下面的文本、问题、选项与答案给出一步一步的推理过程。
文本：人民大会堂和中国历史博物馆雄踞天安门广场东西两侧，广场中央耸立着人民英雄纪念碑和毛主席纪念堂，北有庄严的天安门城楼，南有巍峨的正阳门，广场面积44公顷，为世界最大广场。地处旧城中心的故宫珍藏了历代的珍奇异宝。北海、中南海、景山、天坛、中山公园、劳动人民文化宫、龙潭湖、先农坛、陶然亭、地坛等处，亭榭建筑精美、风景秀丽，是中外人士游类的集中地。
问题：()北有庄严的天安门城楼。
选项：
A:中国
B:中国历史博物馆
C:天安门广场
D:以上选项都不是
答案：C
一步一步的推理过程：

输出：
1. 阅读文本，寻找与问题相关的信息。文本中提到“人民大会堂和中国历史博物馆雄踞天安门广场东西两侧，广场中央耸立着人民英雄纪念碑和毛主席纪念堂，北有庄严的天安门城楼”。
2. 根据文本中的描述，可以确定天安门城楼位于广场的北侧。
3. 问题问的是“()北有庄严的天安门城楼”，需要找到天安门城楼所在的地理位置。
4. 从文本中提取的信息表明，天安门城楼是“北有”的主体，而这个“北有”的参照物是广场。
5. 因此，正确答案是天安门广场的北侧有天安门城楼。
6. 选项中C选项表示“天安门广场”，与文本中的信息一致，所以正确答案是C。
综上所述，答案是C

Figure 1: CoT数据生成示例

在本次评测中，我们选择Qwen1.5-72B-Chat作为直接推理时使用的模型，具体来说，我们首先依靠训练集与验证集数据得到模型在获取答案时的推理过程，prompt设置与模型的输出入图1所示，通过该步骤我们获取模型在推理答案时的思维过程。其次，在对测试集进行推理时结

(c) 2024 中国中文信息学会计算语言学专业委员会

合上下文学习与思维链的策略，在prompt设置中添加同类型数据示例，指导模型先对问题进行分析推理，最后总结出答案。

3.2 基于上下文学习的模型有监督微调

随着模型大小和语料库大小的扩展，大型语言模型显示出上下文学习能力，它们从上下文中的几个示例中学习，大量研究表明大型语言模型可以通过ICL有效执行各种复杂任务。形式上，给定一个查询输入文本x 和一组候选答案$Y = \{y_1, \ldots, y_m\}$，预训练语言模型M 在条件集C 的条件下选择得分最高的候选答案作为预测。集合C 包括一个可选的任务说明I 和K 个示例，因此$C = \{I, s(x_1, y_1), \ldots, s(x_k, y_k)\}$ 或$C = \{s(x_1, y_1), \ldots, s(x_k, y_k)\}$，其中$s(x_k, y_k, I)$ 表示根据任务用自然语言编写的示例。候选答案y_j 的概率可以通过模型M 的整个输入序列的评分函数f 来表示。

$$P(y_j|x) \triangleq f_M(y_j, C, x) \tag{1}$$

最终预测的标签$\hat{y}$ 是具有最高概率的候选答案:

$$y = \arg\max_{y_j \in Y} P(y_j|x) \tag{2}$$

输入：
下面是关于实体识别的单项选择题，请根据下面的文本、问题与选项给出正确的答案。
文本：父亲母亲站在后面，两个姐姐分立两边，而我，我把六岁的自己放在了中间，非常小鸟依人，万千宠爱于一身的样子，真正像一个自幼娇惯父母疼爱的小女孩……我的眼睛湿润了。从彩色喷墨打印机里输出的这张“全家福”，被我当做生日礼物送给了母亲。她刚看了第一眼就哭了。母亲抓着画像，大力地拍着我的肩哽咽：“好女儿，谢谢你，我们终于有了一张全家福了！”
问题：两个姐姐分立()两边。
选项：
A:打印机
B:父亲
C:我
D:以上选项都不是
答案：C

示例数据...

文本：人民大会堂和中国历史博物馆雄踞天安门广场东西两侧，广场中央耸立着人民英雄纪念碑和毛主席纪念堂，北有庄严的天安门城楼，南有巍峨的正阳门，广场面积44公顷，为世界最大广场。地处旧城中心的故宫珍藏了历代的珍奇异宝。北海、中南海、景山、天坛、中山公园、劳动人民文化宫、龙潭湖、先农坛、陶然亭、地坛等处，亭榭建筑精美、风景秀丽，是中外人士游类的集中地。
问题：()北有庄严的天安门城楼。
选项：
A:中国
B:中国历史博物馆
C:天安门广场
D:以上选项都不是
答案：

输出：
答案：C

Figure 2: 上下文学习指令微调数据示例

虽然大型语言模型已经展示了强大的上下文学习能力，但一些研究也表明，通过预训练和上下文学习推理之间的持续训练阶段，这种能力可以进一步增强(Wei et al., 2023; Chen et al., 2022)。因此，我们通过构建上下文学习的指令训练数据来增强大型语言模型的上下文能力，并通过监督式指令微调消除预训练任务与下游上下文学习任务之间的差距。具体来说，我们利用随机选择方法选择同类型的10个示例数据，随后通过人工挑选的方式从中选择5个数据作为prompt中的示例数据，构造有监督的ICL训练数据，然后基于有监督的ICL数据训练Qwen1.5-7B-Chat模型，示例数据如图2所示。

(c) 2024 中国中文信息学会计算语言学专业委员会

3.3 模型投票

投票策略通过集成多个模型或同一模型在不同推理路径上的预测结果，可以有效减少随机误差和模型偏差，提高整体的鲁棒性。由于投票策略能够平衡不同模型的优势，特别是在面对复杂或模糊的输入时，可以综合考虑各个模型的判断，从而得出更加全面和可靠的答案。除此之外，投票机制还可以作为一种正则化手段，防止模型过拟合于特定的数据模式，增强模型在未知数据上的泛化能力。同时在微调训练过程中，我们发现不同模型，甚至同结构的模型在不同训练轮次的验证集预测结果上仍存在差异，我们使用投票策略对模型预测结果进行处理，最后取得了最优的提交成绩。

4 实验

4.1 数据介绍

SpaCE2024更加注重针对大语言模型的空间语义理解能力的测试，宗旨是在一个测试数据集上考察机器中文空间语义理解的综合能力。SpaCE2024不再划分子任务，以选择题的形式考察模型在空间信息实体识别、空间信息角色识别、空间信息异常识别、空间方位信息推理与空间异形同义识别五个层次的空间语义理解能力，数据分布如表1所示。

题目类型	Train			Dev			Test		
	单选	多选	Total	单选	多选	Total	单选	多选	Total
实体识别	937	161	1098	226	24	250	513	87	600
角色识别	1074	19	1093	186	4	190	776	24	800
异常识别	1077	10	1077	40	0	40	530	0	530
空间推理	909	301	1210	468	207	675	1533	537	2070
同义识别	4	1	5	44	11	55	541	139	680

Table 1: SpaCE24数据集数据分布

SpaCE2024 使用准确率（Acc, Accuracy）作为评价指标和排名依据，如公式3所示：

$$Acc = \text{命中正确答案的题数/题目总数} \tag{3}$$

除此之外还使用稳定性来衡量机器表现的稳定程度，不作为排名依据。以比赛题及其对应的干扰题为一组，如公式4所示：

$$\text{稳定性} = \text{组内答案完全一致的组数/总组数} \tag{4}$$

4.2 基于上下文学习与思维链策略的中文空间语义理解模型实验结果

我们首先基于SpaCE2024验证集进行了一系列实验尝试，实验结果如表2所示。

DataSet	Metrics			Accuracy					
	ICL	CoT	Train	Total	实体识别	角色识别	异常识别	空间推理	同义识别
SpaCE24_dev	0-shot	w/o	w	0.5570	0.8560	0.9526	0.8750	0.3170	0.5454
SpaCE24_dev	5-shot	w	w	0.4330	0.6360	0.9316	0.4250	0.2119	0.5091
SpaCE24_dev	5-shot	w	w/o	0.4240	0.6720	0.9263	0.5750	0.1719	0.5455
SpaCE24_dev	5-shot	w/o	w	**0.5686**	**0.8440**	**0.9737**	**0.9250**	**0.3304**	**0.5818**

Table 2: SpaCE24验证集的实验结果

表2展示了在SpaCE24验证集上不同策略的具体得分，其中ICL表示在训练与推理过程中使用上下文学习策略在prompt中加入任务数据示例，CoT表示在训练与推理过程中使用思维链策略在模型输出最后答案前输出推理过程，Train表示是否使用SpaCE24训练集进行模型训练，当不进行模型训练时我们选用的模型为Qwen1.5-72B-Chat，当进行模型训练时我们选用的模型为Qwen1.5-7B-Chat。实验结果表明，在训练和推理阶段使用上下文学习策略与思维链策略对模型效果有较大的提升，值得注意的是，使用思维链策略参与模型训练后会导致模型效果下

(c) 2024 中国中文信息学会计算语言学专业委员会

降。这里我们推测原因为思维链数据是模型根据问题与答案生成的伪推理过程，中间部分步骤存在错误，在训练过程中引入错误的推理过程会影响模型本身的推理能力，抽样检查的思维链推理过程也证明了我们推测的原因。如果有高质量的思维链推理数据即完全正确的推理过程，在模型训练过程中使用思维链数据应当会提升的表现。

4.3 基于投票策略的中文空间语义理解模型实验结果

基于验证集的实验结果，我们在SpaCE2024测试集上进行推理并共提交了6次结果，实验结果如表3所示。

DataSet	Metric	Vote	Accuracy					
			Total	实体识别	角色识别	异常识别	空间推理	同义识别
Baseline	\	w/o	0.4792	0.7509	0.8818	0.6860	0.2196	0.4200
SpaCE24_test	TeleAI_test_1	w/o	0.5991	0.8895	0.9312	0.8440	0.3471	0.5538
SpaCE24_test	TeleAI_test_2	w	0.5958	0.8895	0.9273	0.8480	0.3373	0.5631
SpaCE24_test	TeleAI_test_3	w/o	0.5898	0.8912	0.9364	0.8360	0.3265	0.5523
SpaCE24_test	TeleAI_test_4	w/o	0.5885	0.8947	0.9260	0.8440	0.3255	0.5492
SpaCE24_test	TeleAI_test_5	w	0.5958	0.8895	0.9273	0.8480	0.3373	0.5631
SpaCE24_test	TeleAI_test_6	w	**0.6024**	**0.8947**	**0.9364**	**0.8480**	**0.3471**	**0.5631**

Table 3: SpaCE24测试集的实验结果

表3展示了在SpaCE24测试集上每次提交结果的具体得分，其中Vote表示是否使用了投票策略，我们发现不同策略下的模型，甚至同结构的模型在不同训练轮次的验证集预测结果上存在差异，因此我们挑选了验证集中表现较好的模型作为投票模型。TeleAI_test_2采用多数投票策略来汇总各模型的预测结果，TeleAI_test_5引入了加权投票机制，根据各模型在验证集上的性能为其预测结果分配特定的权重，TeleAI_test_6进一步改进了这一方法，不仅为各模型设置权重，还为不同类型的题目分配权重，从而通过更细致且上下文敏感的聚合过程，提高最终预测的精确度和鲁棒性。由实验结果可以看出投票策略在该任务上可以平衡不同模型的优势，有效减少随机误差和模型偏差，提高整体的性能，我们在测试集上最终取得了0.6024的准确率，与基线模型相比取得了25.71%的相对提升，在实体识别任务上与基线模型相比取得了19.15%的相对提升，在角色识别任务上与基线模型相比取得了6.19%的相对提升，在异常识别任务上与基线模型相比取得了23.62%的相对提升，在空间推理任务上与基线模型相比取得了58.06%的相对提升，在同义识别任务上与基线模型相比取得了34.07%的相对提升。我们的模型在实体识别、角色识别、异常识别与同义识别任务中达到了较高的准确率，但是在空间推理任务上准确率较低，可以看出大模型在推理能力上仍具有较大的提升空间。

5 总结与展望

在本次CCL2024中文空间语义理解评测任务（SpaCE2024）中，我们基于上下文学习与思维链策略精心设计了prompt并结合微调的方式激发大语言模型的空间语义理解能力，构建了一个高效的空间语义理解系统。在最终的评估中，我们队伍提交的系统在测试集实体识别与异常识别题目中排名第一，综合准确率为0.6024，综合排名第一。

参考文献

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens, and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In *Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 255–262.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. In {** SEM 2012*}*: The First Joint Conference on Lexical and Computational*

(c) 2024 中国中文信息学会计算语言学专业委员会

Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation {(SemEval 2012)}, volume 2, pages 365–373. ACL.

Xiao Liming, Sun Chunhui, Zhan Weidong, Xing Dan, Li Nan, Wang Chengwen, and Zhu Fangwei. 2023. Space2022 中文空间语义理解评测任务数据集分析报告(a quality assessment report of the chinese spatial cognition evaluation benchmark). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 547–558.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979.

Liming Xiao, Weidong Zhan, Zhifang Sui, Yuhang Qin, Chunhui Sun, Dan Xing, Nan Li, Fangwei Zhu, and Peiyi Wang. 2023. Ccl23-eval 任务4 总结报告: 第三届中文空间语义理解评测(overview of ccl23-eval task 4: The 3rd chinese spatial cognition evaluation). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 150–158.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 2022. 空间语义理解能力评测任务设计的新思路—space2021 数据集的研制. 语言文字应用, 2:99–110.

CCL 2024

CCL24-Eval任务3系统报告：基于上下文学习的空间语义理解

武洪艳
广东外语外贸大学
信息科学与技术学院
2754976781@qq.com

林楠铠✉
广东工业大学
计算机学院
neakail@outlook.com

曾培健
广东工业大学
计算机学院
lil_ken@163.com

郑伟雄
广东工业大学
计算机学院
1473391854@qq.com

蒋盛益
广州商学院
信息技术与工程学院
jiangshengyi@163.com

阳爱民
广东工业大学
计算机学院
amyang18@163.com

摘要

空间语义理解任务致力于使语言模型能够准确解析和理解文本中描述的物体间的空间方位关系，这一能力对于深入理解自然语言并支持复杂的空间推理至关重要。本文聚焦于探索大模型的上下文学习策略在空间语义理解任务上的有效性，提出了一种基于选项相似度与空间语义理解能力相似度的样本选择策略。本文将上下文学习与高效微调融合对开源模型进行微调，以提高大模型的空间语义理解能力。此外，本文尝试结合开源模型和闭源模型的能力处理不同类型的样本。实验结果显示，本文所采用的策略有效地提高了大模型在空间语义理解任务上的性能。

关键词： 上下文学习 ； 高效微调 ； 样本选择

System Report for CCL24-Eval Task 3: Spatial Cognition Evaluation Based on In-context Learning

Hongyan Wu
Guangdong University of Foreign Studies
School of Information Science and Technology
2754976781@qq.com

Nankai Lin✉
Guangdong University of Technology
School of Computer Science and Technology
neakail@outlook.com

Peijian Zeng
Guangdong University of Technology
School of Computer Science and Technology
lil_ken@163.com

Weixiong Zheng
Guangdong University of Technology
School of Computer Science and Technology
1473391854@qq.com

Shegyi Jiang
Guangzhou College of Commerce
School of Information Technology and Engineering
jiangshengyi@163.com

Aimin Yang
Guangdong University of Technology
School of Computer Science and Technology
amyang18@163.com

Abstract

The Spatial Cognition Evaluation task aims to enable language models to parse and understand spatial relationships between objects described in the text, which is crucial for a deeper understanding of natural language and complex spatial reasoning. In this paper, we investigate the effectiveness of in-context learning for large language

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十三届中国计算语言学大会论文集，第113页-第121页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

models on the spatial cognition evaluation task and propose a demonstration example selection strategy based on option similarity and spatial capability similarity. In this paper, we infuse in-context learning with parameter-efficient fine-tuning to fine-tune the open-source model, improving the spatial cognition capability of the large language model. Moreover, this paper attempts to combine the capabilities of open-source and closed-source models to tackle different types of samples. Experimental results reveal that our proposed strategy effectively improves the performance of the large language model on spatial cognition evaluation tasks.

Keywords: In-context Learning , Parameter-efficient Fine-tuning , Demonstration Example Selection

1 引言

空间范畴是人类认知的一个核心基础,大量空间信息存在于自然语言文本中。在通往人工智能的道路上，空间语义理解是不可绕开的一环。要准确理解文本中表达的空间语义，不仅需要语言知识，还需要使用空间认知能力，构建空间场景，并基于世界知识进行空间方位信息相关的推理。空间语义理解已成为自然语言处理领域中一个热门的研究主题，这在需要理解空间关系的导航系统、问答系统中具有关键的作用。近年来，随着大数据和深度学习技术的不断进步，越来越多的研究者开始研究如何使机器像人类一样能够准确理解自然语言中的空间信息。空间信息提取是理解文本中空间信息的关键，目前的研究主要致力于提取空间元素和空间关系去提升机器的空间语义理解能力。许多自然语言处理技术和机器学习方法已被应用于空间信息提取。例如，条件随机场(CRF)模型(Lafferty et al., 2001)被用于空间元素提取，支持向量机(SVM) (Suykens and Vandewalle, 1999; Roberts and Harabagiu, 2012a)和卷积神经网络(CNN) (Mazalov et al., 2015)模型被用于空间关系提取。各种语言资源，如GloVe (Pennington et al., 2014)、WordNet (Salaberri et al., 2015)和PropBank (Salaberri et al., 2015)也被用于空间信息提取。以ELMO(Peters et al., 2018)和GPT (Yang et al., 2019)为代表的语言模型在建模文本语义关系方面展现出优异的性能，被开发用于命名实体识别和语义角色标注，这使得预训练模型可以很容易应用到空间信息的提取。Shin等人(2020)提出了一个基于BERT的空间信息提取模型用于空间元素提取和空间关系提取。

为了推进空间语义理解任务的发展，第四届中文空间语义理解评测关注于大语言模型的空间语义理解能力测试，SpaCE2024从空间信息实体识别、空间信息角色识别、空间信息异常识别、空间方位信息推理和空间异形同义识别五个层次测试机器中文空间语义理解的能力。这些测试语料来源于不同的领域，涵盖了空间语义理解的不同方面，旨在综合考察机器在理解自然语言中的空间信息方面的能力。本文基于中文空间语义理解任务探索了大模型的上下文学习在该任务上的有效性，我们基于选项相似度与空间语义理解能力相似度为测试样本选择示例。本文将上下文学习融入大模型的高效微调阶段，以提高大模型的空间语义理解能力。此外，本文尝试将开源模型和闭源模型融合用于处理不同难度的样本。实验结果显示，本文所采用的策略有效地提高了大模型在空间语义理解任务上的性能。

2 相关研究

目前，空间语义理解任务要求模型关注于不同层面的空间语义信息。SemEval 2012 (Kordjamshidi et al., 2012)，SemEval 2013和SemEval 2015提出了面向空间语义理解的多个评测，分别关注模型的静态空间和动态空间的语义角色标注能力。SemEval 2012引入了一个主要关注于静态空间关系的空间角色标注任务，SemEval 2013将空间关系扩展到动态，以捕获细粒度的语义。SpaCE中文空间语义理解评测则对模型提出了更高的空间语义理解要求。SpacE 2021提出了三个子任务，分别是空间语义正误判断、空间语义异常归因合理性判断和空间语义判断与归因联合任务。SpacE 2022扩大了任务类型，增加了信息标注任务，要求机器在空间信息异常的文本中识别异常片段，在空间信息正常的文本中进行细粒度的语义角色标注。SpacE 2023则在空间语义异常和空间语义角色标注的基础上增加了对空间场景的关注，考察机器对空间场景异同的判断。

(c) 2024 中国中文信息学会计算语言学专业委员会

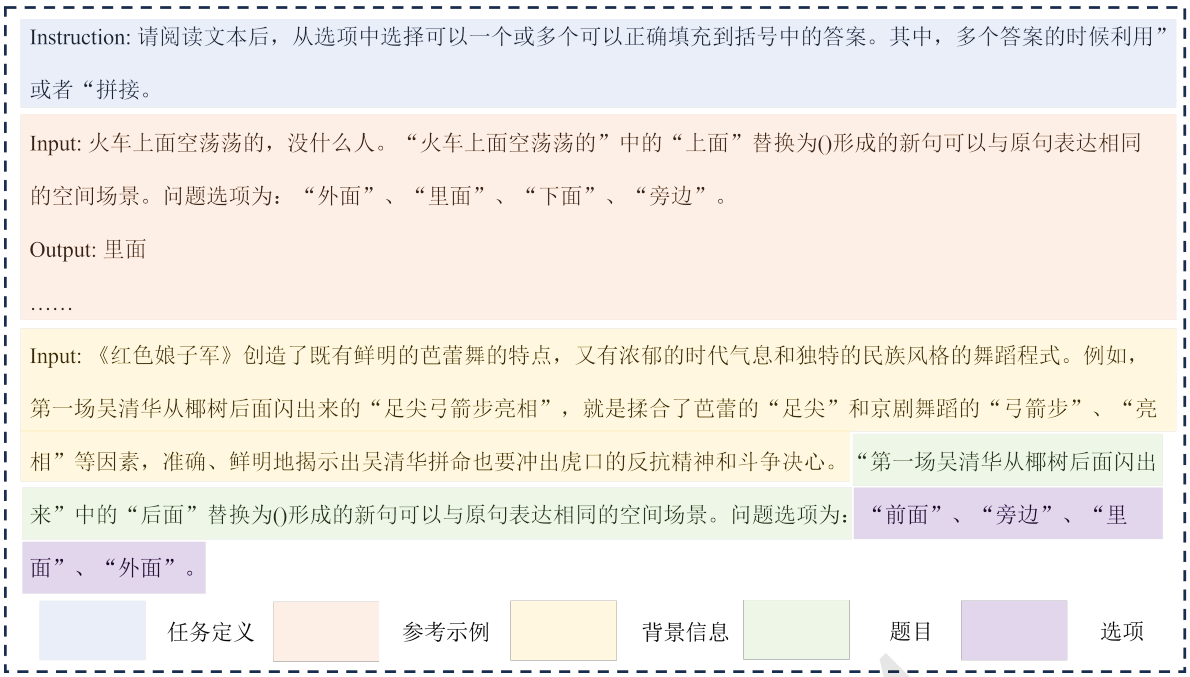


Figure 1: 大模型输入模板

早期的空间语义信息提取主要采用基于机器学习的方法。Nichols和Botros(2015)基于CRF和SVM提出了SpRL-CWW模型。该方法基于CRF使用多种输入特征来提取空间元素，例如使用GloVe的词嵌入、命名实体、词性和依存解析标签。SVM则用于从所有可能的三元组组合中过滤出正确的三元组得到空间关系。D'Souza和Ng(2012b)基于SVM提出了UTD-SpRL模型，该模型采用贪婪特征选择技术生成多个不同的特征并对空间关系相关的参数执行联合检测。一些研究者尝试将多种语言资源用于空间语义理解任务，以补充空间信息。Salaberri等人(2015)提出的X-Space模型使用WordNet中包含的地点、位置、方位等节点信息进行空间元素提取，同时使用PropBank中的参数信息对空间关系进行了分类。

随着深度学习的发展，基于神经网络的方法被开发用于空间信息的提取。Mazalov等人(2015)提出了一个基于卷积神经网络的语义角色标注系统来提取空间角色及其关系，成功地适应于空间信息提取。Dan等人(2020)使用BERT对给定图像的两个实体之间的空间关系进行预测。该模型包含一个由前馈网络实现的空间模型和一个BERT语言模型组成，其中语言模型被作为补充特征来预测图像中不可见的关系。尽管该方法使用了BERT作为语言模型，但仅限于对图像中给定的实体检测关系。Shin等人(2020)提出了BERT空间模型，使用BERT从原始文本中提取空间元素，确定它们对应的空间角色，进一步使用R-BERT对空间角色的关系进行提取，通过两个模块联合有效提高了空间信息提取的性能。本文则基于上下文学习和高效微调探索了大模型在空间语义理解任务上的综合能力，为未来的工作提供了新的研究思路。

3 基于上下文学习的空间语义理解

上下文学习使模型能够在不更新任何参数的情况下从特定任务的示例中学习，它将一些训练样本作为提示添加到推理阶段的测试样本之前。上下文学习的关键在于示例样本的选择，本文基于空间语义理解任务精心设计了样本选择策略，提出根据选项相似度和空间语义理解能力相似度为测试样本选择合适的示例样本。然后，本文将选择得到的示例样本作为前缀添加到测试样本之前构建指令提示，作为大模型的输入。对于开源大模型，本文在高效微调阶段和推理阶段均融合了示例样本作为模型的输入；而对于闭源大模型，本文仅在推理阶段使用示例样本。

3.1 示例样本选择

给定一个包含n个样本的数据集$D = \{(x_1, q_1, t_1, o_1, y_1), ...(x_n, q_n, t_n, o_n, y_n)\}$，对于第$i(i \in$

(c) 2024 中国中文信息学会计算语言学专业委员会

$[1,n]$)个样本，x_i，q_i 和t_i分别代表题目的背景信息、问题以及考察的空间语义理解能力类型，o_i和y_i表示该样本对应的选项和答案，其中，选项o_i包含四个不同的选项文本，即$o_i = \{o_i^1, o_i^2, o_i^3, o_i^4\}$。对于给定的样本$i$以及在数据集$D$中的另一个样本$j$，本文计算两个样本之间的选项相似度$s_o^{ij}$：

$$s_o^{ij} = \frac{|o_i \cap o_j|}{|o_i|}. \tag{1}$$

此外，对于样本i和样本j，如果两个样本考察的空间语义理解能力类型一样，两个样本之间的空间理解能力相似度则为1，否则，其空间能力相似度为0。具体地，两个样本之间的空间语义理解能力相似度s_c^{ij}定义如下：

$$s_c^{ij} == \begin{cases} 1 & t_i = t_j, \\ 0 & t_i \neq t_j. \end{cases} \tag{2}$$

基于上述选项相似度和空间语义理解能力相似度两种不同的相似度，本文进一步将两种相似度相加作为两个样本总的相似度s^{ij}：

$$s^{ij} = s_o^{ij} + s_c^{ij}. \tag{3}$$

对于样本i，本文计算该样本与数据集D中其他的所有样本的相似度，然后选择相似度最大的Q个样本作为示例样本用于上下文学习。

3.2 提示设计

中文空间语义理解任务要求大模型根据输入的信息、题目以及选项，选择正确的答案。本文基于此精心设计了指令提示来引导大模型输出期望的回复。本文设计的提示模板如图1所示，包括五个部分：

(1)任务定义：该部分详细描述了大模型需要完成的具体任务。

(2)参考示例：这里提供了基于选项相似度和空间语义理解能力相似度所选择得到的Q个示例样本，提供了大模型执行空间语义理解任务时的输出示例。

(3)背景信息：这里对应题目的背景信息，机器需要阅读理解后回答问题。

(4)题目：机器需要回答的问题。形式上是一个句中有括号的陈述句。

(5)选项：这里对应题目的选项，在微调过程中作为输入的一部分。

3.3 高效微调

指令微调（Instruction tuning）训练语言模型遵循专门的自然语言指令，从而提高模型的性能和理解能力。这需要在特定任务的数据集上基于提示-回答的形式对模型进行训练。针对开源大模型，本文采用主流的高效微调方法LoRA微调。LoRA微调冻结预先训练的模型权重，并将可训练的低秩矩阵注入到每一层中，从而显著减少需要训练的参数量，从而实现使用更少的计算资源来训练大模型。具体来说，对于使用$W_0 \in \mathbb{R}^{d\times k}$ 作为权重矩阵的线性层，LoRA引入一组低秩矩阵进行训练，分别为$B \in \mathbb{R}^{d\times r}$ 和$A \in \mathbb{R}^{r\times k}$，其中，$k$表示输入的维度，$d$表示输出的维度，$r$是预定义的等级。对于输入$x$，前向传播过程的输出表示为：

$$h = W_0 x + \Delta W x = W_0 x + BAx. \tag{4}$$

在微调过程中，只有矩阵B 和A 被更新，而W_0 保持静态并且不接收梯度更新。同时，本文在微调过程中使用了因果语言建模（Causal Language Modeling, CLM）技术，该方法以自回归方式训练模型，根据给定的输入标记序列$(x_0, x_1, x_2, \ldots, x_{i-1})$ 预测下一个标记x_i，最终使用负对数似然函数来计算训练过程的损失：

$$L_{\mathrm{CLM}}(\Theta) = \mathbb{E}_{x\sim D}\left[-\sum_i \log p(x_i \mid x_0, x_1, \ldots, x_{i-1}; \Theta)\right], \tag{5}$$

其中，Θ表示模型的参数。

(c) 2024 中国中文信息学会计算语言学专业委员会

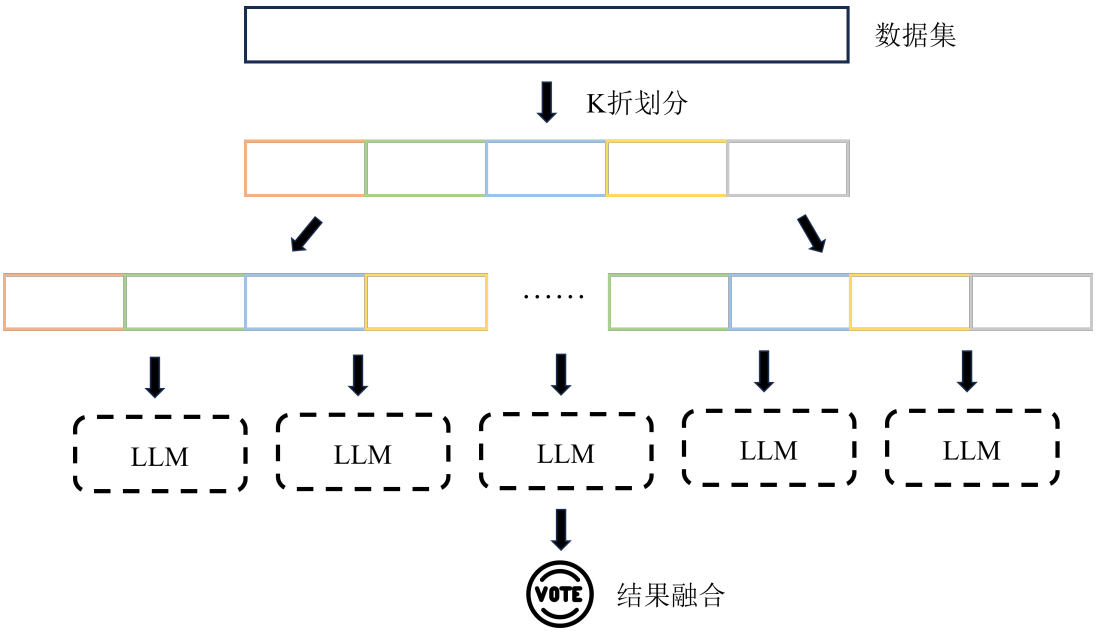


Figure 2: 模型融合流程

4 基于K折数据划分的模型融合

如图2所示，本文采用了K折交叉验证策略来增强模型的泛化能力，具体地，我们首先将整个数据集均匀划分为K个子集。在此过程中，本文选取$K = 5$作为交叉验证的折数。对于划分的K折数据，我们从中随机选取一折数据作为验证集，而其余$K - 1$折的数据则用作训练集。通过这种方式，我们分别训练了K个独立的模型。在模型评估阶段，每个模型都独立地对最终的测试集进行预测。为了得到更为稳健的预测结果，我们采用了集成学习中的投票机制来融合这K个模型的输出，通过投票法确定最终的预测结果。这种方法不仅可以减少模型在特定数据子集上的过拟合风险，同时能够有效提升模型在未知数据上的预测准确度。

空间信息实体识别数据示例：

输入：

{
“**qid**”：“1-train-s-913”，
“**text**”：“上海市浦东新区人民检察院指控，2021年1月7日11时45分许，被告人王某2驾驶牌号为沪CEXXXX的小型面包车沿本区临港大道由东向西行驶至祥凯路东约100米处时(时速大于67公里)，面包车左偏方向驶入中央绿化带，前部撞击路灯杆后向左侧翻，造成面包车前排乘员胡运送医途中死亡，王某2及后排乘员李某受伤。”，
“**question**”：“()前部撞击路灯杆后向左侧翻。”，
“**option**”：{“A”：“路灯杆”，“B”：“绿化带”，“C”：“面包车”，“D”：“以上选项都不是”}
}

输出：

{“**qid**”：“1-train-s-913”，“**answer**”：[“C”]}

Figure 3: 空间信息实体识别数据示例

CCL 2024

Table 1: 数据集分布

训练集	验证集	测试集	总计
4,483	1,210	4,680	10,373

Table 2: 主实验结果

大模型	上下文学习	K折融合	总分	实体识别	角色识别	异常识别	空间推理	同义识别
ChatGLM3	否	是	44.22	64.91	79.48	59.20	25.44	31.69
ChatGLM3	是	否	50.53	74.21	85.20	67.80	31.23	36.00
ChatGLM3	是	是	52.34	73.68	86.62	73.00	32.40	39.69
文心4	是	否	53.89	67.19	93.12	76.60	28.97	56.46
文心4	是	是	55.21	69.82	92.86	76.20	30.68	58.62
ChatGLM3 + 文心4	是	是	56.45	73.68	92.86	76.20	32.40	58.62

5 实验与分析

5.1 实验设置

本文基于RTX 8000 48G GPU完成了所有实验，基于PyTorch框架[0]的代码开发模型。在开源大模型的微调方面，我们采用了支持LoRA微调的LMFlow框架[1]。此外，本次实验中，我们使用的基座模型包括开源大模型ChatGLM3-base[2] (Du et al., 2022) 和闭源大模型文心4[3]。对ChatGLM3-base模型进行微调时，我们将批处理大小设定为1，训练了7个Epoch，学习率设置为5e-5。

5.2 数据集

本次中文空间语义理解任务数据集共10,373条样本，一共约100万字符，训练集、验证集和测试集规模如表1所示。所有样本均以选择题的形式呈现，以空间信息实体识别类型数据为例，如图3所示，每条样本包含qid、text、question、option和answer五个字段，其中qid代表题目编号，采用“能力代号-子集类别-题目类别-题目号”的策略。对于能力代号的表示，1代表实体识别，2代表角色识别，3代表异常识别，4代表空间推理，5代表同义识别。对于子集类别的表示，train代表训练集，dev代表开发集，test代表测试集。对于题目类别的表示，使用s代表答案只有1个的单选题，m代表答案有2个及以上的多选题。text代表该样本的文本材料，即机器需要阅读的文本。question代表机器需要回答的问题，形式上是一个有括号的陈述句。option代表选择题的选项，采用“选项字母-选项内容”键值对的形式，共有四个键-值对。answer代表选择题的答案。

5.3 评估指标

SpaCE2024使用准确率（Accuracy）作为评价指标，对于每个待检查的预测结果，如果与标准答案一致，则认为预测正确。中文空间语义理解任务中的准确率被定义为命中正确答案的题目数量$N_{correct}$占总的测试样本数量N_{total}的比例，计算过程如下：

$$Accuracy = \frac{N_{correct}}{N_{total}}. \tag{6}$$

5.4 实验结果

主实验结果 本文将评测提供的原始训练集和验证集合并后用于模型微调，微调过程进行五折划分与模型融合，实验结果如表2所示。可以发现，基于ChatGLM3融合上下文学习进行高效微调时，模型的性能相比仅使用指令微调时取得了显著提升，从44.22提升至50.53。同时，融合示例样本进行微调后的ChatGLM3在实体识别和角色识别两个类型上取得了优异的性能，分别

[0]https://github.com/pytorch/pytorch
[1]https://github.com/OptimalScale/LMFlow
[2]https://github.com/THUDM/ChatGLM3
[3]https://yiyan.baidu.com/

(c) 2024 中国中文信息学会计算语言学专业委员会

Table 3: 参数探究结果

Q	交叉验证集	测试（K折融合）
-	45.83	44.22
3	48.21	49.47
5	48.61	49.93

Table 4: 不同规模数据下的实验结果

数据	总分	实体识别	角色识别	异常识别	空间推理	同义识别
原始训练集	49.93	71.75	85.84	72.80	28.04	39.38
原始训练集+ 验证集	52.34	73.68	86.62	73.00	32.40	39.69

达到了74.21和85.20。但是，ChatGLM3总体上在空间推理和同义识别两个类别上表现较差，这种现象可能是由于空间推理往往涉及多层次的理解和推断，如物体的相对位置、运动路径等，对模型的推理能力有更高的要求。而同义识别需要对文本的丰富语义有深刻理解，依赖于其语境和上下文信息，涉及细粒度的语义分析，在通用语料训练的大模型则很难捕获细粒度的语义信息。此外，实验结果表明文心4模型在整体表现上比ChatGLM3性能更好，但是在实体识别能力与空间推理能力上表现不如ChatGLM3，因此，本文将文心4和ChatGLM3两个大模型的结果进行融合，分别处理不同空间语义理解能力的题目，以提高系统在中文空间语义理解任务上的表现。总体上，ChatGLM3和文心4在空间推理类型上均表现不佳，表明了大模型在推理任务上仍然面临挑战。

融合策略有效性验证 如表2所示，无论是在开源大模型ChatGLM3还是闭源大模型文心4上，利用五折划分分别构建不同的模型进行推理后融合的效果，均高于不采用融合策略的模型。使用了融合策略后，ChatGLM3的总体评分从50.53提升至52.34，而文心4的总体评分则从53.89提升至55.21，表明我们设计的融合策略有效提升了模型的空间语义理解能力。

参数探究结果 本文以ChatGLM3为实验模型，探究了不同的示例样本数量对于模型性能的影响，结果如表3所示。表中的实验仅采用原始训练集进行五折交叉验证与模型融合，可以发现，不同的示例样本数量Q对于模型的性能具有一定的影响，当选择5个参考示例时的模型性能优于选择3个参考示例时的模型。同时，考虑到随着参考示例的数量增加，所需的训练成本也同步增加，因此，本文采用示例样本数量Q为5作为主要实验设置。

数据规模的影响 本文进一步在ChatGLM3上进行实验，分别探究仅采用原始训练集进行五折划分与模型融合，以及采用原始训练集和验证集进行五折划分和模型融合两种数据规模的效果。在测试集的实验结果如表4所示。可以观察到，增加训练数据量能显著提高模型性能。通过仅添加1000多条样本进行训练，模型的性能已从49.93提高至52.34。这表明，目前的空间语义理解数据集的规模还不足以完全发挥开源大模型在此任务上的潜力。因此，采用数据增强策略以进一步提升模型性能，是未来值得探索的一个研究方向。

6 总结

本文基于上下文学习策略探索了大模型在中文空间语义理解任务上的表现，揭示了大模型在该任务上的潜力。此外，通过基于选项相似度与空间语义理解能力相似度的策略去选择示例样本，研究结果显示，我们提出的样本选择方法能有效提高大模型在中文空间语义理解任务上的性能。实验结果表明，在高效微调阶段融合上下文学习策略能显著提升模型的空间语义理解能力，特别是在开源和闭源模型的综合使用上表现出色。此外，本文的研究不仅提高了模型的实际应用效果，也为未来在空间语义理解领域的研究提供了新的思路和框架。尽管取得了一定的成果，但模型在空间推理类型和同义识别类型上仍存在显著的差距，在接下来的工作中，如何有效提升模型的推理能力和同义区分能力可能是新的重点和难点。空间语义理解仍有待进一步深入研究。未来的工作可以在优化样本选择算法、数据增强策略、探索更多维度的语义关系，以及提升模型在更广泛自然语言处理任务中的泛化能力等方面进行。

(c) 2024 中国中文信息学会计算语言学专业委员会

致谢

本研究受国家社会科学基金项目（No.22BTQ045）资助。

参考文献

Soham Dan, Hangfeng He, and Dan Roth. 2020. Understanding spatial relations through multiple modalities. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2368–2372. European Language Resources Association.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 task 3: Spatial role labeling. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 365–373, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Alexey Mazalov, Bruno Martins, and David Martins de Matos. 2015. Spatial role labeling with convolutional neural networks. In Ross S. Purves and Christopher B. Jones, editors, *Proceedings of the 9th Workshop on Geographic Information Retrieval, GIR 2015, Paris, France, November 26-27, 2015*, pages 12:1–12:7. ACM.

Eric Nichols and Fadi Botros. 2015. Sprl-cww: Spatial relation classification with independent multi-class models. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 895–901. The Association for Computer Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Kirk Roberts and Sanda M. Harabagiu. 2012a. Utd-sprl: A joint approach to spatial role labeling. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 419–424. The Association for Computer Linguistics.

Kirk Roberts and Sanda M. Harabagiu. 2012b. Utd-sprl: A joint approach to spatial role labeling. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 419–424. The Association for Computer Linguistics.

(c) 2024 中国中文信息学会计算语言学专业委员会

Haritz Salaberri, Olatz Arregi, and Beñat Zapirain. 2015. Ixagroupehuspaceeval: (x-space) A wordnet-based approach towards the automatic recognition of spatial information following the iso-space annotation scheme. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 856–861. The Association for Computer Linguistics.

Hyeong Jin Shin, Jeong Yeon Park, Dae Bum Yuk, and Jae Sung Lee. 2020. Bert-based spatial information extraction. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 10–17.

Johan A. K. Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural Process. Lett.*, 9(3):293–300.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

CCL 2024

(c) 2024 中国中文信息学会计算语言学专业委员会

Overview of CCL24-Eval Task 3: The Fourth Evaluation on Chinese Spatial Cognition

Liming Xiao[‡] **Nan Hu**[‡] **Weidong Zhan**[†,*] **Yuhang Qin**[‡] **Sirui Deng**[‡]

Chunhui Sun[†] **Qixu Cai**[‡] **Nan Li**[†]

Department of Chinese Language and Literature, Peking University

Center for Chinese Linguistics, Peking University

[†]{zwd,sch,linan2017}@pku.edu.cn

[‡]{lmxiao,hunan,hezonglianheng,d2020sr,cqx}@stu.pku.edu.cn

Abstract

The Fourth Chinese Spatial Cognition Evaluation Task (SpaCE 2024) presents the first comprehensive Chinese benchmark to assess spatial semantic understanding and reasoning capabilities of Large Language Models (LLMs). It comprises five subtasks in the form of multiple-choice questions: (1) identifying spatial semantic roles; (2) retrieving spatial referents; (3) detecting spatial semantic anomalies; (4) recognizing synonymous spatial expression with different forms; (5) conducting spatial position reasoning. In addition to proposing new tasks, SpaCE 2024 applied a rule-based method to generate high-quality synthetic data with difficulty levels for the reasoning task. 12 teams submitted their models and results, and the top-performing team attained an accuracy of 60.24%, suggesting that there is still significant room for current LLMs to improve, especially in tasks requiring high spatial cognitive processing.

1 Introduction

Spatial expressions have long been a focus of cognitive linguists because they are not only a high-frequency phenomenon in human language but also embody the fundamental mechanisms of how humans perceive the world (Talmy, 1983). In recent years, some NLP evaluation tasks have also sought to explore machines' cognitive semantic understanding capabilities through the lens of spatial expressions by a single task such as spatial role labeling (Pustejovsky et al., 2015; Kordjamshidi et al., 2017) or spatial reasoning (Mirzaee et al., 2021). As for evaluation in Chinese, the **Spatial Cognition Evaluation Task (SpaCE)** has been held for 3 years since 2021 with the aim of comprehensively evaluating machine's capabilities with regard to Chinese spatial semantic understanding by multi-task learning (詹卫东et al., 2022; Xiao et al., 2023a; Xiao et al., 2023b). The latest results indicate that machines' spatial semantic understanding capabilities lags significantly behind the average level of humans, especially in tasks requiring high cognitive processing. Spatial semantic understanding remains a challenging task for NLP systems, even for large language models (LLMs).

In this work, we construct a benchmark that comprehensively assesses LLMs' performance on understanding spatial expressions and conducting spatial reasoning in the following five subtasks: (1) Identifying Spatial semantic Roles (ISR); (2) Retrieving Spatial Referents (RSR); (3) Detecting Spatial semantic Anomalies (DSA); (4) Recognizing Synonymous spatial Expression with different forms (RSE); (5) Spatial Position Reasoning (SPR). SpaCE 2024 has released 10,373 multiple-choice questions (MCQs) of the benchmark, which is available at https://github.com/2030NLP/SpaCE2024. Summarized statistics are shown in Table 1. In summary, the characteristics of the benchmark can be outlined as follows:

- The subtasks are interconnected. They not only share a common evaluation vision about spatial semantics but also emphasize the relevance of data to each other, reflected in the shared contextual information across some tasks.
- Structured synthetic data are used in the reasoning task. These rule-based generated data are 100% accurate and do not require quality auditing.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Sub-Task	Answer Type	#Train	#Dev	#Test		#Total
				Non-Disturb.	Disturb.	
ISR	single	1,074	186	746	30	2,083
	multi	19	4	24		
RSR	single	937	226	489	30	1,948
	multi	161	24	81		
DSA	single	1,077	40	500	30	1,647
RSE	single	4	44	517	30	740
	multi	1	11	133		
SPR	single	909	468	1,509	30	3,955
	multi	301	207	531		
#Total	-	4,483	1,210	4,530	150	10,373

Table 1: Statistics of SpaCE 2024 benchmark.

- The consistency of LLMs is concerned. We set some disturbance data in the test set, by repeating, rephrasing questions, and adjusting the order of options. If an LLM cannot perform well under different conditions of disturbance, it's hard to confidently assert that this LLM has some kind of capability.

2 Task Definition

The five subtasks of SpaCE 2024 have different difficulty levels of cognitive processing, which can be roughly divided into three levels: (1) Basic: ISR. This level requires understanding the lexical meaning, grammatical structure, and semantic structure of a sentence, focusing on local information processing. (2) Moderate: RSR and DSA. This level requires understanding discourse semantics with context and commonsense, focusing on global information processing. (3) Advanced: RSE and SPR. This level requires employing advanced cognitive capabilities based on semantic understanding, such as constructing spatial scenes and conducting spatial reasoning. Details about the five subtasks are introduced in the following.

2.1 Identifying Spatial semantic Roles (ISR)

Identifying semantic roles is a fundamental task in semantic understanding. We assume that a machine that can understand underlying semantics should perform well in identifying semantic roles. SpaCE 2022 proposed an annotation scheme called **STEP** to formalize the semantics of spatial expression. S, T, E and P stand for Spatial entity, Time, Event and Place, respectively. The meaning of a spatial expression can be represented by STEP annotation as "an entity is at a certain place at a certain time via an event". Appendix A presents 14 spatial roles of STEP.

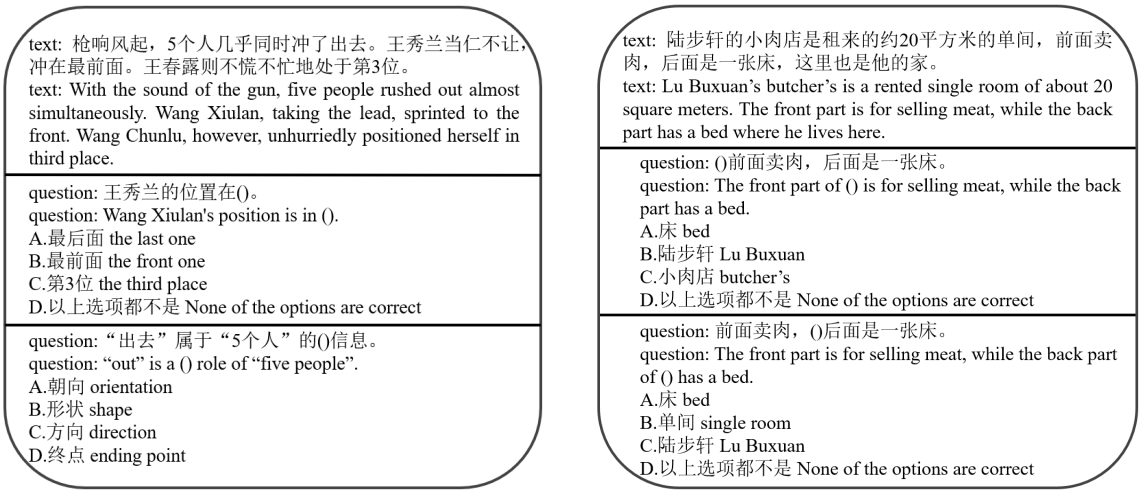
In ISR, machines are required to select the spatial role corresponding to a spatial expression or, conversely, select the spatial expression corresponding to a spatial role, as exemplified in Figure 1(a). The answers of two questions in example are (B) 最前面 and (C) 方向.

2.2 Retrieving Spatial Referents (RSR)

In Modern Chinese, a spatial entity's position is commonly expressed through the combination of a noun as the referent and a following localizer, such as 树下面*(the area under the tree)* and 剧院里面*(the area inside the theatre)*. However, referents can sometimes be omitted and instead be found in a more distant position in the text. In Figure 1(b), there are no referents preceding the localizers 前面*(front)* and 后面*(back)*, yet we can still understand that the meat is sold at the front part of the butcher's and the bed is at the back part of the butcher's.

Since referents play a crucial role in accurately mapping localizers to real-world spatial scenes, the ability to retrieve omitted referents reflects the capability of LLMs in text understanding. In RSR, the

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



(a) ISR example (b) RSR example

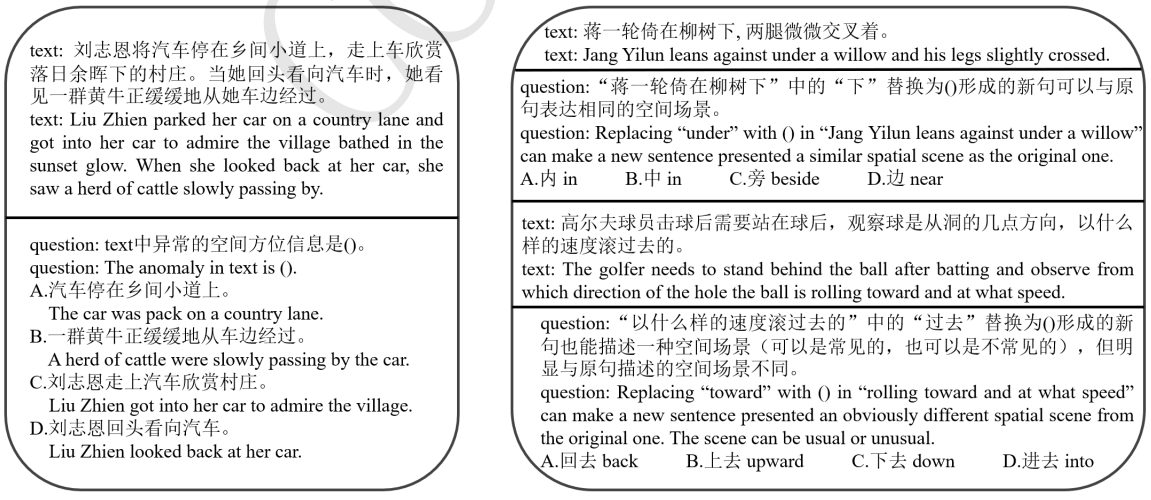
Figure 1: Examples from ISR and RSR training set. English translations are shown for better readability.

machine needs to select one or more entities as the referent(s) of the localizer in question. The answers to two questions in the example are (C) 小肉店 and (B) 单间.

2.3 Detecting Spatial semantic Anomalies (DSA)

Having capability of semantic understanding means being able to detect an expression with semantic anomalies. The text in Figure 2(a) has a spatial semantic anomaly: the protagonist should get out of her car to enjoy the scenery after parking instead of getting in, because she was already inside the car. The following context ”she looked back at her car” also indicates that the protagonist was outside the car.

In DSA, the machine needs to select an expression corresponding to the anomaly in text. The answer of the question in example is (C) 刘志恩走上汽车欣赏村庄. Actually, the expression in option generally don’t have semantic anomalies, but it always conflicts with contextual information in text that makes the anomaly stick out.



(a) DSA example (b) RSE example

Figure 2: Examples from DSA and RSE training set. English translations are shown for better readability.

CCL 2024

2.4 Recognizing Synonymous spatial Expression (RSE)

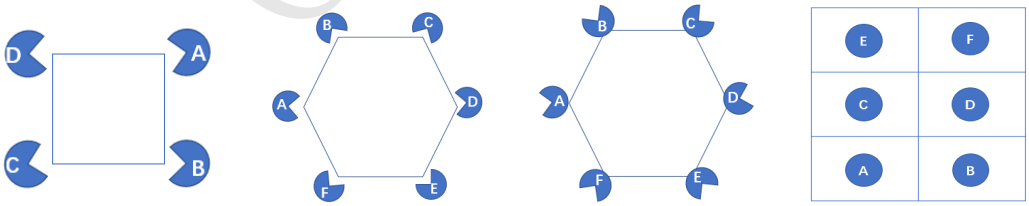
In Modern Chinese, different localizer generally causes to different meaning of the expression. For instance, if we change the localizer 上*(on)* of the sentence 把筷子放在碗上*(Putting the chopsticks on the bowl)* into 边*(beside)*, then the position of the chopsticks will change from the top of the bowl to the area near the bowl. However, there are certain situations where two spatial expressions can have similar meanings even though the localizer is different. For instance, the meaning of both sentence 蒋一轮倚在柳树下 and sentence 蒋一轮倚在柳树上 is *Jiang Yilun leans against a tree*, while the former uses localizer 下*(under)* and the latter uses localizer 上*(on)*. The spatial scenes pictured by these sentences are considered equivalent.

Comprehending such sentence groups needs to utilize commonsense and knowledge to compare the spatial areas of the entity activated by different localizer in the real world. RSE has two kinds of questions. One requires selecting all localizers or directional verbs which ensure that the new sentence after replacement can describe a spatial scene similar to the original one, as exemplified by the top question in Figure 2(b), where the answers are (C) 旁 and (D) 边. The incorrect options are so because the replaced sentence implies "the person leans inside the tree", which is an anomaly. The other type requires to ensure that the two sentences describe different spatial scenes, as seen in the bottom question of Figure 2(b), where the answer is (D) 进去 and the others are incorrect: the scenario described by 滚上去*(rolling upward)* or 滚下去*(rolling down)* is similar to 滚过去*(rolling toward)*, because the trend of the ball's movement are similar and the former is just more detailed in the description of direction; 滚回去*(rolling back)* makes the sentence anomalous because 回去*(back)* implies a return to a previous position but the context suggests batting the ball to a new position.

2.5 Spatial Position Reasoning (SPR)

Spatial Reasoning is a cognitive process based on the construction of mental representations for spatial entities, relations, and transformations, which is necessary for spatial semantic understanding (Clements and Battista, 1992). In SPR, the context is automatically generated based on a preset spatial layout, where entities within the text can form a spatial layout, but their positions must be deduced using the spatial relations provided by the text. Figure 3 shows 4 types of spatial layout schema used in SPR. These layouts contribute to evaluating the comprehension of different spatial relations, as well as the same relation expressed in different ways, in an unified structure. Moreover, a larger number of entities and more advanced layouts indicate a more complex spatial relation, which is conducive to building an evaluation system with varying levels of difficulty.

The example in Figure 4 is an instance of the concentric hexagon layout. First, you need to use the 4 given conditions to infer the position of six statues, as the resolution pictured. Then, find the statues that are not adjacent to Zhang Tianshi: (B) 曹国舅 and (C) 张果老.



(a) square (b) concentric hexagon (c) eccentric hexagon (d) 3×2 matrix

Figure 3: Schema of 4 kinds of spatial layout. Blue circles with letters are spatial entities, with gaps indicating their orientation

3 Dataset

Each data is a four-choice question containing five fields: *qid, text, question, option and answer*. The *qid* field is the unique identity label of the data. For example, *qid* as "3-test-s-1660" points to the 1660^{th}

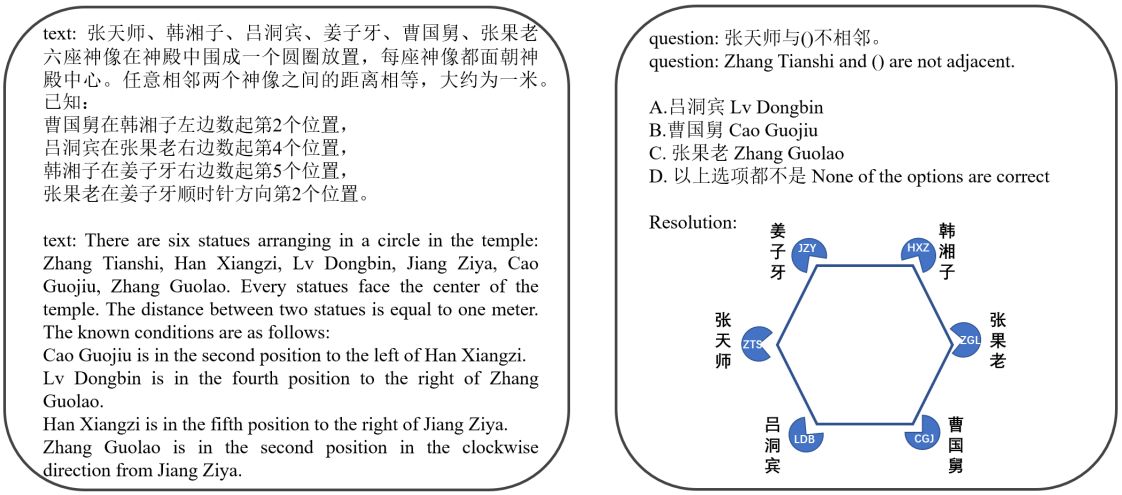


Figure 4: Example from SPR training set. English translations and resolution schema are shown for better readability.

data in the test set, with only one answer, belonging to DSA. We use numbers 1 to 5 to present the five subtasks and use ”s” or ”m” for a single answer or multiple answers, at least 2 up to 4. The *text* and *question* field are string shown above in Figures. Notably, the genre of the text covers general domains such as textbooks and newspapers, as well as specific domains including text of traffic accidents, body movements and geographical encyclopedias. The *option* field is a dictionary with the keys from the letters A to D. The *answer* field is an array containing the letter corresponding to the correct option.

To ensure data quality, data in ISR, RSR, DSA and RSE were generated based on manual annotation and data auditing with the rules that data could only be accepted if two people agreed or one person audited it twice. The SPR data were automatically generated by a spatial knowledge base and an interpretable program, without the need for data auditing.

3.1 Word Replacement for Text

Texts in DSA and RSE were generated by word replacement. We first grouped localizers and directional verbs based on their morphological characteristics and the degree to which they can be grammatically substituted. For example, 上边,下边,前边,后边*(up,down,front,back)* were in one group. The groups also served as options of RSE. Then, using the results of part-of-speech tagging, we replaced the original words in the text with words from the same group. Finally, we manually annotated the replaced sentences to determine if there were any anomalies in the sentences. Data with consistent annotations by two annotators were considered valid. If there were any anomalies, they belonged to DSA, otherwise became a resource of RSE for further annotation about whether the two sentences can picture a similar scene. To generate more sentence pairs for RSE, the members of the SpaCE team needed to construct texts that satisfy the requirement of the given word pair.

3.2 Annotation Extraction for Option

Since 2022, we have recruited nearly 30 students majoring in linguistics each year to annotate STEP information, achieving about 5000 high-quality annotations as a data resource for subtasks. To construct appropriate options for ISR, we extracted annotations that include at least three target roles and then filled in question templates designed for different semantic roles to generate data. RSR took advantage of the discontinuity between the omitted referents and the localizers. We selected discontinuous texts from the annotations and, using the results of part-of-speech tagging, chose nouns as candidate options. For DSA, before word replacement, the text had been annotated with STEP information, making it easy to locate the annotation of the original word and perform the replacement to create a wrong annotation. Together with three other correct annotations of this text, they were sent to GPT-4 to generate natural

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

sentences, thereby forming options related to this text.

3.3 Synthesis Data for SPR

It's too challenging for humans to entirely manually construct a spatial reasoning data, so we try to use synthetic data techniques. Current technology employs LLMs to generate more data, such as data augmentation or to generate complex prompt and narrative texts.(Mukherjee et al., 2023; Josifoski et al., 2023; Eldan and Li, 2023; Dai et al., 2023; Yehudai et al., 2024) For reasoning tasks, LLMs are still unable to produce high-quality data, especially in cases where the reasoning chain is long(Liu et al., 2023). Since the time required for human to audit a reasoning question is much longer than the other tasks, we have decided to use a rule-based method to generate data with transparent intermediate processes. SpartQA is a representative work of the rule-based method, but it only considered 7 types of spatial relations which might no longer be that challenging for LLMs(Mirzaee et al., 2021). To address this, we proposed a program that can automatically generate reasoning questions driven by expert knowledge. Figure 5 illustrates the entire process.

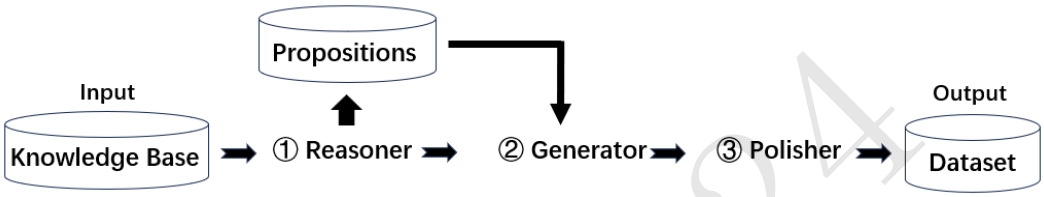


Figure 5: A flowchart for SPR data generation program.

Based on the spatial layout shown in Figure 3, the knowledge base uses templates and reasoning rules to describe the positions of entities within the layout. Each template is a spatial expression that contains slots for entities, like X在Y左边数起第2个位置*(X is in the second position to the left of Y)*, where "X" and "Y" are the slots. Reasoning rules between two templates are logical relations, including equivalence, implication, inclusion, conflict, and reversibility, ensuring that the program has the ability to generate a large number of varying expressions based on a small number of premises. For example, X在Y左边数起第2个位置can deduce X在Y顺时针方向第2个位置*(X is in the second position in the clockwise from Y)* because they are equivalent. We also set up multiple templates as conditions and one template as a conclusion in reasoning rules to expand the program's reasoning ability. For instance, X在Y左边数起第2个位置groups with Z在Y左边数起第1个位置*(Z is in the first position to the left of Y)* can deduce X在Z左边数起第1个位置*(X is in the first position to the left of Z)*. The current knowledge base contains a total of 592 templates written by linguistic experts, expressing more than 30 types of spatial relations.

After loading logical relations and condition-conclusion pairs from knowledge base, the Reasoner needs to input some initial propositions that can form a complete spatial layout and automatically generate more prepositions. Then the Generator will randomly take several prepositions as known conditions and one preposition as a question until these prepositions can form a spatial layout. The entity will be removed from the question as an answer and the other entities will be the options. These extracted propositions are then fed back into the Reasoner to output the intermediate reasoning process. Finally, the Polisher will replace the slots in template like X as a predefined entity like 张天师 *(Zhang Tianshi)* and add a lead in the text to make it conform to natural language.

Given that the knowledge base is constructed correctly, the data generated by this method is guaranteed to be 100 % accurate. It can generate a vast amount of diverse data without limitations if the spatial templates and inference rules in the knowledge base are sufficiently comprehensive. Futhermore, compared to using LLMs for data synthesis, the rule-based method offers strong interpretability, transparency in intermediate processes, and traceable reasoning steps. This is extremely beneficial in assessing the quality and difficulty of spatial reasoning data.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Team	Total	Single	Multi	ISR	RSR	DSA	RSE	SPR
TeleAI	**60.24**	**64.90**	37.45	93.64	**89.47**	**84.80**	56.31	34.71
SHU	59.69	64.34	36.93	91.43	84.91	81.00	54.31	**37.16**
PKU	59.49	63.55	**39.66**	**94.29**	77.19	78.00	**58.77**	37.11
ZUT	56.60	61.71	31.60	91.43	85.26	79.80	49.23	32.11
GDUT	56.47	61.26	33.03	92.86	73.68	76.20	58.62	32.40
BNU	56.20	61.79	28.87	**94.29**	79.65	74.20	56.92	30.64
Knowdee	54.48	59.29	30.95	91.17	75.09	75.40	52.31	30.44
DataGround	53.55	61.69	13.78	90.13	73.33	79.60	51.23	28.58
SU	51.99	56.16	31.60	90.00	80.53	70.20	24.15	34.07
CPIC	48.65	52.91	27.83	86.10	76.67	62.20	40.31	26.03
Insgeek	47.24	51.53	26.27	81.82	67.19	70.00	33.69	27.35
XTU	33.64	36.80	18.21	67.27	49.47	21.60	26.62	21.72
Baseline1 (Fine-tuning)	47.92	54.37	16.38	88.18	75.09	68.60	42.00	21.96
Baseline2 (GPT-4-1106-preview)	46.29	52.94	13.78	85.19	54.39	65.60	45.08	25.00

Table 2: Accuracy of 12 teams (%)

4 Evaluation and Results

We use accuracy as the ranking metric. The higher the accuracy, the better the LLM's capability of comprehensive spatial semantic understanding.

$$Accuracy = \frac{\#Correct}{\#NonDisturbDatas} \tag{1}$$

There are 150 disturbance data in test set that use to evaluate the consistency of LLMs. A group of disturbance data contains an original data, a repeated version, a version with different order of options, and a version with different expression of question. A LLM is considered stable only if all the answers in a group are the same. The higher the consistency score, the less the LLM is affected by the data format, making the evaluation results more reliable.

$$Consistency = \frac{\#StableGroups}{\#DisturbGroups} \tag{2}$$

Totally 30 teams enrolled and 12 teams stick to the end. The top 6 teams won the prize. They all employed LLMs either fine-tuned or utilized prompt engineering. The accuracy results are shown in Table 2. We develop two baselines, one is using prompt engineering in GPT-4-1106-preview[0], the other one is fine-tuning in Qwen1.5-7B-Chat model[1]. Both baselines were tested in zero-shot mode. Most of the teams surpassed our baselines, yet the best total accuracy only slightly exceeded 0.6, suggesting that SpaCE 2024 is quite challenging for current LLMs, especially in RSE and SPR. In contrast, LLMs performed well on tasks from basic level and moderate level, notably the top accuracy was close to 0.9

[0]https://openai.com/index/gpt-4
[1]https://github.com/QwenLM/Qwen1.5

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

in ISR. We also analyzed the performance of questions with single answers versus multiple answers and found that LLMs are much better at handling single-answer questions.

We selected 100 questions from each of the four tasks, excluding SPR, to create a human test set. Each task had 6 human participated. Table 3 shows the performance of the top 6 teams and the average of human accuracy on the human test set. None of a team surpassed human performance. Although accuracy in ISR and RSR is close to human, there is still a gap between LLMs and human on DSA and RSE. For SPR, we believe that spatial reasoning questions tend to reflect individual human differences rather than the overall human level. Additionally, the quality of our synthesis data does not need to be reflected by human scores, as our members have sampled on every type of questions, ensuring no error.

Team	ISR	RSR	DSA	RSE
TeleAI	94.00	92.00	81.00	54.00
SHU	82.00	85.00	72.00	58.00
PKU	91.00	81.00	75.00	60.00
ZUT	86.00	90.00	77.00	49.00
GDUT	86.00	79.00	77.00	54.00
BNU	93.00	83.00	71.00	58.00
Human	**96.00**	**92.33**	**91.50**	**86.33**

Table 3: Accuracy of top 6 teams in Human test set (%)

Table 4 shows the top 6 teams' consistency scores. The results indicate that the performance of LLMs is affected by the question format, which is unrelated to the test content, particularly the order of the options. The 1st was not only the most accurate but also the most stable.

Team	Consistency	Repeated	Reform Question	Switch Option
TeleAI	**84.85**	98.46	85.00	**89.23**
SHU	81.82	95.38	90.00	87.69
PKU	69.70	81.54	85.00	78.46
ZUT	71.21	96.92	**95.00**	72.31
GDUT	77.27	90.77	85.00	84.62
BNU	72.73	98.46	80.00	76.92

Table 4: Consistency scores of top 6 teams (%)

5 Overview of Approaches

Only the top 6 teams submitted their models. Model ensemble with majority voting is the main strategy.

The team from TeleAI company manually selected 5 examples for each subtask to construct in-context learning (ICL) training data, then fine-tuned Qwen1.5-7B-Chat 3 times to do the majority voting. TeleAI set weights on various models and tasks, thereby performing a more nuanced and context-sensitive aggregation to enhance the accuracy and consistency score of the final predictions.

The team from Shanghai University (SHU) recategorized the dataset into four types of questions based on the training size and the difficulty of the tasks. For the questions having large training data with low difficulty, SHU fine-tuned Qwen1.5-7B-Chat, Yi-6B-Chat, and intern2-chat-7B, then voted for the majority as answer. For the structured data in SPR, SHU assumed that it was too difficult and not suitable for training with the easy question, so they fine-tuned the above models again and voted. The remaining questions did not have enough training data to fine-tune well, so SHU ran GPT-4o with Chain of Thought (CoT) and Qwen1.5-110B-Chat with 5-shot to solve them.

The team from Peking University (PKU) ran GPT-4o three times in ISR, RSR, DSA, SPR and conducted a majority vote. For RSE, it ran GPT-4o and Deepseek-chat, meanwhile set another Deepseek

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

as the judge to decide which one was better when two answers were not the same. PKU also set an error correction mechanism to rerun the question in some situations like only outputted one answer in a multi-answer question.

The team from Zhongyuan University of Technology (ZUT) used data augmentation by generating pseudo-labeled data. They used training data to fine-tune Qwen1.5-7B-Chat in different hyperparameter and chose high-score models to predict the unlabeled test data. After majority voting, nearly 3k reliable pseudo-labeled data were added to the training set to fine-tune again until no further improvement in the accuracy of the test set. ZUT also ran GLM-4 in the test set and made model ensemble.

The team from Guangdong University of Technology (GDUT) employed 5-fold cross-validation and majority voting to produce the final prediction by fine-tuning ChatGLM3-6B-base. Another model for ensemble was Ernie-4.0-8k-0329, whose overall score was higher than ChatGLM3 in ISR, DSA and RSE. For ICL, GDUT dynamically selected 5-shot examples by computing the similarity of options between the target test data and the training data in same subtask. The top 5 of similarity would be chosen as examples.

The team from Beijing Normal University (BNU) constructed prompt with one-shot. To select the most representative example for each task, they used Sentence-BERT(Reimers and Gurevych, 2019) to encode all training data into embedding and averaged all embedding to get the centroid of the task. The data closest to the centroid in terms of embedding distance was the most representative.

In general, the models of the top 6 teams were primarily based on ICL and the majority voting strategy. Fine-tuning can bring about a more stable model, and it can improve the performance of a 7B model close to that of GPT-4o, but this improvement is limited to simple tasks. For the two most complex tasks, fine-tuning might not bridge the gap because it's fundamentally related to the reasoning capabilities of the base LLMs.

6 Additional Analysis

Following an analysis based on the results of the top 6 teams, We studied each subtask, the shared text across tasks, and conducted a comparison between SpaCE 2024 and SpaCE 2023.

6.1 Performance of Per-Task

Questions in ISR were labeled as STEP roles. The good performance on most types may be due to the powerful ability of LLMs to capture the correspondence between form and meaning, such as the negation marker of *Factivity* and the preposition marker "到" of *Ending Point*. There were two types of questions where performance needed to be improved: 1) Questions asking for selecting a spatial role. For example, only two teams correctly answered the question "车辆后部"属于"外卖箱"的()信息*("the back of the car" is a () role of "the delivery box")*. Some LLMs may not have a good grasp of STEP, leading to confusion between *Internal Place*, the correct answer, and *Part*. This issue was more serious in the *Direction* type, where LLMs often misidentified directional verbs as other roles. It was possible that the machine's internal knowledge did not classify directional verbs under semantic of *Direction* like STEP did. Therefore, the prompt must include a detailed introduction of STEP to achieve better performance. 2) Questions requiring reasoning ability, including event sequencing questions in *Time*, qualitative distance reasoning questions in type *Distance*, and position reasoning questions similar to SPR in *External Place*.

In RSR, we found that the difficulty of referent retrieval is related to the type of localizers. LLMs performed better on questions where localizers describe up-down, inside-outside, and front-back relations than on those describing left-right and cardinal directions (north, south, east, west). We also observed that LLMs tended to choose entities closer to the target localizer. For example, in the question "()后边架着一架摄像机记录整个授课过程"*("Behind ( ) is a camera recording the entire teaching process"*, only one team correctly chose "classroom" as the answer. The remaining five teams chose the closer option, "blackboard", which is illogical because blackboards in classrooms are commonly mounted on walls, making it impossible to place a camera behind them. Another observation is that LLMs tended to avoid using the subject as the referent. For example, in the question "袁格兵将绳子系在()左边的门柱

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

上，把另一头抛给司机”*”Yuan Gebing tied a rope to the left doorpost of () and threw the other end to the driver”*, the correct answer is the subject ”Yuan Gebing”, but all teams chose ”streetlight” or ”car”, which are inconsistent with the context of the flood rescue situation described in the text.

The DSA's questions were annotated with the amount of textual information needed to infer anomalies. Statistics show that the more information required, the harder the question, and the poorer the LLMs' performance. In texts describing body movements and traffic accidents, embodied experience and traffic knowledge were necessary to infer the anomalies but not included in the text, making it more difficult for LLMs. For example, in texts describing body movements, LLMs performed poorly on questions related to the orientation of the chest in a prone position. Similarly, in texts describing traffic accidents, without common knowledge, such as north-south roads have east-west lanes, it was hard to judge whether two cars will collide. In terms of understanding spatial relations, we observed the same phenomenon as in RSR that LLMs performed worse with left-right and cardinal directions compared to other spatial relations.

RSE had two types of questions. LLMs performed better with synonymous spatial expressions than with those have different meanings, which was consistent with human performance. However, unlike humans, the pairs of words with which LLMs performed better in synonymous spatial expressions were quite different from another type. The former mainly included synonyms, such as ”旁-边” *(beside-next to)* and ”中-内” *(within-inside)*, while the latter mainly included antonyms, such as ”内-外” *(inside-outside)* and ”上面-下面” *(above-below)*. Humans did not show such a significant difference. For example, with common antonyms like ”上-下” (up-down), humans performed well on both types of questions, although antonyms were rarely used in synonymous spatial expressions. It indicates that LLMs' understanding of spatial semantics remains at surface semantics, lacking the capability of modeling spatial scenes case by case.

The difficulty of SPR questions was measured based on the number of actual reasoning conditions used, which were automatically outputted by the reasoning program. The more reasoning conditions required, the longer the reasoning chain, the harder the question, and the worse the performance. For instance, the score of the questions with using 2 conditions was 40 % less than those with only 1 condition. The complexity of spatial layouts and the number of entities also affected LLMs' performance: 1) The more complex the layout, the worse the LLM performs, shown as Square layout >hexagonal layout >matrix layout. 2) LLMs performed better on texts involving 4 entities compared to those with 5 or 6 entities. 3) If a cardinal direction was added to a hexagonal layout, LLMs' performance decreased by approximately 8 %. In terms of spatial relation, we labeled each question with spatial relation based on the localizer used. Statistics show that LLMs perform better in reasoning about opposite and adjacent relation compared to left-right and cardinal directions. Reasoning about up-down relation is even worse.

6.2 Discussion on the Shared Text

20% of the test set questions shared textual information between subtasks. Taking the most basic task, ISR, as a clue, we constructed 801 question pairs in the form of (ISR, RSR/DSA/RSE/SPR). The conditional probability that ISR is correct when the other subtask is correct reached 92.26%, suggesting that the LLMs' ability to complete advanced tasks is based on a fundamental understanding of semantics. On the other hand, the number of pairs where both tasks are incorrect is much fewer than the number of pairs where ISR is incorrect and another one is correct. Other fundamental capabilities of LLMs like reasoning may compensate for this deficiency when the LLMs' semantic understanding is flawed.

6.3 Comparison between 2023 and 2024

SpaCE 2024 inherited ISR, DSA and RSE tasks from SpaCE 2023, providing a basis for comparison. However, the data formats differ between the two years. In 2023, ISR and DSA were sequence labeling tasks, and RSE was a true-false task, while in 2024, they were all multiple-choice questions. To allow a comparison, we selected questions with the same text from both years and used F1 score as a metric for ISR and DSA to be consistent with 2013. For ISR, we only considered the single labeling corresponding to the MCQs as the result of SpaCE 2023.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Year	ISR		DSA		RSE	
	#Num	F1(%)	#Num	F1(%)	#Num	Acc.(%)
2023	92	45.40	40	62.54	46	68.48
2024		95.47		86.25		52.54

Table 5: Per-task Comparison between 2023 and 2024 of SpaCE

As Table 5 shows, the F1 scores in 2023 are the average of two BERT-based models and the accuracy score is the average of two LLM systems. All scores in 2024 are the averages of six LLM systems. The state-of-the-art improved markedly for ISR and DSA, but went down for RSE. This displays the strengths of LLMs compared to BERT-based models, but also indicates the effect of data format. For ISR and DSA, sequence labeling tasks are overly complex in format, preventing the machine's capabilities from being fully demonstrated. For RSE, MCQs are more challenging than true-false questions because true-false questions only require comparing one pair of words, while MCQs provides four pairs, resulting in a lower score. In general, it was the first time that the SpaCE series of tasks employed multiple-choice questions, helping to improve the validity of the evaluation.

7 Conclusion

This paper introduces the fourth evaluation of the Chinese Spatial Cognition Task. The benchmark consists of five subtasks covering 5 dimensions of spatial semantic understanding to comprehensively assess the machine's language comprehension and reasoning capabilities on spatial semantics. In addition to capability assessment, SpaCE 2024 concerns the consistency of LLMs' performance and sets a disturbance set. As for quality control, a novel rule-based method was used in the spatial reasoning task, successfully generating high-quality synthesis data with transparent intermediate processes and difficulty levels.

The accuracy scores across 5 subtasks reveal the following trend: ISR >RSR >DSA >RSE >SPR. Based on a fine-grained analysis, we have gained the following insights into the spatial semantic understanding capabilities of LLMs:

- Regarding the spatial relation, understanding "left and right" and "cardinal directions" are weaknesses of LLMs, which have been observed among tasks.
- LLMs' ability of semantic analysis is comparable to those of humans. However, the semantics of spatial expressions also encode the interactive experiences between humans and other entities, which currently cannot be simulated by LLMs. This limits LLMs' performance in some advanced tasks.
- Spatial reasoning tasks pose a significant challenge for LLMs. Current LLMs do not possess the ability to solve complex reasoning problems. Data fine-tuning is unlikely to improve their performance in spatial reasoning tasks.

In future work, our aim is to generate a large amount of structured, high-quality, and knowledge-controllable data to train LLMs to enhance their semantic understanding and reasoning capabilities. To this end, we will further investigate data synthetic methods to incorporate more spatial relations such as 里*(inside)* and 外*(outside)* into the knowledge base and to extend this method to four other subtasks, exploring a technical path for data synthesis driven by knowledge.

Acknowledgements

We would like to acknowledge the contributions of the members of the SpaCE 2024 team, Jiajun Wang, Dan Xing, Xihao Wang, Zihan Zhang, Xiang Cui, and 25 annotators from Peking University, Tsinghua University, etc. We thank Professor Zhifang Sui and her team members, Yixin Yang, and Jingyuan Ma, for helpful discussions on topics related to this work. This work was supported by the Major

Program of the Key Research Center of the Ministry of Education of Humanities and Social (Grant No.22JJD740004).

References

Douglas H Clements and Michael T Battista. 1992. Geometry and spatial reasoning. *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*, pages 420–464.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132*.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 367–376. Springer.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Leonard Talmy. 1983. How language structures space. In *Spatial orientation: Theory, research, and application*, pages 225–282. Springer.

Liming Xiao, Chunhui Sun, Weidong Zhan, Dan Xing, Nan Li, Chengwen Wang, and Fangwei Zhu. 2023a. Space2022 中文空间语义理解评测任务数据集分析报告(a quality assessment report of the chinese spatial cognition evaluation benchmark). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 547–558.

Liming Xiao, Weidong Zhan, Zhifang Sui, Yuhang Qin, Chunhui Sun, Dan Xing, Nan Li, Fangwei Zhu, and Peiyi Wang. 2023b. Ccl23-eval 任务4 总结报告: 第三届中文空间语义理解评测(overview of ccl23-eval task 4: The 3rd chinese spatial cognition evaluation). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 150–158.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. *arXiv preprint arXiv:2401.14367*.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 2022. 空间语义理解能力评测任务设计的新思路—space2021数据集的研制. 语言文字应用, (02):99–110.

CCL 2024

A STEP scheme

Role	Description	Example
Spatial Entity	the entity in a spatial relation.	桌子上有一杯牛奶。 There is a glass of milk on the table.
External Place	the position of an entity relative to an external referent.	桌子上有一杯牛奶。 There is a glass of milk on the table.
Internal Place	the position of an entity within the referent.	这杯鸡尾酒的上面是红色的。 The top of this cocktail is red.
Starting Point	the initial position of an entity when it moves.	汤姆从宿舍走到了图书馆。 Tome walked from the dormitory to the library.
Ending Point	the final position of an entity when it moves.	汤姆从宿舍走到了图书馆。 Tome walked from the dormitory to the library.
Path	the path taken by an entity when it moves.	汤姆沿着这条街道往南走去。 Tome walked south along the street.
Direction	the position that an entity towards when it moves.	汤姆沿着这条街道往南走去。 Tome walked south along the street.
Orientation	the position that an entity's surface is facing.	汤姆面朝北往南倒着走。 Tome walked backward to the south while facing north.
Part	the piece of a spatial entity.	汤姆把手放在了桌子上。 Tom placed his hand on the table.
Shape	the physical form of a spatial entity.	请大家排成一条直线。 Please line up in a straight line.
Distance	the amount of space between two entities.	两地相隔3公里。 The two places are 3 kilometers apart.
Spatial Event	the event related to a spatial relation.	汤姆把手放在了桌子上。 Tom placed his hand on the table.
Spatial Time	the time that a spatial expression happens.	汤姆在警察到达之前离开了现场。 Tom left the scene before the police arrived.
Spatial Factivity	the state that a spatial expression is a fact.	汤姆不在家。 Tom is not at home.

Table 6: Introduction of STEP scheme. The underline in the example indicates the words corresponding to the roles. English translations are shown for better readability.

CCL 2024

CCL24-Eval任务4系统报告：面向中文抽象语义表示解析的大模型评估与增强

陈荣波，裴振武，白雪峰*，陈科海，张民
哈尔滨工业大学（深圳），计算机科学与技术学院
baixuefeng@hit.edu.cn

摘要

本文介绍了我们在第二十三届中文计算语言学大会中文抽象语义表示解析评测任务中提交的参赛系统。中文抽象语义表示（Chinese Abstract Meaning Representation，CAMR）以一个单根可遍历的有向无环图表示中文句子的语义。本系统选择大语言模型作为解决方案。我们首先系统地评估了当下中文大语言模型在AMR解析任务上的性能，在此基础上基于图融合算法整合性能较高的大模型预测结果，最终得到预测的CAMR图。实验结果表明，1）现有大模型已经具备一定的少样本中文AMR解析能力；2）基于微调中文大模型的AMR解析系统能够取得相较以往最优系统更强的性能；3）图融合算法能够进一步增强基于大模型的CAMR解析系统的性能。

关键词： 中文抽象语义表示；大语言模型；图融合

System Report for CCL24-Eval Task 4: Benchmarking and Improving LLMs on Chinese AMR Parsing

Rongbo Chen, Zhenwu Pei, Xuefeng Bai, Kehai Chen, Min Zhang
School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
baixuefeng@hit.edu.cn

Abstract

This paper introduces our submission system for the Chinese Abstract Meaning Representation (CAMR) parsing evaluation task at the 23rd China National Conference on Computational Linguistics. CAMR represents the semantics of Chinese sentences using a single-rooted, traversable, directed acyclic graph. We choose to build CAMR parsing system based on LLMs. We first systematically benchmark the performance of current Chinese large models on the CAMR parsing task. Based on this, we integrate the prediction results of high-performing large models using a graph ensemble algorithm to obtain the final predicted CAMR graph. The experimental results show that: 1) current large models already possess a certain capability in few-shot CAMR parsing; 2) an AMR parsing system based on fine-tuned Chinese large models can achieve superior performance compared to previous systems; 3) the graph ensemble algorithm can further enhance the performance of a large model-based CAMR parsing system.

Keywords: CAMR , Large language model , Graph ensemble

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

中文抽象语义表示（Chinese Abstract Meaning Representation，CAMR）将词抽象为概念节点，句子中词与词之间的语义关系抽象为有向弧，从而将整个中文句子的语义结构描述为一个单根有向无环图 (Banarescu et al., 2013)。不同于标准的AMR表示范式，CAMR在保留AMR的语义表示能力的同时，还增加了概念对齐以及关系对齐等信息，可以更好地表达中文句子语义。AMR与CAMR解析旨在预测输入文本中对应的语义图 (Bai et al., 2022a; Bai et al., 2022b; Bevilacqua et al., 2021; Cai and Lam, 2020; Flanigan et al., 2014; Konstas et al., 2017; Lyu and Titov, 2018)。受益于自动解析技术的发展，AMR与CAMR已经被广泛应用于机器翻译 (Nguyen et al., 2021)、对话系统 (Bai et al., 2021)、文本摘要 (Liao et al., 2018) 等自然语言处理下游任务领域中。

现有AMR/CAMR解析的方法可以分为以下三类：基于转移的方法（Transition-based Parsing）、基于图的方法（Graph-based Parsing）以及基于序列到序列的方法（Seq2Seq-based Parsing）。其中，基于转移的方法首先将CAMR图转换成一系列“转移动作”，然后通过预测该转移动作序列来一步步地构造AMR图 (Wang et al., 2018)。基于图的方法一般通过“节点预测-关系生成”两阶段的方式进行AMR图预测，包括自回归方式以及非自回归方式 (Zhou et al., 2022; Chen et al., 2022)。基于序列到序列的方法的核心思想是将CAMR图进行序列化，从而以序列生成的形式进行自回归生成 (Huang et al., 2021)。以上三种方法中，基于序列到序列的方法无需人工设计复杂的中间特征，极大降低了任务的复杂度。

近年来，大规模语言模型（Large Language Models，LLMs）在诸多自然语言处理任务上取得了巨大的成功 (Touvron et al., 2023; Zhao et al., 2023; Wang et al., 2023; Ren et al., 2023; Zeng et al., 2022)。得益于巨大的参数规模和训练数据规模，大模型产生了在小型模型中不存在的涌现能力，例如上下文学习、指令遵循和逐步推理等能力 (Zhao et al., 2023)。目前，将大模型应用于下游任务的通用范式是将任务输入和输出拼接并转换为一个文本序列，从而使得大模型能够以与预训练阶段相同（即序列生成）的形式处理下游任务。对于CAMR解析任务，Gao et al. (2023) 首个将中文大语言模型用于CAMR解析，初步揭示了大语言模型在CAMR解析任务上的潜力。尽管如此，Gao et al. (2023) 的性能相对受限，并且目前尚缺乏大模型在CAMR解析任务上的系统研究。

为了充分挖掘中文大模型在CAMR解析任务上的潜力，本文首先在CAMR解析任务上的两种设置下对现有多个中文大模型进行了系统评估，实验设置包括少样本学习和监督学习，评估对象包括ChatGPT、GPT4两个商用模型以及Baichuan-2、LLaMA-3、LLaMA-3-Chinese三个开源中文大模型。在此基础上，本文选取最优的几个解析系统并使用图融合算法 (Hoang et al., 2021) 整合多个输出得到最终结果。实验结果表明，1）现有的商用模型已经具有一定的少样本CAMR解析能力；2）微调后的开源中文大模型能够取得与以往CAMR解析系统可比的性能，同时具有更好的泛化性；3）图融合算法能够进一步提升系统的性能。最终，本文所提出的系统在三个测试集上分别取得了80.96、74.85和66.91的分数。

2 方法

本文使用基于序列到序列的方法进行CAMR解析。首先对CAMR图进行序列化预处理，以训练集中分词句子和序列化CAMR图作为组合，对中文大模型进行训练，在此基础上实现将文本转换成CAMR序列，然后将从多个模型得到的CAMR序列进行后处理，还原成形式合法的CAMR图，最后将多个模型生成的CAMR图进行融合和二次后处理后，得到目标CAMR图。

2.1 CAMR线性化

本文参照CAMRP2023[1]中SUDA[2]的策略，对虚词关系对齐和概念同指进行简化特殊处理。如图 1 所示，我们将虚词关系对齐的表示“:arg0(x8/的) (x7 / 政党)”处理成“:arg0 (x7 / 政党 :ralign (x8 / 的))”，将概念同指的表示“:quant() (x5 / x1)”处理成“:quant (x5 / 一 :coref x1)”。另外，我们还对CAMR图中不具有重要意义的括号、空格和换行等符号进行去除，从而

[1]https://github.com/GoThereGit/Chinese-AMR/tree/main/CAMRP%202023
[2]https://github.com/EganGu/camr-seq2seq

(c) 2024 中国中文信息学会计算语言学专业委员会

简化所得到的CAMR图的线性序列。通过以上方法，我们由原始CAMR图得到可用于大模型处理的线性化CAMR序列。

分词文本：这/种/理念/是/一/个/政党/的/执政/基础/。
分词下标文本：x1_这 x2_种 x3_理念 x4_是 x5_一 x6_个 x7_政党 x8_的 x9_执政 x10_基础 x11_。

CAMR图：(x10 / 基础
:mod() (x9 / 执政-01
:arg0(x8/的) (x7 / 政党
:quant() (x5 / x1)
:cunit() (x6 / 个)))
:domain(x4/是) (x3 / 理念
:mod() (x1 / 这)
:unit() (x2 / 种)))

序列化结果：(x10 / 基础 :mod (x9 / 执政-01 :arg0 (x7 / 政党 :ralign (x8 / 的)
:quant (x5 / 一 :coref x1) :cunit (x6 / 个))) :domain (x3 / 理念
:ralign (x4 / 是) :mod (x1 / 这) :unit (x2 / 种)))

Figure 1: CAMR序列化示例（序列化结果为单行）

2.2 基于中文大模型的CAMR解析

为了系统地激发与评估现有中文大模型在CAMR解析的能力，本文选取了现有5种具有代表性的大模型进行实验，并探索了少样本学习和有监督微调两种设置：对于商用闭源模型，本文使用少样本提示学习来激发模型的CAMR解析能力[3]；对于开源模型，本文使用有监督微调来充分发挥模型的性能。

2.3 基于少样本提示学习的CAMR解析

本文选取ChatGPT和GPT-4两种商用模型进行基于少样本提示学习的CAMR解析。首先构造包含任务指令、示例样本、用户输入的提示，随后将提示作为历史上下文信息输入中文大模型中，最终将大模型的输出作为解析后的CAMR图。本文使用的提示模板如下：

任务指令：你是一个中文抽象语义表示解析系统，根据接下来的几个示例进行学习，随后将用户输入的文本转换为中文抽象语义表示图。
示例样本：示例$_1$；示例$_2$；示例$_3$；示例$_4$；示例$_5$。
用户输入：x1_这 x2_种 x3_理念 x4_是 x5_一 x6_个 x7_政党 x8_的 x9_执政 x10_基础 x11_。

Figure 2: 用于CAMR解析的少样本提示学习模板示例

2.4 基于有监督微调的CAMR解析

为了充分激发开源中文大模型的CAMR解析性能，本文采用监督学习的形式对开源大模型进行全量微调，从而获得用于CAMR解析的专用模型。在大模型的微调过程中，本文将任务指令、分词后的源端文本以及线性化的CAMR序列进行拼接，随后以自回归的方式进行预测。以图 1 中“这种理念是一个政党的执政基础。”为例，构造输入数据：

给定如下分词后的中文文本，输出其所对应的中文抽象语义表示图：“x1_这 x2_种 x3_理念 x4_是 x5_一 x6_个 x7_政党 x8_的 x9_执政 x10_基础 x11_。”

为了引导模型专注于生成CAMR结构，本文取消了对于任务指令、原句内容的损失计算。形式化而言，假设以$\mathcal{D}$代表整个训练数据集，I表示任务指令，x表示原句，y表示线性化的CAMR序列，本方法通过优化以下损失函数进行训练：

[3]由于在预实验中观测到零样本性能较差，本文只对少样本学习结果进行评估。

(c) 2024 中国中文信息学会计算语言学专业委员会

$$\mathcal{L} = -\sum_{\{I,x,y\}\in\mathcal{D}} log\, P(\ y \mid I \oplus x), \tag{1}$$

其中 $\mathcal{L}$ 代表损失函数，$\oplus$ 代表序列拼接操作。本文使用基于梯度下降的方法来进行参数更新。

2.5 后处理

由于模型生成的线性序列并不总是符合CAMR的规范，因此需要对其进行后处理，主要有三个方面，括号补全，节点修正以及特殊关系处理。

- 括号补全：由于模型偶尔会产生括号错误匹配的问题，导致CAMR图中节点和边的关系无法正确表示。如果右括号提前与左括号闭合，则将此右括号置于序列末尾；如果右括号数量不足以匹配左括号，则在序列末尾添加一定数量的右括号使其匹配；如果节点缺少括号，则在最小范围内添加一对括号，使其格式满足规范，例如":op1 x12 / 唱罢"，补全括号后修改为":op1 (x12 / 唱罢)"。
- 节点修正：模型解析预测过程中，可能会产生节点信息错误或者缺失的问题。例如节点的编号与节点的内容不匹配或者缺失，或者节点内容不连续，例如对于连续编号"x1_x2_x3"对应的节点内容中词之间出现多余空格。我们通过规则匹配，补全或重新匹配不正确的节点，使其满足规范。
- 特殊关系处理：包含虚词关系对齐以及概念同指，即":coref"与":ralign"。按照预处理中的对应规则，逆向进行复原。对于虚词关系对齐，搜索到它匹配的父亲节点进行复原；对于概念同指，以同指节点中的一个为核心节点，将标签为编号的节点的标签替换为对应核心节点的标签。

2.6 图融合

得到多个模型的后处理CAMR图后，本文采用图融合算法 (Hoang et al., 2021) 来融合多个预测结果，从而得到更加精确、全面的CAMR图。考虑到普通的图和AMR图的不同点在于AMR的结构中存在节点和边的标签， Hoang et al. (2021) 的工作提出一种有效的启发式算法来计算图融合问题的近似最优解，将图融合的预测问题转化成了图挖掘问题，即在图与图之间寻找最大公共子图。具体而言，图融合算法输入m个AMR图，依次选取每个AMR图作为支点图，对于每个支点图来说，以剩下m-1个图中所包含的节点和边进行投票，对m个支点图选出m个聚合图，最终从这m个聚合图中选出一个最佳的聚合图作为输出。

3 实验

3.1 实验设置

数据：本次评测分为封闭测试和开放测试两个赛道，我们选择参加了开放赛道。本次测评训练集包含16576句数据，开发集A、B分别包含1789和500条数据，测试集A、B、C分别包含1713、1999和2000条数据。其中测试集C旨在考察解析系统在古汉语上的自动解析能力。

基线系统：对于闭源商用模型，本文选取了ChatGPT和GPT-4两个模型进行评估。根据预实验的测试结果，本文选用5-ICL作为少样本测试设置，即示例样本由训练集中随机抽选的5个样本组合而来[4]。对于开源中文大模型，基于其在中文评测数据集（C-EVAL）上的性能，本文选取了Baichuan-2[5]，LLaMA-3[6]，LLaMA-3-Chinese[7]三个开源模型进行有监督微调（S.F.T）。此外，为了进行实验对比，本文选取了基于中文BART模型的CAMR解析系统（BARTseq2seq）作为基线。

参数设置：本文所使用的模型在A800 GPU上进行训练，对于本文所提到的开源模型，训练总批次大小为128，训练轮次为5，最大序列长度为1024，使用AdamW (Loshchilov and Hutter, 2019) 优化器进行优化，学习率搜索空间为{1e-5, 3e-5, 5e-5, 8e-5}。

[4]受资源限制，本文只采用了一组示例样本，没有进行多次随机并评估。
[5]https://github.com/baichuan-inc/Baichuan2
[6]https://github.com/meta-llama/llama3
[7]https://github.com/CrazyBoyM/llama3-Chinese-chat

(c) 2024 中国中文信息学会计算语言学专业委员会

评估指标：本文采用Align_Smatch作为评测指标。曾经作为评测指标的Smatch将每个AMR图转化为三元组的集合，每个集合包含三种数据类型的三元组：表示节点（Instance）的三元组、表示有向弧（Relation）的三元组和表示节点属性（Attribute）的三元组。Align-smatch在Smatch的基础上增添了两种汉语特有的新的数据，概念对齐信息和关系对齐信息 (Xiao et al., 2022)，在评测CAMR图时可以更加客观准确。

3.2 实验结果

不同大模型的CAMR解析性能：表 1 对比了不同模型在测试集TestA上的表现。在少样本测试场景中，ChatGPT模型取得了43.07的F_1分数。而相比之下，GPT-4取得了更高的性能，获得了51.95的F_1分数。以上结果表明现有的商用大模型已经具备一定的CAMR解析能力，为接下来的研究提供了新的思路。此外，通过细粒度分析，我们发现现有的商用大模型在CAMR关系预测上准确率较低（分别是8.62和19.55），原因可能是CAMR的关系标签在大模型的预训练语料中几乎不存在，因此难以预测。在全量微调设置下，现有最强的系统为基于编解码器架构的BARTseq2seq模型。与之相比，基于中文大模型的模型取得了可比或更强的性能。特别地，LLaMA-3-Chinese取得了79.69的F_1分数，这表明基于解码器的大模型仍然具有一定的潜力。另外，通过对比三个基于大模型的系统，可以发现LLaMA-3模型已经具备较强的中文处理能力，继续在中文数据上进行微调未能在CAMR解析任务上带来较大提升。

Setting	Model	Total			Instance	Attribute	Relation
		P	R	F_1	F_1	F_1	F_1
5-ICL	ChatGPT	48.03	39.04	43.07	52.06	77.57	8.62
	GPT-4	55.26	49.01	51.95	61.32	84.41	19.55
S.F.T.	BARTseq2seq	78.63	79.46	79.04	84.94	93.86	64.69
	Baichuan-2	79.64	78.30	78.97	84.77	93.33	64.28
	LLaMA-3	79.79	79.43	79.61	**85.36**	**94.20**	64.72
	LLaMA-3-Chinese	79.69	79.68	**79.69**	**85.36**	94.19	**64.95**

Table 1: 不同模型在测试集TestA上的Align-Smatch得分

图融合性能：我们选取了评分较高的几个模型进行下一步的图融合，以确保最终图融合后的结果效果最好。具体而言，我们选取了表 1 中所有微调后的系统作为图融合候选，并利用不同的学习率构造出另外两种候选模型，总计6个候选模型。表 2 列出了在测试集TestA上我们以模型的不同组合进行图融合后的结果表现。总的来说，在进行图融合过程中，采用的模型数量越多，最终由这些模型预测的结果融合得到的CAMR图的分数也就越高，但存在边际效益递减的情况，随着模型数量增多，平均每个模型带来的提升逐渐降低。另外我们发现，在通过图融合得到输出CAMR图后，进行二次后处理可以使效果小幅度提升。

Setting	#Models	P	R	F_1
Ensemble_1	4	80.43	80.69	80.56
Ensemble_2	5	80.75	80.94	80.84
Ensemble_3	6	80.80	81.11	**80.96**

Table 2: 测试集TestA不同组合图融合后的CAMR图的Align-Smatch得分

最终系统性能：表 3 列出了我们方法的得分以及与往年系统的对比结果。我们的系统在三个测试集上分别获得了80.96、74.85以及66.92的分数。其中，我们的方法在TestA和TestB两组测试集上，达到了和去年测评任务中SUDA的方法（该方法在开放赛道的TestA 和TestB上分别获得了81.30和74.71的分数）相当的水平，表明大语言模型在复杂句子语义图的解析上也可达到良好效果，但依然存在进步空间。另外，相较于去年评测任务中同样采用微调大模型的WestlakeNLP，我们的方法在测试集TestA和TestB上表现较优。此外，表 4 列出了三个测试集的细粒度评测分数，我们注意到TestC的结果相较于测试集TestB和TestC分数偏低，原因可能是TestC由古汉语句子组成，古汉语相较于现代汉语可能在语法结构和语言形式上存在一定的差异，使得模型从现代汉语到古汉语的迁移学习存在较大误差。

(c) 2024 中国中文信息学会计算语言学专业委员会

Model	TestA			TestB			TestC		
	P	R	F_1	P	R	F_1	P	R	F_1
SUDA-HUAWEI$_{22}$	82.16	78.20	80.13	75.52	71.79	73.61	-	-	-
ECNU$_{22}$	73.83	66.05	69.72	66.01	57.71	61.58	-	-	-
BUPT$_{22}$	50.41	42.55	46.15	49.95	42.72	46.05	-	-	-
GDUFE$_{23}$	75.53	75.60	75.56	69.71	67.33	68.50	-	-	-
SJTU$_{23}$	47.41	46.45	46.92	46.44	45.68	46.06	-	-	-
SUDA$_{23}$	80.82	81.79	**81.30**	74.39	75.03	74.71	-	-	-
WestlakeNLP$_{23}$	74.40	70.24	72.26	70.42	68.63	69.52	-	-	-
Ours	80.80	81.11	80.96	75.13	74.57	**74.85**	67.05	66.77	**66.92**

Table 3: 评测结果对比

	Total			Instance	Attribute	Relation	Coref	Ralign
	P	R	F_1	F_1	F_1	F_1		
TestA	80.80	81.11	**80.96**	**86.34**	94.71	**66.96**	1.60%	11.63%
TestB	75.13	74.57	74.85	81.09	89.02	59.94	2.48%	15.42%
TestC	67.05	66.77	66.92	68.52	**95.97**	46.79	3.18%	11.55%

Table 4: 三个测试集的Align-Smatch得分和特殊关系概率

3.3 分析

对于现代文测试集的语义解析，大语言模型依然存在一定的进步空间。在长难句的语义解析上，微调后模型的表现仍然不尽如人意，容易出现解析不完整或者解析偏离的情况，可能是由于幻觉现象以及窗口限制的原因。另外，对于关系的预测这一瓶颈，大模型在发现句子中词与词之间逻辑关系的能力上仍需改进。

如 3.2 所述，在表 4 中可以看出古汉语测试集与现代汉语测试集分数差异较大，差异主要体现在节点（Instance）和关系（Relation）的预测上。在节点属性（Attribute）的预测上，三个测试集均效果不错；在节点预测上相较于现代汉语测试集 A 与 B 上 86.34 和 81.09 的分数，古汉语测试集 C 的分数仅有 68.52；另外在关系预测上，古汉语测试集的分数仅有 46.79，明显低于两个现代汉语测试集。

综合以上，我们推测 TestC 预测效果较差的原因是古汉语和现代汉语句法结构上的差异。例如，古汉语中存在大量虚词，而现代汉语中虚词被简化甚至省略；古汉语的语序存在较多倒装，句法结构较为复杂，而现代汉语结构简洁明了；古汉语中存在一些文字在现代汉语中出现频率较低，仅依靠分词结果很难直接解析其真实词性。以上分析表明，在不充分掌握古汉语语义表示规律的前提下，模型仅依靠分词句子进行的推测较难达到期望结果。

4 结论

在本次 CAMR 解析评测任务中，我们系统地评估了现有中文大模型在CAMR解析任务上的能力，在此基础上使用图融合策略进一步增强基于大模型的CAMR解析系统的性能。实验结果表明，现有的中文大模型已经具备较强的少样本和全样本训练后CAMR解析能力，并且结合图融合技术能够进一步提升系统的性能。此外，相较于现代汉语，古汉语解析仍然是一个比较难的挑战。今后，对于古汉语的 CAMR 解析问题，我们认为一种潜在的解决方案是尝试将古汉语和现代汉语的对齐信息加入训练过程，另外一种是通过自动构造大量古汉语和对应 CAMR 图进行训练来提升解析性能。

致谢

感谢所有审稿人对本文提出的宝贵建议，使本文的内容更加完善和系统。本工作由深圳市高等院校稳定支持计划项目（GXWD20231130104007001）资助。

(c) 2024 中国中文信息学会计算语言学专业委员会

参考文献

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022a. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Xuefeng Bai, Sen Yang, Leyang Cui, Linfeng Song, and Yue Zhang. 2022b. Cross-domain generalization for AMR parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Liang Chen, Bofei Gao, and Baobao Chang. 2022. A two-stage method for Chinese amr parsing. *ArXiv*, abs/2209.14512.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Wenyang Gao, Xuefeng Bai, and Yue Zhang. 2023. System report for CCL23-eval task 2: WestlakeNLP, investigating generative large language models for Chinese AMR parsing. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*.

Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam M. Nguyen, Dzung T. Phan, Vanessa López, and Ramón Fernandez Astudillo. 2021. Ensembling graph predictions for AMR parsing. In *Annual Conferenceon Neural Information Processing Systems*.

Ziyi Huang, Junhui Li, and Zhengxian Gong. 2021. Chinese AMR parsing based on sequence-to-sequence modeling. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chunchuan Lyu and Ivan Titov. 2018. AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Long H. B. Nguyen, Viet H. Pham, and Dien Dinh. 2021. Improving neural machine translation with AMR semantic graphs. *Mathematical Problems in Engineering*.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, A. V. Podolskiy, Grigory Arshinov, A. Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. 2023. Pangu-σ: Towards trillion parameter language model with sparse heterogeneous computing. *ArXiv preprint*, abs/2303.10845.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

CCL 2024

Chuan Wang, Bin Li, and Nianwen Xue. 2018. Transition-based Chinese AMR parsing. In *North American Chapter of the Association for Computational Linguistics*.

Xiao Wang, Wei Zhou, Can Zu, Han Xia, Tianze Chen, Yuan Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, J. Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *ArXiv preprint*, abs/2304.08085.

Liming Xiao, Bin Li, Zhixing Xu, Kairui Huo, Minxuan Feng, Junsheng Zhou, and Weiguang Qu. 2022. Align-smatch: A novel evaluation method for chinese abstract meaning representation parsing based on alignment of concept and relation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *ArXiv preprint*, abs/2210.02414.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

Shilin Zhou, Qingrong Xia, Yang Li, Zhefeng Wang, and Zhenghua Li. 2022. Suda-huawei camr2022 比赛技术评测报告. In *Proceedings of the 21nd Chinese National Conference on Computational Linguistics*.

CCL 2024

CCL24-Eval 任务 4 系统报告：混合 LoRA 专家的中文抽象语义表示解析框架

吴梓浩，尹华*，高子千，张佳佳，季跃蕾，唐堃添
广东财经大学，信息学院，广州
{zehro,GaoZiqian,zhangjiajia,132382851009jyl,TangKuntian}
@student.gdufe.edu.cn
yinhua@gdufe.edu.cn

摘要

本文介绍了我们在第二十三届中国计算语言学大会中文抽象语义表示解析评测任务中提交的参赛系统。抽象语义表示 (Abstract Meaning Representation，AMR) 使用有向无环图对句子进行建模，以语义概念作为节点，关系标签作为边，表示一个句子的语义。我们受到结合语法信息的 AMR 解析研究的启发，提出混合 LoRA(Low-Rank Adaption) 专家的 CAMR 解析框架，该框架包含一个由大型语言模型微调而来的基础 CAMR 解析器和 4 个句类专家和 1 个古汉语 LoRA 专家模型。最终，本文所提出的框架在三个评测数据集中均取得了最好的成绩。

关键词： 中文抽象语义表示；语义解析；专家系统；句类

System Report for CCL24-Eval Task 4: Chinese AMR Parsing framework with Mixture of LoRA Experts

Zihao Wu, Hua Yin*, Ziqian Gao, Jiajia Zhang, Yuelei Ji, Kuntian Tang
Guangdong University of Finance & Economics,School of Informatics,Guangzhou
{zehro,GaoZiqian,zhangjiajia,132382851009jyl,TangKuntian}
@student.gdufe.edu.cn
yinhua@gdufe.edu.cn

Abstract

This paper introduces the system we submitted for the Chinese Abstract Meaning Representation Parsing Evaluation Task at the 23rd Chinese National Conference on Computational Linguistics. Abstract Meaning Representation (AMR) uses a directed acyclic graph to model sentences, with semantic concepts as nodes and relationship labels as arcs to represent the semantics of a sentence. Inspired by the research on AMR parsing that combines grammatical information, we proposed a CAMR parsing framework that mixes LoRA (Low-Rank Adaption) experts, which consists of a basic CAMR parser fine-tuned from a large language model, 4 sentence type experts, and 1 ancient ancient LoRA expert model. In the end, our proposed framework achieved great results.

Keywords: Chinese Abstract Meaning Representation , Semantic Parsing , Expert system , Sentence type

*通讯作者

(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

抽象语义表示 (Abstract Meaning Representation，AMR)(Banarescu et al., 2013) 使用有向无环图对句子进行建模，语义概念作为节点，关系标签作为边，是一种句子级语义表示方法。AMR 已经广泛应用于 NLP 的下游任务中，包括文本摘要 (Nagalavi and Hanumanthappa, 2019)，对话系统 (Bai et al., 2021) 和知识库问答 (Kapanipathi et al., 2021)。由于中英文的词法与句法存在较大差异，中文 AMR(CAMR) 在 AMR 的基础上根据中文的特点进行了扩充和修改 (Li et al., 2019)，主要的改动是引入了文本对齐标注的新方法，AMR 采用单词首字母作为概念节点标签，而 CAMR 在概念节点在标签上使用的是“$xn(n \in N)$”，其中 n 从 1 开始依次递增。句子：“这是什么原因?”的概念节点标签为“x1_ 这 x2_ 是 x3_ 什么 x4_ 原因 x5_?”，并且 CAMR 新增了关系对齐，将关系概念标注在有向弧上，如图 1 所示。这使得将中文句子解析为 CAMR 图时，不能直接采

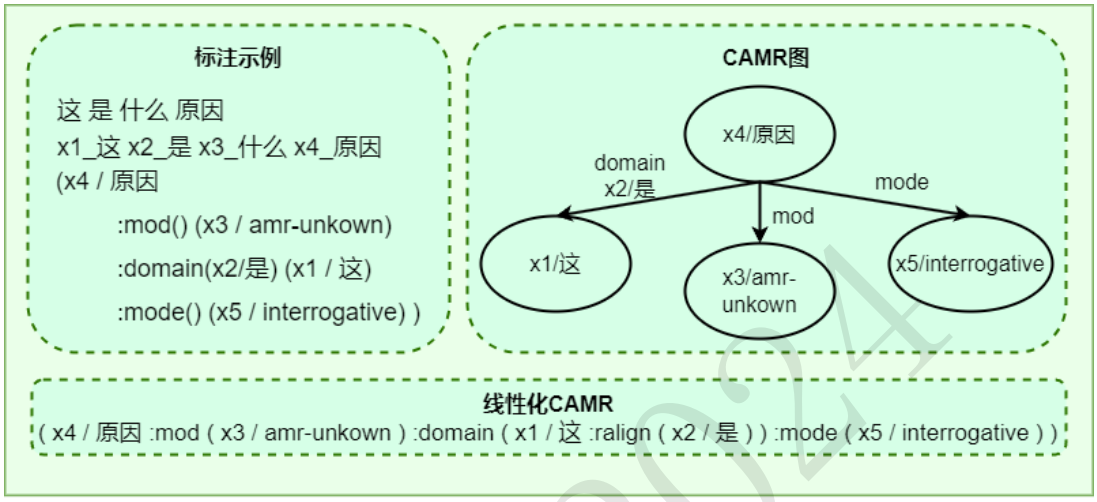


Figure 1: 句子：“这是什么原因?”的 CAMR 示例

用 AMR 解析方法。AMR 解析方法分为四种 (尹华 et al., 2024)：基于图的方法、基于转移的方法、基于形式化的方法和基于 seq2seq 的方法。基于图的方法根据概念识别和关系识别来构建 AMR 图 (Flanigan et al., 2014)；基于转移的方法将输入的句子和依存句法树转换成初始状态，再通过一系列的转换操作完成 AMR 解析 (Wang et al., 2015)；基于形式化的方法通过引入不同的文法、代数方法对图结构数据进行形式化建模，生成中间形式供模型学习，将 AMR 解析任务转换为相关问题求解 (Peng et al., 2015)；实验显示，现在 SoTA(State-of-The-Art) 的 AMR 解析方法是基于 seq2seq 的方法。AMR 图无损线性化的方法首次由Bevilacqua et al. (2021) 提出, 他们基于 BART 构建了 seq2seq 模型 SPRING(Symmetric PaRsIng aNd Generation)，同时实现了 AMR 的解析和生成任务。大语言模型在 NLP 各子任务中展现出卓越的能力，研究者们开始使用更大的基座模型，Lee et al. (2023) 指令微调 FLAN-T5 模型 (Chung et al., 2024)，达到了 AMR 解析的 SoTA。由于 AMR 与语法高度相关，研究者们正不断探索如何在解析器中结合额外的语法信息来提升 AMR 解析性能，最常见的做法是从句法角度，使用额外的模块或者神经网络将句法知识融入到 AMR 解析的过程中，以期望提升 AMR 解析的性能。Sataer et al. (2023) 为模型添加两种额外的句法感知结构将句法知识整合到 PLM(Pre-trained Language Model) 中。Sataer et al. (2024) 使用额外的自注意力模块将不同粒度的文本信息融入到 PLM 中，但过多自注意力模块会带来过多的推理成本。在 CAMRP2022(李斌 et al., 2023) 解析评测任务中，基于图的方法 SUDA-HUAWEI[1] 取得了最佳的成绩，PKU(Chen et al., 2022) 也使用了上述方法。在 CAMRP2023(Xu et al., 2023) 中，Gao et al. (2023) 对 Baichuan-7B 模型[2] 进行全参数微调，达到了同使用 PLM 进行 CAMR 解析工作可比的结果，Yang and Ziming Cheng (2023) 在 ChatGPT3.5(Ouyang et al., 2022) 上进行中文 AMR 的零样本和少样本学习，展现了一定的 CAMR 解析能力。Gu et al. (2023) 训练了一个依存句法和

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

[1]https://github.com/zsLin177/camr

[2]https://github.com/Baichuan-inc/Baichuan-7B

词性联合标注模型，得到对应的句法信息后使用 BiLSTM(Zhang et al., 2015) 将其融入到 PLM 中，但其受限于依存句法信息的质量。当前主流的做法是融合语言学信息来提升 CAMR 解析的准确率。

根据现代汉语语法，中文句子可以按照其语气语用被分为陈述句，疑问句，感叹句和祈使句四种句子类型 (邵敬敏, 2007)。陈述句向听话人传递信息；疑问句询问听话者，希望获取信息；感叹句抒发强烈感情；祈使句则是要求听话人做某事，每种句类通常都有其独特的语言学结构。也有研究者尝试从句类角度进行探索。Yan et al. (2020) 首次分析并标注了一个 CAMR 疑问数据集，但并未涉及解析工作。在 CAMRP2023 中，新增了问句测试集以期望探索 CAMR 解析器对问句的解析能力，但并没有参赛队专门针对这一方面展开研究。

AMR 解析的研究中使用基座模型的参数越来越大，如何在大模型上添加语法信息以提升 AMR 解析性能是一个值得探索的新方向。我们在前期模型 CAPPST(Wu et al., 2023) 的基础上进行扩充与完善，提出了一种全新的 CAMR 解析框架，该框架包含一个由大型语言模型微调而来的基础 CAMR 解析器，并带有 4 个句类和 1 个古汉语 LoRA(Low-Rank Adaption)(Hu et al., 2021) 专家模型，该框架会根据输入来选择性激活专家模型以完成 CAMR 解析。实验结果表明，该框架在不依赖任何其它形式数据和不影响基础解析器性能的情况下，可以有效捕获各类型数据的特征，与未带专家模型的基础解析器相比有明显的性能提升。我们的代码开源在https://github.com/Zehrooo/CAMRP-MoE。

2 方法

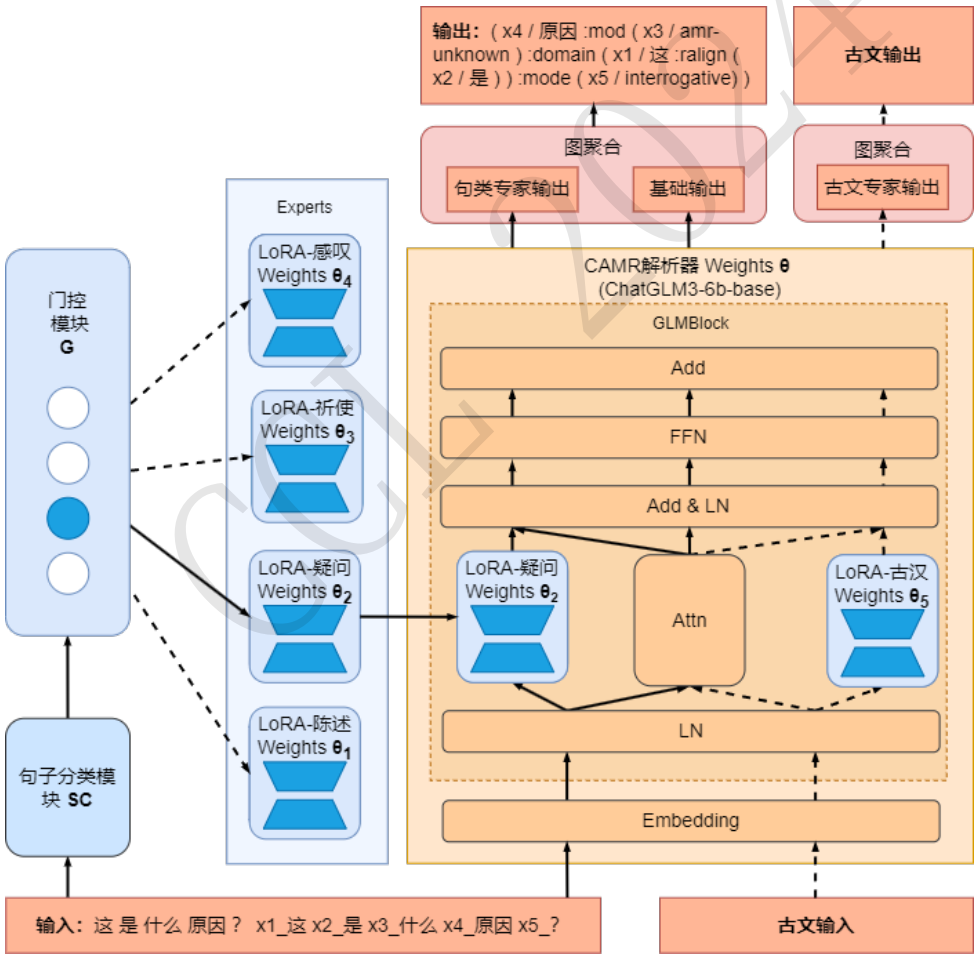


Figure 2: 框架结构示意图

我们把 CAMR 解析视为一种序列到序列的文本生成任务，即输入一个分词后的句子 $S = \{s1, s2, ...sn\}$ 及其带有概念节点标签的句子 $W = \{w1, w2, ...wn\}$，让模型输出 CAMR 图的 DFS 线性化序列 $Y = \{y1, y2, ..., ym\}$，如图 1 所示。线性化的步骤借鉴了Gu et al. (2023) 的处

(c) 2024 中国中文信息学会计算语言学专业委员会

理方法，将虚词和关系对齐处理成额外的边与节点。随后全参数微调 ChatGLM3-6B-base 获得基础 CAMR 解析器。在 CAMR 中每种句类有其独特的结构，表示语气的概念主要有：interrogative 疑问、expressive 感叹、imperative 祈使三种，语气的表示一般通过语气概念和语义关系 mode 来共同表示。对疑问句的处理，是非问、选择问和正反问三种需要用到关系 mode 和疑问概念 interrogative 来表示，例如“:mode (xn / interrogative) ”。同理，祈使、感叹语气也是将其标在分句或者整句的根节点上。基于此特性，我们在基础解析器上训练了 4 个句类 LoRA 专家。为了确保激活正确的 LoRA 专家，还设计了句子分类模块和门控模块，这两个模块共同决定激活哪个 LoRA 专家。此外，针对今年新增的古汉语数据集 (TestC) 也训练了古汉语 LoRA 专家。最后，使用图聚合方法 (Hoang et al., 2021) 将输出结合以获得更好的结果。框架结构示意图如图 2 所示。

2.1 全参数微调 ChatGLM3-6B-base

GLM 采用 Encoder-Decoder 架构 (Du et al., 2022)，与传统 Transformer(Vaswani et al., 2017) 的具体区别如下：GLM 对层归一化和残差连接的顺序进行了调整，使用单层线性层进行输出 token 预测，并用 GeLU(Hendrycks and Gimpel, 2016) 替换了 ReLU 激活函数。ChatGLM3-6B[3]是智谱 AI 和清华大学 KEG 实验室联合发布的对话预训练模型，其基础模型 ChatGLM3-6B-Base 采用了更多样的训练数据、更充分的训练步数和更合理的训练策略，并在语义、推理、代码等不同角度的测评中取得了令人亮眼的表现，同时最大支持 8192 的上下文窗口长度足以满足 CAMR 解析的需求。鉴于 CAMR 解析是复杂的结构化文本生成任务，我们出于以下考虑选择模型：

(1) 不仅需要有良好的文本生成能力，也需要具有良好的文本理解能力。

(2) 具备一定的代码能力，线性化的 CAMR 序列在一定程度上与代码块类似，模型可以原生地完成括号匹配而无需额外后处理。

```
"conversations": [
    {
     "role": "user",
     "content": "这 是 什么 原因 ？ x1_这 x2_是 x3_什么 x4_原因 x5_？ "
    },
    {
     "role": "assistant",
     "content": "( x4 / 原因 :mod ( x3 / amr-unknown ) :domain ( x1 / 这 :ralign ( x2 / 是 )
     ) :mode ( x5 / interrogative) )"
    }
]
```

Figure 3: 用于训练解析器的对话对示例，句子：“这是什么原因?”

如图 3 所示是句子“这是什么原因?”的微调数据示例，为了使模型更能专注所有的输入，在此并未额外添加任何指令。形式化地，将分好词后的句子 S：“这 是 什么 原因 ?”与带有词编号的句子 W：“x1_ 这 x2_ 是 x3_ 什么 x4_ 原因 x5_?”组合作为模型的输入 X，将对应的 CAMR 线性化序列 O：“(x4 / 原因 :mod (x3 / amr-unknown) :domain (x1 / 这 :ralign (x2 / 是)) :mode (x5 / interrogative))”作为模型的输出，单个 GLMBlock 的输出 $Basic_\theta(X)$ 可表示为：

$$X_\theta = X + f^{Attn}(LN(X);\theta) \tag{1}$$

$$Basic_\theta(X) = X_\theta + f^{FFN}(LN(X_\theta);\theta) \tag{2}$$

其中 $f^{Attn}(\cdot)$ 表示自注意力层，$f^{FFN}(\cdot)$ 表示前馈层，$LN(\cdot)$ 表示 LayerNorm 操作，θ 表示各层的权重。

2.2 LoRA 专家

参数高效微调 (Parameter-Efficient Fine-Tuning, PEFT) 提供了以低成本微调 LLM 的方法，LoRA 是其中最具代表性的方法之一，该研究认为在对模型进行微调时，权重的更新具有低秩特性，在训练时冻结其它部分的权重，仅对两个低秩矩阵的权重进行更新从而对模型的权重更新进行约束，以极小的训练成本达到了媲美全参数微调的效果，并且几乎不会引入额外的推理延迟。在 AMR/CAMR 解析中已有研究者尝试使用该方法 (Lee et al., 2023; Yang and Ziming Cheng,

[3]https://github.com/THUDM/ChatGLM3/

(c) 2024 中国中文信息学会计算语言学专业委员会

2023)，但这些研究并未考虑结合语法信息。结合当前句类知识的 AMR 解析研究仍较缺乏的情况，我们认为有必要从句类角度对 CAMR 解析进行研究。混合专家模型 (Mixture of Experts, MoE)(Jacobs et al., 1991) 旨在将多个专家模型的优势结合起来，以提高模型在不同任务下的性能，将 PEFT 与 MoE 结合可以使专家模块的训练成本和模型的推理成本控制在一个可以接受的范围内。

如图 2 所示，在基础解析器上使用 5 种不同类型的数据继续训练对应的 LoRA 专家，分别为陈述句类专家、疑问句类专家、感叹句类专家、祈使句类专家和古汉语专家。句类专家的激活由句子分类模块和门控模块决定。形式化地，带有 LoRA 专家的 GLMBlock 输出 $LE_{\theta i}(X)$ 表示为：

$$X_{\theta i} = X + f^{Attn}(LN(X); \theta_i) \tag{3}$$

$$LE_{\theta i}(X) = X_{\theta i} + f^{FFN}(LN(X_{\theta i}); \theta) \tag{4}$$

其中 θ_i 表示 LoRA 专家 i 的权重。

2.3 句子分类模块

```
"conversations": [
    {
     "role": "system",
     "content": "你是句子分类器，你需要分辨出句子是陈述句、疑问句、感叹句还是祈使句。"
    },
    {
     "role": "user",
     "content": "世界 无奇不有 !"
    },
    {
     "role": "assistant",
     "content": "感叹句"
    }
]
```

Figure 4: 用于训练句子分类模块的对话对示例，句子：“世界无奇不有!”

ChatGLM3-6B 是 ChatGLM3-6B-base 的对话模型，具备一定的对话能力，因此选择对 ChatGLM3-6B 进行 LoRA 微调，期望模型完成句子分类功能。微调数据的实例如图 4 所示，区别于微调 ChatGLM3-6B-base 的数据，此处新增了 system prompt 以约束模型的输出为 4 个句子类别的名称。与上面一致，只计算角色为”assistant” 即期望输出的 loss。形式化地，输入一个分好词后的句子 $S = \{s1, s2, ...sn\}$，模型输出为

$$SC_i(S) = OneHot(type_i) \tag{5}$$

其中 $i \in$ (陈述句，感叹句，祈使句，疑问句)，$OneHot(\cdot)$ 表示进行独热编码。

2.4 门控模块

门控模块根据句子分类模块的输出决定激活哪个 LoRA 专家。本次评测任务新增了古汉语测试集，因此设计了古汉语专家，该 LoRA 专家不受门控模块控制。形式化地，门控模块的输出表示为：

$$Gate_i(SC_i(S)) = \begin{cases} 1, & \text{if } i = \text{argmax}(SC_i(S)) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

其中 $argmax(\cdot)$ 表示获取最大值的索引。

2.5 图聚合

分句和非短句会同时存在一部分的普通结构和另一部分的特殊句类结构，而句类专家专门用于解析特殊句类，对于普通结构的解析性能并不如基础解析器。出于上述原因，通过使用图聚合操作将两个基础解析器 (更均衡) 的输出与两个带有 LoRA 专家 (更特殊) 的 parser 的输出进

CCL 2024

行结合，以期望能得到更好的结果。对于古汉语则使用两个不同训练步长的 LoRA 专家以得到差异化的输出。形式化地，最终的输出表示为

$$Output = \begin{cases} GE(2 \times \text{Basic\_Output}, 2 \times \text{LE\_Output}), & \text{Input} \in \text{Modern Chinese} \\ GE(2 \times \text{LE}_{Anct.}\text{Output}, 2 \times \text{LE}'_{Anct.}\text{Output}), & \text{Input} \in \text{Ancient Chinese} \end{cases} \quad (7)$$

其中 $GE(\cdot)$ 表示图聚合操作 (Graph Ensemble)。

3 实验

3.1 实验设置

根据句类特有的概念节点统计各句类在现代汉语数据集中的分布情况，TestA 和 TestB 的句类分布情况如表 1 所示。其中陈述句类 (Normal) 的数量占比最多，在各数据集占比 79.9%-89.8% 不等，祈使句类 (Imperative) 的数量占比最少，占比 0.2%-1.2% 不等，疑问句类 (Interrogative) 和感叹句类 (Exclamation) 的数量占比相仿，分别是 5.4%-10.5% 和 5.1%-10.8%。由于并未在古汉语解析中使用句类专家，故不对 TestC 进行句类分布统计。

Table 1: 句类在现代汉语数据集 TestA 和 TestB 中的分布情况

Sentence type	Train	Dev	Test A	Test B
Total	16576	1789	1713	1999
Interrogative	901	189	184	129
Exclamation	850	177	186	167
Imperative	36	16	6	25
Normal	14898	1441	1369	1696

全参数微调 ChatGLM3-6B-base 在 8 张 A40 显卡上使用 18365 条数据 (训练集与开发集) 对 ChatGLM3-6B-base 进行全参数微调以得到基础解析器 (Basic Parser)。训练的超参数设置如下：training_step=3000，batchsize=48，learning_rate=1e-6。此外，还尝试使用模型推理 20 万条 THUNews(Sun et al., 2016) 以生成银数据 (silver data) 用作训练，但并未提升基础解析器的性能 (80.47, TestA)。

LoRA **专家** LoRA 专家在单张 A40 显卡上进行训练，经过实验探索发现，使用混合数据训练的 LoRA 专家表现优于使用纯训练集的数据或纯银数据，因此，对于陈述句专家、疑问句专家和感叹句专家，训练数据由 1000 条来自训练集的数据和 2000 条银数据组成，共计 3000 条数据。而由于祈使句的数据较少，祈使句专家的训练数据则由 52 条来自训练集的数据和 948 条银数据组成，共计 1000 条数据。对于古汉语专家，训练数据则是 CAMRP2024 主办方提供的 500 条开发集数据。LoRA 专家训练的超参数设置如下：training_step=500，batchsize=6，learning_rate=1e-6，r=16，alpha=64。

句子分类模块类似地，句子分类模块在单张 A40 显卡上进行训练，训练集使用 CAMR2.0v 中的 4169 条数据。句子分类模块训练的超参数设置如下：training_step=33352(=8epoch)，batchsize=5，learning_rate=1e-6，r=64，alpha=256。对句子分类模块准确度的计算公式如下：

$$\text{Accuracy}(SC) = \frac{1}{n}\sum_{k=1}^{n}\delta(SC_i(S_k) \in A_k) \quad (8)$$

其中 n 是句子的总数，$SC_i(S_k)$ 是分类器对第 k 个句子预测的类别，A_k 是第 k 个句子的实际类别集合，一个句子可能同时是感叹句和祈使句，$delta(\cdot)$ 是一个指示函数，如果 $type_i \in A_i$ 为真，则为 1，否则为 0。

CAMR 解析的评价指标采用 AlignSmatch(Xiao et al., 2022)。

3.2 实验结果与分析

通过对 Chinese-BERT-Large(Cui et al., 2021) 和 ChatGLM-6B 模型进行微调，以完成句子类型多分类任务。表 2 列出了句子分类模块在现代汉语数据集上的准确率，微调后的 Chinese-BERT-Large 在 TestA 数据集上的准确率仅有 54.4%，难以胜任句子分类任务，而 ChatGLM-6B-LoRA

(c) 2024 中国中文信息学会计算语言学专业委员会

Table 2: 句子分类模块在现代汉语数据集上的准确率 (%)

Classifier	ACC	
	TestA	TestB
Chinese-BERT-Large	54.46	\
ChatGLM3-6B-LoRA	95.24	97.34
ChatGLM3-6B-LoRA × 3	**98.10**	**98.54**

的准确率则达到了 95.2%。受到Li et al. (2024) 的启发，聚合多个模型的输出通常能得到更好的性能。因此对三个不同训练步长模型的输出进行投票聚合，实验结果表明，该方法在 TestA 上提升了 2.9% 的准确率，在 TestB 上提升了 1.2% 的准确率。

Table 3: 基础解析器与各专家在三个数据集上的得分比较 (%)

Parser	Sentence Type	Test A			Test B			Test C		
		P	R	F_1	P	R	F_1	P	R	F_1
Basic Parser	All.	80.68	80.81	80.74	75.26	74.77	75.01	58.64	59.16	58.90
Experts		80.78	81.02	80.90	75.12	74.65	74.89	70.51	70.49	70.50
Ensemble Parser		80.87	81.53	**81.19**	74.80	75.75	**75.27**	70.62	72.52	**71.56**
Basic Parser	Norm.	81.05	81.02	81.04	75.22	74.90	75.06			
Norm-Expert		81.12	81.08	81.10	75.11	74.75	74.93			
Ensemble Parser		81.07	81.47	**81.27**	74.73	75.86	**75.29**			
Basic Parser	Excl.	79.11	79.45	79.28	74.68	72.87	73.76			
Excl-Expert.		80.50	80.00	80.25	74.97	73.65	**74.30**			
Ensemble Parser		80.20	81.33	**80.76**	74.27	74.04	74.16			
Basic Parser	Impr.	77.29	72.58	74.86	72.56	71.41	71.98			
Impr-Expert		78.01	75.63	**76.80**	74.64	72.79	73.71			
Ensemble Parser		78.01	75.63	**76.80**	74.94	73.39	**74.16**			
Basic Parser	Intr.	79.31	80.65	79.98	76.83	75.78	76.30			
Intr-Expert		79.88	81.28	80.57	76.45	74.47	75.45			
Ensemble Parser		79.81	82.50	**81.14**	76.28	76.37	**76.32**			

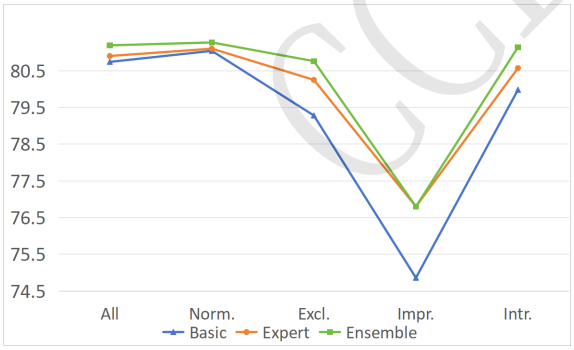


Figure 5: TestA 各得分的折线图

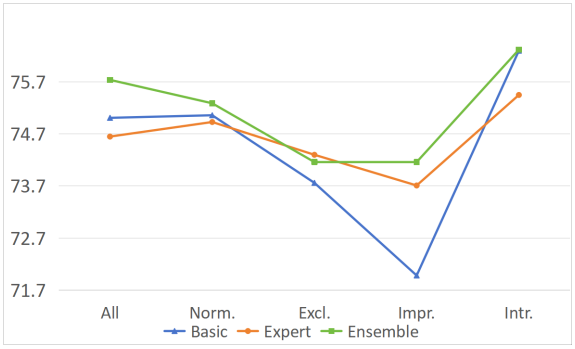


Figure 6: TestB 各得分的折线图

如表 3 所示是基础解析器与各专家在三个数据集上的得分比较。由该表得出两个现代汉语数据集 TestA 和 TestB 的各得分折线图如图 5 图 6 所示，其中横坐标表示不同的解析类别：所有句子类别 (All)、陈述句类 (Norm.)、感叹句类 (Excl.)、祈使句类 (Impr.) 和疑问句类 (Intr.)，纵坐标是 F_1 分数。句类专家和基础解析器 (Basic) 在单张图上共有 10 个得分表现，在 TestA 中，句类专家在所有类别表现得分高于基础解析器的得分，但在 TestB 中，只有感叹句类专家和祈使句类专家的表现优于基础解析器。值得注意的是，句类专家在祈使句类的表现提升最大，在 TestA 与 TestB 中较基础解析器的 F_1 分数分别提升 1.94% 与 1.73%，感叹句类的提升也相当显

著，分别为 0.97% 和 0.54%，陈述句类专家和疑问句类专家在 TestA 中的表现仍有提升，分别提升 0.06% 和 0.59%，但在 TestB 中的表现则不及基础解析器，分别下降 0.13% 和 0.85%。将所有专家的结果进行整合后 (Experts) 与基础解析器进行比较发现，其在 TestA 中的 F_1 分数提升为 0.26%，而在 TestB 中则下降 0.12%，这是由于疑问句类专家和陈述句类专家在 TestB 上表现不佳导致的。古汉语 (Anct.) 专家的表现则相当亮眼，其在 TestC 上的 F_1 分数表现高出基础解析器 11.60%。在进行图聚合 (Ensemble) 操作后的表现大都优于直接使用句类专家或基础解析器的输出，但在 TestB 中感叹句类的表现仍不如句类专家。最后，本文所提出的解析框架在 TestB 的总体表现为 75.27%，不及 TestA 的 81.19%，这或许是由于语料来源不同导致的，TestA 的语料选自宾州中文树库 CTB8.0，来源类别包括新闻网络论坛等，而 TestB 则选自人教版小学语文课本 (李斌 et al., 2023)。其中基础解析器使用与 TestA 同源的语料进行训练，训练 LoRA 专家所使用的银数据则全部来自新闻语料，训练数据来源的差异往往导致模型的泛化能力不足，即便使用大语言模型作为基座，仍无法完全解决这一问题。

Table 4: 评测提交 Align-Smatch 得分对比 (%)

Team	Test A			Test B			Test C		
	P	R	F_1	P	R	F_1	P	R	F_1
BLCU	78.93	78.89	78.91	74.04	74.27	74.16	56.87	58.59	57.72
HITSZ	80.80	81.11	80.96	75.13	74.57	74.85	67.06	66.77	66.91
GDUFE	80.87	81.53	**81.19**	74.80	75.75	**75.27**	70.62	72.52	**71.56**

在本次评测任务中，参加 open 赛道的三支队伍分别是北京语言大学 (BLCU)、哈尔滨工业大学 (深圳)(HITSZ) 和广东财经大学 (GDUFE)，测评结果如表 4 所示。本文所提出带有专家模块的 CAMR 解析框架在本次评测任务三个测试集中均取得了最好的成绩，在现代汉语数据集 TestA 和 TestB 中分别领先第二名 0.23% 和 0.45%，在古汉语测试集上的表现则领先第二名 4.65%，其中 BLCU 和 HITSZ 的参赛系统都是基于现代汉语数据集构建，并未使用主办方所提供的古汉语数据集。但三个参赛队的成绩均不如在现代汉语数据集中的表现，这可能是受基座大模型预训练所用语料的限制，导致其对古汉语的理解能力不如对现代汉语。相较于现代汉语，古汉语的解析表现仍有较大提升空间，亟需更深入的研究与探索。

4 结语

在本次 CAMRP2024 评测任务中，我们首次提出了结合 LoRA 专家的 CAMR 解析器框架，该框架易于扩展、无需过多的训练及推理成本，也不依赖其他形式的数据。实验结果表明，结合专家模块的解析方法可以有效地提高 CAMR 解析器的性能。值得注意的是，在本次评测任务中所有参赛队都使用了大语言模型作为基座模型，与 CAMRP2023 中的Gu et al. (2023) 在 TestA 上的 81.30% 和 TestB 上的 74.71% 相比，参数规模更大的解析器并未带来飞跃性的表现提升，使用何种方法来继续提升 CAMR 解析性能需要更深入的研究与探索。因此，在未来可从以下几个方面进行拓展：增加更多的专家以提高泛化能力。使用结构更多样的专家学习不同形式数据的特征。引入可学习的门控模块，使其可以根据输入自主调整激活专家的数量等。

致谢

本研究受教育部人文社会科学研究青年基金项目 (21YJCZH202) 和广东省法学会法学研究委托课题项目 (GDLS(2024C12)) 资助。

References

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. Semantic representation for dialogue modeling. 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 4430–4445, 2021.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for

(c) 2024 中国中文信息学会计算语言学专业委员会

sembanking. In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, pages 178–186, 2013.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 12564–12573, 2021.

Liang Chen, Bofei Gao, and Baobao Chang. A two-stage method for chinese amr parsing. arXiv preprint arXiv:2209.14512, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70):1–53, 2024.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3504–3514, 2021. doi: 10.1109/TASLP.2021.3124365.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, 2022.

Jeffrey Flanigan, Sam Thomson, Jaime G Carbonell, Chris Dyer, and Noah A Smith. A discriminative graph-based parser for the abstract meaning representation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1426–1436, 2014.

Wenyang Gao, Xuefeng Bai, and Yue Zhang. CCL23-eval 任务 2 系统报告:WestlakeNLP, 基于生成式大语言模型的中文抽象语义表示解析 (system report for CCL23-eval task 2: WestlakeNLP, investigating generative large language models for Chinese AMR parsing). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations), pages 64–69, Harbin, China, August 2023. Chinese Information Processing Society of China. URL https://aclanthology.org/2023.ccl-3.6.

Yanggan Gu, Shilin Zhou, and Zhenghua Li. CCL23-eval 任务 2 系统报告: 基于图融合的自回归和非自回归中文 AMR 语义分析 (system report for CCL23-eval task 2: Autoregressive and non-autoregressive Chinese AMR semantic parsing based on graph ensembling). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations), pages 53–63, Harbin, China, August 2023. Chinese Information Processing Society of China. URL https://aclanthology.org/2023.ccl-3.5.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam Nguyen, Dzung Phan, Vanessa López, and Ramon Fernandez Astudillo. Ensembling graph predictions for amr parsing. Advances in Neural Information Processing Systems, 34:8495–8505, 2021.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. Neural computation, 3(1):79–87, 1991.

(c) 2024 中国中文信息学会计算语言学专业委员会

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue-Nkoutche, et al. Leveraging abstract meaning representation for knowledge base question answering. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3884–3894, 2021.

Young-Suk Lee, Ramón Fernandez Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. Amr parsing with instruction fine-tuned pre-trained language models. arXiv preprint arXiv:2304.12272, 2023.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, and Nianwen Xue. Building a chinese amr bank with concept and relation alignments. Linguistic issues in language technology, 18, 2019.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need, 2024.

Deepa Nagalavi and M Hanumanthappa. The nlp techniques for automatic multi-article news summarization based on abstract meaning representation. In Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, pages 253–260. Springer, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

Xiaochang Peng, Linfeng Song, and Daniel Gildea. A synchronous hyperedge replacement grammar based approach for amr parsing. In Proceedings of the nineteenth conference on computational natural language learning, pages 32–41, 2015.

Yikemaiti Sataer, Zhiqiang Gao, Yunlong Fan, Bin Li, Miao Gao, and Chuanqi Shi. Exploration and comparison of diverse approaches for integrating syntactic knowledge into amr parsing. Applied Intelligence, 53(24):30757–30777, 2023.

Yikemaiti Sataer, Yunlong Fan, Bin Li, Miao Gao, Chuanqi Shi, and Zhiqiang Gao. Hierarchical information matters! improving amr parsing with multi-granularity representation interactions. Information Processing & Management, 61(3):103698, 2024.

Maosong Sun, Jingyang Li, Zhipeng Guo, Zhao Yu, Yabin Zheng, Xiance Si, and Zhiyuan Liu. Thuctc: an efficient chinese text classifier. GitHub Repository, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. A transition-based algorithm for amr parsing. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 366–375, 2015.

Zihao Wu, Hua Yin, Yuelei Ji, Hanlin Wang, Yiliang Lu, and Ziqian Gao. Cappst: Chinese amr parsing with parameter-efficient fine-tuned pre-trained language model for particular sentence types. In Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing, pages 130–134, 2023.

Liming Xiao, Bin Li, Zhixing Xu, Kairui Huo, Minxuan Feng, Junsheng Zhou, and Weiguang Qu. Alignsmatch: A novel evaluation method for chinese abstract meaning representation parsing based on alignment of concept and relation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5938–5945, 2022.

(c) 2024 中国中文信息学会计算语言学专业委员会

Zhixing Xu, Yixuan Zhang, Bin Li, Zhou Junsheng, and Weiguang Qu. Overview of ccl23-eval task 2: The third chinese abstract meaning representation parsing evaluation. In Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations), pages 70–83, 2023.

Peiyi Yan, Bin Li, Tong Huang, Kairui Huo, Jin Chen, and Weiguang Qu. Chinese interrogative sentences annotation and analysis based on the abstract meaning representation. In Proceedings of the 19th Chinese National Conference on Computational Linguistics, pages 77–87, 2020.

Yifei Yang and Hai Zhao Ziming Cheng. System report for ccl23-eval task 2:chinese abstract meaning representation parsing based on large language model. In Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations), pages 41–52, 2023.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation, pages 73–78, 2015.

尹华, 卢懿亮, 季跃蕾, 吴梓浩, and 彭亚男. 抽象语义表示解析方法研究综述. 中文信息学报, 38 (03):1–23, 2024. ISSN 1003-0077.

李斌, 许智星, 肖力铭, 周俊生, 曲维光, and 薛念文. 第二届中文抽象语义表示解析评测. 中文信息学报, 37:33–43, 2023. ISSN 1003-0077.

邵敬敏. 现代汉语通论 (第二版). 上海教育出版社, 2007.

CCL 2024

System Report for CCL24-Eval Task 4: A Two-stage Generative Chinese AMR Parsing Method Based on Large Language Models

Zizhuo Shen, Yanqiu Shao✉, Wei Li
School of Information Science, Beijing Language and Culture University
Beijing 100083, China
blcushzz@gmail.com
yqshao163@163.com
liweitj47@blcu.edu.cn

Abstract

The purpose of the CAMR task is to convert natural language into a formalized semantic representation in the form of a graph structure. Due to the complexity of the AMR graph structure, traditional AMR automatic parsing methods often require the design of complex models and strategies. Thanks to the powerful generative capabilities of LLMs, adopting an autoregressive generative approach for AMR parsing has many advantages such as simple modeling and strong extensibility. To further explore the generative AMR automatic parsing technology based on LLMs, we design a two-stage AMR automatic parsing method based on LLMs in this CAMR evaluation. Specifically, we design two pipeline subtasks of alignment-aware node generation and relationship-aware node generation to reduce the difficulty of LLM understanding and generation. Additionally, to boost the system's transferability, we incorporate a retrieval-augmented strategy during both training and inference phases. The experimental results show that the method we proposed has achieved promising results in this evaluation.

1 Introduction

Semantic parsing is one of the fundamental tasks in the field of Natural Language Processing (NLP). As an important formal method of semantic parsing, Abstract Meaning Representation (AMR) (Banarescu et al., 2013) has attracted widespread attention from researchers in the NLP field. AMR represents the semantics of a sentence in the form of a single-rooted directed acyclic graph (DAG), where nodes are used to represent semantic concepts abstracted from the sentence, and edges are used to represent the semantic relations between concepts. Thanks to the development of AMR technology, researchers have also begun to try to integrate the semantic knowledge in AMR graphs into application systems such as machine translation (Song et al., 2019), text summarization (Liao et al., 2018), and event extraction (Xu et al., 2022), in order to enhance the system's understanding of semantic knowledge and improve system performance.

The data resources and automatic parsing technologies associated with AMR primarily concentrate on English. Li et al. (2021) considering the traits of Chinese, establish the annotation standards and evaluation methods fitting for Chinese AMR (CAMR), significantly advancing the progress of Chinese AMR research.

The existing CAMR automatic parsing technologies can be divided into three categories: transition-based methods (Wang et al., 2018), graph-based methods (Chen et al., 2022), and autoregressive generative methods (Bevilacqua et al., 2022). The transition-based method considers the AMR automatic parsing problem as a sequence-to-action conversion issue. This method usually relies on a complex transition system for action definition and feature extraction used for action prediction. The graph-based method treats the AMR automatic parsing problem as a graph search issue, that is, searching for the highest scoring subgraph from a directed complete graph. Due to the complexity of the AMR automatic parsing task, the graph-based method usually requires the construction of multiple models such as concept recognition and relation recognition to achieve a complete parsing process.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 154-159, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

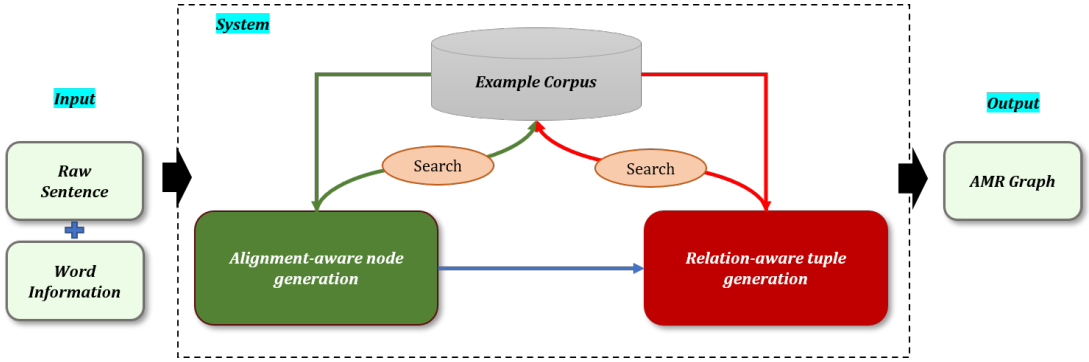


Figure 1: A Brief Flowchart of Our Method.

The autoregressive generative method considers the AMR automatic parsing problem as a sequence-to-sequence generation issue. Compared with the other two methods, the autoregressive generative method has the advantages of simple modeling and easy implementation, and has become the mainstream method for solving AMR automatic parsing problems. Recently, thanks to the powerful generation capabilities of large language models (LLMs) (Zhao et al., 2023), the autoregressive generative methods based on LLMs have also been adopted by more and more researchers and achieved competitive results (Lee et al., 2023).

The existing autoregressive generative methods based on LLMs usually adopt a strategy of generating AMR graphs in a single stage (Gao et al., 2023), such as directly generating linearized sequences of AMR graphs or generating tuples within AMR graphs. Given the complexity of the AMR graph structure, the difficulty of generating AMR graphs in a single stage is significant, which constitutes a huge challenge for the generation ability of LLMs.

In order to reduce the difficulty of generating AMR graphs by LLMs, we decompose the process of generating AMR graphs into two stages. The first stage is **alignment-aware node generation**, and the second stage is **relation-aware tuple generation**. In the first stage, we first group the nodes according to the alignment type between the nodes in the AMR graph and the words in the sentence, and then let the LLM generate different aligned type nodes in groups. In the second stage, we first group the tuples in the AMR graph according to the relation type, and then let the LLM generate tuples of different relation types in groups. Considering that this shared task requires testing the system's ability to automatically parse ancient Chinese sentences into AMR graphs in a zero-shot scenario, we adopt a **retrieval-augmented instruction tuning** strategy to fine-tune the LLM, with the aim of enhancing our system's transferability.

2 Method

Our method can be divided into three parts to descirbe: alignment-aware node generation, relation-aware tuple generation, and retrieval-augmented instruction tuning strategy. Figure 1 shows the flowchart of our method. Our system takes a sentence and the word information in the sentence as input, and after performing two pipeline tasks of alignment-aware node generation and relation-aware tuple generation, it obtains the entire AMR graph as output.

2.1 Alignment-aware Node Generation

To achieve alignment-aware node generation, we categorize the nodes into several types based on the alignment between the nodes in the AMR graph and the words in the sentence (Chen et al., 2022): **general alignment nodes**, **split alignment nodes**, **multi-word alignment nodes**, **and non-alignment nodes**. The general alignment node refers to a node in the AMR graph that can correspond to a word in the sentence. The split alignment node refers to a node in the AMR graph that can correspond to a part of a word in the sentence. The multi-word alignment node refers to a node in the AMR graph that can correspond to multiple words in the sentence. The non-alignment node refers to a node in the AMR graph that cannot correspond to any word in the sentence. Since the generation patterns of nodes with

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 154-159, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Task	Input	Output

Alignment-aware Node Generation	{
	'sentence': '中国将拓宽利用外资渠道',
	'words': {'x1': '中国', 'x2': '将', 'x3': '拓宽', 'x4': '利用', 'x5': '外资', 'x6': '渠道'}
	}
		{
		'non-align': ['(x0, root, -)', '(x10, country, -)'],
		'align': ['(x1, 中国, -)', '(x2, 将, -)', '(x3, 拓宽-01, -)', '(x4, 利用-01, -)', '(x5, 外资, -)', '(x6, 渠道, -)']
		}

Relation-aware Tuple Generation	{
	'sentence': '中国将拓宽利用外资渠道',
	'nodes': {'non-align': ['(x0, root, -)', '(x10, country, -)'], 'align': ['(x1, 中国, -)', '(x2, 将, -)', '(x3, 拓宽-01, -)', '(x4, 利用-01, -)', '(x5, 外资, -)', '(x6, 渠道, -)'] }
	}
		{
		'(top, - , -)': [['(x0, root, -)', '(x3, 拓宽-01, -)']],
		'(arg0, - , -)': [['(x3, 拓宽-01, -)', '(x10, country, -)']],
		'(tense, - , -)': [['(x3, 拓宽-01, -)', '(x2, 将, -)']],
		'(arg1, - , -)': [['(x3, 拓宽-01, -)', '(x6, 渠道, -)'], ['(x4, 利用-01, -)', '(x5, 外资, -)']],
		'(name, - , -)': [['(x10, country, -)', '(x1, 中国, -)']],
		'(mod, - , -)': [['(x6, 渠道, -)', '(x4, 利用-01, -)']],
		}

Table 1: Input and Output Examples for Two Tasks.

the same alignment type are the same, the alignment-aware node generation task can help LLMs learn the generation patterns of nodes, thereby achieving better node generation effects.

The input for the node generation task is a sentence and the words in the sentence, and the output is the nodes in the AMR graph. Therefore, the node generation task can be described as follows: **Given a sentence and the words in the sentence, predict the nodes corresponding to the sentence in the AMR graph**.

For the words in the sentence, we use a dictionary to store the information of the words in the sentence, where the key of the dictionary is the index of the word and the value of the dictionary is the text of the word. For the nodes in the AMR graph, we use a ***dictionary-list*** data structure to store the grouped node information, where the key of the dictionary is the type of the node and the value of the dictionary is the list of all nodes under that type. For each node, we use a triplet to represent all the information of the node: the index of the node, the text of the node, and the index of the co-referential node. An example of the input and output for this task is given in Table 1.

2.2 Relation-aware Tuple Generation

The input for the tuple generation task is a sentence and the nodes in the AMR graph, with the output being the relational tuples in the AMR graph. Therefore, the tuple generation task can be described as follows: **Given a sentence and the nodes in the AMR graph, predict the relational tuples in the AMR graph**.

For the relational tuples in the AMR graph, we use a ***dictionary-list-list*** data structure to store the grouped tuple information. The key of the dictionary is the relational information, and we use a triplet to represent all the information of the relation: the name of the relation, the index of the aligned word of the relation, and the text of the aligned word of the relation. The value of the dictionary is a two-level nested list, where the first level list represents all the node pairs with the same relation, and the second level list represents the parent and child nodes of each node pair. An example of the input and output for this task is given in Table 1.

2.3 Retrieval-augmented Instruction Tuning Strategy

Based on our definitions of the node generation task and the tuple generation task, we can build prompts for these two tasks separately to be used for instruction tuning. The prompts for these two tasks are composed of the following three parts: task description part, example information part, and input information part. The task description part is a natural language description of the current task. The example part is an input-output pair for the current task. The input information part is the input of the sentence

CCL 2024

to be parsed under the current task. To make the examples more representative, we use a vector retrieval model[0] to find examples for each sentence to be parsed. Specifically, we use sentences in the development set to build a vector retrieval database. For a sentence to be parsed, we vectorize it and then search for the sentence with the highest semantic similarity from the vector retrieval database by calculating the cosine similarity as the final example.

In the instruction tuning stage, we denote the entire prompt as X and the corresponding output as O. Each token in O is denoted as o_i. When calculating the loss of the language model, we only calculate the loss for the output part. The calculation formula is shown below:

$$\mathcal{L} = \sum_{i=1} -logP(o_i|X, o_{i-1}). \quad (1)$$

3 Experiments and Analysis

3.1 Experimental Setup

This shared task includes two scenarios: open testing and closed testing. We conduct experiments in the open testing scenario. The LLM we used is the Baichuan2-13B-Base [1] released by Baichuan Intelligent Technology. Due to computational resource limitations, we adopt the LoRA-based method (Hu et al., 2022) for fine-tuning the LLM. The main training parameters used during the fine-tuning process are listed in Table 2.

Parameter Name	Parameter Value
learning_rate	2e-4
lora_rank	64
lora_alpha	16
lora_dropout	0.05
num_train_epochs	30

Table 2: Main parameters used in fine-tuning.

3.2 Experimental Results

This shared task includes a total of three test sets: TestA, TestB, and TestC. Among them, TestA and TestB are modern Chinese test sets, and TestC is an ancient Chinese test set. The AlignSmatch (Xiao et al., 2022) F-scores of all teams on these test sets are shown in Table 3. BLCU is the result of our system.

According to the results on the test set, we find that the performance of our system has a small gap with other systems on TestA and TestB, but a large gap on TestC. This indicates that the transferability of our system in zero-shot scenarios is still limited.

Team	TestA	TestB	TestC
GDUFE	0.8119	0.7527	0.7156
HITSZ	0.8096	0.7485	0.6692
BLCU	0.7891	0.7416	0.5772

Table 3: The results of each team on the three test sets. GDUFE is the result of Guangdong University of Finance and Economics, HITSZ is the result of Harbin Institute of Technology (Shenzhen), and BLCU is our result.

[0]https://huggingface.co/shibing624/text2vec-base-chinese
[1]https://huggingface.co/baichuan-inc/Baichuan2-13B-Base

CCL 2024

3.3 The Impact of the Retrieval-Augmented Strategy

To verify the impact of the retrieval augmented strategy on system results, we design a set of ablation experiments for result analysis. The experimental results listed in Table 4. Through the experimental results, we find that the retrieval augmented strategy indeed improve the system's performance to some extent. It is worth noting that the retrieval augmented strategy significantly improve the system's performance on TestC. This indicates that the retrieval augmented strategy has advantages in zero-shot scenarios.

Method	TestA	TestB	TestC
w/ retrieval-augmented	**0.7891**	**0.7416**	**0.5772**
w/o retrieval-augmented	0.7725	0.7341	0.5366

Table 4: Comparison of results using the retrieval-augmented strategy.

3.4 Error Analysis

Based on the analysis of model prediction results, we summarize several problems currently existing in our system.

- **The generated format is illegal**. Since we use relatively complex data structures to represent the output of each task, there will be some predictions in the system's prediction results that are illegal in the data structure. This inspires us that we may need to design a more reasonable representation form for CAMR tasks.

- **The generated content is illegal**. Since our system uses an autoregressive generative approach, the system will generate some content that does not match the input information or is inconsistent with the CAMR Schema (for example, the generated relation is not a relation in the CAMR Schema). This inspires us to possibly need to design more complex constrained decoding strategies.

- **Lack of relevant knowledge**. Since ancient Chinese texts are usually historical documents, the knowledge contained therein deviates significantly from the knowledge commonly used in modern Chinese. For example, the system make an error in node prediction for the sentence "命子封二百乘以伐京。". It incorrectly predicted the node "北京" based on "京". This inspires us that we may need to pre-train LLM on ancient Chinese texts.

4 Conclusion

In this paper, we propose a two-stage generative CAMR parsing method based on LLMs. We decompose the CAMR task into two pipeline subtasks: alignment-aware node generation and relation-aware tuple generation to reduce the difficulty of understanding the CAMR task for LLMs. Furthermore, we introduce retrieval-augmented strategy during training and inference to enhance the transfer learning capabilities of our system. On three test sets of the CAMR 2024 evaluation task, our system achieve promising results. Through the analysis of experimental results, we find that our system still performs poorly on the ancient Chinese test set, which inspires us to further enhance our system's CAMR auto-parsing ability in low-resource scenarios through pre-training, data augmentation, and other techniques in future works.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. *Linguistic Annotation Workshop,Linguistic Annotation Workshop*, Aug.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2022. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 1256412573, Sep.

Liang Chen, Bofei Gao, and Baobao Chang. 2022. A two-stage method for chinese amr parsing. Sep.

Wenyang Gao, Xuefeng Bai, and Yue Zhang. 2023. System report for ccl23-eval task 2: Westlakenlp, investigating generative large language models for chinese amr parsing. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 64–69.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Young-Suk Lee, RamonFernandez Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. 2023. Amr parsing with instruction fine-tuned pre-trained language models. Apr.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, and Nianwen Xue. 2021. Building a chinese amr bank with concept and relation alignments. *Linguistic Issues in Language Technology*, Aug.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract meaning representation for multi-document summarization. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1178–1190. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, page 1931, Mar.

Chuan Wang, Bin Li, and Nianwen Xue. 2018. Transition-based chinese amr parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Jan.

Liming Xiao, Bin Li, Zhixing Xu, Kairui Huo, Minxuan Feng, Junsheng Zhou, and Weiguang Qu. 2022. Alignsmatch: A novel evaluation method for chinese abstract meaning representation parsing based on alignment of concept and relation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5938–5945, Marseille, France, June. European Language Resources Association.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A two-stream amr-enhanced model for document-level event argument extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5025–5036. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

CCL 2024

Overview of CCL24-Eval Task 4: The Fourth Chinese Abstract Meaning Representation Parsing Evaluation

Zhixing Xu[1,2], Yixuan Zhang[1,2], Bin Li[1,2], Junsheng Zhou[2,3] and Weiguang Qu[2,3]
1. School of Chinese Language and Literature, Nanjing Normal University, China
2. Center for Language Big Data and Computational Humanities, Nanjing Normal University, China
3. School of Computer and Electronic Information, Nanjing Normal University, China
xzx0828@live.com, zyixuan_12@163.com,
libin.njnu@gmail.com, {zhoujs, wgqu}@njnu.edu.cn

Abstract

Abstract Meaning Representation has become a key research area in sentence-level semantic parsing within natural language processing. Substantial progress has been achieved in various NLP tasks using AMR. This paper presents the fourth Chinese Abstract Meaning Representation parsing evaluation, held during the technical evaluation task workshop at CCL 2024. The evaluation also introduced a new test set comprising Ancient Chinese sentences. Results indicated decent performance, with the top team achieving an F_1 of 0.8382 in the open modality, surpassing the previous record at CoNLL 2020 by 3.30 percentage points under the MRP metric. However, current large language models perform poorly in AMR parsing of Ancient Chinese, highlighting the need for effective training strategies. The complex syntax and semantics of Ancient Chinese pose significant challenges. Additionally, optimizing transfer learning techniques to better apply knowledge from Chinese Mandarin to Ancient Chinese parsing is crucial. Only through continuous innovation and collaboration can significant advancements in both Ancient Chinese and Chinese Mandarin AMR parsing be achieved.

1 Introduction

With the growing maturity of morphological and syntactic analysis techniques, natural language processing (NLP) has advanced to the level of semantic analysis. Sentence-level meaning parsing, in particular, has become central to semantic analysis research. To address the challenges of whole-sentence semantic representation and the domain-dependent nature of sentence semantic annotation, Banarescu et al. (2013) proposed a domain-independent whole-sentence semantic representation method called Abstract Meaning Representation (AMR). AMR abstracts the meaning of a sentence using a single-rooted, acyclic, and directed graph, predicting the semantic structure of the targeted sentence. Large-scale corpora have been constructed for AMR, and several international conferences have been held to evaluate AMR semantic parsing tasks.

The conference of CoNLL 2020 featured a cross-lingual track with five languages, and it was the first time Chinese was included. However, parsing Chinese using AMR has its challenges due to significant syntactic and semantic differences between Chinese Mandarin and English. To address these issues, Li et al. (2016) introduced several major changes to develop Chinese Abstract Meaning Representation (Chinese AMR, CAMR), enhancing its ability to parse Chinese effectively. Similar to AMR, the CAMR corpus has begun to take shape and played an important role from CoNLL 2020.

2 Evaluation Task

Our evaluation task is to parse input sentences and output CAMR graphs of the targeted sentences using data from the CAMR corpus. It is noteworthy that the alignment of concepts and relations, as well as additional semantic role labels, have been incorporated into CAMR to better capture the unique characteristics of Chinese. The evaluation task at CoNLL 2020 did not utilize the alignment of concepts and relations. Therefore, to address this issue, our previous CAMRP 2023 evaluation task (Xu et al., 2023)

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

第二十三届中国计算语言学大会论文集，第160页-第171页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

adopted a newly designed metric named ALIGN-SMATCH (Xiao et al., 2022). This metric includes the alignment of concepts and relations, aiming to better evaluate the performance of automatic parsing. And for the sake of consistency and accuracy, ALIGN-SMATCH remains the main metric this year.

Overall, CAMRP 2024 is a follow-up and extension of CAMRP 2023, with key differences including the addition of a blind test set containing 2,177 Ancient Chinese sentences. This extension aims to further test and refine our parsing methods, ensuring robust performance across different eras and styles of Chinese language.

3 Data Set

The CAMR corpus has been constructed through a collaboration between Nanjing Normal University and Brandeis University since 2015 (Li et al., 2016)(Li et al., 2019). Over the past two years, we have utilized these datasets to evaluate the progression of CAMR parsing. To further assess the generalization performance of parsers, we have introduced a new **Dev Set B** and a blind test set **(Test C)**, consisting of 500 and 2,177 Ancient Chinese sentences, respectively. Consequently, our evaluation includes one training set, two development sets, and three test sets, as shown in Table 1.

The Train Set, Dev Set A, and Test A for CAMRP 2024 are sourced from CAMR v2.0, which is available through the Linguistic Data Consortium (LDC)[1]. This dataset, derived from the Chinese Tree Bank 8.0, comprises a total of 20,078 Chinese sentences. Consistently, the dataset includes training, development, and test sets, which have been proven to be of high quality in the evaluation tasks at CAMRP 2022 (Li et al., 2023) and CAMRP 2023.

Data Set	Sentences	Word Tokens
Train Set	16,576	386,234
Dev Set A	1,789	41,822
Dev Set B (Ancient Chinese)	**500**	**9,130**
Test A	1,713	39,228
Test B	1,999	36,940
Test C (Ancient Chinese)	**2,177**	**33,461**

Table 1: Data set distribution

3.1 Data Format

The datasets we offer are available in two different formats: raw text annotations and tuples. Datasets, excluding Dev Set B, Test B and Test C, also come with corresponding dependency analysis results. For detailed information regarding these dependency analyses and to avoid redundancy, please refer to our prior work (Xu et al., 2023).

Figure 1 is an example of a CAMR text representation from the training set, detailed with sentence ID, word tokens, word ID, alignment of concept and relation, and the text annotation of CAMR. All files are encoded in UTF-8. The translation of the original sentence is "命*/command* 子封*/Zifeng* 帅*/lead* 车*/car* 二百*/two hundred* 乘*/unit* 以*/with* 伐*/attack* 京*/capital*," which means "*Command Zifeng to lead 200 chariots to attack the capital.*".

Table 2 is a copy of CAMR tuple representation including sentence ID (sid), source node ID (nid1), source concept (concept1), relation (rel), relation ID (rid), relation alignment word (ralign), target node ID (nid2), and target concept (concept2).

3.2 Ancient Chinese AMR

As the predecessor of CAMRP 2023, the evaluation task this year has introduced a certain amount of sentences of Ancient Chinese, selected from the segmented text of anicent Chinese classic *Zuo Zhuan* (a commentary on the *Spring* and *Autumn* annals of ancient China) processed by the school of Chinese

[1] https://www.ldc.upenn.edu/

第二十三届中国计算语言学大会论文集，第160页-第171页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

```
# ::id export_amr.37 ::2023-06-27 09:37:07
# ::snt 命 子封 帅 车 二百 乘 以 伐 京 。
# ::wid x1_命 x2_子封 x3_帅 x4_车 x5_二百 x6_乘 x7_以 x8_伐 x9_京 x10_。

(x1 / 命-01
    :arg1() (x14 / person
        :name() (x2 / name :op1 x2/子封 ))
    :arg2() (x3 / 帅-02
        :arg0() (x14 / person)
        :arg1() (x4 / 车
            :quant() (x5 / 二百)
            :cunit() (x6 / 乘))
        :arg2(x7/以) (x8 / 伐-01
            :arg0() (x14 / person)
            :arg1() (x9 / 京)))))
```

Figure 1: Sample of CAMR text representation

句子编号 sid	节点编号1 nid1	概念1 concept1	关系 rel	关系编号 rid	关系对齐词 ralign	节点编号2 nid2	概念2 concept2
37	x0	root	:top	-	-	x1	命-01
37	x1	命-01	:arg0	-	-	x13	person
37	x1	命-01	:arg1	-	-	x14	person
37	x1	命-01	:arg2	-	-	x3	帅-02
37	x14	person	:name	-	-	x2	子封
37	x3	帅-02	:arg1	-	-	x4	车
37	x3	帅-02	:arg2	x7	以	x8	伐-01
37	x4	车	:quant	-	-	x5	二百
37	x4	车	:cunit	-	-	x6	乘
37	x8	伐-01	:arg1	-	-	x9	京
37	x8	伐-01	:arg0	-	-	x14	person

Table 2: Sample of CAMR tuples

Language and Literature at Nanjing Normal University. During the annotation process, Yang Bojun's annotated version of annotations on *Zuo Zhuan* (Yang, 1990) was used as a reference. Among these, 500 sentences are for Dev Set B and 2,177 sentences are for Test Set C, focusing on assessing the performance of the parsing system on Ancient Chinese.

Chinese Mandarin has undergone changes in pronunciation, vocabulary, and grammar compared to Ancient Chinese. Therefore, the annotation of Ancient Chinese AMR adds and removes some semantic roles and modifies some predicate argument structures based on the Chinese AMR annotation framework. It also specifies annotation methods for special sentence patterns. Overall, the Ancient Chinese AMR maintains consistency with the Chinese AMR annotation format and includes alignment information for concepts and relations. As shown in Figure 1, the node "二百" (two hundred) is numbered `x5`, completing the concept alignment. The function word "以" (with) is numbered `x7` and is annotated along with the semantic role `:arg2` on the directed arc from node "`x3/`帅" (command) to node "`x8/`伐" (attack), completing the relation alignment.

To better describe the semantic structure of Ancient Chinese, first, the framework of Ancient Chinese AMR has added two concepts to the CAMR system to represent the causative and conative usages, namely "**make**" and "**consider**". As shown in Table 3, the phrase "而速之" in Chinese Mandarin would

第二十三届中国计算语言学大会论文集，第160页-第171页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

New concepts	Example sentence (snippet)	Ancient Chinese AMR
make-01	“...而速之...”	... :arg2(x10/而) (x29 / make-01 :arg1() (x11 / 速 ...
consider-01	”...小人耻失其君...“	... :op1() (x48 / consider-01 :arg0() (x8 / 小人 ...

Table 3: New concepts and examples in Ancient Chinese AMR

be directly translated into “*and fast it* (×)”, which doesn’t make sense at all because the word “速” here in Ancient Chinese often carries a causative meaning. Therefore, its actual meaning should be rendered as “*and make it fast* (✓)”. And similarly, the phrase “小人耻失其君” should be translated as “*the petty man considers losing his lord as a shame* (✓)” instead of “*the petty man shame losing his lord* (×)”.

Second, flexible usage of part of speech is a common linguistic phenomenon in Ancient Chinese. Based on this feature, the Ancient Chinese AMR primarily added annotation rules for nouns used as verbs and nouns used as adverbials. As shown in Figure 2, in the phrase “皆肘之”, the word “肘” (elbow) needs to have the action of striking with the arm supplemented in the actual annotation, with “击-01” (strike) being the root node of the sentence, thus restoring the true semantic expression.

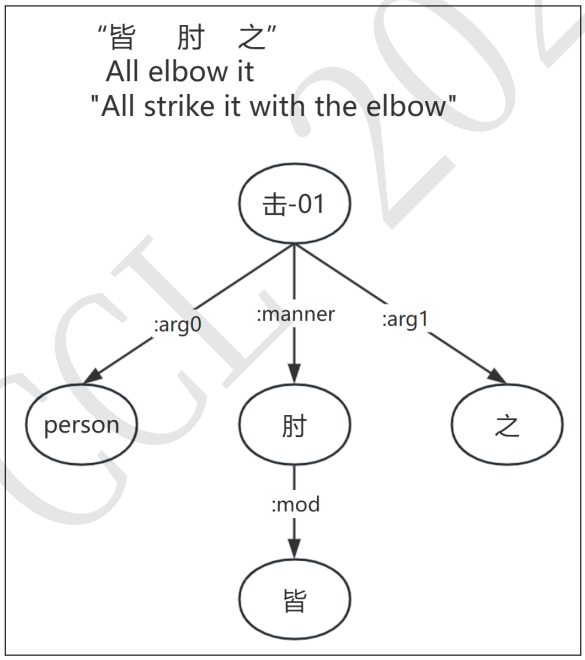


Figure 2: Example of noun used as verb in Ancient Chinese

4 Evaluation Design

In the spirit of innovation and comparison, CAMRP 2024 includes three evaluation metrics and two modalities.

4.1 Evaluation Metrics

- **Smatch** As the most widely-used evaluation metric to compute AMR parsing scores, SMATCH (Cai and Knight, 2013) uses a graph matching algorithm to return precision, recall and F-score. It focuses

第二十三届中国计算语言学大会论文集，第160页-第171页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

mainly on the overlapping of two AMR graphs, specifically on the nodes and edges. Considering that it was originally designed based on the English corpus and serves for English AMR parsing, SMATCH works well when it comes to monolingual AMR comparisons and yet failed to parsing the alignment information in Chinese AMR.

- **MRP** Due to its extensive compatibility, MRP (Oepen et al., 2020) has been utilized as the primary metric in both CoNLL 2019 and CoNLL 2020 for multi-lingual track. It employs a "node-to-node" search strategy to find the maximum match between two semantic graphs. However, for AMR or CAMR parsing evaluation, MRP typically yields higher scores compared to the other two metrics mentioned above, attributed to its relatively lenient scoring method.

- **Align-smatch** According to the changes in Chinese AMR, alignments of concept and relation were introduced into ALIGN-SMATCH (Xiao et al., 2022). Functions word in Chinese, unlike in English, convey a great deal of meaning, and therefore in ALIGN-SMATCH, they are well preserved and the deemed as the reflection between concept nodes. In short, Align-smatch inherits the basics of SMATCH and now can compute concept alignment and relation alignment in a more accurate way, which is indeed necessary in Chinese AMR.

Metric	Node		Edge		Node	
	Concept	Alignment Index	Semantic Role	Alignment Word	Concept	Alignment Index
MRP	帅-02	[5,5]	:arg2	-	伐-01	[16,16]
SMATCH	帅-02	-	:arg2	-	伐-01	-
ALIGN-SMATCH	帅-02	x3	:arg2	以	伐-01	x8

Table 4: Comparison of three metrics regarding alignments

For the three metrics mentioned above, they all share the same essence which is to convert a semantic graph into several sets of triples or tuples, namely "node-edge-node" (mostly). The key difference lies in how they treat alignment information. Table 4 details the comparison of three metrics regarding alignments (example snippet from Figure 1): SMATCH does not provide information for concept alignments or relation alignments, while MRP aligns concepts in a rather cumbersome way (token anchoring period). In contrast, ALIGN-SMATCH can handle both concept and relation alignments effectively. Hence, we consider it as the primary metric for CAMRP 2024, and metrics like MRP and SMATCH are used as references only and serve to reflect any fluctuations or progressions over the past few years.

4.2 Two Modalities

The evaluation task includes Open Modality and Closed Modality:

- **Closed Modality** Participants are required to use the designated training data, test data, and pre-trained model without substitutions. Additionally, we provide dependency analysis results of the training set for each team under Closed Modality. The pre-trained model, HIT_Roberta from Harbin Institute of Technology (Cui et al., 2021), is highly recommended.

- **Open Modality** Participants are permitted to use other pre-trained models and external resources, such as named entities and dependency analysis results, without limitations. However, all resources utilized must be detailed in the final technical report. Manual correction is prohibited in both modalities. Table 5 outlines the requirements for each modality.

5 Evaluation Results

CAMRP 2024 initiates on 1^{st} March, and data sets inclucding train set and dev set are authorized and released via LDC. Test sets are provided on 1st May via our GitHub repository[3]. Participants are to

[3]https://github. com/GoThereGit/Chinese-AMR

CCL 2024

Resources \ Modalities	*Closed*	*Open*
Algorithm	No Limit	No Limit
Pre-trained Model	HIT_Roberta	No Limit
External Resource	Dependency Tree	No Limit
Data Set	Train Set, Dev Set	No Limit
Manual Correction	Not Allowed	Not Allowed

Table 5: Requirements of two modalities

submit their technical report by 25^{th} May and Camera-ready by 25^{th} June. The evaluation task will be hosted as a workshop affiliated with the 23^{rd} China National Conference on Computational Linguistics (CCL 2024) from 26^{th} - 27^{th} July in Taiyuan, China.

5.1 Participants

There are 19 teams enrolled, and yet 3 teams completed the evaluation, resulting in a total of 15 submissions as shown in Table 6 along with other detailed information. All three teams chose the open modality exclusively. This contrasts with the previous year, where the majority of teams opted for the closed modality, with only a few choosing the open modality. Overdue submissions are marked with an asterisk and submissions with manual adjustment (not correction) are marked with a plus sign in Table 6. Each team is listed alphabetically here and throughout the paper.

Team	**Affiliation**	Test A		Test B		Test C	
		closed	*open*	*closed*	*open*	*closed*	*open*
BLCU	Beijing Language and Culture University	0	1^{+}	0	1^{+}	0	1^{+}
GDUFE	Guangdong University of Finance and Economics	0	2	0	2	0	2
HITSZ	Harbin Institute of Technology (Shenzhen)	0	2*	0	2*	0	2*
Total	15	0	5	0	5	0	5

Table 6: Participants information overview

5.2 Overall Results

Results from 3 teams encompassing a total of 15 runs exhibit an unexpected level of parsing performance across a broad spectrum. For the sake of better display and clearer comparison, we accordingly drew 3 tables (Table 7-9) to present all results of three test sets, in open modality and three metrics. *Precision*, *Recall* and *F-score* in each table are abbreviated as P, R and F_1, respectively. Note that Test B was the blind test at CAMRP 2022 and CAMRP 2023, and Test C is the new blind test. For the teams submitted more than two runs, we hereby list their best records. Hyphen "-" marks the team submitted one run only per track. The highest F-score in ALIGN-SMATCH metric per track is in bold font, which would account for a substantial part of final rankings.

In Test A, GDUFE's first run had the highest ALIGN-SMATCH F_1 score (0.8119), indicating superior alignment quality compared to other teams. They also performed well in SMATCH (0.7960) and MRP (0.8382). HITSZ had closely matched F_1 between their runs, with Run 2 slightly improving, demonstrating consistency. BLCU had submitted only one run and had relatively balanced scores.

In Test B, GDUFE again led in ALIGN-SMATCH F_1 (0.7527) and showed strong performance across all metrics. Their second run was slightly lower but consistent. HITSZ showed improvement from Run 1 to Run 2, with notable increases in SMATCH and MRP scores. And BLCU was getting closer than they were in the Test A.

In Test C, GDUFE still had the highest scores, with an ALIGN-SMATCH F_1 of 0.7156 and great results in SMATCH and MRP. Their performance was consistent across two runs. HITSZ showed decent results

(c) 2024 中国中文信息学会计算语言学专业委员会

Team	Run	ALIGN-SMATCH			SMATCH			MRP		
		P	R	F_1	P	R	F_1	P	R	F_1
BLCU	1	0.7893	0.7889	0.7891	0.7701	0.7654	0.7678	0.8176	0.8153	0.8165
	2	-	-	-	-	-	-	-	-	-
GDUFE	1	0.8087	0.8153	**0.8119**	0.7924	0.7996	0.7960	0.8348	0.8417	0.8382
	2	0.8059	0.8154	0.8107	0.7886	0.7997	0.7941	0.8320	0.8417	0.8368
HITSZ	1	0.8075	0.8094	0.8084	0.7913	0.7900	0.7906	0.8343	0.8326	0.8335
	2	0.8080	0.8111	0.8096	0.7927	0.7929	0.7928	0.8364	0.8338	0.8351

Table 7: Results of Test A in open modality

Team	Run	ALIGN-SMATCH			SMATCH			MRP		
		P	R	F_1	P	R	F_1	P	R	F_1
BLCU	1	0.7404	0.7427	0.7416	0.7381	0.7429	0.7405	0.7860	0.7801	0.7831
	2	-	-	-	-	-	-	-	-	-
GDUFE	1	0.7480	0.7575	**0.7527**	0.7462	0.7642	0.7551	0.7862	0.8032	0.7946
	2	0.7462	0.7566	0.7514	0.7438	0.7626	0.7531	0.7846	0.8021	0.7932
HITSZ	1	0.7459	0.7399	0.7429	0.7457	0.7416	0.7437	0.7836	0.7845	0.7841
	2	0.7513	0.7457	0.7485	0.7521	0.7484	0.7502	0.7888	0.7900	0.7894

Table 8: Results of Test B in open modality

with slight variations between runs, especially maintaining strong performance in MRP. BLCU then was a bit left behind and there is room for improvement.

Overall, GDUFE consistently outperformed the other teams across all tests and metrics. Their methods provided the best alignment and overall parsing quality, as reflected in the highest F_1 scores. This indicates their approach to semantic parsing and alignment is highly effective and reliable.

Results vary according to different metrics and test sets. While the training set and the majority of the development sets are in Mandarin Chinese, Test C, which contains 2,177 comparatively long and complex Ancient Chinese sentences, poses the greatest difficulty for scoring:

$$F_1{}^{testA} > F_1{}^{testB} > F_1{}^{testC}$$

The MRP metric, due to its relatively lenient scoring method, yields better results than the other two metrics. Counterintuitively, ALIGN-SMATCH does not have the lowest scores:

$$F_1{}^{mrp} > F_1{}^{align-smatch} > F_1{}^{smatch}$$

This can be attributed to the update of the concept alignment tuples in ALIGN-SMATCH, which generally score easily, resulting in higher scores compared to SMATCH. And what is worth mentioning is that GDUFE has scored a 0.8368 in MRP, which literally outperforms the SOTA at CoNLL 2020 by 3.3 percentage points[4] (Samuel and Straka, 2020).

We are to further discuss more technical details in the subsections below.

5.3 Models and Analysis

The BLCU team introduces a two-stage generative approach for CAMR parsing based on large language models (LLMs). Their method aims to address the complexity inherent in generating AMR graphs by decomposing the task into two stages: alignment-aware node generation and relationship-aware tuple generation. The first stage focuses on generating nodes by grouping them according to their alignment with words in the sentence, while the second stage involves generating relational tuples by grouping them

[4]CAMRP 2024 uses the same Test A as CoNLL 2020.

CCL 2024

Team	Run	ALIGN-SMATCH			SMATCH			MRP		
		P	R	F_1	P	R	F_1	P	R	F_1
BLCU	1	0.5687	0.5859	0.5772	0.5425	0.5683	0.5551	0.6419	0.6697	0.6555
	2	-	-	-	-	-	-	-	-	-
GDUFE	1	0.7062	0.7252	**0.7156**	0.6666	0.6869	0.6766	0.7423	0.7624	0.7522
	2	0.7051	0.7249	0.7149	0.6665	0.6878	0.6770	0.7403	0.7619	0.7510
HITSZ	1	0.6706	0.6677	0.6691	0.6501	0.6426	0.6463	0.7311	0.7280	0.7296
	2	0.6589	0.6728	0.6658	0.6380	0.6519	0.6449	0.7204	0.7351	0.7276

Table 9: Results of Test C in open modality

based on relationship types. To enhance the model's ability to handle zero-shot scenarios, particularly for ancient Chinese texts, the team employs a retrieval-enhanced instruction fine-tuning strategy, which integrates similar example sentences retrieved from the AMR corpus into the instructions.

The alignment-aware node generation process categorizes nodes into several types based on their alignment with sentence words, aiming to predict the nodes corresponding to the sentence in the AMR graph. This involves using a dictionary to store information about the words and nodes, represented as triplets. The relationship-aware tuple generation task then focuses on predicting relational tuples within the AMR graph, storing the tuples in a nested dictionary-list structure to represent the relationships. The retrieval-enhanced instruction fine-tuning strategy constructs data from the AMR corpus for node and tuple generation tasks. By incorporating high-similarity example sentences retrieved through vector retrieval, this strategy aims to improve the system's transferability and performance in parsing.

The HITSZ team leverages large language models and involves a two-step approach as well: systematic evaluation of current Chinese LLMs on the CAMR task, followed by a graph ensemble algorithm to integrate high-performing predictions. They have evaluated both commercial models, ChatGPT (Ouyang et al., 2022) and GPT-4, and open-source models, Baichuan-2[5], LLaMA-3[6], and LLaMA-3-Chinese[7] for their capabilities in few-shot CAMR parsing. Their findings indicated that while current LLMs possess some capacity for few-shot CAMR parsing, fine-tuning these models can significantly improve performance, often surpassing previous best systems. Additionally, the graph ensemble algorithm further enhanced the CAMR parsing performance by combining outputs from multiple high-performing models. The methodology involved a sequence-to-sequence approach for CAMR parsing, where the CAMR graph is linearized for training. The process included a thorough pre-processing step to ensure proper bracket completion, node correction, and handling of special relationships such as co-reference and alignment. The HITSZ team's results demonstrated that their system achieved decent scores on various test sets, highlighting the efficacy of combining LLMs fine-tuning with graph ensemble techniques.

The GDUFE team introduces a novel framework for CAMR parsing that utilizes a mixture of Low-Rank Adaption (LoRA) experts (Hu et al., 2021). Their system comprises a base CAMR parser fine-tuned from a large language model, supported by four sentence type experts and one ancient Chinese LoRA expert model. This framework is designed to leverage the strengths of specialized models for different sentence types, including declarative, interrogative, exclamatory, and imperative sentences, as well as ancient Chinese texts.

In their approach, the base CAMR parser is derived from fine-tuning the ChatGLM3-6B model (Du et al., 2021). The LoRA experts are trained on mixed data, combining training and silver data to enhance performance for specific sentence types. The system also includes a sentence classification module and a gating mechanism to activate the appropriate LoRA expert based on the input sentence type. This allows the framework to effectively handle the diverse linguistic structures found in different sentence types.

The methodology involves treating CAMR parsing as a sequence-to-sequence text generation task. The input sentences, annotated with concept node labels, are transformed into a linearized sequence

[5]https://github.com/baichuan-inc/Baichuan2
[6]https://github.com/meta-llama/llama3
[7]https://github.com/CrazyBoyM/llama3-Chinese-chat

representing the CAMR graph. The fine-tuning process ensures the model can generate accurate CAMR sequences from input text. The sentence classification and gating modules further refine the process by activating the relevant LoRA experts to handle specific linguistic features of the sentences.

Experimental results showed that the integration of LoRA experts significantly boosted the system's ability to parse different sentence types, with particularly impressive results in parsing ancient Chinese texts. The final system, combining outputs through a graph aggregation method, demonstrated superior performance compared to using the base parser alone or individual experts.

5.4 Fine-grained Metrics

In order to better explore the potential of each parsing systems and further promote the development of Chinese AMR parsing, we therefore set several fine-grained metrics. On the base of prior work (Damonte et al., 2017), CAMRP 2024 proposes 7 fine-grained metrics for Chinese AMR parsing, and Table 10 is provided with detailed explanations. **Neg.** computes on semantic roles with *:polarity*, and **Con.** focuses on concepts identification only. **NSF** makes Propbank frame identification without sense, ie, *want-01* / *want-00*. **Reent.** focuses on reentrant arcs or edges. The rest four are specially designed for Chinese AMR parsing. **Imp.** denotes those concept nodes usually ending with *Entity* or *Quantity*, for these concepts are newly asbtracted and generated, not original from the source sentence, namely implicit. **CA** and **RA** are for the precision of concept alignment tuples and relation alignment tuples.

Fine-grained metric		Evaluation object
Neg.	Negations	*:polarity* roles
Con.	Concepts	Concept indentification only
NSF	Non Sense Frames	Propbank frame identification without sense
Reent.	Reentrancies	Reentrant arcs only
Imp.	Implicit	Concepts with suffix such as *Entity*, *Quantity*
CA	Concept Alignment	Concept alignment tuples
RA	Relation Alignment	Relation alignment tuples

Table 10: Seven fine-grained metrics

Team \ Metric	**Neg.**	**Con.**	**NSF**	**Imp.**	**Reent.**	**CA**	**RA**
BLCU	0.7225	0.8637	0.8749	0.8212	0.6059	0.9024	0.4874
GDUFE	**0.7465**	0.8693	**0.8768**	0.8343	0.6378	0.9048	**0.5598**
HITSZ	0.7435	**0.8694**	0.8733	**0.8347**	**0.6458**	**0.9057**	0.5411

Table 11: Subscores of fine-grained metrics in open Test A

Team \ Metric	**Neg.**	**Con.**	**NSF**	**Imp.**	**Reent.**	**CA**	**RA**
BLCU	0.5955	0.8142	0.8139	0.7172	0.6396	**0.8493**	0.4855
GDUFE	**0.6058**	**0.8206**	**0.8191**	**0.7409**	0.6225	0.8426	**0.4910**
HITSZ	0.6015	0.8151	0.8167	0.7129	**0.6597**	0.8372	0.4845

Table 12: Subscores of fine-grained metrics in open Test B

Tables 11-13 display the performance of participants in each track, including three test sets. Generally, subscores in metrics such as **NSF** and **Con.** are notably higher than others. The **Neg.** metric shows variability in difficulty across different test sets. Nearly all subscores in **Reent.** failed to reach 0.65,

CCL 2024

Team \ Metric	Neg.	Con.	NSF	Imp.	Reent.	CA	RA
BLCU	0.3272	0.6218	0.5390	0.7043	0.2787	0.8475	0.1848
GDUFE	0.4829	**0.7747**	0.7303	0.8194	0.3015	**0.8796**	0.3119
HITSZ	**0.5313**	0.7052	**0.7516**	**0.8274**	**0.3149**	0.8788	**0.3449**

Table 13: Subscores of fine-grained metrics in open Test C

highlighting the complexity of the CAMR topology structure and the exceptionally challenging nature of the parsing task. The use of concept alignment annotation in Chinese AMR has clearly had a positive impact, with metrics related to concepts, such as **CA**, often exceeding 0.90. However, **RA** remains the lowest among all metrics, consistent with results from CAMRP 2023.

Notably, the HITSZ team achieved the highest subscore in **Reent.**, attributed to their focus on the graph structure in Chinese AMR. Their specialized approach to co-reference resolution enabled more accurate identification and representation of reentrancies within the AMR graphs.

From the comparison across the three test sets, Test C shows the greatest variability in scores among the teams, likely due to the ancient Chinese corpus used in this test set. This indicates that the transfer learning capability of large models still requires further improvement.

Overall, these fine-grained results illustrate the strengths and weaknesses of each team's approach across different linguistic challenges, highlighting areas such as **Reent.** and **RA** for future improvement, especially in handling ancient Chinese texts.

6 Conclusion and Future Work

This paper provides an overview of the fourth Chinese Abstract Meaning Representation parsing evaluation at CCL 2024. CAMRP 2024 continued with the ALIGN-SMATCH metric to better assess the parsing performance of each participating system. And it was the first time that Ancient Chinese AMR parsing has been introduced into our evaluation series. Three teams submitted their results, each presenting inspiring and motivating work. Some teams advanced prior methods with creative approaches, while others thoroughly explored the capabilities of LLMs. Notably, the GDUFE team achieved a score of 0.8382 in the MRP metric, surpassing the best record from CoNLL 2020 by 3.30 percentage points.

Decent progress has been made in both Mandarin and Ancient Chinese AMR parsing, marked by notable achievements and innovative methodologies. However, relation prediction and alignment remain challenging and act as bottlenecks in the development of Chinese AMR parsing. Despite remarkable breakthroughs in some aspects, leveraging the power of LLMs and maximizing their potential in transfer learning towards Ancient Chinese AMR parsing seem to be critical areas for further improvement. Also, the complexity of semantic relation identification and alignment within AMR structures necessitates focused attention and the development of innovative techniques.

In our future endeavors, we are dedicated to advancing Chinese AMR parsing through comprehensive initiatives. This involves hosting evaluation tasks to facilitate the assessment and benchmarking of parsing models. Additionally, we aim to construct and refine models specifically tailored to the complexities of both Mandarin and Ancient Chinese, thereby propelling the field of semantic analysis forward. By focusing on relation prediction and alignment, we seek to address current challenges and enhance the performance and understanding of Chinese AMR parsing. Through ongoing research, collaboration, and innovation, we aspire to develop robust and accurate parsing models, pushing the boundaries of semantic analysis further.

Acknowledgements

We would like to acknowledge the contributions of the members of the research team, Jinya Lu, Yuan Wen, Yihuan Liu, Peiyi Yan, Liming Xiao, Jin Chen and Pengxiu Lu. We thank them for annotating the corpus of Chinese AMR. And also we extend our appreciation to all the anonymous reviewers who pro-

(c) 2024 中国中文信息学会计算语言学专业委员会

vided thoughtful comments and feedback, helping to refine and strengthen this paper. This research was supported by National Language Commission Project (YB145-41) and National Social Science Foundation of China major project (21&ZD331, 22&ZD262).

References

L Abzianidze, Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajič, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, et al. 2020. Mrp 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Liang Chen, Bofei Gao, and Baobao Chang. 2022. A two-stage method for chinese amr parsing. *arXiv preprint arXiv:2209.14512*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Marco Damonte, Shay B Cohen, and Giorgio Satta. 2017. An incremental parser for abstract meaning representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with chinese amrs. In *Proceedings of the 10th Linguistic Annotation Workshop held in Conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, and Nianwen Xue. 2019. Building a chinese amr bank with concept and relation alignments. *Linguistic Issues in Language Technology*, 18.

Bin Li, Zhixing Xu, Liming Xiao, Junsheng Zhou, Weiguang Qu, and Nianwen Xue. 2023. The second chinese abstract meaning representation parsing evaluation. *Journal of Chinese Information Processing*, 37:33–43.

Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. 2020. Proceedings of the conll 2020 shared task: Cross-framework meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Hiroaki Ozaki, Gaku Morio, Yuta Koreeda, Terufumi Morishita, and Toshinori Miyoshi. 2020. Hitachi at mrp 2020: Text-to-graph-notation transducer. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 40–52.

David Samuel and Milan Straka. 2020. Úfal at mrp 2020: Permutation-invariant semantic parsing in perin. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64.

CCL-2024

Linfeng Song and Daniel Gildea. 2019. Sembleu: A robust metric for amr parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552.

Liming Xiao, Bin Li, Zhixing Xu, Kairui Huo, Minxuan Feng, Junsheng Zhou, and Weiguang Qu. 2022. Alignsmatch: A novel evaluation method for chinese abstract meaning representation parsing based on alignment of concept and relation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5938–5945.

Zhixing Xu, Yixuan Zhang, Bin Li, Junsheng Zhou, and Weiguang Qu. 2023. Overview of CCL23-eval task 2: The third chinese abstract meaning representation parsing evaluation. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 70–83, Harbin, China. Chinese Information Processing Society of China.

Bojun Yang. 1990. *Annotations of the commentary on the spring and autumn annals*. Zhonghua Book Company, Beijing.

CCL 2024

第二十三届中国计算语言学大会论文集，第160页-第171页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

System Report for CCL24-Eval Task 5: Multi-Model Classical Chinese Event Trigger Word Recognition Driven by Incremental Pre-training

Litao Lin[1], Mengcheng Wu[2], Xueying Shen[1], Jiaxin Zhou[1], Shiyan Ou[1,*],

School of Information Management, Nanjing University[1],

College of Information Management, Nanjing Agricultural University[2],

litaolin@smail.nju.edu.cn, wmc@stu.njau.edu.cn, sxy_77@smail.nju.edu.cn,

522023140121@smail.nju.edu.cn, oushiyan@nju.edu.cn

Abstract

This paper addresses the task of identifying and classifying historical event trigger words in Classical Chinese, utilizing both small-scale and large-scale language models. Specifically, we selected the small-scale language model GujiBERT for intelligent processing of classical texts, and the large-scale language model Xunzi-Qwen-14b. Both models underwent continued pre-training and fine-tuning, resulting in GujiBERT-CHED-mlm and Xunzi-Qwen-14b-CHED. For the small-scale language model, we used a BiLSTM as the feature extraction module and a CRF as the decoding module, employing a sequence labeling paradigm to complete the evaluation experiments. For the large-scale language model, we optimized the prompt templates and used a sequence-to-sequence paradigm for evaluation experiments. Our experiments revealed that GujiBERT-BiLSTM-CRF achieved the best performance across all tasks, ranking fourth in overall performance among all participating teams. The large-scale language model demonstrated good semantic understanding abilities, reaching a preliminary usable level. Future research should focus on enhancing its ability to produce standardized outputs.

1 Introduction

Ancient texts contain rich information on historical events, figures, geographical locations, and more, which are of significant value for historical research, cultural heritage preservation, and exploration of historical patterns. Events, as the smallest granular units describing historical knowledge, are critical for the detection and organization of information in ancient texts. This facilitates the enhancement of knowledge services for ancient literature from a digital humanities perspective. Due to the diverse content forms and complex language structures of ancient texts, information extraction typically relies on expert manual annotation and analysis, which is time-consuming and labor-intensive, making it difficult to scale for large volumes of ancient texts. In recent years, the rapid development of natural language processing (NLP), particularly the application of deep learning methods, has provided effective solutions for the automatic detection and identification of events in ancient texts.

The objective of this task is to evaluate and further improve algorithmic models for detecting historical events in ancient texts. Pre-training and fine-tuning are conducted on a prepared dataset of ancient

* Corresponding Author

©2024 China National Conference on Computational Linguistics

Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 172-184, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

historical texts to develop an optimal performance model. This model should possess two key capabilities: first, accurately identifying the corresponding trigger words and their locations, meaning it should precisely pinpoint the words that best represent the occurrence of a historical event in ancient texts and clearly mark their positions within the text; second, correctly determining the event type associated with each trigger word according to the CCL-CHED official classification system for ancient text events, ensuring the accuracy and consistency of detection results.

Based on an extensive survey of research on Chinese ancient text event information extraction, we employed both natural language understanding models and natural language generation models to complete this evaluation task. Specifically, trigger word recognition and classification were treated as a sequence labeling task, with a domain-specific BERT model selected for continued pre-training and fine-tuning to construct the trigger word recognition and classification model. Additionally, to explore the applicability of large language models to this task, a base model tailored for intelligent processing of ancient Chinese texts was selected for continued pre-training and fine-tuning. Comparative testing was conducted to select the optimal prompts, enabling the large language model to generate character indices and trigger word categories for the given text. Figure 1 illustrates the overall technical approach employed in this study. Through a series of experiments, we discovered that smaller language models, exemplified by BERT, outperform in terms of prediction accuracy and the standardization of output. Although large language models have advantages in terms of ease of use, the standardization of their output is challenging to control, necessitating further manual processing. Overall, large language models are less suitable for tasks that require high precision in information extraction.

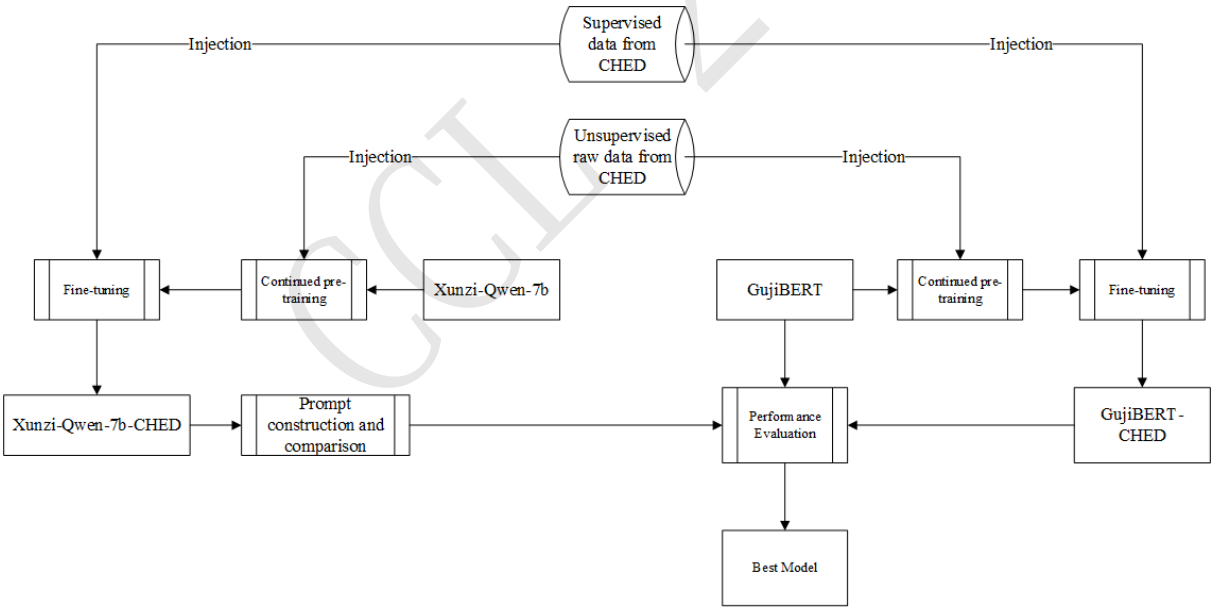


Figure 1: Technical Framework

2 Related Work

Currently, there are three main methods for event extraction from ancient Chinese texts: pattern-matching-based event extraction, statistical machine learning-based event extraction, and deep learning-

based event extraction. The pattern-matching method relies on expert-specified rules to parse sentences, identify trigger words, and categorize event types. The machine learning-based method converts the text into feature vectors and uses classifiers (such as support vector machines and maximum entropy models) for event identification and classification. For instance, Zhangchao Li et al.(2020) studied war events in the “Zuo Zhuan” (Chronicles of Zuo), exploring a combination of rules and machine learning. They used pattern matching to identify war sentences and then employed Conditional Random Fields (CRF) combined with feature models to identify and extract elements of war events. However, traditional statistical model-based methods rely heavily on manual feature engineering, making it difficult to fully extract deep semantic information from the text. With the significant advancements in natural language processing achieved through deep learning, constructing neural network models (such as CNN, RNN, LSTM, etc.) has enabled the automatic learning and extraction of textual features, substantially improving the accuracy and generalization capability of event extraction. Deep learning-based methods include various technical paradigms such as “sequence labeling”, “reading comprehension”, and “sequence-to-sequence”. Additionally, highly domain-specific pre-trained models are used to enhance the learning effectiveness on the text. Representative research based on the sequence labeling paradigm includes a study by Ma et al.(2021), they explored the multi-class automatic classification of canonical trigger words and event sentences using a Bi-LSTM model, achieving an accuracy of 95%. In their research focusing on “Records of the Grand Historian”, Zhongbao Liu et al.(2020) utilized a BERT-LSTM-CRF model to extract historical events from the corpus, achieving an F1 score of 82.3%. In a study by Xuehan Yu et al.(2021) focusing on war sentences from the “Zuo Zhuan” corpus, they compared the performance of five models: GuwenBERT-LSTM, BERT-LSTM, RoBERTa-LSTM, BERT-CRF, and RoBERTa-CRF, for event extraction. Among these models, RoBERTa-CRF achieved an F1 score of 82.1%. research based on the reading comprehension paradigm, such as that by Yu Xuehan et al.(2023), they integrated machine reading comprehension (MRC) into the neural network architecture to achieve event extraction by setting questions. Among these, RoBERTa-MRC_AC and RoBERTa-MRC_MC performed well in event type extraction, achieving F1 scores of 88.2% and 89.2%, respectively, significantly improving the performance. The sequence-to-sequence paradigm involves inputting the sentence to be processed into a generative model, which directly outputs the final result without the need for further manual processing. Representative research includes the study by Zhang et al.(2023) who proposed a generative approach using a knowledge graph-based event generation framework to extract war events from “Records of the Grand Historian”. The trigger word recognition achieved an effectiveness of 71.3%. Wang Y. et al.(2023) used “Records of the Three Kingdoms” as their research corpus and compared sequence labeling models with generative models. They explored the effectiveness of the BERT-BiLSTM-CRF fine-tuned BBCN-SG sequence model and the Stacking-TRN-SG model constructed by integrating three models—T5-SG, RoBERTa-SG, and NEZHA-SG—fine-tuned based on the T5 model, using stacking. The Stacking-TRN-SG model achieved a recall rate of 70.35% in event extraction.

Overall, pattern matching can achieve good results for specific types of texts, but its transferability is poor. Statistical machine learning methods heavily depend on the richness of training samples and perform poorly in parsing sentences that did not appear in the training corpus. Currently, deep learning methods have become mainstream. Among these, models based on the BERT architecture have achieved

CCL 2024

good results under the sequence labeling paradigm. With the rise of transfer learning techniques, methods based on reading comprehension and generative models have also been applied to event information extraction from ancient texts, providing new perspectives for event extraction research. In addition to changing task paradigms and designing, comparing, and analyzing different architectures of deep learning models, many scholars also aim to provide the most basic and foundational optimizations for various natural language processing tasks by constructing domain-specific pre-trained language models (Dongbo Wang et al., 2022). By the end of 2022, the emergence of ChatGPT once again revolutionized the paradigm of natural language processing tasks. Subsequently, large language models continued to innovate, leading to the creation of models tailored for different vertical domains. This includes the "Xunzi" series of models designed for intelligent processing of Classical Chinese. However, as of now, there have been no studies observed that apply large language models to the extraction of trigger words in ancient texts.

Given the extensive research showing that domain-specific continued pre-training of deep pre-trained language models can effectively enhance the model's performance on downstream tasks (Gururangan et al., 2020). Related studies also indicate that sequence labeling models still maintain certain advantages in specific task scenarios. Furthermore, according to Wei et al. (2023), it is evident that within the sequence labeling task paradigm, the BERT-BiLSTM-CRF model achieves the best performance. Considering the current lack of relevant practices involving large language models in the extraction of trigger words from ancient texts, this paper proposes constructing incremental pre-training for BERT specifically for the evaluation task. It combines this with BERT-BiLSTM-CRF to enhance the effectiveness of event information extraction. Additionally, attempts are made to incrementally pre-train and fine-tune large language models, exploring their applicability in extracting trigger words from ancient texts.

3 Domain-adaptive continued pre-training of the model

3.1 Continuing Pre-training Data

This paper uses the contents of the "text" field from the training set provided by CCL-CHED as the continuing pre-training data for all models. Domain-adaptive pre-training is a form of self-supervised learning that only requires domain-specific text input to the model without any manual annotation. In this experiment, using a self-developed Python program, the entire contents of the "text" field from both the train.jsonl and validate.jsonl files in the CHED trigger word recognition task dataset were extracted as training and validation corpora. A total of 5650 valid training sentences were obtained, with an average sentence length of 21.67 characters. The minimum sentence length was 3 characters, and the maximum sentence length was 200 characters. The cumulative distribution of sentences with different numbers of characters is shown in Figure 2.

3.2 Selection and Continuing Pre-training of BERT Model

3.2.1 Selection of Base Model

The SikuBERT series models constructed based on the "Complete Collection of Four Repositories" corpus have laid the groundwork for intelligent processing of Chinese classical texts (Dongbo

CCL 2024

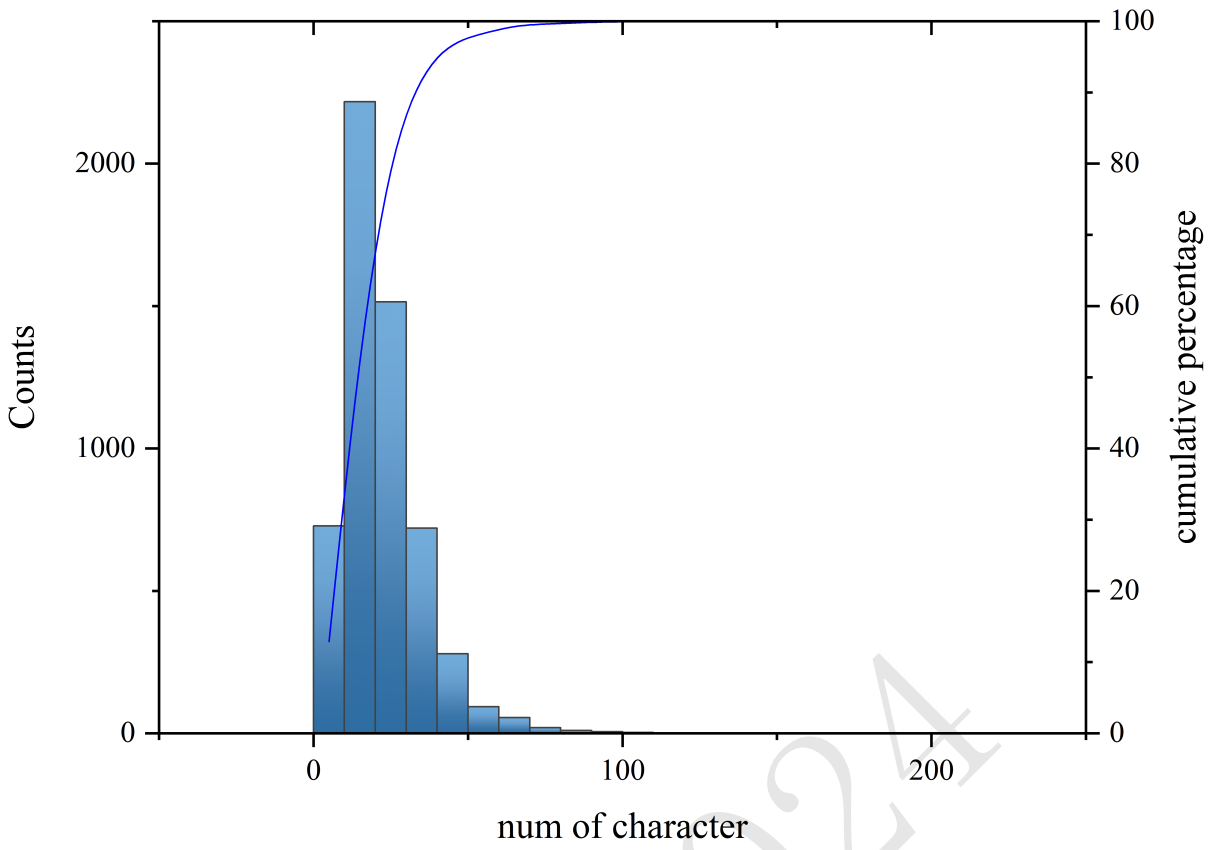


Figure 2: Cumulative distribution plot of the number of sentences with different numbers of characters in the continuing pre-training corpus

Wang et al., 2022). Subsequently, Professor Dongbo Wang's team from Nanjing Agricultural University further utilized approximately 170 million characters of ancient Chinese corpus from the Dazhige website to further pre-train the SikuBERT series models, leading to the development of the GujiBERT and GujiRoberta series models (D. Wang et al., 2023). According to the experimental data in the literature (D. Wang et al., 2023), it is evident that GujiBERT demonstrates relatively superior performance. Based on this, this paper selects GujiBERT as the baseline model for further pre-training.

3.2.2 Continued Pre-training Approach

This paper adopts Masked Language Modeling (MLM) (Devlin et al., 2019) as the task objective during the pre-training stage. MLM involves masking certain characters in the input sentence and requiring the model to predict these characters to learn bidirectional contextual relationships in language. During pre-training, BERT randomly masks 15% of the vocabulary in the input sequence. The masked words are replaced with a special [MASK] token. The objective of pre-training is for the model to predict the masked words based on context. In this way, the model needs to use information from other words in the sentence to infer the masked words, thereby learning deep semantic relationships and contextual dependencies between words. For example, for the sentence "遣内史王谊监六军，攻晋州城。" (Sending the Interior Minister Wang Yi to supervise the six armies to attack Jinzhou city.), after preprocessing by the model, it may become "遣内史王谊监六军，[MASK]晋州城。" (Sending the

Interior Minister Wang Yi to supervise the six armies, [MASK] Jinzhou city.), and the objective of MLM is to predict the character at the “[MASK]” position.

3.2.3 Training Process

During the training process, the preprocessed training data is inputted into the GujiBERT base model at the sentence level for continued training. In the forward pass, the predicted values of each masked word are computed, and the model parameters are adjusted through the calculation of the loss function and backpropagation, enabling the model to learn deep semantic relationships within the corpus. After multiple rounds of tuning and comparison, the final experimental hyperparameters are determined, as shown in Table 1. The model obtained from continued pre-training based on GujiBERT in this paper is named GujiBERT-CHED-mlm.

Experimental parameters	Value
Batch_size	8
Epochs	10
Max_sequence_length	512
Learning_rate	5e-5
optimizer	AdamW
Line_by_line	True

Table 1: Key Hyperparameters for Continued Pre-training for BERT model

3.3 Selection and Continued Pre-training of Large Language Models

3.3.1 Selection of Base Model

The Xunzi language model is a series of large language models tailored for intelligent processing of Classical Chinese texts. It is built upon models such as Qwen, Baichuan, and GLM, and has undergone extensive fine-tuning using a large amount of supervised ancient Chinese corpus data. This model series is capable of comprehensive tasks such as intelligent text indexing, information extraction, poetry generation, translation between classical and modern Chinese, reading comprehension, morphological analysis, and punctuation annotation[1] . In this paper, the Xunzi-Qwen-14b model[2] from this series is selected for continued pre-training and fine-tuning, aiming to construct a large language model suitable for the CCL2024-CHED task.

3.3.2 Continued Pre-training Approach

The Xunzi-Qwen-14b model is further pre-trained using Low-Rank Adaptation (LoRA) (Hu et al., 2021). LoRA fine-tuning is an efficient fine-tuning technique for large-scale language models. It adapts and updates specific parts of the model by introducing low-rank matrices, enabling fine-tuning for specific tasks while maintaining training efficiency and economical memory usage.

[1]https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM

[2]https://www.modelscope.cn/models/Xunzillm4cc/Xunzi-Qwen1.5-14B/summary

CCL 2024

3.3.3 Training Process

To further explore the effectiveness of large language models in this task, the study continues pre-training the Xunzi-Qwen-14b classical Chinese language model using the LLama-factory framework. This aims to enhance the efficiency and performance of event recognition in this task. During the training process, the corpus needs to be preprocessed into a specific format. Training corpus examples are shown in Table 2.

Xunzi-Qwen-14b Continued Pre-training Data Format

```
[
 {"text":"已酉，命习水战于新池。"}
 {"text":"九月，占城国王释利因德缓使蕃词散来。"}
]
```

Table 2: Xunzi-Qwen-14b Continued Pre-training Data Format

The hyperparameters set during the pre-training process of Xunzi-Qwen-14b are shown in Table 3. The model obtained after continuing pre-training on Xunzi-Qwen-14b is referred to as Xunzi-Qwen-14b-CHED in this paper. The perplexity of this model on the validation corpus is 21.99.

Experimental Parameters	Value
stage	pt
finetuning_type	lora
cutoff_len	1024
train_batch_size	1
learning_rate	5e-5
num_train_epochs	3

Table 3: Key Hyperparameters for Continued Pre-training for Xunzi-Qwen-14b

4 Fine-tuning Training and Evaluation of the Model

4.1 Fine-tuning of the BERT Model

4.1.1 Preprocessing of Fine-tuning Training Data

This paper utilizes the training and validation sets provided by CCL-CHED in the three task files for fine-tuning training and performance evaluation. Specifically, according to the event labeling patterns of different tasks, the corpora in each task dataset are converted to the input format required by the sequence labeling model. The B, I, E, S, O role tagging strategy is adopted, where B represents the starting character of a multi-syllable trigger word, E represents the ending character of a multi-syllable trigger word, I represents the remaining characters of a multi-syllable trigger word that are neither at the

CCL 2024

start nor the end, S represents a single-syllable trigger word, and O represents characters in the sentence that do not form trigger words. Based on the B, I, E, S, O role tagging strategy, the number of characters with different role types for each task is shown in Table 4.

Task Name	**Number of Role Tags**
Trigger Word Recognition	5*1+1
Coarse Event Type Identification	5*9+1
Fine-Grained Event Type Identification	5*67+1

Table 4: Number of Character Role Tags for Different Tasks

4.1.2 Experimental Setup

The BERT-BiLSTM-CRF model is built using the PyTorch framework. The pre-trained weights for GujiBERT[3] are obtained from the Hugging Face official website. The computer system used for fine-tuning training of BERT-BiLSTM-CRF operates on Red Hat 4.8.5-44, with a CPU model of 48 Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, and a GPU model of Tesla P100. As shown in Figure 2, the longest sentence in the training corpus contains 200 characters, which is below the maximum sentence length that BERT can handle (512 characters). Therefore, the maximum sentence length for the model is set to the maximum sentence length in each batch of data. The key hyperparameters for fine-tuning training of BERT-BiLSTM-CRF are shown in Table 5.

Experimental Parameters	**Value**
Max sequence length	Pad to length
Batch size	32
Epoch	50
Learning rate	1e-5
LSTM hidden dim	256

Table 5: Key Hyperparameters for fine-tuning

In this experiment, after each training epoch, the model's precision (P), recall (R), and F1 score are computed on the validation set. Only when the F1 score of the new model is the highest, the weights of the model obtained in the current epoch are retained. Thus, in this experiment, the F1 score is used as the criterion for selecting the best model, rather than the lowest loss. The model with the lowest F1 score on the test set is considered the optimal model.

4.2 Fine-tuning of Large Language Models

4.2.1 Instructional Fine-tuning Template

Large language models are subject to various uncontrollable factors, particularly the design of prompts, which have a significant impact on the output results. Faced with various tasks, this study

[3]https://huggingface.co/hsc748NLP/GujiGPT_fan.

CCL 2024

conducted multiple pre-experiments using different prompts. Finally, the prompts listed in Table 6 were determined for instructional fine-tuning and prediction.

Trigger Word Recognition	[{ "instruction": "将下面text古文中触发词和标签以及触发词的位置信息进行自动识别和抽取", "input": "sen_id: 5761, text: 己酉，命习水战于新池。", "output": "{\"sen_id\": 5761, \"text\": \"己酉，命习水战于新池。\", \"events\": [{\"trigger\": \"命\", \"start_offset\": 3, \"end_offset\": 4}]}" }]
Coarse Event Type Identification	[{ "instruction": "将下面text古文中触发词和标签以及触发词的位置信息进行自动识别和抽取", "input": "sen_id: 5761, text: 己酉，命习水战于新池。", "output": "{\"sen_id\": 5761, \"text\": \"己酉，命习水战于新池。\", \"events\": [{\"trigger\": \"命\", \"label\": \"交流\", \"start_offset\": 3, \"end_offset\": 4}]}" }]
Fine-Grained Event Type Identification	[{ "instruction": "将下面text古文中触发词和标签以及触发词的位置信息进行自动识别和抽取", "input": "sen_id: 5761, text: 己酉，命习水战于新池。", "output": "{\"sen_id\": 5761, \"text\": \"己酉，命习水战于新池。\", \"events\": [{\"trigger\": \"命\", \"label\": \"交流-个人交流-诏令-命令\", \"start_offset\": 3, \"end_offset\": 4}]}" }]

Table 6: Example Templates for Task-Specific Prompts

4.2.2 Experimental Setup

The experimental graphical display used is the RTX 8000, and the hyperparameters for instructional fine-tuning are shown in Table 7.

Experimental parameters	Value
stage	stf
finetuning_type	lora
cutoff_len	512
Learning_rate	5e-5
train_batch_size	2
learning_rate	5e-5
num_train_epochs	5

Table 7: Key Hyperparameters for Instructional Fine-tuning of Xunzi-Qwen-14b-CHED

5 Experimental Results and Analysis

5.1 Experimental Results

The experiment uses P (Precision), R (Recall), and F1 as evaluation metrics, with the final results being based on the Micro P, Micro R, and Micro F1 of each predicted label (Kelly, 2009) as the ultimate basis for assessing model performance. For the output results of large language models, this paper converts the outputs into character-level role sequences based on the B, I, E, S, O role labeling strategy, and then calculates Micro P, Micro R, and Micro F1(Kelly, 2009), thereby ensuring the comparability of evaluation results across different models.

5.1.1 Trigger Word Recognition

According to Table 8, the GujiBERT model, which did not utilize unsupervised data from CHED for continued pre-training, performed the best. The performance of large language models was the worst. Upon examining the output results of large language models, it was found that in the test samples with sen_id 8052 and 939, the character interval index of the last trigger word given by the large language models exceeded the sentence length range of the original sentence. Further inspection of the trigger word characters extracted by the large language models revealed that they matched the answers in the validation set. This indicates that large language models possess good semantic understanding capabilities, but their ability to produce standardized output needs improvement. Additionally, BERT based model achieved the maximum F1 score at training epoch 14.

Model Name	Micro P	Micro R	Micro F1
GujiBERT-BiLSTM-CRF	0.799	0.802	0.800
GujiBERT-CHED-mlm-BiLSTM-CRF	0.752	0.767	0.759
Xunzi-Qwen-14b-CHED	0.676	0.665	0.671

Table 8: Trigger Word Recognition Results of Various Models

5.1.2 Coarse Event Type Identification

According to Table 9, the GujiBERT model, which did not use unsupervised data from CHED for continued pre-training, performed the best, while the performance of large language models was

CCL 2024

the worst. By examining the output results of the large language models, it was found that in the test sample with sen_id 917, the character interval index of the last trigger word given by the large language models exceeded the sentence length range of the original sentence. Further inspection of the trigger word characters identified by the large language models revealed that they matched the answers in the validation set. Additionally, BERT based model achieved the maximum F1 score at training epoch 67.

Model Name	Micro P	Micro R	Micro F1
GujiBERT-BiLSTM-CRF	0.837	0.820	0.828
GujiBERT-CHED-mlm-BiLSTM-CRF	0.832	0.779	0.805
Xunzi-Qwen-14b-CHED	0.653	0.650	0.651

Table 9: Trigger Word Recognition and Classification Results at the Coarse Granularity

5.1.3 Fine-Grained Event Type Identification

According to Table 10, the GujiBERT model, which did not use unsupervised data from CHED for continued pre-training, showed the best performance, while the performance of large language models was the worst. Upon examining the output results of the large language models in the fine-grained event extraction task, no issues with output format irregularities were found. Additionally, BERT based model achieved the maximum F1 score at training epoch 25.

Model Name	Micro P	Micro R	Micro F1
GujiBERT-BiLSTM-CRF	0.782	0.762	0.772
GujiBERT-CHED-mlm-BiLSTM-CRF	0.761	0.686	0.722
Xunzi-Qwen-14b-CHED	0.631	0.617	0.624

Table 10: Trigger Word Recognition and Classification Results at the Fine-Grained Granularity

5.2 Results Analysis and Discussion

The experiments found that using GujiBERT as the character encoder followed by a BiLSTM-CRF, the GujiBERT-BiLSTM-CRF model achieved the best performance in all tasks. We speculate that this may be due to the relatively small number of event trigger words in classical Chinese texts and their strong grammatical regularity, which explains why past pattern-matching methods have achieved good results. Additionally, the tasks of trigger word identification and classification are straightforward, not involving the recognition and classification of event arguments and their relationships. Therefore, context-aware sequence labeling models are better able to capture the position and type characteristics of event trigger words. On the other hand, the advantage of large language models lies in their understanding of the overall semantics of sentences, making them well-suited for complex tasks. However, for tasks with simple objectives but strict formatting requirements, these models do not hold as much of an advantage.The experiments also indicate that if the amount of continued pre-training data is too small, the model's performance may not improve and could even decline. This may be due to the insufficient amount of data used for fine-tuning, which is not enough to significantly optimize the large language model. Instead, it might introduce disturbances to the original model, resulting in unstable performance.

CCL 2024

In the experiments, we also found that fewer epochs are needed to train fine-grained event trigger word recognition and classification models compared to training coarse-grained models. Furthermore, the accuracy of fine-grained event trigger word recognition and classification models was even higher than that of the trigger word recognition models. This suggests that for BERT-based sequence labeling models, the number of training epochs required to achieve the maximum F1 score is not positively correlated with the number of role labels. Additionally, more detailed event classification granularity seems to help the model better learn the semantic differences between different characters, thus improving the event extraction task performance.

6 Conclusion and future work

This paper constructs various models for event trigger word extraction and classification through domain-adaptive pre-training and fine-tuning of the small language model GujiBERT and the large language model Xunzi-Qwen-14b. Performance comparisons of different models were conducted through experimental testing. From our experiments, we conclude that small deep pre-trained language models like BERT, which follow a pre-training-fine-tuning approach, are more suitable for fine-grained information extraction and classification tasks.

There is still room for optimization in the sequence labeling paradigm used in this paper: we adopted a detailed role sequence labeling strategy of B, I, E, S, O in our experiments, but trigger words usually consist of only 1 to 2 characters. Thus, some role labels may not be used, which increases the computational load of the CRF module and potentially the classification difficulty. Future research could consider using the simplified B, I, O labeling strategy. Our experiments also show that large language models based on the pre-training-prompt paradigm have superior semantic understanding capabilities but cannot yet generate directly usable results when applied to information extraction tasks. The main challenges in applying large language models lie in the design of prompt templates and further processing of the output content.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.),*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, & Bin Li. (2022). Construction and Application of Pre-trained Models of Siku Quanshu in Orientation to Digital Humanities. *Library Tribune, 42*(6), 31–43.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks.*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users.*Foundations and Trends in Information Retrieval, 3*(1—2), 1–224. https://doi.org/10.1561/1500000012

Ma, X., He, L., Liu, J., Li, Z., & Gao, D. (2021). A Construction Method of the Classification System Oriented to Content Analysis of Ancient Books. *Journal of Library and Information Science in Agriculture, 33*(9), 27–36. https://doi.org/10.13998/j.cnki.issn1002-1248.21-0262

Wang, D., Liu, C., Zhao, Z., Shen, S., Liu, L., Li, B., Hu, H., Wu, M., Lin, L., Zhao, X., & Wang, X. (2023).*GujiBERT and GujiGPT: Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts* (arXiv:2307.05354). arXiv. https://doi.org/10.48550/arXiv.2307.05354

Wang Y., Wang H., Zhu H., & Li X. (2023). Research on the Construction of an Event Recognition Model for Historical Antique Books Based on Text Generation Technology.*Library and Information Service, 67*(3), 119–130. https://doi.org/10.13266/j.issn.0252-3116.2023.03.011

Wei, C., Feng, Z., Huang, S., Li, W., & Shao, Y. (2023). CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection. In M. Sun, B. Qin, X. Qiu, J. Jing, X. Han, G. Rao, & Y. Chen (Eds.), *Chinese Computational Linguistics* (pp. 289–305). Springer Nature. https://doi.org/10.1007/978-981-99-6207-5_18

Xuehan Yu, Lin He, & Jian Xu. (2021). Extracting Events from Ancient Books Based on RoBERTa-CRF. *Data Analysis and Knowledge Discovery, 5*(7), 26–35.

Yu Xuehan, He Lin, & Wang Xianqi. (2023). Research on Event Extraction from Ancient Books Based on Machine Reading Comprehension. *Journal of The China Society For Scientific And Technical Information, 42*(3), 316–326.

Zhang, J., Wei, Y., Zhu, Y., & Wu, B. (2023). Self-adaptive Prompt-tuning for Event Extraction in Ancient Chinese Literature. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191495

Zhangchao Li, Zhongkai Li, & Lin He. (2020). Study on the Extraction Method of War Events in Zuo Zhuan. *Library and Information Service, 64*(7), 20–29. https://doi.org/10.13266/j.issn.0252-3116.2020.07.003

Zhongbao Liu, Jianfei Dang, & Zhijian Zhang. (2020). Research on Automatic Extraction of Historical Events and Construction of Event Graph Based on Historical Records. *Library and Information Service, 64*(11), 116–124. https://doi.org/10.13266/j.issn.0252-3116.2020.11.013

CCL 2024

CCL24-Eval 任务5系统报告：基于增量预训练与外部知识的古文历史事件检测

康文军[1,3]　左家莉[1,2]　胡益裕[2]　王明文[1,2]
[1]江西师范大学　计算机信息工程学院　江西　南昌　330022
[2]江西师范大学　数字产业学院　江西　上饶　334000
[3]南昌理工学院　计算机信息工程学院　江西　南昌　330044
Email: wenjun_kang@qq.com,zjl@jxnu.edu.cn,yiyu_hu@qq.com,mwwang@jxnu.edu.cn

摘要

古文历史事件检测任务旨在识别文本中的事件触发词和类型。为了解决传统pipeline方法容易产生级联错误传播，以及大多数事件检测方法仅依赖句子层面信息的问题，本文提出了一种结合外部信息和全局对应矩阵的联合抽取模型EIGC，以实现触发词和事件类型的精确抽取。此外，本文还整理了一个包含“二十四史”等古汉语文献的数据集，共计约97万条古汉语文本，并利用该文本对BERT-Ancient-Chinese进行增量预训练。最终，本文所提出的模型在三个任务上的总F1值达到了76.2%，验证了该方法的有效性。

关键词： 古文历史事件检测 ； 增量预训练 ； 外部知识

System Report for CCL24-Eval Task 5: Historical Event Detection in Ancient Texts Based on Incremental Pre-training and External Knowledge

Wenjun Kang[1,3]　**Jiali Zuo**[1,2]　**YiYu Hu**[2]　**Mingwen Wang**[1,2]
[1]School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022,China
[2]School of Digital Industry, Jiangxi Normal University, Shangrao, Jiangxi 334000, China
[3]School of Computer and Information Engineering, Nanchang Institute of Technology, Nanchang, Jiangxi 330044, China
Email: wenjun_kang@qq.com,zjl@jxnu.edu.cn,yiyu_hu@qq.com,mwwang@jxnu.edu.cn

Abstract

The task of historical event detection in ancient texts is to identify event trigger words and types in the text. To solve the problem that traditional pipeline methods are prone to cascading error propagation and most event detection methods only rely on sentence-level information, we proposed a joint extraction model, EIGC, which combines external information and global correspondence matrix to achieve accurate extraction of trigger words and event types. In addition, we also compiled a dataset containing ancient Chinese literature, such as the Twenty-four Histories, with a total of about 970,000 ancient Chinese texts, and used the texts to incrementally pre-train BERT-Ancient-Chinese. Finally, the overall F1 value of the proposed model reaches 76.2% on the three tasks, which proves the effectiveness of the method.

Keywords: historical events test in ancient languages , incremental pre-training , external knowledge

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目： 国家自然科学基金(61866018,62266023)
通讯作者： 左家莉

1 引言

事件抽取（Event Extraction）是信息抽取的重要任务之一，旨在从自然语言文本中识别和提取相关事件信息的过程，它在知识图谱构建、阅读理解、问答系统等多个领域有着广泛的应用。具体而言，事件抽取可以被分为两部分：即事件检测任务（触发词识别、事件类型分类）和事件论元抽取任务（事件论元识别和事件论元角色分类）。其中，事件检测任务主要识别事件触发词并判断其所属的事件类型，事件论元抽取则是识别出事件的参与者（如行为主体、对象、时间、地点等）以及它们之间的关系(吴旭等人, 2023)。

古汉语作为中华文明的重要载体，其文献记载了丰富的历史和文化信息。然而，由于古文句法、语义复杂，使用范围小，针对古汉语的信息抽取任务仍然面临着较大挑战。本文基于古文历史事件类型抽取评测任务提供的数据集，对古文历史事件检测任务进行研究。针对传统pipeline方法容易产生级联错误传播，以及大多数事件检测方法仅依赖句子层面信息的问题，本文设计了一个结合外部信息和全局对应矩阵的古文历史事件触发词和类型联合抽取模型（**E**xternal **I**nformation and **G**lobal **C**orresponding Matrix Based Joint Trigger Words and Types of Ancient Historical Events，EIGC），用于古文历史事件检测任务研究。

2 相关工作

经典的事件抽取方法大致包括基于模式匹配和基于机器学习。然而，基于模式匹配需要领域专家设计模板，从而使得该方法的泛化性较低，基于机器学习则涉及复杂的特征工程。随着深度学习技术的不断进步，基于深度学习的方法已经成为当前事件抽取研究的主流方法(王浩畅等人, 2023)。Chen等人(2015)提出一个动态多池化的卷积神经网络（DMCNN），以捕获句子级别的信息，并且动态地保留多个重要信息。Feng等人(2016)提出了一种结合LSTM和CNN的混合模型，该模型通过从特定上下文中提取序列和块级特征来增强语义信息，并利用这些特征来训练多分类器，从而提高了事件检测的性能。

目前，基于深度学习的事件抽取方法可分为pipeline式抽取和联合抽取。在事件检测任务中，pipeline式抽取首先识别出事件的触发词，随后依据这些触发词判断其相应的事件类型。然而，这种方法容易导致错误的级联传播。联合抽取则将这两个子任务建模为一个统一的联合学习框架，这种方法能够充分利用触发词与事件类型之间的潜在联系，进而促进触发词和事件类型抽取效果的相互提升(李华昱等人, 2022)。Chen等人(2012)提出了一种联合抽取的方法，将事件抽取任务分解为两个联合抽取任务，并利用丰富的语言学特征来提取中文事件。贺瑞芳等人(2019)构建了一个基于CRF的多任务学习模型，用于中文事件的联合抽取，有效缓解了分类训练后的语料稀疏问题。Nguyen等人(2012)提出一种基于深度学习中共享的隐藏层表示的联合模型，用以预测实体类型、触发词和事件论元角色，显著提高了中文事件抽取的性能。本文采用全局对应矩阵的方法，分别设计了触发词头尾对齐矩阵和触发词与事件类别对齐矩阵，以实现触发词与事件类型的联合抽取。

大多数现有的事件检测方法主要依赖于句子层面的信息，但在许多情况下，句子层面的信息并不足以提供充分的信息来准确推断出事件类型(李华昱等人, 2022)。因此，为了提高事件检测的准确性，有必要考虑引入上下文信息，以提供更丰富的语境。Lou等人(2021)提出一种多层双向网络（MLBiNet）,以同时捕获篇章级的事件关联和语义信息。Veyseh等人(2021)提出一种策略，动态地从文档中只选择与目标句子最相关的上下文句子，然后将这些句子输入到BERT中进行事件检测，以更有效地捕获长距离文档级上下文信息。为了增强模型对事件类型的理解和推断能力，本文为每条输入文本选择了对应的上下文句子，并设计信息融合模块，以建模输入句子与其上下文之间关系。

自2018年来，Goole、百度等公司发布了包括BERT(Devlin et.al , 2019)、ERNIE(Zhang et.al , 2019)等在内的多种模型，并在事件抽取等几乎所有自然语言处理任务中都表现出色。受基于领域自适应训练思想的启发，一些研究人员考虑使用电子化古籍资源对BERT进行继续预训练，以增强预训练语言模型在特殊领域的处理能力，并取得了一系列优秀的成果，例如GuwenBERT [1]、SikuBERT(王东波等人, 2022)、BERT-Ancient-Chinese(Wang 等人, 2022)等，极大地推动了古籍智能化的发展。本文整理了包括“二十四史”在内的古汉语文献，共计约97万条古汉语文本，并利用该文本对BERT-Ancient-Chinese进行了增量预训练。

[1]https://github.com/Ethan-yt/guwenbert

(c) 2024 中国中文信息学会计算语言学专业委员会

3 模型

3.1 任务定义

古文历史事件类型抽取评测任务的详细描述如下：针对给定的古汉语文本，本任务的目标是识别文本中所有的事件触发词及其对应的事件类型。以句子“进军建德，擒贼帅赵桑干。”为例，其中“进军”一词可视为“派兵到建德”事件的触发词，而“擒”则作为“抓住敌方的将领赵桑干”事件的触发词。因此，在此句子中，触发词分别为“进军”和“擒”，它们分别代表了“军事-备战-出兵”和“军事-作战-俘虏”这两种事件类型。

该评测任务包含两个子任务：（1）触发词识别：此子任务需识别文本中的事件触发词并标注其位置。触发词是指最能代表事件发生的词语，一般为句中的谓语动词（其他句子成分皆可）。（2）事件类型判别：此子任务需参考给定的事件类型体系，为每个触发词确定其所属的事件类型。子任务二进一步细分分为粗粒度（9大类）事件类型判别和细粒度（67小类）事件类型判别两个部分。

3.2 EIGC模型

本文设计了一个结合外部信息和全局对应矩阵的古文历史事件触发词和类型联合抽取模型EIGC（图1），用以建模上述子任务。该模型采用全局对应矩阵的方法，分别设计了触发词头尾对齐矩阵和触发词与事件类别对齐矩阵，以实现触发词与事件类型的抽取。考虑到触发词通常是句中的谓语动词（或其他句子成分），本文在模型中引入了词性信息，并通过字与词性全局对应矩阵的方式来建模二者的关系。此外，通常情况下，由于单个句子的信息不够丰富，只考虑单个句子的信息不足以推断出事件类型，本文还引入了文本上下文信息，以提升模型对输入句子的理解能力。

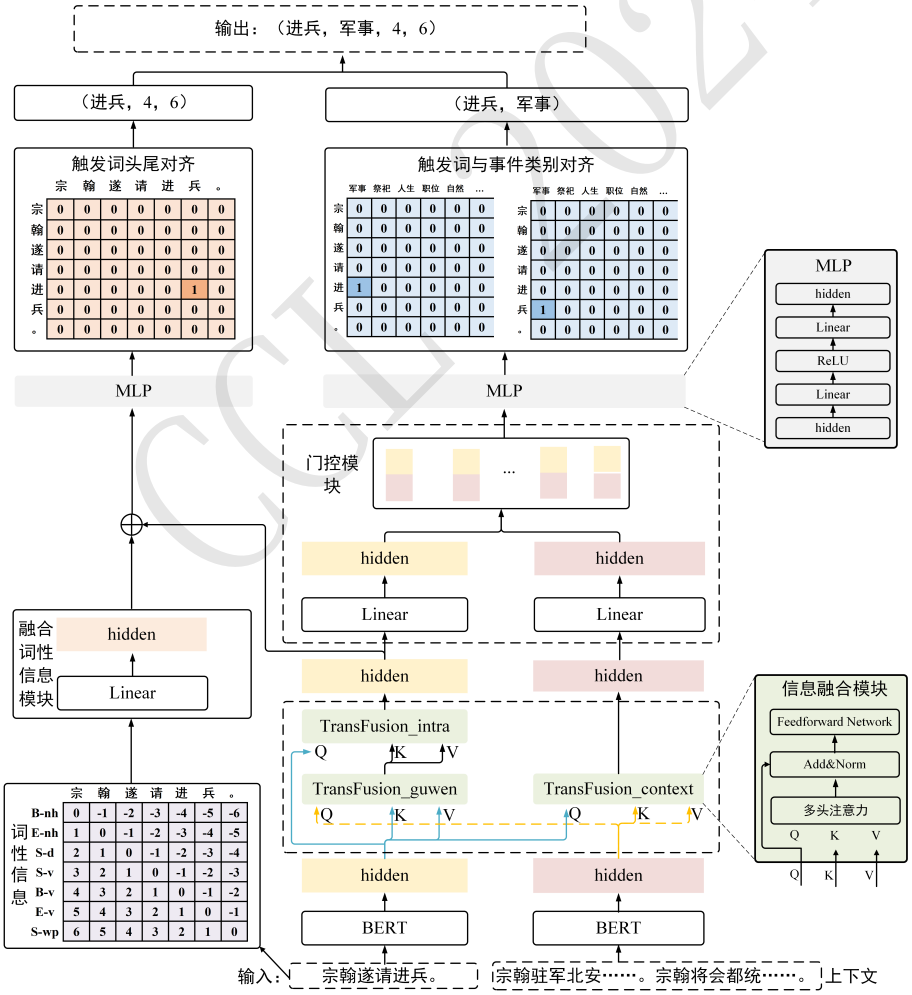


Figure 1: EIGC模型整体架构

(c) 2024 中国中文信息学会计算语言学专业委员会

3.2.1 编码层

对于给定的输入句子S，$S = \{s_1, s_2, \ldots, s_n\}$，$s_n$表示句子$S$中第$n$个字。本文采用BERT作为编码器，以获得$S$中每个字对应的向量表示。

$$\{h_1, h_2, \ldots, h_n\} = BERT(\{s_1, s_2, \ldots, s_n\}) \tag{1}$$

其中，$\{h_1, h_2, \ldots, h_n\}$表示BERT最后一层输出的隐藏层状态，$n$为输入文本的长度。

3.2.2 解码层

在本部分，本文将详细介绍四个主要模块：引入上下文信息模块、融合词性信息模块、触发词头尾对齐以及触发词与事件类别对齐模块。

（1）引入上下文信息模块

对于给定的上下文文本C，$C = \{c_1, c_2, \ldots, c_l\}$，$c_l$表示文本$C$中第$l$个字，本模块同样利用BERT，以获得上下文文本$C$中每个字对应的向量表示。

$$\{h\_c_1, h\_c_2, \ldots, h\_c_l\} = BERT(\{c_1, c_2, \ldots, c_l\}) \tag{2}$$

其中，$\{h_c_1, h_c_2, \ldots, h_c_l\}$表示BERT最后一层输出的隐藏层状态，$l$为上下文句子文本的长度。

信息融合模块：为了建模输入句子与其上下文之间，以及输入句子中字与字之间的关系，本文设计了三个信息融合模块（Transformer-Based Information Fusion Module，TransFusion）（图2），它主要由三个部分组成：多头注意力层、残差网络和前馈神经网络。

$$Q_i, K_i, V_i = W_Q \cdot h_i^Q, W_K \cdot h_i^K, W_V \cdot h_i^V \tag{3}$$

$$Att_i = Softmax\left(Q_i \cdot K_i^T\right) \cdot V_i \tag{4}$$

$$LN(Att_i) = Norm\left(Att_i\right) + Q_i \tag{5}$$

$$FFN(LN(Att_i)) = W_2\left(ReLU\left(W_1 \cdot LN(Att_i) + b_1\right)\right) + b_2 \tag{6}$$

其中，W_Q、W_K、W_V、W_1、W_2、b_1、b_2为可学习参数。在不同信息融合模块，模型分别使用了不同来源的$Q(Query)$、$K(Key)$和$V(Value)$。

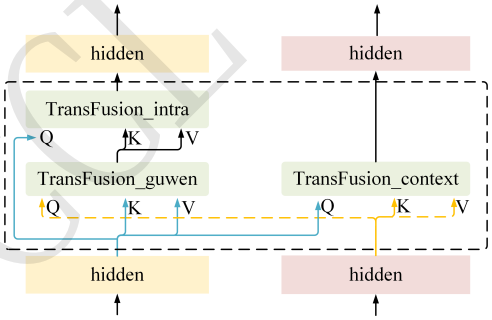


Figure 2: 信息融合层

在获取相应的Q、K、V之后，模型通过以下过程生成对应的隐藏层表示。

$$O_n^1 = \text{TransFusion}_{\text{guwen}}\left(h\_c_l^Q, h_n^K, h_n^V\right) \tag{7}$$

$$O_n^2 = \text{TransFusion}_{\text{context}}\left(h_n^Q, h\_c_l^K, h\_c_l^V\right) \tag{8}$$

$$O_n^3 = \text{TransFusion}_{\text{Intra}}\left(h_n^Q, h\_c_l^K, h\_c_l^V\right) \tag{9}$$

其中，$h\_c_l^Q$、$h\_c_l^K$、$h_c_l^V$分别表示该上下文第l个字对应的$hidden$作为$TransFusion$模块中的Q、K、V，$h\_c_n^Q$、$h\_c_n^K$、$h_c_n^V$分别表示该上下文第n个字对应的$hidden$作为$TransFusion$模块中的Q、K、V。

（2）融合词性信息

(c) 2024 中国中文信息学会计算语言学专业委员会

为了建模输入句子中每个$token$和词性之间的关系，本文设计了输入$token$和对应词性的对齐矩阵，并通过可学习参数让模型学习该矩阵所蕴含的$token$和词性间的关系信息。

$$Pos_{i,j} = I_{i,j}W_{pos}+b_{pos} \tag{10}$$

$$h_{i,j}^{pos} = Concat\left(O_i^3, O_j^3\right) + Pos_{i,j} \tag{11}$$

其中，I为图1中的词性信息矩阵，$I_{i,j}$表示输入句子中第i个$token$和第j个词性的对应，O_i^3和O_j^3为$TransFusion_context$输出的句子中第i个字和第j个字对应的隐藏层表示，$Concat$表示拼接，W_{pos}和b_{pos}为可训练参数。

（3）触发词头尾对齐

为了建模触发词，本文设计了一个基于全局对应矩阵的触发词头尾对齐模块（图3）。该模块的目标是抽取出所有可能的触发词。具体过程如下：对于输入文本S，$S = \{s_1, s_2, \ldots, s_n\}$，该矩阵的维度为$n \times n$，其中每个位置表示一对（触发词开始$token$，触发词结束$token$），该位置上的分数表示存在该（触发词开始$token$，触发词结束$token$）组合的概率，分数越高，概率越大。为了获得所有可能的（触发词开始$token$，触发词结束$token$）组合，本文同样为每个位置的分数设置了阈值，当该位置的分数大于该阈值时，则认为该位置对应的（触发词开始$token$，触发词结束$token$）组合存在，否则认为该组合不是（触发词开始$token$，触发词结束$token$）组合。

$$h_{i,j}^{pos} = MLP(h_{i,j}^{pos}) \tag{12}$$

$$P_{i,j}^{h-t} = \sigma\left(W_{h-t}h_{i,j}^{pos} + b_{h-t}\right) \tag{13}$$

其中，$h_{i,j}^{pos}$为融合词性信息模块输出的隐藏层表示，W_{h-t}和b_{h-t}为可训练参数，σ为$sigmoid$函数，MLP为多层感知机。

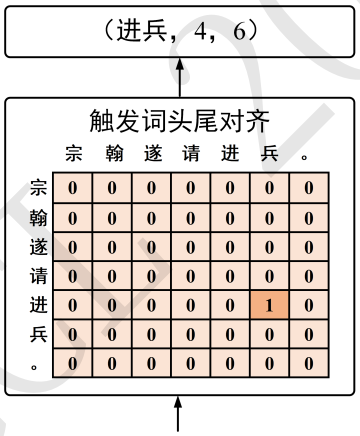


Figure 3: 触发词头尾对齐矩阵

（4）触发词与事件类别对齐

本文将触发词与事件类别对齐细化为两个子任务：触发词开始$token$和事件类别的全局对应，以及触发词结尾$token$和事件类型的全局对应。为此，本文设置了触发词开始$token$和事件类别的全局对应矩阵（图5左半部分）和触发词结尾$token$和事件类型的全局对应矩阵（图5右半部分）。

通过$TransFusion_context$和$TransFusion_guwen$，模型分别获得了输入句子与上下文信息进行交互后的信息的表示和输入句子间进行交互的信息表示。由于上述两种信息对事件类别对齐的贡献不同，我们采用了门控机制（图4）对上述信息进行控制。

$$g = \sigma(W_3 \cdot O_i^3 + W_4 \cdot O_i^2 + b_{gate}) \tag{14}$$

$$u_i = [g \circ \mathrm{O}_i^3] \oplus [(1-g) \circ \mathrm{O}_i^2] \tag{15}$$

其中，O_i^3为输出的隐藏层表示，O_i^2为$TransFusion_guwen$输出的隐藏层表示，W_3、W_4、b_{gate}为可学习参数，$\circ$表示元素相乘。

（c）2024 中国中文信息学会计算语言学专业委员会

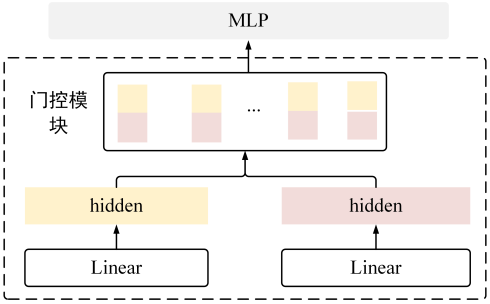


Figure 4: 门控模块

在获得融合后的u_i表示后，我们将其作为触发词与事件对齐矩阵解码部分（图5）的输入。

$$h_{i,j}^{et} = \text{MLP}(u_i^{3\_2}) \tag{16}$$

$$p_{i,j}^{ht} = \sigma\left(h_{i,j}^{et} W_{ht} + b_{ht}\right) \tag{17}$$

$$p_{i,j}^{tt} = \sigma\left(h_{i,j}^{et} W_{tt} + b_{tt}\right) \tag{18}$$

其中，W_{ht}、W_{tt}、b_{ht}、b_{tt}为可训练的参数，MLP为多层感知机。

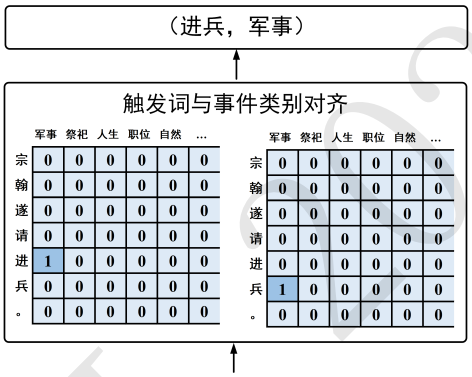


Figure 5: 触发词和事件类型全局对应

3.3 损失函数

该模型采用交叉熵损失函数作为模型的损失函数。模型最终的损失l^{total}包含了l^{et}和l^{ht}两部分。在模型训练过程中，本文将公式中的系数设置为ρ=φ=1。

$$l^{total} = \rho l^{et} + \varphi l^{ht} \tag{19}$$

首先是l^{et}，其包括了l^{eth}、l^{ett}，两者具体计算过程如下:

$$l^{et} = \alpha l^{eth} + \beta l^{ett} \tag{20}$$

$$l^{eth} = \frac{-1}{n \times n^t} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log p_{i,j}^{th} + (1 - y_{i,j}) log(1 - p_{i,j}^{th}) \tag{21}$$

$$l^{ett} = \frac{-1}{n \times n^t} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log p_{i,j}^{tt} + (1 - y_{i,j}) log(1 - p_{i,j}^{tt}) \tag{22}$$

其中，$p_{i,j}^{ht}$表示第i个$token$为触发词开始$token$和第j个事件类型是否存在联系的条件概率；$p_{i,j}^{tt}$表示第i个$token$为触发词结束$token$和第j个事件类型是否存在联系的条件概率。在训练过程中，α和β被设置为0.5。

(c) 2024 中国中文信息学会计算语言学专业委员会

其次是l^{ht}，其为触发词头尾全局对应矩阵对应的损失。

$$l^{ht} = \frac{-1}{n \times n} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log p_{i,j}^{ht} + (1 - y_{i,j}) log(1 - p_{i,j}^{ht}) \tag{23}$$

其中，$p_{i,j}^{ht}$表示第i个$token$为触发词开始$token$以及第j个$token$为触发词结尾$token$的条件概率。在上述公式中，n为输入序列的长度，n^t为定义的事件类型数。

3.4 增量预训练

为确保模型能够充分理解古汉语的表达方式和文化内涵，本文基于东北大学[2]公布的“二十四史”等古汉语文献，共计约97万条古汉语文本，对BERT-Ancient-Chinese进行继续预训练。

在进行增量预训练时，本文沿用了BERT预训练方法中的掩码机制，随机掩码古汉语句子中15%的字符，其中80%的字符被替换为“[MASK]”，10%的字符被替换为词表中的任意一个词，剩余10%保持不变。此外，本文还设置了batch_size为64，文本最大长度为128，epoch为300，以确保训练过程的稳定性和有效性。增量预训练过程如图6所示。

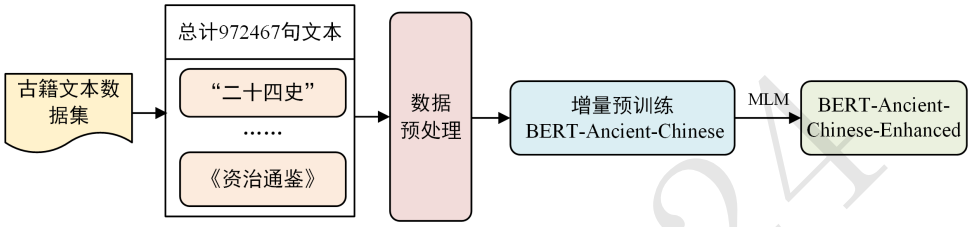


Figure 6: 增量预训练过程图

3.5 词性和上下文的获取与处理

（1）上下文

本文依据“sen_id”和“doc_id”对语料进行了重新排序，以便构建篇章结构，从而获取古文文本的前后上下文。例如“宗翰遂请进兵。”，其前后上下文分别是“宗翰驻军北安，使希尹经略近地，获辽护卫耶律习泥烈，知辽主猎于鸳鸯泺。”和“宗翰将会都统杲于奚王岭。”。上下文的获取与处理如图7所示。

{"sen_id": 7778, "doc_id": 32047, "text": "宗翰驻军北安，使希尹经略近地，获辽护卫耶律习泥烈，知辽主猎于鸳鸯泺。"}
{"sen_id": 7779, "doc_id": 32047, "text": "宗翰遂请进兵。"}
{"sen_id": 7780, "doc_id": 32047, "text": "宗翰将会都统杲于奚王岭。"}

Figure 7: 上下文的获取与处理（“sen_id”表示句子序号，“doc_id”表示文档编号）

（2）词性

本文使用了jiayan工具[3]对古文文本进行词性标注，以获取相应的词性标签。具体而言，对于一个给定的古文文本$S = \{w_1, w_2, w_3 \ldots, w_n\}$，通过使用jiayan进行分词处理，可以得到k个词，即$D = \{d_1, d_2, d_3, \ldots, d_k\}$，其中$d_i$表示第i个词，且D中的每个元素$d_i$均对应一个词性标签(Yu 等人, 2023)。

例如“宗翰遂请进兵。”，经过分词后，我们得到了“宗翰”、“遂”、“请”、“进兵”和“。”这五个词或标点符号，它们各自的词性标签分别为“nh（人名）”、“d（介词）”、“v（动词）”、“v（动词）”和“wp（标点符号）”。详细的词性分类可见jiayan中的词性表。

在实验中，我们采用了“B/M/E/S-词性”的组合标签方式来表示词性，其中B（Begin）表示词汇的第一个汉字，M（Middle）表示词汇中间的汉字，E（End）表示词汇的最后一个汉字，而S（Single）则表示仅包含一个汉字。词性标注示例如图8所示。

[2]https://github.com/NiuTrans/Classical-Modern
[3]https://github.com/jiaeyan/Jiayan

(c) 2024 中国中文信息学会计算语言学专业委员会

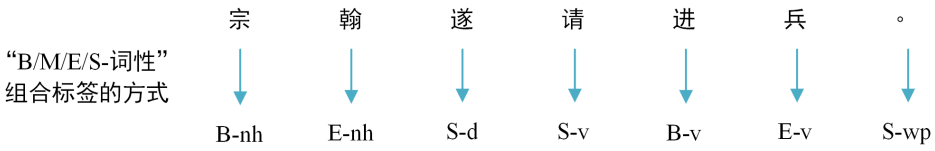


Figure 8: 词性的获取与处理

4 实验

4.1 数据集

本文主要使用古文历史事件类型抽取评测任务提供的数据集[4]对模型进行微调。该数据集共包含8122条古文文本，其中训练集5650条，验证集1218条，测试集1254条。事件类型体系涵盖9大类、67小类。

4.2 实验设置与评估指标

实验设置：本文实验运行环境采用ubutun18.04系统、python3.7、CUDA版本11.3、以及显卡为RTX3090。此外，本文将epoch为100，触发词与事件类别对齐模块和触发词头尾对齐的阈值设置为0.5，训练和测试时batch size分别为8和6，输入文本最大长度为128。

评估指标：触发词识别任务的评估采用两个指标：Exact-match-score和Subset，这两个指标的平均值作为最终的评估结果。事件类型判别任务的评估采用宏平均（macro-average）和微平均（micro-average）两个指标，这两个指标的平均值作为最终的评估结果。

4.3 实验结果

（1）在无答案测试集上的实验结果

如表1所示，本文所提出的模型在所有参赛队伍中排名第二，三个任务上的总F1值达到了76.2%。

排名	任务1	任务1*0.4	任务2	任务2*0.3	任务3	任务3*0.3	总分
1	0.763	0.305	0.842	0.253	0.779	0.234	0.791
2	**0.703**	**0.281**	**0.825**	**0.248**	**0.777**	**0.233**	**0.762**
3	0.700	0.280	0.824	0.247	0.771	0.231	0.758
4	0.672	0.269	0.810	0.243	0.769	0.231	0.742
5	0.690	0.276	0.799	0.240	0.752	0.226	0.741
6	0.674	0.270	0.808	0.242	0.739	0.222	0.734

Table 1: 评测总分排名

（2）在验证集上的消融实验结果

为了验证本文所提出模型中各个模块的有效性，我们在验证集上进行了消融实验。在触发词识别、粗粒度事件类型判别和细粒度事件类型判别三个任务上，我们仅采用了微平均作为评估指标。

如表2所示，在去除增量预训练的消融实验中，三个任务的F1值分别下降了0.6%、1.4%和0.5%。这可能是由于通过在“二十四史”等古汉语文献的数据集上对BERT-Ancient-Chinese进行增量预训练，为模型提供了丰富的基于“二十四史”的领域知识，从而帮助模型能够更好地理解和处理古代汉语中的词汇、语法及表达方式，提升了模型在相关任务上的表现。

在去除词性信息的消融实验中，三个任务的F1值分别下降了0.8%、1.2%和0.5%。相较于在其他三个模块上的效果，触发词识别任务在该模块上下降的最多。这可能是由于触发词通常是句中的谓语动词（或其他句子成分），通过全局对应矩阵来建模输入文本中每个字与词性间的联系，能有效提升模型对古文文本中触发词识别的精度。

在去除门控模块或上下文信息的消融实验中，从粗粒度事件类型判别任务可以看出，当去除门控模块后，F1值下降了0.9%；当去除上下文信息后，F1值下降了0.2%。这可能是由于门控

[4]https://github.com/NLPInBLCU/CHED2024

(c) 2024 中国中文信息学会计算语言学专业委员会

模块能有效地筛选出输入文本和引入的上下文信息中与事件类别判别最相关的信息，从而提升模型对粗粒度事件类型判别的性能。然而，在触发词识别和细粒度事件类型判别任务中，当去除门控模块后，两个任务的F1值基本没发生变化；当去除上下文信息后，两个任务的F1值分别下降了0.6%和0.5%。这可能是由于在这两个任务中，模型已经通过其他模块（如上下文信息和词性信息）充分捕获了必要的信息，从而降低了门控模块对整体性能的影响。

	触发词识别			粗粒度事件类型判别			细粒度事件类型判别		
	P	R	F1	P	R	F1	P	R	F1
EIGC	**0.816**	0.817	**0.817**	0.784	**0.813**	**0.798**	**0.767**	0.768	0.768
-增量预训练	0.803	0.818	0.811	0.778	0.789	0.784	0.756	0.770	0.763
-词性信息	0.798	0.819	0.809	0.766	0.808	0.786	0.753	0.773	0.763
-门控模块	0.814	0.821	0.817	0.776	0.803	0.789	0.766	0.772	**0.769**
-上下文	0.796	**0.826**	0.811	**0.792**	0.801	0.796	0.749	**0.778**	0.763

Table 2: 在验证集上的消融实验结果

5 总结与展望

针对传统pipeline方法容易产生级联错误传播，以及大多数事件检测方法仅依赖句子层面信息而忽略上下文关联性的问题，本文提出了一种融合外部信息和全局对应矩阵的模型EIGC，旨在联合抽取古文历史事件的触发词及类型。实验结果表明，该模型在三个任务上均实现了显著的性能提升，总F1值得分高达76.2%。

未来，我们将探讨如何整合更多外部信息，例如现代文翻译，以进一步提升模型性能。此外，考虑到大型语言模型展现出的强大世界知识掌握能力和上下文理解能力，我们也计划研究如何利用ChatGPT[5]、荀子[6]等大语言模型，以进一步提升古文历史事件检测任务的性能。

参考文献

吴旭, 卞文强, 颉夏青, 孙利娟. 2023. 机器阅读理解式中文事件抽取方法. 计算机工程与应用, 59(16): 93-100.

李华昱, 毕经纶, 闫阳. 2022. 限定域中文事件抽取研究综述. 计算机工程与应用, 58(18): 43-58.

王浩畅,周郴莲,Marius Gabriel PETRESCU. 2023. 基于深度学习的事件抽取研究综述. 软件学报, 34(08): 3905-3923.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 167–176.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A Language-Independent Neural Network for Event Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 66–71.

Chen Chen and Vincent Ng. 2012. Joint Modeling for Chinese Event Extraction with Rich Linguistic Features. In *Proceedings of COLING 2012*, pages 529–544.

贺瑞芳, 段绍杨. 2019. 基于多任务学习的中文事件抽取联合模型. 软件学报, 30(4): 1015-1030.

Nguyen, Trung Minh, and Thien Huu Nguyen. 2019. Joint Modeling for Chinese Event Extraction with Rich Linguistic Features. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6851-6858.

Dongfang Lou, Zhilin Liao, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. MLBiNet: A Cross-Sentence Collective Event Detection Network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4829–4839.

[5]https://chat.openai.com

[6]https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM

(c) 2024 中国中文信息学会计算语言学专业委员会

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Ngo Trung, Bonan Min, and Thien Huu Nguyen. 2021. Modeling Document-Level Context for Event Detection via Important Context Selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5403–5413.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

王东波,刘畅,朱子赫,刘江峰,胡昊天. 2022. SikuBERT与SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究. 图书馆论坛, 42(06): 31-43.

Pengyu Wang and Zhichen Ren. 2022. The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168.

Zijian Yu, Tong Zhu, and Wenliang Chen. 2023. 基于句法特征的事件要素抽取方法(Syntax-aware Event Argument Extraction). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 196–207.

CCL 2024

CCL24-Eval 任务5系统报告：基于大小模型结合与半监督自训练方法的古文事件抽取

付薇薇，王士权，方瑞玉，李孟祥，何忠江，李永翔，宋双永
中国电信人工智能研究院
{fuweiwei, wangsq23, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

摘要

本文描述了队伍“TeleAI”在CCL2024古文历史事件类型抽取评测任务（CHED2024）中提交的参赛系统。该任务旨在自动识别出古代文本中的事件触发词与事件类型，其中事件类型判别被分为粗粒度和细粒度的事件类型判别两部分。为了提高古文历史事件类型抽取的性能，我们结合了大模型和小模型，并采用了半监督自训练的方法。在最终的评估中，我们在触发词识别任务得分0.763，粗粒度事件类型判别任务得分0.842，细粒度事件类型判别任务得分0.779，综合得分0.791，在所有单项任务和综合评分上均排名第一。

关键词： 事件抽取；半监督；自训练

System Report for CCL24-Eval Task 5: Ancient Chinese Text Event Extraction Based on Semi-Supervised Self-Training Method Combining Large and Small Models

Weiwei Fu, Shiquan Wang, Ruiyu Fang, Mengxiang Li, Zhongjiang He, Yongxiang Li , Shuangyong Song
Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd
{fuweiwei, wangsq23, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

Abstract

This article describes the system submitted by the team ”TeleAI” for the CCL2024 Ancient Text Historical Event Type Extraction Evaluation Task (CHED2024). The task aims to automatically identify event trigger words and event types in ancient texts, with event type classification divided into coarse-grained and fine-grained parts. To improve the performance of ancient text historical event type extraction, we combined large and small models and adopted a semi-supervised self-training method. In the final evaluation, we scored 0.763 in the trigger word recognition task, 0.842 in the coarse-grained event type classification task, and 0.779 in the fine-grained event type classification task, with an overall score of 0.791. We ranked first in all individual tasks and the overall score.

Keywords: Event Extraction , Semi-Supervised , Self-Training

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

古文事件抽取是从自然语言文本中识别和提取相关事件信息的过程，这是正确分析古汉语文本、进一步提升事件抽取技术及中国古代历史文本的数字化研究水平的重要步骤。然而，古代事件抽取任务缺乏公开的用于模型训练和评测的数据资源，制约了技术的进一步发展。除此之外，针对古代汉语的信息抽取任务还面临着古文句法语义复杂，使用范围小的复杂挑战。

为了进一步推动古代文本事件抽取任务的研究，研究者依托中国古文历史事件检测数据集（CHED2024）推出古文历史事件类型抽取评测任务，该评测任务共分为三个子任务，分别是触发词识别，识别出文本中触发词及其位置；粗粒度事件类型判别，识别出触发词所属的粗粒度事件类型；细粒度事件类型判别，识别出触发词所属的细粒度事件类型。具体如图1所示。

在本文中，我们使用了一种结合大小模型的半监督自训练方法，用于提升古文事件抽取任务的性能。具体来说，首先，我们利用比赛提供的CHED2024数据得到初始的模型；其次利用该模型在同领域的二十四史未标注古文数据上生成伪标签，利用标签一致性筛选得到的高质量的伪标签数据；最后利用伪标签数据与CHED2024数据进行训练得到最优结果。值得注意的是，在古文事件抽取任务中，由于古文事件抽取任务数据中存在标签不均衡的现象，我们利用大模型强大的泛化能力提高模型在稀缺标签数据上的准确率，弥补小模型在该类数据上的性能损失。

我们的系统在最终线上评测中触发词识别任务得分0.763，粗粒度事件类型判别任务得分0.842，细粒度事件类型判别任务得分0.779，综合得分0.791，排名第一。

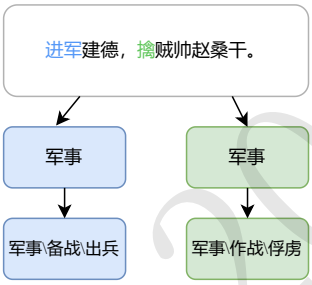


Figure 1: CHED2024任务示例

2 相关工作

事件抽取的主要目标是从文本数据中获取事件信息，事件作为一种特定的信息形式，是指某一特定时间、某一特定地点发生的某一特定事件，涉及一个或多个参与者，通常可以用状态的变化来描述(Doddington et al., 2004)。事件抽取任务旨在将此类事件信息从非结构化的纯文本中提取成结构化的形式，这种形式主要描述现实世界中发生的事件的信息。事件抽取可以作为许多任务的前提，比如信息检索、推荐系统、智能问答、知识图谱构建等，这些任务都依赖于事件抽取任务从文本中提取出的结构化事件信息。事件抽取的主要方法可以分为四种，分别是基于分类任务的事件抽取方法、基于阅读理解任务的事件抽取方法、基于序列标注任务的事件抽取方法和基于序列到结构生成任务的事件抽取方法。

基于分类任务的事件抽取方法通常将事件抽取任务转化为多类别分类问题，该方法将每个文本片段或句子分类为预定义的事件类型(Mekala and Shang, 2020; Guo et al., 2021)。这要求预先定义一组事件类别，并为每个类别构建相应的训练样本。在基于分类的任务中，触发器识别是对一个词是否为触发器进行分类，在确定事件类型后，再将句子中的实体分类为预定义的论元角色。该方法的特点是简单直观，易于实现和解释。它适用于事件类型较少且确定的情况，且对每个事件类型的分类效果较好。然而，它忽略了事件内部结构的细节。

基于机器阅读理解任务的事件抽取方法通过利用机器阅读理解模型来从文本中抽取事件信息(Guo et al., 2020)。该方法要求模型能够理解文本的上下文，定位事件的触发词、事件论元和其他关键信息，并将其结构化表示。在基于机器阅读理解的时间抽取任务中，模型首先需要确定文本所属的事件类型，然后根据事件类型确定要提取的事件论元。首先根据触发词分类确认文本的事件类型，然后为每个论元设计一个问题模式。最后，基于机器阅读理解的事件提取

方法将设计好的问题模式逐个应用于提取模型，根据输入文本获取答案，每个答案对应一个事件论元。基于机器阅读理解任务的事件抽取方法具有捕捉事件上下文信息的优势，能够提供更丰富的事件抽取结果(Chen et al., 2021)。然而，该方法对模型的理解和推理能力要求较高，需要模型具备对文本上下文进行推理和抽象的能力。

基于序列标注任务的事件抽取方法将文本中的每个词标注为事件的不同组成部分，如触发词和论元，常用的序列标注方法包括命名实体识别和关系抽取(Gui et al., 2020)。序列标注任务是基于词级别的多分类任务，旨在基于词级别事件类型进行直接匹配和提取事件论元。它主要涉及两个核心任务：事件类别的识别和分类，以及事件论元的提取。基于序列标注的事件抽取方法简单、快速，能够实现事件类型与事件参数的匹配，而无需使用额外的特征。在基于序列标注的任务中，触发词识别是将文本中一个单词标记为触发词的过程。通过序列标注方法，可以将目标事件从文本中标注出来，使其适用于事件提取任务。对于给定的文本事件类型，通过序列标记模型对事件论元进行标记。序列标注模型输出的序列对文本中的所有单词进行了标注。虽然基于序列标注任务的事件抽取方法能够对文本进行细粒度的标注，捕捉事件的具体信息，但它需要高质量的标注数据和对序列建模的技术支持。

基于序列到结构生成的事件提取方法采用端到端的方式从文本中提取事件信息，它将所有任务统一建模为一个模型，并直接生成事件的触发词和事件论元，这种方法通常使用编码器-解码器结构(Lu et al., 2021)。在基于序列到结构生成的任务中，模型通过统一建模和预测不同的标签，可以实现对事件抽取的端到端处理。这种方法的特点是直接生成事件抽取结果，无需预定义的类别或标签。这种方法能够捕捉文本中复杂的事件结构和上下文信息，并提供结构化的事件抽取结果。然而，基于端到端生成任务的事件抽取方法对大量的训练数据和计算资源要求较高。

在本文中，我们结合大语言模型(Wang et al., 2024)与小语言模型(Liu et al., 2023)来完成评测任务，具体来说，在小模型上我们使用GRTE(Ren et al., 2021)，该方法通过一个迭代的模型来提升对全局特征(Global Feature)的学习，进而提高模型性能；在大模型上我们使用XunziALLM，该模型经过古文领域的持续训练在古文任务上具有较好的表现，可以准确剖析古籍文本的复杂性，进一步提高模型在古文事件抽取任务上的效果。

3 方法

在本文中，我们采取了一种半监督自训练学习的方法，通过融合大小模型的优势，显著提升了古文事件抽取的性能。半监督自训练学习方法如算法1所示，下面我们将在各个小节详细介绍每个模块的具体细节。

Algorithm 1: 半监督自训练学习方法

Input: 有标签数据D_l, 无标签数据D_u, 初始模型M
Output: 训练好的模型M'

用有标签数据D_l 初始化模型M;
$D'_u \leftarrow D_u$;
while 未达到停止准则 **do**
  在$D_l \cup D'_u$ 上训练模型M;
  使用模型M 为无标签数据D'_u 生成伪标签;
  基于一致性检验筛选高质量的伪标签数据;
  $D'_u \leftarrow$ 选取的高质量伪标签数据;
在$D_l \cup D'_u$ 上训练最终模型M';
return M';

3.1 模型初始化

在古文事件抽取任务中，由于古文事件抽取任务数据中存在标签不均衡的现象，我们利用大模型强大的泛化能力提高稀缺样本的准确率，弥补小模型在该类数据上的性能损失。具体来说，我们使用CHED2024训练集初始化大模型与小模型，其中小模型部分经过对比我们选择在古文事件抽取中表现最好的GuwenBERT作为基座模型，我们将事件抽取任务转换为关系抽取任务，使用GRTE方法通过对全局特征的学习提高从古文文本中抽取事件触发词与细粒度事件

类型的准确率。大模型部分我们选择经过古籍数据微调的Xunzi-Baichuan-7B作为基座模型，构造直接从原始文本中获取事件触发词与细粒度事件类型的指令。

3.2 伪标签生成

为了获取更多与古文事件抽取相关的高质量训练数据，我们首先从网络上获取了开源的《二十四史》原文，其次通过规则方式筛选出17w无标签数据，再次利用初始化的大小模型对无标签数据进行预测，最后利用标签一致性检验策略从中过滤筛选出4w条高质量伪标签数据。

标签一致性检验主要通过比较大模型和小模型对于同一条数据预测标签是否一致来判断该标签是否可靠，同时为了缓解训练数据中类别不均衡问题，我们提高了大模型预测结果中稀有类别标签数据的采纳比例。通过一致性检验可以筛选出高质量的伪标签，减少低质量标签对模型训练的负面影响，利用多次预测的标签一致性能够捕捉更可靠的样本信息，从而提升模型的泛化能力，通过有效过滤掉噪声数据和错误标签可以确保模型训练数据的纯净度。

3.3 训练策略

为了更好的利用数量较少的高质量CHED2024数据，我们在训练过程中首先使用伪标签数据作为训练集，CHED2024数据作为验证集，当模型在伪标签数据上达到稳定状态时，切换到高质量的CHED2024数据。先使用伪标签数据再使用真实标签数据训练策略可以有效提高模型性能。先使用伪标签数据进行训练，有助于提高模型的泛化能力和对数据的理解；后用真实标签数据进行训练可以使模型更快地收敛并获得更好的性能。同时这种方法还能够降低对有标签数据的依赖，可以最大程度地利用有限的有标签数据资源。

除此之外，我们还在伪标签数据与CHED2024数据中通过上下文增强策略提高模型对于稀缺样本预测的准确率，具体来说，对于稀缺样本，我们从训练集中将其与前后若干条数据进行拼接，然后将其加入训练数据。

4 实验

为了提高古文事件抽取任务上模型的性能表现，我们针对CHED2024数据集进行了若干实验，值得注意的是，本小节实验结果均基于CHED2024验证集中细粒度事件类型判别任务计算得到。

4.1 基座模型对比

为了寻找到最适合处理古籍领域文本的基座模型，我们首先在CHED2024数据集上分别尝试了若干开源预训练语言模型，实验结果如表1示。

Model	BERT	RoBERTa-base	SIKU-BERT-base	SIKU-RoBERTa-base	BTfhBER	Guwen-NER-RoBERTa-base
CHED2024_dev	0.7032	0.7287	0.7735	0.798	0.793	**0.828**

Table 1: CHED2024数据集上开源预训练语言模型的实验结果

表1展示了在CHED2024验证集上在GRTE结构下，不同开源预训练语言模型的性能表现，其中的F1得分为该开源预训练语言模型的最高Micro-f1得分。从实验结果可以看出Guwen-NER-RoBERTa-base在CHED2024数据集上表现最佳，因此在本次评测中我们选择该模型作为后续实验的基座模型。

4.2 GRTE VS GPLINKER

我们在CHED2024数据集对GRTE方法与GPLINKER方法上进行对比，利用5折交叉训练的方式在训练集上进行训练，在验证集上的的实验结果如表2所示：

我们将事件抽取的任务改为关系抽取的任务形式，表2展示了GRTE方法与GPLINKER方法在CHED2024数据集上微调后的实验结果，实验结果表明基于GRTE方法的模型性能优于GPLINKER方法，在Micro-f1上获得了4.6%的相对提升，我们在后续的实验中选用GRTE作为抽取方法。

CCL 2024

Method	GRTE	GPLINKER
Fold-0-Micro-f1	0.765	0.768
Fold-1-Micro-f1	0.789	0.774
Fold-2-Micro-f1	0.824	0.704
Fold-3-Micro-f1	0.847	0.827
Fold-4-Micro-f1	0.802	0.798
Average	0.828	0.791

Table 2: CHED2024数据集上GRTE与GPLINKER的实验结果

4.3 CHED2024实验结果

为了利用大量的无标注数据，我们基于半监督自训练的方式获取了高质量的伪标签数据，并通过上下文增强的方式提高稀缺样本的准确率，为了证明伪标签数据与上下文增强策略的有效性，我们基于伪标签数据与CHED2024数据进行了实验，实验结果如表3所示。

Metric	Micro-f1	Macro-f1
CHED2024_dev	0.828	0.723
CHED2024_dev+Context	0.831	0.752
CHED2024_dev+Pseudo	0.904	0.768

Table 3: CHED2024验证集上模型微调的实验结果

表3展示了CHED2024验证集上Guwen-NER-RoBERTa-base+GRTE方法的实验结果，其中CHED2024_dev表示只使用CHED2024训练集与验证集，CHED2024_dev+Context表示在CHED2024数据集上利用稀缺样本的上下文对其进行数据增强，CHED2024_dev+Pseudo表示使用CHED2024与通过半监督自训练方法获取的高质量伪标签数据。基于上下文对数据集稀缺样本进行增强，具体来说，对于稀缺样本，我们从训练集中将其与前后若干条数据进行拼接，然后将其加入训练数据。实验结果表明加入上下文增强后的稀缺样本数据可以在Micro-f1上获得0.3%的相对提升，在Macro-f1上获得4.0%的相对提升。我们还利用半监督自训练的方式从大量无标注数据中获取高质量的伪标签数据提高模型最终的性能。在对于伪标签数据的利用上，我们先利用伪标签数据作为训练集，CHED2024作为验证集，当模型在伪标签数据上达到稳定状态时，切换到CHED2024数据集。实验结果表明该方法在验证集上Micro-f1获得了9.2%的相对提升，在Macro-f1上获得了5.2%的相对提升。

Task	**Rank**	**Exact**	**Subset**	**Score**
Trigger Word Recognition	1	0.636	0.889	0.763
Task	**Rank**	**Macro-f1**	**Micro-f1**	**Score**
Coarse-grained Event Type Classification	1	0.839	0.845	0.842
Fine-grained Event Type Classification	1	0.743	0.814	0.779

Table 4: CHED2024测试集上模型的实验结果

表4展示了我们队伍在最终测试集上的实验结果，其中触发词识别任务得分为0.763，排名第一；粗粒度事件类型判别任务得分为0.842，排名第一；细粒度事件类型判别任务得分为0.779，排名第一，综合得分0.791，排名第一，具体计算过程如公式1所示:

$$\text{Total\_Score} = \text{Task\_1} \times 0.4 + \text{Task\_2} \times 0.3 + \text{Task\_3} \times 0.3 \quad (1)$$

5 总结与展望

在本次CCL2024古文事件抽取评测中，我们利用大小模型结合与半监督自训练的方式从开源的无标注数据中获取到高质量的伪标签数据，然后依靠CHED2024伪标签数据提高模型在古文事件抽取上的表现，我们“TeleAI”队伍提交的系统在三项任务中均取得第一名的成绩，综合得分为0.791，排名第一。然而，本次比赛由于时间因素我们未对伪标签数据进行多轮迭代，未来可以通过多轮迭代的方式持续提高伪标签数据的质量，进一步提高模型性能。

参考文献

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12666–12674.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Tao Gui, Jiacheng Ye, Qi Zhang, Zhengyan Li, Zichu Fei, Yeyun Gong, and Xuan-Jing Huang. 2020. Uncertainty-aware label refinement for sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2316–2326.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020. 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.

Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. 2021. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12929–12936.

Shixuan Liu, Chen Peng, Chao Wang, Xiangyan Chen, and Shuangyong Song. 2023. icsberts: Optimizing pre-trained language models in intelligent customer service. *Procedia Computer Science*, 222:127–136.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2646–2656.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.

CCL 2024

Overview of CCL24-Eval Task 5: Classical Chinese Historical Event Detection Evaluation

Zhenbing Feng[1,2] , Wei Li[1], Yanqiu Shao[1,2,*]
Information Science School, Beijing Language and Culture University [1]
Language Resources Monitoring and Research Center [2]
15 Xueyuan Road, HaiDian District, Beijing, 100083
zbfengblcu@163.com, liweitj47@blcu.edu.cn, yqshao163@163.com

Abstract

Event detection involves identifying and extracting event information from natural language texts. The complex syntax and semantics of Classical Chinese, coupled with its limited usage, pose significant challenges for information extraction tasks on classical Chinese texts. At the 23rd China National Conference on Computational Linguistics (CCL 2024), we launched an evaluation task focused on the extraction of historical events from Classical Chinese. We used our constructed Classical Chinese Historical Event Logical Schema to identify event triggers and classify event types. The evaluation utilized the Classical Chinese Historical Event Detection Dataset (CHED), annotated from *The Twenty-Four Histories* corpus, with the aim of enhancing event extraction technologies and advancing the digital study of classical Chinese historical texts. The evaluation included two subtasks and attracted 28 teams, with 15 teams submitting valid results. In the subtask of trigger identification, the best-performing system achieved an Exact match score of 63.6%. In the subtasks of coarse-grained and fine-grained event type classification, the top systems achieved F1-scores of 84.5% and 81.4%, respectively.

1 Introduction

Event detection is the process of identifying and extracting relevant event information from natural language texts. Given the complex syntax and semantics of classical Chinese, coupled with its limited usage scope, the task of information extraction for classical Chinese texts remains a significant challenge.

Constructing high-quality datasets tailored to specific domains is essential for event detection tasks. While several high-quality Event Detection (ED) datasets exist for English and modern Chinese, including ACE 2005 (Walker et al., 2006), LEVEN (Yao et al., 2022), MAVEN (Wang et al., 2020), PoE (Li et al., 2022), and DuEE (Li et al., 2020), classical Chinese lacks such datasets due to its semantic complexity and unique historical context. Large-scale datasets in English and modern Chinese are not directly applicable to classical Chinese ED.

Research in deep learning for event detection has explored historical event detection in classical Chinese texts, such as the study of war events in Shiji and ZuoZhuan (Dang, 2021) (Jiuming Ji, 2015) (Zhongbao et al., 2020). However, there still remain challenges such as sparse training data, complex text structures, semantic ambiguity. These issues highlight the need for constructing specific datasets and refining text features. To address these critical challenges and enhance the accuracy and efficiency of classical Chinese event detection , we have developed the classical Chinese Historical Event Dataset (CHED) (Congcong et al., 2023). This dataset is intended to serve as a benchmark for advancing the development of event detection algorithms in classical Chinese historical texts.

The Language and Culture Computing Laboratory (**LCC-Lab**) of Beijing Language and Culture University conducted the historical event type extraction evaluation as part of the CCL 2024 conference. The task aimed to assess the performance of algorithms in detecting event types in classical texts and to promote research and development in classical Chinese event extraction technology. The competition

* Corresponding Author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 201-209, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

attracted 28 teams, with 15 submitting valid results. The best-performing systems achieved a match score of 76.3% in Trigger Identification and F1-score of 84.5% and 81.4% in fine and coarse Event Type Classification tasks, respectively.

The evaluation included two subtasks: Trigger Identification and Event Type Classification, corresponding to the two steps of event type detection. We summarize the two main features as follows:

(1) Event Type Schema and Dataset [1]. The evaluation used the classical Chinese Historical Event Logical Schema with 9 major categories and 67 subcategories. The CHED dataset, based on *The Twenty-Four Histories* corpus, includes 8,122 valid event instances, providing a robust data benchmark for event detection in classical Chinese texts.

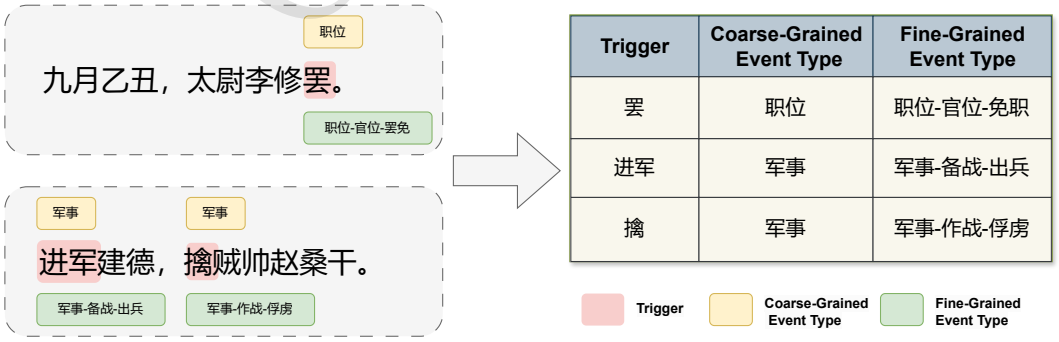
(2) Task Setup. Trigger Identification focused on marking event triggers, while Event Type Classification assigned event types at coarse and fine granularities. This setup ensures a comprehensive assessment of the system's capabilities in accurately identifying event triggers and tests the precision of event type detection algorithms at different granularities, making the evaluation criteria more rational.

2 Task Description

In this section, we will provide a detailed description of the two subtasks.

Subtask 1: Trigger Identification. This subtask involves identifying event triggers within the text and marking their locations. Our dataset is constructed based on the principle of minimal triggers, primarily using single-syllable words that best represent the occurrence of events within the text. The goal is to ensure that the identified triggers are concise yet accurately indicative of the events described. As shown in Figure 1, in the sentence "九月乙丑，太尉李修罢。" (*In September of Yi Chou, General Li Xiu was dismissed.*), the word "罢" (ba) means *dismiss*. Therefore, the trigger in this sentence is "罢" (ba).

Subtask 2: Event Type Classification. In this subtask, each identified trigger word is classified into an event type based on a predefined event type schema. The classification is conducted at both coarse-grained and fine-grained levels, allowing for a detailed understanding of the event's nature. Coarse-grained event types include 9 major categories, and fine-grained event types include 67 subcategories. As shown in Figure 1, in the sentence "进军建德，擒贼帅赵桑干。" (*Advancing to Jiande, capturing the enemy leader Zhao Sanggan*), the word "进军" (advance) could represent the event of dispatching troops to Jiande, and the word "擒" (capture) could represent the event of capturing the enemy leader Zhao Sanggan. Therefore, the trigger "进军" (advance) corresponds to the coarse-grained event type "军事" (*Military*) and the fine-grained event type "军事-备战-出兵" (*Military - Prepare_for_war - Send_troops*). The trigger "擒" (capture) corresponds to the coarse-grained event type "军事" (*Military*) and the fine-grained event type "军事-作战-俘虏" (*Military - Combat - Capture*).



Trigger	Coarse-Grained Event Type	Fine-Grained Event Type
罢	职位	职位-官位-免职
进军	军事	军事-备战-出兵
擒	军事	军事-作战-俘虏

Figure 1: Diagram of classical Chinese Historical Event Detection

[1]Event Type Schema and Dataset is released on https://github.com/NLPInBLCU/CHED2024

3 Data Description

3.1 Event Logical Schema Construction

The construction of an event type schema in a given context should meet the criteria of comprehensive coverage, precise granularity, and high accuracy. As shown in Figure 2 , the initial construction was based on word frequency statistics and semantic clustering of the translated corpus, and it was finalized through trial annotation and expert evaluation.

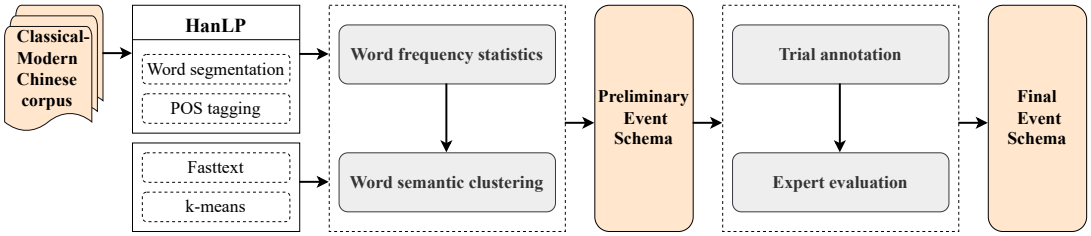


Figure 2: The diagram illustrates the complete process for constructing the event schema.

Word Frequency Statistics. High-frequency words in a text often reflect its main content and central themes, closely related to event types. We selected translated works of *The Twenty-Four Histories* from NiuTrans and used HanLP for word segmentation and part-of-speech tagging, followed by word frequency analysis. High-frequency words such as "进攻" (attack) were identified as potential event types for historical events in classical Chinese.

Semantic Clustering. Further analysis was conducted to classify words with similar semantics automatically. We used Fasttext to generate vector representations for each word and applied the k-means clustering algorithm to group words with high semantic similarity.

Trial Annotation. To evaluate event coverage in texts, we randomly selected 15 documents from the Benji and Liezhuan sections of each book in *The Twenty-Four Histories*. Based on trial annotation results, we modified and merged some event types.

Expert Evaluation. To ensure accuracy and avoid subjective bias, experts and students with backgrounds in linguistics and computer science evaluated our event types. This process led to the final event schema for CHED, including 9 major categories and 67 subcategories. Figure 3 illustrates the complete structure of the military category, which is one of the 9 major categories. The 9 major categories of events include ***Life***, ***Position***, ***Communication***, ***Movement***, ***Ritual***, ***Military***, ***Law***, ***Economy*** and ***Nature***.

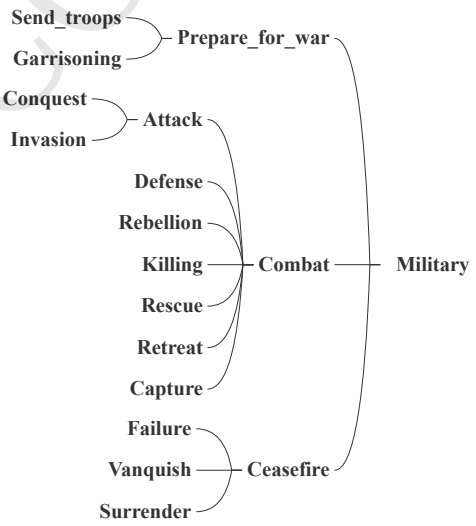


Figure 3: The diagram shows the complete hierarchical structure of the ***Military*** category, including 13 fine-grained event types.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 201-209, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.2 Dataset

Through two rounds of annotation, we refined the event type schema, resolved annotation discrepancies, and improved annotation consistency and accuracy. Both annotators were graduate students with backgrounds in linguistics and had received coursework or training in classical Chinese.

We randomly selected 2 to 3 complete volumes from the Benji and Liezhuan sections of each book in *The Twenty-Four Histories*. In total, we selected 61 volumes, covering 61 historical figures and 13,159 sentences for annotation. This effort resulted in the creation of the CHED dataset, which contains 8,122 valid event instances. The distribution of data across different historical books as well as train set, dev set and test set is shown in the table 1.

Number	History Book	Train Set	Dev Set	Test Set	Total
1	Records of the Grand Historian	521	96	98	715
2	Records of the Three Kingdoms	343	86	66	495
3	Book of Zhou	483	118	113	714
4	Book of Liang	38	7	14	59
5	Book of Wei	174	32	24	230
6	Book of Later Han	281	58	61	400
7	Book of Song	75	10	24	109
8	Book of Southern Qi	81	21	30	132
9	Book of Han	417	100	81	598
10	Book of Jin	236	44	44	324
11	Book of Chen	101	21	27	149
12	Book of Northern Qi	141	27	21	189
13	Ming History	249	42	47	338
14	History of the Southern Dynasties	238	52	68	358
15	Book of Sui	205	43	37	285
16	New History of the Five Dynasties	528	117	123	768
17	History of Song	302	49	67	418
18	Book of Northern History	227	60	49	336
19	New Book of Tang	70	25	33	128
20	Old Book of Tang	175	51	49	275
21	History of Liao	98	18	14	130
22	History of Yuan	271	68	78	417
23	History of Jin	175	32	36	243
24	Old History of the Five Dynasties	221	41	50	312
-	**Total**	**5650**	**1218**	**1254**	**8122**

Table 1: Chinese Historical Texts Dataset: Distribution Across Train Set, Dev Set, and Test Set.

4 Evaluation Metrics

This section introduces the evaluation metrics for the two subtasks and the comprehensive assessment.

4.1 Subtask 1: Trigger Identification

Participants will be evaluated on the test set. They must submit a result file containing the position information of triggers within each sentence. This subtask accounts for 40% of the overall score.

The Subset Match Score is calculated as follows:

$$\text{Subset Match Score} = \frac{\sum_{i=1}^{N} \text{SubsetMatch}(S_i, T_i)}{N} \tag{1}$$

where N is the total number of entries, S_i is the set of standard triggers for entry i, T_i is the set of submitted triggers for entry i, and SubsetMatch$(S_i, T_i) = 1$ if all elements of S_i are present in T_i, otherwise SubsetMatch$(S_i, T_i) = 0$.

The Exact Match Score is calculated as follows:

$$\text{Exact Match Score} = \frac{\sum_{i=1}^{N} \text{ExactMatch}(S_i, T_i)}{N} \tag{2}$$

where N is the total number of entries, S_i is the set of standard triggers for entry i, T_i is the set of submitted triggers for entry i, and ExactMatch$(S_i, T_i) = 1$ if S_i is equal to T_i, otherwise ExactMatch$(S_i, T_i) = 0$.

The total score is calculated as:

$$\text{Total Score} = \frac{\text{Subset Match Score} + \text{Exact Match Score}}{2} \tag{3}$$

4.2 Subtask 2: Event Type Classification

Participants must submit a result file containing each trigger and its predicted event type. Subtask 2 is evaluated at both a coarse-grained level (9 major categories) and a fine-grained level (67 subcategories). The evaluation will use both macro-average and micro-average metrics, and this subtask accounts for 60% of the overall score. For **coarse-grained evaluation** , the metrics are calculated based on 9 major categories. For **fine-grained evaluation**, the metrics are calculated based on 67 subcategories, using the same formulas as coarse-grained metrics.

Macro-Average Metrics:

$$\text{Macro Precision} = \frac{1}{L}\sum_{l=1}^{L}\frac{TP_l}{TP_l + FP_l} \tag{4}$$

$$\text{Macro Recall} = \frac{1}{L}\sum_{l=1}^{L}\frac{TP_l}{TP_l + FN_l} \tag{5}$$

$$\text{Macro F1} = \frac{1}{L}\sum_{l=1}^{L}\frac{2 \times TP_l}{2 \times TP_l + FP_l + FN_l} \tag{6}$$

Micro-Average Metrics:

$$\text{Micro Precision} = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L}(TP_l + FP_l)} \tag{7}$$

$$\text{Micro Recall} = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L}(TP_l + FN_l)} \tag{8}$$

$$\text{Micro F1} = \frac{2 \times \text{Micro Precision} \times \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}} \tag{9}$$

Total Score Calculation for Subtask 2:

$$\text{Total Score for Subtask 2} = \frac{\text{Macro F1} + \text{Micro F1}}{2} \tag{10}$$

4.3 Comprehensive Assessment

$$\text{Overall Score} = 0.4 \times \text{Task 1} + 0.3 \times \text{Coarse-grained of Task 2} + 0.3 \times \text{Fine-grained of Task 2} \tag{11}$$

CCL 2024

5 Evaluation Results

5.1 Participants and Team Scores

The evaluation attracted a diverse range of participants. A total of 28 teams registered for this evaluation, including 17 teams from academic institutions, 2 teams from commercial organizations, and 9 independent teams or individuals. In the end, 15 teams submitted their results. We released the training and validation datasets for each task on March 5th, and published the answerless test set for the evaluation task on April 5th. All participating teams submitted their result sets before April 25th. The final scores and rankings were announced on May 10th, and some teams' scores were subject to appeal. The scores of the participating teams are shown in the table 2.

Rank	Team	Task 1		Task 2 - coarse		Task 2 - grain		Total Score
		Exact	Subset	Ma-F1	Mi-F1	Ma-F1	Mi-F1	
1	TeleAI	**63.60**	88.90	83.90	**84.50**	74.30	**81.40**	79.12
2	JXNU	63.10	77.50	82.20	82.80	75.50	79.90	76.18
3	SXU	62.00	78.00	82.10	82.60	74.80	79.30	75.82
4	NJU	60.40	77.50	80.90	81.00	75.10	78.60	74.22
5	SCAU	56.90	79.40	80.90	80.70	72.30	78.10	74.11
6	CAS	58.50	77.00	78.70	81.10	70.10	77.60	73.36
7	CPIC	62.80	72.00	77.20	80.30	70.60	74.50	72.53
8	BUPT	60.80	73.60	78.10	78.60	67.70	77.30	72.48
9	XZK	55.50	77.00	77.50	77.10	68.00	72.70	70.02
10	BLCU-1	57.90	70.70	75.60	78.00	60.10	73.50	68.58
11	SEU	54.80	70.00	70.10	78.20	58.10	74.10	68.04
12	XJNU	50.20	60.30	69.90	73.90	54.90	71.20	62.59
13	ZZU	45.40	52.20	66.90	65.60	38.90	48.80	49.10
14	BIT	26.00	47.60	37.80	73.40	18.90	57.70	42.49
15	BLCU-2	0.00	0.00	46.10	51.30	18.80	45.90	28.17

Table 2: **Final Rankings and Scores of Participating Teams in the Evaluation** (Unit: %). Scores for Subtask 1: Trigger Identification, Subtask 2: Coarse-Grained Event Type Classification, and Subtask 2: Fine-Grained Event Type Classification. Please refer to Section 4 for detailed score calculation. Total Score is calculated with a weighting of 4:3:3.

5.2 Methodology and Analysis

We ultimately received system reports from six participating teams. This subsection will summarize and analyze the methods used by the top five ranked teams based on final scores in the evaluation.

TeleAI used a semi-supervised self-training method that combined large and small models. The large language model (Wang et al., 2024) improved accuracy on scarce labeled data, while the small language model (Liu et al., 2023) (Ren et al., 2021) provided flexibility and precision for specific tasks. They generated high-quality pseudo-labels from unlabeled data, initially training the model on labeled data and then creating and filtering pseudo-labels through consistency checks. This method effectively increased training data quantity and diversity, enhancing the model's performance in recognizing event triggers and classifying event types in classical Chinese texts.

JXNU used the EIGC model to jointly extract information by combining external knowledge and a global correspondence matrix. They incrementally pre-trained the BERT-Ancient-Chinese model (Wang and Ren, 2023) on approximately 970,000 classical Chinese texts to enhance its understanding of classical literature. During event extraction, the model incorporated both textual information and external semantic knowledge to improve the accuracy of event trigger and type identification. The EIGC model

used a global correspondence matrix for joint extraction of triggers and event types, employing BERT encoding to incorporate contextual information and integrating part-of-speech data(Yu et al., 2023). This approach effectively increased the accuracy of event extraction in classical Chinese texts and demonstrated broad applicability in handling complex texts .

SXU combined the pre-trained ANCIENT-BERT model (Lample et al., 2016) , optimized specifically for classical Chinese, with Conditional Random Fields (CRF) to enhance the automatic recognition of event triggers and their types in classical texts. ANCIENT-BERT effectively captures the semantic information of traditional and rare characters, while the CRF layer helps find the optimal annotation path globally. They conducted experiments comparing the performance of using ANCIENT-BERT alone versus the combination with CRF. Additionally, they explored the introduction of pointer networks and various loss functions, such as Focal Loss, for further optimization.

NJU used both small-scale and large-scale language models to automatically identify and classify historical event triggers in classical Chinese texts. They selected the small-scale model GujiBERT (Wang et al., 2023), optimized for classical Chinese, and the large-scale model Xunzi-Qwen-14b (Hu et al., 2021). The models employed sequence labeling and sequence-to-sequence evaluation paradigms, respectively, and were pre-trained and fine-tuned for the task. Experimental results showed that GujiBERT, combined with BiLSTM and CRF, outperformed the large-scale model in terms of accuracy and standardized output, especially in high-precision information extraction tasks.

SCAU employed a method based on multi-granularity contrastive learning and Gaussian distribution embeddings. By combining coarse-grained entity-level contrastive learning with fine-grained token-level contrastive learning, they enhanced both the accuracy and efficiency of event detection. They employed the BERT-base-chinese model and applied a CRF schema to refine the learning process. A Gaussian transformation network was employed to create Gaussian embeddings for tokens, assuming that the token semantic representations are distributed according to a Gaussian distribution. Non-linear activation functions were used to produce embeddings for the mean and variance of the Gaussian distribution, and contrastive loss was applied to optimize the relationships between positive and negative samples, further enhancing the model's performance.

Rank	Team Name	Model Used
1	TeleAI	Guwen-NER-RoBERTa-base
2	JXNU	EIGC
3	SXU	ANCIENT-BERT
4	NJU	GujiBERT, Xunzi-Qwen-14b
5	SCAU	BERT-base-chinese

Table 3: Models Used by Top 5 Teams

The participating teams in the CCL24-Eval Task 5 employed various methods to improve trigger identification and event type classification. The table 3 shows the models used by each team. Common approaches included the use of BERT-based models, such as ANCIENT-BERT and BERT-Ancient-Chinese, tailored for understanding classical Chinese texts, and CRF for sequence labeling tasks to identify event triggers and classify event types.

Their optimization was achieved through techniques like semi-supervised learning, incremental pretraining, and the integration of external knowledge. For instance, TeleAI employed pseudo-label generation, JXNU utilized a global correspondence matrix, and SXU combined ANCIENT-BERT with CRF to enhance data diversity, contextual understanding, and optimal annotation paths. Furthermore, NJU demonstrated the effectiveness of using both small and large language models and SCAU applied multi-granularity contrastive learning with Gaussian embeddings. These approaches leverage diverse model architectures and sophisticated learning techniques, aiming to improve both accuracy and efficiency.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 201-209, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

6 Conclusion and Future Work

This paper presents an overview of the evaluation of historical event type extraction from classical Chinese texts. Utilizing the classical Chinese Historical Event Detection dataset (CHED) and the constructed event schema, this evaluation offers a solid data foundation. The evaluation is divided into two subtasks: trigger word identification and event type classification, with the latter further divided into coarse and fine granularity levels. This detailed approach provides a comprehensive reflection of model performance.

A total of 28 teams registered for the competition, with 15 teams submitting valid results and 6 teams submitting system reports. Many participating teams used pre-trained language models (such as BERT) and optimized them through semi-supervised learning, incremental pre-training, and integration of external knowledge to adapt to the event extraction tasks in the field of classical Chinese literature.In the trigger word identification task, the best performing system achieved an Exact match score of 63.6% , while in the fine-grained and coarse-grained event type classification tasks, the best performances reached F1-scores of 84.5% and 81.4%, respectively. Overall, the challenges in event extraction from classical Chinese texts remain rooted in the syntactic and semantic complexity of the language. Enhancing models' ability to understand the deep semantics and subtle contextual nuances of classical Chinese is crucial.

In the future, we will continue to expand the resources for event analysis to provide classical Chinese data support, and continue to conduct evaluation tasks related to classical Chinese event research . We hope that the iterations of future evaluation tasks can continue to promote technological progress in the field of classical Chinese event detection, thereby facilitating the inheritance of classical cultural classics and the advancement of digital humanities.

Acknowledgements

This research project is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (24YCX167), the National Natural Science Foundation of China (62306045, 61872402), Beijing Language and Culture University's School-level Project (Special Fund for the Basic Scientific Research Business Fee of Central Universities) (18ZDJ03), Beijing Language and Culture University's Phoenix Tree Innovation Platform Project (21PT04).

References

Wei Congcong, Feng Zhenbing, Huang Shutan, Li Wei, and Shao Yanqiu. 2023. CHED: A cross-historical dataset with a logical event schema for classical Chinese event detection. In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 875–888, Harbin, China, August. Chinese Information Processing Society of China.

Jianfei Dang. 2021. Research on knowledge extraction method of chinese classics based on deep learning. Master's thesis, North University of China.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Nan Li Jiqing Sun Jiuming Ji, Jinhui Chen. 2015. Effect analysis of chinese event extraction method based on literatures. *Journal of Modern Information*, 35(12)(3-10).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 534–545. Springer.

Qian Li, Jianxin Li, Lihong Wang, Cheng Ji, Yiming Hei, Jiawei Sheng, Qingyun Sun, Shan Xue, and Pengtao Xie. 2022. Type information utilized event detection via multi-channel gnns in electrical power systems. *CoRR*, abs/2211.08168.

Shixuan Liu, Chen Peng, Chao Wang, Xiangyan Chen, and Shuangyong Song. 2023. icsberts: Optimizing pre-trained language models in intelligent customer service. In Chrisina Jayne, Danilo P. Mandic, and Richard J. Duro, editors, *International Neural Network Society Workshop on Deep Learning Innovations and Applications, INNS DLIA@IJCNN 2023, Gold Coast, Australia, 23 June 2023*, volume 222 of *Procedia Computer Science*, pages 127–136. Elsevier.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2646–2656. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Pengyu Wang and Zhichen Ren. 2023. The uncertainty-based retrieval framework for ancient chinese CWS and POS. *CoRR*, abs/2310.08496.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and Xiyu Wang. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *CoRR*, abs/2307.05354.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huihan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, Jiaxin Peng, Wenjun Zheng, Shiquan Wang, Bingkai Yang, Xuewei He, Zhuoru Jiang, Qiyi Xie, Yanhan Zhang, Zhongqiu Li, Lingling Shi, Weiwei Fu, Yin Zhang, Zilu Huang, Sishi Xiong, Yuxiang Zhang, Chao Wang, and Shuangyong Song. 2024. Telechat technical report. *CoRR*, abs/2401.03804.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale chinese legal event detection dataset. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 183–201. Association for Computational Linguistics.

Zijian Yu, Tong Zhu, and Wenliang Chen. 2023. (syntax-aware event argument extraction). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 196–207.

Liu Zhongbao, Dang Jianfei, and Zhang Zhijian. 2020. Research on automatic extraction of historical events and construction of event graph based on historical records. *Library and Information Service*, 64:116–124.

CCL 2024

System Report for CCL24-Eval Task 6: A Unified Multi-Task Learning Model for Chinese Essay Rhetoric Recognition and Component Extraction

Qin Fang, Zheng Zhang, Yifan Wang, Xian Peng*
National Engineering Research Center of Educational Big Data,
Central China Normal University, Wuhan, China
{fangqin, zhangz, wangyifan0122}@mails.ccnu.edu.cn
pengxian@ccnu.edu.cn

Abstract

In this paper, we present our system at CCL24-Eval Task 6: Chinese Essay Rhetoric Recognition and Understanding (CERRU). The CERRU task aims to identify and understand the use of rhetoric in student writing. The evaluation set three tracks to examine the recognition of rhetorical form, rhetorical content, and the extract of rhetorical components. Considering the potential correlation among the track tasks, we employ the unified multi-task learning architecture to fully incorporate the inherent interactions among the related tasks to improve the overall performance and to complete the above 3 track tasks with a single model. Specifically, the framework mainly consists of four sub-tasks: rhetorical device recognition, rhetorical form recognition, rhetorical content recognition, and rhetorical component extraction. The first three tasks are regarded as multi-label classification tasks, and the last task is regarded as an entity recognition task. The four tasks leverage potential information transfer to achieve fusion learning. Finally, the above four sub-tasks are integrated into a unified model through parameter sharing. In the final evaluation results, our system ranked fourth with a total score of 60.14, verifying the effectiveness of our approach.

Keywords: Multi-task learning , Rhetoric Recognition , Text Classification , Entity Recognition

1 Introduction

In the learning process of primary and secondary school students, rhetorical devices are not only a core component of reading comprehension and writing skills but also an indispensable element in shaping excellent literary works. Identifying and understanding the use of rhetoric in students' compositions can significantly enhance their expressive skills and guide them in producing higher-quality narratives and descriptions.

The CERRU task systematically defines the fine-grained rhetorical types found in primary and secondary school compositions, as detailed in Table 1. Evaluation in this task requires participating teams to not only identify rhetorical devices within sentences but also to conduct fine-grained classification of rhetorical form and content, and to provide the object and content of each rhetorical description. To achieve this, three tracks were established: rhetorical form recognition, rhetorical content recognition, and rhetorical component extraction. The task provided 634 training set examples, 225 validation set examples, and 5,000 test set examples.

Given the small number of training samples and the potential correlation between the track tasks, we have adopted a unified multi-task learning architecture to fully incorporate the inherent interactions among these related tasks, aiming to enhance learning efficiency and prediction accuracy (Zhang and Yang, 2021). By combining all tasks into a single model, we reduce computation and enable simultaneous completion of the three track tasks (Chen et al., 2021). Specifically, our framework consists of four

*Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 210-216, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

修辞手法 Rhetorical Device	修辞形式类型 Form Type	修辞内容类型 Content Type
比喻 Simile	明喻, 暗喻, 借喻 Explicit, Implicit, Borrowed	实在物, 动作, 抽象概念 Concrete Objects, Actions, Abstract Concepts
比拟 Analogy	名词, 动词, 形容词, 副词 Noun, Verb, Adjective, Adverb	拟人, 拟物 Personification, Objectification
夸张 Hyperbole	直接夸张, 间接夸张, 融合夸张 Direct, Indirect, Integrated	扩大夸张, 缩小夸张, 超前夸张 Amplification, Diminishment, Anticipatory
排比 Parallelism	成分排比, 句子排比 Component, Sentence	并列, 承接, 递进 Coordination, Succession, Gradation

Table 1: The fine-grained rhetoric types in form and content

main sub-tasks: rhetorical device recognition, rhetorical form recognition, rhetorical content recognition, and rhetorical component extraction. The first three tasks are treated as multi-label classification tasks, while the last task is handled as an entity recognition task. Initially, we employ a transformer-based pre-trained language model as a shared feature encoder to represent sentences. Subsequently, the four tasks leverage potential information transfer to achieve fusion learning. Finally, these sub-tasks are integrated into a unified model through parameter sharing.

Additionally, we experimented with five different mainstream transformer-based pre-trained language models as backbone networks to assess their performance on the task. Given that multi-task learning requires optimizing models for multiple objectives, we also experimented with four different loss weighting schemes to approach the optimal performance of the model.

In this paper, our contributions can be summarized in three main aspects:

(1) We propose a multi-task learning framework that integrates related subtasks, enhancing interactions between them. This approach allows us to use a single model to complete all three track tasks effectively.

(2) We compare the performance of five different pre-trained language models as backbone networks and explore four weighting methods to optimize the model's performance.

(3) Our proposed framework achieved fourth place in the CCL24-Eval Task 6 (Chinese Essay Rhetoric Recognition and Understanding, CERRU) with a total score of 60.14, demonstrating the effectiveness of our method.

2 Methodology

To fully leverage the potential correlation between each task, we employ a multi-task learning framework. This approach can be seen as an inductive knowledge transfer method that improves generalization by sharing domain information across complementary tasks. By learning multiple tasks using shared representations, insights gained from one task can aid in learning the others (Caruana, 1997). Additionally, to further enhance the model's generalization ability, we incorporate adversarial training methods during model training.

2.1 Model Architecture

Figure 1 illustrates an overview of the framework. During the training phase, each task has its corresponding objective function, and all task-specific training data are used to jointly train the model in a bottom-up order.

Shared Feature Encoder The shared feature encoder focuses on mapping the input tokens to distributed semantic representations, which are shared across four downstream subtasks. To better capture and summarize the semantics of a given sentence, we adopt a pre-trained language model based on Transformer (Vaswani et al., 2017) as the shared feature encoder and fine-tune it based on the joint loss function of multi-task learning.

Given an input sentence $X = \{x_1, x_2, \cdots, x_n\}$, we first insert special tokens $[CLS]$ at the beginning and

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 210-216, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

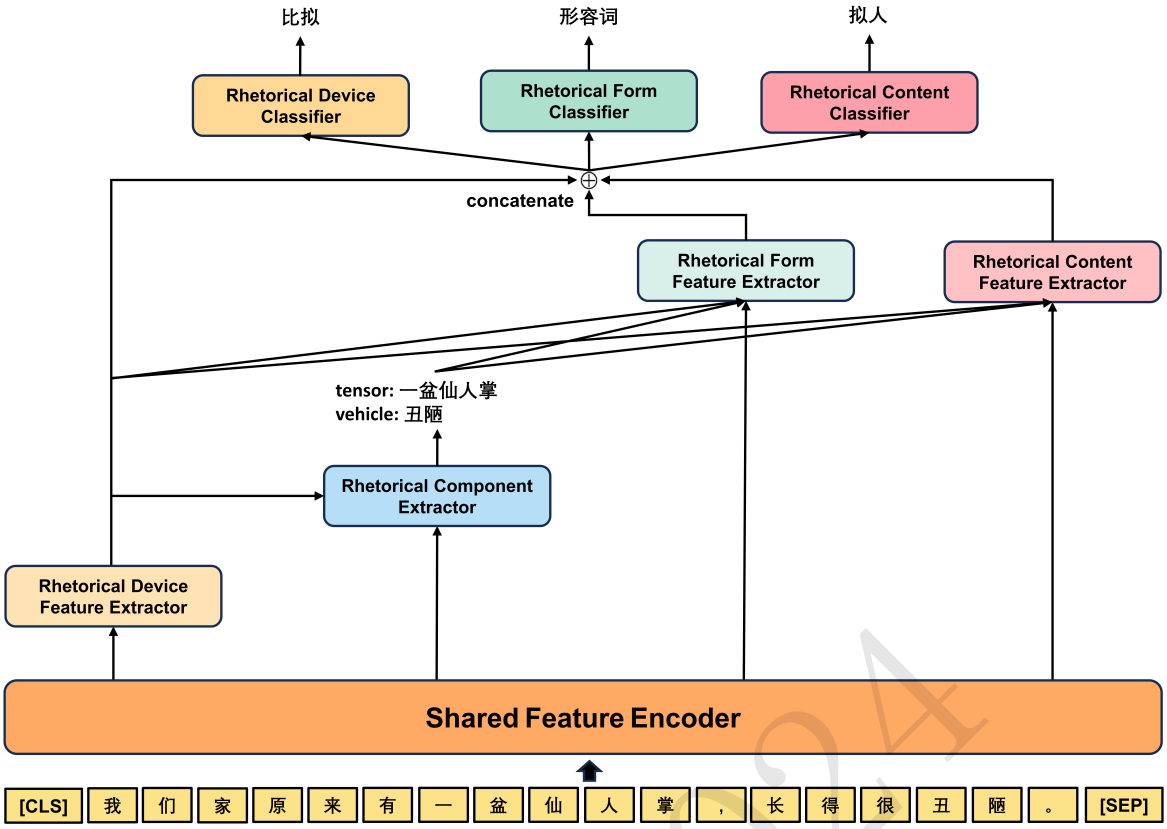


Figure 1: Model Architecture

$[SEP]$ at the end. The processed sequence is then input into the shared feature encoder. Subsequently, the shared feature encoder generates semantic representations for each token, with the output represented as $O^{encoder}$.

Rhetorical Device Feature Extractor Due to the use of an improved BERT-based pre-trained model in the shared feature encoder, which captures rich contextual information through Bidirectional Encoder Representations (BERT) (Devlin et al., 2018), we further utilize a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) network to enhance sentence representations and extract comprehensive features of rhetorical devices. BiLSTM leverages both forward and backward LSTM directions for feature extraction, capturing semantic features across contexts to obtain more comprehensive feature information. The final feature output of the rhetorical device feature extractor is denoted as F^{device}.

Rhetorical Component Extractor To enhance the extraction of rhetorical components from sentences, we concatenate rhetorical device features with semantic representations. Specifically, F^{device} is simply appended at the beginning of $O^{encoder}$, resulting in $[F^{device}, O^{encoder}]$. Although other concatenation methods were not considered, this straightforward approach effectively integrates the potential information from rhetorical device features, enhancing the module's performance. Subsequent experimental results have validated the effectiveness of this method.

For accurate entity boundary identification, we employ Efficient GlobalPointer (Su et al., 2022), a span-based entity recognition method. Efficient GlobalPointer uses two modules to detect the start and end positions of entities within a sentence, allowing for the classification of sentence subsequences as named entities. Figure 2 illustrates a matrix corresponding to two types of entities in the sentence. Compared to GlobalPointer, Efficient GlobalPointer achieves comparable performance with fewer parameters.

Rhetorical Form and Content Feature Extractor The architectures of the rhetorical form extractor and the rhetorical content extractor are identical. The specific process begins with local context

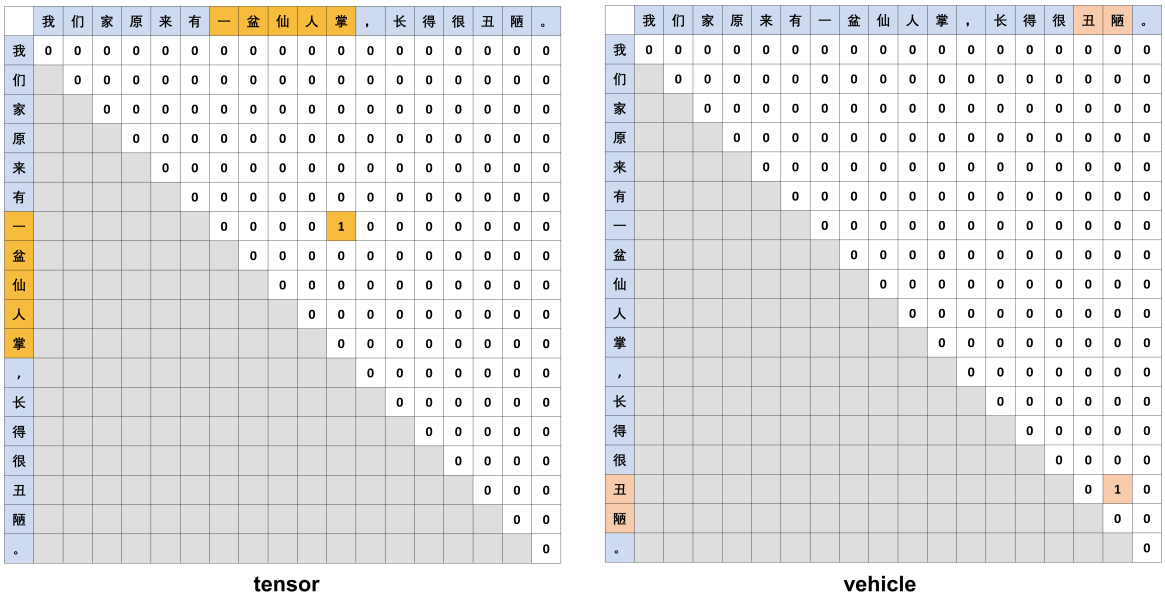


Figure 2: Entity recognition decoding structure based on GlobalPointer

enhancement on the contextualized word representations based on the entity information output by the rhetorical component extractor. Subsequently, rhetorical device features are concatenated to fuse with the enhanced representations. Feature extraction is then performed using multi-head self-attention and BiLSTM to capture comprehensive feature information. Finally, each feature extractor outputs F^{form}(for rhetorical form) and $F^{content}$(for rhetorical content).

Classifier The rhetorical device classifier, rhetorical form classifier, and rhetorical content classifier share the same architecture. Recognizing the correlation between rhetorical devices, forms, and content, we concatenate the outputs of the three rhetorical feature extraction modules, denoted as $[F^{device}, F^{form}, F^{content}]$. Initially, these inputs are cross-fused and feature-extracted using BiLSTM. Subsequently, the classification results are produced through a fully connected network.

2.2 Adversarial Training

In the field of natural language processing, adversarial training is employed as a regularization method to improve a model's generalization performance. We incorporated FGM (Miyato et al., 2016) to add adversarial training to our model. FGM introduces an adversarial attack in the direction opposite to the gradient during backpropagation in the embedding layer, thereby inducing adversarial training effects. This training method not only enhances the model's generalization ability but also improves its robustness.

2.3 Loss Function

Overall, the input sentence X is encoded by the shared feature encoder. The contextual outputs are then used to compute four tasks with task-specific labels Y_i for i = 1, 2, 3, 4. We jointly optimize the integrated loss during training as follows:

$$\mathcal{L}_{tot}(\mathbf{X}, \mathbf{Y}_{1:4}) = \sum_{i=1}^{4} \lambda_i \mathcal{L}_i(\mathbf{X}, \mathbf{Y}_i) \tag{1}$$

where $\mathcal{L}_i$ represents the cross-entropy loss for each task, and λ_i is the weighting factors that balance the contribution of each task's loss to the overall loss. The overall model loss can be heavily influenced by one task due to the varying loss magnitudes across different tasks, causing other tasks to have less impact on the learning process of shared network layers. To mitigate this, it is crucial to choose appropriate

weights to balance the training of each task, ensuring all tasks contribute effectively to the model's improvement.

In our experiments, we used four weighting schemes to determine the weights suitable for our model. These include equal weighting(λ_i=1), weight uncertainty (Kendall et al., 2017), random loss weight (Lin et al., 2021) and dynamic weight average (Liu et al., 2019). Readers can refer to the earlier citations for implementation details of these methods.

3 Experiments

3.1 Experiment Setting

During training, we use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2e-5 and a batch size of 16. The maximum sequence length is set to 256, and the maximum number of epochs is 30. The random seed is set to 1018. Additionally, we employ a learning rate warm-up strategy where the number of warm-up steps is 10% of the total number of training steps.

3.2 Experimental Setup

For the shared feature encoder, we used five mainstream transformer-based pre-trained language models as the backbone network. These include Chinese-MacBERT-Large (Cui et al., 2020), Chinese-RoBERTa-WWM-Ext-Large (Cui et al., 2019), StructBERT-Large-Zh (Wang et al., 2019), Erlangshen-DeBERTa-v2-710M-Chinese (Zhang et al., 2022), and ERNIE-3.0-Xbase-Zh (Sun et al., 2021). For each backbone network, we experimented with four weighting methods as described in Section 2.3. Their performance on the validation set is shown in Table 2.

Backbone Network	Weighting Scheme	Score			
		Track 1	Track 2	Track 3	Total
Chinese-MacBERT-Large	Equal Weights	**48.08**	51.59	63.65	54.44
	Uncert. Weights	46.28	52.51	63.44	54.08
	Random Loss Weight	46.67	**53.85**	**63.73**	**54.75**
	Dynamic Weight Average	43.92	50.31	63.51	52.58
chinese-roberta-wwm-ext-large	Equal Weights	46.09	52.73	63.86	54.23
	Uncert. Weights	46.28	52.24	**65.35**	54.62
	Random Loss Weight	41.50	50.17	63.63	51.77
	Dynamic Weight Average	**46.34**	**53.98**	63.66	**54.66**
StructBERT-Large-Zh	Equal Weights	**49.44**	**55.10**	62.24	**55.59**
	Uncert. Weights	45.66	51.49	62.51	53.22
	Random Loss Weight	46.44	52.39	64.43	54.42
	Dynamic Weight Average	45.45	53.19	**64.50**	54.38
Erlangshen-DeBERTa-v2-710M	Equal Weights	43.65	50.99	66.90	53.85
	Uncert. Weights	42.30	48.53	66.09	52.31
	Random Loss Weight	**48.01**	**52.61**	67.58	**56.06**
	Dynamic Weight Average	41.63	50.01	**68.07**	53.23
ERNIE-3.0-Xbase-Zh	Equal Weights	43.95	53.17	66.36	54.50
	Uncert. Weights	45.95	**54.78**	**67.40**	**56.04**
	Random Loss Weight	**46.14**	53.27	66.70	55.37
	Dynamic Weight Average	45.98	54.15	66.83	55.65

Table 2: The performance of each backbone network with different weighting methods on the validation dataset. The best performing combination of backbone network and weighting is highlighted in bold. The top validation scores for each metric are annotated with boxes.

CCL 2024

3.3 Experimental Results

We selected the optimal weighting method for each pre-trained model based on their performance on the validation set, identifying the top five performing models. Subsequently, during the last five epochs of training, we applied the Stochastic Weight Averaging (SWA) (Izmailov et al., 2018) method to these models for evaluation on the test set. For the final evaluation, we employed model ensemble voting. Our approach achieved a fourth-place ranking in the final evaluation results. The scores of the top five teams and baseline are presented in Table 3.

Team	Track 1	Track 2	Track 3	Score
Team1	61.30	62.29	75.28	66.29
Team2	59.20	60.92	77.96	66.03
Team3	53.77	60.15	68.26	60.72
Our team	50.86	55.81	73.75	60.14
Team5	51.48	55.11	69.51	58.70
Baseline	45.66	56.89	20.85	41.13

Table 3: The final evaluation results of the top five teams and Baseline

3.4 Results Analysis

Based on the evaluation results, our team's overall performance is close to the third position. Specifically, our performance across different tracks is as follows: on Track 3, our results significantly exceed the baseline (52.9 points) and the third-place score (5.49 points), whereas on Track 1, our performance is moderate, just slightly above the baseline (5.2 points), and on Track 2, our performance is below the baseline (1.08 points).

The experimental results demonstrate that our adopted multi-task learning model architecture achieves better generalization capability through shared representations across tasks. In particular, the model shows significantly enhanced entity recognition capabilities on Track 3.

However, our performance in classification tasks on Track 1 and Track 2 is relatively average. This could be attributed to the multi-task learning model needing to concurrently optimize loss functions across multiple sub-tasks, a design that prevents multi-task learning from achieving the optimal performance on each sub-task as in single-task learning. Nevertheless, the advantage of multi-task learning lies in its ability to use fewer model parameters to accomplish learning and inference efficiently for multiple sub-tasks with a single model.

In conclusion, our experimental results highlight both the potential and limitations of multi-task learning in terms of cross-task generalization capability.

4 Conclusion

In this paper, we propose a unified multi-task learning framework for CCL24-Eval Task 6 (CERRU), to enhance feature fusion and interaction among subtasks, and achieve a single model capable of completing all subtasks. We experimented with various pre-trained language models and weighting methods, and further improved our experimental results through model voting. Our experiments demonstrate that our proposed approach achieves good results in this evaluation.

However, there are still several shortcomings in this system. In the future, we plan to enhance the model's performance through data augmentation and domain-specific pre-training. Additionally, we intend to explore more suitable weighting methods for this evaluation.

References

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

Chen, S., Zhang, Y., and Yang, Q. (2021). Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., and Hu, G. (2019). Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*.

Kendall, A., Gal, Y., and Cipolla, R. (2017). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Lin, B., Ye, F., Zhang, Y., and Tsang, I. W.-H. (2021). Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *Trans. Mach. Learn. Res.*, 2022.

Liu, S., Johns, E., and Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880.

Loshchilov, I. and Hutter, F. (2017). Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Miyato, T., Dai, A. M., and Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Su, J., Murtadha, A., Pan, S., Hou, J., Sun, J., Huang, W., Wen, B., and Liu, Y. (2022). Global pointer: Novel efficient span-based approach for named entity recognition. *ArXiv*, abs/2208.03054.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L., and Si, L. (2019). Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Zhang, J., Gan, R., Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., and Chen, C. (2022). Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Zhang, Y. and Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.

CCL 2024

CCL24-Eval任务6系统报告：中小学作文修辞识别与理解

赵亮
洪恩完美未来科技有限公司
zhaoliang802652@ihuman.com
余浩
洪恩完美未来科技有限公司
yuhao602137@pwrd.com

武伟轩
洪恩完美未来科技有限公司
wuweixuan801519@ihuman.com
鲁文斌
洪恩完美未来科技有限公司
luwenbin@ihuman.com

摘要

本技术报告是对2024CCL评测任务(中小学作文修辞识别与理解评测)的一种解决方案。在中小学生的学习过程中，修辞手法不仅是阅读理解和写作技巧的核心组成部分，同时也是塑造优秀文学作品的不可或缺的元素。识别并理解学生作文中的修辞使用，可以帮助学生提高作文表达能力，指导学生更高质量的叙述和描写。对修辞的识别目前属于自然理解领域比较困难的任务，因为需要用到人类领域的大量先验知识，而且很多时候不同的修辞之间的边界还是模糊的。我们通过lora技术直接微调基于qwen-chat-7B的大语言预训练模型，来进行修辞类别的识别。

我们的主要创新技术点为：基于相同的输入输出数据来构造多条训练数据提升算法表现；分级分层来进行修辞的判断，先进行大的修辞类别判断，再把大的修辞类别做为输入对修辞的子类别进行判断；针对修辞成分抽取的任务，直接输出对应的结果文本，再对应回原文本进行位置检索，而不是直接输出索引下标。

关键词： 大语言预训练模型；修辞判断；文本分类

System Report for CCL24-Eval Task 6: Identification and Understanding of Rhetoric in Primary and Secondary School Essays

Liang zhao
ihuman
zhaoliang802652@ihuman.com
WeiXuan Wu
ihuman
wuweixuan801519@ihuman.com
Hao Yu
ihuman
yuhao602137@pwrd.com
WenBin Lu
ihuman
luwenbin@ihuman.com

Abstract

This technical report presents a solution to the 2024 CCL evaluation task (Rhetorical Device Recognition and Understanding in Primary and Secondary School Essays). In the learning process of primary and secondary school students, rhetorical devices are not only core components of reading comprehension and writing skills but also indispensable elements in shaping excellent literary works. Recognizing and understanding the use of rhetoric in student essays can help students improve their writing skills and guide them to produce higher-quality narratives and descriptions. Identifying rhetorical devices is currently a challenging task in the field of natural understanding because it requires extensive human domain knowledge, and often the boundaries between different rhetorical devices are blurred. We utilize LoRA technology to fine-tune

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

a large pre-trained language model based on Qwen-Chat-7B to recognize categories of rhetorical devices.

Our main innovative technical points are: constructing multiple training data instances based on the same input-output data to improve algorithm performance; using a hierarchical approach to judge rhetorical devices by first determining broad categories of rhetoric and then using these broad categories as input to determine subcategories of rhetoric; and for the task of extracting rhetorical components, directly outputting the corresponding result text and then mapping it back to the original text for position retrieval, rather than directly outputting index positions.

Keywords: Large Pre-trained Language Model , Rhetorical Judgment , Text Classification

1 引言

1.1 引言

本报告主要是提出一种帮助中小学教师进行修辞识别的大模型自动识别算法。熟练使用修辞是中小学生写作的一项重要而基本的要求，所以能正确高效的对中小学生作文的修辞手法进行准确的识别，可以为语文老师提供十分重要的帮助。本报告做的三个任务分别是修辞形式分类，修辞内容分类，和修辞成分抽取。

我们首先会进行修辞的大类别识别，主要是中小学常见的四个修辞类别：比喻，拟人，排比，夸张。然后我们会对修辞形式类型进行细粒度的识别，比喻包括：明喻、暗喻、借喻；比拟包括：名词、动词、形容词、副词；夸张包括:直接夸张、间接夸张、融合夸张；排比包括：句子排比、成分排比。再然后我们会对修辞内容类型进行细粒度的识别，比喻包括：实在物、动作、抽象概念；比拟包括：拟人、拟物；夸张包括：扩大夸张、缩小夸张、超前夸张；排比包括：并列、承接、递进。最后，进行修辞成分的抽取，修辞成分由连接词、连接对象、连接内容三个部分组成。我们通过利用大模型算法对这三个任务进行训练，得到了最终适合的结果。

1.2 主要困难

比喻和比拟的近似。很多时候比喻和比拟之间的界限是十分模糊的。比如说图1中的多个例子能准确区分是比喻还是比拟实际上是比较困难的。

夸张的判断。很多时候一个句子是否是夸张需要大量结合具体知识甚至很多时候需要结合具体说话场景，夸张的判断对语境的理解和要求是比较高的，如图2这里例举了一些比较难判断的具体例子。

修辞成分的抽取。修辞成分给的输出直接是对应的索引数值，直接从输入的文本，转成输出的索引是非常抽象的。但是如果输出的是文本内容，则即使只差一个字符也无法对应到原来的文本内容中去，所以修辞抽取这个任务的最终输出的方式也是一个比较困难的问题。

1.3 数据集分析

全部数据集包括训练集、验证集、测试集，每个集合条数的分布可以参照表1。训练集和测试集中大的修辞类别和不同的子类别所占的比例可以参照表2和表3。可以看出不同修辞类别尤其是子类别之间的数量相差是十分巨大的，尤其是“排比”和“夸张”两个类别中很多子类别的数目是十分的稀疏的，这种分布特性也非常增大了分类的难度。

为了进行较为充分的实验比对，本文主要实验数据为线下测试集的统计，线上效果最终为任务1：53.77，任务2：60.15，任务3：68.26。

2 技术方案

我们的整体技术方案是使用lora(Hu et al. [2021])方法对qwen-chat-7B(Bai et al. [2023])大语言模型进行针对性的微调，lora方法已经集成到peft库中(Han et al. [2024])。我们做3个任务使用的是同一个语言大模型，实际上包括了4个子任务：修辞大类别分类，修辞形式分类，修辞内容分类，以及修辞成分抽取。

(c) 2024 中国中文信息学会计算语言学专业委员会

```
{
    "id": 1229,
    "conversations": [
      {
        "from": "user",
        "value": "给你一个句子，```她个子高挑，是班上最高的女生，却又很瘦，远看就像一根纤细的竹竿。```已知使用了“比拟”的修辞手法,请你对这个句子进行判断修辞具体的“内容类型”
      },
      {
        "from": "assistant",
        "value": "拟物"
      }
    ]
  },
  {
    "id": 1265,
    "conversations": [
      {
        "from": "user",
        "value": "请找到下面句子中全部的修辞手法，```奋斗在一线的，大多是青年，你们总要接过先辈们手中的火炬，进继发展先辈们的精神与品格。```"
      },
      {
        "from": "assistant",
        "value": "['比喻']"
      }
    ]
  }
```


Figure 1: 相近的比喻和比拟的示例


```
{
    "id": 1360,
    "conversations": [
      {
        "from": "user",
        "value": "请找到下面句子中全部的修辞手法，```外祖父是方圆几十里最出名的人物，二胡声总能把鸟儿们引来与他一同伴奏。```"
      },
      {
        "from": "assistant",
        "value": "['夸张']"
      }
    ]
  },
  {
    "id": 3424,
    "conversations": [
      {
        "from": "user",
        "value": "给你一个句子，请你找到句子中所有的修辞手法。给你的句子是```我怕得是欢喜的回到家后，人都不在，好像只要我回家慢一点，就可以让家中有人等待着我。```"
      },
      {
        "from": "assistant",
        "value": "['夸张']"
      }
    ]
  }
```


Figure 2: 比较难判断的夸张的示例

CCL 2024

数据集	数目
训练集	约700
验证集	约300
盲测集	约5000

Table 1: 数据集统计

修辞形式	数目
明喻	193
暗喻	129
借喻	76
名词	12
动词	131
形容词	20
副词	7
直接夸张	56
间接夸张	28
融合夸张	24
成分排比	67
句子排比	48

Table 2: 修辞形式的统计分布

修辞内容	数目
实在物	233
动作	33
抽象概念	131
拟人	145
拟物	19
扩大夸张	98
缩小夸张	9
并列	84
承接	5
递进	26

Table 3: 修辞内容的统计分布

2.1 预处理阶段

针对“修辞形式”分类和“修辞内容”分类，首先我们把数据拆成两层，第一层针对大的修辞进行分类，第二层则针对两个子任务进行分类。在对第二层进行分类的时候，把第一层分类的输出直接当作输入，举个例子，输入是：修辞的“形式类型”指的是对一个修辞手法进行更细粒度的修辞分类，给你一个句子，已知使用了“比拟”的修辞手法,请你基于这个修辞手法，对这个句子进行判断修辞具体的“形式类型”，给你的句子是它枝干虬劲有力，高大挺拔，宛如一位镇守在诸陪葬坑前的披坚执锐的秦朝战士。对应的输出为：“名词”。

“修辞成分抽取”则把索引转化成对应的文本进行生成来增加预测的准确率和稳定性，举个例子，输入是：一个句子是“白衣天使把我们带进了真正的春天。”，请判断具体的“修辞成分”，分别是“描写对象、连接词、描写内容”。对应的输出为：[None, None, “真正的春天”]。

2.2 一级修辞大类别的分类

直接利用lora的技术finetune语句的输出进行分类，不再使用额外的分类网络，从而使学习更加集中。先构建一个prompt，在相应位置上填充对应的句子，输出对应的修辞大分类的序列，可以为零个、一个、或者多个，如果为多个大分类则进行全排列有多个相同训练权重的对应输出。

为了增强算法的稳定性，同时进一步优化算法的泛化能力，采用多种prompt开头的方法对其进行针对性的优化，即一个相同的输入数据使用多种prompt来构建多个输入数据来增强模型对应的综合能力。原始输入为“ 请找到下面句子中全部的修辞手法，火红的夕阳下，那只金色的翅膀直直地伸向天空，犹如一块金色的墓碑，这是老狼紫岚的墓。”，改变后的输入为一个集合{“ 请找到下面句子中全部的修辞手法，火红的夕阳下，那只金色的翅膀直直地伸向天空，犹如一块金色的墓碑，这是老狼紫岚的墓。”, “ 给你一个句子，请你找到句子中所有的修辞手法。给你的句子是火红的夕阳下，那只金色的翅膀直直地伸向天空，犹如一块金色的墓碑，这是老狼紫岚的墓。”, ... }，然后对输出结果进行判断。做预测的时候则随机选择一个prompt进行输出。

2.3 基于一级修辞分类的二级修辞子类别的分类

和一级修辞大类别的分类一样，直接利用lora的技术finetune语句的输出进行分类，不再使

(c) 2024 中国中文信息学会计算语言学专业委员会

用额外的分类网络。同样如同“预处理阶段”所说的那样，把第一层分类的输出直接当作输入来辅助二级修辞子类别的分类。

为了增强算法的稳定性，也同样采用多种prompt开头的方法对其进行针对性的优化。并且两个任务“修辞形式”和“修辞内容”，都会把第一阶段的一级修辞大类别输出的结果当作前置条件做为输入，当在训练阶段是直接把真实的一级修辞大类别做为输入这样会使得相对的训练误差降低，然后在预测阶段由于没有准确的一级于此分类就把第一阶段的预测结果来当作输出。通过不同参数和prompt的组合在验证集上的反复调试，我们最终得到了相对最优的训练结果。

2.4 修辞成分抽取的文本和索引之间的转化

修辞成分抽取的任务，输入部分还是会针对同一个输入数据利用多个prompt来构造多条有差异性的输入数据，输出部分我们把索引在原文中进行了定位得到了原本的输出内容。然后训练数据变成了文本输入和文本输出，从而降低了训练的难度。值得注意的是，在预测阶段，预测输出的文本可能与原文不完全一致，我们采用了子串对比编辑距离的方式，只要输出文本和输入文本的子串之间的编辑距离小于一定的值，则认为完全一致，直接把相应的子串当作结果。

最后我们再根据确定的子串在原文中的索引，来得到最终的输出结果。在实验中我们同样尝试了不同的最大编辑距离阈值，最终我们选择了2做为最大可接受的编辑距离。具体的结果可以参见实验篇章。

3 实验结果

针对第二章的技术方案我们在官方提供的数据集上进行了充分的实验，具体的实验结果可以参照表4和表5。

如表4所示，在修辞成分抽取的任务中，我们也实验了多个不同最大编辑距离的阈值来进行比较。最终我们确定了把最大编辑距离设置为2，可以得到相对最优的输出结果。

如表5所示，最优模型为复合了多个句子做为输入数据进行训练得到的，我们针对线上数据集的预测也和线下的预测方式和相关参数是保持相应一致的。最终我们确立每组输入数据配合3个promot进行组合，可以训练得到相对最优模型。

最大编辑距离的阈值 \ 输入的prompt数目	1	2	3	4	5	6
1	0.705	0.711	0.713	0.720	0.709	0.711
2	0.741	0.747	0.753	0.749	0.748	0.752
3	0.732	0.731	0.736	0.729	0.719	0.721
4	0.730	0.729	0.741	0.719	0.731	0.732

Table 4: 修辞成分抽取实验结果统计

任务 \ 输入的prompt数目	1	2	3	4	5	6
task1	0.559	0.562	0.590	0.588	0.581	0.583
task2	0.551	0.558	0.571	0.572	0.566	0.569

Table 5: 修辞形式分类和修辞内容分类实验结果统计

4 总结

我们通过实验证明了针对中小学作文修辞识别和理解任务的几个方法的有效性：对单条输入数据使用不同prompt扩充为多条训练数据来构造训练集；分级分层来进行修辞判断的方法，即先进行大的修辞类别判断，再把大的修辞类别做为输入对修辞的子类别进行判断，来优化算法输出；针对修辞成分抽取的任务，直接输出对应的结果文本，再对应回原文本进行位置检索，同时使用控制最大编辑距离的方法来检查位置的合理性。

(c) 2024 中国中文信息学会计算语言学专业委员会

在中小学作文修辞识别和理解这个任务中后续还有很大的算法效果提升空间，希望本报告的方法能帮助更多的相关研究人员进行相关算法的优化和改进。

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

CCL 2024

System Report for CCL24-Eval Task 6: Essay Rhetoric Recognition and Understanding Using Synthetic Data and Model Ensemble Enhanced Large Language Models

Jinwang Song, Hongying Zan*, Kunli Zhang
School of Computer and Artificial Intelligence, Zhengzhou University
Zhengzhou, 450001
jwsong@gs.zzu.edu.cn,iehyzan@zzu.edu.cn,ieklzhang@zzu.edu.cn

Abstract

Natural language processing technology has been widely applied in the field of education. Essay writing serves as a crucial method for evaluating students' language skills and logical thinking abilities. Rhetoric, an essential component of essay, is also a key reference for assessing writing quality. In the era of large language models (LLMs), applying LLMs to the tasks of automatic classification and extraction of rhetorical devices is of significant importance. In this paper, we fine-tune LLMs with specific instructions to adapt them for the tasks of recognizing and extracting rhetorical devices in essays. To further enhance the performance of LLMs, we experimented with multi-task fine-tuning and expanded the training dataset through synthetic data. Additionally, we explored a model ensemble approach based on label re-inference. Our method achieved a score of 66.29 in Task 6 of the CCL 2024 Eval, Chinese Essay Rhetoric Recognition and Understanding(CERRU), securing the first position.

Keywords: Rhetoric Recognition, Large Language Models, Model Ensemble, Synthetic Data

1 Introduction

Essay writing is a crucial means of assessing students' language proficiency and logical thinking skills. The development and application of natural language processing (NLP) technologies have significantly advanced the field of education. Typically, grading essays demands substantial time and effort from teachers. By applying automation technologies to this process, we can alleviate the workload on teachers, allowing them to focus more on instruction and student guidance. Rhetorical devices are an essential component of essays, making it necessary to automate the extraction and classification of these devices as a key dimension in the automated assessment of writing quality.

Previous work in the fields of machine learning and deep learning has extensively explored the identification and classification of rhetorical devices in writing. For instance, (Xiaoxi et al., 2018) used convolutional neural networks and support vector machines to identify metaphors in both Chinese and English datasets. (Hu et al., 2017) employed sequential models to recognize metaphors in text, while (Liu et al., 2018) adopted a multi-task learning approach to classify rhetorical sentences and extract their rhetorical components. Additionally, (Li and Li, 2022; Iqbal et al., 2023) utilized BERT models to identify metaphors in compositions by overseas Chinese students.

The recognition and extraction of rhetorical devices can be fundamentally categorized as tasks of text classification and entity recognition, both of which have been extensively studied. The introduction of the GPT series (Radford et al., ; Radford et al., 2019; Brown et al., 2020) has sparked significant interest in LLMs within the NLP community, marking the advent of the era of large language models (Zhao et al., 2023). Following the release of the LLaMA series (Touvron et al., 2023), the open-source community for LLMs has flourished. In this new era, some studies have utilized the in-context learning capabilities of large models for text classification (Sun et al., 2023). Furthermore, (Wang et al., 2023) demonstrated the effectiveness of supervised instruction fine-tuning of large language models for information extraction tasks.

*Corresponding author

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 223-231, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Against this backdrop, we explored the use of generative large language models for rhetorical recognition and extraction in the CCL24-Eval Task 6 Chinese Essay Rhetoric Recognition and Understanding(CERRU). For this task, we processed the dataset into fine-tuning instructions and applied parameter-efficient fine-tuning methods to the Yi(Young et al., 2024) and Qwen1.5(Bai et al., 2024) models with supervised instruction fine-tuning. To further enhance model performance, we incorporated multi-task learning methods and augmented the training set with synthetic data generated by LLMs. Finally, we developed a model ensemble approach involving re-inference for Hierarchical rhetorical classification tasks. Our method achieved a final score of 66.29 in CCL 2024 Eval Task 6 CERRU, ranking 1st.

2 Chinese Essay Rhetoric Recognition and Understanding

CCL 2024 Eval Task 6: Chinese Essay Rhetoric Recognition and Understanding includes three tracks:

Track 1: This track classifies rhetorical devices in each sentence at a coarse-grained level into five categories: metaphor, simile, hyperbole, parallelism, and no rhetoric. Additionally, each rhetorical category is further classified into subcategories based on form, resulting in a total of 4 coarse-grained categories and 12 fine-grained subcategories.

Track 2: Similar to Track 1, this track classifies rhetorical devices in each sentence at a coarse-grained level into the same five categories. However, the fine-grained classification is based on content, resulting in 4 coarse-grained categories and 11 fine-grained subcategories.

Track 3: This track focuses on identifying rhetorical components within sentences, specifically conjunction, tenor, and vehicle.

Overall, Tracks 1 and 2 fall under hierarchical text classification tasks, while Track 3 is an entity extraction task.

3 Methodology

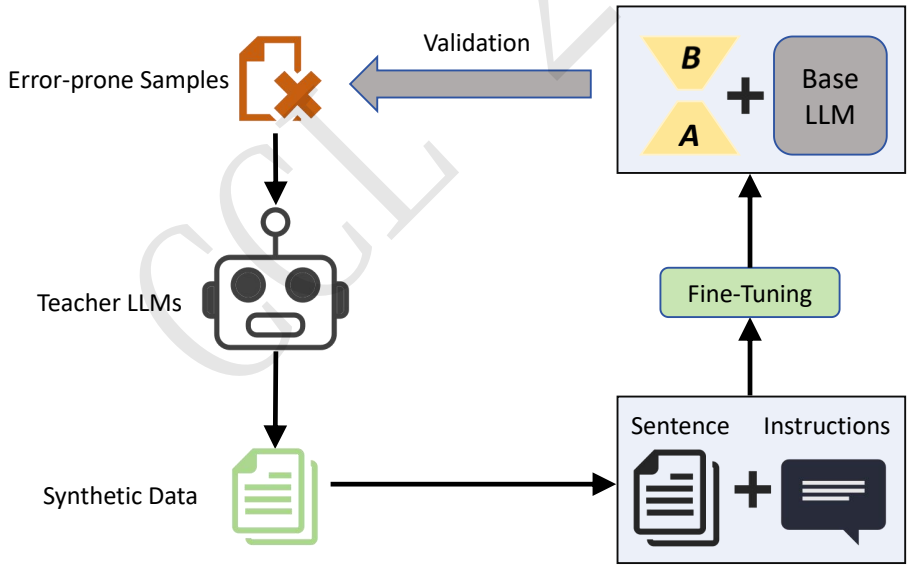


Figure 1: The process of generating synthetic data

3.1 Parameter-Efficient Instruction Fine-Tuning

Despite the evaluation task encompassing both classification and entity recognition, these tasks can be unified under an instruction-based format within the generative fine-tuning framework for LLMs. To fine-tune LLMs with limited hardware resources, we employed the LoRA (Hu et al., 2021) method for instruction fine-tuning of the base LLM.

3.2 Multi-Task Learning

We believe that in the rhetorical classification task, not only should the LLM be informed of the rhetorical category to which a sentence belongs, but it should also identify the specific entities within the sentence that determine this category. This approach aids the LLM in better understanding and analyzing rhetorical categories within sentences. The same principle applies to the task of rhetorical entity extraction.

It is important to note that the datasets for the three tracks in the evaluation task contain identical text, differing only in their respective annotations. Therefore, implementing multi-task learning in our experiments was straightforward: we simply combined the instruction datasets from the three tracks and performed fine-tuning on this mixed dataset.

3.3 Synthetic Data

The limited number of annotated training samples provided for the evaluation task constrained further improvements in model performance. Inspired by the LLM2LLM method (Lee et al., 2024), we recorded error-prone samples in track 1&2 from the validation set during the fine-tuning process. As shown in Figure 1, we then used a more powerful LLM as a teacher model to generate synthetic data based on these error-prone samples.

3.4 Model Ensemble Based on Label Re-Inference

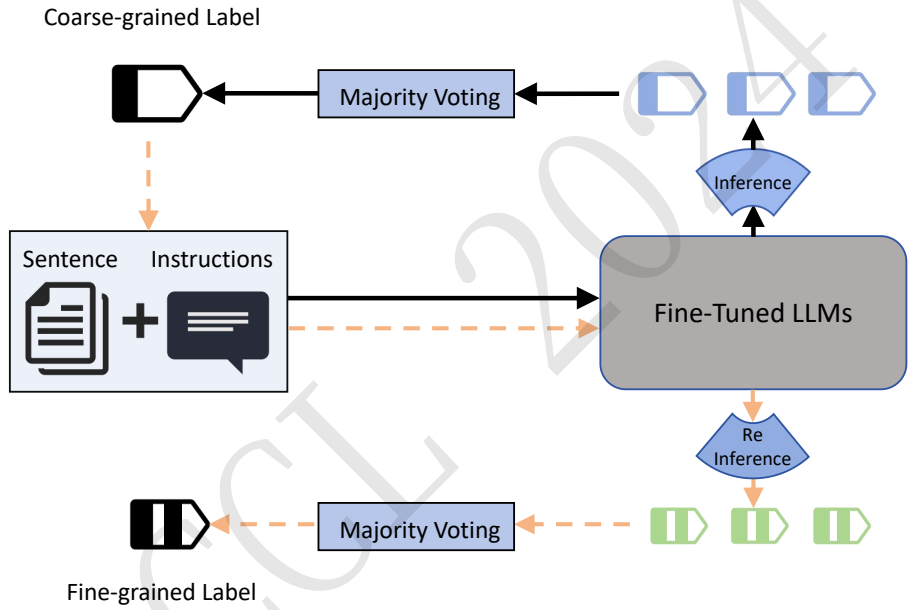


Figure 2: Model ensemble based on label re-inference

To further enhance model performance on track 1&2, we explored a model ensemble approach for the task of classifying coarse-grained and fine-grained labels using LLMs. This process is shown in Figure 2. Assume there are K fine-tuned LLMs, with the parameters of the k-th LLM denoted as θ_k. For a given sentence x, we add instructions to form $Instruction(x)$, which is then input into θ_k for inference, yielding a coarse-grained label $y_k^{(h)}$(**H**igh-level label) and a fine-grained label $y_k^{(l)}$(**L**ow-level label), along with their respective possible label sets $\mathcal{C}^{(h)}$ and $\mathcal{C}^{(l)}$:

$$(y_k^{(h)}, y_k^{(l)}) = \underset{(c^{(h)},c^{(l)})\in\mathcal{C}^{(h)}\times\mathcal{C}^{(l)}}{\arg\max} P\left(y_k^{(h)} = c^{(h)}, y_k^{(l)} = c^{(l)} \mid Instruction(x)\ , \theta_k\right) \tag{1}$$

Next, we perform majority voting on these coarse-grained label results:

$$y_{ens}^{(h)} = \arg\max\left(\sum_{k=1}^{K} \delta(y_k^{(h)} = c_j^{(h)})\right) \tag{2}$$

where $j = \{1, 2, ..., |\mathcal{C}^{(h)}|\}$, and δ is the indicator function, which equals 1 when $y_k^{(h)} = c_j^{(h)}$ and 0 otherwise. Given that generative LLMs predict the next token based on the given sequence, and that the LLM has learned the constraints between coarse-grained and fine-grained labels during the fine-tuning phase, there is no need for additional fine-tuning. Provided $y_{ens}^{(h)}$ is not null, we append $y_{ens}^{(h)}$ to the instruction and re-input it into the LLM for re-inference under the constraint of the coarse-grained label:

$$y_k^{(l\_re)} = \underset{c^{(l)} \in \mathcal{C}^{(l)}(y_{ens}^{(h)})}{\arg\max} P\left(y_k^{(l\_re)} = c^{(l)} \mid Instruction(x) + y_{ens}^{(h)}, \theta_k\right) \tag{3}$$

We then conduct majority voting on the re-inferred fine-grained labels:

$$y_{ens}^{(l)} = \arg\max\left(\sum_{k=1}^{K} \delta(y_k^{(l\_re)} = c_j^{(l)})\right) \tag{4}$$

where $j = \{1, 2, ..., |\mathcal{C}^{(l)}|\}$. The resulting $(y_{ens}^{(h)}, y_{ens}^{(l)})$ constitutes the final classification result for the sentence x.

4 Experiments

4.1 Dataset

The experimental dataset is derived from CCL 2024 Task 6 and includes three tracks. Each track consists of 634 samples in the training set, 225 samples in the validation set, and 5000 samples in the test set.

We processed the datasets into instruction formats, examples of the fine-tuning data can be found in the appendix. For training samples with multiple answers, we separated them using "\n".

4.2 Models

In our experiments, we used several LLMs for instruction fine-tuning: Yi-6B-Base, Qwen1.5-7B-Base, Qwen1.5-14B-Base, and Qwen1.5-32B-Base. We observed that, in most cases, larger LLMs tend to achieve better performance.

For generating synthetic data, we employed Qwen-max-0403 and Qwen-max-0428 as teacher models, accessed via API. The prompt used for calling the teacher models can be found in the appendix.

4.3 Settings

We used the AdamW optimizer with β_1 and β_2 set to (0.9, 0.999) and a weight decay of 1e-2. The initial learning rate was 3e-5, with cosine annealing applied to decay the learning rate to 6e-6 after 600 steps. The batch size was 8, and gradient checkpointing was enabled. The gradient norm was clipped to a maximum of 1.0.

For the LoRA hyperparameters, we used $r = 72$ and $\alpha = 612$, which is equivalent to $\alpha = 72$ with rsLoRA enabled. As noted in (Kalajdzievski, 2023), the scaling factor $\frac{\alpha}{r}$ in the original LoRA implementation (Hu et al., 2021) is too aggressive, leading to gradient collapse issues when using larger LoRA ranks. Adjusting the scaling factor to $\frac{\alpha}{\sqrt{r}}$ can mitigate this problem. Therefore, we adopted a larger LoRA α while using the original LoRA scaling factor.

During training, evaluation was performed on the validation set every 5 steps. We used the evaluation script provided by the organizers to compute the F1 score and saved the checkpoint with the highest score.

For inference, we employed beam search with (temperature=0.5, num_beams=3) to predict the target sequence.

The experiments were conducted on 1∼5 Nvidia A30 GPUs. The frameworks used for the experiments were Pytorch and HuggingFace Transformers.

CCL 2024

Model	track1		track2		track3	
	val	test	val	test	val	test
Yi-6B-Base	49.08	53.72	52.12	48.63	67.11	68.00
Qwen1.5-7B-Base	49.83	56.29	52.59	53.63	68.35	-
Qwen1.5-14B-Base	**52.06**	**56.68**	**57.21**	**60.58**	69.98	70.97
Qwen1.5-32B-Base	51.66	-	55.18	-	**73.17**	**74.20**

Table 1: Model comparison

4.4 Model Performance Comparison

We tested several LLMs, fine-tuning them on the training set for each track. The results are presented in Table 1, where "val" indicates the validation set score and "test" indicates the score obtained after submitting the test set results.

Overall, larger models generally performed better. However, we observed that Qwen1.5-32B-Base did not outperform the smaller 14B model on the validation sets for Track 1&2. Ultimately, we selected Qwen1.5-14B-Base for subsequent experiments in Track 1 and Track 2, and Qwen1.5-32B-Base for the task in Track 3.

4.5 Evaluation Results

Table 2 shows the test set scores under different methods.

Track	Model&Method	Test Score
track1	Qwen1.5-14B-Base	56.68
	+MultiTaskLearning+ModelEnsemble	59.12
	+MultiTaskLearning+ModelEnsemble+SyntheticData+ValData	**61.30**
track2	Qwen1.5-14B-Base	60.58
	+MultiTaskLearning+ModelEnsemble	61.72
	+MultiTaskLearning+ModelEnsemble+SyntheticData+ValData	**62.29**
track3	Qwen1.5-32B-Base	74.20
	+MultiTaskLearning	74.61
	+MultiTaskLearning+PostProcessing	**75.28**

Table 2: Evaluation results

For Tracks 1&2, scores improved after applying multi-task learning and label re-inference model ensemble methods, with a more significant improvement in Track 1.

Subsequently, we used Qwenmax-0403 and Qwen-max-0428 as teacher models and generated synthetic data using error-prone samples from the original validation set as seeds. During this process, we found that the teachers provided highly repetitive answers for the same error-prone samples. Therefore, we manually filtered out the highly repetitive synthetic samples. Additionally, to ensure that the synthetic data was similar in length to the original samples, we removed synthetic samples that significantly deviated in length from the original samples based on their length ratio. This resulted in nearly 200 synthetic data samples.

Moreover, it is generally accepted that the validation set, being most similar in distribution to the original training set, is a crucial resource. We ultimately combined the validation set with 100 synthetic samples into the training set, leaving the remaining synthetic data as the validation set. This approach further improved our scores, reaching 61.30 for Track 1 and 62.29 for Track 2.

For Track 3, multi-task learning provided a slight score improvement.

We observed that the task definition for "exaggeration" rhetoric did not include conjunctions. However, the model often included conjunctions in the prediction results. Therefore, we implemented a simple post-processing step: if a sentence was identified as exaggeration rhetoric in both Track 1 and Track 2, we removed the conjunctions from the Track 3 prediction. Similarly, if a sentence was labeled

CCL 2024

as "no rhetoric" in both Track 1&2, we also set it to "no rhetoric" in the Track 3 prediction. This final post-processing step increased the score to 75.28.

5 Conclusion

In this paper, we explored the use of large language models (LLMs) for the tasks of essay rhetoric recognition and understanding. By employing instruction fine-tuning, multi-task learning, synthetic data augmentation, and a model ensemble method based on label re-inference, we significantly improved the model's performance. Ultimately, our approach achieved a score of 66.29 in Task 6 of the CCL 2024 evaluation, CERRU, securing the first place. This outcome demonstrates the tremendous potential of LLMs in the automated processing of complex natural language tasks.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2024. Introducing qwen1.5.

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. *Advances in neural information processing systems*, 33:1877–1901.

Xiao Hu, Wei Song, Lizhen Liu, Xinlei Zhao, and Chao Du. 2017. Automatic recognition of simile based on sequential model. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, volume 2, pages 410–413.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Sehrish Iqbal, Mladen Rakovic, Guanliang Chen, Tongguang Li, Rafael Ferreira Mello, Yizhou Fan, Giuseppe Fiorentino, Naif Radi Aljohani, and Dragan Gasevic. 2023. Towards automated analysis of rhetorical categories in students essay writings using bloom's taxonomy. pages 418–429. Association for Computing Machinery (ACM).

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipali, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. 2024. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.

Chunhong Li and Yongquan Li. 2022. A new method of figurative rhetoric recognition based on automated essay scoring of the oversea chinese students' instructional composition corpus. In *Proceedings of the 6th International Conference on Digital Technology in Education*, pages 107–111.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Brussels, Belgium, October-November. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Huang Xiaoxi, Li Hanyu, Wang Rongbo, Wang Xiaohua, and Chen Zhiqun. 2018. Recognizing metaphor with convolution neural network and svm. *Data Analysis and Knowledge Discovery*, 2(10):77–83.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendix

A Examples of the fine-tuning data

A.1 track1:

{
"instruction": "请你识别出以下句子中的修辞类别，这是一个层次多标签问题，共有4大类和12小类。若你认为有多个结果，请使用换行符隔开。\n句子：他变的行尸走肉,成天郁郁寡欢。\n选项：比喻：明喻，暗喻，借喻。比拟：名词，动词，形容词，副词。夸张：直接夸张，间接夸张，融合夸张。排比：成分排比，句子排比。",
"output": "夸张-直接夸张"
}

A.2 track2:

{
"instruction": "请你识别出以下句子中的修辞内容类型，这是一个层次多标签问题，共4大类11小类：他变的行尸走肉,成天郁郁寡欢。\n选项：比喻：实在物，动作，抽象概念。比拟：拟人，拟物。夸张：扩大夸张，缩小夸张，超前夸张。排比：并列，承接，递进。",
"output": "夸张-扩大夸张"
}

A.3 track3:

{
"instruction": "抽取出下列学生作文文本中的修辞成分。抽取结果格式为：\"连接词：xxx|描写对象：xxx|描写内容：xxx\"，若有多个结果，请使用换行符隔开。\n（1）针对比喻修辞，对于明喻形式，修辞成分包括连接词（喻词）、描写对象（本体）和描写内容（喻体）；对于暗喻形式，修辞成分包括描写对象（本体）和描写内容（喻体）；对于借喻形式，修辞成分包括描写内容（喻体）；\n（2）针对比拟修辞，不论形式如何，修辞成分都包括描写对象（比拟对象）和描写内容（比拟内容）；\n（3）针对夸张修辞，不论形式如何，修辞成分都包括描写对象（夸张对象）和描写内容（夸张内容）；\n（4）针对排比修辞，不论形式如何，修辞成分都包括连接词（排比项或排比标记）。\n文本：他变的行尸走肉,成天郁郁寡欢。\n抽取结果：",
"output": "描写对象：他|连接词：无|描写内容：变的行尸走肉"
}

B System prompt for calling the teacher LLMs

system_prompt = '''现在你是一个经验丰富的语文知识助手。
有一道题目是：以句子作为基本单位，将每个句子中使用的修辞手法按粗粒度分类成比喻、比拟、夸张、排比以及无修辞类，同时每类修辞进一步从形式角度细粒度分类。
所有类别和解释如下：
一、比喻：
从形式划分：

CCL 2024

（1）明喻：本体、喻体、喻词都出现，通过使用像\像"\好像"\如"\仿佛"等比喻词来连接本体和喻体。如：仙人掌像几个绿色的手掌，有大有小，有粗有细。
（2）暗喻：仅本体、喻体出现，而没有\像"\好像"\如"\仿佛"这些比喻词。如：若要想要种子茁壮成长，少不了机会这一养料。
（3）借喻：仅喻体出现。如：一把弯刀挂在天上。
从内容划分：
（1）实在物：本体是可见、可触、可想的实在物体。如：一把弯刀挂在天上。
（2）动作：本体是某种动作、行为或事件。如：丁尽粳冬的一夜雨声，敲起了春耕的锣，擂响了播种的鼓。
（3）抽象概念：本体是抽象概念，如爱、时间、勇气等。如：时间就是金钱。
二、比拟：
从形式划分：
（1）名词：用写人的名词写物/用写物的动词写人或其他物。如：我看到这辆车久历风尘，实在高寿。
（2）动词：用写人的动词写物/用写物的动词写人或其他物。如：杜鹃花在风中摇曳，向人们展示它优美的舞姿。
（3）形容词：用写人的形容词写物/用写物的形容词写人或其他物。如：湖水愈发温柔，愈发安详。
（3）副词：用写人的副词写物/用写物的副词写人或其他物。如：高粱红了脸，羞答答地低下头微笑。
从内容划分：
（1）拟人：把非人当作人写。如：湖水愈发温柔，愈发安详。
（2）拟物：把非A的某物当作 A写，A非人。如：我到了自家的房外，我的母亲早已迎着出来了，接着便飞出了八岁的侄儿宏儿。
三、夸张：
从形式划分：
（1）直接夸张：直接对事物进行夸张。如：我的爱比所有加起来还要多。
（2）间接夸张：夸大另一样东西来夸大某事物。如：一海洋的水都洗不干净他的手。
（3）融合夸张：借助其他修辞进行夸张，通常是比喻。如：馋虫上身的我也管不了那么多了，拿着这钱就去了小商铺。
从内容划分：
（1）扩大夸张：向大、多、长或高等夸张。如：长长的队伍没有尽头，链接着五湖四海，千山万岭。
（2）缩小夸张：向小、少、短或低等夸张。如：随便你什么时候仰头看，只能看见巴掌大的一块天。
（3）超前夸张：把后出现的事说到先出现的事之前。如：还没回家就已经闻到香味了。
四、排比：
从形式划分：
（1）成分排比：指一个句子中的某些成分，如主谓宾定状等，通过重复的形式排列在一起。如：她的枕，她的床，她的房间，已经空了。
（2）句子排比：排比项可单独成句。如：它坚强的意志让我感动，它从不服从命运的安排，它与命运斗争，它触动了我的心灵。
从内容划分：
（1）并列：排比项顺序改变不影响语义通顺。如：工地上没人唱歌，没人跳舞，没人摔跤，没人吹牛皮，没人闹哄哄地赌饭吗。。。
（2）承接：排比项之间有先后逻辑顺序，如时间、程度、发展状况等，不能改变顺序。如：如果我没记错的话，他不就是三岁扎小辫、五岁穿花裤、九岁还吃奶的那个密级生么？
（3）递进：各排比项表达的含义、情感等层层递进，不能改变顺序。如：我与她深一脚浅一脚重新往黑暗里，往天塌地陷的前面闯，往一个几乎毫无希望的绝境里闯。
五、无修辞。

CCL 2024

你发现有一些同学分不清一些句子所属的类别，需要更多的例子。现在需要你的帮助。我会给出一个例句，你需要参考例句，创造一个与其粗细类别都相同的新句子。
要求：
1.你给出的答案的类别只限定在上面给出的解释中。
2.确保不要出现错误。
3.确保新句子的类别与例句一致。若例句给出的答案有多个，你也需要保证新句子的答案个数一致，每个答案逗号隔开。
4.若例句中没有用到修辞，你需要给出一个没有用到任何修辞手法的句子。
5.创造与例句意义相似的句子，但不要复制原文
6.尽量使你给出的句子贴近学生真实写作风格。
7.若例句中为单一类别，请你尽量保证给出的新句子只包含与例句相同的修辞手法，尽量避免混入额外的修辞手法。
8.新句子长度尽量与例句相近。
9.发挥你的创造力。

一定要保证输出格式为如下格式：
<句子>：[句子]
<粗类别>：[粗类别]
<形式>：[形式]
<内容>：[内容]
'''

CCL 2024

CCL24-Eval任务6系统报告：基于深度学习模型的中小学作文修辞识别与理解评测

李晨阳[1,2]，张龙[1,2]，郑秋生[1,2]
[1]中原工学院 前沿信息技术研究院，河南 郑州 450007
[2]河南省网络舆情监测与智能分析重点实验室，河南 郑州 450007
2312826399@qq.com

摘要

在中小学生的学习进程中，修辞手法是阅读和写作技巧的核心，也是优秀文学作品的关键元素。然而，识别与理解学生文章中的修辞使用需要大量的人工，为教师的作文评估和教学提出了挑战。最近的研究开始使用计算机技术来自动评审作文，其中修辞的使用是评估的重要部分。本文介绍了我们在第二十三届中文计算语言大会中中小学作文修辞识别与理解评测中的所用的参赛方法。在本次评测中，我们针对不同任务，分别使用了传统模型分类模型和大模型，再利用伪标签、数据增强等方法提升模型性能。实验结果表明，我们的方法取得了较为先进的效果。

关键词： 预训练模型 ； 数据增强 ； 半监督学习

System Report for CCL24-Eval Task 6: Assessment of Rhetoric Recognition and Understanding in Primary and Secondary School Essays Based on Deep Learning Models

Chenyang Li[1,2], Long Zhang[1,2],Qiusheng Zheng[1,2]
[1]Frontier Information Technology Research Institute,
Zhongyuan University of Technology, Zhengzhou 450007 China
[2]Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou China
2312826399@qq.com

Abstract

In the learning process of primary and secondary school students, rhetorical devices are at the core of reading and writing skills, and are key elements of excellent literary works. However, identifying and understanding the use of rhetoric in students' essays requires a lot of manual effort, posing a challenge for teachers in essay evaluation and instruction. Recent research has begun to use computer technology to automatically evaluate essays, where the use of rhetoric is an important part of the assessment. This paper introduces our methods used in the evaluation of rhetorical recognition and understanding in primary and secondary school essays at the 23rd Chinese Computational Linguistics Conference. In this evaluation, we employed both traditional classification models and large models for different tasks, and further enhanced model performance through pseudo-labeling and data augmentation. Experimental results indicate that our methods achieved relatively advanced performance.

Keywords: pretrained models , data augmentation , semi-supervised learning

1 引言

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

修辞手法在中小学生的学习过程中占据了重要的地位，它是阅读理解和写作技巧的核心，也是塑造优秀文学作品的重要元素。然而，识别和理解修辞的使用需要大量人工成本，对教师的作文评估和教学提出了挑战。目前，随着教育的发展和网络的普及，许多研究者和机构开始探索利用计算机技术来实现作文的自动评改，其中修辞手法的应用是重要考量因素。作文的修辞使用被视为反映学生文采和语言表达能力的重要指标，对于评估作文质量，指导学生提升表达能力有重要意义。一些现有的研究采用对齐策略等规则，从句子结构，语义信息等语言学特征角度进行排比和比喻等修辞手法的识别；而另一些工作则将修辞理解视为构成抽取任务。然而，这些研究通常存在以下问题：①针对不同的修辞类别进行独立识别，缺乏泛用性；②识别粒度粗，缺乏多层次，细粒度的修辞类型定义；③缺乏对不同修辞类型的修辞对象和内容的定义，无法为学生作文提供全面的指导意见。在这样的背景下，第二十三届中国计算语言学大会发布了中小学作文修辞识别与理解评测。该评测的数据集源自以汉语为母语的中小学生考试作文，包括记叙文和议论文等多种文体。任务目标是系统地定义中小学作文中出现的细粒度修辞类型，从修辞形式和内容两个方面进行识别，并对修辞的使用评分。对于本评测的任务1和任务2，我们采用传统模型进行分类，并使用伪标签、数据增强等方法来提升性能，对于任务3，我们把该信息抽取任务视为生成任务，并引入Qwen大模型(Bai et al., 2023)进行生成。

2 相关工作

文本分类在自然语言处理和文本挖掘中具有重要的作用，通过不断学习文本特征进行预测分类，在各个方面的研究中都具有十分重要的意义和研究价值 (Minaee et al., 2021)。传统的文本分类是基于机器学习方法 (Cheng, 2020)，包括支持向量机、决策树、朴素贝叶斯等，但这些方法都只解决了词汇层面的问题，无法有效学习和反映语句之间的语义相关性和深层语义特征。近年来，深度学习技术在计算机视觉和自然语言处理领域都取得了显著的进展。在自然语言处理任务中，基于深度学习的文本分类模型备受关注和研究，如CNN (Wan et al., 15) (Wang et al., 2017)、RNN (Le et al., 2017)、GNN (Yao et al., 2018)、Attention (Kim et al., 2018)和预训练模型。它们在文本分类等自然语言处理任务中都表现出了优秀的效果。特别是预训练模型，在预训练时就已经接触了大量的文本数据，因此能学习到更加丰富的语义信息，其在文本分类等任务中也具有更高的准确性和泛化能力。

半监督学习是近年来新兴的一种智能学习范式，其利用未标记的数据提高模型性能 (Rizve et al., 2021)，传统的监督学习方法需要使用有标签的数据来建模。然而，在现实世界中给训练数据打标签可能需要昂贵的代价，或者耗费大量的时间。从领域专家那里获得这些标签数据有隐含的成本，例如有限的时间和财务资源。对于涉及使用大量类标签进行学习且有时具有相似性的应用程序尤其如此。半监督学习（SSL）模型能够允许模型在其监督学习中集成部分或者全部的未标签数据来解决这一固有的瓶颈。其目标是通过这些新标记的数据最大化模型的学习性能，同时最小化标注数据的成本。

模型融合是一种训练多个模型并进行融合的方法，旨在通过融合模型结果来超越单个模型的表现。常用的模型的融合方法共有三种，第一种是投票法，适用于分类任务，即对多个学习模型的预测结果过进行投票，以少数服从多数的方式来确定最后的结果，还可以根据人工设置或者根据模型评估分数来设置权重。第二种是平均法，适用于回归和分类任务，即对于学习模型的预测概率进行平均。第三种是交叉融合法，主要思路就是把原始的训练集先分成两部分，例如按9：1划分训练集和测试集，在第一轮训练时，使用训练集训练多个模型，然后对测试集进行预测，在第二轮训练时，直接用第一轮训练的模型在测试集上的预测结果作为新特征继续训练。

3 实现方法

如图1所示，我们先对数据进行预处理，然后对当前主流深度学习模型进行评估，选出表现较好的模型作为我们的基线模型，最后采用伪标签、数据增强等方法来提升分数。

3.1 任务1、2实现方法

任务一与任务二皆为嵌套多标签分类任务，由于这两个任务使用方法一样，我们以任务一为例进行介绍。我们分别对其训练粗粒度和细粒度两个模型，最后在进行粗细结果的融合。

(c) 2024 中国中文信息学会计算语言学专业委员会

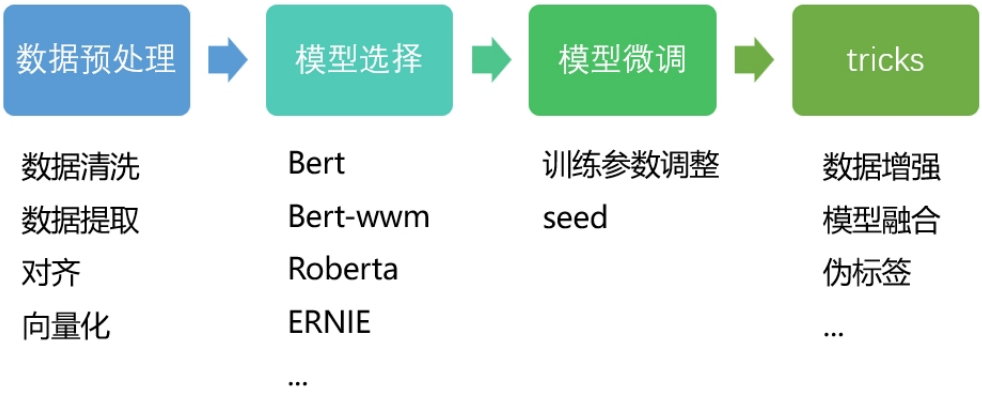


图 1: 模型的数据预处理、预测以及后处理过程

3.1.1 模型选择

如表3.1.1所示，我们选取了当前主流的预训练模型，包括Bert，Bert的变体模型和Ernie，再次基础上对每个模型进行微调。从直观来讲，网络模型越大，层数越深，学习能力越强大，因此我们的Ernie模型 (Sun et al., 2019)选择了20层网络结果的Ernie3.0-xbase进行测试，同时我们后面提出的Ernie均指Ernie3.0-xbase。经过验证，我们发现ernie的效果最高，于是我们把ernie作为我们的基线模型。

模型	Bert_base[0]	Bert_wwm[1]	Bert_wwm_ext[2]	Ernie3.0_xbase[3]	Rebort_wwm_ext[4]
验证集F1	31.1	32.8	31.8	33.4	32.1

表 1: 各模型验证集F1值

3.1.2 数据增强

为了提升F1分数，我们对每个类别的总数和每个类别对应的F1值进行了比较，如图2所示，我们发现了数据分布不平均，部分标签对应的数据量过少的问题，例如属于明喻这一类别的数量有147条，而属于副词这一类别的数量只有4条。这也导致副词这一类别的f1值为0。为了提升该类别的f1分数，我们采用数据增强的方式来使得模型学习到更多的相关特征。采用同义词替换、随机词插入和相似句生成等方法对这些标签的数据进行了数据增强，增强例子如图3所示。在对数据总量增加了三倍之后，我们使用增强后的数据对基线模型进行了重新训练。结果表明，无论是在验证集还是测试集上，模型的性能都有了显著的提升。

3.1.3 基于半监督学习的伪标签生成

伪标签方法来自于半监督学习，其核心思想是借助无标签的数据来提升有监督过程中的模型性能。由半监督学习生成伪标签的过程如图4所示，伪标签方法主要是将模型对无标签的测试数据的预测结果重新加入到训练集中，从而增大数据量以提升模型效果(Li et al., 2023)。这种方法适用于模型精度较高的情况。我们在预测过程中设置了一个阈值0.98。在预测测试集过程中，模型输出概率值大于该阈值的数据都被保存作为伪标签。这些伪标签在我们看来都是接近完全正确的数据。每次的伪标签都接近1000条。这表明我们的数据集在原本数量的基础上又增加了1000条数据。经过多轮的伪标签训练后，筛选出的伪标签内容会越来越接近，最终模型达到了拟合的状态，此时在进行后续的伪标签方法已经不能够再提升测试集的f1值。于是我们采用模型融合的方法来进一步提升f1分数。

[0]https://huggingface.co/google-bert/bert-base-chinese
[1]https://huggingface.co/hfl/chinese-bert-wwm
[2]https://huggingface.co/hfl/chinese-bert-wwm-ext
[3]https://huggingface.co/nghuyong/ernie-3.0-xbase-zh
[4]https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

(c) 2024 中国中文信息学会计算语言学专业委员会

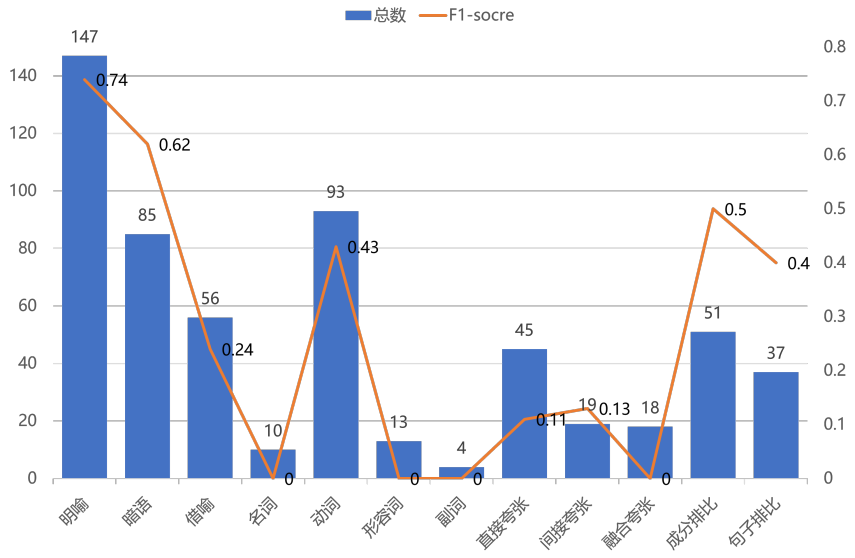


图 2: 各类别对应的训练集数量

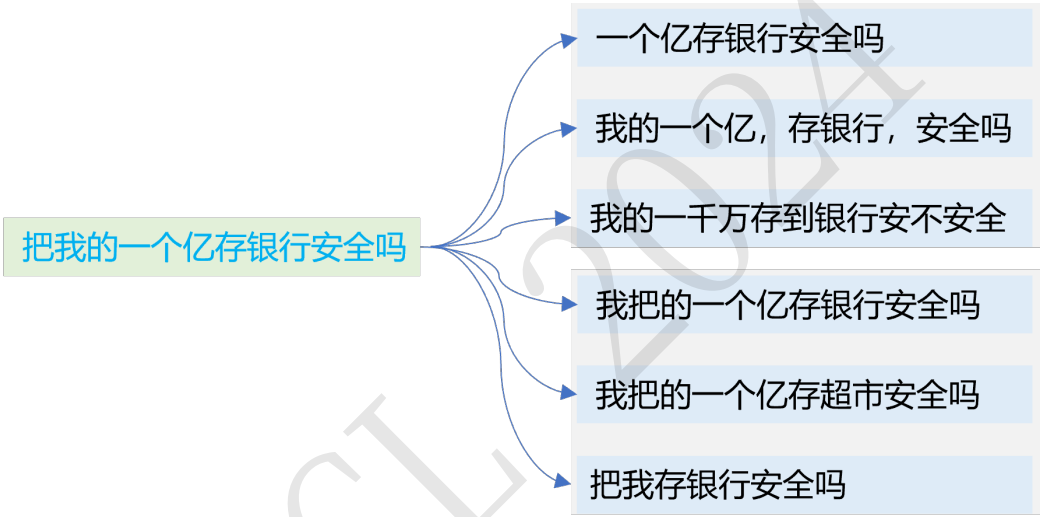


图 3: 数据增强样例

3.1.4 模型融合

关于模型融合，周志华教授(2021)的《机器学习》一书中提到：模型融合要好而不同，即模型差异性越大，融合效果越好。我们从两方面来增加差异化，一是使用不同的两个模型，bert和ernie。二是重新划分训练集和验证集来改变模型输入。再使用这两个模型对测试集进行预测时，我们没有直接输出预测的类别，而是输出了每个类比的概率，然后使两个模型预测的类别概率进行等权相加，得出模型融合后预测的新概率，最后选取具有最大概率的那一个类别作为预测结果。

3.2 任务3实现方法

任务3为信息抽取任务，我们最初尝试使用传统在的信息抽取类模型来完成，但效果并不理想，继而转向了语言大模型，把其作为文本生成任务来实现。

3.2.1 数据格式化及模型选择

我们首先尝试了使用prompt模板的ChatGLM4，发现生成的答案除了格式不规范外还会出现很多原句中没有出现过的词，这显然也是不符合结果要求的。于是我们采用微调大模型的方式来实现。首先是对数据进行格式化处理。如图所示，我们把该任务组为序列生成任务，

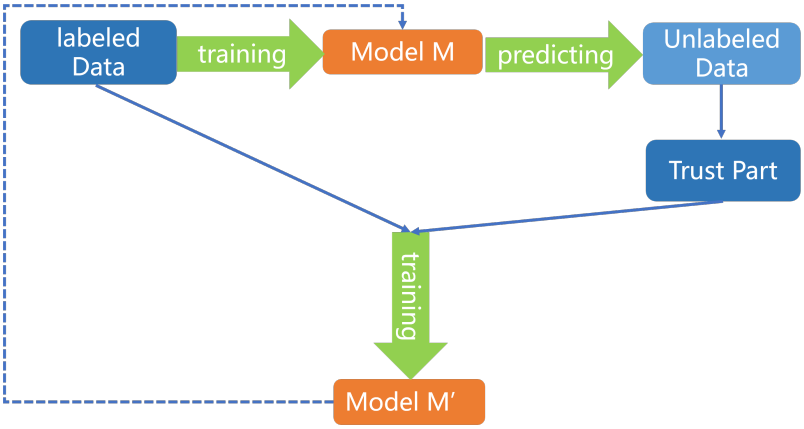


图 4: 基于半监督学习的伪标签生成

而不是直接生成符合本任务的数字答案。如图5所示，我们把原句作为输入，把连接词、描写对象、描写内容3个属性作为输出，其间用“#”进行分隔,如果某个属性为空，则用”null”表示，如果3个属性都为空，则用”null#null#nul”表示。最后针对输出的结果在输入句子的基础上进行位置对比而输出符合任务格式的数据。我们尝试了ChatGLM6b和Qwen7b等开源大模型，采用qlora(Dettmers et al., 2024)和p-tuning(Liu et al., 2022)等方式针对本任务数据进行微调。我们发现Qwen7b效果较好，且当使用qlora进行微调时，把秩的大小设置为64要明显好于其他参数。于是把其当作基线模型进行后续的测试。


```
{
  "paragraphId": 2,
  "sentenceId": 13,
  "sentence": "他变的行尸走肉,成天郁郁寡欢。",
  "componentList": [
    {
      "conjunction": null,
      "conjunctionBeginIdx": null,
      "conjunctionEndIdx": null,
      "tenor": "他",
      "tenorBeginIdx": 0.0,
      "tenorEndIdx": 0.0,
      "vehicle": "变的行尸走肉",
      "vehicleBeginIdx": 1.0,
      "vehicleEndIdx": 6.0
    }
  ]
```

```
{
  "conversations": [
    {
      "from": "user",
      "value": "他变的行尸走肉,成天郁郁寡欢。"
    },
    {
      "from": "assistant",
      "value": "null#他#变的行尸走肉"
    }
  ]
},
```


图 5: 左为原始数据格式，右为格式化后格式

3.2.2 伪标签

在本任务中，我们同样对数据进行分析，并发现数据总体数量是比较少的，属于小样本任务。为了提升后续的分数，我们同样采用任务一中的数据增强方法，以此增加数据集的数量。在增加了1倍数据后，经过训练后发现在其结果在验证集上有了很大的提升，但在测试集上却有所下降，出现了过拟合的状态，可能是由于数据增强的方式破坏了原本的数据分布导致，

于是我们采取了同样能增加训练集数量的伪标签方法。伪标签方法适用于模型精度比较高的情况，此时我们的模型准确率已经达到69的分数，故可以采用该技巧。由于大模型无法想传统分类模型那样输出预测的概率，故不能想任务一那样设置阈值，于是我们结合模型融合的经验，我们利用微调后的ChatGLM和Qwen进行预测，并去预测结果相同的部分加入到训练集中重新训练。每次的伪标签数量大概在1500条，这表明我们的数据集在原本数量的基础上又增加了1500条数据。结果表明，在使用伪标签后，评价指标有所提升。

4 实验结果

如下表所示，分别列出了3个任务的实验结果。任务1和任务3展示了我们在基线模型的基础上，使用了数据增强和伪标签等方法后的线上F1值，相对任务的基线模型分数45.66，我们有了很大的提升。针对于任务3，在使用大模型之前，我们也尝试使用了类似Bert的传统分类模型，然而，传统模型在测试集上的表现要远低于基线模型的结果。这一结果表明基座模型对于推理效果有着显著的影响，而大模型在预训练获得的世界知识和涌现能力对空间语义理解能力任务有着重要帮助。我们在对大模型微调时也面临着一个普遍问题，即幻觉现象(Huang et al., 2023)。当模型生成的文本不遵循原文或者不符合事实时，我们就认为模型出现了幻觉，尽管我们在训练集中标注的答案都是在输入文本中出现过的，但模型在结果预测时仍会出现除输入文本以外的词，为此，我们暂时只采用正则表达式来过滤掉这些无效答案。

模型	线上F1值
Baseline	45.66
Ernie3	44.72
Ernie3+数据增强	46.16
Ernie3+数据增强+伪标签	49.83
Ernie3+数据增强+伪标签+bert_wwm(模型融合)	51.48

表 2: 任务1结果

模型	线上F1值
Baseline	56.89
Ernie3	49.78
Ernie3+数据增强	53.56
Ernie3+数据增强+伪标签	54.25
Ernie3+数据增强+伪标签+bert_wwm(模型融合)	55.11

表 3: 任务2结果

表 4: 任务3结果

模型	线上F1值
Baseline	20.85
ChatGLM4+promptChatGLM-4[5]	24.68
ChatGLM6b[6]	57.00
Qwen1.5-7bQwen1.5-7b[7]	63.30
Qwen7b-1.5+伪标签	69.51

表 5: 任务2结果

[5]https://chatglm.cn
[6]https://github.com/THUDM/ChatGLM3
[7]https://github.com/QwenLM/Qwen1.5

(c) 2024 中国中文信息学会计算语言学专业委员会

5 总结

通过大量实验发现，对于本任务数据集，大多数模型预测的结果分数相近，并且由于本任务数据集数规模比较小，因此数据增强方法伪标签方法都可以增加数据集的规模，提升测试集的F1分数，并增加模型的泛化性。使用多轮伪标签方法后，后续筛选得出的伪标签几乎不会有变化，导致模型的性能不再有提升。这时可以采用模型融合技术，取差异较大的多个模型，分别学习不同的输入，使得多个模型之间学到的知识尽量不同，这样使得多个模型可以更好的融合，提高性能。进一步优化方面，针对训练集存在的过拟合问题，可以考虑在划分训练集和验证集时进行数据均衡。使用模型融合时，可以先采用五折交叉验证法来训练多个模型，然后再对多个预测结果取平均。

参考文献

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*,363:366 – 374.

Chenyang Li, Long Zhang, Qiusheng Zheng, Zhongjie Zhao, and Ziwei Chen. User preference prediction for online dialogue systems based on pre-trained large model. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 349–357. Springer, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Jiang Cheng. 2020. Research and implementation of Chinese long text classification algorithm based on deep learning. *University of the Chinese Academy of Sciences (Institute of artificial intelligence, Chinese Academy of Sciences)*.

Le, H. T., Cerisara, C., and Denis, A. 2017. Do convolutional networks need to be deep for text classification? *arXiv preprint arXiv:1707.04108*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. 2021. Deep learning based text classification: a comprehensive review. *ACM computing surveys (CSUR)*,54(3), 1-40.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. 2021. In defense of pseudo-labeling:an uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Seonhoon Kim, Jin Hyun Hong, inho Kang, and nojun kwak. 2019. Semantic sense matching with densely connected recurrent and co-attentive information. *Proceedings of the AAAI Conference on Artificial Intelligence*,33(01), 6586-6593.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., and Cheng, X. 2016. A deep architecture for semantic matching with multiple positive sense representations. *Proceedings of the AAAI Conference on Artificial Intelligence*,30(1). https://doi.org/10.1609/aaai.v30i1.10342.

Wang, Z., Hamza, W., and Florian, R. 2017. Bilateral multi-perspective matching for natural language sentences. *In procedures of the twenty Sixth International Joint Conference on artistic intelligence*, ijcai-17, pages 4144 – 4150.

Yao, L., Mao, C., and Luo, Y. 2019. Graph revolutionary networks for text classification. *In Proceedings of the AAAI conference on artificial intelligence*,Vol. 33, No. 01, pp. 7370-7377

(c) 2024 中国中文信息学会计算语言学专业委员会

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, danxiang Zhu, Hao Tian, and Hua wuErnie. 2019. Enhanced representation through knowledge. *arXiv preprint arXiv:1904.09223*.

Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.

CCL 2024

CCL24-Eval 任务6系统报告：人类思维指导下大小模型协同决策的中文修辞识别与理解方法

王雯[1], 汤思怡[1],于东[1,*], 刘鹏远[1,2]
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
wangwenblcu@gmail.com, tangsiyi0805@gmail.com
yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

摘要

CCL24-Eval任务6提出了一个多层次、细粒度中小学作文修辞识别与理解任务。针对任务特点，本文提出了人类思维指导下大小模型协同决策的中文修辞识别与理解方法。该方法根据人类在面对修辞识别和理解任务时的处理思路，将任务顺序重新定义，并分别选取大小语言模型，使每个步骤的实现效果均达到局部最优，以局部最优达到整体任务的最优效果。结果表明，本文提出的方法能够有效对修辞进行识别与理解，在三个赛道上相较于Baseline方法分别提升了13.54、4.03、57.11。

关键词： 大语言模型；修辞识别；成分抽取

System Report for CCL24-Eval Task6: A Chinese Rhetoric Recognition and Comprehension Approach for Collaborative Decision-Making in Large and Small Models Guided by Human Thinking

Wen Wang[1],Siyi Tang[1],Dong Yu[1,*], Pengyuan Liu[1,2]
1.Faculty of Computer Science,Beijing Language and Culture University,Beijing,100083
2.National Language Resources Monitoring and Research Center for Print Media,Beijing,100083
wangwenblcu@gmail.com, tangsiyi0805@gmail.com
yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

Abstract

CCL24-Eval Task 6 presents a multi-level, fine-grained elementary and middle school composition rhetoric recognition and comprehension task. Aiming at the task characteristics, this paper proposes a Chinese rhetoric recognition and comprehension method with collaborative decision-making between large and small models under the guidance of human thinking. The method redefines the order of the task according to the human thinking when facing the task of rhetoric recognition and comprehension, and selects the large and small language models respectively, so that the realization effect of each step reaches the local optimization, and the local optimization achieves the optimal effect of the overall task. The results show that the method proposed in this paper can effectively recognize and comprehend rhetoric, and improves 13.54, 4.03, and 57.11 compared with the baseline method on the three tracks, respectively.

Keywords: LLM , Rhetorical Recognition , Component Extraction

*为通讯作者
基金项目： 教育部人文社科规划项目（23YJAZH184）；北京语言大学梧桐创新平台(中央高校基本科研业务费)(21PT04)；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果（24YCX114）
©2024 中国计算语言学大会根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

修辞识别和理解是一个具有挑战性的任务，其目的是在探究和理解文本的修辞结构和语义信息后识别文本使用的修辞类型，并对其中的修辞成分进行抽取。修辞识别和理解任务主要应用于机器阅读理解、作文自动评分等具体问题中，旨在对语文和作文教学提供帮助。在《义务教育语文课程标准》中，对中小学作文训练提出了明确要求，要求学生能够对修辞手法进行恰当地运用(李娜, 2017)。这一要求的出现，使得修辞自动识别和抽取的意义更加显著。然而，之前的研究工作大多集中在对特定修辞大类的识别(武阗阗et al., 2023)，粒度较粗且泛用性差。同时，部分将修辞理解任务与成分抽取任务等同的研究工作(赵琳玲et al., 2021; 郭英豪, 2024)，缺少对修辞全面的定义，无法对作文评价给出全面的指导。

CCL24-Eval任务6提出了一个覆盖细粒度、多层次的中小学作文修辞识别与理解任务。对于层次多且复杂的任务，传统的修辞识别方法并不能很好解决问题。随着ChatGPT等大语言模型的出现，科研人员发现使用海量数据训练出的大语言模型可以有效理解文本的语义，并且对逻辑关系具有良好的分析能力(崔亿萍, 2023)。然而，大语言模型在某些特定领域上存在认知缺陷(Zhang et al., 2023)，知识幻象问题也较为突出(Li et al., 2024)。基于此，针对CCL24-Eval任务6，我们提出了一种人类思维指导下大小模型协同决策的中文修辞识别与理解方法。

2 任务定义

CCL24-Eval任务6旨在对修辞进行多层次、细粒度的识别以及对修辞对象及内容的抽取。针对当前修辞识别泛用性差、粒度较粗、定义不明确等缺点，该任务提出了三个赛道：修辞形式类型识别、修辞内容类型识别、修辞成分抽取。三个赛道分别对比喻、比拟、夸张和排比修辞的共性内容进行统一的识别与抽取。每个赛道的评价标准将不同粒度、不同成分的识别效果进行调和，使结果更贴近中小学生作文修辞识别的真实需求。

然而，人类通过自己的认知能力和思维对修辞进行判断的顺序并非按照赛道1、赛道2、赛道3的顺序依次进行。人类拥有分层级的个人语言能力系统，最基础最核心的是认知力，为观念意识到达了某层级的程度；然后是思维力，对应具体事件策划及分析的程度；再上层是逻辑力，即语言要素组成的逻辑程度(张凯, 2024)。廖巧云的研究为语义修辞的生成(廖巧云, 2018b)和理解机制(廖巧云, 2018a)构建了不同框架，同样表明了修辞需要读者调动相应的语言能力来理解修辞要素之间的关系，进而才能理解语义。因此，人类面对修辞识别等任务时的思维过程也基本符合上述语言能力系统中从低到高的层级顺序，即先对修辞手法进行粗粒度的分类，然后在对修辞成分的抽取和分析的同时，对修辞成分的关系、性质等方面有一个具体的认识。根据人类的思维过程，我们重新定义CCL24-Eval任务6三个赛道中的任务顺序，如图1所示，蓝色代表处理顺序1，绿色代表处理顺序2，黄色代表处理顺序3。

图 1: 重定义后的任务

3 数据扩充

CCL24-Eval任务6发布的数据集[0]中包含375条比喻句、148条比拟句、103条夸张句、107条排比句和125条无修辞句，为提高评测数据集的多样性和规模，提升模型的训练效果和泛化能力，我们利用Li et al. (2022)公开的CLGC中文文采评估语料库[1]，并爬取了部分网络数据，采用大语言模型自动标注和人工校对的方法，扩充了基础评测数据集，扩充过程如图2所示。最终形成了比喻、比拟、夸张、排比四种修辞各1000条语料的扩充评测数据集，数据集情况如附录表5所示。

[0]https://github.com/cubenlp/CERRU
[1]https://github.com/blcunlp/CLGC/

(c) 2024 中国中文信息学会计算语言学专业委员会

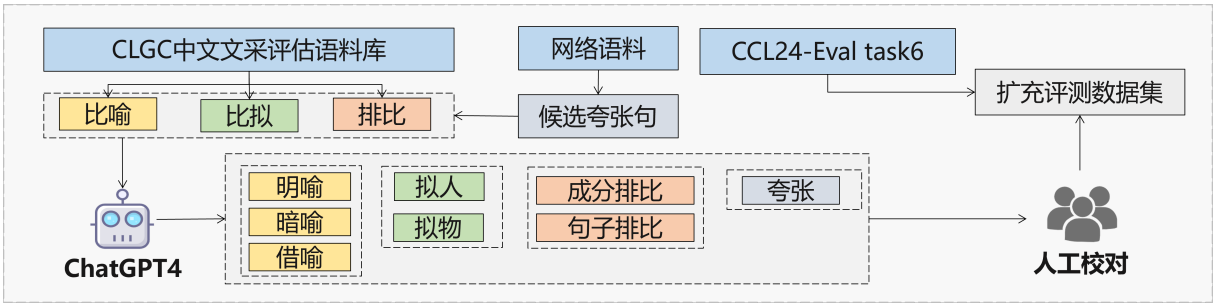


图 2: 评测数据扩充过程

在图2中可以看到，我们从CLGC中文文采评估语料库中抽取比喻、比拟、排比标签的句子；并以“夸张句”为关键词爬取夸张句网站[2345]，再从小说阅读网[6]爬取1000篇中文小说，抽取文本中带有“夸张”关键词的句子及该句子的前后两句作为候选夸张句。选取ChatGPT4作为自动标注模型，输入抽取的比喻、比拟、夸张、排比句以及候选夸张句进行标注。如果句子具有比喻修辞，则按照成分显隐性细分为明喻、暗喻、借喻；如果句子具有比拟修辞，则按照喻体性质细分为拟人和拟物；如果句子具有排比修辞，则按照排比项成分细分为成分排比和句子排比；如果句子具有夸张修辞，则纳入夸张句语料。ChatGPT4标注完毕后，招募三名语言学硕士研究生进行人工校对，提高评测数据质量。

4 方法

4.1 方法概述

修辞手法的识别和成分抽取任务的多个层次之间有着极高的关联性，即特定的细粒度分类只存在于相应的粗粒度分类下，上一层次的识别准确程度极大程度影响了下一层次的识别与抽取的效果。因此，我们根据该任务层次关联的特点提出了如图3的处理流程，在流程的每个节点上都试图达到局部最优的效果，并以此提高任务整体的性能。此外，任务的处理顺序可能会对效果产生影响，我们调整了任务顺序进行探究，具体情况如A.11所示。

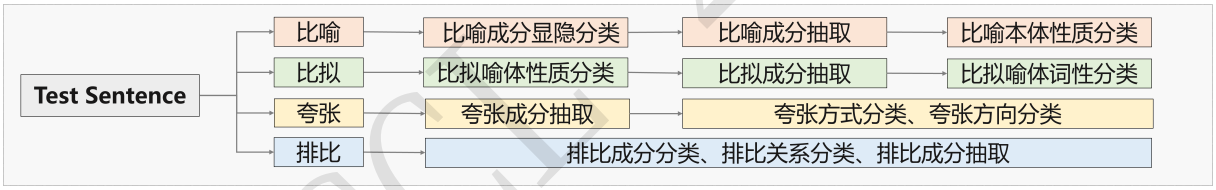


图 3: 方法概况

4.2 修辞的粗粒度识别

我们使用章节3中对CCL24-Eval任务6数据集扩充后带有比喻、比拟、夸张、排比四种修辞粗粒度标签的数据集训练了BERT模型。然而，一个测试句子可能会使用一种或多种修辞手法。因此，我们将BERT输出的概率作为Softmax输入到ChatGPT4中，指导大语言模型对句子中包含的修辞手法作出最终的判断。对于每个测试句子，通过上述BERT-LLMSoftmax方法预测后会被分入到其相应的粗粒度修辞分类中，以待进行后续的识别与抽取任务。BERT-LLMSoftmax方法使用的Prompt模板详见附录A.1。

4.3 比喻的识别与理解

比喻修辞的识别与理解流程如图4所示。首先，利用扩充评测数据集训练BERT模型，如果测试句子在通过LLM Softmax后被标注为比喻句，则再利用扩充评测数据集中具有明喻、

[2] https://www.pinyudu.com/jiaju/2341
[3] https://www.baihuawen.cn/yuedu/zhaichao/20125.html
[4] https://www.ruiwen.com/word/duanyidiandekuazhangju.html
[5] https://wenku.baidu.com/view/264da81551ea551810a6f524ccbff121dd36c53e.html
[6] https://www.readnovel.com/

(c) 2024 中国中文信息学会计算语言学专业委员会

暗喻、借喻标签的语料训练BERT模型，将测试句子分为明喻、暗喻、借喻三类。接着，构建比喻-形式样本库，样本库中包括明喻、暗喻、借喻三种形式的句子各10条，每条均标注了本体、喻词、喻体等相应修辞成分。采用Few-shot提示策略，设计比喻修辞成分抽取Prompt模版并随机采样，利用ChatGPT4模型，以抽取出比喻的不同修辞成分。抽取明喻类别修辞成分的Prompt模板如附录A.2所示，其他细粒度分类的Prompt模板与之类似。然后，构建比喻-本体分类样本库，其中包含实在物、动作、抽象概念三个类别的样本各10条。采用Few-shot策略，设计Prompt模板并随机采样，利用ChatGPT4对本体内容的分类进行判断。针对比喻句本体内容进行分类的Prompt模板，具体内容如附录A.3所示。

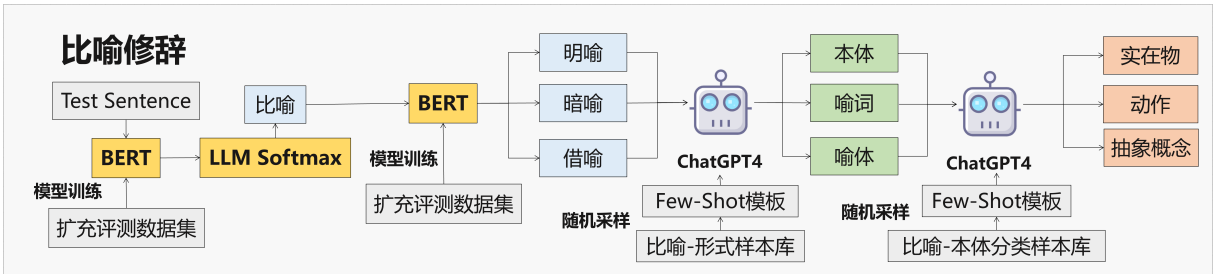


图 4: 比喻修辞处理流程

4.4 比拟的识别与理解

图5展示了比拟修辞的识别与理解流程。首先，利用扩充评测数据集训练BERT模型，如果测试句子在通过LLM Softmax后被标注为比拟句，则再利用扩充评测数据集中具有拟人、拟物标签的语料训练BERT模型，将测试句子分为拟人、拟物两类。接着，构造比拟-成分样本库，其中拟人、拟物两种喻体性质的句子各10条，每个样本均标注了比拟对象和比拟内容。在比拟修辞成分抽取步骤中，采用Few-shot提示策略，设计Prompt模板并随机采样，将该Prompt模板注入ChatGPT4进行成分抽取。附录A.4展示了拟人句子进行成分抽取的Prompt示例，拟物分类的Prompt与之类似。同样，对于比拟内容和其对应的词性分类也构建比拟-喻体词性样本库，包含了标有名词、动词、形容词、副词四种分类的比拟内容示例各10条，使用随机采样的Few-shot策略的Prompt模板，在大语言模型的辅助下将比拟内容分类，对比拟内容进行分类的Prompt模板则如附录A.5所示。

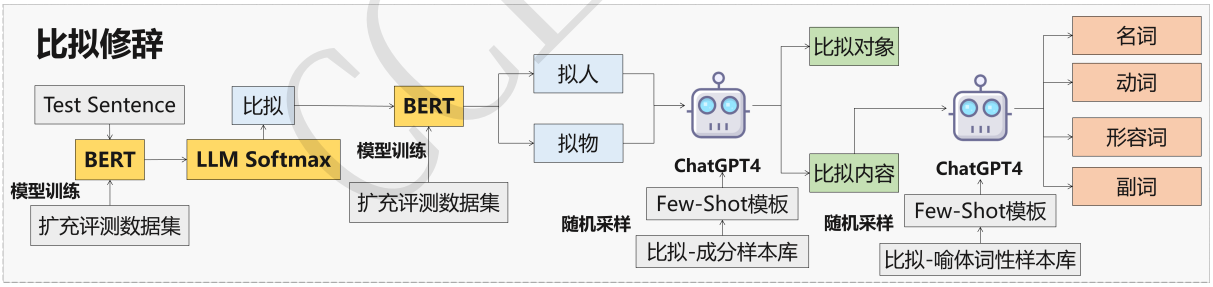


图 5: 比拟修辞处理流程

4.5 夸张的识别与理解

夸张修辞识别的任务特点与其他修辞存在差异，根据人类思维指导，先进行夸张修辞的成分抽取，再根据其成分的性质和形式进行对应的分类，故夸张的识别与理解流程与其他修辞手法有所不同。基于夸张修辞分析的客观情况，夸张的识别与理解任务均使用大语言模型完成。如图6所示，夸张修辞的样本库共有3个，夸张-成分样本库中包含标注了夸张对象和夸张内容的夸张句子样本共10条，夸张-形式样本库中包含直接夸张、间接夸张、融合夸张三种夸张形式的样本各10条，夸张-方向样本库中则包含扩大夸张、缩小夸张、超前夸张三种夸张方向的样本各10条，其中后两个样本库中同时标注了句子中的夸张对象和夸张内容。在夸张成分抽取、夸张形式识别、夸张方向识别三个任务中，分别从其对应的样本库中随机采样出样本，作

(c) 2024 中国中文信息学会计算语言学专业委员会

为Few-shot策略中的示例样本，设计Prompt模板，使用ChatGPT4完成各个阶段的任务。三个任务所使用的Prompt模板，具体如附录A.6、A.7和A.8所示。

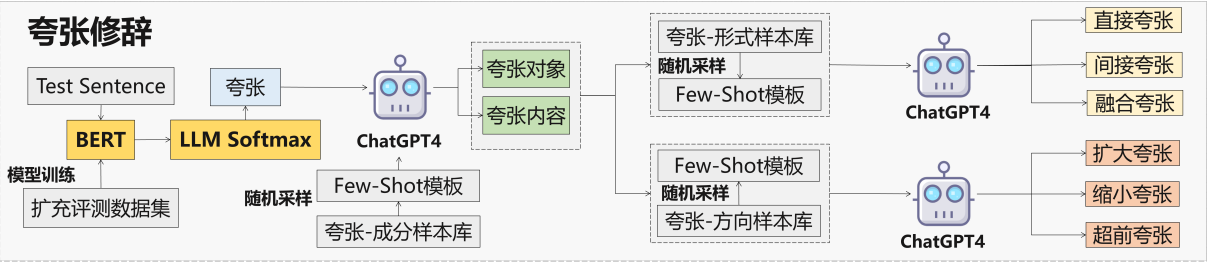


图 6: 夸张修辞处理流程

4.6 排比的识别与理解

排比修辞的识别与理解流程如图7所示。首先，利用扩充评测数据集训练BERT模型，如果测试句子在通过粗粒度识别后被标注为排比句，则可以同时进行排比成分分类、排比关系分类、排比成分抽取。利用标有成分排比、句子排比标签的扩充评测数据集语料训练BERT模型进行排比成分分类。构建排比-成分和排比-关系样本库，排比-成分样本库中包含10条标有排比项的排比句，排比-关系样本库中则包含并列、承接、递进三种排比项关系的句子各10条。采用Few-shot策略，利用ChatGPT4对排比项成分进行抽取和排比句内的逻辑关系判断。具体步骤如图7所示。其中针对排比成分抽取、排比关系判断的Prompt模板可见附录A.9、A.10。

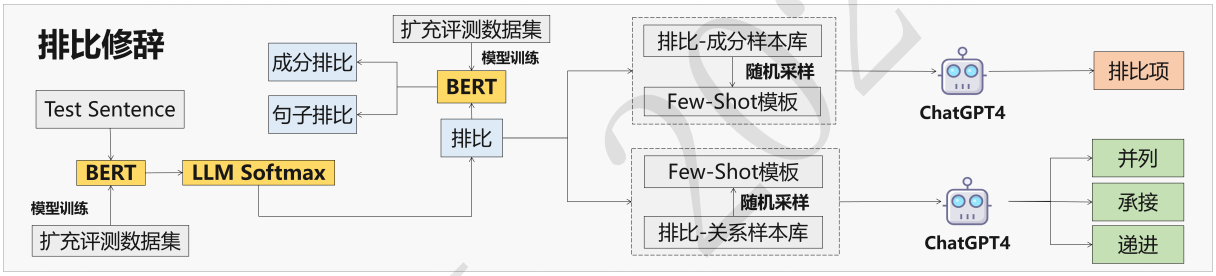


图 7: 排比修辞处理流程

5 实验与结果

5.1 修辞粗粒度识别

在修辞的粗粒度（比喻、比拟、夸张、排比）识别中，对比使用CCL24-Eval任务6训练数据训练BERT模型进行分类、使用大语言模型直接进行分类、使用扩充评测数据集训练BERT模型进行分类三种方法，其中使用扩充评测数据集训练的BERT模型识别效果最好，我们使用该模型输出的概率作为Softmax对ChatGPT4进行指导，以判定句子最终的修辞标签，识别效果较直接分类有显著提高。不同方法的具体识别效果如表1所示。

任务	方法	训练数据	测试数据	P	R	F1
修辞粗粒度识别	BERT	training.json	validation.json	0.37	0.38	0.38
	LLM	-	validation.json	0.50	0.45	0.42
	BERT	扩充评测数据集	validation.json	0.62	0.65	0.63
	BERT_LLMSoftmax	扩充评测数据集	validation.json	**0.70**	**0.73**	**0.70**

表 1: 不同方法在修辞粗粒度识别上的效果

5.2 修辞细粒度识别

每个测试句子经过粗粒度的修辞识别过程后，会被分入到比喻、比拟、夸张、排比四种修辞分类中。在此之后，对比喻、比拟、排比三种修辞类型，进行修辞的细粒度识别，即

(c) 2024 中国中文信息学会计算语言学专业委员会

对于比喻修辞，按照修辞成分的显隐性分为明喻、暗喻、借喻；对于比拟修辞，按照喻体的性质分为拟人、拟物；对于排比修辞，按照排比项的成分分为成分排比和句子排比。针对每种修辞的细粒度分类识别，本文分别进行了实验，对比了扩充评测数据集训练的BERT模型与ChatGPT4的识别效果。如表2所示，使用BERT的F1值均优于大语言模型。

任务	粗粒度分类	细粒度分类	LLM(P/R/F1)	BERT(P/R/F1)
修辞成分抽取	比喻	明喻/暗喻/借喻	0.31/0.30/0.30	**0.74**/**0.72**/**0.72**
	比拟	拟人/拟物	**0.85**/0.74/0.78	0.80/**0.88**/**0.83**
	排比	成分排比/句子排比	**0.64**/0.56/0.51	0.57/**0.57**/**0.57**

表 2: 不同方法在修辞细粒度识别上的效果

5.3 修辞成分抽取

依据章节4中介绍的方法，本文使用ChatGPT4进行四种修辞的成分抽取。对于比喻修辞，抽取本体、喻词、喻体；对于比拟修辞，抽取比拟对象、比拟内容；对于夸张修辞，抽取夸张对象、夸张内容；对于排比修辞，抽取排比项。在上述步骤中，我们均对比了Few-shot和思维链(Chain-of-thought，CoT)两种不同策略的抽取效果。具体效果可见表3，使用Few-shot策略的效果均优于CoT策略。

任务	粗粒度分类	连接词	描写对象	描写内容	CoT(F1)	Few-shot(F1)
修辞成分抽取	比喻	喻词	本体	喻体	0.39	**0.73**
	比拟	-	比拟对象	比拟内容	0.18	**0.53**
	夸张	-	夸张对象	夸张内容	0.27	**0.58**
	排比	排比项	-	-	0.16	**0.96**

表 3: 不同策略在修辞内容抽取上的效果

5.4 修辞成分分类

在抽取出修辞成分之后，进行修辞成分的分类。对于比喻本体，分为实在物、动作、抽象概念；对于比拟成分，分为名词、动词、形容词、副词；对于夸张形式，分为直接夸张、间接夸张、融合夸张；对于夸张方向，分为扩大夸张、缩小夸张、超前夸张；对于排比项之间的关系，分为并列、承接、递进。在修辞成分分类的步骤中，采用Few-shot和CoT策略分别设计相应的Prompt。对比两种策略，可以发现，Few-shot有更高的F1值，具体情况可见表4。

粗粒度分类	成分	成分分类	CoT(P/R/F1)	Few-shot(P/R/F1)
比喻	本体	实在物/动作/抽象概念	0.60/0.58/0.56	**0.63**/**0.67**/**0.62**
比拟	比拟成分	动词/名词/形容词/副词	0.71/0.83/0.69	**0.75**/**0.92**/**0.78**
夸张	形式	直接/间接/融合	0.28/0.29/0.29	**0.57**/**0.54**/**0.54**
夸张	方向	扩大/缩小/超前	**0.98**/0.75/0.82	0.83/**0.97**/**0.89**
排比	关系	并列/承接/递进	0.39/0.44/0.25	**0.61**/**0.89**/**0.66**

表 4: 不同方法在修辞成分分类上的效果

6 结论

针对CCL24-Eval任务6中小学作文修辞识别与理解任务，本文提出了人类思维指导下大小模型协同决策的中文修辞识别与理解方法。该方法通过人类在开展修辞识别和理解任务时的思考思路，重新定义了任务顺序。在完成任务的过程中使用大小模型协同决策的策略，充分利用了大小模型各自的优势，以每个步骤的局部最优来达到整体最优的效果。在本次评测任务中，本文的方法在赛道一、赛道二、赛道三上的分数分别为59.20、60.92、77.96，相比于Baseline结果高出了13.54、4.03、57.11。未来，我们将进一步优化模型的架构，以提升协同效果和模型的综合性能，并将其他类型修辞手法纳入自动识别与理解范围，适应更多样化的应用场景。

第二十三届中国计算语言学大会论文集，第240页-第252页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

参考文献

Yi Li, Dong Yu, and Pengyuan Liu. 2022. Clgc: A corpus for chinese literary grace evaluation. In *Language Resources and Evaluation Conference*.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

崔亿萍. 2023. 大模型在自然语言处理的应用和研究. 中国科技期刊数据库工业*A*, pages 57–61.

廖巧云. 2018a. 语义修辞的生成机制研究. 外语教学, 39(03):10–15.

廖巧云. 2018b. 语义修辞的识解机制. 现代外语, 41(01):1–11+145.

张凯. 2024. 中文修辞能力核心要素计算与智能评价标准构建——基于中小学语文教材修辞语料. 昆明学院学报, 46(02):23–34.

李娜. 2017. 小学高年级语文作文教学的评价研究. 小学生*(*教学实践*)*, (05):1.

武阗阗, 宋子尧, 韩旭, 程苗苗, 巩捷甫, 王士进, and 宋巍. 2023. 学生议论文中的比喻论证作用分析. 中文信息学报, 37(10):158–166.

赵琳玲, 王素格, 陈鑫, 王典, and 张兆滨. 2021. 基于词性特征的明喻识别及要素抽取方法. 中文信息学报, 35(01):81–87+95.

郭英豪. 2024. 面向高考鉴赏类问题的比喻与夸张识别及要素抽取技术研究. Master's thesis, 山西大学.

CCL 2024

A 附录

类别	数量	平均句长	示例
比喻	1000	38.02	他沉默着，像一个木偶似的站立在林青史的面前。
比拟	1000	29.18	这会儿，你看，小草含着泪珠儿，在泣在愁。
夸张	1000	30.22	一个芝麻大的官放个屁，到了你们这儿也是8级地震！
排比	1000	44.50	夏天的夜晚，是那么的宁静，那么的美丽，那么凉爽。
无修辞	1000	27.01	他是一名善良而又喜欢帮助同学的少先队员。

表 5: 扩充评测数据集情况

A.1 BERT LLMSoftmax Prompt模板

##任务描述##
请识别测试句子所具有的修辞，其中可能包含比喻、比拟、夸张、排比和无修辞中的一种或多种修辞。
-比喻：抓住和利用不同事物的相似点，用另一个事物来描绘所要表现的事物。特征为1.一般由三部分组成，即：本体、喻词、喻体。2.本体和喻体必须是性质不同的两类事物。3.在比喻句中，喻体必须出现。4.本体和喻体之间必须有相似点。
-比拟：把物当作人写，或把人当作物写，或把甲物当作乙物来写。特征是1. 把物“人化”，或把人“物化”，或“甲物乙物化”；2.存在本体和拟体，其中本体必须出现，而拟体一般不出现。
-夸张：故意言过其实，通过对客观的人、事物做扩大或缩小或超前的描述。特征为：1.故意把客观事物说得“小，少，低，弱，浅”；2. 故意把事物说得“大、多、高、深、强”；3. 把后出现的说成先出现，把先出现的说成后出现。
-排比：把三个或三个以上结构相同或相似、语气一致、意思密切关联的句子或句子成分排列起来，使内容和语势增强。特征为：1.排比必须三个或三个以上结构相同或相似、意义相关、语气一致的词组或句子排列成串，形成一个整体。2.排比结构、长度等大致相似、没有类似的对称效果，字数不做严格要求。
-无修辞：即不具有任何修辞手法。
##参考分类##
按照BERT模型分类，测试句子具有的修辞标签概率为：{预测1}>{预测2}>{预测3}>{预测4}>{预测5}，该预测概率仅作参考，并非真实确定标签。
##测试句子##
{test sentence}

A.2 比喻修辞抽取明喻形式成分Prompt模板

##任务描述##
测试句子是比喻中的明喻形式，请从中抽取出描写对象（本体）、喻词和描写内容（喻体）。喻词有像、好像、好比、好似、恰似、如、有如、犹如、仿佛、如同。
##参考分类##
【示例1】
Sentence：其中赛龙舟最有特色，那争先恐后的激烈场面就像屈原强烈的爱国精神一样，几千年来一直激励着我们。
Tenor：争先恐后的激烈场面
Conjunction：像
Vehicle：屈原强烈的爱国精神
【示例2】
Sentence：回家后我们围绕着她聚成环，但假期结束后，我们又如同蒲公英一般随风而去，飞向远方。
Tenor：我们
Conjunction：如同
Vehicle：蒲公英

第二十三届中国计算语言学大会论文集，第240页-第252页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

【示例3】
Sentence：一片片枯叶好似仙女下凡，飘然而至。
Tenor：一片片枯叶
Conjunction：好似
Vehicle：仙女
##测试句子##
请按照以上结构，抽取以下测试句子，如果有多个喻词、描写对象（本体）和描写内容（喻体）需要一并写出：
Sentence：{test sentence}

A.3 比喻修辞本体分类Prompt模板

##任务描述##
测试句子是比喻句，本体即被比喻的事物，请判定该句子的本体属于实在物、动作、抽象概念。
-实在物：本体是可见、可触、可想的实在物体；
-动作：本体是某种动作、行为或事件；
-抽象概念：本体是抽象概念，如爱、时间、勇气等。
##参考分类##
【示例1】
Sentence：他是一个男生，大大的眼睛仿若桃花竞相开放，明媚缠绵。
tenor：大大的眼睛
Content：实在物
解释：大大的眼睛是可见的实在物。
【示例2】
Sentence：人生是一颗多味的“秀逗”糖，会越吃越甜！
tenor：人生
Content：抽象概念
解释：人生是抽象概念。
【示例3】
Sentence：雨水流在它脸上流过，像是少女在哭泣。
tenor：雨水流在它脸上流过
Content：动作
解释：雨水流在它脸上流过是动作。
【示例4】
Sentence：我们之间仿佛隔了厚厚的屏障。
tenor：没有明确指出
Content：抽象概念
解释：句子的本体没有直接说明，需要进一步理解语义，本体应该是两个人之间的情感或沟通上的隔阂。
##测试句子##
Sentence：{test sentence}

A.4 比拟修辞抽取拟人形式成分Prompt模板

##任务描述##
测试句子是比拟中的拟人形式，请从中抽取出修辞成分，包括描写对象（比拟对象）和描写内容（比拟内容）。
##参考示例##
【示例1】
Sentence：当暴风雨来临时，任凭风吹雨打，把小草打得东倒西歪，它都没有屈服，依然牢固地扎根在泥土里，顽强地生长着。
Tenor：小草
Vehicle：屈服

CCL 2024

Tenor：小草
Vehicle：顽强地
【示例2】
Sentence：天上已有七八个星在闪烁，一切显得那么神秘，渐渐的，月亮不再羞答答的了。
Tenor：月亮
Vehicle：羞答答的
【示例3】
Sentence：人们只看见电火的愤怒和山洪的疯狂。
Tenor：电火
Vehicle：愤怒
Tenor：山洪
Vehicle：疯狂
##测试句子##
请按照以上结构，抽取以下测试句子，如果有多个描写对象（比拟对象）和描写内容（比拟内容）需要一并写出：
Sentence：{test sentence}

A.5 比拟修辞比拟成分词性标注Prompt模板

##任务描述##
测试句子是比拟句，请判断标志比拟成分(vehicle)的词性为名词、动词、形容词和副词中的哪一种。
##参考示例##
【示例1】
Sentence：我静静地哭着，寒风无情地吹着你的绒毛，也让我感到了一丝寒意。
tenor：寒风
vehicle：无情地吹着你的绒毛
词性：副词
【示例2】
Sentence：你骄傲自满，尾巴翘上天了。
tenor：你
vehicle：尾巴翘上天了
词性：名词
【示例3】
Sentence：我们家原来有一盆仙人掌，长得很丑陋。
tenor：仙人掌
vehicle：丑陋
词性：形容词
【示例4】
Sentence：她似乎衣袋里全装着天真，一掏出来就可以用。
tenor：天真
vehicle：一掏出来就可以用
词性：动词
注意：只需回答比拟成分词性分类，不需要进行解释。
##测试句子##
Sentence：{test sentence}
Tenor：{tenor}
Vehicle：{vehicle}

A.6 夸张修辞抽取成分Prompt模板

##任务描述##
测试句子是夸张句，请从中抽取出修辞成分，包括描写对象（夸张对象）和描写内容（夸张内

CCL 2024

容）。
##参考示例##
【示例1】
Sentence：可怕的雷鸣震裂了天空。
Tenor：雷鸣
Vehicle：震裂了天空
【示例2】
Sentence：常德盛恨不得把一天掰成两天来过。
Tenor：常德盛
Vehicle：恨不得把一天掰成两天来过
【示例3】
Sentence：听到这句话，我仿佛感到血管里的血一下子凝固了。
Tenor：血管里的血
Vehicle：一下子凝固了
##测试句子##
请按照以上结构，抽取以下测试句子，如果有多个描写对象（夸张对象）和描写内容（夸张内容）需要一并写出：
Sentence：{test sentence}

A.7 夸张句判断夸张方向Prompt模板

##任务描述##
测试句子是夸张句，请判断句子夸张方向属于扩大夸张、缩小夸张还是超前夸张。
-扩大夸张：故意把事物说得“大、多、高、深、强”，对事物形象、性质、特征、作用、程度等加以扩大；
-缩小夸张：故意把客观事物说得“小，少，低，弱，浅”，对事物形象、性质、特征、作用、程度等加以缩小；
-超前夸张：把后出现的说成先出现，把先出现的说成后出现。
##参考示例##
【示例1】
Sentence：可怕的雷鸣震裂了天空。
Tenor：雷鸣
Vehicle：震裂了天空
Content：扩大夸张
【示例2】
Sentence：这个足球场只有巴掌那么大。
Tenor：这个足球场
Vehicle：只有巴掌那么大
Content：缩小夸张
【示例3】
Sentence：还没回家就闻到香味了。
Tenor：还没回家
Vehicle：就闻到香味了
Content：超前夸张
##测试句子##
Sentence：{test sentence}
Tenor：{tenor}
Vehicle：{vehicle}

A.8 夸张句判断夸张形式Prompt模板

##任务描述##
测试句子是夸张句，请判断句子夸张形式属于直接夸张、间接夸张还是融合夸张。

CCL 2024

-直接夸张：直接对某物进行夸张
-间接夸张：夸大另一样东西来夸大某事物
-融合夸张：借助其他修辞进行夸张
##参考示例##
【示例1】
Sentence：可怕的雷鸣震裂了天空。
Tenor：雷鸣
Vehicle：震裂了天空
Form：直接夸张
解释：在这句话中，“雷鸣”被直接夸张地描述为能“震裂天空”，以强调雷声的巨大和恐怖。
【示例2】
Sentence：常德盛恨不得把一天掰成两天来过。
Tenor：常德盛
Vehicle：恨不得把一天掰成两天来过
Form：间接夸张
解释：在这句话中，常德盛因为某种原因（如工作繁忙或时间不够用）而希望把一天掰成两天来过。这种夸张的表达间接地表现出他对时间的极度需求或对事情的重视。
【示例3】
Sentence：不愿意去上学的孩子像蜗牛爬行一样去学校。
Tenor：不愿意去上学的孩子
Vehicle：像蜗牛爬行一样去学校
Form：融合夸张
解释：在这句话中，“不愿意去上学的孩子”被比作“蜗牛爬行一样去学校”，这是将比喻和夸张结合起来使用。通过比喻，形象地夸张了孩子走得非常慢，生动地表达了孩子对上学的抗拒和拖延。
注意：只需回答夸张形式，不需要进行解释原因。
##测试句子##
Sentence：{test sentence}
Tenor：{tenor}
Vehicle：{vehicle}

A.9 排比修辞抽取排比项成分Prompt模板

##任务描述##
测试句子是排比句，请抽取该句子的排比项。
-排比项：多次重复出现的那些共同词、句式等。
##参考示例##
【示例1】
Sentence：微风吹过，花瓣随风飘落，有的落在头上，有的落在肩上，有的落在地上。
Conjunction：有的落在
【示例2】
Sentence：勿忘昨天的苦难辉煌，无愧今天的使命担当，不负明天的伟大梦想。
Conjunction：天的
【示例3】
Sentence：劳动，让我们的双手更灵活，让我们的身体更健康，让我们的生活更美好。
Conjunction：让我们的
##测试句子##
Sentence：{test sentence}

A.10 排比修辞判断排比项关系Prompt模板

##任务描述##
测试句子是排比句，请判断句子的排比项关系是并列、承接还是递进。

(c) 2024 中国中文信息学会计算语言学专业委员会

-并列：排比项顺序改变不影响语义通顺；
-承接：排比项之间有先后逻辑顺序，如时间、程度、发展状况等，不能改变顺序；
-递进：各排比项表达的含义情感等层层递进，不能改变顺序。
##参考示例##
【示例1】
Sentence：微风吹过，花瓣随风飘落，有的落在头上，有的落在肩上，有的落在地上。
Content：并列
【示例2】
Sentence：勿忘昨天的苦难辉煌，无愧今天的使命担当，不负明天的伟大梦想。
Content：承接
【示例3】
Sentence：劳动，让我们的双手更灵活，让我们的身体更健康，让我们的生活更美好。
Content：递进
##测试句子##
Sentence：{test sentence}

A.11 任务顺序对于效果的影响

本文提出的人类思维指导下大小模型协同决策的中文修辞识别与理解方法在解决任务过程中，将任务拆解并重新排序。然而，任务顺序的不同也可能对最终的效果产生影响。因此，我们在不同修辞大类下，探究了其他顺序的效果。

对于比喻修辞，本文方法的处理顺序为比喻成分显隐分类、比喻成分抽取、比喻本体性质分类；探究顺序为比喻成分显隐分类、比喻本体性质分类、比喻成分抽取。比喻本体性质分类任务的Prompt依赖于比喻成分抽取任务结果，调换任务顺序后仅对比喻本体分类任务产生影响。对于比拟修辞，本文方法的处理顺序为比拟喻体性质分类、比拟成分抽取、比拟喻体词性分类；探究顺序为比拟喻体性质分类、比拟喻体词性分类、比拟成分抽取。比拟成分词性分类任务的Prompt依赖于比拟成分抽取任务结果，调换任务顺序后仅对比拟成分词性分类任务产生影响。对于夸张修辞，本文方法的处理顺序为在夸张成分抽取后，同时对夸张方式、夸张方向进行分类；探究不依赖夸张成分抽取结果，直接对夸张方式和夸张方向进行分类，调换任务顺序后仅对夸张方式分类和夸张方向分类任务产生影响。对于排比修辞，三个任务之间没有依赖关系，因此不需要调整任务顺序进行探究。修辞的任务顺序对于效果的具体影响如表6所示，可以看出，本文方法中的任务顺序具有更高的P、R、F1值，验证了本文提出方法的有效性。

粗粒度分类	成分	成分分类	探究顺序 P/R/F1	本文方法 P/R/F1
比喻	本体	实在物/动作/抽象概念	0.27/0.27/0.25	**0.63/0.67/0.62**
比拟	比拟成分	动词/名词/形容词/副词	0.34/0.50/0.27	**0.75/0.92/0.78**
夸张	形式	直接/间接/融合	0.30/0.31/0.30	**0.57/0.54/0.54**
夸张	方向	扩大/缩小/超前	0.65/0.47/0.53	**0.83/0.97/0.89**

表 6: 调整任务顺序探究结果

(c) 2024 中国中文信息学会计算语言学专业委员会

Overview of CCL24-Eval Task6: Chinese Essay Rhetoric Recognition and Understanding (CERRU)

Nuowei Liu[1], Xinhao Chen[1], Yupei Ren[1,2], Man Lan[1,2]*, Xiaopeng Bai[2,3], Yuanbin Wu[1,2], Shaoguang Mao[4], Yan Xia[4]

[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]Microsoft Research Asia

{nwliu, 51215901006, ypren}@stu.ecnu.edu.cn
{mlan, ybwu}@cs.ecnu.edu.cn, xpbai@zhwx.ecnu.edu.cn
{shaoguang.mao, yanxia}@microsoft.com

Abstract

Rhetoric is fundamental to the reading comprehension and writing skills of primary and middle school students. However, current work independently recognize single coarse-grained categories or fine-grained categories. In this paper, we propose the CCL24-Eval Task6: Chinese Essay Rhetoric Recognition and Understanding (CERRU), consisting of 3 tracks: (1) Fine-grained Form-level Categories Recognition, (2) Fine-grained Content-level Categories Recognition and (3) Rhetorical Component Extraction. A total of 32 teams registered to participate in CERRU and 9 teams submitted evaluation results, with 7 of these teams achieving an overall score that surpassed the baseline.

1 Introduction

In the learning process of primary and middle school students, rhetoric is not only a core component of reading comprehension and writing skills, but also an indispensable element in shaping excellent literary works. Recognizing and understanding the use of rhetoric in students' essays can help improve their expressive abilities in writing. However, this requires a significant amount of manual effort, posing challenges to teachers in term of essay assessment and instruction. With the development of education and the widespread availability of the Internet, many researchers have begun to explore the use of computer technology for automatic grading of essays (Rudner et al., 2006), where the use of rhetoric is a crucial part of teachers' essay grading.

The use of rhetoric in essays reflects the level of literacy grace and language expression ability (Guo et al., 2018), which is significant for helping teachers assess the quality of essays and guide students in improving their expressive skills. In recent years, research on the recognition of rhetoric in essays often employs alignment strategies and other rules to perform coarse-grained recognition of rhetoric such as parallelism and metaphor from the perspectives of sentence structure and semantic information (Niculae, 2013; Song et al., 2016) or designs model structures specifically to recognize simile (Liu et al., 2018; Zeng et al., 2020). These efforts independently recognize different major rhetorical categories such as metaphor, personification, hyperbole and parallelism, lacking universality. On the other hand, they are coarse-grained and lack fine-grained definitions of rhetorical categories. Furthermore, beyond recognizing rhetorical categories in sentences, some researches treat the understanding of rhetoric as a component extraction task, for example, extracting the tenor and the vehicle from metaphorical sentences (Wang et al., 2022). These researches lack definitions for the rhetorical subjects and contents of other rhetorical devices, and thus cannot provide systematic and comprehensive guidance and feedback on the essays of elementary and middle school students.

Therefore, to address the aforementioned challenges, we propose the CCL24-Eval Task6: **C**hinese **E**ssay **R**hetoric **R**ecognition and **U**nderstanding (CERRU). The dataset for the evaluation originates from

*Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

examination essays written by elementary and middle school students whose native language is Chinese. The genres of these essays include narrative and argumentative writing, among others. Our task settings systematically define the fine-grained rhetorical categories found in these essays, recognizing them from both form level and content level based on the linguistic definitions of rhetoric (Li, 2020). Furthermore, we define the subjects and contents of each rhetorical category, which aids teachers in understanding the use of rhetoric at the sentence level in student essays. It also supports elementary and middle school students in practicing appropriate rhetorical techniques in their writing.

CERRU categorizes rhetorical devices into metaphor, personification, hyperbole and parallelism, and further subdivides these four rhetorical categories into fine-grained categories. As shown in Figure 1, CERRU includes 3 tracks, which are

- Track1: Fine-grained Form-level Categories Recognition
- Track2: Fine-grained Content-level Categories Recognition
- Track3: Rhetorical Component Extraction

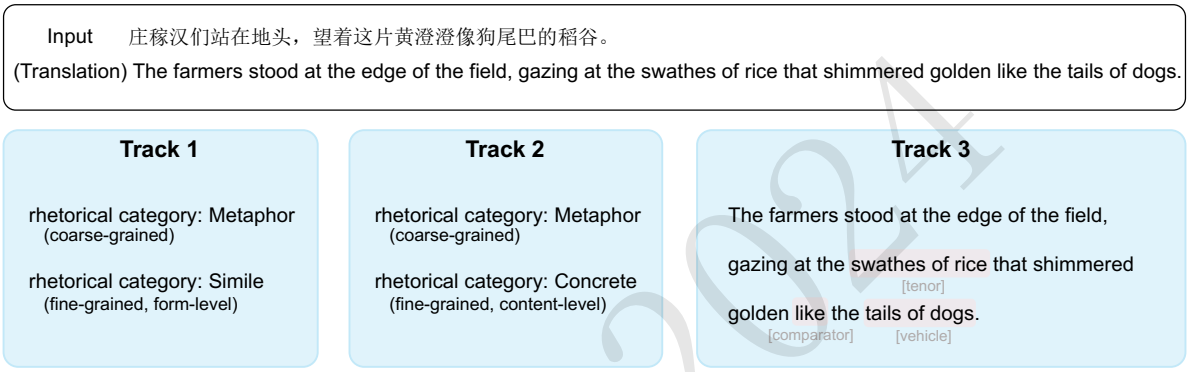


Figure 1: An example of CERRU.

2 Task Descriptions

2.1 Track1: Fine-grained Form-level Categories Recognition

Track1 uses sentences as basic units and categorizes the rhetorical devices into four coarse-grained categories: metaphor, personification, hyperbole, parallelism. As shown in Table 1, each category is further subdivided into fine-grained form-level categories.

- For metaphor, it is subdivided into simile, metaphor and metonymy.
- For personification, it is subdivided into noun, verb, adjective and adverb.
- For hyperbole, it is subdivided into direct hyperbole, indirect hyperbole and mixed hyperbole.
- For parallelism, it is subdivided into structure parallelism and sentence parallelism.

Metaphor			Personification				Hyperbole			Parallelism	
Simile	Metaphor	Metonymy	Noun	Verb	Adjective	Adverb	Direct Hyperbole	Indirect Hyperbole	Mixed Hyperbole	Structure Parallelism	Sentence Parallelism

Table 1: The relationship between coarse-grained categories and fine-grained form-level categories.

Track1 is a multi-label classification problem, involving predicting the coarse-grained rhetorical category and fine-grained form-level category used in a given sentence.

2.2 Track2: Fine-grained Content-level Categories Recognition

Similar to track1, track2 uses sentences as basic units and categorizes the rhetorical devices into four coarse-grained categories: metaphor, personification, hyperbole, parallelism. As shown in Table 2, each category is further subdivided into fine-grained content-level categories.

- For metaphor, it is subdivided into concrete, action and abstract.
- For personification, it is subdivided into personification and anthropomorphism.
- For hyperbole, it is subdivided into amplification, understatement and prolepsis.
- For parallelism, it is subdivided into coordination, subordination and gradation.

Metaphor			Personification		Hyperbole			Parallelism		
Concrete	Action	Abstract	Personification	Anthro-pomorphism	Ampli-fication	Understatement	Prolepsis	Coor-dination	Subor-dination	Gradation

Table 2: The relationship between coarse-grained categories and fine-grained content-level categories.

Track2 is a multi-label classification problem, involving predicting the coarse-grained rhetorical category and fine-grained content-level category used in a given sentence.

2.3 Track3: Rhetorical Component Extraction

Rhetorical components include the described object in the given sentence and the specific content of the description. Extracting these components helps understanding students' use of rhetoric, reflecting their language expression skills. As shown in Table 3, track3 uses sentences as basic units and categorizes the rhetorical components in the sentences into connector, object and content.

- For metaphor-simile, the rhetorical components include comparator, tenor and vehicle. For metaphor-metaphor, the rhetorical components include tenor and vehicle. For metaphor-metonymy, the rhetorical components include vehicle.
- For personification, regardless of form-level category, the rhetorical components include personification object and personification content.
- For hyperbole, regardless of form-level category, the rhetorical components include hyperbole object and hyperbole content.
- For parallelism, regardless of form-level category, the rhetorical components include parallelism marker.

Rhetorical Component	Metaphor			Personification	Hyperbole	Parallelism
	Simile	Metaphor	Metonymy			
Connector	Comparator	-	-	-	-	Parallelism Marker
Object	Tenor	Tenor	-	Personification Object	Hyperbole Object	-
Content	Vehicle	Vehicle	Vehicle	Personification Content	Hyperbole Content	-

Table 3: Rhetorical components of different fine-grained form-level categories.

CCL 2024

3 Datasets

3.1 Dataset Annotation

CERRU collects the eesays used in our dataset from essays written by primary and middle school students for their exams. The collected data covers various geners of writing, such as character and scene description.

During the process of dataset annotation, four annotators participated, including undergraduates and postgraduates majoring in linguistics. First, preliminary annotation guidelines were established. Second, the four annotators jointly pre-annotated 50 essays. After completing the pre-annotation, the inter-annotator agreement of the annotation results was checked, and the annotation guidelines were further revised based on the results. Finally, each of the four annotators formally annotated about 140 essays, totaling 503 essays. Specifically, the last 20 essays annotated by Annotator A were identical to the first 20 essays annotated by Annotator B, and so on. The overlapped annotations were used to check the inter-annotator agreement of the formal annotation results.

3.2 Dataset Statistics

Track1, track2 and track3 share the same training set, validation set and test set while each track has distinct annotations. Track1 and track2 focus on fine-grained form-level and content-level categories respectively while track3 focus on rhetorical components. The size of dataset in shown in Table 4 and the portion of the test set used for evaluation constitutes approximately 10% of the entire test set.

#Training set	#Validation set	#Test set
634	225	5000

Table 4: Statistics of dataset used in CERRU.

4 Evaluation Metrics

In this section, we introduce the metrics used in CERRU. F_1 refers to macro-F1 score in track1, track2 and track3. The overall score of CERRU is the arithmetic mean of track1, track2 and track3.

4.1 Track1: Fine-grained Form-level Categories Recognition

As displayed in Equation 1, the overall F1 score of track1 is comprised of two parts: the F1 score of coarse-grained categories and fine-grained form-level categories.

$$F_1 = 0.3 \times F_1^{\text{rhetorical}} + 0.7 \times F_1^{\text{form}} \tag{1}$$

where $F_1^{\text{rhetorical}}$ denotes the F1 score of coarse-grained categories and F_1^{form} denotes the F1 score of fine-grained form-level categories.

4.2 Track2: Fine-grained Content-level Categories Recognition

As displayed in Equation 2, the overall F1 score of track2 is comprised of two parts: the F1 score of coarse-grained categories and fine-grained content-level categories.

$$F_1 = 0.3 \times F_1^{\text{rhetorical}} + 0.7 \times F_1^{\text{content}} \tag{2}$$

where $F_1^{\text{rhetorical}}$ denotes the F1 score of coarse-grained categories and F_1^{content} denotes the F1 score of fine-grained content-level categories.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.3 Track3: Rhetorical Component Extraction

As displayed in Equation 3, the overall F1 score of track3 is comprised of three parts: the F1 score of connectors, the F1 score of objects and the F1 score of contents.

$$F_1 = \frac{1}{3} \times F_1^{\text{connector}} + \frac{1}{3} \times F_1^{\text{object}} + \frac{1}{3} \times F_1^{\text{content}} \tag{3}$$

where $F_1^{\text{connector}}$, F_1^{object} and F_1^{content} denotes the F1 score of connectors, objects and contents respectively.

5 Baselines

In this section, we introduce the baseline approaches used in CERRU and the scores on track1, track2 and track3.

For track1 and track2, we take both the tasks as multi-label classification problems and fine-tune RoBERTa [1] (Liu et al., 2019) on the training set. A Dropout (Srivastava et al., 2014) layer and a linear layer are concatenated to RoBERTa, and the output after applying sigmoid function is used to represent the probabilities of each category in the given sentence. For track3, we take the task as named entity recognition and fine-tune RoBERTa on the training set. A Dropout layer and a linear layer are concatenated to RoBERTa. Furthermore, we utilize the IOB tagging format (Ramshaw and Marcus, 1999) to tag the comparator, tenor, vehicle, personification object, personification content, hyperbole object, hyperbole content and parallelism marker. The output from RoBERTa after applying argmax function is represented as an entity tag on each token. Subsequently, the consecutive "`B-`" prefix tag and "`I-`" prefix tag are combined to represent the corresponding rhetorical components.

As shown in Table 5, we report the baseline scores on both the validation set and the test set for reference.

Track	F1 (on validation set) (%)	F1 (on test set) (%)
Track1	38.11	45.66
Track2	35.28	56.89
Track3	21.29	20.85

Table 5: Baseline results on the validation set and the test set.

6 Results

In this section, we first discuss the overall results, including the statistics of the participating teams and their scores on each track (See Section 6.1). Considering the correlation between different tracks, most of the teams choose to combine the dataset from different tracks for joint training. Therefore, we then discuss the approaches they use respectively (See Section 6.2 - Section 6.6). Finally, an overall analysis will be discussed in Section 6.7.

6.1 Overall Results

For CCL24-Eval Task6, a total of 32 teams registered to participate in CERRU. Utimately, 9 teams submitted evaluation results and obtained valid scores, with 7 of these teams achieving an overall score that surpassed the baseline. Details are listed in Table 6.

Furthermore, the statistics on the usage of LLMs, external data and data augmentation methods by the top 5 teams based on their overall scores are listed in Table 7.

[1] https://huggingface.co/uer/chinese_roberta_L-12_H-768

Team Name	Track1 (%)	Track2 (%)	Track3 (%)	Score (%)
Zhengzhou University (ZZU)	61.30	62.29	75.28	66.29
Beijing Language and Culture University (BLCU)	59.20	60.92	77.96	66.03
iHuman Inc.	53.77	60.15	68.26	60.72
Central China Normal University (CCNU)	50.86	55.81	73.75	60.14
Zhongyuan University of Technology (ZUT1)	51.48	55.11	69.51	58.70
Zhongyuan University of Technology (ZUT2)	51.48	55.82	57.00	54.77
Institute of Computing Technology (ICT)	50.23	52.78	54.22	52.41
baseline	**45.66**	**56.89**	**20.85**	**41.13**
Individual Team	40.00	52.66	-	37.84
Jiangxi Normal University (JXNU)	39.60	39.13	-	33.19

Table 6: Scores of the participating teams. ”-” indicates that the team did not submit evaluation results on the track, and the overall score is calculated based on the baseline.

Team Name	LLMs	External Data	Data Augmentation
ZZU	✓	✗	✓
BLCU	✓	✓	✗
iHuman Inc.	✓	✗	✗
CCNU	✗	✗	✗
ZUT1	✓	✗	✓

Table 7: Statistics on the usage of LLMs, external data and data augmentation methods. ”LLMs” indicates whether to use Large Language Models. ”External Data” indicates whether data outside the provided dataset for CERRU is used. ”Data Augmentation” indicates whether any augmentation is performed on the provided dataset for CERRU.

CCL 2024

6.2 Team ZZU

ZZU employ LoRA (Hu et al., 2021) method for instruction fine-tuning Yi (Young et al., 2024) and Qwen1.5 (Team, 2024). Noticing that the three tracks share the same training set, validation set and test set, differing only in the respective annotations, they combine the instruction datasets from the three tracks and perform multi-task fine-tuning on the mixed dataset. Moreover, inspired by the LLM2LLM method (Lee et al., 2024), they record error-prone samples in track1 and track2 from the validation set during the fine-tuning process, using a more powerful LLM as a teacher model to generate synthetic data based on these error-prone samples. Additionally, to further enhance model performance, they explore a model ensemble approach to classify coarse-grained and fine-grained categories using LLMs.

6.3 Team BLCU

To expand the dataset, BLCU first adopt GLGC (A Corpus for Chinese Literary Grace Evaluation) (Li et al., 2022), a publicly available corpus, and some online data as the external data. Then, they propose an approach for Chinese rhetoric recognition and understanding with collaborative decision-making between large and small language models under the guidance of human thinking. They redefine the order of tasks and select the large and small language models in the specific process to reach the local optimization at each step. In particular, they use BERT (Devlin et al., 2018) to output the probabilities of each category in a given sentence and employ GPT-4 (Achiam et al., 2023) to predict the result using the output after applying the softmax function.

6.4 Team iHuman Inc.

iHuman Inc. directly employ LoRA (Hu et al., 2021) for fine-tuning Qwen-7B (Bai et al., 2023). For track1 and track2, they first predict the coarse-grained categories of each given sentence and then predict

the corresponding fine-grained categories of the given sentence. To enhance the robustness of their approach, multiple prompts are pre-defined. For track3, noticing that the predicted output may not be exactly the same as in the given sentence, they use a substring comparison method based on edit distance. Particularly, when the edit distance between the output and a substring of the input sentence is less than a certain threshold, they consider them to be identical and directly use the corresponding substring as the result.

6.5 Team CCNU

CCNU employ the unified multi-task learning architecture to fully incorporate the correlation between the three tracks. First, they use the Transformer (Vaswani et al., 2017) pre-trained model as shared feature encoder to represent the sentences. The framework they propose consists of four sub-tasks: rhetorical device recognition, form-level category recognition, content-level category recognition and rhetorical component extraction, which enhance each other's fusion learning. Finally, the aforementioned sub-tasks are integrated into a unified model through parameter sharing.

6.6 Team ZUT1

ZUT1 employ an ensemble model combining BERT (Devlin et al., 2018) and ERNIE (Sun et al., 2019) for track1 and track2. Furthermore, a data augmentation approach is used to enable the model to learn more relevant features from the imbalanced dataset. In particular, they apply methods such as synonym replacement, random word insertion and similar sentence generation to the labeled data. Additionally, they add the prediction generated by the model on unlabeled data back into the training set, thereby increasing the size of training set to enhance the performance of the model. For track3, they use ChatGLM-6B (Zeng et al., 2022) and Qwen-7B (Bai et al., 2023) with QLoRA (Dettmers et al., 2024) fine-tuning method to extract the rhetorical components from the given sentence.

6.7 Overall Analysis

Overall, the teams using LLMs perform better on most tracks compared to those using other approaches while CCNU also achieve a competitive performance. Additionally, the use of external data and data augmentation methods also significantly improves the performance. Most of the teams use LoRA or QLoRA to fine-tune the LLMs, while the methods of data augmentation vary between the teams. Furthermore, several teams improve the overall performance by effectively defining new sub-tasks and rearrange the order in which these sub-tasks are addressed.

7 Conclusion

In this paper, we propose the CCL24-Eval Task6: **C**hinese **E**ssay **R**hetoric **R**ecognition and **U**nderstanding (CERRU), consisting of 3 tracks: (1) Fine-grained Form-level Categories Recognition, (2) Fine-grained Content-level Categories Recognition and (3) Rhetorical Component Extraction. A total of 32 teams registered to participate in CERRU and 9 teams submitted evaluation results and obtained valid scores. Furthermore, we discuss the approaches used by the top 5 teams based on their overall scores. The results demonstrate that the usage of LLMs and data augmentation methods help improve the overall scores.

Acknowledgements

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004) and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jingjin Guo, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. 2018. Attention-based bilstm network for chinese simile recognition. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 144–147. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipali, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. 2024. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.

Yi Li, Dong Yu, and Pengyuan Liu. 2022. Clgc: A corpus for chinese literary grace evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5548–5556.

Qingrong Li. 2020. *Modern Practical Chinese Rhetoric*. BEIJING BOOK CO. INC. In Chinese.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vlad Niculae. 2013. Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 110–114.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Lawrence M Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to identify sentence parallelism in student essays. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 794–803.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Qwen Team. 2024. Introducing qwen1.5, February.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaoyue Wang, Linfeng Song, Xin Liu, Chulun Zhou, and Jinsong Su. 2022. Getting the most out of simile recognition. *arXiv preprint arXiv:2211.05984*.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9515–9522.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

CCL 2024

System Report for CCL24-Eval Task 7: Assessing Essay Fluency with Large Language Models

Haihong Wu,Chang Ao,Shiwen Ni
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{ haihongw@mail.ustc.edu.cn, c.ao@siat.ac.cn,sw.ni@siat.ac.cn }

Abstract

With the development of education and the widespread use of the internet, the scale of essay evaluation has increased, making the cost and efficiency of manual grading a significant challenge. To address this, The Twenty-third China National Conference on Computational Linguistics (CCL2024) established evaluation contest for essay fluency. This competition has three tracks corresponding to three sub-tasks. This paper conducts a detailed analysis of different tasks, employing the BERT model as well as the latest popular large language models Qwen to address these sub-tasks. As a result, our overall scores for the three tasks reached 37.26, 42.48, and 47.64.
Keywords: Large Language Models, Assessing Essay Fluency, Fine-tuning

1 Introduction

CCL points out that with the development of education, the quantity of essay texts is gradually increasing. The cost and efficiency of manually grading essays have become a major challenge. Many research institutions aim to provide objective, accurate, and timely scoring and feedback by analyzing the language, content, and structure of essays. Among these factors, the fluency of expression is an important aspect of essay evaluation for teachers. Therefore, this competition revolves around evaluating the fluency of essay texts. There are three tracks in this competitionincluding:

Track 1: Identification of sentence types in primary and secondary school essays. the identification of incorrect sentence types in primary and secondary school essays is a multi label classification problem, which predicts which types of incorrect sentences a sentence is. The sick sentence type label contains both lexical and syntactic errors. In this evaluation task, a total of 5 coarse-grained error types and 14 fine-grained error types were defined.

Track 2: Revision of sentence types in primary and secondary school essays. the task of rewriting incorrect sentences in primary and secondary school essays is a text generation task, which involves inputting incorrect sentences and outputting modified ones.

Track 3: Evaluation of fluency in primary and secondary school essays. The fluency rating task for primary and secondary school essays is a multi classification task, which involves inputting an essay and outputting its level of fluency. This evaluation task defines three fluency levels: excellent, average, and failing.

Three tracks provide more basis for evaluating the fluency of primary and secondary school essays and offering higher-quality assessments.

2 Model and Methods

2.1 Model

In track1 and track2, fine-tuning the Qwen 7b model has proven highly effective, leveraging its superiority in processing Chinese texts to achieve excellent results. In addition, the BERT model (Devlin et al. (2018)) has a much smaller number of parameters than large language models, but it performs extremely

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 262-268, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

well in classification tasks. In track 3, the BERT large model was employed for the classification task of essay fluency, and the results demonstrated the excellent performance of the BERT model as well.

2.1.1 Qwen

The Qwen-1.5-7B-Chat (Bai et al. (2023)) model is a large-scale language model based on the transformer (Vaswani et al. (2017)) architecture, designed primarily for conversational applications. With a parameter count of 7 billion, it belongs to the class of state-of-the-art transformer-based models, known for their ability to handle complex natural language understanding and generation tasks. The Qwen-1.5-7B-Chat model is built on the transformer architecture, which enables it to capture long-range dependencies in text data efficiently. The model is trained on diverse datasets encompassing various topics and language styles, ensuring robust language understanding and generation capabilities across different contexts.

2.1.2 Bert

The BERT-Large model (Devlin et al. (2018)) is a high-capacity language model based on the transformer architecture, designed for a wide range of natural language processing tasks. With a parameter count of approximately 340 million, it belongs to the class of state-of-the-art transformer-based models, renowned for their proficiency in handling intricate language understanding and generation tasks. The BERT-Large model leverages the transformer architecture's capability to efficiently capture long-range dependencies in text data. It is pre-trained on extensive datasets covering diverse topics and language styles, ensuring robust language understanding and generation abilities across various contexts.

To incorporate a formulaic description into this paragraph, we can highlight the pre-training process mathematically:

$$\mathbf{M}_{\text{BERT-Large}}(\theta) = \arg\min_{\theta} \sum_{(x,y)\in\mathcal{D}_{\text{pre}}} \mathcal{L}_{\text{BERT}}(\mathbf{M}_{\text{BERT-Large}}(x;\theta), y)$$

Here, $\mathbf{M}_{\text{BERT-Large}}(\theta)$ represents the BERT-Large model with parameters θ. $\mathcal{D}_{\text{pre}}$ denotes the pre-training dataset consisting of diverse texts. $\mathcal{L}_{\text{BERT}}$ is the loss function that measures the discrepancy between the model's predictions $\mathbf{M}_{\text{BERT-Large}}(x;\theta)$ and the true labels y.

This formulation emphasizes that the BERT-Large model is trained by minimizing the loss function $\mathcal{L}_{\text{BERT}}$ over a comprehensive pre-training dataset $\mathcal{D}_{\text{pre}}$, ensuring its readiness to comprehend and generate language across varied domains and styles.

2.2 Fine-tuning

Fine-tuning (Friederich (2017)) is a term commonly used in the fields of machine learning (Jordan and Mitchell (2015)) and deep learning (LeCun, Bengio, and Hinton (2015)). It refers to the process of further optimizing and adjusting a pre-trained model to enhance its performance on a specific task. Full-Parameter fine-tuning of large language models leverages the vast amounts of knowledge embedded in the pre-trained model while adapting it to perform better on specific tasks, such as text classification, question answering (Lv et al. (2023)). Previous work has also shown that fine-tuning the BERT model yields superior performance on multi-class classification tasks (Sun et al. (2019)).

The fine-tuning process can be described mathematically as follows:

1. Pre-trained Model: Let $\mathbf{M}_{\text{pre}}(\theta)$ be the pre-trained model with parameters θ. This model has been trained on a large corpus of general data.

2. Task-specific Dataset: Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the task-specific dataset, where x_i represents the input data and y_i represents the corresponding labels.

3. Loss Function: Define a task-specific loss function $\mathcal{L}(\mathbf{M}_{\text{pre}}(x_i;\theta), y_i)$ that measures the difference between the model's predictions and the true labels.

4. Optimization Objective: Fine-tuning aims to minimize the loss function over the task-specific

dataset. This can be expressed as:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(\mathbf{M}_{\text{pre}}(x_i;\theta), y_i)$$

We conducted fine-tuning with training data on the Qwen-1.5-7B-Chat model. This fine-tuning process enhances the model's performance on specific tasks by adapting its parameters to the target dataset. Given the Qwen model's generally strong capability in understanding Chinese, we decided to fine-tune the Qwen model in the track1 and track2 to achieve assessment and modification of essay sentences.

Let $\mathbf{M}_{\text{Qwen}}$ denote the Qwen-1.5-7B-Chat model. The fine-tuning process can be formulated as:

$$\theta^*_{\text{Qwen}} = \arg\min_{\theta_{\text{Qwen}}} \sum_{i=1}^{N} \mathcal{L}_{\text{Qwen}}(\mathbf{M}_{\text{Qwen}}(x_i;\theta_{\text{Qwen}}), y_i)$$

Fine-tuning with training data on the BERT-Large model enhances its performance on specific tasks by customizing its parameters to suit the track3 dataset. This process adapts the model's learned representations to better align with the nuances and intricacies of the particular task at hand, resulting in improved task-specific performance and accuracy. Therefore, we decided to fine-tune the BERT-Large-Chinese model in the third track to achieve fluency assessment of essays.

Let $\mathbf{M}_{\text{BERT}}$ denote the BERT-Large model. The fine-tuning process for the BERT-Large model can be formulated as:

$$\theta^*_{\text{BERT}} = \arg\min_{\theta_{\text{BERT}}} \sum_{i=1}^{N} \mathcal{L}_{\text{BERT}}(\mathbf{M}_{\text{BERT}}(x_i;\theta_{\text{BERT}}), y_i)$$

In summary, the fine-tuning process involves adapting the parameters θ of pre-trained models $\mathbf{M}_{\text{pre}}$ to minimize the task-specific loss function over the dataset $\mathcal{D}$, thereby improving the model's performance on specific tasks.

2.3 multi-prompt context learning

Multi-prompt context learning, as discussed in Chen et al. (2024), is an innovative approach in natural language processing (NLP) that aims to enhance a model's capability to understand and process textual contexts. Traditional NLP models often operate with a single context prompt or focus on a single task, which can limit their ability to fully grasp the nuances of complex textual information. In contrast, multi-prompt context learning introduces multiple context prompts, enabling the model to better comprehend and adapt to various facets of the input data.

In the inference phase of the track2, a multi-prompt approach was employed. This method involves providing the model with several prompts or questions, which allows it to generate more comprehensive and diverse responses. By leveraging multiple prompts, the model can investigate different dimensions of the input data, leading to richer and more accurate inferences. The theoretical underpinning of this approach can be expressed with the following formalism:

Let $P = \{p_1, p_2, \ldots, p_n\}$ denote the set of prompts provided to the model, where p_i represents an individual prompt. The model's response R can be described as a function f of the input data D and the set of prompts P:

$$R = f(D, P) = f(D, \{p_1, p_2, \ldots, p_n\})$$

Each prompt p_i explores a different aspect of the input data D, contributing to the overall understanding and response generation. The individual responses r_i corresponding to each prompt p_i can be aggregated to form a final comprehensive response R:

$$r_i = f(D, p_i) \quad \text{for} \quad i = 1, 2, \ldots, n$$

CCL 2024

The aggregation function g combines these individual responses r_i to produce the final response R:

$$R = g(r_1, r_2, \ldots, r_n)$$

This aggregation can be achieved through various techniques such as averaging, voting, or more sophisticated ensemble methods. The multi-prompt approach thereby allows the model to harness diverse perspectives from the input data, enhancing the overall quality and accuracy of its inferences.

By integrating multiple prompts, the model can effectively:

1. Capture a wider range of contextual information.
2. Mitigate the risk of bias associated with single-prompt models.
3. Improve robustness and reliability of the generated responses.

In summary, multi-prompt context learning represents a significant advancement in NLP, facilitating more nuanced and precise understanding of textual data through the use of multiple, complementary prompts. This approach not only broadens the scope of information the model can process but also enhances its adaptability to varied and complex contexts.

3 Experiments

During the experiment, we used the transformer library to load the Qwen-7b and Bert-large-chinese pre-training model, trained the model with the datasets announced by the competition organizer, and optimized the model parameters through backpropagation and gradient descent algorithms. The hyperparameter settings during model training are shown in table1 and table2.

Hyperparameter	Value
Train Epochs	3.0
Learning Rate	5×10^{-5}
Batch Size	16

Table 1: Hyperparameters of Training Qwen

Hyperparameter	Value
Train Epochs	15
Learning Rate	1×10^{-5}
Batch Size	8

Table 2: Hyperparameters of Training BERT

3.1 Evaluation Results for Track 1

Metric	Score
Micro-F1	48.84
Macro-F1	25.68
Coarse-grained Micro-F1	58.06
Fine-grained Micro-F1	39.62
Coarse-grained Macro-F1	36.25
Fine-grained Macro-F1	15.10
Score	37.26

Table 3: Evaluation Metrics Scores

3.1.1 metric results

The Micro-F1 score of 48.84 indicates a moderate level of overall accuracy, considering both precision and recall across all instances. In contrast, the Macro-F1 score of 25.68 is significantly lower, suggesting that the model's performance varies widely across different classes, with some classes potentially having much poorer performance.

The Coarse-grained Micro-F1 score of 58.06 is higher than the overall Micro-F1 score, indicating that the model performs better when evaluated on broader categories. Similarly, the Coarse-grained Macro-F1 score of 36.25 is higher than the overall Macro-F1 score, reinforcing the idea that the model's performance is more consistent at a higher level of categorization.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 262-268, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The Fine-grained Micro-F1 score of 39.62 and the Fine-grained Macro-F1 score of 15.10 are both lower than their coarse-grained counterparts. This suggests that the model struggles with finer distinctions between categories, which is a common issue in complex classification tasks.

The overall score of 37.26 provides a single summary metric, but it should be interpreted in the context of the more detailed metrics. It reflects the model's average performance but does not capture the nuances revealed by the other scores.

3.1.2 Analysis

The evaluation metrics reveal a model that performs moderately well overall but has significant room for improvement, especially in distinguishing fine-grained categories. Future work should focus on improving the model's ability to handle finer distinctions between classes, as well as addressing any imbalances in class performance that are suggested by the low Macro-F1 scores.

3.2 Evaluation Results for Track 2

Metric	Score
EM	13.93
Bert PPL	15.82
Levenshtein	1.90
BLEU-4	89.60
BertScore	97.34
Precision	45.22
Recall	27.25
F0.5	39.95
Score	42.48

Table 4: Evaluation Metrics Scores

3.2.1 metric results

The result is shown on table 2,The evaluation metrics provide a comprehensive insight into the model's performance. The Exact Match (EM) score of 13.93 indicates that the model's output exactly matches the reference answers only a small fraction of the time, highlighting a need for improvement in generating fully correct answers. The perplexity (Bert PPL) score of 15.82 suggests moderate fluency and coherence in sentence generation, but there is room for refinement. A Levenshtein distance of 1.90 reflects a high similarity between the model's outputs and the reference answers, though not perfectly aligned. The BLEU-4 score of 89.60 demonstrates the model's excellent ability to capture the phrase structures of the reference answers. A BertScore of 97.34 shows that the model's generated answers are semantically very close to the reference, indicating strong contextual understanding. However, a precision score of 45.22 reveals that less than half of the model's generated answers are correct, and a recall score of 27.25 indicates significant omissions of correct answers. The F0.5 score of 39.95, which weighs precision more heavily, further underscores the need for improvement in balancing accuracy and completeness in the model's outputs.

3.2.2 analysis

The model performs well in semantic understanding and generating coherent text (with high BertScore and BLEU-4 scores), but there is significant room for improvement in exact matching and accurate generation (as indicated by lower EM and Precision scores). The relatively high Levenshtein score suggests a degree of similarity between the model's outputs and the reference answers, but they are not exactly the same. The perplexity score indicates moderate generation capability but not optimal performance. Overall, the model shows promise in generating high-quality text but needs to enhance its accuracy and comprehensive coverage of reference answers.

CCL 2024

Finally, the overall score of 42.48 further highlights that while the model has some strengths, particularly in semantic similarity and structure capture, it requires significant enhancements in precision and recall to achieve better performance.

3.3 Evaluation Results for Track 3

Metric	Score
ACC	48.59
Precision	46.42
Recall	43.07
F1	43.44
QWK	0.1338
AvgScore	47.64

Table 5: Evaluation Metrics Scores

3.3.1 metric results

The evaluation metrics indicate that the model's overall performance is moderate but has significant room for improvement. The accuracy (ACC) of 48.59 means that slightly less than half of the model's predictions are correct, pointing to a need for better overall accuracy. The precision of 46.42 shows that less than half of the model's positive predictions are correct, suggesting a fair number of false positives that need reduction. With a recall of 43.07, the model correctly identifies a significant portion of the actual positive cases, but it still misses many true positives. The F1 score, which balances precision and recall, is 43.44, reflecting the need for improvement in both areas to achieve better overall effectiveness. The Quadratic Weighted Kappa (QWK) score of 0.1338 indicates poor agreement between the model's predictions and the true labels beyond random chance, highlighting the necessity for significant enhancements in prediction quality. Finally, the average score (AvgScore) of 47.64 summarizes the model's performance across different metrics, suggesting that while the model performs moderately well in some aspects, it generally underperforms and requires improvements in various areas to achieve better overall effectiveness.

3.3.2 analysis

The model demonstrates moderate performance with an accuracy of 48.59, indicating that slightly less than half of its predictions are correct. The precision of 46.42 and recall of 43.07 suggest that the model produces a fair number of false positives and misses a significant number of true positives. The F1 score of 43.44 highlights the need for balanced improvements in both precision and recall. The low QWK score of 0.1338 points to poor agreement with the true labels, indicating that the model's predictions are often inaccurate beyond what could be attributed to random chance.

Finally, the average score of 47.64 underscores the overall moderate performance of the model, suggesting that substantial enhancements are needed across all evaluated metrics to achieve better accuracy, precision, recall, and agreement with true labels.

4 Conclusion

In this experiment, we conducted a comprehensive assessment of the fluency of essays and validated it through a competition format. The results indicate that large language models perform exceptionally well in text correction for evaluating essay fluency, while BERT models are more effective in essay fluency classification. In future research, we will further explore evaluation methods for essay fluency, integrating theories from linguistics, psychology, and related fields, in order to seek more scientifically effective evaluation metrics and methods.

CCL 2024

References

Bai, Jinze et al. (2023). *Qwen Technical Report*. arXiv: 2309.16609 [cs.CL].

Chen, Haoran et al. (2024). "Multi-prompt alignment for multi-source unsupervised domain adaptation". In: *Advances in Neural Information Processing Systems* 36.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Friederich, Simon (2017). "Fine-tuning". In: *The Stanford encyclopedia of philosophy*.

Jordan, Michael I and Tom M Mitchell (2015). "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245, pp. 255–260.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

Lv, Kai et al. (2023). "Full parameter fine-tuning for large language models with limited resources". In: *arXiv preprint arXiv:2306.09782*.

Sun, Chi et al. (2019). "How to fine-tune bert for text classification?" In: *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*. Springer, pp. 194–206.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

CCL 2024

System Report for CCL24-Eval Task 7: Multi-Error Modeling and Fluency-Targeted Pre-training for Chinese Essay Evaluation

Jingshen Zhang†, Xiangyu Yang†, Xinkai Su, Xinglu Chen, Tianyou Huang, Xinying Qiu‡
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China
{20211003207,20221050039,20210602169,20221003046,20231010040}@gdufs.edu.cn
xy.qiu@foxmail.com

Abstract

This system report presents our approaches and results for the Chinese Essay Fluency Evaluation (CEFE) task at CCL-2024. For Track 1, we optimized predictions for challenging fine-grained error types using binary classification models and trained coarse-grained models on the Chinese Learner 4W corpus. In Track 2, we enhanced performance by constructing a pseudo-dataset with multiple error types per sentence. For Track 3, where we achieved first place, we generated fluency-rated pseudo-data via back-translation for pre-training and used an NSP-based strategy with Symmetric Cross Entropy loss to capture context and mitigate long dependencies. Our methods effectively address key challenges in Chinese Essay Fluency Evaluation.

1 Introduction

With the growing integration of smart education and deep learning technologies, automated text evaluation systems have become increasingly critical. These systems aim to accurately and efficiently assess students' compositions, reduce teachers' workload, and provide instant feedback for error correction and writing improvement. The China National Conference on Computational Linguistics (CCL-2024) has presented the Chinese Essay Fluency Evaluation (CEFE) as a public assessment task. This task focuses on three primary text evaluation strategies: error sentence type recognition, error sentence correction, and essay fluency evaluation, offering an in-depth research direction for the field of automatic text evaluation.

This system report provides an overview of our work on the CEFE evaluation task, highlighting the different strategies employed for each track:

- For Track 1, Error Sentence Type Recognition, we analyzed two fine-grained errors, utilized a binary classification model for prediction optimization, compared and selected training corpora, and trained a coarse-grained model based on the Chinese Learner 4W corpus.
- For Track 2, Error Sentence Correction, we adopted a strategy that involved constructing a pseudo-dataset containing sentences with multiple error types to enhance model performance.
- For Track 3, Essay Fluency Evaluation, we achieved first place by employing back-translation techniques to construct pseudo-data with triple-labeled fluency ratings for pre-training and adapting an NSP-based strategy to effectively utilize contextual information and avoid long sequence dependencies.

The remainder of this report is structured as follows: Section 2 presents related research in the field of Chinese composition fluency evaluation. Sections 3, 4, and 5 detail

† Equal contribution, ‡ Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 269–277, Taiyuan, China, July 25 – 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

our methodologies, experiments, and results for Tracks 1, 2, and 3, respectively. Finally, Section 6 concludes the report with an analysis of our findings, discusses the limitations of our work, and potential future research directions. Our code and data are available at https://github.com/astro-jon/ccl2024-coherence.

2 Related Research

The field of Chinese composition fluency evaluation has gained significant attention from researchers in recent years, with the three different track directions involved in this review being popular topics for related research.

2.1 Error Sentence Type Recognition

Error sentence type recognition has been a focus of many studies. Zhang et al., (2020) combined Graph Convolutional Networks (GCN) and Transformers for Chinese grammatical error detection, leveraging the strengths of both architectures to improve performance. Wang et al. (2023) proposed a multi-granularity approach for Chinese grammar error detection and correction, utilizing character-level, word-level, and sentence-level information to enhance the model's ability to identify and correct various types of errors.

2.2 Error Sentence Correction

Error sentence correction has also received significant attention in recent research. Li et al. (2022) proposed a Sequence-to-Action (S2A) module that combines source and target sentences as inputs to automatically generate token-level action sequences for predicting editing operations, effectively integrating the advantages of sequence-to-sequence (seq2seq) models and sequence-tagging models to mitigate the overcorrection problem and improve the performance of the syntactic error correction task. Wu and Wu (2022) introduced a new framework for Chinese grammatical error correction that addresses both spelling and grammar errors, utilizing a two-stage approach that first corrects spelling errors and then focuses on grammatical error correction. Zhou et al. (2023) proposed decoding interventions to improve seq2seq grammatical error correction models, focusing on enhancing the decoding process to generate more accurate and fluent corrections.

2.3 Essay Fluency Evaluation

In addition to error correction, the flow of the text is an equally critical factor in measuring the quality of the text. Mesgar and Strube (2018) proposed a neural local coherence model for text quality assessment that captures the flow of semantic connections between neighboring sentences based on the most similar semantic states and encodes the pattern of changes in text-perceived coherence. Qiu et al. (2022) explored the potential of coherence and syntactic features in neural models for automatic essay scoring, combining syntactic feature dense embedding with the BERT model and investigating the joint model of coherence, syntactic information, and semantic embedding. Sheng et al. (2024) proposed a novel non-referential coherence measure called BB Score, which is based on Brownian Bridge Theory and evaluates text coherence by measuring the ordered and coherent interactions between sentences.

3 Track 1: Error Sentence Type Recognition

3.1 Methodology

Our methodology for Track 1 employed a hierarchical approach as illustrated in Figure 1:

1. **Token-level error identification:** The approach starts by identifying errors at the token level rather than the sentence level. This step covers various types of errors, including Character-Level Errors (such as missing or incorrect characters), Component Incompleteness (e.g., missing subject), Component Redundancy (e.g., redundant subject), and Component Mismatch (e.g., verb-object mismatch).

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 269-277, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

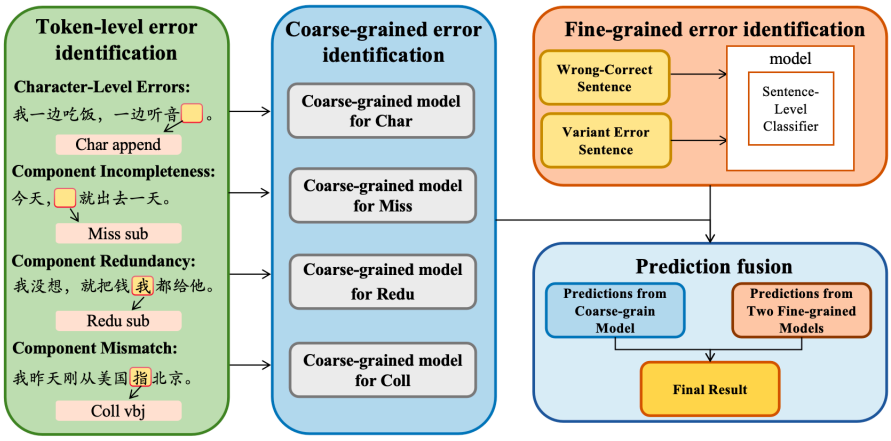


Figure 1: Methodology for Track1

2. **Coarse-grained error modeling**: Four separate coarse-grained error models are constructed to handle different error types: a model for Character-Level Errors (Char), a model for Component Incompleteness (Miss), a model for Component Redundancy (Redu), and a model for Component Mismatch (Coll).

3. **Fine-grained error categorization**: This step involves further refining the predictions from the coarse-grained models by categorizing and identifying specific fine-grained error types within each broad category. Two approaches are employed for this purpose:

   - **Wrong-Correct Sentence**: Sentences with errors are matched with their corrected counterparts, and the model learns to distinguish between erroneous and correct sentences.
   - **Variant Error Sentence**: Sentences containing specific error types (e.g., misordering) are paired with sentences containing other error types, and the model learns to differentiate between them.

4. **Prediction fusion**: The predictions from the four coarse-grained models and the two fine-grained models (Wrong-Correct Sentence and Variant Error Sentence) are combined to generate the final error identification results. This fusion step ensures that the insights from all the models are integrated to produce the most accurate and comprehensive error analysis.

3.2 Experiment and Results

We used the CSED (8,682 sentences) (Sun et al., 2023) and Chinese Learner 4W (39,989 sentences) (Lu et al., 2020) corpus for pseudo-data construction due to limited official data without token-level labels. For training our coarse-grained models, we trained each model on the respective datasets mentioned above instead of merging them. Subsequently, we selected the better one for each coarse-grained model.

The coarse-grained models were trained on the combined corpora using the chinese-electra-180g-base-discriminator model (Cui et al., 2021) , with a maximum length of 512, 30 epochs, batch size 32, and learning rate 2e-5. We compared uniform (25% each) and full corpus distribution strategies.

For the fine-grained binary classification models, we used chinese-roberta-wwm with maximum length 512, 30 epochs, batch size 2, and learning rate 1e-5. Precision, recall, and micro F1 evaluated performance.

We successfully trained four coarse-grained models on the 4W corpus using non-repetitive pseudo-data construction. For the challenging misordering and redundancy error types, we

CCL 2024

trained fine-grained models on sentence pairs contrasting the target error with others using the public corpus. This approach achieved our best score of 36.47.

4 Track 2: Error Sentence Correction

4.1 Methodology

For Track 2, we observed that the original training set contained sentences with multiple error types, whereas previous pseudo-data construction methods from Wang et al. (2023) introduced only one error type per sentence. To better match the original data, we proposed constructing a pseudo-dataset containing sentences with varying numbers of error types with the following steps:

- Apply Wang et al.'s method to introduce single error types into correct sentences.
- Randomly select 1/5 of those single-error sentences.
- From the selected sentences in Step 2, randomly select another 1/5 and introduce a second error type.
- Repeat Step 3, selecting 1/5 from the previous iteration, to create sentences with up to four error types.

The proportion of constructed data can be estimated using the following formula:

$$Percent_i = C_3^{i-1}(1-p)^{4-i}p^{i-1} \tag{1}$$

where i indicates the number of error types, and p represents the selection parameter, which is set to 1/5 in this context to match the distribution of the original dataset.

Thus we created a diverse pseudo-dataset with sentences containing varying numbers of error types, better reflecting real-world erroneous data. We then trained the real_learner_bart_CGEC encoder-decoder model (Zhang et al., 2023) on this multi-error pseudo-dataset to enhance its ability to correct sentences with numerous errors.

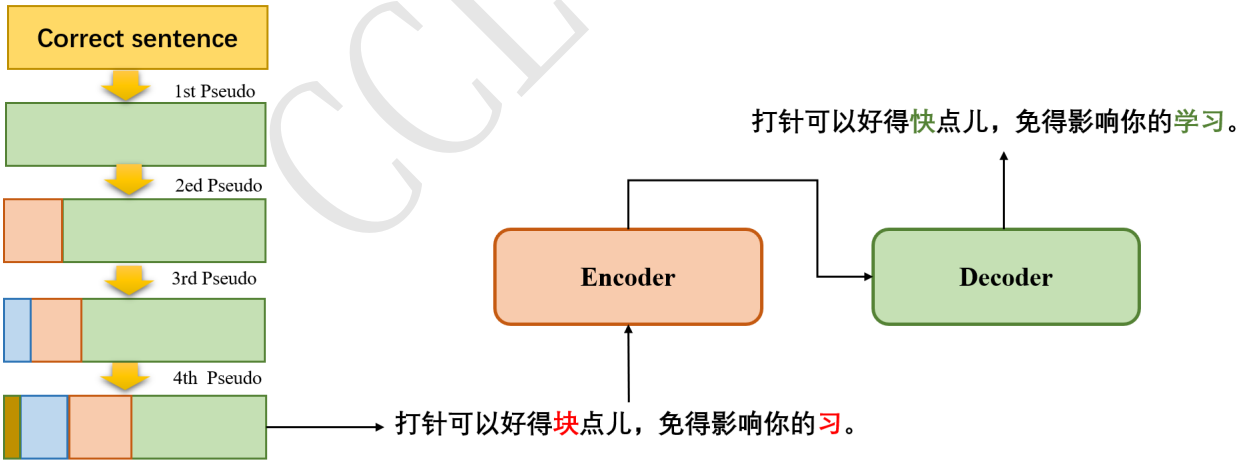


Figure 2: Methodology for Track 2

4.2 Experiment and Results

We used the Chinese Learner 4W Corpus (Lu et al., 2020) for our constructed corpus. We utilized the constructed pseudo-dataset containing numerous errors to train the Bart-base model and subsequently evaluated its performance on the validation set. The evaluation metrics were

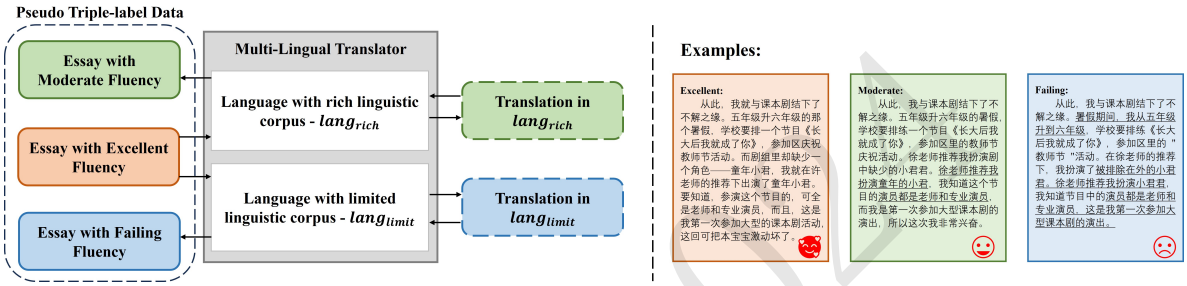
CCL 2024

based on the calculations provided by the official Track 2 guidelines. The best result on the test set for Track 2 was achieved using the real_learner_bart_CGEC model proposed by Zhang et al. (2023), obtaining a final score of 41.09.

5 Track 3: Essay Fluency Evaluation

5.1 Methodology

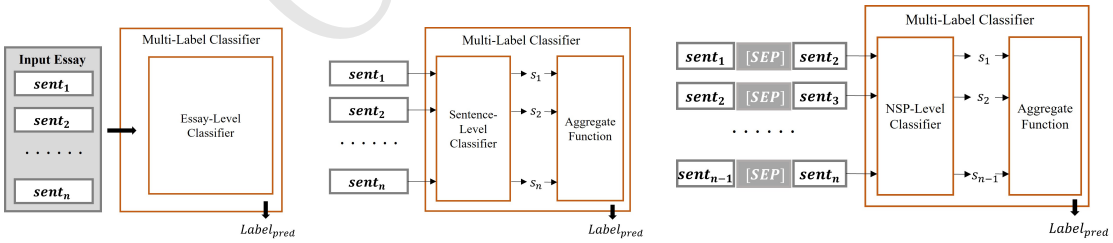
For Track 3, we addressed two key challenges: limited training data and modeling document-level inputs. To augment the scarce labeled data, inspired by Lu et al., (2021), we employed back-translation to construct pseudo-data with triple fluency ratings, as shown in Figure 3(a). Essays back-translated using a resource-rich language, $lang_{rich}$, were labeled as moderately fluent, while those using a resource-poor language, $lang_{limit}$, were labeled as failing fluency. The original essays acted as excellent fluency examples. This back-translated corpus was used for pre-training. And we select English as $lang_{rich}$ and Japanese as $lang_{limit}$ for English has more translation training corpus than Japanese.



(a) Framework based on Back-translation (b) A specific example of back-translation results

Figure 3: (a) Framework based on Back-translation; (b) A specific example of back-translation results

To capture contextual information while avoiding long sequence issues, inspired by Qiu et al. (2022), we adapted an NSP-based training strategy illustrated in Figure 4(c). Instead of inputting the entire essay (Figure 4a) or individual sentences (Figure 4b), we input pairs of neighboring sentences joined by [SEP] tokens. An average aggregation function combined the sentence pair predictions into a final essay fluency score.



(a) Essay-Level Classifier (b) Sentence-Level Classifier (c) Our training strategy: NSP-Level Classifier

Figure 4: Comparison of different training patterns: (a) Essay-Level Classifier, (b) Sentence-Level Classifier, (c) Our training strategy: NSP-Level Classifier

To provide robustness to potential labeling noise from the pseudo-data, we optimized the Symmetric Cross Entropy (SCE) loss (Wang, 2019) defined in Equation 1. SCE incorporates Reverse Cross Entropy (Equation 2) in addition to the standard Cross Entropy (Equation 3),

with tunable hyperparameters and balancing the two components.

$$\ell_{ce} = -\sum_{k=1}^{K} q(k|x) \log(p(k|x)) \tag{2}$$

$$\ell_{rce} = -\sum_{k=1}^{K} p(k|x) \log(q(k|x)) \tag{3}$$

$$\ell_{sce} = \mu\ell_{ce} + \beta\ell_{rce} \tag{4}$$

where ℓ_{ce} is the cross-entropy loss $q(k|x)$ is the ground truth class distribution conditioned on sample x $p(k|x)$ is the predicted distribution over labels by the classifier ℓ_{rce} is the reverse cross-entropy loss ℓ_{sce} is the symmetric cross-entropy loss μ and β are tunable hyperparameters

Other key aspects included oversampling (Appendix D) to handle the imbalanced label distribution and pre-training the RoBERTa model on the back-translated data before fine-tuning on the actual task.

5.2 Experiments and Results

We randomly selected 43 essays with perfect scores from zuowenwang[1] as our fluency excellence examples. We utilized Chinese-roberta-wwm-ext[2] as our base model. Based on ablation experiments (Appendix E), we fixed the hyperparameters μ as 0.1 and β as 1 for the Symmetric Cross Entropy (SCE) loss. With this configuration, we achieved the state of the art on the test set with the score of 51.96 for Track 3.

6 Conclusions

This report presented our approaches and results for the three tracks of the Chinese Essay Fluency Evaluation (CEFE) task at CCL-2024.

For Track 1, we employed a hierarchical method combining token-level error identification, coarse-grained modeling on the Chinese Learner 4W corpus, fine-grained binary models, and prediction fusion to handle both broad and specific error types.

In Track 2, constructing pseudo-data with multiple error types per sentence improved performance in correcting real-world sentences compared to previous single-error methods.

Our Track 3 approach, which achieved first place, utilized back-translated pseudo-data with triple fluency labels, an NSP-based strategy to incorporate context while mitigating long sequence issues, and Symmetric Cross Entropy loss for increased robustness.

By addressing challenges such as limited data, error diversity, long-range dependencies, and label noise, our methods contribute to advancing intelligent assessment of Chinese essays. Potential future directions include cross-lingual generalization, few-shot learning to reduce annotation requirements, and generating more detailed feedback to further enhance student learning.

7 Acknowledgements

This work is partially supported by Guangzhou Science and Technology Plan Project (202201010729), and Guangdong Social Science Foundation Project (GD24CWY11). We thank the reviewers for their helpful comments and suggestions.

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. *Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing.*pages 3504-3514.

[1]https://www.zuowen.com/xsczw/fanwen/

[2]https://huggingface.co/hfl/chinese-roberta-wwm-ext

Jiquan Li, Junliang Guo, Yongxin Zhu, Xin Sheng, Deqiang Jiang, Bo Ren, and Linli Xu. 2022. *Sequence-to-action: Grammatical error correction with action guided sequence generation. In Proceedings of the AAAI Conference on Artificial Intelligence.*pages 10974-10982.

Dawei Lu, Xinying Qiu, and Yi Cai. 2020. *Sentence-level readability assessment for L2 Chinese learning.. Chinese Lexical Semantics: 20th Workshop, CLSW 2019.*pages 381-392.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. *An unsupervised method for building sentence simplification corpora in multiple languages. In Findings of the Association for Computational Linguistics: EMNLP 2021.*pages 227–237.

Mohsen Mesgar, and Michael Strube. 2018. *A neural local coherence model for text quality assessment. In Proceedings of the 2018 conference on empirical methods in natural language processing.*pages 4328-4339.

Xinying Qiu, Shuxian Liao, Jiajun Xie, and Jianyun Nie. 2022. *Tapping the Potential of Coherence and Syntactic Features in Neural Models for Automatic Essay Scoring. In 2022 International Conference on Asian Language Processing (IALP).*pages 407-412.

Zhecheng Sheng, Tianhao Zhang, Chen Jiang, and Dongyeop Kang. 2024. *BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence. In Proceedings of the AAAI Conference on Artificial Intelligence.*pages 14937-14945.

Bo Sun, Baoxin Wang, Yixuan Wang, Wanxiang Che, Dayong Wu, Shijin Wang, and Ting Liu. 2023. *Csed: A chinese semantic error diagnosis corpus. arXiv preprint arXiv:2305.05183.*

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. *Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV).*pages 322–330.

Yixuan Wang, Yijun Liu, Bo Sun, and Wanxiang Che. 2023. *System Report for CCL23-Eval Task 8: Chinese Grammar Error Detection and Correction Using Multi-Granularity Information. In Proceedings of the 22nd Chinese National Conference on Computational Linguistics.*pages 271-281.

Xinyu Wu, and Yunfang Wu. 2022. *From spelling to grammar: A new framework for chinese grammatical error correction. arXiv preprint arXiv:2211.01625.*

Jinhong Zhang. 2020. *Combining GCN and Transformer for Chinese Grammatical Error Detection. arXiv preprint arXiv:2105.09085.*

Yue Zhang, Bo Zhang, Haochen Jiang, Zhenghua Li, Chen Li, Fei Huang and Min Zhang. 2023. *NaSGEC: a Multi-Domain Chinese Grammatical Error Correction Dataset from Native Speaker Texts. arXiv preprint arXiv:2305.16023.*

Houquan Zhou, Yumeng Liu, Zhenghua Li, Min Zhang, Bo Zhang, Chen Li, Ji Zhang and Fei Huang. 2023. *Improving Seq2Seq grammatical error correction via decoding intervention.* In *Findings of the Association for Computational Linguistics: EMNLP 2023,* Singapore.

Appendix A Track 1: Choice of Training Dataset for Coarse-grained Error Modeling

Corpus/Strategy	Char F1	Coll F1	Miss F1	Redu F1
CSED/repeatable	0.388	0.249	0.237	0.292
CSED/non-repetitive	0.401	0.302	0.311	0.320
4W/repeatable	**0.529**	0.513	0.255	**0.501**
4W/repetitive	0.499	**0.632**	**0.298**	0.496

Table 1: F1 scores for models with different corpora and strategies

Table 1 shows the performance of the two corpora on the four models under different strategies. The experimental analysis shows that the models trained on the 4W corpus constructed by Chinese learners generally outperform those trained on the CSED corpus. When comparing corpus strategies, the uniform allocation strategy performs better in both corpora, while the

repetition strategy may lead to overfitting. Therefore, we selected the 4W corpus with the uniform allocation strategy (i.e. **4W/repetitive**) as the best solution for Track1 coarse-grained model training.

Appendix B Track 1: Choice of Strategy for Fine-grained Error Categorization

Table 2 presents the corpus samples, the respective number of sentences constructed using two different corpus construction methods, and the F1 scores obtained from the corresponding trained models.

Method	Fine-Grained	Label	Sentence Number	F1
Wrong - Correct	Misorder	0 1	50 50	24.5
	Redundancy of other constituents	0 1	194 194	38.4
Wrong - Variant Error	Misorder	0 1	50 51	64.9
	Redundancy of other constituents	0 1	194 196	50.4

Table 2: The binary classification model training corpus and the corresponding F1 values for both methods. The ratio of 0 and 1 in both methods is 1:1.

Based on the F1 scores of the two strategies, we decided to train with the second strategy of **Variant Error Sentence**, which involves using a corpus that consists of both specified error types and a variety of other error types.

Appendix C Track 2: Validation of Pseudo-data Construction Method

Strategy	EM	BLEU	$F_{0.5}$	B.S.	Leven	PPL_{BERT}	Avgscore
Bart-base + 1 error	1.0	86.86	21.7	96.89	0.91	2.72	47.99
Bart-base + 2 errors	2.0	86.64	24.35	96.92	1.16	2.67	48.65
Bart-base + 3 errors	3.0	86.82	25.95	96.96	1.26	2.65	49.27
Bart-base + 4 errors	3.0	86.70	27.29	97.03	1.31	2.64	49.55

Table 3: The table presents the results from the pseudo dataset containing numerous errors on the validation set.

The proposed methodology, which involves constructing a pseudo-dataset with sentences containing multiple error types and training an encoder-decoder model on this dataset, proves to be effective in enhancing the performance of the sentence rewriting model. The experimental results (Table 3) demonstrate the superiority of this approach compared to training on sentences with only a single error type.

Appendix D Track 3: Sampling Strategy for Back-translation

We observe that the distribution of multi-label quantities is imbalanced (Table on the left of Figure 5), for which we adopt the oversampling strategy (figure on the right of Figure 5).

Appendix E Track 3: Choice of Parameters for SCE Loss

Table 4 presents the ablation studies evaluating different μ and β parameter values for the SCE loss:

CCL 2024

Label	Nums	Percent.
Excellent (优秀)	12	12%
Moderate (一般)	45	45%
Failing (不及格)	43	43%

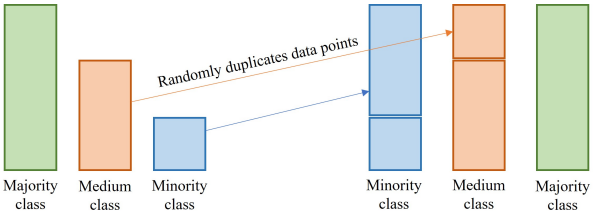


Figure 5: The table on the left displays the distribution of the multi-label quantities. And the figure on the right illustrates the oversampling strategy.

μ	β	Acc	F1	QWK	AvgScore
1	0	50.0	36.67	62.26	45.79
1	1	50.0	37.23	69.51	47.52
0.5	1	60.0	64.44	79.59	66.14
0.1	1	**70.0**	**74.6**	**85.85**	**75.47**
0.05	1	60.0	64.44	79.59	66.14
0.01	1	**70.0**	65.56	71.7	68.12

Table 4: The ablation studies for searching the optimal parameters in validation. Notice that: when β is set to 0, cross-entropy loss is employed during training.
We decided that $\mu = 0.1$ and $\beta = 1$ provided the best balance between standard and reverse cross-entropy for robust training on the potentially noisy pseudo-data labels.

CCL 2024

CCL24-Eval任务7系统报告：中小学作文语法错误检测、病句改写与流畅性评级的自动化方法研究

田巍
北京慧点科技有限公司，北京，中国
tianweia@mail.taiji.com.cn

摘要

本研究旨在提高中小学生作文评改的质量和效率，通过引入先进的自然语言处理模型进行作文病句检测、纠正和流畅性评分，并分别针对三个具体的任务进行了模型构建。在任务一中，提出语法错误替换方法进行数据增强，接着基于UTC模型对语病类型进行识别。在任务二中，融合了预训练的BART模型和SynGEC策略进行文本纠错，充分利用了BART的生成能力和SynGEC的语法纠错特性。任务三中，基于TextRCNN-NEZHA模型进行作文流畅性的评级，构建了一个能够综合语义信息的分类器。经评测，本文提出的方法在任务一和任务二中均位列第一，任务三位列第二，即提出的方法可以有效地识别病句类型和纠正作文中的病句，并给出合理的作文流畅性评级。

关键词： UTC ； BART ； SynGEC ； TextRCNN-NEZHA

System Report for CCL24-Eval Task 7: Research on Automated Methods for Grammar Error Detection, Malapropism Revision, and Fluency Grading in Primary and Secondary School Compositions

Wei Tian
Beijing Smartdot Technology Co., Ltd, Beijing, China
tianweia@mail.taiji.com.cn

Abstract

This study aims to improve the quality and efficiency of essay grading for elementary and middle school students by introducing advanced natural language processing models to detect, correct malformations in sentences, and score fluency in essays. Also, models were built for three specific tasks. In Task 1, a method of grammatical error replacement was proposed for data augmentation, followed by the identification of types of language maladies based on the UTC model. In Task 2, the integration of the pretrained BART model and SynGEC strategy was implemented for text correction, fully utilizing BART's generative capabilities and SynGEC's grammatical correction features. In Task 3, the essay's fluency grading was conducted based on the TextRCNN-NEZHA model, building a classifier that integrates semantic information. Upon evaluation, the methods proposed in this paper ranked first in Tasks 1 and 2 and second in Task 3, demonstrating the effectiveness of the proposed methods in identifying and correcting malformations in sentences, as well as providing a reasonable fluency rating for essays.

Keywords: UTC , BART , SynGEC , TextRCNN-NEZHA

第二十三届中国计算语言学大会论文集，第278页-第284页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

在教育领域不断发展和网络普及的背景下，作文评价面临着劳动成本高和效率低的挑战，从而促使研究人员和机构去探索使用计算机技术进行作文的自动评改。本任务目的在于帮助学生更加明确地识别自己写作中的问题，并为他们的作文修改提供具体的指导，评测任务包括以下三个具体方面：

（1）中小学作文病句类型识别：识别并分类作文中可能出现的各种病句类型，为病句的改正提供基础。

（2）中小学作文病句改写：通过自动化技术，针对识别出的病句进行有效的改写，以提升作文的流畅性和整体质量。

（3）中小学作文流畅性评级：综合作文的词汇使用、结构组织等方面，给出作文流畅性的综合评定，以指导学生如何提高写作能力。

为解决上述问题，文本在任务一中，基于UTC模型(Lou et al., 2023)进行两个阶段的训练；在任务二中，融合了预训练的BART模型(Lewis et al., 2019)和SynGEC策略(Zhang et al., 2022)进行文本纠错；在任务三中，基于TextRCNN-NEZHA模型进行作文流畅性的评级。

2 任务定义

任务一的病句识别，定义为一个多标签分类问题，即对于给定的一句话，预测其包含的病句错误类型。评测任务将病句错误分为两个层次：首先是5类粗粒度错误类型，它们分别为字符级错误，成分残缺型错误，成分赘余型错误，成分搭配不当型错误和不合逻辑错误；其次是14种细粒度错误类型，包括缺字漏字、错别字错误、缺少标点、错用标点、主语不明、谓语残缺、宾语残缺、其他成分残缺、主语多余、虚词多余、其他成分多余、语序不当、动宾搭配不当、其他搭配不当等，病句类型样例如表 1所示。

type	**text**
错别字错误	祥子再度沉论。
缺字漏字	虽然是件小事，可却能影孩子一生。
缺少标点	“聚如一团火，散是满天星”
错用标点	口说无凭，古人云：”以史为鉴“。
主语不明	现在在南京，已多年未见奶奶和姑姑了。
.....	.....
其他搭配不当	你还可以点开微信，无论多远，只要他有网，他都能收到。

Table 1: Task1 病句类型样例

任务二的病句改写，定义为文本生成任务，即给定一句话，在确保原文意图不发生改变的前提下，为中小学生作文中的错误句子提出最小化的修改方案，如表 2所示。自动对错误句子进行修正对于帮助学生理解写作中的问题、提升写作水平具有重要意义。

source	**target**
这些精神，往往会令人感到受益匪浅。	这些精神，往往会令人受益匪浅。
祥子再度沉论。	祥子再度沉沦。
我们秦谁可再会。	我们秦淮河再会。
她是多么地坚强啊！	她是多么的坚强啊！
驰车来到宏村，已是薄暮。	我驰车来到宏村，已是薄暮。

Table 2: Task2 病句改写样例

任务三的流畅性评价，定义为单标签分类问题，即给定一篇作文，根据作文的流畅度，进行打分，分数类别有优秀、一般和不及格，如表 3所示。这样可以方便教师进行最终评分，减

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

轻教师批改作文的负担，为教师提供一个更加高效和直观的方式来评估学生的写作能力，同时也使学生能够更清晰地了解自己在写作上的表现。

text	level
现实足此岸，理想是彼岸，中间隔着湍急的河流，...	优秀
她的个子不高，眼睛也不大，但是炯炯有神，...	中等
阳光是普遍的，如果没有阳光便就不可能，...	不及格

Table 3: Task3 流畅度评价样例

3 模型

3.1 UTC模型

UTC通过统一语义匹配方式USM（Unified Semantic Matching）来将标签和文本的语义匹配能力进行统一建模。这种方法可以更好的理解多类别文本数据，并从中筛选出正确的类别，如图 1所示。

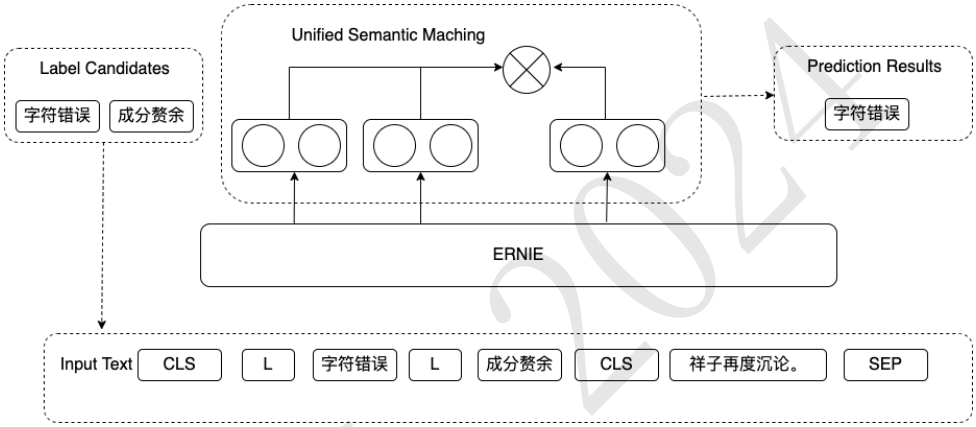


Figure 1: UTC模型架构

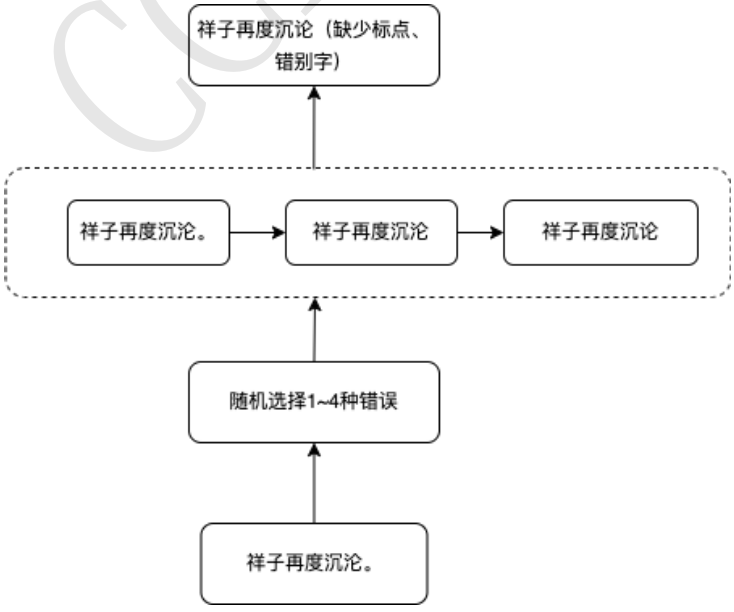


Figure 2: 语法错误替换方法

CCL 2024

在Task1任务中，训练集和验证集的数量分别为1237句和104句，但测试集为9040句，在小样本的训练背景下，为增强模型的鲁棒性，我们提出一种语法错误替换方法，对正确语句进行随机的改错来使得模型具有好的泛化能力，如图 2所示，对“祥子再度沉沦。”这句话，从14种错误类型中，随机选择1-4种错误，比如错别字和缺少标点，然后依次进行错别字和缺少标点的改错，最终的语句为“祥子再度沉论”，通过上述的错误替换方法对正确的语句进行数据增强，使用UTC模型进行微调，可以在低资源情况下使得模型获取不错的纠错效果，将提出的语法错误替换方法开源在https://github.com/TW-NLP/CGED_DAT。

面对Task1的病句类型识别任务，本文采用两个阶段训练策略，第一个阶段是使用语法错误替换方法，对开源的公共数据集进行随机的改错，并基于UTC模型来进行微调。第二个阶段，使用提供的真实数据集来进行第二个阶段的微调，使得模型可以更好的检测出真实的应用场景中出现的语法错误。

3.2 BART-SynGEC模型

BART模型的工作机制概括为两个主要步骤，首先，通过对输入文本执行一系列的预处理操作（如删除部分文本、重排句子等），然后，模型试图使用这种破坏的版本来重建原始文本。因为BART模型的预训练任务与文本纠错任务较为相似，为此采用BART作为基础模型。为了有效的捕捉语法结构，因此融合SynGEC方法，有效的将输入句子的句法结构注入BART模型的编码器部分的方法，使得模型可以更好的生成无语法错误的句子，模型框架如图 3所示。

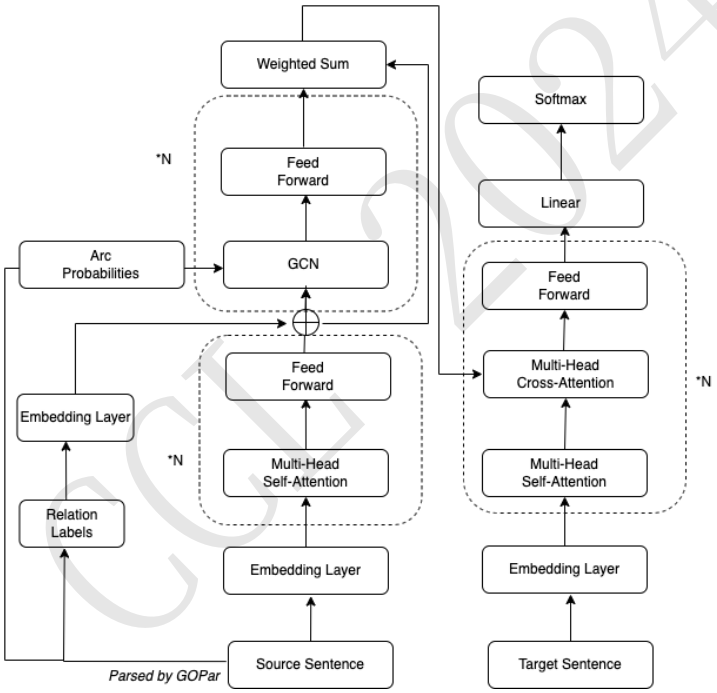


Figure 3: BART-SynGEC模型架构

面对Task2的病句改写任务，本文采用两个阶段的训练策略，第一阶段使用开源的Lang8数据集(Zhao et al., 2018)和HSK数据集(Zhang et al., 2009)进行模型的训练，其中Lang8包含1,220,906条数据，HSK有15,6870条数据，使用通用的数据进行训练，使得模型在通用领域拥有较好的纠错效果。第二阶段，使用Task2提供的训练集进行微调，更好的纠正在中小学生写作的场景下所出现的错误。

3.3 TextRCNN-NEZHA模型

NEZHA模型(Wei et al., 2019)是基于BERT进行了一系列经过验证的改进，其中包括相对位置编码、全字掩码策略、混合精度训练以及训练模型时的LAMB优化器(You et al., 2019)，为了更好的获取整个文章的语义信息，提出TextRCNN-NEZHA模型，如图 4所示。

(c) 2024 中国中文信息学会计算语言学专业委员会

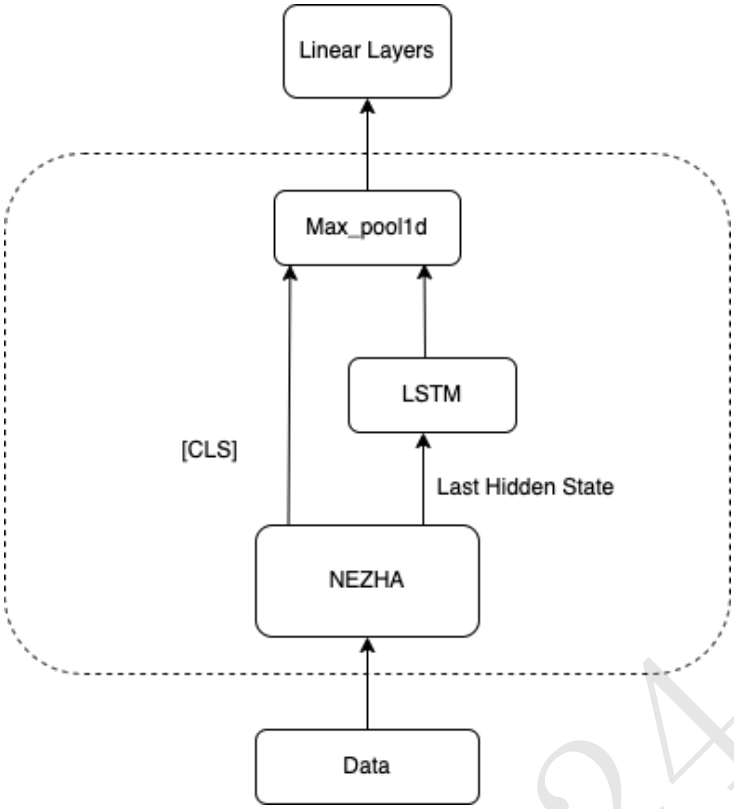


Figure 4: TextRCNN-NEZHA模型架构

面对Task3的流畅度评价任务，TextRCNN-NEZHA模型通过学习[CLS]和词的语义信息，更好的对句子进行语义编码，来捕捉句子流畅度信息。

4 实验

4.1 实验环境

实验所使用的是Ubuntu20.04.2LTS操作系统，用Python作为开发语言，采用Paddle和PyTorch深度学习开发框架。实验采用的CPU为Intel酷睿i7-9700F，GPU为NVIDIA V100，显存32G。

任务一的UTC模型在训练过程中选用AdamW作为优化器，在第一阶段训练的batch-size设置为32，学习率设定为2e-5，maxlen设置为512，epoch为20。第二个阶段训练的batch-size设置为8，学习率设置为1e-5,epoch为20。任务二的BART-SynGEC模型，采用Adam作为优化器，batch-size设置为4，学习率设置为1e-5，maxlen设置为512，epoch为10。任务三的TextRCNN-NEZHA模型，采用AdamW作为优化器，batch-size设置为4，学习率为3e-5，maxlen设置为512，epoch为15。

4.2 评测指标

任务一，粗粒度病句识别分数和细粒度病句识别分数，具体计算方式如下。

$$\text{Score}_{\text{track1}} = 0.5 * F_1^{\text{Course-grained}} + 0.5 * F_1^{\text{Fine-grained}} \quad (1)$$

任务二采用EM(Exact Match)、Bert PPL、与input的编辑距离、BLEU-4、BERTScore以及评估指标(参考MuCGEC)，最终实际排名将综合考虑上述所有指标得到AvgScore，计算方式如下。

$$Score_{track2} = (EM + BLEU + BERTScore)/4 - Levenshtein - PPL_{BERT} \quad (2)$$

任务三采用准确率（Accuracy，Acc）、精确率（Precision，P）、召回率（Recall，R）、F1值（Macro F1）、Quadratic weighted Kappa（QWK）来评估中小学作文流畅性评

(c) 2024 中国中文信息学会计算语言学专业委员会

级的分类效果，计算方式如下。

$$Score_{track3} = 0.5 * F_1 + 0.2 * QWK + 0.3 * ACC \tag{3}$$

4.3 任务一

针对中小学作文病句类型识别任务，首先基于BERT模型(Devlin et al., 2018)来完成文本分类任务，Micro-F1和Macro-F1的得分为46.36、22.88。为了更好的进行病句的识别，本文使用训练数据对UTC模型进行微调，Micro-F1和Macro-F1的得分为51.83、25.46，相比于BERT分别提升5.47和2.58。为了提升模型的鲁棒性，本文基于语法错误替换方法来进行两个阶段的微调，先在伪数据上使用UTC进行微调，然后用提供的训练数据进行第二阶段微调，最后Micro-F1和Macro-F1的得分为56.42、40.54，相比于BERT提升10.6和17.66，相比于UTC提升4.95和15.08，证明了本文基于UTC和语法错误替换方法的有效性，评测结果如表 4所示。

Model	Micro-F1	Macro-F1	Score
baseline（BERT-base-zh）	46.36	22.88	34.62
UTC	51.83	25.46	38.64
UTC+数据增强	56.42	40.54	48.48

Table 4: Task1实验结果

4.4 任务二

针对病句改写任务，首先基于BART模型在训练数据进行训练，最终得分为35.71，为增强模型的鲁棒性，使用伪数据和Nasgec数据集进行训练，并融合SynGEC进行语法纠错，最后在训练数据集进行微调，相比于BART模型，得分提高11.3，证明了所提方法的有效性，任务二评测结果如表 5所示。

Model	EM	PPL	LS	BLEU	Bert	Pre	Recall	F0.5	Score
BART	8.18	15.72	2.85	85.67	96.85	27.21	23.67	26.42	35.71
BART-SynGEC	19.32	15.68	1.63	91.28	97.86	55.15	33.45	48.81	47.01

Table 5: Task2实验结果

4.5 任务三

针对流畅性评价任务，首先基于BERT进行分类，由于训练集的数量为100条，且测试集为1855条，面对小样本分类数据，我们首先选用TextRCNN-NEZHA模型来作为基础模型，接着引入MulDrop策略(Inoue, 2019)，并使用DCE(Li et al., 2019)来进行损失的计算，评测结果如表 6所示。

Model	Acc	Pre	Recall	F1	QWK	AvgScore
Baseline(BERT-base-zh)	45.52	41.42	39.15	38.57	0.1543	44.48
TextRCNN-NEZHA	51.66	67.98	41.29	38.40	0.2008	46.71
TextRCNN-NEZHA+DCE	52.17	44.27	42.63	41.30	0.2330	48.63
TextRCNN-NEZHA+DCE+MulDrop	51.41	46.24	45.64	45.62	0.2330	50.56

Table 6: Task3实验结果

通过对比，本文提出的TextRCNN-NEZHA+DCE+MulDrop模型相比于Baseline，在F1和QWK分别提升7.05、0.0787，证明所提方法的有效性。

5 总结

针对CCL2024的Task7所提出的三个不同任务，本文提出三种不同的解决方案，在任务一中，采用语法错误替换方法进行数据的增强，并基于UTC进行模型的训练。

(c) 2024 中国中文信息学会计算语言学专业委员会

在任务二中，采用BART-SynGEC方法进行语法纠错。在任务三中，采用TextRCNN-NEZHA+DCE+MulDrop来提升模型的鲁棒性，经评测，所提出的方法可以有效地识别病句类型和纠正作文中的病句，并能给出合理的作文流畅性评级。

参考文献

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, number 11, pages 13318–13326.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022. SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser. *arXiv preprint arXiv:2210.12484*.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. NEZHA: Neural contextualized representation for Chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855*.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II 7*, pages 439–445.

Baolin Zhang. 2009. Features and functions of the HSK dynamic composition corpus. *International Chinese Language Education*, 4:71–79.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training BERT in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

CCL 2024

System Report for CCL24-Eval Task 7: Prompting GPT-4 for Chinese Essay Fluency Evaluation

Dan Zhang, Thuong Hoang, Ye Zhu
Deakin University
{dan.zhang thuong.hoang, ye.zhu}@deakin.edu.au

Abstract

This report presents the methodology and results of utilizing GPT-4 for CCL24-Eval Task 7 of Chinese Essay Fluency Evaluation (CEFE). The task is divided into three tracks: Identification of Error Sentence Types, Rewriting Error Sentences, and Essay Fluency Rating. We employed a few-shot prompt engineering to guide GPT-4 in performing this task. Our approach integrated fine-grained error analysis with advanced NLP techniques to provide detailed, actionable feedback for students and teachers. Despite some successes, particularly in generating semantically similar and syntactically relevant corrections, our analysis revealed significant challenges, especially in multiple-label classification and the accurate identification of error types. The report discusses these findings and suggests areas for further improvement.

1 Introduction

The rapid development of education and the widespread use of the Internet have significantly increased the volume of essay evaluations. This growth poses significant challenges in terms of the cost and efficiency of human evaluation. To overcome these challenges, researchers and institutions have increasingly integrated Artificial Intelligence (AI) and Machine Learning (ML) algorithms into educational settings. Various NLP tasks have emerged, including automated item generation (Zou et al., 2022), grammar error detection and correction (Lu et al., 2022), reading and text complexity assessment, writing analysis and feedback (Jia et al., 2022), and automated writing evaluation.

For essay evaluation systems, particularly those targeting primary and middle school students' writing, the goal is to provide objective, accurate, and timely evaluations by analyzing various aspects of essays, including language use, content, and structural coherence. One of the critical aspects of essay evaluation is the fluency of expression, which reflects not only the smoothness and normative use of language but also the overall writing proficiency and ability to convey ideas clearly.

Essay fluency is essential for assessing and improving writing quality. It encompasses several linguistic features, such as sentence length, lexical complexity, and sentence structure (Yang et al., 2019). Current approaches to essay fluency evaluation typically involve scoring based on these linguistic features, treating it as a grammar correction task to identify and correct spelling and grammatical errors, or regarding it as a faulty sentence detection task to determine the presence of errors. However, these methods often consider essay fluency assessment as an isolated natural language processing (NLP) task, lacking systematic integration across multiple levels and perspectives. Moreover, existing studies tend to define grammatical error types in broad categories like redundancy, omission, misuse, and disorder, which are insufficiently detailed for providing precise feedback.

To advance the field of grammatical error recognition and correction in examination essays written by native Chinese-speaking primary and secondary school students, the China National Conference on Computational Linguistics (CCL-2024) has included Chinese Essay Fluency Evaluation (CEFE) as one of its

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 285-293, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

shared tasks. This task systematically classifies text errors at various granularities, provides human-annotated data, and proposes three tracks that cover error detection and correction, sentence rewriting, and essay fluency rating.

To address this task, we employed few-shot prompting techniques and leveraged the advanced generation capabilities of GPT-4 (OpenAI, 2023) to make predictions across the three tracks. Our approach integrates fine-grained error analysis with advanced NLP techniques to deliver detailed, actionable feedback for students and teachers. Specifically, we used prompt engineering (Dai et al., 2023) to instruct GPT-4 to rewrite erroneous sentences with minimal modifications. Additionally, we guided GPT-4 to rate the overall fluency of essays. Ultimately, our goal is to improve the quality and fluency of student essays, providing them with insights necessary for continuous improvement in their writing skills.

2 Task Description

The CEFE shared task systematically defines fine-grained error types that affect essay fluency and offers corrective suggestions. This fine-grained approach helps students understand their writing issues more clearly and assists teachers in quickly gauging students' writing levels, facilitating more effective writing instruction. The evaluation task involves a comprehensive analysis of essay fluency from multiple linguistic angles: lexical, syntactic, and semantic, and provides modification suggestions. The task is divided into three tracks:

Track 1: Identification of Error Sentence Types This track focuses on recognizing both coarse-grained and fine-grained errors commonly found in student essays. It approaches the issue from two perspectives: character-level and component-level errors. There are four types of coarse-grained grammatical errors (Shen et al., 2023): Character-Level Error (CL), Incomplete Component Error (IC), Redundant Component Error (RC), and Incorrect Constituent Combination Error (ICC). Additionally, there are fourteen fine-grained error types, which provides a more detailed understanding of specific errors. Additionally, there are fourteen fine-grained error types, providing a more detailed understanding of specific errors. The detailed error types are listed in the table 1 below.

Coarse-grained Error	**Fine-grained Error**
Character-level	Missing Word Typo Missing Punctuation Punctuation Misuse
Incomplete Component	Subject unknown Predicate Incompleteness Object Incompleteness Other Incompleteness
Redundant Component	Subject Redundancy Function Word Redundancy Other Redundancy
Incorrect Constituent Combination	Improper Word Order Verb-object Mismatch Other Mismatch

Table 1: Coarse-grained and Fine-grained Error Types

Track 2: Rewriting Error Sentences This track involves rewriting incorrect sentences in primary and secondary school compositions. The challenge is to provide minimal modifications to the erroneous sentences while preserving the original semantics. The revisions should make as few changes as possible, focusing on correcting errors without altering the intended meaning.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 285-293, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Track 3: Essay Fluency Rating This track involves assessing the overall fluency of essays. For this evaluation, essays are rated based on three levels of fluency: Excellent, Average, and Poor. These tracks collectively aim to enhance the understanding and evaluation of essay fluency, providing actionable insights for both students and educators.

3 Methodology

Track 1: Identification of Error Sentence Types To effectively identify error types, we designed a series of prompts tailored to leverage the capabilities of GPT-4. We also combined few-shot learning (Brown et al., 2020) in the prompt engineering (Chung et al., 2024). Our prompt design process involved the following steps:

- Initial Exploration: We began by generating a diverse set of example prompts to assess GPT-4's initial performance in identifying error types. These examples included sentences with known errors annotated with both coarse-grained and fine-grained error types.
- Prompt Refinement: Based on initial results, we refined our prompts to enhance clarity and specificity. Each prompt was designed to clearly define the task, provide context, and include labeled examples.
- Few-Shot Learning: We utilized few-shot learning by incorporating several annotated examples within the prompts. This approach helped GPT-4 understand the task requirements and improved its ability to generalize from the examples provided.
- Iterative Testing and Optimization: We iteratively tested and optimized our prompts using feedback from each iteration to fine-tune the examples and instructions.

This process ensured that the prompts effectively guided GPT-4 in accurately identifying error types.

The final few-shot learning-based prompt we used for track 1 is in Appendix Figure1:

Track 2: Rewriting Error Sentences To achieve effective sentence rewriting, we designed a series of prompts tailored to guide GPT-4 in producing accurate and semantically consistent corrections. The design process (Chung et al., 2024) included the following steps:

- Contextual Examples: We incorporated contextually relevant examples within the prompts to demonstrate how erroneous sentences should be corrected. These examples showcased minimal yet precise modifications, emphasizing the importance of maintaining original semantics. By illustrating corrections in context, we helped GPT-4 understand the nuances of effective rewriting.
- Clear Instructions: Each prompt provided clear and concise instructions specifying the need to make minimal changes while retaining the original meaning of the sentences. This guidance helped GPT-4 focus on correcting errors without altering the intended message, ensuring that the corrections were both accurate and meaningful.
- Few-Shot Learning: We utilized few-shot learning by including multiple annotated examples within each prompt. This approach allowed GPT-4 to learn the task requirements and apply them to new unseen sentences effectively.
- Iterative Refinement: Through iterative testing and refinement, we continuously improved our prompts to enhance GPT-4's performance. Feedback from each iteration was used to adjust the examples and instructions, ensuring the model generated accurate and semantically consistent corrections. This iterative approach allowed us to fine-tune the prompts for optimal results.

By following these steps, we ensured that GPT-4 was effectively guided in rewriting erroneous sentences, producing corrections that were both accurate and semantically consistent.

The final few shot learning prompt we used for track 2 is in Appendix Figure 2:

CCL 2024

Track 3: Essay Fluency Rating To effectively rate essay fluency, we designed a series of prompts that provided clear instructions and examples. Our prompt design process included the following steps:

- Definition of Fluency Levels: We provided detailed descriptions of the three fluency levels (Excellent, Average, and Poor) within the prompts. This helped GPT-4 understand the criteria for each category.
- Annotated Examples: We included multiple annotated examples of essays rated at different fluency levels. These examples served as references, illustrating the characteristics of each fluency level.
- Few-Shot Learning: By incorporating several examples within each prompt, we utilized few-shot learning to help GPT-4 generalize the rating criteria across different essays.
- Iterative Refinement: We iteratively tested and refined our prompts to enhance clarity and effectiveness. Feedback from each iteration was used to adjust the examples and instructions, ensuring optimal performance.

The final few shot learning prompt we used for track 3 is in Appendix Figure 3:

4 Experiment Results and Analysis

Track 1 Result In Track 1, GPT-4 (OpenAI, 2023) encountered significant challenges in error identification, leading to incomplete and empty predictions:

The model failed to recognize certain error types, resulting in missing annotations. This indicates that GPT-4 struggled with multiple label classification tasks, particularly in accurately identifying all relevant error types within a sentence.

As for the empty predictions, there were instances where the model outputted empty results for error types, even when multiple errors were present. This further highlights GPT-4's difficulty in handling tasks requiring multiple labels.

The performance issues observed indicate that GPT-4 is not well-suited for multiple label classification tasks, especially in the context of identifying fine-grained grammatical errors.

We summarize the key challenges and insights for track 1: Complexity of Multiple Labels:

- Multiple label classification requires the model to identify several error types within a single sentence, which appears to be a challenging task for GPT-4.
- Prompt Sensitivity: The model's performance was highly sensitive to the phrasing and structure of the prompts. Despite iterative refinement, the prompts may not have been effective enough to guide the model accurately.
- Error Types Granularity: The fine-grained nature of error types required a deep understanding of grammatical rules and context, which GPT-4 struggled to achieve consistently.

These findings suggest that further refinement and alternative approaches are needed to improve GPT-4's performance in multiple label classification tasks.

Track 2 Result Our evaluation process for sentence rewriting involved comparing the corrected sentences generated by GPT-4 against reference annotations.

To evaluate the performance of our model, we utilized several metrics: Exact Match (EM), BERT Perplexity (Bert PPL), Levenshtein distance, BLEU-4, BERTScore (Zhang et al., 2019), Precision, Recall, and F0.5. An aggregated score, AvgScore, was calculated using a specific formula.

Exact Match (EM) measures the percentage of sentences that match the reference sentence exactly. BERT Perplexity (Bert PPL) (Devlin et al., 2018) indicates the fluency of the generated text based on the perplexity score from a BERT model. Levenshtein Distance measures the number of single-character

CCL·2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 285-293, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

edits required to change the generated sentence into the reference sentence. BLEU-4 (Papineni et al., 2002) evaluates the n-gram precision up to 4-grams, indicating the similarity between the generated and reference sentences. BERTScore (Zhang et al., 2019) uses BERT embeddings to measure the semantic similarity between the generated and reference sentences. Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall is the ratio of correctly predicted positive observations to all observations in the actual class. F0.5 is a weighted harmonic mean of precision and recall that gives more weight to precision.

The AvgScore was calculated using the following formula:

$$\text{AvgScore} = \frac{\text{EM} + \text{BLEU-4} + \text{F0.5} + \text{BERTScore}}{4} - \text{Levenshtein} - \text{Bert PPL}$$

The results are shown in table 2.

Model	EM	Bert PPL	Levenshtein	BLEU-4	BERT Score	Precision	Recall	F0.5	Score
Baseline model	9.67	3.26	1.46	88.72	97.28	37.33	20.53	32.08	52.22
Best Rank model	11.5	2.91	2.7	88.00	97.37	38.75	25.72	35.19	52.41
GPT-4	3.78	12.29	12.66	70.75	91.89	0.41	1.06	0.46	16.77

Table 2: Comparison of performance on Track 2.

The low precision score indicates that a significant portion of the model's corrections are not correct. A recall score greater than 1 suggests potential issues with the calculation or interpretation, as recall should typically be between 0 and 1. The F-0.5 score, which prioritizes precision, is also low, reinforcing the conclusion that the model's corrections are often inaccurate.

Performance analysis

- The performance analysis in Track 2 reveals strengths in generating semantically similar and syntactically relevant corrections, as indicated by high BLEU-4 and BERTScore values.
- However, the model struggles with exact matches and precision, suggesting that further refinement is needed to improve the accuracy of corrections.
- The overall AvgScore of 16.77 highlights a moderate performance, indicating room for improvement in future iterations.

Track 3 Result Our evaluation process for the essay fluency rating task involved comparing the fluency ratings generated by GPT-4 against reference annotations. The evaluation employs several metrics: Accuracy (ACC), Precision, Recall, Macro F1, and Quadratic Weighted Kappa (QWK). An aggregated score, AvgScore, is calculated using a weighted combination of these metrics. The AvgScore is calculated using the following formula:

$$\text{AvgScore} = 0.5 * \text{F1} + 0.2 * \text{QWK} + 0.3 * \text{ACC}$$

QWK is first normalized to the [0,1] range before calculating the weighted score.

The results are shown in the below table 3.

We achieved an accuracy of 48.59%, indicating that nearly half of the essay fluency ratings were correctly classified. This reflects a moderate level of correctness in the model's predictions.

The precision score of 55.79% suggests that over half of the predicted fluency ratings were correct. This high precision indicates that the model is good at identifying correct positive instances, meaning the model's predictions are reliable when it does identify a positive case.

CCL 2024

Model	ACC	Precision	Recall	F1	QWK	AvgScore
Baseline model	45.52	41.42	39.15	38.57	0.1543	44.48
Best Rank model	51.66	49.89	46.54	47.42	0.1818	51.03
GPT-4	48.59	55.79	40.09	39.25	0.1202	45.40

Table 3: Comparison of Performance on Track 3.

The recall score of 40.09% shows that the model identified about 40% of the actual positive instances. While this is lower than the precision score, it indicates that there is room for improvement in capturing all relevant positive cases.

The F1 score of 39.25 reflects a balanced consideration of precision and recall. This score indicates that the model maintains a reasonable trade-off between precision and recall, providing a holistic measure of its classification performance.

Quadratic Weighted Kappa (QWK):

- The QWK score of 0.1202, normalized to 0.5601, indicates a moderate agreement with human ratings.
- This suggests that the model's predictions align reasonably well with human judgments, though there is still room for improvement.

AvgScore:

- The AvgScore of 45.40 is a composite measure reflecting the overall performance across all metrics.
- This score shows that GPT-4 model performs fairly well but also highlights areas where further refinement could lead to better results.

Performance analysis

- The performance analysis in Track 3 demonstrates a relatively strong model for assessing essay fluency, with high precision and a moderate level of accuracy and agreement with human ratings.
- However, the lower recall score indicates that the model could benefit from strategies aimed at capturing more positive instances.

5 Conclusion

The results of our evaluation indicate that while GPT-4 shows potential in tasks such as rewriting sentences and rating essay fluency, it faces significant challenges in accurately identifying error types, particularly in multiple-label classification tasks. Our approach demonstrated strengths in generating semantically similar corrections and achieving moderate agreement with human ratings. However, these findings highlight the need for further refinement and alternative strategies to improve the model's accuracy and overall performance in these tasks.

References

Xinshu Shen, Hongyi Wu, Xiaopeng Bai, Yuanbin Wu, Aimin Zhou, Shaoguang Mao, Tao Ge, and Yan Xia. 2023. *Overview of CCL23-Eval Task 8: Chinese Essay Fluency Evaluation (CEFE) Task*. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 282–292.

Bowei Zou, Pengfei Li, Liangming Pan, and Aiti Aw. 2022. *Automatic true/false question generation for educational purpose*. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 61–70.

CCL 2024

Yiting Lu, Stefano Bannò, and Mark Gales. 2022. *On assessing and developing spoken 'grammatical error correction' systems*. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 51–60.

Qinjin Jia, Yupeng Cao, and Edward Gehringer. 2022. *Starting from "zero": An incremental zero-shot learning approach for assessing peer feedback comments*. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 46–50.

Yiqin Yang, Li Xia, and Qianchuan Zhao. 2019. *An automated grader for Chinese essay combining shallow and deep semantic attributes*. *IEEE Access*, volume 7, pages 176306–176316. IEEE.

OpenAI, R. 2023. *GPT-4 Technical Report. arXiv 2303.08774*. *View in Article*, volume 2, number 5.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. *Can large language models provide feedback to students? A case study on ChatGPT*. In *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint arXiv:1810.04805*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. *BERTScore: Evaluating text generation with BERT*. *arXiv preprint arXiv:1904.09675*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. *Language models are few-shot learners*. *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and others. 2024. *Scaling instruction-finetuned language models*. *Journal of Machine Learning Research*, volume 25, number 70, pages 1–53.

A Appendix

CCL 2024

中小学作文病句类型识别是一个多标签分类问题。每条句子可能包含一个或多个错误类型。
本任务旨在分析每条句子，并精确地识别出每个病句对应的粗粒度和细粒度错误类型。

病句类型定义：
病句错误类型包括词法、句法、语义错误，具体分为四个粗粒度错误类型和十四个细粒度错误类型如下：

字符级错误：包含 缺字漏字，错别字错误，缺少标点，错用标点。
成分残缺型错误：包含 主语不明，谓语残缺，宾语残缺，其他成分残缺。
成分赘余型错误：主语多余，虚词多余，其他成分多余。
成分搭配不当型错误：语序不当，动宾搭配不当，其他搭配不当。

请参考以下示例对输入的中小学生作文句子进行分析，识别并标注出现的每一种病句类型。

需要同时注意句子可能存在的多种错误，并给出相应的粗粒度和细粒度错误类型。只输出预测结果
输出格式应包括两个列表："CourseGrainedErrorType" 和 "FineGrainedErrorType"。
示例：
Example 1
Example 2

作文句子="{sentences}"

Figure 1: The final few-shot learning-based prompt used for track 1.

中小学作文病句改写任务的目标是对中小学生作文中出现的错误句子进行最小化修改，
同时确保语义保持不变。这是一个文本生成任务，输入为错误的句子，输出为修改后的句子。

操作指南：
请仔细分析输入的错误句子，并进行适当的修改以纠正错误，确保修改后的句子语义保持一致。
输出应仅包含修改后的句子（revisedSent），不包含其他信息。

示例：
Example 1
Example 2

作文句子="{sentences}"

请确保所有修改尽可能地简洁，并严格遵循语法规则，以提供清晰、准确的改写。

Figure 2: The final few-shot learning prompt used for track 2.

CCL 2024

任务描述：
你是一名特级中小学语文老师。中小学作文流畅性评级任务是一个多分类任务，目的是对一篇作文的流畅性进行评级。
流畅性是指文本的语言表达是否通顺、自然，以及文章结构是否合理。

流畅性等级：

优秀：作文语言表达极为流畅，逻辑清晰，段落过渡自然，无明显语法错误或不自然表达。

一般：作文整体表达尚可，存在一些语法错误或表达不自然，逻辑和结构表现平均，影响阅读体验。

不及格：作文在语言表达上非常不流畅，逻辑不清晰或结构混乱，存在多处语法错误或不自然的表达，明显影响理解。

操作指南：
输入：一篇中小学生作文。
输出：根据作文的流畅性，给出相应的评级（优秀、一般、不及格）。只输出`essay_score_level`

示例输入和输出：
`Example 1`
`Example 2`

参照以上示例，对以下作文进行分类，只输出对`essay_socre_level`的预测{优秀，一般，不及格}：

作文句子="`{essay}`"

Figure 3: The final few-shot learning-based prompt used for track 3.

CCL 2024

CCL24-Eval任务7系统报告：基于大模型数据增强的作文流畅性评价方法

彭倩雯[1]，高延子鹏[1]，李晓青[1]，闵凡珂[1]，李明锐[1]，王志春[1,2]，刘天昀[3]
[1] 北京师范大学人工智能学院
[2] 智能技术与教育应用教育部工程研究中心
[3] 中国科学院信息工程研究所
{qwpeng,yanzipenggao,xiaoqingli,minfanke,mingruili}@mail.bnu.edu.cn
zcwang@bnu.edu.cn , liutianyun@iie.ac.cn

摘要

CCL2024-Eval任务7为中小学生作文流畅性评价（Chinese Essay Fluency Evaluation，CEFE），该任务定义了三项重要且富有挑战性的问题，包括中小学作文病句类型识别、中小学作文病句改写、以及中小学作文流畅性评级。本队伍参加了评测任务7的三项子任务，分别获得了45.19、43.90和45.84的得分。本报告详细介绍本队伍在三个子任务上采用的技术方法，并对评测结果进行分析。

关键词： 作文流畅性评价 ； 数据增强 ； 语言模型

System Report for CCL24-Eval Task 7: Essay Fluency Evaluation Method Based on Large Model Data Augmentation

Qianwen Peng[1],Yanzipeng Gao[1],Xiaoqing Li[1],Fanke Min[1],Mingrui Li[1]
Zhichun Wang[1,2],Tianyun Liu[3]
[1] School of Artificial Intelligence,Beijing Normal University
[2] National Engineering Laboratory for Cyberlearning and Intelligent Technology
[3] Institute of Information Engineering，Chinese Academy of Sciences
{qwpeng,yanzipenggao,xiaoqingli,minfanke,mingruili}@mail.bnu.edu.cn
zcwang@bnu.edu.cn , liutianyun@iie.ac.cn

Abstract

The CCL2024-Eval Task 7 focuses on Chinese Essay Fluency Evaluation (CEFE) for primary and secondary school students. This task encompasses three significant and challenging problems: identifying sentence errors in essays, rewriting erroneous sentences, and rating the fluency of the essays. Our team participated in all three sub-tasks (tracks) of Task 7, achieving scores of 45.19, 43.90, and 45.84. This report provides a detailed account of the technical methods employed by our team for each sub-task and analyzes the evaluation results.

Keywords: Essay Fluency Evaluation , Data Augmentation , Language Model

1 任务概述

作文流畅性指的是作文语句的通顺程度和语言使用的规范程度，是体现作文质量的重要方面。CCL24-Eval任务7为中小学生作文流畅性评价，设立三个子任务，包括：

（1）中小学作文病句类型识别：识别作文中不同的病句类型。

（2）中小学作文病句改写：改写作文中的病句使其成为正确句子。

（3）中小学作文流畅性评级：对作文流畅性作三等级评价。

评测任务提供了基于以汉语为母语的中小学生考试作文构建的测试数据，要求参赛者从词法、句法、语义等多角度对作文流畅性进行分析。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

2 子任务一：中小学作文病句类型识别

2.1 任务描述

中小学作文病句类型识别任务本质上是一个多标签分类任务，其目标是针对给定的作文病句，预测该病句所属的一个或多个类型。病句类型包括4大类14小类，如表1所示。

Table 1: 中小学作文病句类型分类

粗粒度错误类型	细粒度错误类型
字符级错误	缺字漏字、错别字错误、缺少标点、错用标点
成分残缺型错误	主语不明、谓语残缺、宾语残缺、其他成分残缺
成分赘余型错误	主语多余、虚词多余、其他成分多余
成分搭配不当型错误	语序不当、动宾搭配不当、其他搭配不当

2.2 数据准备

为了训练模型实现对病句的准确分类，本队伍通过以下步骤构建训练数据集。

（1）开放作文数据获取：本队伍从中小学作文网[0]等网站，获取中小学生优秀作文234篇，包含句子共计10000句。

（2）基于GPT-3.5-Turbo的作文生成：构建如表2所示的提示，调用GPT-3.5-Turbo(OpenAI, 2023)模仿中小学生身份生成作文240篇。

Table 2: 中小学作文生成提示

类别	提示
叙事	你是一个初二年级的学生，现在请以“人生”为主题，写20篇叙事作文，要求内容充实，语句通顺。
记人	你是一个三年级的学生，现在请以“身边的英雄”为主题，写20篇记人作文，要求内容充实，语句通顺。

（3）基于大语言模型的病句生成：构建如表3的提示，调用GPT-3.5-Turbo 、文心一言3.5[1]等语言模型从给定的正确句子出发生成指定类型的病句；该步骤共生成病句19万条，其中由GPT-3.5-Turbo生成的病句占比60%，由文心一言生成的病句占比40%，覆盖表1中所有错误类型，并在各个细粒度错误类型上均匀分布。

Table 3: 提示示例

细粒度错误类型	提示
错别字错误	你是一个语文老师，你的任务是将句子中的一个字改成错别字，你可以从两个角度入手，一是把这个字改成它的形似字，二是把这个字改成它的同音字。注意不要原句返回也不要改变句子的长度，现在你试试句子：
错用标点	你是一个语文老师，你的任务是将句子中标点符号换成错误的，试试句子：
动宾搭配不当	现在你是一名语文不好的小学生，写出的句子经常出现中文语法错误。特别的，你只会写出“动宾搭配不当”的句子。动宾搭配不当指的是在句子中，动词（动作）和宾语（动作的对象）之间的搭配不合语法规则或语义习惯，导致句子意思不清或逻辑混乱。现在将给你一个句子，请修改为动宾搭配不当的句子。正确句子：

[0] http://www.zuowen.com/
[1] https://yiyan.baidu.com/

(c) 2024 中国中文信息学会计算语言学专业委员会

Table 4: 语言模型生成病句示例

细粒度错误类型	语言模型生成病句
错别字错误	她的手小小的，她的膊搏胖胖的，她的腿短短的，好像我家的布娃娃。
错用标点	今天是爸爸的生日；我决定亲手种一颗向日葵送给他，作为礼物。
动宾搭配不当	在家里和家人一起种植花的时候，我感觉特别开心和幸福。

（4）格式转换：将“原句-病句”的数据格式转换为“病句-错误类型标签”的数据格式，标签根据生成病句时使用的提示确定。

（5）数据增强：本队伍还使用了一种误差不变的数据增强方法，通过替换命名实体来增强数据的多样性。在本次测评任务中，这种数据增强方法可以提高模型的识别语病错误能力。

（6）数据清洗：由于数据集中存在噪声，需要对其进行清洗。其具体步骤如表5所示。

Table 5: 数据清洗步骤

步骤	实现
去除特殊字符和重复标点	正则表达式
统一使用中文标点	编写替换函数
去除输入长度与输出长度差异过大的文本	设置差异阈值5
去除长度过短的文本	设置最小长度阈值10

2.3 模型构建及训练

针对作文病句类型识别任务，本队伍构建了基于BERT的病句分类模型，使用训练数据和评测任务数据集对模型进行了微调训练。给定句子s，设$\text{BERT}_{\text{CLS}}(s)$为模型输出层[CLS]的隐式向量，病句分类模型可表示为：

$$f(s) = \text{MLP}(\text{BERT}_{\text{CLS}}(s)) \tag{1}$$

模型在微调训练过程中，采用了Adam优化器对BCEWithLogitsLoss损失函数进行优化。为提高模型分类性能，本队伍基于评测任务提供的验证集，搜索确定了每个标签分类概率的最优阈值；以0.001为步长，从0.5到0.6搜索并计算模型得分，选择得分最高的阈值作为模型各标签最终的分类阈值。基于上述模型及分类方法，本队伍在测试集上取得了45.19的得分。

2.4 对抗训练

本队伍尝试通过FGM对抗训练技巧 (Miyato et al., 2016)以提升模型的鲁棒性。其具体过程如算法1所示

Algorithm 1 FGM对抗训练流程

1: **for** 每个输入向量x **do**
2: 计算x 的前向损失、反向传播得到梯度$g1$
3: 根据嵌入矩阵的梯度计算出噪声扰动r，并加到当前嵌入上，相当于$x + r$
4: 计算$x + r$ 的前向损失，反向传播得到对抗的梯度，累加到$g1$上得到梯度$g2$
5: 将嵌入恢复为第一步时的值
6: 根据梯度$g2$对参数进行更新
7: **end for**

实验结果表明，FGM对抗训练在作文病句类型识别任务中对模型表现并没有明显增益效果。本队伍对此分析：对于作文病句类型识别任务，文本中存在大量噪声和对于分类并无帮助的内容。因此，通过对抗训练引入增强扰动，理论上应该能够增强模型的抗干扰性。但由于本队伍数据增强阶段已加入了有效的扰动，故使用FGM对抗训练技巧无法进一步提高模型的性能。

(c) 2024 中国中文信息学会计算语言学专业委员会

3 子任务二：中小学作文病句改写

3.1 任务描述

中小学作文病句改写任务本质上是一个文本生成任务，目标是针对输入的病句，输出改正后的句子。任务要求在保持语义不变的前提下，为中小学生作文中的错误句子提供最小化修改方案。

3.2 数据准备

为了训练模型实现对病句的准确修改，本队伍通过组合以下数据集构建混合训练数据。

（1）评测任务训练集：评测任务训练集来源于中小学作文数据，规模约1300句。

（2）大模型生成的病句数据集：在中小学作文病句类型识别任务中使用语言模型生成的数据的基础上，通过数据处理和数据清洗等步骤，得到适用于中小学作文病句改写任务的19万条句子。

（3）正确句子数据集：评测任务测试集中包含无语病的正确句子，这要求限制模型对正确句子的修改。为改善模型的过纠问题，本队伍构建了一个正确句子数据集，规模约10000句，数据来源于中小学作文病句类型识别任务中获取的开放数据集。

（4）开放语法纠错数据集：开放语法纠错数据集包括苏州大学和阿里巴巴达摩院联合发布的MuCGEC (Zhang et al., 2022b)中文语法纠错评测数据集、YACLC汉语学习者文本多维标注数据集 (Yingying Wang, 2021)等，规模约360万句。开放数据集和评测任务数据集相比，前者每一错误句子中包含的错误数目远少于后者。

综上，本队伍数据集整体构成如表6所示。

Table 6: 数据集信息

数据集	数量	平均长度
评测任务训练数据	1,336	46.33
生成病句数据	140,000	35.07
正确句子数据	10,000	36.53
开放纠错数据	3,648,481	30.84

3.3 模型构建及训练

针对中小学作文病句改写任务，本队伍采用大模型文本纠错方法，使用训练数据对基础模型进行了指令微调，构建了基于Qwen-14B(Bai et al., 2023)的病句改写模型。针对大模型普遍存在的过纠问题，本队伍通过设计具体的提示模版来强化模型的泛化能力，如表7所示。在训练过程中，本队伍探索了不同的参数设定，其在任务中的表现结果如表8所示。最终，本队伍在测试集上取得了43.90的成绩。

Table 7: 病句修改提示模板

类型	示例
身份假设	你是一个中小学语文老师
任务描述	你需要将下面的句子改为正确的、你需要修改下面的句子、你需要改正病句
具体要求	保证输出句子和原句之间较小的距离、请尽可能做较小的改动、保证修改后句子的流畅度、请尽可能让修改后的句子流畅

(c) 2024 中国中文信息学会计算语言学专业委员会

Table 8: 不同参数的实验结果

序列长度	学习率	迭代次数	BERT PPL	Levenshtein	BlEU-4	BERTscore
1024	5e-5	3.00	2.70	1.66	0.90	0.98
1024	5e-5	10.00	2.60	3.82	0.99	0.99
90	5e-4	10.00	2.59	3.61	0.98	0.99
1024	5e-4	10.00	2.59	3.76	0.99	0.99

3.4 大模型病句改写性能分析

本队伍对比了不同模型在病句改写方面的效果，包括Qwen-7B、Qwen-14B-Chat、Qwen1.5-14B-Chat 和基于BART (Lewis et al., 2019) 的序列标注模型。在纠错效果方面，本队伍观察到基于BART 的序列标注模型在处理字符级错误方面表现突出，而Qwen 模型则在涉及语义错误、如语序不当等复杂情况下表现更为出色，能够识别并修改BART 模型无法处理的错误。如表9所示，在处理病句‘这根本是一座不可能翻去的山。”’时，Qwen 系列模型能够根据句意将“翻去”修正为“翻越”，并且还能识别出句末的符号错误，而BART 则只能将“翻去”修改为“翻过去”。

Table 9: 大模型病句改写性能

原句	这根本是一座不可能翻去的山。”
Qwen-7B	这根本是一座不可能翻去的山。
Qwen-14B-Chat	这根本是一座不可能翻越的山。
Qwen1.5-14B-Chat	这根本是一座不可能翻越的山。
BART	这根本是一座不可能翻过去的山。”

在结果评分层面，由于大模型的幻觉问题，在处理细粒度纠错任务时会出现过纠现象，不必要的过度改写导致得分始终无法提升。但当考虑到大型模型的规模和参数量达到一定程度时，模型涌现出更多能力，具有更多的潜能来捕捉文本的语义信息和上下文关系。这种额外的能力使它们能够更准确地识别和纠正复杂错误，包括语义错误、逻辑错误等，而这些错误可能会超出传统文本纠错模型如BART的处理范围。因此，大型模型在病句改写任务中展现出了更为广阔的潜力，有望推动文本生成和修正领域的发展。

4 子任务三：中小学作文流畅性评级

4.1 任务描述

中小学作文流畅性评级任务本质上是一个多分类任务，目标是针对给定的作文，预测作文在流畅性方面所属的等级。本次评测任务共定义了三个流畅性等级：优秀、一般、不及格。

4.2 数据准备

考虑到存在语病的作文流畅性普遍较差，本队伍在中小学作文流畅性评级任务中额外训练了中小学作文病句识别模型。针对中小学作文病句识别模型，本队伍预先对数据进行处理，将评测任务训练集中无语病的句子作为正例，存在语病的句子作为负例。数据集规模为2600句，数据集格式：输入为作文句子，输出为有语病或无语病两种标签。

4.3 模型构建及训练

针对作文流畅性评级任务，本队伍构建了基于BERT的分类模型，使用训练数据对模型进行了微调训练。给定作文文段s，设$\mathrm{BERT}_{\mathrm{CLS}}(s)$为模型输出层[CLS]的隐式向量，作文流畅性评价模型可表示为：

$$f(s) = \mathrm{MLP}(\mathrm{BERT}_{\mathrm{CLS}}(s)) \tag{2}$$

(c) 2024 中国中文信息学会计算语言学专业委员会

本队伍在构建作文流畅性评级模型时，选择了中文BERT-WWM(Lewis et al., 2019)预训练模型来初始化模型的编码器部分，并利用官方数据集对模型进行了精细的训练。本队伍测试了不同学习率设置，发现过高或过低的学习率均会对模型的性能产生负面影响，结果如表10。

Table 10: 不同学习率的实验结果

学习率	Precision	Recall	F1 Score	Overall Score
5e-5	71.43	68.18	69.77	69.00
5e-4	50.00	45.45	47.62	47.44
5e-6	40.91	40.91	40.91	39.49

在解码过程中，模型有时会忽视文段中的句子语病。为了改善这一状况，本队伍针对性地设计了一套如算法2所示的解码后处理策略。通过本策略，模型能够更有效地纠正模型在解码过程中忽视语病的问题。

Algorithm 2 解码后处理

初始标签预测$\hat{y} = \mathcal{M}(s)$，检测语病数量n_e
if $\hat{y} =$ 优秀 $\wedge\, n_e > \tau$ **then**
    if $n_e \leq \tau_{\text{一般}}$ **then**
        最终标签$\hat{y} \leftarrow$ 一般
    else
        最终标签$\hat{y} \leftarrow$ 不及格
    end if
end if

具体而言，本队伍使用病句识别模型，判断文段中是否存在语病；使用基于BERT的分类模型，判断作文流畅性评级。若输入作文被初始评定为优秀，但其中存在语病，则根据存在语病的句子数量，将其评级降低为一般或不及格。为保证后处理策略的有效性，本队伍在验证集上对解码后处理模型的阈值进行搜索，确定其阈值为3。这意味着，如果一篇初始被评为优秀的作文中被识别的语法错误数量超过3，则会被调整评级为不及格；如果存在语法错误但数量未超过3，则会被调整评级为一般。基于上述模型及分类方法，本队伍在测试集上取得了45.84的得分。

5 结语

在CCL2024-Eval任务7-中小学作文流畅性评价评测任务中，针对三个子任务本队伍分别提出了中小学作文病句类型识别模型，中小学作文病句改写模型和中小学作文流畅性评级模型，并且提出了针对中小学作文数据集稀缺问题的大模型辅助数据生成方法，还尝试使用了一些额外的性能提升技术，例如对抗训练、多轮微调等。实验结果表明，本队伍提出的多种策略均可以使模型性能得到有效的提升，最终三个子任务得分分别为45.19、43.90、45.84。通过这三个子任务，本队伍实现了多维细粒度的自动作文评价，有效提高了作文评分的可解释性和反馈的丰富程度。如今，作文的自动流畅性评价具有明确的研究意义和应用场景。本队伍的方法推进了自动批改在课堂教学中的应用，可以有效辅助教师进行课堂教学。但是，要想真正实现让机器像人一样去欣赏和批判写作，包括对文章的立意思辨、篇章结构等方面进行评价依然是非常困难的。如何持续提高机器的对作文的审美能力和鉴别水平依然是开放问题。

6 致谢

资助本工作项目：科技创新2030 -“新一代人工智能”重大项目（2021ZD0113000），国家自然科学基金项目（62276026）。

(c) 2024 中国中文信息学会计算语言学专业委员会

参考文献

J. Achiam, S. Adler, S. Agarwal, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Izumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. *arXiv preprint arXiv:1909.00502*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online, April. Association for Computational Linguistics.

OpenAI. 2023. Gpt-3.5-turbo. Accessed: 2024-06-18.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51, Melbourne, Australia, July. Association for Computational Linguistics.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 25–35, Suzhou, China, December. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918. Association for Computational Linguistics.

Liner Yang Yijun Wang Xiaorong Lu Renfen Hu Shan He Zhenghao Liu Yun Chen Erhong Yang Maosong Sun Yingying Wang, Cunliang Kong. 2021. Yaclc: A chinese learner corpus with multidimensional annotation. *arXiv preprint arXiv:2112.15043*.

Y. Zhang, H. Jiang, Z. Bao, et al. 2022a. Mining error templates for grammatical error correction. *arXiv preprint arXiv:2206.11569*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022b. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July. Association for Computational Linguistics.

(c) 2024 中国中文信息学会计算语言学专业委员会

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

J. Zhou, C. Li, H. Liu, et al. 2018. Chinese grammatical error correction using statistical and neural models. *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II*, pages 117–128.

CCL 2024

Overview of CCL24-Eval Task 7: Chinese Essay Fluency Evaluation (CEFE) Task

Xinlin Zhuang[1], **Xinshu Shen**[1], **Hongyi Wu**[1], **Man Lan**[1,2*], **Xiaopeng Bai**[2,3], **Yuanbin Wu**[1,2], **Aimin Zhou**[1,2], and **Shaoguang Mao**[4]

[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]Microsoft Research Asia
{xinlinzhuang,xinshushen,hongyiwu}@stu.ecnu.edu.cn
{mlan,ybwu,amzhou}@cs.ecnu.edu.cn,xpbai@zhwx.ecnu.edu.cn
{shaoguang.mao}@microsoft.com

Abstract

This paper presents a detailed review of Task 7 in the CCL24-Eval: the second **C**hinese **E**ssay **F**luency **E**valuation (CEFE). The task aims to identify fine-grained grammatical errors that impair readability and coherence in essays authored by Chinese primary and secondary school students, evaluate the essays' fluency levels, and recommend corrections to improve their written fluency. The evaluation comprises three tracks: (1) Coarse-grained and fine-grained error identification; (2) Error sentence rewriting; and (3) Essay Fluency Level Recognition. We garnered 29 completed registrations, resulting in 180 submissions from 10 dedicated teams. The paper discusses the submissions and analyzes the results from all participating teams.

1 Introduction

Education is an enduring and evolving journey, continuously reshaping itself, particularly in the wake of the Internet's proliferation and the development of Large Language Models (LLMs) (Zhao et al., 2023), which has also significantly expanded the scope of Chinese essay evaluation. The marked increase in the volume of essays requiring evaluation has highlighted concerns regarding the cost-effectiveness and efficiency of manual essay corrections, positioning these issues as salient factors in contemporary educational methodologies. In light of these challenges, a growing number of scholars and educational institutions have initiated investigations into the feasibility of utilizing computer technologies for Automated Essay Correction (AEC) (Rudner et al., 2006; Ramesh and Sanampudi, 2022). These relative methods fulfill a twofold purpose. **Firstly**, it facilitates the provision of objective, precise, and timely feedback by analyzing various dimensions of an essay, such as language, content, and structure, and addressing inherent writing challenges. This, in turn, potentially enhances students' comprehension of their writing difficulties, thereby improving their overall writing competencies. **Secondly**, it enables educators to more accurately assess students' writing proficiency and offer more focused instructional support, furthering the educational advancement of students. In practical educational settings, a critical component that teachers assess during essay evaluation is the **fluency** of expression. This aspect reflects the essay's coherence and grammatical accuracy, offering insights into the writer's proficiency and ability to convey ideas effectively. Improving fluency is essential for enhancing the accuracy of essay evaluations and raising the authors' writing standards.

Nevertheless, the current evaluation of essay fluency at primary and secondary education levels faces significant challenges: **1) Lack of detailed criteria** Most existing assessments focus broadly on overall essay quality without delving deeply into fluency. There is a noticeable absence of systematic criteria, which hampers a thorough understanding and development of students' writing skills. **2) Limited interpretability** Previous studies often approach fluency as merely a scoring endeavor, providing only an aggregate score or rating. Alternatively, they treat it as a basic Grammatical Error Correction (GEC) task (Gong et al., 2021; Tsai et al., 2020). Such approaches predominantly target simple grammatical

*Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 302-310, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

mistakes, reviewing them through simple corrections like additions, deletions, and modifications, which limits comprehensive feedback on the text's flow and structure. However, these methodologies typically overlook the analysis of specific grammatical error types and fail to specify the exact nature of the errors encountered. Providing students with detailed descriptions of error types and suggested corrections is beneficial as it enables them to recognize their mistakes, refine their essays, and prevent the recurrence of these errors in future writing. **3) Scarcity of authentic data from primary and secondary educational contexts** There is a notable shortage of public datasets for studying essay fluency among Chinese primary and secondary school students. Previous research relying on GEC often utilizes rule-based or inter-language datasets from learners of Chinese as a second language. However, the errors in compositions written by native Chinese students are more diverse and require a deeper understanding of complex grammatical structures. Figure 1 presents examples of sentences taken from essays written by primary and secondary school students. These excerpts showcase various errors along with recommended corrections. Typically, each sentence contains multiple types of errors that extend beyond simple spelling mistakes, illustrating the complexity of issues found in student writing.

Chinese Sentence	English Translation
Sentence：我一共种了两株在阳台上，我平时见不到它们，只有在周末才能望上几眼。 ErrorType：语序不当、主语多余 RevisedSentence：我在阳台上一共种了两株，平时见不到它们，只有在周末才能望上几眼。	Sentence：I planted two plants in total on the balcony. I can't see them usually, only catch a glimpse of them on weekends. ErrorType：Inappropriate Word Order,Subject Redundancy RevisedSentence： I on the balcony planted two plants in total, andcan't see them usually, only catch a glimpse of them on weekends.

Figure 1: An example of our task. In modern Chinese, adverbials are typically positioned between the subject and the predicate rather than at the end of the sentence, thereby leading to an 'Inappropriate Word Order' error. Moreover, in the first two short sentences, there is a problem of 'Subject Redundancy' where the subject 'I' is repeated unnecessarily.

These shortcomings in current methodologies highlight the need for a more detailed and nuanced approach that not only identifies fine-grained errors but also delivers specific, actionable feedback to students. It underscores the importance of utilizing composition data from the authentic writing contexts of primary and secondary school students to better address their specific learning needs. Motivated by this, we present the CCL24-Eval task [0]: ***Chinese Essay Fluency Evaluation*** (**CEFE**), which aims to identify and correct errors that impede the fluency of writing in essays by primary and secondary school students and to assess the overall fluency level of an essay. Compared to Task 8 of the CCL23-Eval (Shen et al., 2023), we have introduced a new track and removed an original one, taking into account their practical application and relevance. This task featured three tracks: (1) *Coarse-grained and fine-grained error identification*; (2) *Error sentence rewriting*; (3) *Essay fluency level recognition*, aiming at providing a higher-quality evaluation of fluency in primary and secondary school essays.

This task attracted 29 teams to sign up for the competition, and in the end, we received 180 submissions from 10 teams. The task description is presented in Section 2. We describe the data we used in this task in Section 3. We explain baselines used for each track and list participants' information and results from their submissions and provide a more in-depth discussion in Section 5.

2 Task Description

Our evaluation is structured into three distinct tracks, each crafted to tackle specific aspects of identifying and correcting errors in essays written by primary and secondary school students. These tasks are designed to shed light on the common types of errors these students make, offering a basis for focused enhancements in their writing skills.

[0] https://github.com/cubenlp/2024CCL_CEFE

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 302-310, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2.1 Track 1: Coarse-Grained and Fine-Grained Error Identification

Track 1 is dedicated to identifying types of grammatical errors in compositions from primary and secondary school students. Traditional methods often overlook these errors, failing to specifically highlight the diverse writing challenges students face. This track addresses the issue through two analytical lenses: character-level and component-level errors. We define four broad categories of grammatical errors: *Character-Level Error **(CL)***, *Incomplete Component Error **(IC)***, *Redundant Component Error **(RC)***, and *Incorrect Constituent Combination Error **(ICC)***. Additionally, we have outlined fourteen fine-grained error types, offering a deeper insight into the potential mistakes in student writing. Given the developmental stage of primary and middle school students, their compositions often contain multiple errors within the same sentence, making this task particularly challenging. Consequently, this track is structured as a multi-label classification task. Overall, Track 1 includes a total of 4 coarse-grained error types and 14 fine-grained error types. Detailed descriptions and examples of each error type are provided on the competition homepage, and the specific category definitions are as follows:

Character-Level Error (CL) includes four fine-grained error types: Word Missing (**WM**), where a word in a commonly used fixed collocation is missing from the sentence and needs to be added; Typographical Error (**TE**), where there are typos in the sentence that need to be revised or deleted; Missing Punctuation (**MP**), where punctuation is missing from the sentence and needs to be added; and Wrong Punctuation (**WP**), where the punctuation used in the sentence is wrong and needs to be revised or deleted.

Redundant Component Error (RC) is composed of three fine-grained error types: Subject Redundancy (**SR**), which occurs when a complex adverb is followed by a repeated subject referring to the same entity, and the modification is to delete one subject; Particle Redundancy (**PR**) refers to the redundant use of particles, which should be deleted during editing; Other Redundancy (**OR**) refers to any redundant elements not covered by the previous types, which should also be deleted in modification.

Incomplete Component Error (IC) consists of four fine-grained error types: Unknown Subject (**US**), which occurs when the sentence lacks a subject or the subject is unclear, and the solution is to add or clarify the subject; Predicate Missing (**PM**) refers to a sentence lacking verbs, which may be corrected by adding predicates; Object Missing (**OBM**) means that a sentence lacks an object, and the solution is to add an object; Other Missing (**OTM**) refers to other missing components besides the incomplete subject, predicate, and object, which may be corrected by adding the missing components except for the subject, predicate, and object.

Incorrect Constituent Combination Error (ICC) includes three fine-grained error types: Inappropriate Verb-Object Collocation (**IVOC**) refers to the predicate and object not being properly matched, and may be corrected by replacing either the predicate or object with other words; Inappropriate Word Order (**IWO**) means that the order of words or clauses in the sentence is unreasonable, and may be corrected by rearranging some words or clauses; Inappropriate Other Collocation (**IOC**) refers to any element in the sentence not covered by the previous types being improperly matched, and may be corrected by replacing it with other words.

2.2 Track 2: Error Sentence Rewriting

Track 2 focuses on the rewriting of incorrect sentences in compositions by primary and secondary school students. The main challenge of this track is to devise a minimal modification strategy for these erroneous sentences, ensuring that the original meaning is preserved. The corrections should involve as few changes as necessary because over-modifying can obscure the original errors, making it difficult for students to recognize and learn from their mistakes. This is crucial for teachers to better understand the writing challenges their students face, and to aid in improving their writing skills. It emphasizes the importance of maintaining the student's original thought process while steering them towards grammatical accuracy and clearer expression.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 302-310, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2.3 Track 3: Essay Fluency Level Recognition

Track 3 is designed to assess the overall fluency of an entire essay by examining the organization of words, sentences, and paragraphs. By evaluating the fluency of essays, this approach offers teachers a more efficient and intuitive method to assess students' writing skills. It also provides students with a clearer understanding of their writing performance. Defined as a multi-label classification task, Track 3 categorizes essays into three fluency levels: *excellent*, *average*, and *failing*. This classification helps in distinguishing essays based on their coherence and structural organization, allowing for targeted feedback that can guide improvements in students' writing abilities.

3 Datasets

In an effort to enhance research on essay fluency among primary and secondary school students, we meticulously annotated a dataset with fine-grained grammatical error types and provided corresponding corrections that impact sentence fluency. We developed the second **C**hinese **E**ssay **F**luency **E**valuation (**CEFE 2.0**) dataset based on CEFE 1.0 dataset (Shen et al., 2023), aiming to offer in-depth insights into the typical grammatical errors students make. This fine-grained dataset not only helps in identifying common mistakes but also supports the development of targeted teaching strategies to improve writing skills in young learners.

3.1 Data Collection

The foundational material for our dataset was derived from actual essays written by primary and secondary school students during their examinations. These compositions span a variety of genres, including character and scene descriptions, chosen for their genuine representation of students' writing skills. This choice of data source is pivotal as it captures the authentic richness of real-world writing scenarios. Exam essays particularly offer unfiltered insights into the writing abilities, habitual patterns, and prevalent errors among students in these educational stages. The diversity and complexity of the errors and required revisions found in these essays mirror the real challenges students face, providing a robust basis for our research. By utilizing these authentic compositions, our findings and proposed solutions remain highly relevant and directly applicable to improving student writing, thereby maximizing the impact of our work.

3.2 Data Annotation

The annotation team was composed of four undergraduate students, four postgraduate students specializing in language-related disciplines, and four expert reviewers with backgrounds in Chinese teaching. This diverse group was responsible for identifying error types and suggesting sentence revisions, adhering to the principle of **minimal changes**. Prior to beginning their tasks, all annotators underwent a training session designed to familiarize them with the specific annotation guidelines. The annotation process was structured as follows: an initial annotation was conducted collaboratively by an undergraduate and a postgraduate student. Subsequently, expert reviewers performed a verification pass to confirm the accuracy and reliability of the annotations, making necessary adjustments where needed. The annotated data was organized into five groups, and the team conducted weekly online meetings to discuss prevalent issues and refine their approach. This comprehensive process not only focused on pinpointing specific errors but also on providing actionable correction suggestions. Such a dual approach enhances the clarity of the feedback and equips students with the tools they need to improve their writing skills effectively.

3.3 Data Statistics

This section delineates the distribution of training, validation, and test datasets for each track. Given the prevalent scarcity of annotated data in real-world contexts, participants are tasked with developing robust models for assessing sentence fluency using a limited dataset. The test dataset includes both correct and intentionally flawed sentences, with a portion of the data reserved for blind evaluation. The statistics for our dataset are presented in Table 1.

CCL 2024

	Train Set	Dev Set	Test Set
Track 1	1,000	100	2,000
Track 2	1,000	100	2,000
Track 3	100	10	2,000

Table 1: The statistics of the **CEFE 2.0** dataset. The number for Track 1 and Track 2 corresponds to individual sentences, whereas for Track 3, it represents entire essays.

4 Evaluation Metrics

Different evaluation metrics are utilized across the various tracks of the task; however, the calculations for precision and recall remain consistent throughout. **Precision** is calculated as the ratio of correctly identified instances to the total number of instances identified by the model. Conversely, **Recall** is calculated as the ratio of correctly identified instances to the total number of instances labeled in the ground truth. The **F1-score**, frequently employed in binary or multi-class classification tasks, represents the harmonic mean of precision and recall, and is computed using the formula: $F_1 = \frac{2PR}{P+R}$.

4.1 Track1: Coarse-Grained and Fine-Grained Error Identification

The total score of Track1 is composed of two parts: coarse-grained and fine-grained wrong sentence identification score. The specific calculation method is as follows:

$$Score_{Track1} = 0.5 * F_1^{Coarse-grained} + 0.5 * F_1^{Fine-grained} \quad (1)$$

Specifically, precision (P), recall (R), and micro F_1 are used to evaluate the recognition effect of coarse and fine-grained wrong sentence types (See details in Section 2.1).

4.2 Track2: Error Sentence Rewriting

Due to the variety of rewriting outcomes, we assess the results of the model from two distinct perspectives:

Comparison with Gold References: We utilize three evaluation metrics: 1) **Exact Match (EM)**: This metric calculates the percentage of sentences generated by the model that perfectly align with the gold standard references. 2) **Edit Metrics** (proposed by MuCGEC) (Zhang et al., 2022): This method transforms error-correct sentence pairs into a series of operations and compares these operations produced by the model against the correct references, subsequently calculating precision, recall, and $F_{0.5}$ scores. 3) **BLEU** (Papineni et al., 2002): This metric assesses the N-gram overlap between sentences generated by the model and the correct references, providing a measure of linguistic similarity.

Correctness and Reasonableness of Results: We also apply three metrics to evaluate the rewritten sentences: 1) **Perplexity (PPL)**: Utilizing BERT (version bert-base-chinese) (Kenton and Toutanova, 2019), this metric gauges the fluency and predictability of the rewritten sentences. 2) **Levenshtein Distance**: This measures the edit distance between the rewritten sentence and the original, aiming to achieve accurate corrections with minimal edits to maintain clarity in understanding the nature of the errors. 3) **BERTScore** (Zhang et al., 2020): This score quantifies the semantic similarity between the rewritten and original sentences, ensuring the corrections maintain contextual integrity.

These metrics are subsequently weighted to compute a final score, effectively balancing various aspects of quality in the rewritten sentences:

$$Score_{Track2} = (EM + BLEU + F_{0.5} + BERTScore)/4 - Levenshtein - PPL_{BERT} \quad (2)$$

4.3 Track 3: Essay Fluency Level Recognition

To evaluate the classification performance of elementary and secondary school essay fluency ratings, we employ a range of metrics: Accuracy (Acc), Precision (P), Recall (R), Macro F1, and Quadratic

CCL 2024

Weighted Kappa (QWK). Given that the QWK ranges from [-1, 1], we first normalize it to the interval [0, 1] before incorporating it into the weighted final score calculation. The composite score for track three is calculated as follows:

$$Score_{Track3} = 0.5 * F_1 + 0.2 * QWK + 0.3 * Acc \quad (3)$$

5 Results and Analysis

5.1 Baselines

We provide the results of our baseline models as a reference. For Track 1 and Track 3, we fine-tuned BERT (version *bert-base-chinese*) (Kenton and Toutanova, 2019) over the corresponding training datasets for 5 epochs, utilizing batch sizes ranging from 16 to 24, a learning rate of 2×10^{-5}, and employing the Adam optimizer. For Track 2, we fine-tuned BART (version *bart-base-chinese*) (Lewis et al., 2020) on the training dataset for 5 epochs, with a fixed batch size of 16, a learning rate of 2×10^{-5}, and the AdamW optimizer. Detailed results of these baseline models are provided in Section 5.

5.2 Results

In our competition, a total of 10 teams submitted their final results. The basic information about them are detailed in Table 2. Table 3 displays the final results of the participating teams in this competition. The average score across the three tracks will serve as the final score for the current team.

ID	Team Name	Organization	Track 1	Track 2	Track 3
1	Smartdot	Smartdot Technologies Co.,Ltd.	✓	✓	✓
2	BNU	Beijing Normal University	✓	✓	✓
3	ZUT-POLab	Zhongyuan University of Technology	✓	✓	✓
4	GDUFS-CL	Guangdong University of Foreign Studies	✓	✓	✓
5	SIAT-UI	Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences	✓	✓	✓
6	ZZU1	Zhengzhou University	✗	✓	✗
7	CPIC	China Pacific Insurance(Group) Co., Ltd.	✓	✓	✓
8	KUST	Kunming University of Science and Technology	✓	✓	✓
9	DeakinAI	Deakin University	✓	✓	✓
10	ZZU2	Zhengzhou University	✓	✓	✗
Total Number		29	27	28	26

Table 2: The basic information of the participants with a total of 29 teams, where 27 teams for Track 1, 28 teams for Track 2 and 26 teams for Track 3.

6 Participant Systems

In this task, the competing teams employed a variety of approaches to detect and correct errors in student essays and to assign grades based on fluency. This section will provide an overview of the methods that were successful in each track. The unique approaches of each team demonstrate the diversity of methods that can be exploited in automated essay scoring and present various potential directions for future research.

6.1 Track1: Coarse-Grained and Fine-Grained Error Identification

For Track 1, ***Smartdot*** implemented a two-stage fine-tuning strategy, initially utilizing publicly available datasets to augment data through grammar error substitutions and fine-tuning using the expanded dataset with the UTC model (Universal Text Classification). Subsequently, they leveraged the training data provided by the competition for a second phase of fine-tuning. ***ZUT-POLab*** first augmented the training data through operations such as insertion, substitution, and deletion, and then fine-tuned the ERNIE 1.0 model using the expanded dataset. ***GDUFS-CL*** analyzed two fine-grained errors, utilizeda binary

Rank	Team Name	Track 1	Track 2	Track 3	Score
1	Smartdot	48.48	52.41	50.56	50.48
2	BNU	45.19	43.90	45.84	44.98
3	ZUT-POLab	41.66	43.49	46.98	44.04
4	GDUFS-CL	36.47	41.09	51.96	43.17
5	SIAT-UI	37.26	42.48	47.64	42.46
6	ZZU1	-	46.06	-	41.72
7	CPIC	35.42	45.32	33.38	38.04
	baseline	34.62	30.00	44.48	36.37
8	KUST	34.19	24.87	45.71	34.92
9	DeakinAI	0.00	16.77	45.40	20.72
10	ZZU2	0.00	0.00	-	14.83

Table 3: Final scores of the participating teams. "-" indicates that the team did not submit evaluation results on the track, and the overall score is calculated based on the baseline.

classification model for prediction optimization, compared and selected trainingcorpora, and trained a coarse-grained model based on the Chinese Learner 4W corpus.

Some teams also leveraged Large Language Models (LLMs) to assist in their tasks. ***BNU*** first undertook data augmentation, scraping essays from websites such as primary and secondary school essay networks. They then used LLMs like GPT-3.5 to generate specific types of flawed sentences starting from given correct sentences. After constructing the data, they fine-tuned BERT using both the augmented and provided training data. ***SIAT-UI*** utilized the released task data to fine-tune the Qwen1.5-7b-chat model.utilized the released task data to fine-tune the Qwen1.5-7b-chat model.

6.2 Track2: Error Sentence Rewriting

For Track 2, ***Smartdot*** proposed a two-stage strategy: in the first stage, they pre-trained BART using pseudo-native data and the NaSGEC dataset, and incorporated SynGEC for grammatical error correction. In the second stage, they fine-tuned the model on the training dataset provided by the competition. ***ZZU1***, based on the BART model, proposed employing multi-pass decoding within the sequence-to-sequence framework to iteratively refine the corrections from the previous round, and additionally introduced an early stopping mechanism to reduce computational costs. ***ZUT-POLab*** proposed a diffusion generative model where, during the forward process, each step's text is encoded using ERNIE 1.0, and in the reverse process, the text modeling capabilities of ERNIE are utilized to progressively decode the masked tokens.

BNU utilized LLMs to address this task. They initially constructed hybrid training data by enhancing the quality of the data created for Track 1 through data processing and cleaning efforts, and by integrating open-source datasets such as MuCGEC. Subsequently, they fine-tuned the Qwen-14B model based on the crafted hybrid data and designed specific prompt templates to reinforce the model's generalization ability.

6.3 Track 3: Essay Fluency Level Recognition

For Track 3, ***GDUFS-CL*** employed back-translation techniques to construct pseudo-data with triple-labeled fluency ratings for pre-training and adapting an NSP-based strategy to effectively utilize contextual information and avoid long sequence dependencies. ***Smartdot*** The Smart team initially selected the TextRCNN-NeZha model as their foundational model. They then introduced the MulDrop strategy and employed the DCE loss for the classification of essay fluency levels. ***SIAT-UI*** utilized the data provided by the competition to fine-tune the Bert-Large-Chinese model. ***ZUT-POLab*** first employed operations such as insertion, substitution, and deletion to augment the training data, and then fine-tuned the ERNIE 3.0 model using the expanded dataset. ***BNU*** employed a multitask learning approach, where, in addition to the primary task of essay fluency level recognition, they also trained an auxiliary model for the identification of grammatical errors in these essays to support the fluency grading.

7 Conclusions and Future Work

This paper presents an overview of the CCL24-Eval Task the second ***C**hinese **E**ssay **F**luency **E**valuation* (**CEFE**). We conduct this evaluation using our meticulously annotated **CEFE 2.0** dataset. The evaluation is divided into three distinct tracks: (1) Coarse-grained and fine-grained error identification; (2) Error sentence rewriting; and (3) Essay Fluency Level Recognition. We received a total of 29 completed registration forms, culminating in 180 submissions from 10 participating teams. In addition, we provide a comprehensive analysis and summary of the methodologies employed by the participants, which will contribute to future research in this field of natural language processing. The findings indicate that the employment of LLMs and the application of data augmentation techniques contribute to enhancing the aggregate scores.

Acknowledgements

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004) and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. Iflyea: A chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Lawrence M Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Xinshu Shen, Hongyi Wu, Xiaopeng Bai, Yuanbin Wu, Aimin Zhou, Shaoguang Mao, Tao Ge, and Yan Xia. 2023. Overview of CCL23-eval task 8: Chinese essay fluency evaluation (CEFE) task. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 282–292, Harbin, China, August. Chinese Information Processing Society of China.

Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S Chang. 2020. Lingglewrite: a coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

CCL 2024

System Report for CCL24-Eval Task 8: A Two-stage Prompt-Based Strategy for CRMUS Track 1

Mosha Chen
Hangzhou Guangran Digital Technology Co., Ltd, Hangzhou, China
chenmosha@holoflow.cn

Abstract

Large Language Model (LLM) has sparked a new trend in Natural Language Processing, and an increasing number of researchers have recognized the potential of using LLM to unify diverse NLP tasks into a text-generative manner. To explore the potential of LLM for the children's stories domain, CCL2024 has released the **C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories (CRMUS) task. This paper presents a straightforward yet effective two-stage prompt-based strategy for the CRMUS Track 1. In the initial stage, we use the same prompt to obtain responses from GPT-4, ERNIE-4, and Qwen-Max. In the subsequent stage, we implement a voting mechanism based on the results from the first stage. For records with inconsistent outcomes, we query GPT-4 for secondary confirmation to determine the final result. Experimental results indicate that our method achieved an average score of 79.27, securing first place in the closed domain among ten participating teams, thereby demonstrating the effectiveness of our approach.

1 Introduction

With the popularization of ChatGPT, Large Language Model (LLM) has motivated an increasing trend in both industry and academia; an increasing number of researchers are exploring the potential of LLMs (Steven et al., 2023; Chen and Si, 2024; Zhang et al., 2023; Wu et al., 2023; He et al., 2023) across various domains, leading to the proposal of new paradigms for NLP tasks. Following the trend, the **C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories (CRMUS) evaluation task[0] is introduced in CCL2024, which aims to evaluate Chinese pre-trained language models and large language models from multiple perspectives in terms of commonsense reasoning and moral understanding on the children's education domain. In order to investigate different techniques in the field of LLMs, the CRMUS task provides two tracks: prompt-based and fine-tuning of LLM parameters.

The purpose of the prompt-based track in the CRMUS task is to assess LLM's potential in story commonsense reasoning and moral understanding. The types of commonsense involved in the commonsense reasoning task cover a wide range of aspects: temporal commonsense, spatial commonsense, biological commonsense, physical commonsense, and social commonsense. Given the comprehensive reasoning capabilities of LLMs, we hypothesized that a prompt-based approach would effectively fit the CRMUS task. To this end, we propose a straightforward yet effective two-stage prompt engineering pipeline:

- In the first stage, we use a uniform prompt to obtain responses from three advanced commercial LLMs: GPT-4, ERNIE-4, and Qwen-Max.
- In the second stage, we adopted a majority voting strategy for the LLM responses from the first step. For the inconsistent results, we query GPT-4 for a secondary confirmation, with a slightly different prompt from the first step, which narrows down the range of options using only the options returned from the first step. This secondary confirmed choice is chosen as the final submission result.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

[0]https://github.com/SXU-YaxinGuo/CRMU

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 311-319, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Our experimental results demonstrate that our method achieved an average score of 79.27, ranking first in the closed domain among ten participating teams. This confirms the effectiveness of our approach. Furthermore, our method leverages the strengths of multiple LLMs and a novel two-stage voting mechanism, offering a robust solution for commonsense reasoning and moral understanding tasks without requiring extensive fine-tuning. This innovation lies in the efficient combination of multiple models' strengths and the strategic use of prompt engineering to enhance performance in the CRMUS task.

2 Related Word

2.1 Prompt Engineering

Recent research has highlighted that the use of prompts can substantially enhance the performance of pre-trained language models in various downstream applications(Brown et al., 2020). This has sparked significant interest in the effective construction of prompts. Manual prompt creation, however, is labor-intensive and not always feasible. To address this, Shin(2020) proposed a discrete word space search algorithm that leverages downstream application training data. While this approach outperforms manual prompt design, it is limited by the weak expressive capability of discrete prompts, resulting in only modest improvements in downstream tasks. To overcome these limitations, some researchers have introduced prompt-tuning methods(Lester et al., 2021) that optimize continuous prompt vectors through gradient backpropagation. These methods have demonstrated considerable performance gains; however, since parameter fine-tuning is not allowed in the prompt-based track, we adopted the approach of manually designing prompts. Unlike these methods, our approach not only involves manually designed prompts but also integrates responses from multiple LLMs through a novel voting mechanism, enhancing the robustness and accuracy of the final results.

2.2 In-context Learning

The release of GPT-3(Radford et al., 2019), OpenAI's former state-of-the-art large language model, has significantly drawn the research community's attention to a novel area: In-Context Learning (ICL). The authors demonstrated that GPT-3, a self-supervised pre-trained model, can effectively perform new tasks without prior specific training, simply by giving a manually designed prompt that includes an optional task description and a few example demonstrations. This groundbreaking capability has spurred extensive research exploring various facets of ICL. The performance of ICL has been demonstrated to be highly sensitive to the selection of demonstration examples(Li et al., 2023). To address this issue, Rubin(2022) proposed methods for learning to retrieve suitable demonstration examples. Li(2023) introduced a series of techniques to enhance demonstration selection performance. Levy(2023) focused on selecting diverse demonstrations to improve in-context compositional generalization. Furthermore, Qin(**?**) developed an iterative approach that selects diverse examples yet closely correlates with the test sample for ICL demonstrations. Our work differs from ICL methods as we do not rely on example demonstrations for task performance. Instead, we focus on leveraging the diverse capabilities of multiple LLMs and a structured voting strategy, which we believe offers a more direct and reliable approach to addressing the CRMUS task.

3 Our Methods

Due to the comprehensive reasoning capabilities of LLM, we believe the prompt-based method could fit the CRMUS task well. We propose a simple yet effective two-stage prompt-based pipeline for the evaluation tasks. The overall processing pipeline, illustrated in Figure 1, applies to the Commonsense Reasoning (CR) subtask. The Moral Understanding (MR) subtask follows the same pipeline except for the prompt.

3.1 First Stage: Query LLMs

In the first stage, we select three of the most advanced commercial LLMs—GPT-4, ERNIE-4, and Qwen-Max—as our testbeds. The same prompt is applied to each LLM.

CCL 2024

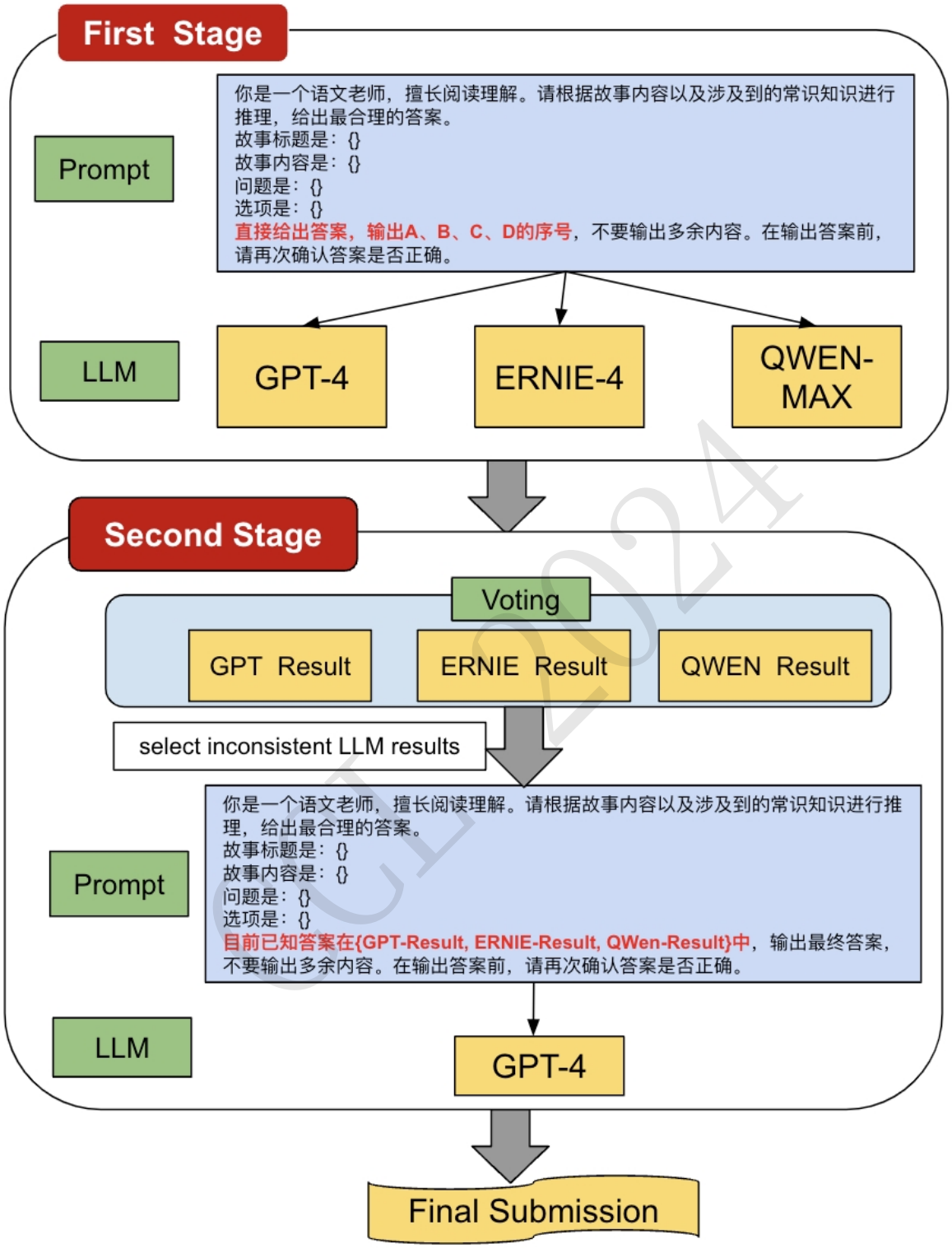


Figure 1: Two-stage prompt-based pipeline for the Common Reasoning (CR) subtask

```
你是一个语文老师，擅长寓言阅读理解。请根据故事内容以及涉及到的常识知识进行推理，给出最合理的答案。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

直接给出答案，输出A、B、C、D的序号，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 2: CR subtask prompt for the first stage

```
你是一个语文老师，擅长寓言阅读理解。请基于给定的故事，从4个候选答案中选择最恰当的、最符合故事情节的寓意。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

直接给出答案，输出A、B、C、D的序号，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 3: MU subtask prompt for the first stage

3.1.1 Prompt Design

Our designed prompts adhere to conventional prompt elements, including persona (e.g., a language teacher proficient in reading comprehension tasks), instruction (e.g., complete commonsense reasoning or moral understanding tasks for fables), input (e.g., story title, story content, questions, and options), and constraints (e.g., output only the answer option labels). We believe that current LLMs are sufficiently powerful to handle common understanding and reasoning tasks; hence, we did not invest effort in selecting demonstrations that are important in ICL. We have established different prompts for the CR and MU tasks. Please refer to Figures 2 and 3 respectively.

The results returned by LLMs will be further processed to extract the answer label.

3.1.2 Post-Processing

Although we have explicitly asked the LLMs to return the option label only, LLMs always return additional information, such as explanations. Observing the output from LLMs, we have identified a common pattern: in most instances, these models first present an answer label, followed by an explanation. Consequently, we have adopted a simple heuristic rule: traverse the output result from the beginning, and if the current character being traversed corresponds to one of the option labels A˜D, return the matched label as the post-processed result. It is important to note that this rule is not universally applicable. For example,

```
你是一个语文老师，擅长寓言阅读理解。请根据故事内容以及涉及到的常识知识进行推理，给出最合理的答案。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

目前已知答案在{}几个选项中，给出最终的答案。
输出{}的序号即可，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 4: CR subtask prompt for the second stage

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 311-319, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

```
你是一个语文老师，擅长寓言阅读理解。请基于给定的故事，从4个候选答案中选择最恰当的、最符合故事情节的寓意。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

目前已知答案在{}几个选项中，给出最终的答案。
输出{}的序号即可，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 5: MU subtask prompt for the second stage

models like QWen occasionally follow a different pattern, analyzing each option before providing a final answer. In such cases, our designed heuristic rule fails to apply.

3.2 Second Stage: Vote & Secondary LLM Confirmation

After obtaining results from three LLMs, we conduct a majority vote on the results. For instances where the voting results are inconsistent, we have further designed a second-stage prompting strategy to resolve the discrepancies.

In the development set, we observed that GPT's results significantly outperformed those of ERNIE and QWEN for the CR task. In the MU task, ERNIE's results were slightly better than GPT's, but the advantage was marginal. Therefore, we use GPT's results as the primary basis for voting, categorized into the following four scenarios (assuming the results returned by GPT, ERNIE, and QWEN are denoted as A, B, and C, respectively).

- **Case 1**: If all three models return the same result ($A = B = C$), we adopt this result.
- **Case 2**: If GPT's result matches one other model's result ($A = B$ or $A = C$), we adopt GPT's result.
- **Case 3**: If all three models return different results ($A \neq B \neq C$), we use the second-stage prompting strategy to resolve the discrepancy.
- **Case 4**: If ERNIE and QWEN's results match and differ from GPT's result ($B = C \neq A$), we consider the second-stage prompting strategy to resolve the discrepancy.

We define inconsistencies as cases where the results returned by GPT do not match other LLMs' outcomes, specifically cases 3 and 4. To address these inconsistencies, we have designed a secondary prompt strategy to request GPT once more to obtain the final result. The prompts in the second phase are essentially the same as those in the first phase, with the only difference being the set of answer options. In the first phase, the answer options consist of the initial four complete options from the test set. In the second phase, the prompts include answer options derived from the responses from the three LLMs in the first phase. The second step prompt for the CR & MU tasks is illustrated in Figures 4 and 5.

The outcome generated by GPT in the second step undergoes post-processing using the same method described in Section 3.1.2. The post-processed result is then chosen as the final submission answer.

4 Experiments

4.1 Datasets and Evaluation Metrics

The classic fable used in the CRMUS task was manually collected from web sources. The questions and answers for the CR subtask were manually annotated. A combination of automated construction and manual annotation was employed for the MU subtask. Overall, the annotation quality is high.

Table 1 presents the development and test data statistics.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 311-319, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

split	subtask type	records number
dev set	CR	400
dev set	MU	252
test set	CR	1,692
test set	MU	1,056

Table 1: Statistics of the dev & test set

	CR Acc	MU Acc
GPT-4	88.00	71.33
ERNIE-4	82.67	72.67
QWen-MAX	82.00	70.67

Table 2: Accuracy for 150 randomly selected dev set records

The evaluation metric for both subtasks is accuracy. The final evaluation score of the competing model is calculated as the weighted average of all evaluation metrics, defined as:

$$\text{Score} = 0.4 \times \text{Acc}_1 + 0.6 \times \text{Acc}_2$$

where Acc_1 represents the accuracy of the CR subtask, and Acc_2 represents the accuracy of the MU subtask.

4.2 Experimental Setup

We use the three most advanced commercial LLMs – GPT-4[1], ERNIE-4[2], and Qwen-Max[3] – as our testbed. We directly utilized the API of the large models, with all model parameters set to the official API default values. Our experimental code is released at https://github.com/Holoflow/CCL2024-CRMUS-Track1.

4.3 Experimental Results

Considering the cost of LLMs, we evaluated[4] the performance of the three LLMs using only 150 randomly selected records from each task in the development set. Specifically, we randomly selected 150 records from the CR subtask and 150 records from the MU subtask. The accuracy of the three LLMs on the dev set is shown in Table 2.

From the dev set, we can draw a preliminary conclusion: GPT's results significantly outperformed those of ERNIE and QWen for the CR task. In the MU task, ERNIE's results were slightly better than GPT's, but the advantage was marginal. This also inspired us to base the voting strategy in the second stage primarily on GPT's results.

Table 3 represents the detailed result for the test set[5]. It is important to note that the **voting** method in the table is not the voting strategy described in Section 3 of this paper, but rather a majority voting strategy. In cases where the results of the three LLMs are inconsistent, the CR subtask uses the results from GPT-4, while the MU task uses the results from ERNIE-4.

It should also be noted that Qwen's experimental results are lower than those of the actual situation. This is because, during the post-processing stage, Qwen's result pattern differs slightly from that of the other two LLMs as explained in 3.1.2.

4.4 Analysis

Based on the experimental results from the dev and test sets, we can draw the following conclusions:

[1]Model snapshot used is gpt-4-turbo-2024-04-09
[2]Model snapshot used is Ernie-4.0-8K-0329
[3]Model snapshot used is qwen-max-0403
[4]We used the official evaluation script to run the results locally
[5]The results are extracted from the public dashboard: http://cuge.baai.ac.cn/#/ccl/2024/crmus

	CR Acc	MU Acc	Weight Score
GPT-4	84.46	68.37	74.80
ERNIE-4	78.30	69.13	72.80
QWen-MAX	76.83	66.57	70.68
Voting	84.99	71.21	76.72
Voting & Secondary Confirmation	86.52	74.43	79.26

Table 3: Results on the test set

- The MU tasks are more challenging than the CR tasks because CR tasks focus on reasoning based on objective facts, whereas MU tasks emphasize the understanding of subjective meanings. In the development set, our analysis of MU tasks revealed the presence of many ambiguous options for human beings. This also explains why the task organizers assigned a higher weight to the MU tasks in the evaluation metrics.
- The distribution of the dev set and the test set is significantly different. The same method shows a noticeable decrease in performance on the test set, with an average drop of 3-4 percentage points. It is speculated that the task organizers included more challenging data in the test set.
- The performance of the three LLMs is consistent across both the dev and test sets.
- A simple majority voting strategy does not significantly improve the performance of the CR subtask, with only a 0.5 percentage point increase compared to the best single model (GPT-4). However, it significantly improves the performance of the MU subtask, with an increase of more than 2 percentage points compared to the best single model (ERNIE-4).
- The secondary prompt enhanced strategy shows a significant improvement over the voting strategy in both the CR and MU subtasks, further validating the effectiveness of our approach.

5 Conclusion

This work proposes a simple yet effective two-stage prompt-based strategy for the CRMUS Track 1. In the first stage, we employ the same prompt to obtain responses from GPT-4, ERNIE-4, and Qwen-Max. We adopt a voting and secondary prompt-based confirmation strategy in the second stage. The experimental results demonstrate that our method achieves an average score of 79.27, ranking first in the closed domain among ten participating teams, thus confirming the effectiveness of our approach. These results further validate the advantages of LLMs in story reasoning and moral understanding, showcasing their potential in addressing complex tasks in the children's education domain. The successful application of our method underscores the robustness and reliability of combining multiple LLMs and utilizing a strategic voting mechanism.

Our approach demonstrates that even with a simple prompt-based strategy, significant improvements can be achieved by leveraging the diverse strengths of different LLMs. The integration of a secondary confirmation step ensures higher accuracy and consistency in the results, which is particularly beneficial for tasks requiring nuanced understanding and reasoning. We believe that LLMs will play an increasingly important role in children's education, offering sophisticated tools for story reasoning and moral comprehension that can adapt to the needs of young learners.

6 Future Work

In the future, we will continue to explore the potential of large models in CRMUS tasks from two main directions:

- Investigating More Effective Prompt Optimization Strategies: We aim to enhance the performance of general LLMs in CRMUS tasks, particularly in the Moral Understanding (MU) sub-tasks. This involves developing and refining prompt engineering techniques to better capture the nuances of

commonsense reasoning and moral judgment. By optimizing prompt design and employing advanced strategies for prompt adaptation, we can improve the LLMs' ability to understand and respond to complex educational scenarios more effectively.

- Developing Specialized Domain Models: We will focus on creating models specifically tailored to children's educational contexts. These specialized domain models will be designed to better adapt to the unique requirements of CRMUS and similar tasks, ensuring that the models can provide more accurate and contextually appropriate responses. By training models on datasets that reflect the specific language, themes, and moral considerations relevant to children's stories, we can enhance the applicability and effectiveness of LLMs in educational settings.

By pursuing these directions, we hope to push the boundaries of what LLMs can achieve in the field of children's education, contributing to the development of intelligent educational tools that support and enhance learning experiences for young learners.

7 Acknowledgements

We greatly thank all anonymous reviewers for their helpful comments. Thanks to Shanxi University and Hefei University of Technology for providing the high-quality CRMUS dataset. We also thank Min Zhou, Ye Yang, Wenyi Jiang, and Zhenjun Wang for helpful discussions.

References

Moore, Steven and Tong, Richard and Singh, Anjali and Liu, Zitao and Hu, Xiangen and Lu, Yu and Liang, Joleen and Cao, Chen and Khosravi, Hassan and Denny, Paul and Brooks, Chris and Stamper, John. 2023. *Empowering Education with LLMs - The Next-Gen Interface and Content Generation*. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pages:32-37.

Chen, Yuetian and Si, Mei. 2024. *Reflections & Resonance: Two-Agent Partnership for Advancing LLM-based Story Annotation*. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pages:13813-13818.

Zhang, Dell and Petrova, Alina and Trautmann, Dietrich and Schilder, Frank. 2023. *Unleashing the Power of Large Language Models for Legal Applications*. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pages:5257–5258.

Wu, Shijie and Irsoy, Ozan and Lu, Steven and Dabravolski, Vadim and Dredze, Mark and Gehrmann, Sebastian and Kambadur, Prabhanjan and Rosenberg, David and Mann, Gideon. 2023. *BloombergGPT: A large language model for finance*. ArXiv preprint ArXiv: 2303.17564.

Kai He and Rui Mao and Qika Lin and Yucheng Ruan and Xiang Lan and Mengling Feng and Erik Cambria. 2023. *A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics*. ArXiv preprint ArXiv: 2310.05694.

Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda and Agarwal, Sandhini and Herbert-Voss, Ariel and Krueger, Gretchen and Henighan, Tom and Child, Rewon and Ramesh, Aditya and Ziegler, Daniel and Wu, Jeffrey and Winter, Clemens and Hesse, Chris and Chen, Mark and Sigler, Eric and Litwin, Mateusz and Gray, Scott and Chess, Benjamin and Clark, Jack and Berner, Christopher and McCandlish, Sam and Radford, Alec and Sutskever, Ilya and Amodei, Dario. 2020. *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, pages:1877-1901.

Shin, Taylor and Razeghi, Yasaman and Logan IV, Robert L. and Wallace, Eric and Singh, Sameer. 2020. *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages:4222-4235.

Lester, Brian and Al-Rfou, Rami and Constant, Noah. 2021. *The Power of Scale for Parameter-Efficient Prompt Tuning*. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,pages:3045-3059.

Radford, Alec and Wu, Jeffrey and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya. 2019. *Language Models are Unsupervised Multitask Learners*. OpenAI Blog 1(8), 9.

Li, Xiaonan and Lv, Kai and Yan, Hang and Lin, Tianyang and Zhu, Wei and Ni, Yuan and Xie, Guotong and Wang, Xiaoling and Qiu, Xipeng. 2023. *Unified Demonstration Retriever for In-Context Learning*. ArXiv preprint ArXiv: 2305.04320.

Rubin, Ohad and Herzig, Jonathan and Berant, Jonathan. 2022. *Learning To Retrieve Prompts for In-Context Learning*. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages:2655-2671.

Levy, Itay and Bogin, Ben and Berant, Jonathan. 2023. *Diverse Demonstrations Improve In-context Compositional Generalization*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages:1401-1422.

CCL 2024

CCL24-Eval 任务8系统报告：基于指令微调与数据增强的儿童故事常识推理与寓意理解研究

于博涵，李云龙，刘涛，郑傲泽，张坤丽，咎红英
郑州大学/ 河南省郑州市
{alexyu010120, lylyun, taoliu01, zhengaoze}@gs.zzu.edu.cn
{ieklzhang, iehyzan}@zzu.edu.cn

摘要

尽管现有语言模型在自然语言处理任务上表现出色，但在深层次语义理解和常识推理方面仍有提升空间。本研究通过测试模型在儿童故事常识推理与寓意理解数据集（CRMUS）上的性能，探究如何增强模型在复杂任务中的能力。在本次任务的赛道二中，本研究使用多个7B以内的开源大模型（如Qwen、InternLM等）进行零样本推理，并选择表现最优的模型基于LoRA进行指令微调来提高其表现。除此之外，本研究还对数据集进行了分析与增强。研究结果显示，通过设计有效的指令格式和调整LoRA微调参数，模型在常识推理和寓意理解上的准确率显著提高。最终在本次任务的赛道二中取得第一名的成绩，该任务的评价指标Acc值为74.38，达到了较为先进的水准。

关键词： 儿童故事问答 ； 大语言模型 ； 指令微调 ； 数据增强

System Report for CCL24-Eval Task 8:Research on Commonsense Reasoning and Moral Understanding in Children's Stories Based on Instruction Fine-Tuning and Data Augmentation

Bohan Yu, Yunlong Li, Tao Liu, Aoze Zheng, Kunli Zhang, Hongying Zan
Zhengzhou University / Zhengzhou City, Henan Province
{alexyu010120, lylyun, taoliu01, zhengaoze}@gs.zzu.edu.cn
{ieklzhang, iehyzan}@zzu.edu.cn

Abstract

Despite the impressive performance of existing language models in natural language processing tasks, there remains significant potential for improvement in deep semantic understanding and commonsense reasoning. This study investigates methods to enhance model capabilities in complex tasks by evaluating their performance on the Children's Story Commonsense Reasoning and Moral Understanding Dataset (CRMUS). For Track 2 of this task, we employed several open-source models with fewer than 7 billion parameters (e.g., Qwen, InternLM) for zero-shot reasoning, and selected the best-performing model for instruction fine-tuning using LoRA to enhance its performance. Additionally, we conducted a thorough analysis and enhancement of the dataset. Our findings demonstrate that designing effective instruction formats and adjusting LoRA fine-tuning parameters significantly improves the accuracy of models in commonsense reasoning and moral understanding. Consequently, we achieved first place in Track 2, with an evaluation metric (Acc) score of 74.38, representing a notable advancement.

Keywords: Children's story question answering , Large language model , Instruction Tuning , Data augmentation

第二十三届中国计算语言学大会论文集，第320页-第326页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

当前，自然语言处理领域对儿童故事问答这一新兴任务展现出浓厚兴趣。此任务旨在深化对儿童故事的理解与推理，为教育领域提供高效工具，促进学生理解力与语言表达技能的评估与提升。核心挑战在于深入解析给定故事与问题，检验模型对故事情节与常识的整合理解能力。具体而言，常识推理部分需依据故事情节与隐含常识，从多个选项中甄选最佳答案；寓意理解则聚焦于捕捉故事寓意，选出最贴合情节的选项。

在第二赛道中，本研究运用多种开源大型语言模型，通过开发集上的多轮零样本测试，确定InternLM2(Cai et al., 2024)为指令微调的基础模型。首先，采用固定微调策略，选取最优LoRA(Hu et al., 2021)配置。确定最佳LoRA参数后，通过不同微调模块的组合优化效果。鉴于常识推理数据类型的不平衡分布，本研究实施数据增强策略。初步利用ChatGPT生成超过200条常识推理示例，经人工审查精选137条高质量数据。将这些数据纳入开发集进行微调，显著提升了Acc指标。针对常识推理与寓意理解，各选取一组最佳参数组合，分别实现CR与MU Acc指标72.87与75.38，综合Acc指标达74.38，在本次评测中取得第一名。

2 方法

2.1 指令微调

指令微调有效强化了模型在特定任务上的表现。本研究通过精准提取数据集关键信息，构建提示模板并与数据融合，使模型深化学习儿童故事领域的专业知识，进而提升对儿童故事的语义理解与特征辨识能力。此外，经指令微调后的模型能更准确地遵循指定格式作答，极大地简化了从模型输出中抽取答案的过程。

基于零样本测试的结果，本研究在指令微调环节选用了两款以中文为主的大型语言模型——Qwen1.5-Chat-7B(Bai et al., 2023)与InternLM2-Chat-7B。其中，InternLM2-Chat-7B担任主微调模型，而Qwen1.5-Chat-7B则用于辅助验证最优LoRA参数。通过实验不同LoRA参数与微调模块搭配，最终锁定两组配置，分别对应常识推理（CR）与寓意理解（MU）的最佳Acc指标。

2.2 LoRA

由于参数规模巨大，微调整个大语言模型需要很高的成本。当用于特定任务的训练时，参数高效微调方法只需要微调少量关键参数，就能达到甚至超过全参微调的性能。其中具有代表性的是低秩适配（Low-Rank Adaptation，LoRA）方法，在冻结预训练模型权重的基础上，独立训练一个低秩分解矩阵，然后与预训练模型权重合并，方法如图1所示。

将预训练权重矩阵记为$W_0 \in R^{d\times d}$，低秩矩阵记为$\Delta W = BA$，其中$B \in R^{d\times r}$，$B \in R^{r\times d}$，d和r分别是预训练权重矩阵和低秩矩阵的秩，并且$r \ll d$，矩阵A和B分别通过随机高斯分布和零初始化，包含了可训练的参数（通常来自注意力层）。在推理阶段，使用两部分矩阵融合后的参数，如公式1所示。

$$h = W_0 x + \Delta W x \tag{1}$$

r作为一个超参数，代表了可训练参数量的规模，具体大小需要根据训练数据集大小和特点确定，此外，ΔW进一步通过超参数α缩放，决定低秩矩阵参数影响推理的程度。这种方法极大减少了内存需求，并且训练出的参数具有很强的表征能力，更适合用于特定任务的微调。

3 实验设置

3.1 数据集介绍

本研究所用数据集源于“CCL 2024 Task8儿童故事常识推理与寓意理解评测”。常识推理子任务的问题及答案经人工精心标注，而寓意理解子任务则结合自动构建与人工校验完成。常识推理涵盖社会、生物、时间、空间及物理等多元常识类型，部分题目甚至融合多种类型。数据集细分为开发集与测试集，前者含652条记录，包括400条常识推理与252条寓意理解数据；后者则拥有2768条，分别为1692条常识推理和1056条寓意理解实例。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

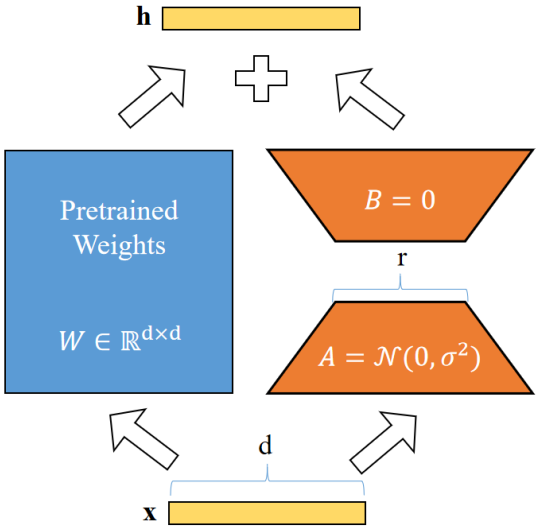


Figure 1: LoRA参数高效微调原理图

具体而言，常识推理子任务的开发集与测试集每条数据包含：标识符（id）、标题（title）、故事内容（story）、问题描述（question）、候选选项（options）、正确答案（answer）及常识类别（type）。值得注意的是，测试集的“answer”字段为空。寓意理解子任务的开发集与测试集未标注常识类型，其余属性与常识推理子任务一致。

3.2 评价指标

对于常识推理与寓意理解两项子任务，统一采用Acc指标衡量性能。参赛模型的最终评价得分由下式计算得出，即所有相关指标的加权平均：

$$Score = 0.4 \times Acc_1 + 0.6 \times Acc_2 \tag{2}$$

此处，Acc_1代表常识推理子任务的回答准确度，而Acc_2则对应寓意理解子任务的准确率。

提示模板样例

系统提示：
请根据\"生物常识、物理常识\"和一个儿童故事来做一道常识推理单项选择题。\n请你一步一步思考并直接给出答案。你将从A, B, C, D中选出正确的答案，并写在【答案】和<eoa>之间。\n完整的题目回答的格式如下：\n【答案】 ... <eoa>\n请你严格按照上述格式作答。\n儿童故事和题目如下：

用户输入：
故事:一只公鸡在田野里为自己和母鸡们寻找食物。他发现了一块宝玉，便对宝玉说：\"若不是我，而是你的主人找到了你，他会非常珍惜地把你捡起来；但我发现了你却毫无用处。我与其得到世界上一切宝玉，倒不如得到一颗麦子好。\"\n题目:关于公鸡对宝玉的看法，下列选项描述正确的是？\nA. 宝玉太硬了，不好吃 B. 主人非常喜欢吃宝玉 C. 宝玉不是食物，但自己可以拿去卖钱 D. 宝玉不是食物，不能吃

期望输出：
【答案】 D <eoa>

Figure 2: 指令数据示例

(c) 2024 中国中文信息学会计算语言学专业委员会

3.3 数据预处理

鉴于原始数据集包含若干非必要字段，如id、title、domain等，本研究对寓意理解（MU）数据集进行了精简，仅保留story、question、options与answer字段。相比之下，常识推理（CR）任务因涉及特定类型常识，故在保留前述字段的基础上，额外保存type字段，确保模型能依据常识类别提供更为精确的回应。

提示模板的设计借鉴了GAOKAO-Bench(Zhang et al., 2023)的研究，最终定制的指令模板详情见图2。

3.4 数据增强

鉴于开发集内常识推理任务各类常识分布的不均衡性，本研究引入数据增强技术予以应对。表1展示了各常识类别的样本量，显而易见，空间与物理常识远少于社会及生物常识。模型在开发集上的推理结果显示，其在时间、空间与物理常识的表现欠佳，推测原因可能与这些类型数据稀缺有关。为此，本研究首先借助ChatGPT生成逾200条常识推理示例，再经人工严格筛选，最终精选137条优质数据。增强后的各类常识数量详情参见表2。

常识类型	社会常识	生物常识	时间常识	空间常识	物理常识
常识数量	196	90	67	37	37

Table 1: 开发集上各常识类型的数量

常识类型	社会常识	生物常识	时间常识	空间常识	物理常识
常识数量	212	138	110	51	67

Table 2: 扩充开发集后各常识类型的数量

4 实验流程

4.1 实验参数设置

本研究的实验参数配置详列于表3。所有实验采用PyTorch深度学习框架执行，微调工作依托于LLaMA Factory(Zheng et al., 2024)框架，并在配备一张4090与一张A40的硬件环境下运行。

模型参数	参数值
训练轮数	5
学习率	5e-5
截断长度	1536
Batch size	1
Optimizer	AdamW
Warmup ratio	0.1
Lr scheduler	Cosine
Gradient accumulation steps	8

Table 3: 实验参数

4.2 模型选择

为确保在性能优越的模型基础上开展微调，本研究挑选了Baichuan2-Chat-7B(Yang et al., 2023)、Qwen-Chat-7B、Qwen1.5-Chat-7B、Yi-Chat-6B(Young et al., 2024)与InternLM2-Chat-7B，在原始开发集（含400条常识推理[CR]数据与252条寓意理解[MU]数据）上执行零样

(c) 2024 中国中文信息学会计算语言学专业委员会

本推理。表4呈现的实验成果表明，所有测试模型在CR与MU任务上的Acc指标均超越主办方提供的基准线。综合考量各模型的Acc得分后，最终选定InternLM2-Chat-7B作为微调工作的基准模型。

模型	CR	MU	Overall
Baichuan2-Chat-7B	43.5	42.8	43.1
Qwen-Chat-7B	45.5	40.4	42.4
Qwen1.5-Chat-7B	62.5	38.0	47.8
Yi-Chat-6B	50.3	44.4	46.8
InternLM2-Chat-7B	65.7	52.7	57.9
Baseline	31.2	33.2	32.4

Table 4: 各模型在开发集上零样本推理的表现

4.3 LoRA参数选择

LoRA的超参数设定对实验成效具有显著影响。本研究参照LoRA论文建议，于原始开发集上实施多组超参数微调，并在测试集上评估结果，详情见表5。其中，Wqkv构成基本微调模块。实验显示，当Rank设为64、Alpha为16时，模型表现欠佳；调整至LoRA论文推荐的Rank：Alpha比例1：2后，模型性能显著提升。尤其当Rank等于256、Alpha设定为512时，模型效能达到顶峰。

除在InternLM-Chat-7B上探索最优参数外，本研究亦采用Qwen1.5-Chat-7B在原始开发集上执行多轮微调试验，测试集上的结果列于表6。综合考量InternLM-Chat-7B与Qwen1.5-Chat-7B的实测表现，最终决定采用Rank=256、Alpha=512的配置。

Rank	Alpha	Target	CR	MU	Overall
64	16	Wqkv	65.90	52.75	58.01
256	512	Wqkv	70.39	71.21	70.88
512	512	Wqkv	69.73	71.40	70.73
512	1024	Wqkv	69.56	70.45	70.09

Table 5: InternLM-Chat-7B的不同LoRA超参数设置

Rank	Alpha	Target	CR	MU	Overall
32	64	Wqkv	34.86	67.04	54.17
64	128	Wqkv	64.42	68.46	66.84
128	256	Wqkv	65.95	69.31	67.97
256	512	Wqkv	65.36	69.6	67.9

Table 6: Qwen1.5-Chat-7B的不同LoRA超参数设置

4.4 LoRA微调模块选择

依据4.3章节中选定的最优LoRA超参数，本研究进一步探索不同模块组合的效果。InternLM2-Chat-7B模型可供微调的组件涵盖Wqkv、W1、W2、W3、Wo，其中Wqkv被视为基础微调模块，其余则为可选附加模块。将Wqkv与任一额外模块搭配进行微调实验，所得测试集上的结果详载于表7。

审视表7可发现，Wqkv与W1的组合展现出最优效用。鉴于模块叠加可增扩微调参数规模，从而增强模型效能，本研究特增设一组实验，测试Wqkv联合W1、W2的效果。相应结果收录于表8。

(c) 2024 中国中文信息学会计算语言学专业委员会

Rank	Alpha	Target	CR	MU	Overall
256	512	Wqkv	70.39	71.21	70.88
256	512	Wqkv,W1	72.1	71.78	71.91
256	512	Wqkv,W2	71.21	72.15	71.77
256	512	Wqkv,W3	70.68	72.34	71.68
256	512	Wqkv,Wo	70.98	72.15	71.68

Table 7: 在原始开发集上加入单一模块的微调

Rank	Alpha	Target	CR	MU	Overall
256	512	Wqkv,W1,W2	72.87	71.78	72.22

Table 8: 在原始开发集上加入W1、W2模块

模型在整合W2后，展现出性能的显著提升，这一进展激励了对更多模块组合的探索。本研究在扩充后的数据集上进行了多轮精细的微调实验。根据表9中的实验结果，数据增强虽使CR任务的表现略有下降，却提升了MU任务的性能。分析认为，CR任务涉及的推理过程较为繁复，现有增强数据的复杂程度可能不足以完全匹配其需求；相反，MU任务侧重于直接从文本中提取答案，因此，增强的CR信息有效地深化了MU情境下文本的信息层次。

Rank	Alpha	Target	CR	MU	Overall
256	512	Wqkv,W1,W2	69.9	74.14	72.44
256	512	Wqkv,W2,W3	70.8	74.9	73.26
256	512	Wqkv,W1,W3	70.56	75.38	73.45
256	512	Wqkv,W2,Wo	71.57	72.91	72.37
256	512	Wqkv,W1,W2,W3	71.04	72.25	71.76
256	512	Wqkv,W1,W2,Wo	71.74	72.15	71.99
256	512	Wqkv,W2,W3,Wo	67.96	71.02	69.8
256	512	Wqkv,W1,W2,W3,Wo	71.74	71.59	71.65

Table 9: 在扩充后的开发集上微调更多模块

通过对不同模块进行实验对比，观察到模型性能并未随微调参数数量的增加而持续上升。令人意外的是，当所有可微调模块均被纳入时，模型表现出现下滑。基于此，本研究最终采纳了表8中记录的最优CR结果，以及表9中所列的最佳MU结果，具体详情参见表10。

CR	MU	Overall
72.87	75.38	74.38

Table 10: 最佳分数组合

5 结论

本研究深入分析并评估了语言模型在处理复杂语义理解，特别是常识推理与寓意理解任务时的能力提升途径。在第二赛道的竞赛中，我们选择了Qwen1.5-Chat-7B与InternLM2-Chat-7B等模型作为研究对象。通过参数优化与数据增强技术的应用，发现InternLM2-Chat-7B在上述两项任务中展现出最优性能。实验成果证实，精心设计的提示模板结合LoRA微调参数调整及数据增强策略，能显著提高模型的常识推理精度和寓意理解水平，有力地促进了语言模型在

复杂语义理解和常识推理领域的发展。这一研究为后续在自然语言理解领域，尤其是儿童故事分析方向，奠定了坚实的基础，提供了宝贵的指导思路。

参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

CCL 2024

System Report for CCL24-Eval Task 8: Exploring Faithful and Informative Commonsense Reasoning and Moral Understanding in Children's Stories

Zimu Wang[1,3], Yuqi Wang[1,3], Nijia Han[1], Qi Chen[2], Haiyang Zhang[1], Yushan Pan[1], Qiufeng Wang[1], Wei Wang[1†]

[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
[2]School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University
[3]Department of Computer Science, University of Liverpool
{Zimu.Wang19,Yuqi.Wang17,Nijia.Han23}@student.xjtlu.edu.cn
{Qi.Chen02,Haiyang.Zhang,Yushan.Pan}@xjtlu.edu.cn
{Qiufeng.Wang,Wei.Wang03}@xjtlu.edu.cn

Abstract

Commonsense reasoning and moral understanding are crucial tasks in artificial intelligence (AI) and natural language processing (NLP). However, existing research often falls short in terms of faithfulness and informativeness during the reasoning process. We propose a novel framework for performing commonsense reasoning and moral understanding using large language models (LLMs), involving constructing guided prompts by incorporating relevant knowledge for commonsense reasoning and extracting facts from stories for moral understanding. We conduct extensive experiments on the Commonsense Reasoning and Moral Understanding in Children's Stories (CRMUS) dataset with widely recognised LLMs under both zero-shot and fine-tuning settings, demonstrating the effectiveness of our proposed method. Furthermore, we analyse the adaptability of different LLMs in extracting facts for moral understanding performance.

1 Introduction

Proficiency in acquiring reasoning abilities, such as arithmetic, commonsense, and symbolic reasoning, plays an essential role in artificial intelligence (AI) and natural language processing (NLP) (Wang et al., 2023). Unlike arithmetic reasoning, which involves manipulating numbers, and symbolic reasoning, which involves interpreting logic and symbols, commonsense reasoning encompasses counterfactual, abductive, and monotonic reasoning (Ashida and Sugawara, 2022). It is crucial for language understanding and enables humans to navigate daily situations seamlessly (Sap et al., 2020). Applications of commonsense reasoning include text classification (Wang et al., 2019), question answering (Mihaylov and Frank, 2018), and natural language generation (Chen et al., 2019).

Commonsense reasoning is typically framed as a multiple-choice format, where the goal is to determine the plausibility of candidate answers. This approach mirrors how people often consider several plausible choices based on a given situation and their thought processes (Figure 1) (Ashida and Sugawara, 2022). Previous research has focused primarily on utilising pre-trained language models (PLMs) and conducting the reasoning process based on factual time and space information (Talmor et al., 2019), human behaviours (Zhang and Choi, 2021; Emelin et al., 2021), and story texts (Ashida and Sugawara, 2022). Recently, with the development of large language models (LLMs) that have shown remarkable performance in a range of natural language understanding and reasoning tasks (Peng et al., 2023; Na et al., 2024), they have also been leveraged to enhance the reasoning process (Wang and Zhao, 2023; Bian et al., 2024; Krause and Stolzenburg, 2024). Similar to commonsense reasoning, moral understanding is the process of comprehending the moral of the given context from multiple candidates.

Despite the progress made, existing research in commonsense reasoning and moral understanding faces significant challenges, particularly regarding the *unfaithfulness* and *uninformativeness* of the reasoning process. Current LLM-based methods primarily follow the in-context learning (ICL) paradigm

†Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 327-335, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

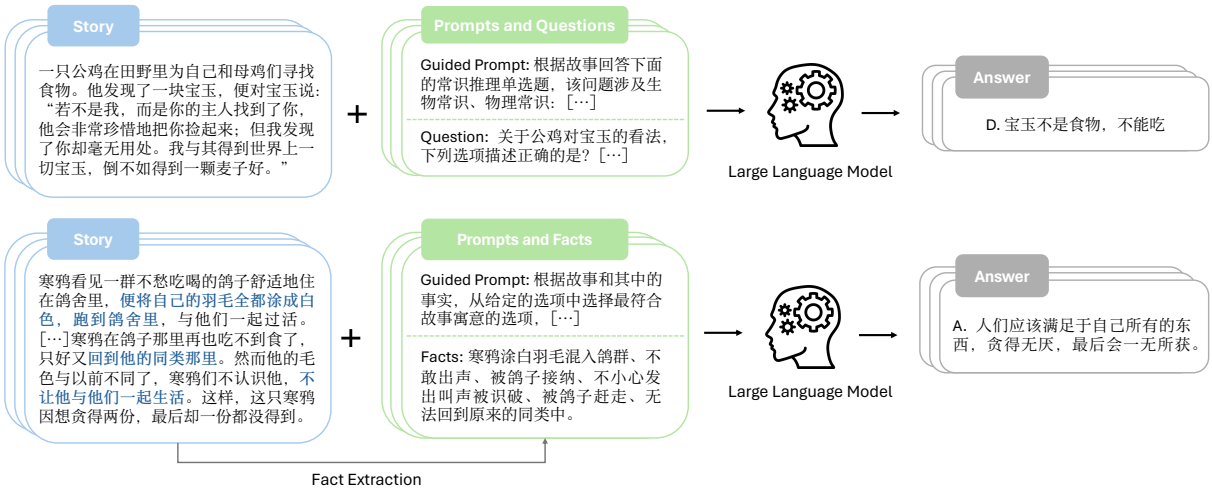


Figure 1: Overall framework of our proposed method to conduct *faithful* and *informative* commonsense reasoning and moral understanding.

(Brown et al., 2020), which conditions the models on a natural language instruction and/or a few demonstrations (Qiao et al., 2023; Wang and Zhao, 2023). However, commonsense reasoning usually involves various types of knowledge applied in different stories and questions, such as temporal, spatial, biological, physical, and social knowledge; and the facts described in the stories, such as their plots, characters, and events, are highly related to the moral that the authors intend to impart. Though effective, the aforementioned information is usually disregarded in previous research.

Motivated by this phenomenon, we design a novel framework to conduct commonsense reasoning and moral understanding, making the reasoning process more *faithful* and *informative*. Unlike the previous work that utilises external knowledge, such as knowledge bases (Mitra et al., 2020; Bian et al., 2021), search engines (Talmor et al., 2021), and the knowledge generated by LLMs (Liu et al., 2022), we construct guided prompts for the two tasks, as shown in Figure 1. For commonsense reasoning, we incorporate the related knowledge concerning the story and the question into the prompt, and for moral understanding, we extract the facts contained in the stories as additional supervision.

We conduct extensive experiments on the Commonsense Reasoning and Moral Understanding in Children's Stories (CRMUS) dataset with the widely recognised LLMs: GLM-3, GLM-4, Moonshot, and Yi-34B for zero-shot prompting and ChatGLM3-6B (Zeng et al., 2023), InternLM2-7B (Cai et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Yi-6B (01.AI et al., 2024) for fine-tuning. Experimental results demonstrate the effectiveness of guided prompts for both commonsense reasoning and moral understanding. Among the models, GLM-4 and InternLM2-7B achieve the best performance in zero-shot and fine-tuning settings, respectively. Furthermore, we conduct additional experiments to analyse the adaptability of extracted facts from different LLMs for moral understanding performance.

The key contributions of this work are summarised as follows:

- We propose a novel framework for commonsense reasoning and moral understanding using LLMs, making the process becomes *faithful* and *informative*.
- We perform extensive experiments on widely recognised LLMs to demonstrate the effectiveness of the proposed method.
- We conduct additional experiments on moral understanding and analyse the adaptability of extracted facts on different LLMs on this task.

2 Background

Commonsense reasoning has received considerable attention over the past decade. Recent research highlights the substantial improvements in this area by incorporating additional knowledge, broadly falling

End-to-End Prompt for Commonsense Reasoning:
根据故事回答下面的单项选择题，只给出答案即可： [Translation: Please answer the following multiple-choice question based on the story, providing only the answer:] 故事[Story]：{story_text} 问题[Question]：{question} 选项[Options]：{options} 答案[Answer]：
End-to-End Prompt for Moral Understanding:
根据故事从给定选项中选择最符合故事说明的寓意的选项，只给出答案即可： [Translation: Please select the option that best matches the moral from the given choices based on the story, providing only the answer:] 故事[Story]：{story_text} 选项[Options]：{options} 答案[Answer]：

Table 1: End-to-end prompts for the commonsense reasoning and moral understanding tasks, in which {story_text}, {question}, and {options} refer to the story context, the question, and the candidate options, respectively.

into two categories. The first approach involves augmenting the task with external knowledge graphs, such as ConceptNet (Speer et al., 2017) and FreeBase (Bollacker et al., 2008). Noteworthy methods like KAGNet (Lin et al., 2019) and GRF (Ji et al., 2020) operate by reasoning over the links connecting different entities and relationships within the knowledge graphs. However, it is worth mentioning that commonsense knowledge includes a wide range of facts and scenarios that exceed the capacity of a single knowledge graph with a specific schema (Yu et al., 2022).

The second approach focuses on leveraging the internal knowledge of LLMs, which are trained on massive datasets to generate task-specific knowledge. For instance, Zhou et al. (2021) employs self-talk procedures (Shwartz et al., 2020) and inquiry-based discovery learning to generate implicit commonsense before response generation. Similarly, Qin et al. (2020) and Zhao et al. (2023) generate plausible explanations for commonsense reasoning by incorporating future context in decoding algorithms and using posterior regularisation for constraint enforcement. Additionally, Paranjape et al. (2021) prompts GPT2-XL (Radford et al., 2019) for inference using generated contrastive explanations. Furthermore, studies like those discussed by Liu et al. (2022) and Cao and Jiang (2024) emphasise the improvements in commonsense reasoning, even in zero-shot scenarios, through the incorporation of LLM-generated knowledge. Unlike the previous work, we construct guided prompts with knowledge highly related to the stories, which are more faithful to the story contexts and have higher generalisability.

3 System Overview

Following the overall framework illustrated in Figure 1, in this section, we describe the design of the system to conduct *faithful* and *informative* commonsense reasoning and moral understanding in detail.

3.1 Problem Definition

We define our commonsense reasoning and moral understanding tasks as follows. Given a story context $S = \{w_1, w_2, \ldots, w_M\}$ (M is the number of words within the story), a question q, and a list of candidate answers $A = \{a_1, a_2, \ldots, a_N\}$ (N is the number of candidate answers), the aim of the tasks is to select the answer $a^* \in A$ that matches the question q with respect to the story S most. To make the reasoning process *faithful* and *informative*, we add the related knowledge K, a subset of a pre-defined list $\mathcal{K}$, containing the knowledge related to the story and the question (e.g. temporal, spatial, and social knowledge) for commonsense reasoning and the facts that happened in the stories $F = \{f_1, f_2, \ldots, f_P\}$ (P is the number of facts in the story) for moral understanding into the prompt, in which the list of facts F is extracted by an LLM, which could be GLM-4, Moonshot, and Yi-34B.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 327-335, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

PROMPT FOR FACT EXTRACTION:
根据故事内容，抽取故事中与寓意有关的事实，只给出答案即可，以顿号分隔： [Translation: Based on the story content, extract the facts relevant to the moral of the story, providing only the answers separated by serial commas:] 故事[Story]：{story_text} 答案[Answer]：

Table 2: Fact extraction prompt for the moral understanding task, in which {story_text} refers to the story context.

GUIDED PROMPT FOR COMMONSENSE REASONING:
根据故事回答下面的常识推理单项选择题，只给出答案即可，该问题涉及{reasoning_type}： [Translation: Answer the following commonsense reasoning multiple-choice question based on the story, providing only the answer. The question involves {reasoning_type}:] 故事[Story]：{story_text} 问题[Question]：{question} 选项[Options]：{options} 答案[Answer]：
GUIDED PROMPT FOR MORAL UNDERSTANDING:
根据故事和其中的事实从给定选项中选择最符合故事说明的寓意的选项，只给出答案即可： [Translation: Based on the story and its facts, select the option that best matches the moral of the story from the given choices, providing only the answer:] 故事[Story]：{story_text} 事实[Facts]：{extracted_facts} 选项[Options]：{options} 答案[Answer]：

Table 3: Guided prompts for the commonsense reasoning and moral understanding tasks, in which {story_text}, {question}, {reasoning_type}, {extracted_facts}, and {options} refer to the story context, the question, the related knowledge, the facts extracted by the LLMs, and the candidate options, respectively.

3.2 End-to-End Prompt Construction

We first construct a prompt to conduct end-to-end commonsense reasoning and moral understanding and consider it as our baseline, as shown in Table 1. It includes an instruction for the target task, a story context, a question, and four candidate options. Because all the questions for moral understanding are the same (i.e. "*Which of the following options best matches the moral of the story?*"), we incorporate the requirement of performing moral understanding into the instruction for that task. The prompt ends with the word "*Answer:*", which asks the language models to answer the question.

3.3 Guided Prompts Construction

We design separate guided prompts for commonsense reasoning and moral understanding with respect to the task characteristics, which are explained as follows:

Commonsense Reasoning Commonsense reasoning usually includes multiple types of knowledge, such as temporal, spatial, biological, physical, and social knowledge, to support the reasoning process. Therefore, we incorporate the relevant knowledge into the prompt, which is obtained from the golden annotations, and we ask the model to consider this information when making accurate predictions, as shown in Table 3.

Moral Understanding For moral understanding, we consider the facts in the stories—such as their plots, characters, and events—as additional supervision, since these elements are often connected to the moral that the author conveys. By analysing the essential facts within the story, LLMs can gain a deeper understanding of the message the author intends to impart. The construction of guided prompts for moral understanding includes two procedures: fact extraction and guided prompt construction.

CCL 2024

Model	End-to-End Prompt		Guided Prompt	
	Accuracy (CR)	Accuracy (MU)	Accuracy (CR)	Accuracy (MU)
GLM-3	70.45	56.91	72.10 (↑ 1.65)	56.91 (↑ 0.00)
GLM-4	**83.92**	**65.91**	**84.28** (↑ 0.36)	**66.38** (↑ 0.47)
Moonshot	<u>77.01</u>	<u>65.44</u>	<u>78.66</u> (↑ 1.65)	<u>65.06</u> (↓ 0.38)
Yi-34B	74.65	59.38	74.53 (↑ 0.12)	59.09 (↓ 0.29)

Table 4: Performance of LLMs on commonsense reasoning (CR) and moral understanding (MU) under the *zero-shot* setting, in which the best and the second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

Model	End-to-End Prompt		Guided Prompt	
	Accuracy (CR)	Accuracy (MU)	Accuracy (CR)	Accuracy (MU)
ChatGLM3-6B	53.72	61.08	53.49 (↓ 0.23)	62.03 (↑ 0.95)
InternLM2-7B	**70.63**	**71.50**	**70.80** (↑ 0.17)	**70.83** (↓ 0.67)
Qwen1.5-7B	66.96	<u>69.60</u>	66.96 (↑ 0.00)	<u>69.60</u> (↑ 0.00)
Yi-6B	<u>67.85</u>	65.25	<u>67.02</u> (↓ 0.83)	63.16 (↓ 2.09)

Table 5: Performance of LLMs on commonsense reasoning (CR) and moral understanding (MU) under the *fine-tuning* setting, in which the best and the second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

Initially, we extract the factual details from the stories by prompting LLMs, the prompt for which is shown in Table 2. Similar to the prompt for reasoning, the prompt for fact extraction includes an instruction and a story, and it ends with the word "*Answer:*" to ask the models to answer the question. Once the facts are extracted, they are incorporated into the end-to-end prompt as additional supervision, and the guided prompt is used to ask the models to conduct the moral understanding process, as depicted in Table 3.

4 Experiments

4.1 Experimental Setup

We conducted experiments on widely recognised LLMs under both zero-shot prompting and fine-tuning settings. Under the zero-shot setting, we experimented on GLM-3[0] (`glm-3-turbo`), GLM-4 (`glm-4`), Moonshot[1] (`moonshot-v1-8k`), and Yi-34B[2] (`yi-34b`), which were accessed by calling their official APIs. During the reasoning process, we set the temperature as 0 to stabilise the output of the models. We also fine-tuned four LLMs, including ChatGLM3-6B (Zeng et al., 2023), InternLM2-6B (Cai et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Yi-6B (01.AI et al., 2024), which were accessed from the Hugging Face[3] repository. During the fine-tuning process, we set the number of epochs as 20, the learning rate as $5e-5$, the batch size as 2, and the gradient accumulation steps as 8, and we adopted LLaMA-Factory (Zheng et al., 2024) for efficient fine-tuning with the LoRA strategy (Hu et al., 2022). All experiments were conducted on a single NVIDIA A10 Tensor Core GPU.

4.2 Experimental Results

Main Results Tables 4 and 5 present the main experimental results of commonsense reasoning and moral understanding under zero-shot and fine-tuning settings, in which we utilised GLM-4 as the model

[0] https://open.bigmodel.cn/
[1] https://platform.moonshot.cn/
[2] https://platform.lingyiwanwu.com/
[3] https://huggingface.co/

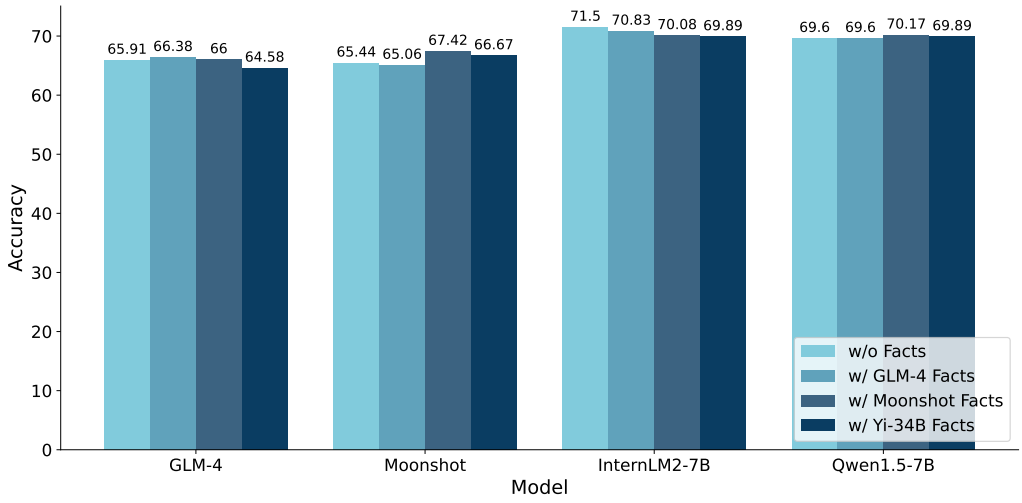


Figure 2: Effects of facts extracted from different LLMs in the moral understanding task.

to extract the facts that occur in the stories. Among all experiments, GLM-4 and IntenLM2-7B achieved the best performance under zero-shot and fine-tuning settings, regardless of the utilisation of the guided prompts. After fine-tuning the smaller-sized LLMs, such as InternLM-7B and Qwen1.5-7B, the models could perform comparable to or even better performance than the larger-scale LLMs, such as GLM-4 and Moonshot, in moral understanding; however, there was still room for improvement in terms of commonsense reasoning.

We also observed the effectiveness of guided prompts in commonsense reasoning and moral understanding. Regarding commonsense reasoning, the use of guided prompts led to performance improvements across nearly all models, indicating that incorporating knowledge successfully enhanced the commonsense reasoning process. However, for moral understanding, guided prompts proved beneficial specifically for GLM-4; thus, further investigations are needed to assess the generalisability of the method in moral understanding.

Effects of the Extracted Facts The main experimental results of guided prompts for moral understanding were remarkable—typically, the facts extracted from the stories should closely align with the moral intended by the authors. It was observed that these facts, extracted by GLM-4, significantly benefited GLM models. This correlation underscores the importance of aligning the models utilised for fact extraction with those used for moral understanding prediction. To further investigate the relationship between fact extraction and moral understanding, we conducted additional experiments using GLM-4, Moonshot, and Yi-34B for fact extraction, and subsequently employing GLM-4, Moonshot, InternLM2-7B, and Qwen1.5-7B for predicting moral understanding.

We presented the experimental results in Figure 2, which substantiated our earlier hypothesis. Generally, using the same LLM for both fact extraction and moral understanding prediction (e.g. GLM-4) contributed significantly to model performance. Interestingly, despite Moonshot initially performed worse than GLM-4 in our main experiments, its performance improved notably when employed it for fact extraction from the stories. This underscores the efficacy of guided prompts in enhancing moral understanding. The impact of extracted facts varied for InternLM2-7B and Qwen1.5-7B, highlighting how different LLMs affect moral understanding performance based on the extracted facts.

5 Conclusion and Future Work

We introduced a novel framework for commonsense reasoning and moral understanding with LLMs, aiming to ensure a *faithful* and *informative* reasoning process. Specifically, we developed guided prompts that integrate relevant knowledge for commonsense reasoning and the facts that happened in the stories extracted by LLMs for moral understanding. We conducted extensive experiments on the CRMUS

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 327-335, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

dataset with widely recognised LLMs under both zero-shot and fine-tuning settings and demonstrated the effectiveness of our proposed method. We further analysed the adaptability of extracted facts of different LLMs on moral understanding. In the future, we will make the guided prompts more diverse, incorporating more useful features to guide the reasoning process. We will also transfer our method on more reasoning tasks to test the generalisability of our proposed method.

Acknowledgements

This research is funded by the Postgraduate Research Scholarship (PGRS) at Xi'an Jiaotong-Liverpool University, contract number FOSA2212008, and partially supported by 2022 Jiangsu Science and Technology Programme (General Programme), contract number BK20221260.

References

01.AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Mana Ashida and Saku Sugawara. 2022. Possible stories: Evaluating situated commonsense reasoning under multiple possible scenarios. In *Proceedings of COLING*, pages 3606–3630, October.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. *Proceedings of AAAI*, 35(14):12574–12582, May.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD*, page 1247–1250.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, volume 33, pages 1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Rui Cao and Jing Jiang. 2024. Knowledge generation for zero-shot knowledge-based VQA. In Yvette Graham and Matthew Purver, editors, *Findings of EACL*, pages 533–549, March.

Jiaao Chen, Jianshu Chen, and Zhou Yu. 2019. Incorporating structured commonsense knowledge in story completion. *Proceedings of AAAI*, 33(01):6244–6251, Jul.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of EMNLP*, pages 698–718, November.

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

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of EMNLP*, pages 725–736, November.

Stefanie Krause and Frieder Stolzenburg. 2024. Commonsense reasoning and explainable artificial intelligence using large language models. In *Artificial Intelligence. ECAI 2023 International Workshops*, pages 302–319.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of EMNLP-IJCNLP*, pages 2829–2839, November.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of ACL*, pages 3154–3169, May.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of ACL*, pages 821–832, July.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2020. How additional knowledge can improve natural language commonsense question answering?

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of ACL-IJCNLP*, pages 4179–4192, August.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of ACL*, pages 5368–5393, July.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of EMNLP*, pages 794–805, November.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of ACL (Tutorial)*, pages 27–33, July.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of EMNLP*, pages 4615–4629, November.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of AAAI*, 31(1), Feb.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158, June.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Proceedings of the NeurIPS (Datasets and Benchmarks)*, volume 1.

CCL 2024

Yuqing Wang and Yun Zhao. 2023. Gemini in reasoning: Unveiling commonsense in multimodal large language models.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. *Proceedings of AAAI*, 33(01):7208–7215, Jul.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, Suparna De, and Amir Hussain. 2023. Fusing external knowledge resources for natural language understanding techniques: A survey. *Information Fusion*, 92:190–204.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of EMNLP*, pages 4364–4377, December.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *Proceedings of ICLR*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of EMNLP*, pages 7371–7387, November.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of ACL*, pages 14883–14896, July.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models.

Pei Zhou, Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Think before you speak: Learning to generate implicit knowledge for response generation by self-talk. In *Proceedings of NLP4ConvAI*, pages 251–253, November.

CCL 2024

CCL24-Eval 任务8系统报告：基于提示工程和思维链的提示词构造

罗允[*]
北京交通大学
计算机科学与技术学院
23120391@bjtu.edu.cn

冯毅[*]
北京交通大学
计算机科学与技术学院
21112027@bjtu.edu.cn

景丽萍[†]
北京交通大学
计算机科学与技术学院
lpjing@bjtu.edu.cn

摘要

儿童故事常识推理与寓意理解评测任务旨在从常识推理和寓意理解两个任务多角度评价中文预训练语言模型和大型语言模型的常识推理和故事理解能力，这考察了模型的常识储备能力以及对文本内容的深入理解能力，因此极具挑战性。随着大语言模型的发展，其卓越的指令跟随能力显著提升了自然语言处理任务的效率和效果。然而，这也对提示词的设计提出了更高的要求，因为提示词的质量直接影响了大模型的表现和预测结果的准确性。因此，设计有效的提示词变得尤为重要，不仅需要理解任务的具体需求，还要具备对语言模型的深入认识和灵活运用能力。本文针对儿童故事常识推理与寓意理解评测赛道一的两个任务，提出了一种基于提示工程的提示词构造方法。首先，我们提出了一种基于融合提示工程、思维链的通用提示词构建框架；然后，我们针对具体的任务调整对应的提示词模板；最后，结合语言模型使用这些提示词进行结果预测。在本次评测中，我们的方法在赛道一的封闭数据条件下获得了第三名的成绩，这验证了我们方法的有效性，并展示了其在自然语言理解领域的应用潜力。

关键词： 提示工程；思维链；少样本学习；上下文学习

System Report for CCL24-Eval Task 8: Prompt Construction Based on Prompt Engineering and Chain-of-Thought

Yun Luo[*]
Beijing Jiaotong University
School of Computer Science and Technology
23120391@bjtu.edu.cn

Yi Feng[*]
Beijing Jiaotong University
School of Computer Science and Technology
21112027@bjtu.edu.cn

Liping Jing[†]
Beijing Jiaotong University
School of Computer Science and Technology
lpjing@bjtu.edu.cn

Abstract

Evaluation on Commonsense Reasoning and Moral Understanding in Children's Stories (CRMU), aims to evaluate the commonsense reasoning and story comprehension abilities of Chinese pre-trained language models and large language models from multiple perspectives, focusing on both Commonsense Reasoning (CR) and Moral Understanding (MU). This task tests the models' ability to retain commonsense knowledge and their deep understanding of textual content, making it extremely challenging. With the development of large language models, their excellent instruction-following capabilities have significantly improved the efficiency and effectiveness of natural language

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
[*]相同贡献
[†]通讯作者

processing tasks. However, this also raises the bar for prompt design, as the quality of the prompts directly affects the model's performance and the accuracy of the predictions. Therefore, designing effective prompts has become crucial, requiring not only an understanding of the specific task requirements but also an in-depth knowledge of and flexibility in using language models. In this paper, we propose a prompt construction method based on prompt engineering for the two tasks in Track 1 of the Children's Story Commonsense Reasoning and Implication Understanding Evaluation. First, we introduce a general prompt construction framework that integrates prompt engineering and chain of thought reasoning. Then, we adjust the corresponding prompt templates for each specific task. Finally, we use these prompts in conjunction with the language model to generate prediction results. In this evaluation, our method achieved third place under the closed data conditions of Track 1, demonstrating the effectiveness of our approach and its potential applications in the field of natural language understanding.

Keywords: Prompt Engineering , Chain of Thought , Few-shot Learning , In-context Learning

1 引言

近年来，随着预训练模型的兴起及比例定律 (Kaplan et al., 2020; Bahri et al., 2021)的提出，语言模型的规模界限不断被刷新。xAI公司推出的Grok-1模型，以其庞大的参数量，成为了迄今为止最大的开源语言模型。尽管这些语言模型在特定任务上的表现已经接近甚至超越了人类水平，但它们在处理需要推理能力和深层语义理解的文本任务时，仍显示出一定的局限性。

常识推理和寓意理解是自然语言处理领域的两大核心研究方向。常识推理聚焦于让模型具有人类一样的理解和运用广泛的常识性知识的能力，这些常识通常涵盖生物、物理、时间概念以及社会习俗等诸多维度(Davis and Marcus, 2015; Liu and Singh, 2004; Talmor et al., 2018)。该领域的目标在于使模型能够结合常识性知识，更好地模拟人类进行分析与推理。寓意理解的核心在于使模型能够敏锐地识别并理解隐藏在文本表层含义之下的深层语义。这意味着模型需要具备与人类相近的文本理解能力，能够捕捉到文本中的隐喻、讽刺等复杂语言现象 (Tanasescu et al., 2018; del Pilar Salas-Zárate et al., 2017)。因此，常识推理和寓意理解的研究不仅对于推动语言模型的理解能力至关重要，而且对于提升模型在复杂语言环境中的表现和适应性具有深远的影响。

当前已经有工作通过使用预训练模型来完成这些任务，比如Liu等人 (Liu et al., 2021)提出可以先从语言模型中生成与问题相关的常识知识，然后在回答问题时将这些知识作为额外的输入。Ling等人 (Ling et al., 2023)受到人类解释中的对比性质的启发，使用语言模型完成解释提示，这些提示根据证明正确答案所需的关键属性来对比不同的选项。通过将这些解释作为条件来指导模型的决策。这些研究采用了少样本学习的策略，通过为模型提供少量的常识性知识和学习样例，使模型能够学习到相关的知识和回答的格式。然而，在针对当前评测任务的应用中，这种方法展现出了一定的局限性。具体而言，由于本次评测任务中的每条数据均包含较长的故事文本、故事相关的问题以及多个选项，与常规问答任务中的数据相比，文本长度显著增长。当面临过长文本时，这些方法可能会遇到模型输入长度的限制问题，或者因文本信息过于复杂而导致信息遗忘，从而影响模型的性能。

因此，针对这些问题，本研究提出了一种结合提示工程与思维链技术的构建提示词策略，旨在利用大模型进行常识推理和寓意理解任务。为了充分激发大规模模型的潜在能力，本文建议采用提示工程的方法构建提示词框架，该框架涵盖任务描述、模型在任务中的角色定位、执行任务时需要注意的事项，以及评价结果的标准。在本研究中，我们采取了一种综合方法，将少样本学习技术 (Brown et al., 2020; Wang et al., 2020)与思维链技术相结合。为了应对过长输入内容可能导致的模型对提示词信息的遗忘问题，我们在少样本学习部分没有选择堆砌案例——而是为每个任务只提供一个具有代表性的案例。通过向模型展示问题的正确答案，并引

(c) 2024 中国中文信息学会计算语言学专业委员会

导模型对问题的各个选项进行深入的分析，以此作为提示词中的学习案例。最终，按照提示工程的要求，设计提示词的格式并利用分隔符进行引导，以促使模型生成规范且精确的答案。这种方法不仅提高了答案的准确率，保证了答案的格式规范性，也便于后续的处理与分析工作。

本文针对儿童故事常识推理与寓意理解评测赛道一的两个任务，提出了一种基于提示工程的提示词构造方法。鉴于两个任务在模型能力需求上的差异——常识推理任务侧重于模型结合常识性知识进行逻辑推断，而寓意理解任务则强调模型通过语言表层深入挖掘深层的语义内涵——我们率先构建了一个融合提示工程与思维链的通用提示词框架。随后，针对两个任务的具体特点，我们对提示词模板进行了精细化的调整，以确保模型能够运用相应的能力来完成任务。最后结合语言模型使用这些提示词进行结果预测。在本次评测中，我们的方法在赛道一的封闭数据条件下取得了第三名的成绩，充分验证了本研究所提出方法的有效性。

2 相关工作

2.1 常识推理

常识推理研究侧重于如何更好地运用外部知识来进行推理。之前的工作大多聚焦于如何更好地使用常识知识构建知识图谱，并使用神经网络结合知识图谱完成推理。Moghimifar等人 (Moghimifar et al., 2020)提出当时的方法面对没有见过的情况时无法识别多样化的隐含社群关系，从而无法估计正确的推理路径。因此设计了一种名为COSMO（Conditional SEQ2SEQ-based Mixture model）的条件序列到序列混合模型，该模型能够动态地生成多样化的内容，并用于形成即时的动态知识图谱，以支持常识推理。Sap等人 (Sap et al., 2019)构建了ATOMIC，这是一个包含推理知识的知识图谱，其中的信息以if-then的形式呈现事件、心理状态和角色之间的关系。常识推理通常依赖于预先构建的常识知识图谱，旨在通过结构化的方式整合和表示广泛的常识性信息，从而为推理过程提供必要的知识支持。然而，当面临全新的、之前未曾遇到过的问题形式或者特定的常识领域时，这类方法往往会展现出其局限性。

2.2 寓意理解

传统的机器阅读理解任务往往将关注的重心放在机器的推理能力上。比如Li等人 (Li et al., 2022a)提出了一种神经-符号方法，该方法通过在代表文本单元之间逻辑关系的图上传递消息来预测答案，在处理需要逻辑推理的机器阅读理解任务上显示出有希望的结果。而寓意理解在此基础上还需要对深层语义的理解能力。Guan等人 (Guan et al., 2022)通过构造寓意故事数据集STORAL并在传统的模型上进行了广泛的实验以展示此类任务的难度，并提出一种检索增强的算法，通过从训练集中检索相关概念或事件作为额外的指导来提高模型性能。本文的方法与Guan的方法相似，但我们的任务更注重使用提示工程设计提示词来对模型进行指导。

2.3 上下文学习

当预训练模型达到相当规模时，其将展现出优秀的上下文学习能力（亦称情境学习） (Min et al., 2022)。具体而言，针对一个预先充分训练的庞大模型，当面临迁移至全新任务时，无需对模型进行繁琐的微调操作。仅需提供数个简明的输入-输出对示例，该模型便能理解并适应新任务的具体要求，从而展现出其高度的适应性和学习能力 (Dong et al., 2022)。

提示工程是指设计与优化输入给人工智能模型的提示词，以确保模型能够更好地根据提示词生成预期的内容。比如，Kong等人 (Kong et al., 2023)提出了一种策略性设计的角色模拟提示方法，通过设计特定的角色扮演提示来引导大型语言模型进行推理。这些提示旨在激发模型扮演特定角色或实体的能力，从而模拟出更接近真实场景的交互和问题解决过程。基于此，我们也在提示词中加入了角色扮演的部分，与Kong等人做法的不同之处在于我们为完成常识推理此类任务设计了多个角色，激发模型全面地思考问题的能力。

对于一些较为复杂的任务，比如算术、常识和符号推理等任务上，让模型直接生成最后的答案效果可能会很差，这种情况下我们可以使用思维链来激发模型的推理能力。Wei等人 (Wei et al., 2022b)受到之前使用形式语言生成中间步骤以及模型从上下文中可以进行少样本学习等工作的启发，提出了思维链的方法。通过在提示词中给出推理的中间步骤，来引导模型在之后的生成过程中生成中间步骤来更好地完成推理任务，在GSM8K任务上使用PaLM 540B结合思维链的方法实现了当时最先进的性能。在完成选择题时，分析选项-排除错误选项-得出正确选

(c) 2024 中国中文信息学会计算语言学专业委员会

项是一种产生中间步骤的方法，因此我们在提示中加入了由大语言模型生成的对各个选项的分析，来鼓励大模型分步解决问题。

在过去的一年中，大型语言模型技术取得了显著的学术与工业进展，市场上涌现出了一系列卓越的大型语言模型产品。以清华大学研发的GLM-4、百度公司精心打造的文心大模型4.0，以及OpenAI公司推出的GPT-4为例，这些模型不仅在文本生成领域展现出了卓越的能力，它们在机器翻译、命名实体识别和主题抽取等任务中也表现出了优秀的性能。鉴于此，本研究旨在探究将优秀的大型语言模型与精心设计的提示词相结合，在处理复杂任务时可能取得的效果。

3 任务描述

儿童故事常识推理与寓意理解评测分为常识推理和寓意理解两个子任务，旨在评价模型的常识推理与故事理解的能力。数据集的规模如下表所示：

数据集	开发集	测试集	总计
常识推理(CR)	400	1692	2092
寓意理解(MU)	252	1056	1308

Table 1: 数据集规模展示

常识推理：常识推理子任务的问题和答案由人工标注，问题涉及到的常识类型包含社会常识、生物常识、时间常识、空间常识以及物理常识。如Figure 1所示：

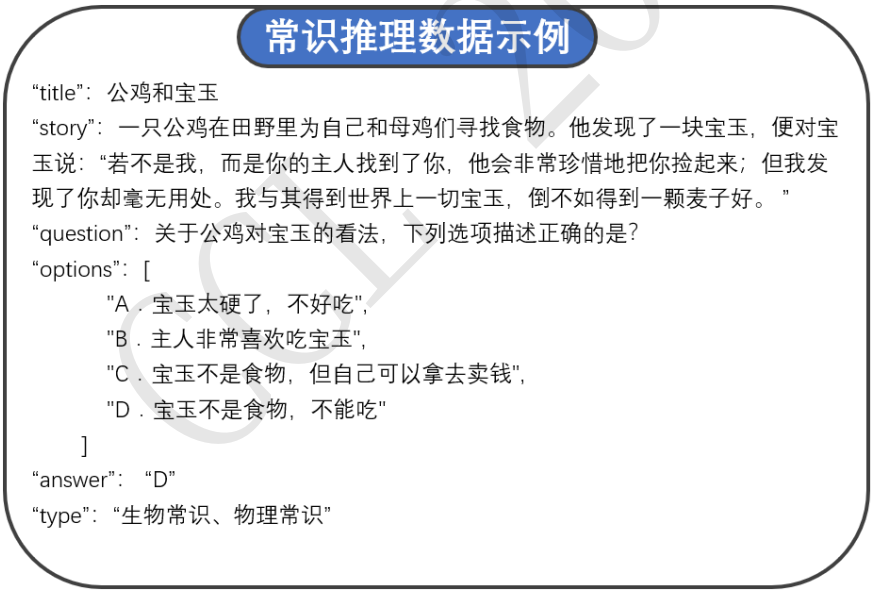


Figure 1: 赛道一常识推理数据示例

想要正确地回答图一中的问题，首先需要有一定的生物知识，即宝玉是一种矿物，它并不在公鸡的食谱上；其次，具备的物理常识让我们知道，宝玉质地坚硬，公鸡不吃宝玉并非是因为宝玉不好吃，而是其不能吃，因此对于公鸡来说宝玉还不如一粒麦子。这样可以推理出选项D是正确答案。

寓意理解：寓意理解子任务的问题和答案采用自动构建和人工标注结合的方式。题目一般要求从四个候选选项中选择最符合故事情节的寓意，如Figure 2所示：

(c) 2024 中国中文信息学会计算语言学专业委员会

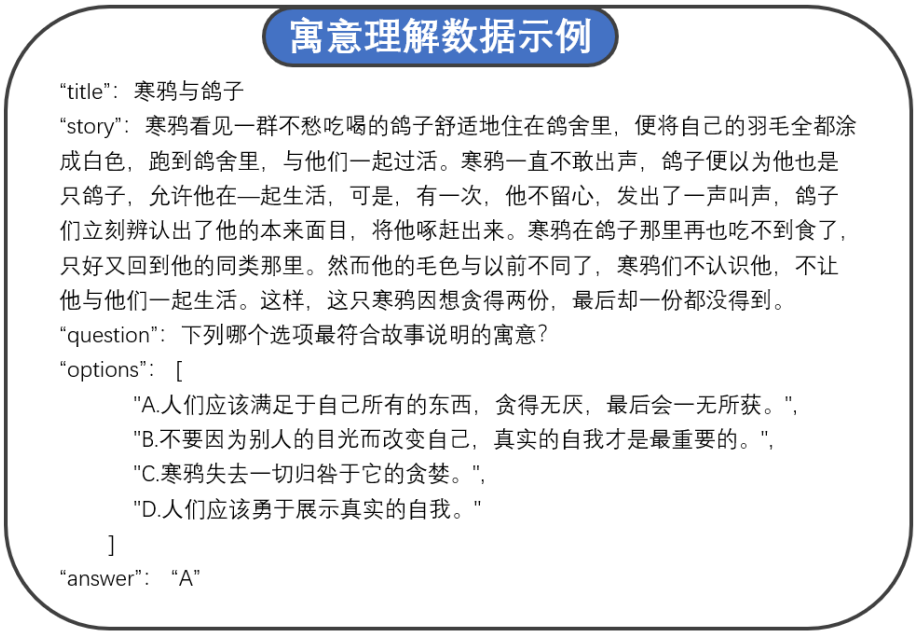


Figure 2: 赛道一寓意理解数据示例

要想推理出正确答案，首先需要理解故事。在这个故事里，寒鸦因为自己的贪念，在看见鸽子舒适的生活后伪装自己混入鸽群，在被鸽子识破之后失去了一切，最后甚至不被自己的族群所接受。然后，结合文章开始进行推理——B选项虽然涉及到“改变自己”这一故事情节，但与本文的主题无关；C选项只是重复了情节，浮于文本的表面而没有触及深层的语义；D选项与B选项类似，虽然“真实的自我”与故事的文本相关，但并非是故事的寓意。寓言故事往往通过虚构情节和以动物作为主人公来向阅读者传达哲理或警示，这则寓言故事正是在告诫我们知足是一种美德，贪婪可能会葬送我们所拥有的一切。因此，A选项是正确的选项。

4 提示词设计方法

4.1 明确任务和角色

在本研究中，我们在提示词的初始部分明确指出了大模型所执行任务的具体名称。随后，遵循OpenAI在其官方网站上发布的关于提示词设计的方法，我们进一步指导大模型在任务中应扮演的角色及其需完成的具体任务。以故事寓意理解任务为例，我们期望大模型不仅能够深入理解故事内容，还能进行批判性分析，并最终做出决策。因此，我们将这些要求明确地体现在对大模型角色的设定上：

在提示词的开篇明确任务以及角色：

#任务名称#：儿童寓言故事理解题（选择最佳寓意）
#扮演角色#：
故事解读者：精确理解故事的情节及其所要传达的核心思想。
批判性分析者：分析故事背后的深层含义，并批判性地评估各个选项与故事的一致性。
决策者：在理解故事的基础上，做出最符合故事意旨的选择。

4.2 任务详细指导

在明确了任务要求和角色定位之后，本研究进一步阐述了完成该任务所需遵循的指导原则。类比于教育领域中，教师在考试前向学生传授解题技巧，本研究也在提示词设计中向大模型传达了执行任务的关键策略。以故事寓意理解任务为具体案例，研究者作为教育者的角色，

(c) 2024 中国中文信息学会计算语言学专业委员会

基于Oakhill等人 (Oakhill et al., 2014)提出的儿童在阅读理解中所需的能力，我们提炼出了三条实用的指导原则。这些原则被纳入到提示词中，以指导大模型在执行任务时的策略选择：

在注意事项中为模型提供任务的详细指导：

#完成任务的注意事项#：
1. 理解故事重点：要精准把握故事所要表达的重点，进而推断其蕴含的深层道理。
2. 深入理解而非仅看表面：不能仅仅围绕故事的表面内容进行分析。寓言故事的目的通常是以简单的故事传递深层的人生或道德理念，因此，需要超越故事的字面意义，挖掘更深的含义。
3. 确保答案选项与故事的关联性：正确的选项必须与故事内容紧密相关，且需要围绕故事的核心思想进行论述，避免选择那些与故事无关或偏离故事主旨的答案。

4.3 任务标准定义

在机器翻译任务中，Li等人 (Li et al., 2022b)提出可以通过修改提示词来使模型生成的翻译中包含某些词汇或者风格更符合要求。基于此，本研究进一步将评价标准纳入提示词设计之中，以期优化模型输出的质量和风格。参考Yang等人 (Yang and Klein, 2021)在可控文本生成中从主题符合程度、文本质量以及多样性三方面来全面地评价受控文本的质量，我们也提出了对任务完成程度的评价角度。以故事寓意理解任务为例，我们认为准确性、深度、关联性、适宜性与一致性是衡量任务完成程度的五个维度，因此我们将其整合进提示词，旨在引导大模型生成更符合预期的翻译结果：

在提示词中明确任务完成的标准：

#完成任务的标准#：
1. 准确性：所选答案必须精确反映故事的核心寓意或教训。
2. 深度：答案需要体现对故事深层次意义的理解。
3. 关联性：选择的答案必须与故事情节和主题直接相关，且能恰当地体现故事的教育意义。
4. 适宜性：确保所选寓意适合儿童的认知水平，并能为其提供有价值的教育意义。
5. 一致性：在类似的测试场景中，所选答案应保持一致的评判标准和解释逻辑。

4.4 融入思维链与少样本策略

对于多步骤的推理问题，可以通过思维链技术让大模型将较为复杂的问题分解成可以一步步解决的子问题，然后再依次求解来提高模型的推理效果。OpenAI的研究人员发现大模型的推理能力能够通过思维链获得较大的提升，在与运动有关的常识推理上，运用了思维链的PaLM (Chowdhery et al., 2023)表现甚至超过了运动爱好者。

因此，在本研究中，我们采用的提示词设计融合了少样本学习和思维链策略：首先，我们从基线模型在训练集中表现不佳的题目中精选了一个具有代表性的例子。我们认为，挑选此类题目将激励模型进行更深入的思考。其次，受到思维链理论的启发，我们认为在多项选择题中对每个选项进行详尽解释是一种有效的问题分解策略。基于这一理念，我们决定利用GLM-4模型，结合问题和答案，对这道题的四个选项进行详尽分析，并将其整合为提示词的一部分，旨在提升模型的推理和决策能力：

在提示词中融入思维链和少样本学习的策略：

#示例#：
##寓言故事##：
（寒鸦与鸽子）寒鸦看见一群不愁吃喝的鸽子舒适地住在鸽舍里，便将自己的羽毛全都涂成白色，跑到鸽舍里，与他们一起过活。寒鸦一直不敢出声，鸽子便以为他也是只鸽子，允许他在一起生活，可是，有一次，他不留心，发出了一声叫声，鸽子们立刻辨认出了他的本来面目，将他啄赶出来。寒鸦在鸽子那里再也吃不到食了，只好又回到他的同类那里。然而他的毛色与以前不同了，寒鸦们不认识他，不让他与他们一起生活。这样，这只寒鸦因想贪得两份，最后却一份都没得到。
##问题##：
下列哪个选项最符合故事说明的寓意?
A．不要为了一时的利益而放弃自己的原则和尊严。
B．贪婪只会让人得到短暂的快乐，却失去长久的幸福。
C．虚伪和伪装最终会被识破，因而失去信任和尊重。
D．真诚和坦率的态度才是与他人建立真正联系的关键。
##答案##：选B
##分析##：选项B是最符合这个寓言故事的寓意，因为它直接指出了贪婪行为所带来的后果。故事中的寒鸦因为追求短期的利益而伪装自己，最终不仅失去了新的生活环境，也被自己原本的同类排斥。这显示了寒鸦的贪婪导致了短暂的满足后的长期不幸，体现了贪婪可能带来的瞬间快乐和长远的失落。因此，选项B“贪婪只会让人得到短暂的快乐，却失去长久的幸福”最准确地反映了故事的主旨。其他选项虽然也可以从某种程度上解释故事中的某些方面，但没有选项B那样直接和全面地反映了故事的核心教训。选项A侧重于原则和尊严，选项C关注于伪装和识破，选项D强调真诚和坦率，而B选项则是直接指向了故事的主题——贪婪的后果，这是故事最核心的寓意。因此，B选项是对这个故事寓意最恰当的表达。

4.5 运用分割符

在本研究中，提示工程的设计原则贯穿于我们提示词的构建过程。OpenAI在其发布的提示词设计指南中建议，应恰当地使用分隔符来区分输入的不同部分，比如使用三重引号（”””标题”””）来明确标识。此外，为了便于后续处理中答案的提取，我们遵循提示工程的原则设计了格式，引导模型首先输出答案，随后提供相应的分析：

在提示词中善用分隔符引导模型：

##寓言故事##：题目：寒鸦与鸽子
故事：寒鸦看见一群不愁吃喝的鸽子舒适地住在鸽舍里.....
##问题##：
问题：下列哪个选项最符合故事说明的寓意?
选项：
A．不要为了一时的利益而放弃自己的原则和尊严。
B．贪婪只会让人得到短暂的快乐，却失去长久的幸福。
C．虚伪和伪装最终会被识破，因而失去信任和尊重。
D．真诚和坦率的态度才是与他人建立真正联系的关键。
##答案##：
##分析##：

(c) 2024 中国中文信息学会计算语言学专业委员会

5 实验

5.1 提示词设计与模型选择

我们按照在第4章提供的方法，针对常识推理和寓意理解这两个任务，精心构建了相应的提示词。由于我们在提示词设计中融入了思维链技术，因此对模型的规模提出了一定的要求：根据Wei等人 (Wei et al., 2022b)的发现，思维链提示是一种取决于模型尺度的涌现能力 (Wei et al., 2022a)。对于大多数参数量小于10B的小型模型，思维链提示会导致模型性能的损害。只有在与参数量较大的模型一起使用时才会产生性能提升。我们在一些模型上进行了评测，包括ERNIE-speed-128k、deepseek-chat、Yi-34B-chat、glm4-air以及GPT-4。我们的实验结果如下表所示：

队伍	常识推理	寓意理解	总计
Baseline	0.688	0.561	0.612
Team-1	0.865	**0.744**	**0.793**
Team-2	0.834	0.734	0.774
RENIE-speed-128k	0.657	0.562	0.600
deepseek-chat	0.826	0.645	0.717
Yi-34B-chat	0.727	0.602	0.652
glm4-air	0.768	0.412	0.554
Our Method（GPT4）	**0.869**	0.708	0.773

Table 2: 封闭数据的赛道一评测提交结果对比

根据Table 2所展示的数据，GPT-4模型结合本研究采用的方法在赛道一的两个子任务中均展现出了优秀的性能。特别是在常识推理任务上，我们的方法在赛道一上获得了第一名的成绩，这充分地证明了我们的方法的有效性。

5.2 实验结果分析

为了深入探究在本研究所提出的提示词中哪些组成部分对任务性能具有显著影响，本研究进一步开展了消融实验。该实验旨在系统地评估和比较各个组成部分对模型性能的具体贡献。消融实验涵盖了以下关键要素：任务与角色（TR）、详尽的指导（DG）、任务的标准（CT）以及学习样本（SL）。以下是消融实验的结果：

提示词	常识推理	寓意理解	总计
Our Prompt	**0.869**	0.708	0.773
-TR	0.861	**0.721**	**0.777**
-DG	0.856	0.705	0.766
-CT	0.865	0.715	0.775
-SL	0.832	0.688	0.745

Table 3: 消融实验结果对比

由Table 3的实验结果，我们得到了以下结论：

(1) 针对常识推理任务，本研究设计的提示词在实验中取得了显著的高分，这一结果表明，所设计的提示词能够有效地促进模型对常识知识的理解和应用。提示词的设计在提升模型对常识推理任务的处理能力方面发挥了关键作用；

(2) 在寓意理解任务上，本研究设计的提示词得分略低于“-TR”以及“-CT”。经分析，这一现象可能源于提示词中角色和评价标准的定义过于复杂，导致模型未能充分集中注意力于深层语义的理解，从而影响了其性能表现；

(3) 在常识推理和寓意理解两个子任务中，“-SL”的得分普遍较低，这一结果突显了在提示词中融入任务样例的必要性。这表明，任务样例的加入对于提升模型在相关任务上的表现具有显著影响。

第二十三届中国计算语言学大会论文集，第336页-第345页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

6 总结

本文针对儿童故事常识推理与寓意理解评测赛道一的两个任务，提出了一种基于提示工程的提示词构造方法。我们首先提出了一种融合提示工程和思维链的通用提示词构建框架，然后针对具体任务对提示词模板进行了调整，最后结合语言模型使用这些提示词进行结果预测。在本次评测中，我们的方法在赛道一的封闭数据条件下取得了第三名的成绩，证明了该方法的有效性。通过这项研究，我们展示了提示工程在提升大语言模型性能方面的重要性，尤其是在处理复杂自然语言理解任务时。此外，我们的方法不仅强调了提示词设计的关键性，还展示了在实际应用中对任务需求和模型特性的深刻理解和灵活运用的必要性。然而，我们也意识到当前提示词设计方法存在的局限性，尤其是在需要人工设计模板且步骤较为繁琐的情况下。展望未来，我们计划继续改进提示词构建方法，以期实现更高效的自动化设计流程。同时，我们也期待将这一方法应用于更广泛的自然语言处理任务中，以探索其更深远的应用潜力。我们相信，随着技术的不断进步和研究的深入，提示工程将为自然语言处理领域带来更多创新和突破。

参考文献

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*.

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. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Miguel Ángel Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. *arXiv preprint arXiv:2204.09438*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Xiao Li, Gong Cheng, Ziheng Chen, Yawei Sun, and Yuzhong Qu. 2022a. Adalogn: Adaptive logic graph network for reasoning-based machine reading comprehension. *arXiv preprint arXiv:2203.08992*.

Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. 2022b. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590.

Chen Ling, Xuchao Zhang, Xujiang Zhao, Yifeng Wu, Yanchi Liu, Wei Cheng, Haifeng Chen, and Liang Zhao. 2023. Knowledge-enhanced prompt for open-domain commonsense reasoning. In *1st AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

第二十三届中国计算语言学大会论文集，第336页–第345页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Farhad Moghimifar, Lizhen Qu, Yue Zhuo, Mahsa Baktashmotlagh, and Gholamreza Haffari. 2020. Cosmo: Conditional seq2seq-based mixture model for zero-shot commonsense question answering. *arXiv preprint arXiv:2011.00777*.

Jane Oakhill, Kate Cain, and Carsten Elbro. 2014. *Understanding and teaching reading comprehension: A handbook*. Routledge.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Chris Tanasescu, Vaibhav Kesarwani, and Diana Inkpen. 2018. Metaphor detection by deep learning and the place of poetic metaphor in digital humanities. In *The thirty-first international flairs conference*.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*.

CCL 2024

Overview of CCL24-Eval Task 8: Evaluation of Commonsense Reasoning and Moral Understanding in Children's Stories

Guohang Yan[1], Feihao Liang[1], Yaxin Guo[1], Hongye Tan[1,2,⋆], Ru Li[1,2], Hu Zhang[1]
[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
{202222407055, 202322408029, 202112407002}@email.sxu.edu.cn
{tanhongye, liru, zhanghu}@sxu.edu.cn

Abstract

This paper provides a comprehensive review of the the CCL24-Eval Task 8: ***C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories*(**CRMUS**). This task has designed two sub-tasks, which aim to assess the commonsense reasoning and implicit meaning comprehension capabilities of Large Language Models(LLMs). We heve received registration forms from 33 teams, 15 of which submitted final results that exceeded the baseline score. We present the results of the top 5 teams and our analysis of these results.

1 Introduction

Stories are essential reading material in education, often containing rich knowledge, vivid plots, memorable characters, and profound implicit meanings. They serve as important vehicles for the dissemination of knowledge, cultural inheritance, and value shaping. Story comprehension requires models not only to understand plots based on social, physical, and other common knowledge, but also to analyze character relationships, intentions, and behaviors, and to infer the profound meanings conveyed by the story (Tomasulo et al., 2012; Pelletier and Beatty, 2015; Dorfman and Brewer, 1994). It is suitable for evaluating the cognitive abilities of LLMs.

Therefore, We have constructed a new challenging story comprehension dataset **CRMUS** (***C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories*), and designed two sub-tasks based on the cognitive process of human comprehension of stories. Moreover, we organized CCL24-Eval Task 8, **CRMUS**. This evaluation is divided into two tracks: (1) Track 1 allows the use of commercial LLMs through prompt learning; (2) Track 2 allows the use of open-source LLMs through fine-tuning, but the model parameters must not exceed 7 billion. In the end, we received registration forms from 33 teams, of which 15 submitted final results that exceeded the baseline we provided. We found that although LLMs already possess certain text comprehension and reasoning abilities, they still perform poorly in deep semantic comprehension and reasoning tasks that extend beyond the surface meaning of the text, such as commonsense reasoning and implicit meanings comprehension.

The task description is presented in Section 2. The dataset we constructed for this task in Section 3. In Section 4, we provide baselines for two sub-tasks. We discuss the metrics used to rank participant submissions in Section 5. In Section 6, we list participants' information and results from their submissions and provide a more in-depth discussion. We introduce the methods of excellent teams in Section 7. Finally, We conclude the paper in Section 8.

2 Task Description

We designed the following two sub-tasks to evaluate the commonsense reasoning and implicit meaning comprehension abilities of LLMs.

Commonsense Reasoning(CR) Based on a given story and associated commonsense questions, the sub-task requires selecting the correct answer. This sub-task requires the model to reason and answer

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 346-352, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

questions using commonsense knowledge (usually implicit) derived from the story. The questions are in multiple-choice format, each including a question and four options.

Moral Understanding(MU) Based on a given story, select the most appropriate and relevant moral from multiple candidate options that best fits the story plot. This sub-task is a multiple-choice question with four moral options.

There are two tracks set for this evaluation, each containing the two sub-tasks mentioned above. Track 1 allows the use of ChatGPT, GPT-4, ERNIE Bot, and other commercial LLMs through prompt learning; Track 2 allows the use of open-source LLMs such as LLaMA-2 and Qwen-1.5 through fine-tuning, with model parameters not exceeding 7 billion.

3 Datasets

3.1 Dataset Construction

This task uses classic fable stories, manually collected from website[1], as raw materials and meticulously annotates them to construct the **CRMUS** dataset.

Annotation Process We annotates data through the following steps:

- **Preparation** We have developed an annotated outline that includes task definitions and examples. Based on this outline, we invited 10 graduate students with NLP-related knowledge from our team to participate in the annotation process. To enhance efficiency and quality, annotators first independently annotate the same story, then summarize the issues encountered during annotation, and refine the annotation outline accordingly.

- **Initial Annotation** For the commonsense reasoning sub-task, to ensure diversity of problems, at least two annotators are required to propose a minimum of 4 questions for each story and provide corresponding options. The questions should encompass various commonsense types, such as society, biology, time, space, and physics. Additionally, to more effectively highlight the model's limitations, annotators must identify the commonsense types relevant to each problem and provide a detailed explanation for the answers. For the moral understanding sub-task, we use the sentences of story implicit meanings as the correct answers and require annotators to provide three different implicit meanings as incorrect answers. Additionally, we request annotators to annotate two additional questions for each story using LLMs via prompt learning. Specifically, to enhance the diversity of options, annotators are required to create prompt templates, utilize various LLMs to generate multiple implicit meanings based on the story, and then filter and rewrite them to align more closely with the story's existing one. These implicit meanings are used as candidate answers for the remaining two questions.

- **Quality Control** We adopt a cross-checking approach to process the collected data. For the commonsense reasoning task, examiners are required to rate each question on a scale from 0 (unqualified) to 2 (excellent) and make modifications or add additional annotations to some questions as necessary. Finally, non-annotators will conduct secondary verification and remove any non-conforming data. For the moral understanding sub-task, inspectors carefully review each option and modify or re-annotate those that do not meet the requirements.

Finally, we adjusted the distribution of correct answers in the dataset to randomly and evenly spread them across options A, B, C, and D.

3.2 Data Samples

Each example in the development and test sets of the commonsense reasoning sub-task includes the following information: ID, title, story, question, options, answer, and commonsense type. The moral understanding sub-task includes the same information except for commonsense type. Specific examples are detailed in Figure 1:

[1] https://m.thn21.com/Article/chang/3306.html

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 346-352, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.3 Data Statistics

The questions and answers of the commonsense reasoning sub-task are manually annotated, while the moral understanding sub-task uses a combination of automatic generation and manual annotation. The types of common knowledge involved in the commonsense reasoning task include social, biological, temporal, spatial, and physical commonsense. The specific counts of different questions are detailed in Table 1. (Note: Some questions involve multiple types of commonsense)

Commonsense Type	Number
Social	1048
Biological	426
Temporal	308
Spatial	259
Physical	178

Table 1: Number of questions for each commonsense type

The number of question contained in each file is shown in Table 2.

Sub-task	Dev Set	Test Set	Total
Commonsense Reasoning	400	1692	2092
Moral Understanding	252	1056	1308

Table 2: Dataset size of **CRMUS**

4 Baseline

Track 1 utilizes the commercial LLM GLM-3-Turbo from Zhipu AI as the baseline model. Track 2 employs the LLaMA-2 open-source model, chinese-alpaca-2-7b-hf, fine-tuned with a Chinese corpus. For details of the baseline system, refer to the description available at website[1].

5 Evaluation Metrics

The final evaluation **Score** of the participating model is the weighted average of the accuracy of the answers in each sub-task. The specific calculation method is as follows:

$$Score = 0.4 * Acc_1 + 0.6 * Acc_2 \quad (1)$$

Specifically,
Acc_1 = the accuracy of answers for the commonsense reasoning sub-task
Acc_2 = the accuracy for the moral understanding sub-task.

6 Results and Analysis

Table 3 and Table 4 respectively present the top five official rankings of the two tracks, based primarily on the **Score**. Teams in Track 1 and Track 2 surpassed the baseline model scores. It is observed that the overall Score of Track 1 teams surpasses that of Track 2 teams, highlighting the advantage of commercial closed-source LLMs over open-source LLMs with parameters under 7B.

Based on the model proposals submitted by participating teams, it was observed that most teams employ the prompt design strategy to prompt LLMs to identify commonsense knowledge within the story,

[1] https://github.com/SXU-YaxinGuo/CRMU

Team Name	Organization	Rank	Score	CR score	CR score
Arabian Nights	South China Normal Univerity Shandong University	1	85.13	82.86	86.65
holoflow	Individual	2	79.27	86.52	74.43
AIAYN	Beijing Jiaotong University	3	77.66	86.05	72.06
XCZL	China Telecom	4	77.39	83.39	73.39
ZZU_NLP	Zhengzhou University	5	75.66	84.46	69.79
Basiline	-	-	61.15	68.79	56.06

Table 3: Track 1 results (Unit: %)

Team Name	Organization	Rank	Score	CR score	MU score
ytkj	Huazhong University of Science and Technology	1	80.82	66.96	90.06
ZZU_NLP	Zhengzhou University	2	74.38	72.87	75.38
Arabian Nights	South China Normal Univerity Shandong University	3	73.85	59.34	83.52
zyy	Shanghai University	4	71.42	70.74	71.88
XJTLU-DKE	Xi'an Jiaotong-liverpool University	5	71.22	70.8	71.5
Basiline	-	-	32.4	31.15	33.24

Table 4: Track 2 results (Unit: %)

perform commonsense reasoning, and select appropriate morals that align with the narrative. Strategies include assigning specific roles to LLMs, establishing "task completion precautions," defining "task completion standards," and similar approaches.

Certain participating teams performed fine-tuning experiments on open-source LLMs using methods like LoRA(Hu et al., 2021), selecting optimal parameters and fine-tuning modules to enhance LLMs' performance in tasks related to commonsense reasoning and implicit meaning comprehension. Overall, while these teams explored novel and interesting approaches and achieved some results, the innovativeness of these techniques was limited. They focused on activating the capabilities of LLMs for specific tasks without fundamentally enhancing the models' innate ability in commonsense reasoning and deep semantic comprehension.

7 Participant Systems

This evaluation includes two tracks. Track 1 primarily assesses the performance of different commercial models in tasks related to commonsense reasoning and implicit meaning comprehension, alongside evaluating the efficacy of various prompt strategies in enhancing model capabilities. Track 2 focuses on investigating whether open-source LLMs with limited parameter sizes can enhance their commonsense reasoning and implicit meaning comprehension abilities through pretraining and fine-tuning. Presented below are the technical approaches adopted by select outstanding teams across two tracks.

Track 1

In Track 1, ***holoflow*** proposes a straightforward yet effective two-stage prompt engineering.

• Initially, they used identical prompts to obtain responses from three advanced commercial LLMs: GPT-4, ERNIE-4, and Qwen-Max, respectively.

• Subsequently, they implemented a majority voting strategy for the LLM responses obtained in the first step. In cases of inconsistency, they queried GPT-4 for secondary confirmation using a slightly modified prompt compared to the first step, narrowing down the options to those returned initially. The choice confirmed in this secondary phase was selected as the final submission result.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 346-352, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The experimental results demonstrate that their method achieved the final Score of 79.27, placing first in the closed dataset of Track 1 among 10 submitted results, thereby confirming its effectiveness. The results further validate the efficacy of the prompt-based approach in addressing the CRMUS task.

Track 2

ZZU_NLP secured first place in the closed data of Track 2. Their approach primarily involved designing effective prompt templates, fine-tuning LoRA parameters, and utilizing data augmentation techniques.

In the instruction fine-tuning stage, they chose two LLMs mainly in Chinese, namely Qwen1.5-Chat-7B (Bai et al., 2023) and Internlm2-Chat-7B (Cai et al., 2024). Among them, Internlm2-Chat-7B is the main fine-tuning model, and Qwen1.5-Chat-7B is the auxiliary model to verify the optimal LoRA parameters. By testing the combination of different LoRA parameters and fine-tuning modules, it was ultimately determined that two sets of parameters can provide the optimal Acc indicators for CR and MU, respectively.

In the process of conducting commonsense reasoning on the development set, they found that the model performed poorly in terms of temporal, spatial, and physical knowledge, and speculated that this may be due to the small amount of data for several commonsense types. Therefore, they used data augmentation methods to address the issue of uneven distribution of different commonsense types in the CRMUS dataset. They created over 200 commonsense reasoning data using ChatGPT, and then manually reviewed and screened 137 high-quality data. The data was then expanded to the development set for fine-tuning, resulting in an improvement in the accuracy.

8 Conclusion and Future Work

This paper presents an overview of the CCL24-Eval Task 8, i.e., ***C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories*(**CRMUS**). This evaluation is conducted using our meticulously annotated CRMUS dataset. These tasks are designed to assess LLMs' ability to understand and reason about commonsense knowledge in stories, as well as their capacity to capture the deep semantics and implicit meanings within the stories. We received a total of 33 completed registration forms, of which 15 teams submitted the final results that exceeded the baseline we provided. Additionally, we offer a comprehensive analysis and summary of the methodologies employed by the participants, which will inform and guide future research in the field of natural language processing.

Finally, we believe that this evaluation remains challenging for LLMs, primarily due to the models' insufficient semantic understanding and reasoning abilities. In the future, we will continue to explore and enhance the CRMUS dataset, aiming to further improve its scale and quality. We also aim to explore additional forms of commonsense and reasoning Q&A, as well as moral examination methods, to better evaluate the commonsense reasoning and deep semantic understanding abilities of LLMs.

9 Acknowledgements

We would like to acknowledge the contributions of the other members of Hongye Tan's team and thank them for annotating the CRMUS corpus. We appreciate the support from Scientific and Technological Innovation 2030 - "New Generation Artificial Intelligence"(2020AAA0106100) and the National Natural Science Foundation of China(62076155). Lastly, we extend our appreciation to the CCL Evaluation Committee for their support.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Marcy H Dorfman and William F Brewer. 1994. Understanding the points of fables. *Discourse Processes*, 17(1):105–129.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Janette Pelletier and Ruth Beatty. 2015. Children's understanding of aesop's fables: relations to reading comprehension and theory of mind. *Frontiers in Psychology*, 6:146239.

Daniel J Tomasulo, James O Pawelski, et al. 2012. Happily ever after: The use of stories to promote positive interventions. *Psychology*, 3(12):1189.

CCL 2024

title	The Crow And The Jug		乌鸦喝水
story	The crow was extremely thirsty and flew to a large water jar. There wasn't much water in the jar, but he tried his best but still couldn't drink it. So he exerted all his strength to push, trying to topple the jar and pour out water, but the large water jar couldn't be pushed. At this moment, the crow remembered the method he had used before and threw stones into the water jar. As the number of stones increased, the water in the jar gradually increased. Finally, the crow happily drank water and quenched its thirst.		乌鸦口渴得要命，飞到一只大水罐旁，水罐里没有很多水，他想尽了办法，仍喝不到。于是，他就使出全身力气去推，想把罐推倒，倒出水来，而大水罐却推也推不动。这时，乌鸦想起了他曾经使用的办法，用口叼着石子投到水罐里，随着石子的增多，罐里的水也就逐渐地升高了。最后，乌鸦高兴地喝到了水，解了口渴。
CR sample	question	What else can a crow throw into a jar to drink water in the story?	文中乌鸦还可以将什么东西丢到罐子里来喝到水?
	options	A. Stone Lion B. Table Tennis C. Leaves **D. Glass beads**	A．石狮子 B．乒乓球 C．树叶 **D．玻璃珠**
MU sample	question	Which of the following options **best** corresponds to the implicit meaning of the story?	下列哪个选项**最**符合故事隐含的寓意?
	options	A. Merely relying on past experience without emphasizing thinking and innovation is insufficient. **B. Intelligence and wit can sometimes be more effective than brute force.** C. While strong physical strength can resolve challenges, wisdom also relies on strength. D. Teamwork can sometimes overcome difficulties.	A.不注重思考和创新，而只依赖过去的经验是不可行的。 **B.聪明机智有时比蛮力更为有效。** C.强大的力量才能解决困境，智慧也得依 靠力量。 D.有时团队协作能够克服困难。

Figure 1: Samples of **CRMUS**

CCL 2024

System Report for CCL24-Eval Task 9:Chinese Vision-Language Understanding Evaluation

Jiangkuo Wang, Linwei Zheng, Kehai Chen*, Xuefeng Bai, Min Zhang

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

{220110927,220110604}@stu.hit.edu.cn

{chenkehai,baixuefeng,zhangmin2021}@hit.edu.cn

Abstract

This paper introduces our systems submitted for the Chinese Vision-Language Understanding Evaluation task at the 23rd Chinese Computational Linguistics Conference. In this competition, we utilized X^2-VLM and CCLM models to participate in various subtasks such as image-text retrieval, visual grounding, visual dialogue, and visual question answering. Additionally, we employed other models to assess performance on certain subtasks. We optimized our models and successfully applied them to these different tasks.

1 Introduction

In today's era of information explosion, multimodal understanding and interaction between vision and language have become increasingly important. With the rapid development of deep learning technology, vision-language pre-training models can establish associations between images and text, demonstrating powerful performance.

This competition encompasses various subtasks, spanning image-text retrieval, visual grounding, visual dialogue, and visual question answering. These tasks not only assess the model's comprehension and processing abilities with multimodal data but also evaluate its adaptability and robustness across diverse application scenarios. Through participation in these tasks, our goal is to validate the effectiveness of different vision-language pre-training models and delve into their potential in practical deployments.

The image-text retrieval task requires the model to retrieve relevant images based on textual descriptions or find matching text based on images. This demands the model to efficiently encode and align image and text features. The visual grounding task requires the model to accurately locate target objects in images, testing its fine-grained feature extraction capabilities. The visual dialogue task involves understanding the content of images and generating natural dialogues based on context. The visual question answering task requires the model to answer questions based on image content, assessing its visual understanding and language generation capabilities.

In this competition, we successfully completed the above tasks by optimizing and adjusting multiple models, including the X^2-VLM model. This paper provides a detailed introduction to our research methods, experimental process, and results, and summarizes the performance and achievements of the models in each task. Additionally, we will share the challenges and solutions encountered during the competition, aiming to offer references for future research and applications.

2 Methodology

The main models discussed in this paper include X^2-VLM (Zeng Y et al.), CCLM (Zeng Y et al.), Chinese CLIP (Redford et al.), and OFA (Wang P et al.). X^2-VLM is a general pre-training model capable of handling tasks that combine vision and language. CLIP, proposed

*Corresponding author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 353–363, Taiyuan, China, July 25 – 28, 2024.
Volume3: Evaluations

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

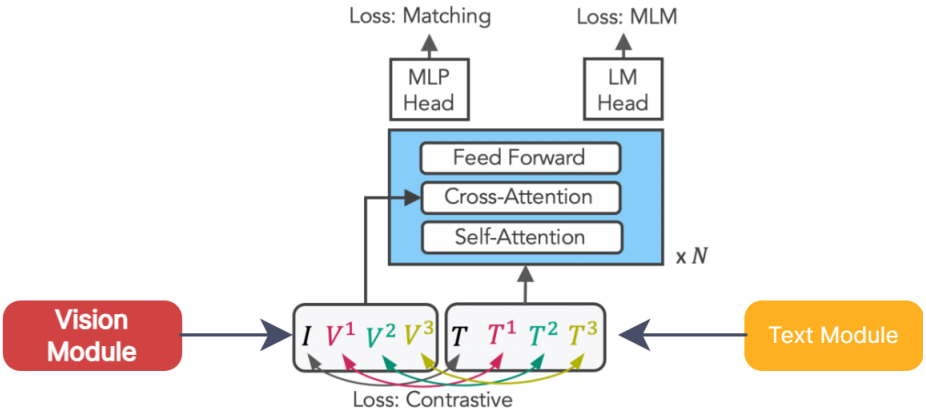


Figure 1: The overall structure of the X^2-VLM model

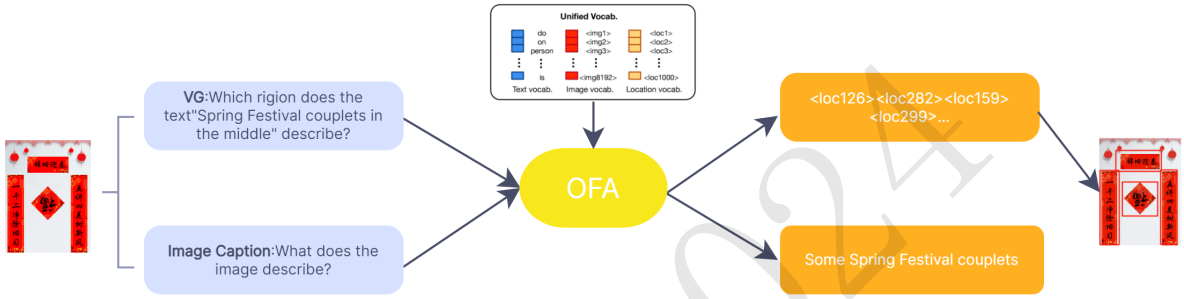


Figure 2: The overall structure of the OFA model

by OpenAI, is a multimodal model that can embed images and text into the same semantic space.

2.1 Model Architecture

Existing research on vision-language alignment generally falls into two categories: coarse-grained and fine-grained. Coarse-grained approaches use convolutional neural networks (He et al.) or vision transformers (Alexey et al.) to encode overall image features (Huang, Kim,Li rt al.). However, these methods struggle with fine-grained vision-language alignments, such as at the object level, from noisy image-text pairs that are typically weakly correlated (Huo et al.). To achieve fine-grained alignment, many methods use pre-trained object detectors as image encoders (Hao et al. Lu et al. Gan et al. Chen et al.). However, object detectors produce object-centric features that cannot encode relationships among multiple objects and can only recognize a limited number of object categories.

X^2-VLM is a unified model with vision, text, and multimodal fusion modules, all based on the Transformer architecture (as shown in Figure 1). We use three types of data for vision-language pre-training: object labels on images (Lin,Shao et al.) such as "man" or "backpack," region annotations on images (Kuznetsova et al.) such as "boy wearing backpack," and text descriptions for images such as "The first day of school gives a mixed feeling to both students and parents." The fusion module integrates vision and text features through cross-attention mechanisms. During pre-training, the modules act as encoders, and the text and fusion modules are adaptable for generative tasks. The model handles various data types, including image-text pairs, video-text pairs, and image annotations. It aligns visual concepts with textual descriptions and localizes them within images. This architecture facilitates unified encoding for both images and videos, leveraging pre-training to enhance understanding across modalities.

CCL 2024

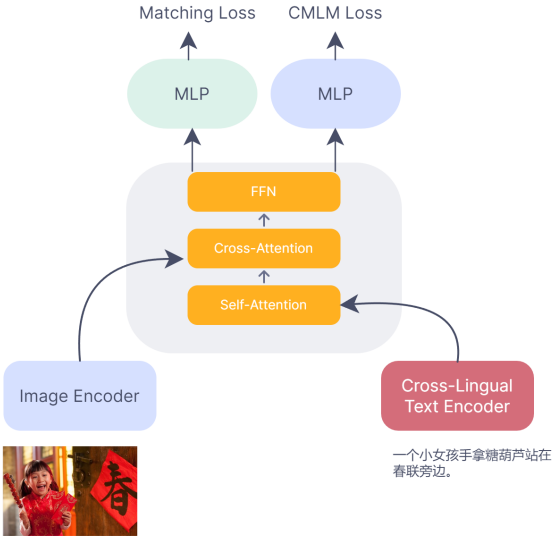


Figure 3: The overall structure of the CCLM model

The Cross-view Language Modeling (CCLM) framework combines cross-lingual and cross-modal pre-training using shared architecture and objectives (as shown in Figure 2). It consists of a Transformer-based image encoder, a cross-lingual text encoder, and a fusion model. The image encoder (Dovitskiy et al.) splits images into patches and embeds them, while the text encoder processes text inputs. The fusion model integrates text and image features through cross-attention. CCLM aligns representations of paired inputs in a common semantic space, sharing input-output formats, architectures, and training objectives. It uses contrastive, matching, and conditional masked language modeling losses to maximize sequence and token-level mutual information between inputs.

OFA is a unified Seq2Seq framework designed to integrate input/output modalities, architectures, and tasks. The model uses ResNet for visual feature extraction and byte-pair encoding for text processing. It employs a unified vocabulary for text, images, and objects (as shown in Figure 3). The architecture is based on the Transformer encoder-decoder framework, incorporating self-attention, feed-forward networks, and cross-attention layers. OFA supports multi-task and multimodal learning, including tasks such as visual grounding, image captioning, and visual question answering. It leverages large-scale pre-training datasets (Wei, Sanh et al.) and optimizes performance using cross-entropy loss and a Trie-based search strategy. Compared to models that rely on much larger paired datasets (Wang et al. Yuan et al.), OFA achieves better performance in various vision and language downstream tasks.

2.2 Innovative Text Data Augmentation

To further enhance the performance of our models in the image-text retrieval (ITR), visual dialogue (VD), and visual question answering (VQA) subtasks, we introduced an innovative text data augmentation strategy. Leveraging the advanced capabilities of the ChatGPT large language model, we performed various augmentation techniques on the textual data in our dataset. These techniques included synonym replacement, random insertion, random deletion, and random swapping of words within the text descriptions, showcasing our novel approach to enhancing textual data diversity.

Synonym Replacement Our method involved using ChatGPT to identify and replace words in the text with their synonyms. This innovative approach generated diverse versions of the same text, improving the model's ability to generalize across different expressions of the same concept. For example, the sentence "The boy is playing with a ball" could be augmented to "The child

is playing with a sphere."

Random Insertion For random insertion, we utilized ChatGPT to insert contextually relevant words at various positions within the text. This technique, innovative in its contextual relevance, augmented sentences like "The boy is playing with a ball" to "The energetic boy is playing with a ball happily," thus increasing data variety.

Random Deletion In random deletion, we employed ChatGPT to randomly remove words from the text while maintaining the overall meaning. This method enhances the model's robustness by forcing it to infer missing information. For example, "The boy is playing with a ball" could be augmented to "The boy playing a ball."

Random Swapping Random swapping involved using ChatGPT to randomly exchange the positions of words within the text, creating syntactically varied sentences that convey the same meaning. This technique improved the model's flexibility in understanding different word orders, such as augmenting "The boy is playing with a ball" to "Playing with a ball, the boy is."

Implementation Details The text augmentation was automated using ChatGPT's API. We systematically applied these innovative techniques to the entire dataset, ensuring contextually appropriate synonyms, semantically related insertions, comprehensible deletions, and structured swaps. This novel integration significantly increased the diversity of our textual data, improving the training process and enhancing the model's performance across tasks.

Overall, our innovative use of ChatGPT for text data augmentation demonstrated a unique and effective strategy, significantly boosting the robustness and generalization capabilities of our models.

2.3 Data Preprocessing

To enhance the model's performance across various tasks, we conducted the following preprocessing operations:

- **Visual Question Answering Task:** We augmented the text input of the model by incorporating historical question-answer pairs and special tokens. This addition provides additional contextual information, aiding the model in better understanding the input text and generating accurate answers.

- **Visual Dialogue Task:** Similarly, we enriched the text input by including historical dialogue pairs and special tokens. This ensures that the model considers the context comprehensively during the dialogue process, leading to more informed predictions.

- **Data Augmentation:** For all tasks, we applied diverse data augmentation techniques to the images, including random cropping, flipping, rotation, and color jittering. These operations increase data diversity, mitigate overfitting, and improve the model's generalization capability.

- **Answer List Processing:** In visual question answering and visual dialogue tasks, we modified the answer list content without removing duplicates. This approach prioritizes frequently occurring answers in the list, thereby enhancing response accuracy and diversity.

2.4 Multi-Task Learning Approach

To enhance our models' performance and robustness, we employed an innovative multi-task learning (MTL) strategy. This allows our models to learn from multiple related tasks simultaneously, improving generalization across image-text retrieval (ITR), visual dialogue (VD), and visual question answering (VQA).

CCL 2024

Multi-Task Learning Framework Our MTL framework trains a single model on multiple tasks concurrently. Shared layers learn common representations, while task-specific layers capture nuances for each task. This reduces overfitting and improves overall performance.

Joint Loss Function We designed a joint loss function that combines losses from each task:

$$L = \lambda_1 L_{ITR} + \lambda_2 L_{VD} + \lambda_3 L_{VQA}$$

where λ_i are weights for each task's loss. These weights were tuned to balance the contributions during training.

Training Procedure Our training procedure involved:

- **Data Preparation:** We prepared a unified dataset with samples for ITR, VD, and VQA tasks.
- **Model Initialization:** Models were initialized with pre-trained weights.
- **Joint Training:** We trained the model using the joint loss function, updating shared and task-specific layers based on combined gradients.
- **Hyperparameter Tuning:** Extensive tuning was conducted to balance task losses.

Innovative Impact Our MTL approach leverages shared knowledge to enhance model performance, setting a new standard for integrating multiple tasks in a unified framework. This demonstrates substantial advancements in vision-language understanding.

By employing this MTL strategy, we achieved significant improvements in developing robust and efficient multimodal models.

2.5 Model Fusion

In our image-text retrieval task, we used an innovative model fusion approach to enhance accuracy and robustness. We combined four models: X^2-VLM, CCLM, Chinese CLIP, and OFA, leveraging their unique strengths.

First, we independently trained and evaluated each model on our dataset. We then extracted retrieval results from each model for a given image and combined these results using a weighted voting method. Each model's contribution was weighted based on its performance metrics (e.g., accuracy and recall) from the validation phase. This ensured that the most reliable models had a greater influence on the final results.

For each image, each model generated a ranked list of text descriptions. We calculated a weighted score for each text description across all models, selecting the descriptions with the highest aggregated scores as the best matches. This comprehensive approach improved retrieval performance and robustness.

The fusion process involved:

- **Model Training and Evaluation:** Independently training and evaluating each model on the image-text retrieval task.
- **Result Extraction:** Generating ranked lists of text descriptions for each image from each model.
- **Weighted Voting:** Applying a weighted voting mechanism based on individual model performance.
- **Score Aggregation:** Aggregating scores for each text description across all models.
- **Final Selection:** Choosing the top-ranked text descriptions as the most relevant matches.

This model fusion approach significantly improved retrieval performance and ensured robust results by mitigating individual model weaknesses.

CCL 2024

Model Usage Strategy In addition to utilizing the X^2-VLM and CCLM models to tackle all five subtasks, we also leveraged the Chinese CLIP model and the OFA model for certain tasks to evaluate their performance in multimodal scenarios. Specifically, we employed a model fusion approach for the visual image-text retrieval subtasks. By combining the strengths of X^2-VLM, CCLM, Chinese CLIP, and OFA through a weighted voting mechanism, we enhanced the accuracy and robustness of our results. This fusion strategy allowed us to leverage the unique capabilities of each model, leading to improved performance across these key subtasks.

3 Experiments

3.1 Dataset Description

Our experiments utilized the Chinese image-text multimodal understanding evaluation dataset provided by the organizers. The dataset includes images from 15 main categories and 92 subcategories, manually curated to reflect elements commonly found in Chinese cultural contexts or everyday life.

3.2 Experimental Setup

We structured our experiments into several key steps, all conducted on four NVIDIA RTX A6000 GPUs:

1. **Data Preprocessing:** We enhanced text data diversity using innovative text data augmentation techniques, including synonym replacement, random insertion, random deletion, and random swapping with ChatGPT. Additionally, we applied diverse image augmentation techniques like random cropping, flipping, rotation, and color jittering to improve data diversity and model generalization.

2. **Model Selection and Architecture Optimization:** We selected four models: X^2-VLM, CCLM, Chinese CLIP, and OFA, optimizing their architectures for performance and efficiency. This included integrating multi-task learning frameworks.

3. **Model Training:** Each model was independently trained using the prepared training and validation sets. Multi-task learning strategies were employed to improve generalization and robustness by allowing models to learn from multiple related tasks simultaneously.

4. **Hyperparameter Tuning:** Extensive hyperparameter tuning was conducted to optimize model performance, adjusting learning rates, batch sizes, and other critical parameters.

5. **Model Fusion:** Post-training, we combined the strengths of X^2-VLM, CCLM, Chinese CLIP, and OFA using a weighted voting mechanism. This ensured that the most reliable models had a greater influence on the final results, enhancing accuracy and robustness.

6. **Result Analysis:** We analyzed the experimental results, comparing the performance of different models across each subtask. The analysis highlighted the strengths and weaknesses of each model and assessed the impact of our innovative techniques, providing insights for further optimization and future research.

The experimental results are shown in Tables 1 through 5.

CCL 2024

Table 1: Experimental Results for Text Retrieval

Model	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Model Fusion	66.5	88.9	92.3
X^2-VLM	66.8	88.2	93.3
CCLM	59.9	85.4	91.3
OFA	58.2	80.3	87.9
CLIP	54.3	77.6	84.1

Table 2: Experimental Results for Image Retrieval

Model	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Model Fusion	48.5	78.9	87.3
X^2-VLM	48.7	77.4	87.0
CCLM	43.6	73.5	83.4
OFA	42.3	74.1	80.6
CLIP	45.1	73.9	86.3

Table 3: Experimental Results for Visual Question Answering

Model	**Accuracy (%)**
X^2-VLM	54.4
CCLM	58.3

Table 4: Experimental Results for Visual Grounding

Model	**IoU (%)**
X^2-VLM	55.7
CCLM	44.6
OFA	47.6

Table 5: Experimental Results for Visual Dialog

Model	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
X^2-VLM	29.3	42.5	49.4
CCLM	34.3	48.5	54.7

CCL 2024

3.3 Experiment Analysis

The X2VLM model excels in the image-text retrieval and visual grounding subtasks due to its unified architecture integrating vision, text, and fusion modules with Transformers. This enables effective cross-attention between text and vision features, associating visual concepts with text across images, videos, and annotations for comprehensive understanding. Its multi-grained training optimizes alignments and localizations, improving comprehension and localization. Additionally, the use of various loss functions enhances multimodal understanding and performance.

The Cross-View Language Modeling (CCLM) model performs strongly in VQA and VD subtasks due to its cross-lingual and cross-modal pre-training framework. Using a shared Transformer-based architecture, CCLM aligns image-text and text-translation pairs in a common semantic space. This approach maximizes mutual information between different data views through contrastive loss, matching loss, and conditional masked language modeling loss, enhancing sequence-level and token-level understanding. By sharing input-output formats and optimizing mutual information, CCLM effectively integrates visual and linguistic information, leading to superior performance in multimodal tasks.

Additionally, our text data augmentation strategy using ChatGPT further boosts model generalization and robustness by creating diverse textual inputs.

4 Conclusion

In the Chinese Vision-Language Understanding Evaluation task of the 23rd Chinese Computational Linguistics Conference, we designed and submitted a multi-model system to participate in several subtasks, including image-text retrieval, text retrieval, visual question answering, visual localization, and visual dialogue. We primarily utilized the X^2-VLM, CCLM, Chinese CLIP, and OFA models, leveraging their strengths to enhance performance in each subtask.

The experimental results demonstrate that the X^2-VLM model excelled in image-text retrieval and visual grounding tasks, particularly showcasing strong capabilities in aligning image and text features. The CCLM and OFA models also exhibited good performance in specific tasks. Through model optimization and data preprocessing, we successfully enhanced the system's performance in each subtask.

In future work, we plan to further optimize these model architectures and explore additional data augmentation and preprocessing methods to further improve model performance. Additionally, we aim to apply these models to a broader range of multimodal tasks to verify their applicability in diverse scenarios.

Overall, the outcomes of this competition underscore the potential and advantages of these multi-model systems in tackling complex multimodal tasks, offering valuable insights and guidance for future research and applications.

CCL 2024

Acknowledgements

Thank you to all the reviewers for their valuable suggestions, which have greatly improved the content of this paper.The work was supported by the National Natural Science Foundation of China under Grant 62276077, Guangdong Basic and Applied Basic Research Foundation (2024A1515011205), and Shenzhen College Stability Support Plan under Grants GXWD20220811170358002 and GXWD20220817123150002

Reference

Zeng Y, Zhang X, Li H, et al. 2023. *X^2-VLM: All-in-one pre-trained model for vision-language,* tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zeng Yan, et al. 2022. *Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training.* arXiv preprint arXiv:2206.00621 .

Radford A, Kim J W, Hallacy C, et al. 2021.*Learning transferable visual models from natural language supervision*International conference on machine learning. PMLR, 2021: 8748-8763.

Wang P, Yang A, Men R, et al. 2022 *Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework*International Conference on Machine Learning. PMLR, 2022: 23318-23340.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.2016. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. *An image is worth 16x16 words: Transformers for image recognition at scale. In International* Conference on Learning Representations, 2020.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu.2020. *Pixel-bert: Aligning image pixels with text by deep multi-modal transformers.* arXiv preprint arXiv:2004.00849

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. *Vilt: Vision-and-language transformer without convolution or region supervision.* In International Conference on Machine Learning, pages 5583–5594. PMLR, 2021.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021.*Align before fuse: Vision and language representation learning with momentum distillation.* Advances in Neural Information Processing Systems, 34, 2021.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. *Wenlan: Bridging vision and language by large-scale multi-modal pre-training.* arXiv preprint arXiv:2103.06561, 2021.

Hao Tan and Mohit Bansal.2019. *LXMERT: Learning cross-modality encoder representations from transformers.* In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514.

CCL 2024

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee.2019. *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. In Advances in Neural Information.Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019,NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 13–23, 2019.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. *Visualbert: A simple and performant baseline for vision and language*. arXiv preprint arXiv:1908.03557, 2019.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu.2020. *Large-scale adversarial training for vision-and-language representation learning*. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020.*Uniter: Universal image-text representation learning. In European conference on computer vision*, pages 104–120. Springer, 2020.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020.*Oscar: Object-semantics aligned pre-training for vision-language tasks*. In European Conference on Computer Vision, pages 121–137. Springer, 2020.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao.2021. *Vinvl: Revisiting visual representations in vision-language models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5579–5588, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014.*Microsoft coco: Common objects in context. In European conference on computer vision*, pages 740–755. Springer, 2014.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019.*Objects365: A large-scale, high-quality dataset for object detection*. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 8429–8438. IEEE, 2019. doi: 10.1109/ICCV.2019.00852.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset,Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al.2018. *The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale*. arXiv preprint arXiv:1811.00982, 2018.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.2017. *Visual genome: Connecting language and vision using crowdsourced dense image annotations*. International journal of computer vision, 123(1):32–73, 2017.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai,and Quoc V Le. 2021.*Finetuned language models are zero-shot learners*. arXiv preprint arXiv:2109.01652, 2021.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021.*Multitask prompted training enables zero-shot task generalization*. arXiv preprint arXiv:2110.08207, 2021.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021.*Simvlm: Simple visual language model pretraining with weak supervision.* ArXiv, abs/2108.10904, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021.*Florence: A new foundation model for computer vision.* ArXiv, abs/2111.11432, 2021.

CCL 2024

CCL24-Eval任务9总结报告：中文图文多模态理解评测

王宇轩[1]，刘议骏[2]，万志国[1]，车万翔[2]
[1]之江实验室，杭州，311121
[2]哈尔滨工业大学，哈尔滨，150001
yxwang@zhejianglab.com, yijunliu@ir.hit.edu.cn
wanzhiguo@zhejianglab.com, car@ir.hit.edu.cn

摘要

中文图文多模态理解评测任务旨在从多角度评价中文图文多模态预训练模型的图文多模态建模和理解能力。本任务共包括五个子任务：图片检索、文本检索、视觉问答、视觉定位和视觉对话，最终成绩根据这五个任务的得分综合计算。本文首先介绍了任务的背景和动机，然后从任务介绍、评价指标、比赛结果、参赛方法等方面介绍并展示了本次评测任务的相关信息。本次任务共有11支队伍报名参赛，其中3支队伍提交了结果。

关键词： 中文；多模态；图文检索；视觉问答；视觉定位；视觉对话

Overview of CCL24-Eval Task 9: Chinese Vision-Language Understanding

Yuxuan Wang[1], Yijun Liu[2], Zhiguo Wan[1], Wanxiang Che[2]
[1]Zhejiang Lab, Hangzhou, 311121
[2]Harbin Institute of Technology, Harbin, 150001
yxwang@zhejianglab.com, yijunliu@ir.hit.edu.cn
wanzhiguo@zhejianglab.com, car@ir.hit.edu.cn

Abstract

The Chinese Vision-Language Understanding Task aims to evaluate the vision-language modeling and understanding capabilities of Chinese vision-language pre-training models from multiple perspectives. This task includes five sub-tasks: image retrieval, text retrieval, visual question answering, visual grounding, and visual dialogue. The final score is calculated based on the combined results of these five tasks. This paper first introduces the background and motivation of the task, then presents and demonstrates relevant information about this task from aspects including task description, evaluation metrics, submitted results, and participant methods. A total of 11 teams registered to participate in this task. And 3 teams eventually submitted their results.

Keywords: Chinese , Multimodality , Image-text retrieval , Visual question answering , Visual grounding , Visual dialog

1 背景和动机

近年来，英文图文数据集经历了快速发展，从最基本的图像描述任务开始。继该领域的MS-COCO (Lin et al., 2014)和Flickr30K (Young et al., 2014)图片描述数据集之后，大量涵盖各种任务的英文图文数据集相继出现，这些任务包括视觉问答 (Antol et al., 2015; Goyal

第二十三届中国计算语言学大会论文集，第364页–第371页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

et al., 2017)、视觉推理 (Suhr et al., 2017; Suhr et al., 2019; Zellers et al., 2019)、视觉定位 (Kazemzadeh et al., 2014; Mao et al., 2016)、视觉蕴涵 (Xie et al., 2019)、视觉关系检测 (Plummer et al., 2015)和视觉对话visual dialogue (Das et al., 2017)等。这些英文图文数据集的出现对英文图文多模态预训练模型的评价体系建立起到了重要作用，也同时推动了该领域的发展。

在其他语言上，近年来也有不少工作尝试在这些数据集的基础上构建非英语的图文数据集。例如，MS-COCO数据集就被扩展到了德语、法语 (Rajendran et al., 2016)、日语 (Yoshikawa et al., 2017)和中文 (Li et al., 2019)上。这些语言上的图文数据集或者是直接将原始标注中的英语翻译成目标语言，或者是使用目标语言在MS-COCO的图片上重新进行标注。然而，无论通过上述哪一种方法构建的数据集，都使用了原始英文图文数据集中来自西方文化背景的图片。研究表明，使用这种包含文化偏置的数据会严重限制模型在很多其他语言和文化中的表现。(Stock and Cissé, 2018; DeVries et al., 2019; Liu et al., 2021) 同理，在这类包含文化偏置的图片基础上构建的中文图文数据集也无法客观准确地评价目前的中文图文预训练模型。

针对该问题，我们组织了本次中文图文多模态理解评测任务，从收集图片流程开始，严格控制图片为中国文化环境中具有代表性或日常生活常见的内容，并选择了5个重要且具有代表性的图文多模态理解任务：

- 图片检索（Image Retrieval）：基于给定的文本描述从若干候选中检索出对应图片。
- 文本检索（Text Retrieval）：基于给定的图片从若干候选中检索出对应的文本描述。
- 视觉问答（Visual Question Answering）：基于给定的图片用短语回答问题。
- 视觉定位（Visual Grounding）：基于给定的图片和文本描述找出图片中对应的实体。
- 视觉对话（Visual Dialog）：基于给定的图片和对话历史从若干候选中选出最合适的回复文本。

本次评测旨在通过上述5个子任务从图文表示对齐、图文理解和推理、图片细节理解和图片整体理解等多个角度对中文图文预训练模型进行评价。

2 评测任务及数据

2.1 图片类别及图片收集

本任务包含15大类、92小类的图片，具体图片类别如表 1所示。

保证数据集中的图片具有中文文化代表性，是本次评测在构建数据集时十分重要的一点。我们使用百度众包从互联网上采集图片，并且在培训过程中着重强调了收集的图片内容为中国文化环境中具有代表性或日常生活常见的。此外，为了保证后续任务具有足够挑战性，我们对每个类别的图片都采集了两个不同的子集。其中第一个子集中的图片必须包含该图片类别的至少2个实体，该子集用于后续视觉问答和视觉定位任务的标注，这是为了保证在视觉定位任务的标注过程中能让模型区分同一类别的不同实体，而非更简单的区分不同类别的实体。而第二个子集中的图片必须包含3到5个不同类别的实体，该子集用于后续视觉对话任务的标注，这是为了保证在视觉对话任务的标注中对话内容足够丰富。此外，这两个子集上都会进行图文检索任务的标注。

为了进一步确定收集到的图片的质量，我们对通过外包收集来的图片进行了二次筛查，以确保这些图片都具有中文文化代表性，并且各子类中的图片都符合对应子类的要求。通过二次筛查，我们过滤掉了大部分不符合要求的图片。为了进一步确保标注质量，在后续标注过程中，当标注人员发现图片不符合上述要求时，我们也允许标注人员跳过当前图片。

2.2 图文检索

图文检索包括两个子任务，分别是图片检索和文本检索，其中图片检索任务定义为给定一句文本，要求模型从候选图片集合中选出最相关的10张图片，并对它们进行排序。而文本检索任务定义为给定一张图片，要求模型从候选文本集合中选出最相关的10个文本，并对它们进行

第二十三届中国计算语言学大会论文集，第364页-第371页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

大类	图片类别
动物	大熊猫, 牛, 鱼, 狗, 马, 鸡, 鼠, 鸟, 人, 猫
食物	火锅, 米饭, 饺子, 面条, 包子
饮品	奶茶, 可乐, 牛奶, 茶, 粥, 酒
衣服	汉服, 唐装, 旗袍, 西装, T恤
植物	柳树, 银杏, 梧桐, 白桦, 松树, 菊花, 牡丹, 兰科, 莲, 百合
水果	荔枝, 山楂, 苹果, 哈密瓜, 龙眼
蔬菜	小白菜, 马铃薯, 大白菜, 胡萝卜, 花椰菜
农业	锄头, 犁, 耙, 镰刀, 担杖
工具	汤勺, 碗, 砧板, 筷子, 炒锅, 扇子, 菜刀, 锅铲
家具	电视, 桌子, 椅子, 冰箱, 灶台
运动	乒乓球, 篮球, 游泳, 足球, 跑步
庆典	舞狮, 龙舟, 国旗, 月饼, 春联, 花灯
教育	铅笔, 黑板, 毛笔, 粉笔, 原子笔, 剪刀
乐器	古筝, 二胡, 唢呐, 鼓, 琵琶
艺术	书法, 皮影, 剪纸, 秦始皇兵马俑, 鼎, 陶瓷

Table 1: 图片类别列表

排序。图文检索任务的数据样例如图1a所示。该任务主要评价的是模型的图文表示对齐能力。该任务在标注时，我们要求每个标注者对给定的图片写5句不同的文本进行描述，同时要求不同文本之间的重复应少于30%。

2.3 视觉问答

视觉问答任务定义为给定一张图片及一个与图片内容相关的问题，要求模型根据图片使用短语给出答案。视觉问答任务的数据样例如图1b所示。该任务主要评价的是模型的图文理解和简单推理能力。在数据集中，每张图片对应3个问题和答案对。该任务在标注时，我们要求每个标注者对给定的图片写3个问题，并用尽可能简洁的短语给出正确答案。

2.4 视觉定位

视觉定位任务定义为给定一张图片及一个描述图中实体的短语，要求模型在图中画出该实体对应的边界框（bounding box）。视觉定位任务的数据样例如图1c所示。该任务主要评价的是模型的图片细节理解和区分能力。为了使该任务更具挑战性，该任务数据集中每张图片都对应若干句描述同一类实体不同个体的文本。比如例子中的两个文本就分别描述了图中各不相同的两个皮影。这样的任务设置方式，显然比让模型分辨图片中的皮影和人这种不同类别的实体更具挑战性。该任务标注分为两个阶段，第一阶段要求标注者将图片中我们定义的92类实体都用边界框框出来。第二阶段，我们将标注者分为两组，其中第一组标注者为图片中与图片类别实体相同的每个实体写一句自然语言描述文本，用以将其和其他实体区分开。（例如，对于一张在狗类别中的图片，第一组标注者要为图中的每条狗都写一句描述，用以区分当前描述的狗和其他的狗。）而另一组标注者则只能看到图片和描述文本，这些标注者要根据文本在图中点出文本描述的实体。如果实体在正确的边界框内，则视为该标注正确，否则视为标注错误，要求其他标注者重新标注。

2.5 视觉对话

视觉对话任务定义为给定一张图片、一段对话历史和一个问题，要求模型根据对话历史和当前的问题从100个候选答案中选出可能性最高的10个答案并对它们进行排序。（这100个候选答案是参考Das等人 (2017)的工作，从所有答案中选出的）视觉对话任务的数据样例如图 1d所示。该任务主要评价的是模型的图片整体理解、对话历史理解、指代消解和文本生成的综合能力。视觉对话任务与视觉问答任务的区别主要有两点：一是视觉问答要求模型用尽可能简单的短语回答问题，而视觉对话则要求模型回复完整的句子；二是视觉问答一般是较为直观的问题，不涉及历史信息，而视觉对话任务还要求模型对历史对话信息有所理解，里面还涉及到指

(c) 2024 中国中文信息学会计算语言学专业委员会

1.桌子中间摆放着火锅

2.两种口味的火锅摆放在木质的桌子上

3.一个辣的和一个菌汤锅底的火锅放在桌上

4.火锅四周摆满了涮火锅用的蔬菜、肉、丸子等食材

5.桌子中间摆放着两个口味的火锅，周围的陶瓷碗里盛放着涮火锅用的食材

(a) 图文检索

Q: 龙舟划向什么方位?

A: 右方

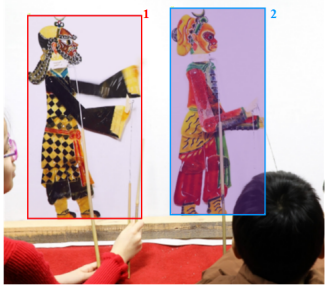
Q: 有几支队伍在划龙舟?

A: 5

Q: 大多数人的姿势是站立还是坐着?

A: 坐着

(b) 视觉问答



1.戴眼镜女孩手里拿着的皮影

2.短发男孩手里拿着的皮影

(c) 视觉定位

Caption: 蓝色桌垫上有许多食物

Q1: 桌上都有哪些食物?

A1: 食物中有鸡蛋、包子、小菜、馒头和粥

Q2: 桌上的粥是哪种粥?

A2: 桌上的粥是黑米粥

.....

Q10:桌面上的鸡蛋有几个?

A10:桌面上有两个鸡蛋

(d) 视觉对话

Figure 1: 评测中各子任务数据样例.

代消解的能力。显然，视觉对话是对模型各项能力的一个综合评价，相比视觉问答对于模型能力有更高的要求。该任务在标注时，我们将标注者分为两组，其中第一组能看到当前图片，而第二组只能看到一句描述当前图片的文本（来自图文检索标注）。第二组标注者要对第一组标注者就图片内容进行提问，以此尽可能想象出当前的图片，而第一组标注者需要根据图片据实回答。每张图片要求进行10轮问答。

2.6 数据统计

任务	训练集	开发集	测试集
图文检索	17,920	3,116	8,973
视觉问答	43,086	7,713	21,507
视觉定位	28,950	5,196	14,497
视觉对话	39,750	6,510	20,360

Table 2: 各子任务数据集样本数量统计

表2中列出了本次任务中各子任务数据集中样本数量的统计。其中图文检索任务中列出的是图片的数量，每张图片对应了5个文本描述。视觉对话列出的是对话轮次数量，该任务中每张图片对应10轮对话，我们将每轮对话及其历史作为一个样本。

3 评价指标

本次评测中，针对不同子任务，我们选择了不同的评价指标。具体来说，对于图片检索、文本检索及视觉对话，由于都是从候选集合中选出目标并进行排序，因此我们使用前1/前5/前10的召回率（$R@1/R@5/R@10$）作为评价指标，计算方式如下：

$$R@N = \frac{\text{正确结果在前}N\text{个出现的样本数}}{\text{总样本数}}$$

对于视觉问答任务，我们使用预测结果的正确率（Accuracy）作为评价指标，计算方式如下：

$$Accuracy = \frac{\text{预测正确的样本数}}{\text{总样本数}}$$

对于视觉定位任务，我们使用预测边界框和正确边界框的重叠度（Intersection over Union，IoU）作为评价指标，计算方式如下：

$$IoU = \frac{\text{正确边界框和预测边界框重叠部分面积}}{\text{正确边界框面积} + \text{预测边界框面积} - \text{二者重叠部分面积}}$$

最终，我们计算上述所有指标的宏平均值作为最终排名的依据。

4 提交结果

参赛队伍	TR			IR			VQA	VG	VD			AVG
	R@1	R@5	R@10	R@1	R@5	R@10	Acc	IoU	R@1	R@5	R@10	
江南大学	40.7	61.3	66.0	28.5	46.4	50.7	39.0	48.9	**30.1**	**44.8**	**51.9**	46.2
哈工大（深）	**43.8**	**69.4**	**78.4**	28.7	52.9	63.8	47.9	10.8	23.9	37.5	44.1	45.6
北语	41.9	68.9	78.1	**28.9**	**54.1**	**65.1**	**55.6**	**49.3**	29.5	43.8	50.7	**51.4**

Table 3: 本次评测所有提交队伍结果，其中TR表示文本检索任务，IR表示图片检索任务，VQA表示视觉问答任务，VG表示视觉定位任务，VD表示视觉对话任务，AVG表示各指标平均值，Acc表示正确率，IoU表示重叠度。参赛队伍中哈工大（深）表示哈尔滨工业大学（深圳），北语表示北京语言大学。

本次评测任务共有11支队伍报名参赛，由于任务较为困难，最终只有3支队伍提交了结果，分别是江南大学、哈尔滨工业大学（深圳）和北京语言大学。本次任务评测阶段采取盲测方式，即在测试阶段我们将没有答案的5个子任务的测试集问题及对应图片发给各支队伍，各支队伍使用各自训练好的模型预测出答案之后统一提交给评测组织方。然后由我们统一对各队伍提交的结果计算各子任务得分，并将所有指标的平均值作为最终排名的依据。上述三支队伍提交的结果见表3，各指标上最高的结果用加粗字体标出。可以看出，三支队伍各自在部分子任务上取得了最好成绩，其中，江南大学队伍在视觉对话子任务上取得最好成绩，哈尔滨工业大学（深圳）队伍在文本检索子任务上取得最好成绩，而北京语言大学队伍在图片检索、视觉问答和视觉定位三项子任务上都取得了最好成绩。最终，北京语言大学队伍凭借51.4的平均分数取得了本次评测的综合最好成绩。

同时，值得注意的是，三支参赛队伍在五个子任务上得到的结果的绝对值都比较低，尤其是其中的图片检索、视觉问答、视觉定位和视觉对话几个任务。这种结果一方面证实了该评测任务具有较大挑战性，另一方面也说明目前的图文预训练模型在面对中文图文理解问题时，仍然有巨大的进步空间。

5 方法概述

本节中我们将分别介绍三支队伍所使用的方法。

5.1 江南大学队伍

江南大学队伍采取了直接使用任务提供的训练集在X^2VLM (Zeng et al., 2022)预训练模型上针对每个子任务单独进行微调的方法。X^2VLM模型采用模块化架构，包括视觉编码器、文本编码器和多模态融合模块，所有模块基于Transformer架构。其特点是能够在预训练过程中同时学习多粒度的视觉-语言对齐和定位，支持图像-文本和视频-文本任务的统一预训练，并且具有高适应性，可以通过替换文本编码器适应不同语言或领域的任务。与CLIP (Radford et al., 2021)主要学习全局图像和文本特征不同，X^2VLM还学习对象和区域级别的细粒度特征对齐，展示了更强的多任务处理能力和灵活性。组织方提供的X^2VLM模型[1]是在中文的图文对数据上进行预训练得到的。

5.2 哈尔滨工业大学（深圳）队伍

而哈尔滨工业大学（深圳）队伍也采用了类似的微调策略，但他们同时还对比了多个支持中文的图文预训练模型上微调的结果，具体包括：

[1]https://github.com/zengyan-97/X2-VLM

(c) 2024 中国中文信息学会计算语言学专业委员会

- CCLM (Zeng et al., 2023)：该模型是一个跨视图语言建模框架，旨在统一跨语言和跨模态的预训练，其特点在于同时处理多模态数据（如图像-字幕对）和多语言数据（如平行句对），通过条件掩码语言建模、对比学习和匹配目标，最大化不同视图之间的互信息，从而将它们对齐到一个共同的语义空间。

- 中文CLIP (Yang et al., 2022)：该模型是一个专门针对中文的视觉-语言基础模型，通过两阶段预训练方法实现。在第一阶段中，模型采用锁定图像编码器，仅优化文本编码器的方式进行训练；在第二阶段，解锁图像编码器，进行对比学习，从而使整个模型适应中文数据集。

- OFA (Wang et al., 2022)：该模型是一种任务无关和模态无关的框架，旨在通过一个简单的序列到序列学习框架统一架构、任务和模态。OFA在预训练和微调阶段采用基于指令的学习，不需要为下游任务添加额外的任务特定层。OFA模型能够有效地转移到未见过的任务和领域，具有出色的零样本学习能力和域适应能力。

其中，中文CLIP模型只在图文检索任务中使用，而OFA模型只在图文检索和视觉定位任务中使用。根据开发集上的结果，他们最终选取X^2VLM模型用于图文检索、视觉定位任务，CCLM模型用于视觉问答、视觉对话任务。

5.3 北京语言大学队伍

北京语言大学队伍采用了拉近真实图片与答案导向图片的中文图文多模态理解增强方法。他们首先使用VisCPM-Paint模型 (Hu et al., 2023)的文生图功能，利用文本标注信息生成答案导向的图片，从而对训练数据进行扩充。然后，生成的图片与原始图片一起送入模型进行微调，对不同子任务设置不同的训练目标，拉进生成的图片与真实图片距离。

具体来说，在第一步文生图阶段，该队伍针对不同子任务使用不同的提示词进行生成。例如，对于图文检索任务，将每张图片对应的五个描述文本拼接起来作为提示词来生成和原始图片类似的图片。对于视觉问答任务，将每张图片对应的3个问题和答案组合，来生成符合所有问题、答案的图片。对于视觉对话任务，将每张图片对应的10轮问答中的答案进行组合，来生成符合对话的图片。

在第二步精调阶段，该队伍针对不同任务采用了不同的精调模型架构。针对图文检索任务，采用的精调模型通过共享参数的视觉编码器编码真实图片和答案导向图片，并使用余弦向量损失和KL散度损失分别拉近它们的表示和相似度矩阵。最后，将这些损失与原本的匹配损失和对比学习损失共同作为优化目标进行多模态微调。对于视觉问答和对话任务，该队伍计算真实图片和答案导向图片结果的下一个符号预测损失，并通过KL散度损失拉近它们输出解码器的概率分布。而对于视觉定位任务，则通过损失函数和KL损失缩小答案导向图片和原图的预测边界框与目标边界框之间的差异。

6 总结

本次任务针对目前中文上缺少全面的，不包含西方文化偏置图片的图文多模态评测数据集的问题，开展了包含图片检索、文本检索、视觉问答、视觉定位和视觉对话等五个子任务的中文图文多模态理解评测。本次评测吸引了11支来自学术界和工业界的队伍报名，但由于任务难度较大，最终只有3个队伍提交了结果。提交结果的队伍使用本任务提供的数据对多个现有的中文图文预训练模型进行了评测，同时还对基于文生图的数据增广方法进行了探索。从方法上来看，提交了结果的队伍主要方法集中在使用主办方提供的数据进行精调上。从最终结果来看，所有队伍的结果的绝对值都比较低，这一方面反映了本评测任务的挑战性，另一方面也说明目前中文图文预训练模型在实际的、没有西方文化偏置的图片上的性能仍然比较弱，还有较大进步空间。该结果也说明，解决文化偏置问题是未来中文图文预训练模型的一个重要发展方向。本次评测虽然包括了5个子任务，但在图片使用上实际不同任务之间是有比较多重叠部分的，即很多图片都有不止一种任务的标注。这种特性可以较好地支持多任务学习，但遗憾的是，在参赛队伍提交的方案中，我们没有发现对于不同任务之间关系及图文预训练模型中多任务学习方法的探索。我们认为在未来，这种图文多任务学习是一个值得探索的方向。

第二十三届中国计算语言学大会论文集，第364页-第371页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

参考文献

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *CoRR*, abs/2308.12038.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Trans. Multim.*, 21(9):2347–2360.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

(c) 2024 中国中文信息学会计算语言学专业委员会

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 171–181. The Association for Computational Linguistics.

Pierre Stock and Moustapha Cissé. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 504–519. Springer.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 217–223. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: contrastive vision-language pretraining in chinese. *CoRR*, abs/2211.01335.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 417–421. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *CoRR*, abs/2211.12402.

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5731–5746. Association for Computational Linguistics.

CCL 2024

System Report for CCL24-Eval Task 9: Bridging the Gap between Authentic and Answer-Guided Images for Chinese Vision-Language Understanding Enhancement

Feiyu Wang[1], Wenyu Guo[1], Dong Yu[1*], Chen Kang, Pengyuan Liu[1,2]
1.Faculty of Computer Science, Beijing Language and Culture University, Beijing, 100083
2.National Language Resources Monitoring and Research Center for Print Media, Beijing, 100083
wfy_0502@163.com, xk17guowenyu@126.com, yudong@blcu.edu.cn
kangchen@blcu.edu.cn, liupengyuan@pku.edu.cn

Abstract

The objective of the Chinese Vision-Language Understanding Evaluation (CVLUE) is to comprehensively assess the performance of Chinese vision-language multimodal pre-trained models in multimodal modeling and understanding across four tasks: Image-Text Retrieval, Visual Question Answering, Visual Grounding, and Visual Dialog. To enhance the models' performance across various multimodal tasks, this paper propose a multimodal information understanding enhancement method based on answer-guided images. Firstly, we propose task-specific methods for answer-guided image generation. Secondly, the authentic and answer-guided images are fed into the model for multimodal fine-tuning, respectively. Finally, training objectives are set for different tasks to minimize the gap between the answer-guided images and authentic images, thereby supervising the results produced by the authentic images utlizing answer-guided images. The experimental results demonstrate the effectiveness of the proposed method.

1 Introduction

The Chinese Vision-Language Understanding Evaluation (CVLUE) aims to assess Chinese vision-language multimodal pre-trained models from multiple perspectives, including Image-Text Retrieval (ITR), Visual Question Answering (VQA), Visual Grounding (VG), and Visual Dialog (VD). This comprehensive evaluation is designed to measure the multimodal modeling and understanding capabilities of these models. Exploring the diverse dimensions of Chinese multimodal pre-trained models not only refines modeling strategies and optimization algorithms, but also significantly enhances the models' ability in multimodal information comprehension and interaction. Researchers can also gain a deeper insight into their practical applications within real-world Chinese contexts.

To enhance the capabilities of Chinese vision-language multimodal pre-trained models across various multimodal tasks, we propose a method to strengthen Chinese multimodal comprehension by bridging the gap between authentic and answer-guided images. Firstly, we propose a method to generate answer-guided images for each task. Specifically, for the ITR task, we generate images according to image captions utilize Chinese text-to-image generation models. The synthetic images effectively highlight key textual information through attributes such as color, quantity, and orientation. For the VQA and VD tasks, we adopt an image generation approach based on questions and answers, integrating answer information into the image generation process. For the VG task, images with grounding information are used as answer-guided images. Secondly, both the answer-guided images and authentic images are fed into the model, respectively. Thirdly, we bridge the gap between authentic and answer-guided images during the training process, thereby enabling the answer-guided images to supervise and elevate the results yielded by authentic images. Experimental results across various tasks demonstrate the effectiveness of our proposed method in enhancing Chinese multimodal understanding.

*Corresponding author: Dong Yu.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 372-381, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 BackGround

2.1 X^2-VLM

X^2-VLM (Zeng et al., 2022a) represents a modular architecture multimodal vision-language model, which is trained through a unified framework to learn multi-grained visual-language alignments. This model enhances its comprehension of weak-correlated image-text pairs by associating visual concepts such as objects, regions, and images with text descriptions, facilitating the execution of various image-text tasks without the need for additional image annotations. Moreover, X^2-VLM exhibits commendable cross-lingual and cross-domain adaptability. By replacing text encoders tailored for specific languages or domains, it effectively adapts to image-text tasks across different languages and domains without the necessity for related pre-training, demonstrating its potential and flexibility in multimodal applications. Thus, we have selected X^2-VLM as the pre-trained model for this Chinese image-text multimodal evaluation.

2.2 VisCPM

VisCPM (Hu et al., 2023) is a family of open-source large multimodal models, which support multimodal conversational capabilities (VisCPM-Chat model) and text-to-image generation capabilities (VisCPM-Paint model) in both Chinese and English, achieving the state-of-the-art performance among Chinese open-source multimodal models. VisCPM is trained based on the large language model CPM-Bee with 10B parameters, fusing visual encoder (Muffin) and visual decoder (Diffusion-UNet) to support visual inputs and outputs. In this evaluation, we utilize the VisCPM-Paint model to generate images according to Chinese prompts related to ITR, VQA and VD tasks.

3 Participating System

This section provides a detailed description of the methods and strategies employed in our evaluation. Our method can be devided into two stages: the generation of answer-guided images and the multimodal fine-tuning for each task. To obtain answer-guided images, we design different prompts for text-to-image generation according to different tasks. We also set specific training objectives to accommodate the characteristics of each task during the multimodal fine-tuning stage.

3.1 Answer-guided Image Generation

In the ITR, VQA and VD tasks, we employ the VisCPM-Paint model for image generation and adopt distinct text-to-image generation prompt strategies for different tasks.

As shown in Table 1, for the ITR task, we concatenate the five different captions associated with each image in the training set to form the prompt for text-to-image generation. For the VQA task, as indicated in Table 2, we combine the questions and answers corresponding to each image in the training set using GPT-3.5 (Ouyang et al., 2022). The output sentences of GPT-3.5 are used as prompts for image generation. For the VD task, considering that the majority of images in the training set for the VD task overlap with those in the ITR task, we choose to use the images generated for the ITR task as the answer-guided images for the VD task, while the rest of the images that are not included in the ITR training set are generated using prompts crafted manually from dialogues.

For the VG task, given the particularity of the task, we don't employ the method of text-to-image generation to obtain answer-guided images. As shown in Figure 1, we annotate the bounding boxes of each phrase directly on the authentic images based on the provided bounding box data in the training set. The images are annotated on the authentic images then utilized as the answer-guided images for this task.

3.2 Multimodal Fine-tuning

For different tasks, we design specific training objectives for multimodal fine-tuning. Given the discrepancies in distribution, color, and orientation information between answer-guided images and authentic images, and considering that answer-guided images contain the visual information required by multimodal tasks from text-image pre-trained model, we propose a method to bridge the gap between

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 372-381, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

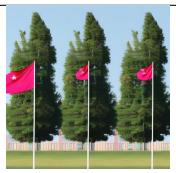
Task	Train Set	Text-to-image Generation Prompt	Generated images
ITR	"caption": ["旗杆上有三面红旗","天空下有三根旗杆，每根旗杆上都挂着一面红旗"， "三面红旗挂在旗杆上， 旗杆下面还有一些树"]	旗杆上有三面红旗，天空下有三根旗杆，每根旗杆上都挂着一面红旗，三面红旗挂在旗杆上，旗杆下面还有一些树。	
VQA	{"question": "有几只大熊猫？"， "answer": "2" }, {"question": "大熊猫在户外吗？"， "answer": "是"}, {"question": "周围有植物吗？"， "answer": "有" }	图中共有两只大熊猫，大熊猫在户外，周围有植物。	
VD	"dialogues":[{"question": "这个人坐在海边做什么？"， "answer": "这个人坐在海边在欣赏风景、放松。"}, {"question": "这个人的表情是怎样的？"， "answer": "看不到表情但是她很放松。"}, {"question": "这个人坐在海边身边有什么？"， "answer": "身边有海浪，桌凳，酒水等其他景观。"} {"question": "这个人是男人还是女人？"， "answer": "是一个很漂亮的女人。"}]	一个很漂亮的女人坐在海边欣赏风景，身边有海浪，桌凳，酒水等其他景观。虽然看不到表情但是她很放松。	

Table 1: Prompt examples for text-to-image generation.

Prompt	请你把下面的问答转化为陈述句：有几只大熊猫？3；有两只熊猫是抱在一起的吗？是；大熊猫在人群的哪边？前边。 图中共有三只大熊猫，其中有两只是抱在一起的，而这些大熊猫位于人群的前边。 请你把下面的问答转化为陈述句：有几只大熊猫？2；大熊猫在户外吗？是；周围有植物吗？有。
GPT-3.5	图中共有两只大熊猫，大熊猫在户外，周围有植物。

Table 2: Declarative sentence generation for the VQA task.

authentic images and answer-guided images during the training process through targeted training objectives. This approach enables the model to learn from answer-guided images during the training process and generate more qualified results during the inference process without answer-guided images.

For the input text T, authentic image X, and answer-guided image Y, we employ the text encoder to encode the text T into text embedding $\mathbf{t}$, and we employ a shared-parameter visual encoder to encode the authentic image X and the answer-guided image Y into visual embedding $\mathbf{x}$ and $\mathbf{y}$ respectively. We then define the similarity between the authentic image and the text, as well as the similarity between the answer-guided image and the text, as follows:

$$s(X,T) = g_x(\mathbf{x}_{\text{cls}})^\top g_t(\mathbf{t}_{\text{cls}}), \tag{1}$$

$$s(Y,T) = g_y(\mathbf{y}_{\text{cls}})^\top g_t(\mathbf{t}_{\text{cls}}), \tag{2}$$

where $\mathbf{t}_{\text{cls}}$ is the output [CLS] embedding of the text encoder, and $\mathbf{x}_{\text{cls}}$ and $\mathbf{y}_{\text{cls}}$ are the output [CLS] embedding of the visual encoder. g_t, g_x and g_y are transformations that map the [CLS] embeddings to normalized lower-dimensional representations. Based on it, when the batch size is N, we calculate the in-batch authentic image-to-text and text-to-image similarity as:

$$p^{\text{x2t}}(X) = \frac{\exp(s(X,T)/\tau)}{\sum_{i=1}^{N}\exp(s(X,T^i)/\tau)}, \tag{3}$$

$$p^{\text{t2x}}(T) = \frac{\exp(s(X,T)/\tau)}{\sum_{i=1}^{N}\exp(s(X^i,T)/\tau)}, \tag{4}$$

```
{
  "ref_id": 24577,
  "image": "images/train/A/4-dog/A-4-1.png",
  "text": "离货架最近偏白色的狗",
  "width": 837,
  "height": 551,
  "bbox": [
    474.8822507242,
    176.489921287,
    152.690886778,
    176.552619846
  ]
},
```


(a)visual grounding training set

(b)authentic image

(c)answer-guided image

Figure 1: Answer-guided image obtaining for the VG task.

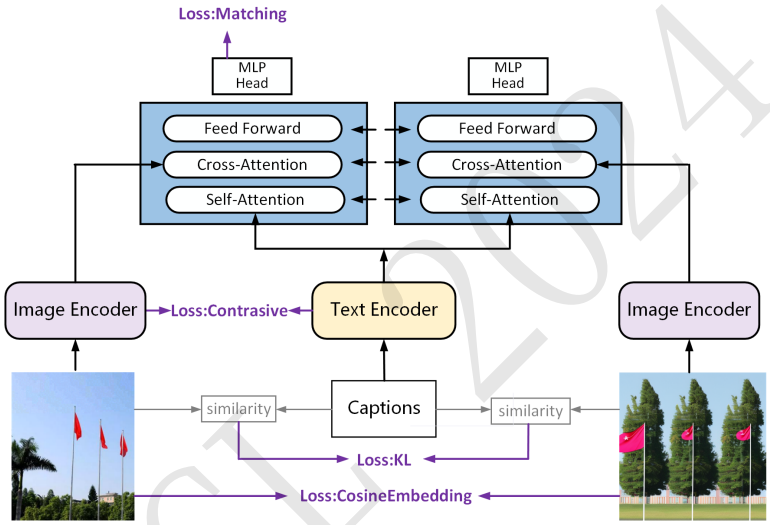


Figure 2: The model architecture of the ITR task.

Similarly, the answer-guided image-to-text similarity is defined as follows:

$$p^{\mathrm{y2t}}(Y) = \frac{\exp(s(Y,T)/\tau)}{\sum_{i=1}^{N}\exp(s(Y,T^i)/\tau)}, \tag{5}$$

where τ is a learnable temperature parameter.

For the ITR task, we use the model architecture in Figure 2. Following previous works (Zeng et al., 2022b; Zeng et al., 2023), in order to align authentic images with texts, we employ contrastive loss as the training objective during fine-tuning. Let $u^{\mathrm{x2t}}(X)$ and $u^{\mathrm{t2x}}(T)$ denote the ground-truth one-hot similarity, and the contrastive loss is defined as the cross-entropy H between $\mathbf{p}$ and $\mathbf{u}$:

$$\mathcal{L}_{\mathrm{cl}} = \frac{1}{2}\mathbb{E}_{X,T\sim D}\left[H(\mathbf{u}^{\mathrm{x2t}}(X), \mathbf{p}^{\mathrm{x2t}}(X)) + H(\mathbf{u}^{\mathrm{t2x}}(T), \mathbf{p}^{\mathrm{t2x}}(T))\right], \tag{6}$$

We also utilize the matching loss to ascertain the alignment between an image and its corresponding text:

$$\mathcal{L}_{\mathrm{match}} = \mathbb{E}_{X,T\sim D}\left[H(\mathbf{u}_{\mathrm{match}}, \mathbf{p}_{\mathrm{match}}(X,T))\right], \tag{7}$$

CCL 2024

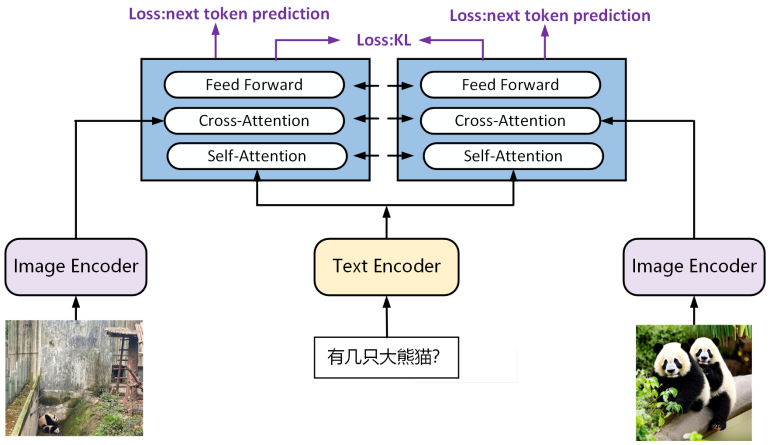


Figure 3: The model architecture of the VQA and VD task.

where $\mathbf{p}_{\text{match}}$ denotes the matching probability of image-text pair predicted by the model, $\mathbf{u}_{\text{match}}$ is a 2-dimensional one-hot vector representing the ground-truth label.

Innovatively, we employ cosine embedding loss and Kullback-Leibler (KL) divergence loss to eliminate the discrepancies between authentic and answer-guided images. Firstly, we use the cosine embedding loss to eliminate the visual representation gap. The loss is calculated as follows:

$$\mathcal{L}_{\text{cos}} = \text{CosineEmbeddingLoss}\left[g_y(\mathbf{y}_{\text{cls}}) \| g_x(\mathbf{x}_{\text{cls}})\right], \tag{8}$$

Secondly, because similarity matrices play a key role in calculating the contrastive loss and matching loss, inspired by previous works (Fang and Feng, 2023; Zhou and Long, 2023; Guo et al., 2023; Zhang et al., 2023; Fang et al., 2022), we introduce the Kullback-Leibler (KL) divergence loss to bridge the gap between the authentic and answer-guided image-to-text similarity matrices:

$$\mathcal{L}_1 = \text{KL}\left[p^{\text{y2t}}(Y) \| p^{\text{x2t}}(X)\right], \tag{9}$$

Finally, we employ the contrastive loss, the matching loss, the cosine embeeding loss and the KL divergence loss as combined objectives for optimization during the multimodal fine-tuning stage:

$$\mathcal{L}_{\text{itr}} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{cos}} + \gamma\mathcal{L}_1, \tag{10}$$

where λ and γ are hyperparameters that control the contribution of the cosine embedding loss and the KL divergence loss.

For the VQA and VD tasks, as proposed in Figure 3, we utilize the next token prediction loss to train our model. Specifically, we denote the correct answer sentence as $u = (u_1, ..., u_M)$. The loss functions of the authentic and answer-guided images are calculated respectively:

$$\mathcal{L}_{\text{x}} = -\sum_{j=1}^{M} \log p(u_j \mid u_{<j}, t, x), \tag{11}$$

$$\mathcal{L}_{\text{y}} = -\sum_{j=1}^{M} \log p(u_j \mid u_{<j}, t, y), \tag{12}$$

Innovatively, because the prediction probabilities have a key impact on calculating the next token prediction loss, we utilize KL divergence loss to enhance the prediction consistency produced by both types of images at the decoder side:

$$\mathcal{L}_2 = \sum_{j=1}^{M} \text{KL}\left[\mathbf{p}(u_j|u_{<j}, t, y) \| \mathbf{p}(u_j|u_{<j}, t, x)\right], \tag{13}$$

Finally, the training objective of the VQA and VD task can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{x}} + \mathcal{L}_{\text{y}} + \mathcal{L}_2, \tag{14}$$

For the VG task, the fine-tuning model architecture is similar to the VQA or VD task. The bounding box of the given entity i is defined as $\mathbf{b}^i = (lx, ly, w, h)$. The authentic image X and answer-guided image Y are sent into the model to predict the bounding box of the entity, respectively. The predicted bounding boxes are as follows:

$$\hat{\mathbf{b}}_x^i(X, T^i) = \text{Sigmoid}(\text{MLP}(\mathbf{cx}_{\text{cls}}^i)), \tag{15}$$

$$\hat{\mathbf{b}}_y^i(Y, T^i) = \text{Sigmoid}(\text{MLP}(\mathbf{cy}_{\text{cls}}^i)), \tag{16}$$

where Sigmoid is for normalization, MLP denotes multi-layer perceptron, $\mathbf{cx}_{\text{cls}}^i$ is the [CLS] embedding of the fusion module given the features of X (the authentic image) and T (the description of the entity), and $\mathbf{cy}_{\text{cls}}^i$ is obtained the same way from Y (answer-guided image) and T.

Following the previous work (Zeng et al., 2022b), we employ the same loss function in the pre-training stage to minimize the discrepancy between the predicted bounding box from the authentic image and the target bounding box:

$$\mathcal{L}_{\text{bbox}} = \mathbb{E}_{(X,T^i)\sim D}\left[L_{\text{iou}}(\mathbf{b}^i, \hat{\mathbf{b}}^i) + \|\mathbf{b}^i - \hat{\mathbf{b}}^i\|_1\right], \tag{17}$$

We utilize the KL divergence loss to minimize the distance between the predicted bounding boxes generated from authentic and answer-guided images:

$$\mathcal{L}_3 = \text{KL}\left[\hat{\mathbf{b}}_y^i(Y, T^i) \| \hat{\mathbf{b}}_y^i(X, T^i)\right], \tag{18}$$

Finally, the training objective of the VG task is as follows:

$$\mathcal{L}_{\text{vg}} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_3. \tag{19}$$

4 Experiment

Dataset We conduct experiments on the dataset provided by the organizer[0] for both fine-tuning and testing stages. The dataset includes 15 major categories and 92 subcategories of images. The collection of images is carried out manually according to the categories, and there is a strict requirement that the content of the images must be representative of the Chinese cultural environment or commonly seen in daily life.

Pre-trained Models We utilize the CCLM-X^2VLM-base[1] to initialize our model. We employ BEiT-2 (Peng et al., 2022) as our image encoder and XLM-RoBERTa-base (Conneau et al., 2020) as our text encoder. Additionally, for tasks requiring image-text generation, we employ the VisCPM-Paint model[2] based on the specified image-text generation prompts.

Systems Settings We set the hyperparameters λ and γ introduced in the ITR task to 0.5, while the rest of parameters follow the parameters set in the baseline model provided by the organizer. In terms of computing resources, we fine-tune the model on 2 V100 for the ITR task and on 4 A100 for tasks involving VQA, VD, and VG during the fine-tuning phrase. During the testing phrase, all tasks are tested on 4 A100 to obtain results.

We utilize the same system setting in the baseline system and our model. The answer-guided images and consistency training objects are removed in the baseline system. The experimental results on the validation datasets are shown in Table 3. By adopting the answer-guided images and our proposed

[0] https://github.com/WangYuxuan93/CVLUE/tree/main

[1] https://lf-robot-opensource.bytetos.com/obj/lab-robot-public/x2vlm_ckpts_2release/cclm_x2vlm_base.th

[2] https://huggingface.co/openbmb/VisCPM-Paint/blob/main/pytorch_model.bin

CCL 2024

Task	TR			IR			VQA	VD			VG
Metrics	R@1	R@5	R@10	R@1	R@5	R@10	ACC	R@1	R@5	R@10	IoU
Baseline	56.5	83.7	89.8	39.7	67.7	78.4	52.4	28.6	41.2	47.2	48.6
OURS	**57.0**	**83.9**	**90.4**	**39.8**	**68.0**	**78.8**	**53.6**	**28.6**	**41.4**	**47.8**	**49.3**

Table 3: The experimental results on the validation dataset.

Task	TR			IR		
Metrics	R@1	R@5	R@10	R@1	R@5	R@10
OURS	**57.0**	**83.9**	**90.4**	39.8	**68.0**	**78.8**
-remove KL	55.7	82.1	89.9	38.9	67.0	77.5
-remove COSINE	55.9	82.6	89.6	39.5	67.2	78.2
-replace COSINE with L2	56.6	83.0	90.0	**40.0**	67.6	78.2

Table 4: Ablation study on different training objectives in the ITR task.

Task	TR			IR			VQA	VD			VG
Metrics	R@1	R@5	R@10	R@1	R@5	R@10	ACC	R@1	R@5	R@10	IoU
OURS	**57.0**	**83.9**	**90.4**	**39.8**	**68.0**	**78.8**	**53.6**	**28.6**	**41.4**	**47.8**	**49.3**
-NOISE	52.7	76.9	84.4	39.7	67.2	78.3	52.2	27.4	40.8	47.0	48.9
-RANDOM	53.2	77.9	85.0	39.3	67.4	78.0	52.5	27.4	40.2	46.1	48.4

Table 5: Ablation study on different answer-guided images in all tasks.

consistency training objectives during the fine-tuning phase, the performance across various tasks shows a certain degree of improvement. The improvements show the effectiveness of the answer-guided images. The proposed training objects in bridging the gap between authentic and answer-guided images are able to boost the comprehension of Chinese multimodal contexts.

Ablation Study To further prove the effectiveness of our proposed methods, we conduct the following set of ablation experiments: 1) Ablation study on training objectives in the ITR task; 2) Ablation study on answer-guided images.

As shown in Table 4, we conduct studies on the training objectives in the ITR task to assess the impact of different loss functions on model performance, which involves the removal of the KL divergence loss or the cosine embedding loss, and the substitution of the cosine embedding loss with L2 loss. The drop of results indicates the effectiveness of our proposed training objectives, it also proves the effectiveness of our proposed training method in improving prediction consistency and mitigating the representation disparity.

As shown in Table 5, we conduct studies to assess the effectiveness of answer-guided images generated using our method in all tasks. We compare our model's performance with two regularization methods: NOISE and RANDOM. NOISE means using noise vectors as the answer-guided image representations. RANDOM means shuffling the correspondences between answer-guided images and textual queries in the training set. We observe a obvious decline using two regularization methods, which shows that the semantic information in the answer-guided images play a key role in our proposed method.

5 Related Work

Multimodal Pre-training Mutlimodal pre-training research includes the acquisition and clean of large-scale multi-modal data and the design of network architectures and pre-training objectives and so on. We focus on the design of pre-training objectives. CLIP (Radford et al., 2021) is trained on contrastive learning loss, which is widely used in dual-modality. Unicoder-VL (Li et al., 2020) utilizes the visual-linguistic matching loss to extract the positive and negative image-sentence pairs and predict

whether the given sample pairs are aligned or not. Unicoder-VL also uses the masked object classification loss to predict the object category of the masked image regions. UNITER (Chen et al., 2020) uses the word-region alignment loss which targets at explicitly achieves the fine-grained alignment between the multimodal inputs. E2E-VLP (Xu et al., 2021) use the image-text generation model to generate text based on a given image. Ling (Ling et al., 2022) uses the multimodal sentiment prediction loss to enhance the pre-trained models by capturing the subjective information from vision-language inputs. Image-conditioned denoising autoencoding is adopted in XGPT (Xia et al., 2020) to align the underlying image-text using an attention matrix. LXMERT (Tan and Bansal, 2019) uses the masked object regression loss to regress the masked feature or image regions. In our method, we use the contrastive loss, the image-text matching loss, the next token prediction loss and other task-specific prediction losses to train our model.

Chinese Text-to-image Generation The mainstream Chinese diffusion image generation models are derived from further training based on stable-diffusion (Rombach et al., 2022). Some researchers replace the CLIP text encoder with a bilingual encoder or Chinese encoder Taiyi-CLIP (Wang et al., 2022), Chinese-CLIP (Yang et al., 2022), and Alt-CLIP (Chen et al., 2023b), followed by pre-training for text-image matching on a Chinese text-image dataset. Some researchers train on a Chinese text-image dataset for text-to-image generation and obtain the Chinese version of the diffusion image generation model Taiyi-diffusion (Wang et al., 2022) and Alt-diffusion (Ye et al., 2024). ERNIE-ViLG 2.0 (Feng et al., 2023) embarked on training a Chinese diffusion model from scratch using Chinese image-text pairs. In the era of LLMs, PaLI (Chen et al., 2023a) develops a 17B multilingual language-image model based on 10B image-text pairs spanning 100 languages. MultiFusion (Bellagente et al., 2023) discovers that the multilingual language model can help cross-lingual transfer in text-to-image generation. We use the VisCPM (Hu et al., 2023) model, demonstrating that the zero-shot transfer performance of multilingual multimodal models can surpass that of models trained on Chinese multimodal data.

6 Conclusion

In the Chinese Vision-Language Understanding Evaluation task, we propose a system that enhances Chinese text-image multimodal understanding by bridging the gap between authentic images and answer-guided images. Firstly, we propose an image generation module, utilizing the text-to-image generation model or answer information from the train set to generate answer-guided images for different tasks. Secondly, we send the authentic and answer-guided images into the model respectively during the fine-tuning stage. Thirdly, we design specific training objectives for different tasks to encourage the representation or prediction consistency between the two images. Our proposed method improves the performance of our model over the baseline model across all the tasks on the validation set.

Acknowledgements

This work is funded by the Humanity and Social Science Youth foundation of Ministry of Education (23YJAZH184) and the Fundamental Research Funds for the Central Universities in BLCU (No.21PT04).

References

Marco Bellagente, Manuel Brack, Hannah Teufel, Felix Friedrich, Björn Deiseroth, Constantin Eichenberg, Andrew Dai, Robert Baldock, Souradeep Nanda, Koen Oostermeijer, Andrés Felipe Cruz-Salinas, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. 2023. Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023a. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023b. Altclip: Altering the language encoder in CLIP for extended language capabilities. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8666–8682. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15864–15881. Association for Computational Linguistics.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao Tian, Hua Wu, and Haifeng Wang. 2023. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10135–10145. IEEE.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023. Bridging the gap between synthetic and authentic images for multimodal machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2863–2874. Association for Computational Linguistics.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2149–2159. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

CCL 2024

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, and Ming Zhou. 2020. Xgpt: Cross-modal generative pre-training for image captioning.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 503–513. Association for Computational Linguistics.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: contrastive vision-language pretraining in chinese. *CoRR*, abs/2211.01335.

Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. 2024. Altdiffusion: A multilingual text-to-image diffusion model. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 6648–6656. AAAI Press.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022a. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *CoRR*, abs/2211.12402.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022b. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *CoRR*, abs/2211.12402.

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5731–5746. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Yucheng Zhou and Guodong Long. 2023. Improving cross-modal alignment for text-guided image inpainting. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3445–3456, Dubrovnik, Croatia, May. Association for Computational Linguistics.

CCL 2024

CCL24-Eval 任务10系统报告：维沃手语数字人翻译系统

何俊远，刘鑫，杨牧融，李小龙，黄旭铭，滕飞，陈晓昕，付凡
维沃移动通信有限公司
{hejunyuan, liuxin.rgzn, murong.yang, xiaolong.xlli}@vivo.com
{huangxuming, fei.teng, xiaoxin.chen, fan.fu}@vivo.com

摘要

本文介绍了我们在第二十四届中国计算语言学大会手语数字人翻译质量评测中提交的参赛系统。本次评测任务旨在评测手语数字人将汉语翻译成中国手语方面的自然性和准确性。本文介绍的手语数字人翻译系统首先通过手语翻译算法将汉语文本翻译成手语文本，然后将手语文本对应的手语动作单元运用动作融合算法合成为自然、完整的手语数字人动作，同时借助面部驱动算法将口型、表情等非语言元素自然地融入手语合成中，实现带微表情的和唇形同步的手语数字人。最终，我们在官方手语数字人翻译质量的人工评测集上取得了3.513的综合评分，获得了该任务第一名的成绩[1]。

关键词： 手语数字人；手语翻译；动作融合；唇形同步

System Report for CCL24-Eval Task 10: vivo Sign Language Avatar Translation System

Junyuan He, Xin Liu, Murong Yang, Xiaolong Li, Xuming Huang,
Fei Teng, Xiaoxin Chen, Fan Fu
vivo Mobile Communication Co., Ltd
{hejunyuan, liuxin.rgzn, murong.yang, xiaolong.xlli}@vivo.com
{huangxuming, fei.teng, xiaoxin.chen, fan.fu}@vivo.com

Abstract

This paper introduces the competition system we submitted for the sign language avatar translation quality evaluation at the 24th China National Conference on Computational Linguistics. The goal of the evaluation task was to assess the naturalness and accuracy of the sign language avatars in translating Chinese into Chinese Sign Language. The sign language avatar translation system described in this paper first translates Chinese text into sign language text using sign language translation algorithms, then synthesizes the corresponding sign language action units into natural, complete sign language avatar actions using action fusion algorithms, and naturally non-verbal elements such as lip shapes and facial expressions into the sign language synthesis with the help of facial driving algorithms, achieving sign language avatar figures with nuanced facial expressions and synchronized lip shapes. Ultimately, our system achieved a comprehensive score of 3.513 in the official sign language avatar translation quality manual evaluation test set and won first place in this task.

Keywords: Sign Language Avatar , Sign Language Translation , Action Fusion , Lip Sync

[1]https://github.com/ann-yuan/QESLAT-2024

(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

手语数字人（Sign Language Avatars，SLA）通过模拟手语动作来实现实时的手语转译，是当前小语种自然语言处理的重要任务之一，它可有效克服听障人士面临的交流障碍，提高该群体的社会参与度和沟通效率。得益于一些手语数据集的开源，比如德国手语数据集RWTH-PHOENIX-Weather 2014(Koller et al., 2015)、美国手语数据集American Sign Language Lexicon Video Dataset(Athitsos et al., 2008)，国内外提出了一些手语生成（Sign Language Production，SLP）模型(Stoll et al., 2018; Saunders et al., 2020; Stoll et al., 2020; Huang et al., 2021)来将手语文本翻译成连续的手语视频流。尽管手语生成技术在全球范围内得到了发展和应用，但特定于中国手语的研究进展却因缺乏专门的数据资源而显得滞后，国内规模较大的手语翻译数据集有CSL-Daily(Zhou et al., 2021)，这种情况也限制了中国手语翻译系统的创新与实现。

当前实现SLA模拟手语动作的核心步骤通常包括汉语转手语、手语合成和通过渲染引擎来获得最终的手语数字人视频序列。由于手语拥有独特的规则和语法体系，这使得汉语与手语之间的相互转译变得极为复杂。此外，手语合成的关键点包括获取手语句子中每个Gloss[2]对应的姿势、关节旋转角度等信息、为相邻的Gloss生成自然的过渡动作和同步展示口型和表情这些面部特征。其中中国手语常用的Gloss所对应的手势、关节旋转角度等信息可以通过动作捕捉技术来获得。虽然这种方法的成本依旧很高，但其优势在于不需要针对特定的手语句子进行数据采集，具有良好的扩展性。此外，采用动作捕捉技术还能确保最终通过渲染引擎合成的SLA效果达到较高的水准。

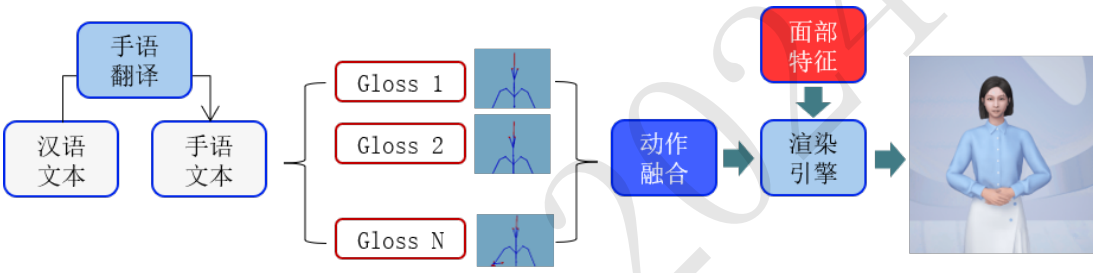


Figure 1: 手语数字人系统流程图

本次评测任务针对手语数字人将汉语翻译成中国手语的语法准确性、自然性、可读性以及文化适应性进行评估。如图1所示，为了保证SLA模拟手语动作的效果，在本次的评测任务中，我们设计了一个包含了手语翻译、动作融合和面部驱动三个主要模块的手语数字人系统。其中手语翻译模块参考了开源的多语言语言模型（Multilingual Random Aligned Substitution Pre-training，mRASP）(Lin et al., 2020)，通过预训练mRASP以及数据增强方法来实现汉语到手语的精准翻译。而动作融合模块借鉴Slot(2007)的工作提出了一种多帧插值平滑算法来使得手语Gloss之间的动作过渡得更加自然流畅，并参考Saunders et al.(2022)的思路设计了一个基于Transformer(Vaswani et al., 2017)的数字人手语合成速度调节模型来改变不同语境下手语Gloss的速度。最后，设计了一个面部驱动模块来把口型和表情等非语言信息融入到手语合成中，提高了手语合成的可懂率。实验结果表明，我们的方法可有效提高SLA模拟中国手语动作的准确性和自然性。

在下文中，我们先后在第二节和第三节介绍参赛系统所涉及到的手语翻译算法和动作融合算法，随后在第四节介绍面部驱动算法。介绍完我们的参赛系统后，我们在第五节和第六节给出实验结果与分析。最后，我们在第七节进行总结并展望未来的研究工作。

2 手语翻译算法

本节主要介绍我们的手语翻译算法。手语翻译是低资源翻译任务，平行语料极度稀缺，标注成本高。我们使用回译(Sun et al., 2020)的方法生成伪双语平行语料，缓解语料稀缺的问题。

[2]类似于汉语词汇，Gloss是手语词汇的唯一标识，如：家-家庭-房子-房，父亲-爸爸，水①，水②。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

另一挑战是手语相比汉语更加精炼省略得多，主要是因为手语利用空间特征传达信息。因此，虽然手语的词汇和句子数量较少，但其精炼的细节都在空间特征里，这使得手语和汉语的翻译任务更加复杂。同时由于手语词汇有限，在翻译时针对手语动作需要结合上下文进行转译。基于此我们设计了汉语-手语翻译算法的数据策略和模型策略：

- 使用预训练翻译模型提升低资源翻译任务翻译效果；
- 基于双语平行数据词对齐概率构建中文词与手语词的映射关系；
- 归纳手语语法规则，生成伪平行语料缓解语料稀缺问题；
- 构建模型对手语相同打法候选词进行排序，提升翻译质量。

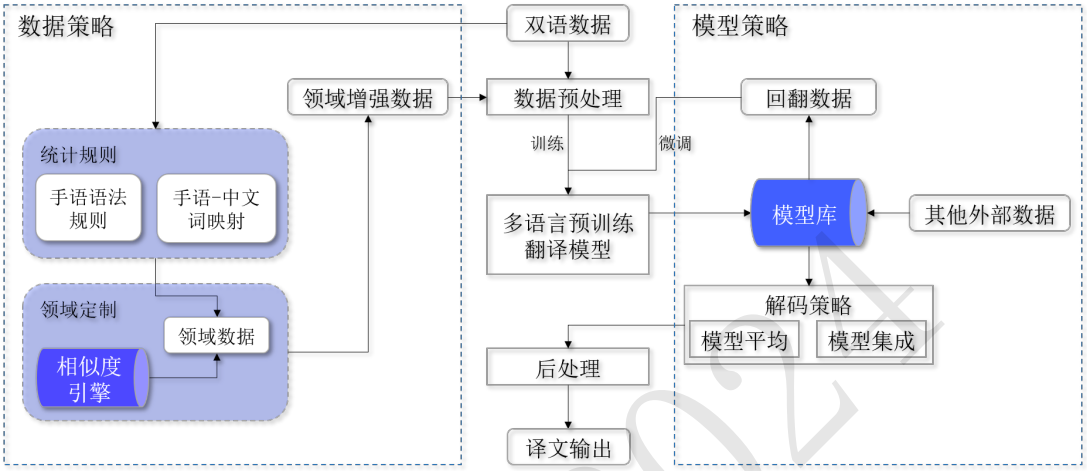


Figure 2: 手语翻译算法方案

本节提到的手语翻译算法使用的是多语言语言模型（mRASP）(Lin et al., 2020)，加入低资源的手语平行语料进行预训练，以便于利用资源较多的语种的先验知识来提升在手语方向上的翻译效果，并基于此预训练模型产生大量伪标签数据作扩充数据，以此训练回翻模型，然后通过迭代的方式持续提升模型性能。图2为我们的手语翻译算法方案。

具体方案的实现上，双语语料是采用了100万语料数据，但数量远远不够支撑模型训练，我们制定了数据策略和模型策略来增强。数据策略指的是从手语和汉语的语法规则以及一些固定的映射关系入手，统计汇总大量规则，基于规则可以用汉语句子构造一些简单的手语Gloss。领域定制指的是如果某些领域的数据特别少，可以定向化召回一些相关领域的示例，构造手语Gloss。模型策略就是利用预训练模型的能力，用100万数据微调之后，通过回翻的手段（汉语转Gloss）来做数据增强。

最终，我们的手语翻译算法在自测集上达到了手语词可懂率：汉语-手语87%，手语-汉语84%；通过句子纠错和上下文语义实现句子词准率70.5%。

3 手语数字人动作融合算法

手语合成的核心是保证语义传达准确，同时手语数字人的动作自然不生硬，贴近真人。这就要求手语句子中每个Gloss的核心动作完整且没有冗余动作，同时相邻Gloss对应的动作衔接自然。此外，不同语境下的同一个或不同的手语动作都要有节奏感，贴近真实的用户场景。为了实现这些目标，必须开发合适的手语合成算法来确保手语动作细节处的平滑过渡和手语动作节奏的恰当调整。本节将介绍手语数字人系统所涉及的动作融合算法，包括多帧插值平滑算法和基于Transformer(Vaswani et al., 2017)的数字人手语合成速度调节算法。

3.1 多帧插值平滑算法

动作混合（Motion Blending）通常在动画、游戏开发和电影制作中使用，指的是将两个或多个动画片段平滑地过渡合并，以创造出一个连贯且自然的动作序列。Slot(2007)提出了一种通

(c) 2024 中国中文信息学会计算语言学专业委员会

用的方法来混合两种动作使这两种动作之间的变化尽可能逼真，即利用时间扭曲自动确定两种动作之间的时序，并应用动作对齐来控制动作的方向。我们借鉴该方法提出了一种多帧插值平滑算法来为手语句子中相邻的Gloss生成连贯且自然的过渡动作。

构建手语动作库：针对每一个中国手语常用的Gloss，先通过动作捕捉技术从专业的手语老师中收集到该Gloss每一帧的姿势、关节旋转角度等信息。接着人工删除每个Gloss的冗余帧，即完整保留每个Gloss帧序列的核心动作和适当保留Gloss帧序列前后的部分过渡动作。最后这些处理好的Gloss帧序列构成一个完整的手语动作库。

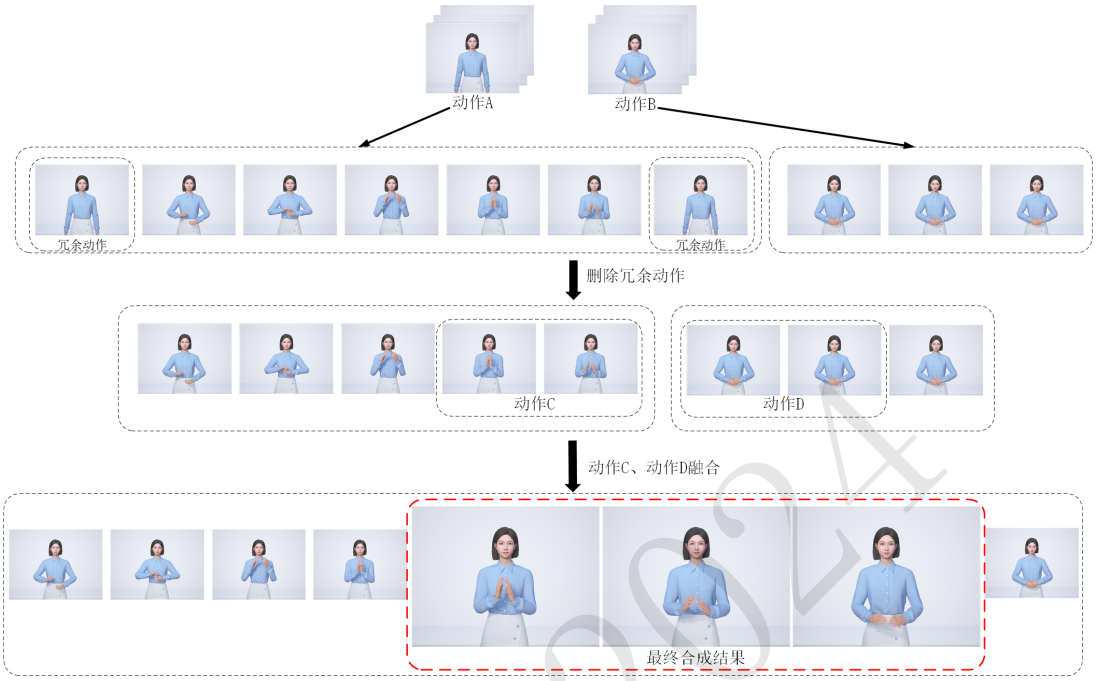


Figure 3: 多帧插值平滑算法流程图

	C+0	C+1	C+2	C+3	C+4	C+5
D+0	20069.	21106. 5	21902. 6	22374. 4	22476. 6	22409. 5
D+1	21148. 4	20894. 9	21487. 6	21860. 8	21949	21969
D+2	19907. 8	20409.	20838	21126. 4	21216. 7	21299. 5
D+3	19699. 3	20018. 5	21227. 1	21573.	21649. 6	22074. 2
D+4	19779. 5	19952. 3	20164. 5	20367. 8	20440. 2	20597. 8
D+5	19714. 6	19799. 5	19957. 2	20139. 3	20212.	20386. 2
D+6	19459. 5	19511. 9	19654. 7	19835. 9	19912. 9	20088. 1
D+7	19282	19326. 8	19470. 2	19656. 3	19737. 4	19908. 2

Figure 4: 动作C第m帧与动作D第n帧所有骨骼关节的空间距离累加结果

确定动作混合轨迹：在第2节中，通过手语翻译模块，我们将汉语文本转换为相应的手语文本Gloss。对于每个得到的手语Gloss，我们从预先构建好的手语动作库中选取相应的帧序列来合成动作。如图3所示，当合成手语动作时，我们需要组合两个帧序列：前者为帧序列A，后者为帧序列B。为了确保合成动作准确传达手语语义，我们提取帧序列A末尾的一定长度的帧作为序列C，并从帧序列B的开头选取一定长度的帧作为序列D。接下来，我们利用BVH（Biovision Hierarchy，BVH）动作文件中的数据，计算帧序列C和帧序列D中每一帧的骨骼关节空间坐标。这些数据包括骨骼关节在空间中的初始姿态坐标和每一帧对应的旋转信息。通过计算骨骼关节的空间位置，我们可以进一步计算在两个动作帧之间的骨骼关节间的空间距离。该空间距离通过累加所有关节之间的距离来计算得出。最后可获得如图4所示的表格，表内单元格数据为动作C第m帧与动作D第n帧所有骨骼关节的空间距离累加的结果。为确保合成手语动作的流畅性和连贯性，必须对参与合成的动作帧进行细致选择。具体而言，在融合动作A和动作B时，每一帧融合生成的动作帧都需要从动作C和动作D的帧集合中挑选出一对空间

(c) 2024 中国中文信息学会计算语言学专业委员会

位置差异最小的对应帧。而融合生成的每一动作帧，主要将从动作C和动作D中选取出来的动作帧进行四元数差值，差值不是直接计算两帧的平均值，而是需要通过弹簧阻尼的方式去实现，类似于一个缓入缓出的曲线，最终合成的动作中，动作C的权重逐渐由1到0，动作D的权重逐渐由0到1。该操作旨在找到一条从动作C初始帧$(C+0)$至动作D终末帧$(D+n)$的融合路径，沿途保持空间变化的最小化。为平衡合成动作对原始动作A和B的依赖程度，并避免偏向任一动作，我们对图4所示的帧选择策略施加一个约束：即限制连续在一直线方向（横向或竖向）上选择的单元格数量，如最多允许连续选择3个单元格。该规定需灵活应用，以满足实际动作合成的需求，并最终实现在终结帧$(D+n)$上获得平滑的合成轨迹。

3.2 基于Transformer的数字人手语合成速度调节算法

对于上述依靠每个手语Gloss对应的手语动捕数据来进行数字人手语合成的策略，优点是只需获取常见的手语Gloss的动捕数据，此时每个手语Gloss对应的动作的速度是相对一致的。而对于不同手语句子中的同一个手语Gloss对应的动作，若其速度可根据手语句子的特定语境进行动态调整来使得数字人的动作更加自然和容易理解，则可以缓解经过多帧插值平滑算法合成的手语数字人视频序列整体韵律节奏单一的问题。

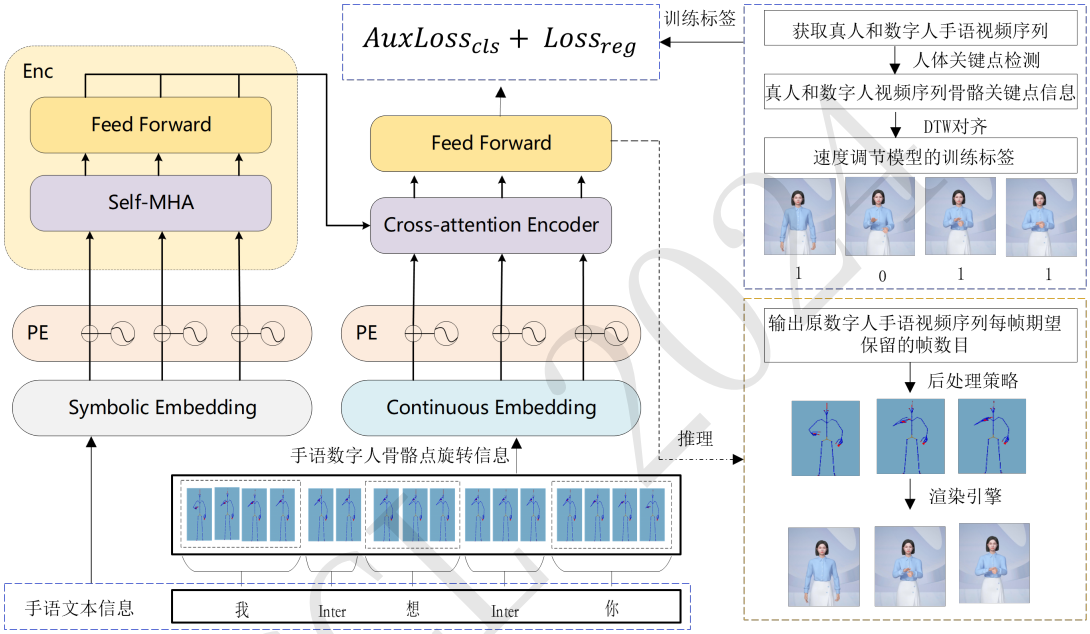


Figure 5: 数字人手语合成速度调节模型架构图

我们参考Saunders et al.(2022)的思路提出了一种基于Transformer的数字人手语合成速度调节算法，使得合成的数字人视频序列拥有贴近真实场景的韵律节奏，以促进用户的理解。其模型架构如图5所示，训练模型时的输入包含手语文本中的Gloss信息、手语数字人视频序列中的骨骼点旋转信息和手语数字人视频序列的训练标签。速度调节模型会先通过Cross-attention Encoder模块来融合手语文本信息和手语数字人视频序列中每一帧所包含的骨骼点旋转信息，它能辅助模型学习到手语句子中关键的上下文信息。最后在预测头部分，包含了一个回归损失和一个分类损失。

获取训练标签： 具体的，如图5所示，假设通过多帧插值平滑算法获得手语文本Gloss对应的手语数字人视频序列为X_u，而该手语句子对应的真人视频序列为T_r，首先使用开源的BlazePose(Bazarevsky et al., 2020)人体骨骼关键点检测技术，从X_u和T_r的每帧中提取出手部和头部的三维骨骼关键点信息。接着进一步利用这些关键点信息和每个Gloss在X_u及T_r中的起始和结束时间，通过动态时间规整（Dynamic Time Warping, DTW）(Müller, 2007)算法对齐两个视频序列。通过这种对齐，我们可以精确匹配X_u中每一帧在真实韵律节奏下所对应的帧数目K_t。K_t的值是一个离散标量，其中0表示该帧被忽略，1表示保留该帧，而大于1的值表示此时需要K_t帧来调整X_u的韵律。通过上述处理，获得的K_t值序列作为速度调节模型训练的标签。

(c) 2024 中国中文信息学会计算语言学专业委员会

设计类别感知均方误差损失函数: 我们提出了一个新的损失函数，称为类别感知均方误差损失函数（Classification Aware MSE Loss），用于动态调整多元回归任务中不同特征的权重，可有效缓解特征不平衡的问题。假定W_j表示权重占比，j表示X_u中的第j帧，C_j代表分类头输出的第j帧的分类预测值，即预测保留第j帧的数量，T_j是第j帧的真实标签，即实际需要保留的第j帧数量，则有

$$W_j = \frac{\max(C_j, T_j) + 1}{\min(C_j, T_j) + 1} \tag{1}$$

在模型训练过程中，由于C_j的输出结果是变化的，当预测值C_j接近真实标签T_j时，权重W_j接近1。若预测值C_j与真实标签T_j之间的差距增大，W_j则相应增大，以便使模型在训练过程中更加关注那些与真实标签差异较大的特征。因此，W_j能够在训练中动态调整不同特征的权重占比。

对于分类子网络，为了获取更准确的分类预测值C_j，我们采用了标签平滑的交叉熵损失函数$AuxLoss_{cls}$，其中K是类别的总数，n表示X_u中包含的总帧数，t_{ij}是第j帧真实标签对应的独热编码中的第i个元素，c_{ij}是模型预测得到的概率分布中第j帧的第i个元素，ϵ是标签平滑的平滑参数，则有

$$AuxLoss_{cls} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{K}\left((1-\epsilon)t_{ij} + \frac{\epsilon}{K}\right)\log(c_{ij}) \tag{2}$$

对于回归子网络，我们采用了类别感知均方误差损失函数$Loss_{reg}$，其中n表示X_u中包含的总帧数，R_j表示回归头输出的第j帧的回归结果，即预测需要保留的第j帧数量，则有

$$Loss_{reg} = \frac{1}{n}\sum_{j=1}^{n} W_j \cdot (R_j - T_j)^2 \tag{3}$$

模型推理时的后处理: 在推理阶段，模型将输出原手语数字人视频序列每一帧需要保留的帧数目K_p。为了确保动作过渡的平滑度，采用了如下的后处理操作：

- 当$K_p = 0$时，相关帧将被舍弃；
- 当$K_p = 1$时，该帧被直接保留；
- 当$K_p = 2$时，保留当前帧，并通过对当前帧与后续帧进行非线性插值操作以生成一个新的帧；
- 当$K_p = 3$时，首先对当前帧与其前一帧进行非线性插值以产生一个新的帧，继而保留当前帧，并对当前帧与后续帧执行非线性插值操作以获取另一新的帧；
- 当$K_p \geq 4$时，首先对当前帧与其前一帧执行非线性插值以产生一个新的帧，随后复制当前帧两次，并对当前帧与后续帧执行非线性插值操作产生一个额外新的帧。

最后，我们使用渲染引擎对经过后处理的手语数字人视频序列中每一帧对应的动捕数据进行渲染处理，生成了流畅且节奏感恰当的手语数字人视频序列。

4 手语数字人面部驱动算法

本节主要介绍我们的面部驱动算法。手语表达要实现高准确率和高可懂率，除了保证手语动作的正确和流畅，口型和表情等非手控信息也尤为重要(Von Agris et al., 2008)。打手语时附有口型，口型准确，能辅助听障人士理解动作；打手语时口不动，让用户感受不到数字人在与他聊天。打手语时准确呈现表情，表情生动，适当夸张，提升真实感；打手语时无表情，形象呆板，容易让听障人士产生歧义。

本节介绍我们所设计的面部驱动算法，图6为具体算法架构。涉及非语言信息，主要包括口型驱动和表情驱动两个模块。在高质量口型合成数据缺乏的情况下，我们舍弃了Audio2Face(Karras et al., 2017; Tian et al., 2019)的端到端方案，采用参考Jali(Edwards et al., 2016)的可解释性强，可配置化程度更高Pipline方案。

(c) 2024 中国中文信息学会计算语言学专业委员会

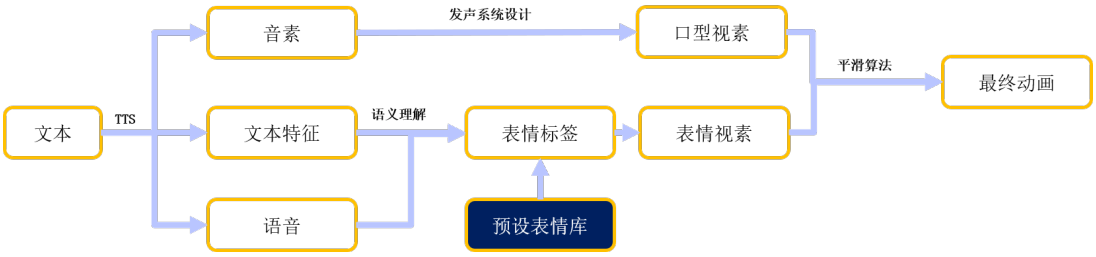


Figure 6: 面部驱动算法架构

4.1 口型驱动算法

我们将文本作为口型驱动的源头。将文本时间序列转换成音素时间序列，并通过口型驱动算法，生成嘴型和面部微表情系数，再通过驱动引擎实现数字人的面部动作的精准驱动。该此方法与语音无关，只与文本内容相关，不受语音特性变换影响。为了实现手语文本语音和数字人口型播报的同步，需要获取时间同步的口型BS（BlendShape，BS）参数以驱动数字人。

将第2节自然文本经过手语翻译算法得到的手语文本Gloss，再经过vivo自研文本语音合成TTS系统生成语音和对应音素，保证语音和音素序列是时间同步的。文字中每个字的多个音素p，以及每个音素的持续时间共同构成语音文字对齐信息。

对音素序列进行分组，同时每组音素设定一个重要性权重w_1，让对口型影响大的音素具有更大的权重。将语音文字信息中的每个音素持续时间进行分析，将输入文本时长按时间间隔T等分为多个。

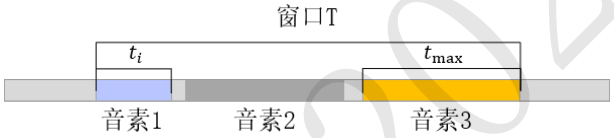


Figure 7: 音素窗口示意图

图7为音素窗口示意图。在时间窗口T中计算音素的密集程度。第i个音素的密集程度w_2计算方式为

$$w_{2i} = (\frac{t_i}{t_{max}} + \frac{t_i}{T})/(n+1) \tag{4}$$

其中t_i表示当前第i个音素持续时间，t_{max}表示时间窗口T中最长的音素持续时间，n表示时间窗口T中。通过这个权重，可以丢弃密集度高但是对口型影响较小的音素，避免唇形的抖动问题。音素参考拼音类型可以分为声母和韵母，其中韵母可以分为单韵母、复韵母、前鼻韵母、后鼻韵母、整体认读音节、三拼音节。对三拼音节和整体认读音节进行拆分，拆分为前四种韵母的组合。最后音素序列分为声母、单韵母、复韵母、前鼻韵母、后鼻韵母，同时每组音素设定一个重要性权重，让对口型影响大的音素具有更大的权重。一般的，对于重要性权重，设定声母、单韵母、复韵母、前鼻韵母、后鼻韵母的权重分别为（1.0，0.9，0.6，0.5，0.5）。最终根据语音对齐信息得到重要性权重w_1和音素的密集度权重w_2。

然后进行数字人口型映射，将音素和数字人口型BS参数进行逐一映射，设计合理的数字人发音系统，使得数字人单一口型和现实人物一致。为了方便起见，我们将一个口型对应的一组BS参数叫做视素v(Bear and Harvey, 2017)，则可以将音素序列转换成视素序列。由于音素的个数是有限且独立的，我们首先使用常见的面部工具（如LiveLinkFace），通过人工采集获取不精确的音素到BS参数映射关系。然后专业的3D建模工程师人工调整虚拟人的BS参数，完成音素到视素的精确表达。最终将对应的音素序列P转变为驱动参数序列V，对于每帧i来说，视素驱动参数：

$$v_i = min(S(p_i) * w_{1i} * w_{2i}, 1.0) \tag{5}$$

(c) 2024 中国中文信息学会计算语言学专业委员会

其中w_{1i}为音素的重要性权重，w_{2i}为音素密集度权重，S为音素到视素的映射关系。根据前述得到的离散音素序列，得到相应的视素参数序列。

为了解决离散视素间的平滑过渡，我们使用SG（Savitzky-Golay，SG）时间序列平滑算法(Schafer, 2011)对视素的不同参数分别进行插值和平滑，SG平滑算法是在一个滑动窗口内，进行多项式拟合。该方法可以实现自然的口型切换，更良好的交互体验。

字平滑策略：将每个汉字持续的时间（即一组音素对对应的时序序列）作为一个平滑窗口，将V分为多个不同的词窗口V_i，对于每个V_i应用SG算法，保证每个字对应的口型自然。

句平滑策略：将每个字平滑后的视素序列再次应用SG算法对整句话进行平滑，保证唇形运动自然。

$$V = SG\underbrace{((SG(v_1), \ldots, SG(v_i))}_{m} \tag{6}$$

其中m表示该序列中字的个数。最终得到平滑的视素序列V_s，用于驱动手语数字人口型。根据文本信息，生成数字人模型的口型变化序列和语言，然后驱动数字人模型的口型变化，使其与现实人物的语音保持一致。

4.2 表情驱动算法

对文本特征进行提取后，通过语义理解得到情感标签。我们提前制作了数字人通用的预设表情库，将得到的情感标签与预设表情进行一一映射。表情驱动算法同样是基于视素进行数字人面部微表情的驱动，预设表情同样是可以认为是视素。将表情视素和口型视素直接相加，再通过平滑算法进行平滑，即可实现带微表情的唇形同步数字人面部驱动。

最终我们的面部驱动算法集成了口型驱动和表情驱动能力，支持中英文所有发音的口型，支持13种表情，主观评测口型一致性大于90%。

5 自测结果

本节主要介绍我们的自测方法和评测结果。手语是小语种，语料库和相关公开评测集有限，难以自动化评测。我们采用自构建手语语料库的形式，并聘请相关手语专家进行人工评测，确保我们的手语数字人能够较好的满足听障人群的需求。

5.1 评测方式

为了对整个手语数字人将汉语翻译成中国手语方面的自然性和准确性进行评估，我们先采集了约6000句通用场景下包含书面语文本及其对应的手语转写文本的句子，然后精选其中的约1500句包含常用Gloss的句子，覆盖日常生活、工作学习等场景，构成自测专用的手语语料库，并基于该语料库对我们系统的手语合成结果进行人工评测打分。我们邀请多位手语专家进行人工评测，以手语语法的准确性、表达的自然性和可读性以及是否满足听障人士理解为主要标准。

通过我们的手语数字人系统获取到手语语料库中每个手语句子对应的手语数字人视频，统计每一个手语数字人视频翻译正确的词占该手语句子全部评测词的比值，取平均后作为手语词可懂率；统计每一个手语数字人视频的翻译得分，取平均后作为手语句可懂得分。

5.2 评测标准

对于汉语-手语翻译的评测标准，参考中英翻译的评价指标，我们定义0-5分的可懂率，我们认为3分是一个大致能够还原原意的标准。以下是具体的评测细则：

- 0分：合成的手语与原文完全不对应；
- 1分：看了手语不知所云，仅个别词语合成正确，无语义和逻辑；
- 2分：手语与原文小部分符合；有漏译、误泽或严重语法错误；
- 3分：手语大致表达了原文的意思，但对译文整体理解影响不大；
- 4分：手语基本较流畅地表达了原文的意思，但不影响整体理解；
- 5分：手语准确且完整地表达了原文信息，表达流畅自然。

第二十三届中国计算语言学大会论文集，第382页-第392页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

5.3 评测结果

我们基于自构建的通用场景评测集，并邀请多位手语专家进行人工评测，最终的评测结果准确率指标为：

- 手语词可懂率96.82%，手语句可懂得分4.13分（满分5分）；
- 自然程度主观评价：流畅度与真实感优秀；
- 口型与表情：支持中英文所有发音的口型，支持13种表情。

6 官方结果与分析

CCL24-Eval 手语数字人翻译质量评测以手语语法的准确性、表达的自然性和可读性以及是否满足聋人理解为主要标准，综合考虑手势清晰度、流畅性、与汉语原文的语义一致性等[3]。手语数字人翻译质量的人工评测包括四个主要指标：手语语法准确性、自然性、可读性以及文化适应性。综合得分是根据每个单项指标的得分与其相应的权重系数计算得出的加权和。评测重点关注手语数字人在准确表达手语的能力，强调自然性和可读性，同时考量其文化适应性。

Table 1: CCL24-Eval 手语数字人翻译质量的综合评测结果

排名	队伍	得分	单位
1	维沃手语数字人团队	3.513	维沃移动通信有限公司
2	team 1	2.447	-
3	team 2	2.119	-
4	team 3	1.806	-

Table 2: CCL24-Eval 手语数字人翻译质量不同维度的评测结果

专家组评分				
队伍	准确性	自然性	可读性	文化适应性
维沃手语数字人团队	3.50	3.75	3.43	2.36
team 1	2.21	2.36	2.00	1.79
team 2	2.61	2.25	2.21	2.07
team 3	1.04	2.14	1.68	1.75
采集组评分				
队伍	准确性	自然性	可读性	文化适应性
维沃手语数字人团队	3.39	3.43	2.86	2.43
team 1	1.11	2.07	1.86	1.29
team 2	2.14	2.07	1.96	1.11
team 3	0.25	1.79	1.54	0.82
普通组评分				
队伍	准确性	自然性	可读性	文化适应性
维沃手语数字人团队	4.14	4.14	3.75	3.11
team 1	1.75	2.89	2.43	2.50
team 2	2.93	2.89	2.89	2.57
team 3	1.79	2.86	2.54	2.04

[3]https://github.com/ann-yuan/QESLAT-2024

(c) 2024 中国中文信息学会计算语言学专业委员会

表1展示了CCL24-Eval 不同队伍（前4名队伍）的手语数字人翻译质量综合评测结果，其中，我们所开发的手语数字人系统在本次评测中表现卓越，其成绩较第二名领先超过1分，说明本文方法在手语合成任务上有着更为优越的性能，能保证手语数字人将汉语翻译成自然、准确和可读性强的中国手语，并且能被更多的聋人群体所理解和接受。

表2为CCL24-Eval 官方提供的专家、采集和普通组在手语语法准确性、自然性、可读性以及文化适应性四个评测指标上不同队伍的手语数字人翻译质量评测结果，我们提出的手语数字人系统在专家、采集和普通组所评测的准确性、自然性、可读性以及文化适应性这四个指标均取得了本次评测的最佳成绩。其中，本文提出的手语翻译算法可获取更符合中国手语词序规则的手语文本，而动作融合算法则将手语文本对应的每个手语动作单元融合成流畅和手势形态更精准清晰的手语数字人动作，同时面部驱动算法将非言语元素自然地融入翻译中，这些都使得我们的手语数字人系统在手语语法准确性、自然性和可读性均取得较好的评测成绩。但在文化适应性上相比其它指标有1分左右的差值，具体原因是本文提出的手语翻译算法在翻译过程中有时并未考虑到文化差异和特定的社会语境，如礼貌用语、行业术语等，同时我们提出的表情驱动算法当前只支持13种表情，有时无法映射出某些手语句子对应的情感色彩。而较低的文化适应性也会影响到整个手语数字人系统在自然性和可读性上的表现，比如部分面部表情未自然地融入翻译中时会显现出生硬、不自然的状态和在不同语境下的适应性未被考虑时会影响其可读性。

7 结语

本文以提高听障人士和健听人士的双向沟通效率为动机，为广大听障群体提供沟通上的便利，提出了结合手语翻译算法、动作融合算法和面部驱动算法的手语数字人翻译系统。本文提出的系统在手语专家的评测下达到了手语词可懂率96.82%，手语句可懂得分4.13分（满分5分）的优秀性能，同时手语语义传达较为准确，手语数字人动作自然不生硬，贴近真人。并且该模型在“CCL24-Eval 手语数字人翻译质量评测”任务上取得了综合得分第一的成绩。本文的不足之处在于没有在更细粒度的手语动作特征上进行动作融合，现在无论平滑过渡还是韵律节奏都是以词动作为粒度进行融合，未来为了更贴近真人，应该拆分到左右手、手位置等更细的粒度。同时手语数字人的手语动作不只是局限于手部动作，身体姿态也能传递很多信息，以及如何实现更加生动的面部驱动效果仍是值得研究的方向。

参考文献

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Helen L Bear and Richard Harvey. 2017. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95:40–67.

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December.

(c) 2024 中国中文信息学会计算语言学专业委员会

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151.

Ronald W Schafer. 2011. What is a savitzky-golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4):111–117.

Kristine Slot. 2007. Motion blending. *Copenhagen University. Department of Computer Science*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Yibo Sun, Duyu Tang, Nan Duan, Yeyun Gong, Xiaocheng Feng, Bing Qin, and Daxin Jiang. 2020. Neural semantic parsing in low-resource settings with back-translation and meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8960–8967.

Guanzhong Tian, Yi Yuan, and Yong Liu. 2019. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *2019 IEEE international conference on Multimedia & Expo Workshops (ICMEW)*, pages 366–371. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

CCL 2024

CCL24-Eval任务10系统报告：结合LLM与3D动画技术的手语数字人系统

杨阳[1,2]，张颖[2]，黄锴宇[2]，徐金安[2,*]
1. 果不其然无障碍科技（苏州）有限公司
2. 北京交通大学
{y.yang, novzying, kyhuang, jaxu}@bjtu.edu.cn

摘要

手语翻译（Sign Language Translation, SLT）系统作为一种重要的辅助技术，为听障人士提供了与他人沟通的有效途径。然而，传统手语翻译系统在准确性、流畅性差等方面存在问题。本文提出了一种结合大语言模型（Large Language Model, LLM）和3D动画技术（3D Animation Technology）的手语翻译系统，旨在克服这些局限，提高翻译的准确性和流畅性。本文详细介绍了系统的设计与实现过程，包括提示词设计、数据处理方法以及手语数字人翻译系统的实现。实验结果表明，采用LLM方法在手语翻译中能够生成较为自然和准确的结果。在标准评估和人工评估的两种评估方法下，本系统在大多数情况下能够较好地完成手语翻译任务，性能优于传统方法。本文的研究为进一步改进手语翻译系统提供了有益的参考和启示。

关键词： 手语数字人；手语翻译；大语言模型

System Report for CCL24-Eval Task 10: A Sign Language Avatar System Integrating LLM and 3D Animation Technology

Yang Yang[1,2], Ying Zhang[2], Kaiyu Huang[2], Jinan Xu[2,*]
1. GoBetterStudio
2. Beijing Jiaotong University
{y.yang, novzying, kyhuang, jaxu}@bjtu.edu.cn

Abstract

Sign Language Translation (SLT) systems serve as a crucial assistive technology, providing an effective means of communication for individuals with hearing impairments. However, traditional SLT systems face challenges in terms of accuracy and fluency. This paper proposes a novel SLT system that combines Large Language Models (LLM) and 3D Animation Technology to address these limitations and enhance translation accuracy and fluency. The paper provides a detailed account of the system's design and implementation process, including prompt design, data processing methods, and the implementation of the sign language digital human translation system. Experimental results demonstrate that the LLM-based approach can generate more natural and accurate translations in SLT. Under both standard and human evaluations, this system performs the SLT tasks better than traditional methods in most cases. This research offers valuable insights and references for further improving SLT systems.

Keywords: Sign Language Avatar , Sign Language Translation , Large Language Model

*通讯作者/Corresponding author

第二十三届中国计算语言学大会论文集，第393页-第404页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

手语是一种复杂的视觉语言，主要用于听障人群的交流。它不仅包括手部动作，还涉及面部表情和身体姿态等多方面的表达。这些特性使得手语的翻译和生成面临巨大挑战。近年来，随着人工智能（Artificial Intelligence, AI）和自然语言处理（Natural Language Processing, NLP）技术的快速发展，开发高效的手语翻译系统成为可能。手语数字人（Sign Language Avatars）可以模拟手语动作，为聋人提供实时或非实时的翻译服务，不仅能极大地提升听障人士的交流能力，还能促进社会的包容性和无障碍环境的建设。

手语数字人翻译属于手语生成（Sign Language Production, SLP）领域，即以口语作为源语言，生成相应的手语表达(Yin et al., 2021)。现有的手语生成方法主要分为以下几类，各自存在明显的局限性:

- 拼接手语图片：通过从现有语料中标记图片并进行拼接来生成翻译结果。其局限性在于，准确性受限于语料的质量和覆盖范围，难以应对复杂的手语表达。
- 姿态估计技术(Zuo et al., 2024; Forte et al., 2023)：这一关键方法通过分析输入文本来解码并确定手部姿态信息，如手部的位置、形状和运动轨迹。然而，这种方法需要大量高质量的手语视频数据来训练姿态估计模型，数据的质量和多样性直接影响生成效果。
- 生成式模型（如扩散模型）(Baltatzis et al., 2024; Saunders et al., 2020a)：这些模型通过不断预测下一帧图像，以生成连续的手语动作。尽管能够生成高质量的连续图像序列，但其计算复杂度高，难以实现实时翻译。

在实际应用中，选择合适的方法并在现有方法基础上进一步提高手语生成的准确性和自然性，是未来研究的重要方向。

本研究提出了一种创新性手语翻译系统，该系统融合LLM的文本处理能力与3D动画技术的表现力，具备深度语义理解，能够深入分析复杂的语言结构和上下文信息，生成更为准确和自然的手语翻译。此外，系统设计了多样化的上下文学习样本，确保翻译结果在语义和语境上的双重精准，满足多样化的交流场景需求。3D动画技术的应用，进一步丰富了手语的视觉呈现，创造出既流畅又逼真的手语动作，这些动作不仅在视觉上吸引人，更在表达上贴近自然手语的非言语特征，如节奏、力度和表情。系统具备实时交互与反馈的能力，确保用户输入能够得到即时响应，有效提升了沟通的效率和用户体验。多模态融合技术的应用，整合了手部动作、面部表情和身体语言，实现了全面的交流表达，使手语翻译更加生动和直观。系统的可扩展性与兼容性设计，也使其能够灵活地集成到不同的平台和设备中，为用户提供了便捷的接入和使用体验。

综上所述，本文研究在手语翻译技术领域提供了一种全新的视角和解决方案，不仅推动了技术的发展，也为听障社群带来了更加丰富和便捷的沟通方式，有助于构建一个更加包容和无障碍的交流环境。

2 数据处理

数据预处理是开发手语数字人翻译系统的关键步骤之一。高质量的数据能显著提升模型性能，进而提高翻译准确性。本文在数据预处理过程中主要包括数据准备和数据格式化两个部分，这些步骤确保了模型能够有效地学习和翻译手语。

2.1 数据集

本研究严格遵循评测标准，所采用的手语语料库均来源于经过评测任务[1]授权的数据库，如表 1所示，包括“XMU_CSL”、“BUU_CSL”以及“ZZSZY_CSL”。这些语料库覆盖了日常生活中的各类典型情境，如医疗沟通、客户服务、交通指示和购物交流等多个关键领域。通过这种跨领域的数据使用策略，本文确保了数据的全面性和多样性，为手语生成模型的训练与评估提供了坚实基础。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

[1] https://github.com/ann-yuan/QESLAT-2024

Dataset	Lang.	Sentences	版权所有者
XMU_CSL	CSL	500	厦门大学
BUU_CSL	CSL	500	北京联合大学
ZZSZY_CSL	CSL	74	株洲手之声信息科技有限公司

Table 1: 本研究中使用的手语语料库概览，包括语料库名称、使用的语言（CSL代表中国手语）、句子数量以及版权所有者。表格展示了三个主要的语料库，共有1074个句子。

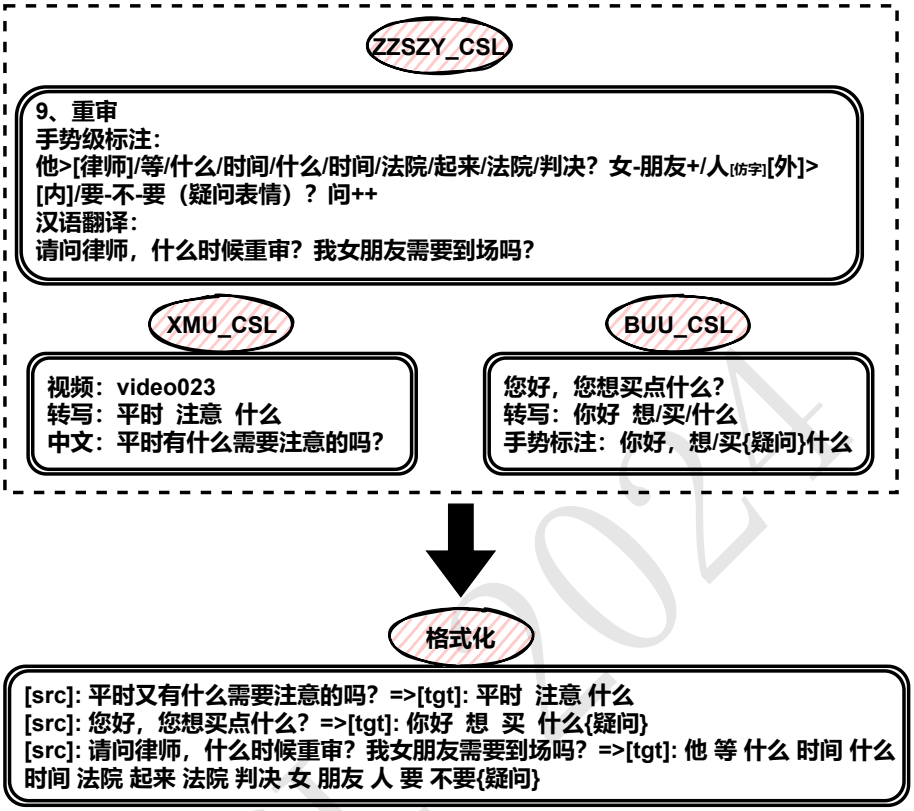


Figure 1: 数据格式化流程图，展示了如何将不同标注风格的语料库内容转换为统一格式。

如图 1所示，这些语料库不仅包含手语记录，还包含丰富的标注信息。每个条目都由中文源文本及其对应的Gloss组成。Gloss作为一种标准化的手语文本表现形式，为每个具体的手语动作提供了精确描述。这种一对一的映射关系极大地方便了从文本到手势的直接转换，提升了模型训练的效率和准确性。例如，中文句子“平时有什么需要注意的吗？”在Gloss中可能被分解为“平时”、“注意”和“什么”，这种分解不仅体现了手语的构词特点，也为理解和生成手语提供了结构化的视角。

如Muller(2023)所述，Gloss存在以下缺点：

- 信息丢失：Gloss并非手语的完整表示，缺乏手语的非手动通道（如面部表情和身体语言）和三维空间使用等语言线索。图 1中，“XMU_CSL”是词级标注，只标注出了手语动作。为了减轻信息丢失的问题，“BUU_CSL”加入了表情信息标注，非手部动作标记等。“ZZSZY_CSL”则更进一步，使用了更加详细的手势级标注，同样包括表情信息和非手部动作标记等。

- 不一致性：不同语料库中的Gloss转写标准差异很大，导致不同语料库或跨语言的Gloss不可比。如图 1所示，“XMU_CSL”、“BUU_CSL”以及“ZZSZY_CSL”存在不同的标注风格。

(c) 2024 中国中文信息学会计算语言学专业委员会

- 资源密集：Gloss的转写过程需要由专家语言学家完成，非常劳动密集，如果简单应用一些数据增强方法，容易导致翻译错误。
- 非实际应用：Gloss是语言学工具，并非聋人社区中确立的书写系统，手语用户通常在日常生活中不阅读或书写Gloss，所以很难从其他来源获取Gloss。

尽管存在这些缺点，Gloss也有许多优势，这些优势使其在手语翻译系统中仍然具有重要作用：

- 文本兼容性：Gloss作为手语的语义标签，在英语中通常由口语单词的大写基本形式组成，汉语中则由字词组成，能够无缝融入现有的机器翻译(Machine Translation)流程。
- 简化处理：由于Gloss是文本形式的，现有的机器翻译方法只需最少修改就能应用。
- 易于理解：对于机器翻译研究者来说，Gloss提供了一种相对容易理解手语的方式，因为它们以文本形式呈现。

利用这些高质量的手语语料库，本文旨在深入探索并推动手语翻译技术的发展，为听障群体提供更加准确、自然且高效的交流辅助工具。这些资源的多样性和丰富性，为本研究的深度分析和模型优化提供了坚实的数据支持。

2.2 数据格式化

数据格式化是数据处理流程中的关键步骤，涉及将原始数据转换成适合LLM进行上下文学习的形式。这一过程对于确保模型能够有效学习Gloss的翻译规则至关重要。

首先，将每条数据中的中文源文本与其对应的Gloss进行配对，形成源文本和目标文本对。这些文本对构成了模型训练的基本单元，使得模型能够在上下文中学习从中文到Gloss的映射。然后，本文利用生成的源文本和目标文本对，构建模型的上下文学习样本。

具体细节如图 1所示。为了确保所有数据的格式统一，便于模型批量处理，本文将不同来源的语料处理为相同格式的数据。在源文本和目标文本中添加特殊标记，以帮助模型识别源目语言的边界。例如，在源文本前添加“[src]”标记，在目标文本前添加“[tgt]”标记。在处理表情时，本文将表情块置于手语块之后，手语和表情视为一个同步块，同步块表示手语和表情在客户端会同时启动。

通过上述步骤的数据预处理，确保了数据的高质量和一致性，为后续模型的训练和评估提供了坚实基础。数据预处理不仅提升了模型性能，也为手语翻译系统的实际应用奠定了基础。本文将在未来的工作中继续扩展数据集，探索更加复杂的手语表达形式，并进一步优化数据预处理流程，以应对更复杂的手语翻译任务。

3 方法

为有效解决手语识别中的歧义性和边界模糊性问题，本文提出了一种结合LLM和3D动画技术的手语翻译系统，为听障者提供更加准确和便捷的沟通工具。系统整体架构如图 2所示，主要包括推理阶段和可视化阶段。在推理阶段，系统首先读取用户的中文输入，并结合知识库语料与LLM将中文翻译为Gloss。随后，手语数字人模块生成手语的视觉表示。系统的核心在于利用LLM结合上下文信息动态生成准确的Gloss。在可视化阶段，系统以多模态方式展示结果，使用户能够直观理解和验证手语转换结果。这一流程不仅提高了手语翻译的准确性和效率，而且通过生动的视觉表现增强了交流的自然性和直观性。

3.1 Gloss推理阶段

在推理阶段，系统主要集成了知识库语料与LLM。LLM的翻译任务提示词设计如表 2所示，包含任务描述、示例和翻译任务三部分(Agrawal et al., 2022; Vilar et al., 2023; Zhang et al., 2023)。任务描述提供了手语翻译任务的背景信息，解释了[src]和[tgt]的含义，即从中文原文翻译到手语的文本表示形式——Gloss，以及Gloss在机器学习模型中的重要性。示例部分展示了人类专家编撰的翻译例子，帮助模型理解如何进行翻译，“example”是占位符，代表上下文学习样本。翻译任务部分明确说明需要翻译的具体中文句子，即占位符“chinese”，并提示模型生成对应的Gloss。这种设计旨在提高模型对手语翻译准确性和自然性的理解，进而提升翻译质量。

(c) 2024 中国中文信息学会计算语言学专业委员会

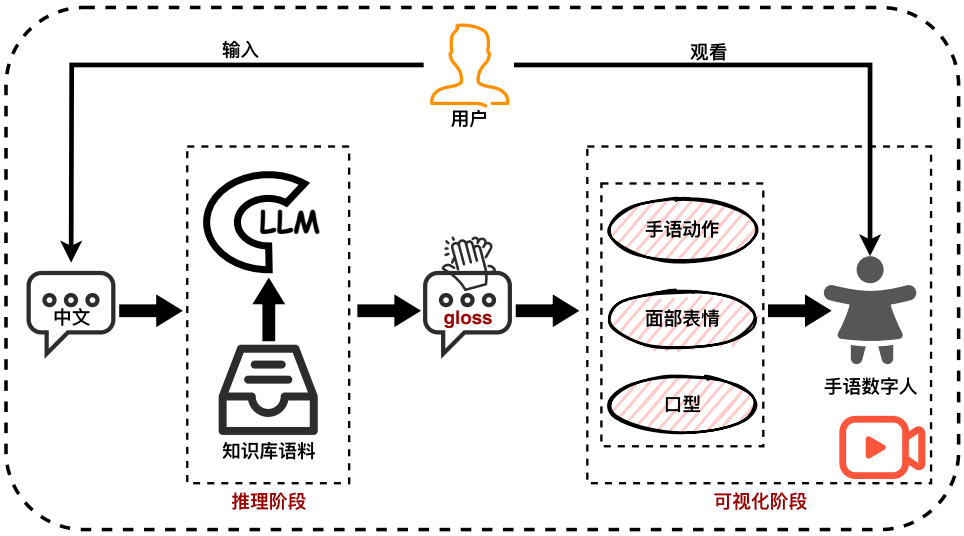


Figure 2: 手语翻译系统的架构示意图，包括两个核心阶段：推理阶段和可视化阶段。

分类	提示词
任务描述	你是一个手语翻译专业助手。需要将[src]翻译到[tgt]，[src]中是中文，[tgt]中是对应的gloss，gloss是手语的文本表示， 对于机器学习模型来说，gloss是非常友好的一种表示。
示例	下面给出了人类专家编撰的例子，你需要仔细学习，然后给出最后一行翻译的gloss。{example}
翻译任务	[src]: {chinese} ⇒ textnormal{[tgt]}:

Table 2: 用于LLM翻译任务的提示词设计。

如表 3所示，本文充分利用LLM出色的文本处理能力，将处理后的数据输入到模型中，使其能够深刻理解知识库语料并生成最终的Gloss。除此之外，本文对比了包括文心4.0、通义Plus和星火Pro在内的多个模型，旨在评估它们在翻译准确性和自然性方面的表现。通过这种比较，可以洞察不同模型在理解和转换自然语言到手语Gloss序列的能力上的差异。整个过程涉及复杂的语义转换和上下文理解，不仅提升了翻译的准确性，还增强了翻译的自然性，使手语翻译系统在实际应用中更加实用和可靠。

3.2 可视化阶段

在可视化阶段，将LLM输出的Gloss视为一系列具体的指令，指导数字人的行为表现。客户端采用了一个创新的3通道动画状态机设计，包括手语通道、表情通道和口型通道，以实现动作、表情和口型的同步播放。

3D手语动作库的构建 本文在构建3D手语动作库的过程中，严格遵循了国家标准手语词典的规范，精心制作了所有国家标准手语词典中的手语动作文件，总计超过8000个GLTF格式的动作文件。为了确保翻译的准确性和一致性，本文将每个动作与相应的Gloss进行了一一对应。此外，本文对每个动作进行了详细的标注，包括手势起始位置、结束位置、动作时长以及手势轨迹等信息，以便于在实际应用中能够精确调用和展示。

为了进一步提高手语动作库的实用性和覆盖面，本文还扩展了国家标准手语词典的内容，增加了常用的专业术语和新兴词汇的手语动作。同时，针对不同手语使用者的习惯差异，本文提供了多种手语动作的变体，以满足不同用户的需求。

手语表现 在手语表现阶段，系统首先对Gloss进行解析，将其拆分为手语动作、表情和口

(c) 2024 中国中文信息学会计算语言学专业委员会

	Source	Reference	LLM	Inference
Train	两斤多点，行不行?	两斤 多 一点 行 不行	文心4.0 通义Plus 星火Pro	两斤 多 一点 行 不行 两斤 多 一点 行 不行 两斤 多 一点 行 不行
	是的，我需要手语翻译。	是 手翻 我 需要	文心4.0 通义Plus 星火Pro	是 手翻 我 需要 是 手翻 我 需要 是 手翻 我 需要
Test	您的外卖到了，请尽快领取，谢谢。	外卖 到 拿 快 谢谢	文心4.0 通义Plus 星火Pro	外卖 你 到 快 取 谢谢 您的外卖 到 请 快领 谢谢 外卖 到 你 快 领取 谢谢
	具体需要做哪些检查呢?	具体 检查 哪些	文心4.0 通义Plus 星火Pro	具体 检查 哪些 具体 检查 哪些 需要 检查 具体 哪些 需要
	如果发现有诈骗情况，要及时报警	如果 发现 骗子 有 打电话 报警 快	文心4.0 通义Plus 星火Pro	如果 发现 诈骗 有 报警 及时 如果 发现 诈骗 情况 要 报警 发现 诈骗 情况 及时报警
	请问您需要将行李托运至何处?	问问 你 行李 托运 到 哪里	文心4.0 通义Plus 星火Pro	你 行李 托运 哪里 请问 问 你 行李 托运 哪里 行李 托运至何处 你 需要

Table 3: 不同LLM在处理手语翻译任务时的推理结果对比。其中Source列展示原始的中文句子，Reference列提供这些句子的标准标注结果，Inference列展示各个LLM生成的推理结果。

型，分别对应动画状态机的三个通道。手语动作的过渡采用了动画混合技术，该技术通过插值计算在不同动作之间生成平滑过渡，使得手语动画更加自然流畅。具体而言，动画混合技术能够在动作切换时计算中间帧，从而避免生硬的动作切换，提升视觉连贯性。

表情部分使用预先制作的常见情绪表情，这些表情存储在YAML格式的配置文件中，通过morph通道控制数字模型的面部表情变化，确保面部表情的自然性和准确性。对于口型部分，本文采用ARKit BlendShape技术，口型数据来源于Peng(2023)的工作结果。该模型通过对输入句子的情感和TTS语音特征进行分析，生成对应的口型变化数据，尽可能保证口型动作与手语内容的高度一致和同步。

可视化阶段与用户交互环节共同构成一个闭环系统，旨在提供高质量的多模态交流体验。手语翻译系统通过多模态方式展示结果，使用户能够直观地理解和验证手语的转换结果。整个系统的设计和实现，不仅提升了手语翻译的准确性和自然性，还增强了用户的互动体验，为手语使用者提供了一个更加便捷和高效的交流工具。

4 结果

4.1 实验评估

数据集划分与评估方法 为了评估不同LLM在手语翻译中的表现，本文将数据集划分为训练集和测试集。考虑到商业LLM的token限制，本文随机选择了语料库中100条数据作为训练集（即上下文学习样本），其余数据作为测试集。这种划分方式确保了在有限的资源下能够充分训练模型并评估其泛化能力。本文选择了三种不同的知名商业大模型进行评估，通义千问(Bai et al., 2023)，百度文心(Sun et al., 2021)，讯飞星火[2]，并从训练集和测试集中各选择了10句，通过LLM翻译成Gloss，对结果进行评估。评估指标包括BLEU-1（B1）、BLEU-2（B2）、BLEU-3（B3）、BLEU-4（B4）和ROUGE评分，这些指标能够综合反映模型在翻译质量上的表现。

[2] https://gitee.com/iflytekopensource/iFlytekSpark-13B

(c) 2024 中国中文信息学会计算语言学专业委员会

自动评测结果 表 4展示了不同LLM在训练和测试阶段的表现。评估采用了BLEU-1至BLEU-4和ROUGE指标，这些指标衡量了模型生成文本与参考文本之间的相似度。表格中列出了中国部分知名商业大模型的表现。从总体上看，各模型的表现均较好，显示了本文系统的通用性。然而，在测试阶段，文心4.0的表现最为优异，特别是在生成测试集的Gloss序列方面具有显著优势。这表明文心4.0在实际应用中具有更高的可靠性和准确性。评估训练集样本的目的是验证LLM在上下文学习中的有效性，而测试集样本的评估则用于测试LLM在实际应用中的泛化能力。文心4.0在所有评测指标上均优于其他模型，特别是在生成测试集序列Gloss时表现尤为突出。这可能是因为文心4.0在语义理解和上下文处理上有更强的能力。

LLM	Train					Test				
	B1	B2	B3	B4	ROUGE	B1	B2	B3	B4	ROUGE
星火3.5	75.94	75.32	74.53	71.41	78.33	32.98	19.61	7.70	4.18	33.56
星火Pro	82.88	81.45	80.57	76.77	83.71	35.02	25.94	21.29	14.42	37.74
通义Plus	79.31	73.84	70.34	67.03	82.55	30.99	14.25	5.70	3.07	34.73
通义Turbo	87.93	82.33	78.58	74.95	87.31	40.79	15.72	5.93	3.08	37.23
文心4.0	**96.30**	**94.73**	**93.59**	**91.22**	**96.67**	**66.67**	**46.06**	**31.85**	**18.79**	**63.20**
文心3.5	83.64	74.67	69.67	63.35	83.22	56.56	30.47	20.62	13.40	47.67

Table 4: 不同大型语言模型（LLM）在训练集和测试集上的自动评测结果。

组别	准确性	自然性	可读性	文化适应性
专家组	2.1	2.25	2.21	2.07
采集组	2.14	2.07	1.96	1.11
普通组	2.93	2.89	2.89	2.57

Table 5: 人类评审对手语数字人翻译质量的评分结果，分别从准确性、自然性、可读性和文化适应性四个维度进行评估。

人类评测结果 表 5展示了手语数字人翻译质量的人工评测结果，采用5分制，包括四个主要指标：手语语法准确性、自然性、可读性以及文化适应性，各项指标的得分反映了数字人在手语表达上的综合表现。评测重点关注手语数字人在准确表达手语的能力，强调自然性和可读性，同时兼顾文化适应性。从结果可以看出，普通组在所有指标上得分最高，表明该系统在实际应用环境中的表现较好。专家组和采集组的评分略低，可能是因为他们对手语有更高的专业要求。这些结果表明本文的系统在实际应用中具有较高的用户接受度，但仍有提升空间，对于理解手语翻译系统在实际应用中的用户接受度和改进方向具有重要意义。

4.2 结果分析

训练集推理的完美表现 所有模型在训练集上的推理表现都非常优秀，这说明它们在训练过程中成功地记住了训练数据的语境和模式。具体来说，训练集中的每个句子在不同模型生成的推理结果中都能高度一致，几乎没有误差。这表明模型在面对已知数据时，能够很好地应用其学习到的知识进行准确的推理。

测试集推理结果的差异 在测试集上，不同模型推理结果出现了显著的差异。具体可分为以下3类：

- Gloss不一致：例如，在句子“您的外卖到了，请尽快领取，谢谢。”的推理中，不同模型使用的Gloss略有不同。文心4.0的结果是“外卖 你 到 快 取 谢谢”，而通义Plus的结果是“您的外卖 到 请 快领 谢谢”。尽管从中文角度看，这些Gloss在语义上相近，但在3D数字人手语生成时，这种不一致可能导致手语翻译的不准确。

(c) 2024 中国中文信息学会计算语言学专业委员会

- Gloss顺序不一致：在句子“请问您需要将行李托运至何处？”的推理中，文心4.0的结果是“你 行李 托运 哪里 请问”，而通义Plus的结果是“问 你 行李 托运 哪里”。Gloss顺序的不同会影响手语翻译的准确性，因为手语的语序与口语可能不同，需要特别的注意。

- Gloss缺失或多增：例如在句子“请问您需要将行李托运至何处？”的推理中，文心4.0的结果是“你 行李 托运 哪里 请问”，而星火Pro的结果是“行李 托运至何处 你需要”。这里可以看到星火Pro的推理结果中多了一些Gloss，而文心4.0则缺失了一些Gloss。Gloss的缺失或多增会影响到句子的完整性和准确性。

评测任务的综合得分分析 在表 6中展示了不同参赛队伍在专家组、采集组和普通组的得分情况。表格中列出了四个组别的评分结果：A组，B组，D组以及本研究提出的方法（标记为“Ours”）。每组得分反映了各队伍的数字人在表达手语语义时的有效性。在总共14支参赛队伍中，包括本研究在内的4支队伍能够有效表达手语语义。根据专业评审的打分，本研究的方法在所有参赛者中排名第二，显示出较高的翻译准确性和用户接受度，尤其在专家组和普通组中得分较高。这表明本方法在处理实际应用中的手语翻译任务时具有较好的通用性和准确性。然而，与最佳表现的模型相比，仍存在一定差距，特别是在采集组中的表现需要进一步提升。

可能的改进方向 为了进一步提升LLM在gloss生成上的一致性和准确性，以及3D手语数字人的表现力，以下是一些潜在的改进方向：

- 同步参考标准化手语词库：模型在推理时需同步参考一个权威的手语词库，确保生成的Gloss与标准化手语一致。这有助于避免因语义相近而选择错误的Gloss，从而提高翻译的准确性。

- 优化Gloss顺序：通过扩充上下文学习样本库，并增加不同语境和复杂句子的训练样本，使模型能够学习到正确的Gloss顺序。此外，制定基于手语专家的建议和手语语法知识的翻译规则，以指导模型生成符合手语习惯的Gloss序列。

- 减少Gloss的缺失或多增问题：加强对句子结构的理解，通过强化语法校验机制，确保生成的句子结构完整且语义准确。同时，在生成过程中引入纠错机制，以及时发现并修正Gloss的缺失或多增问题(Yano and Utsumi, 2021)。

- 加强手语韵律的建模：在Gloss推理阶段，引入关于手语韵律的信息，如手势的节奏和力度(Inan et al., 2022)，以使数字人的手语表现更贴近真实的手语交流。

- 增强表情和口型的同步：通过增加表情和口型数据的训练量，并应用先进的同步算法，进一步优化表情和口型的同步技术，确保其能够准确反映手语的情感和意图。

- 语境理解的增强：利用向量库检索技术，提供与输入语句语义相近的上下文学习样本，以增强模型对语境的理解能力。这种技术可以帮助模型更好地捕捉到语句的深层含义和上下文联系，从而提高手语Gloss的生成能力。

总体来看，文心4.0在自动评测和人工评测中均表现出色，特别是在实际应用中的表现上显示出显著优势。这表明，结合LLM和3D动画技术的手语翻译系统在提高手语翻译准确性和用户体验方面具有巨大潜力。

组别	A	B	Ours	D
专家组	3.250	2.098	2.321	1.705
采集组	3.009	1.580	1.830	1.143
普通组	3.777	2.375	2.839	2.313

Table 6: 不同参赛队伍在手语翻译评测任务中的综合得分。

第二十三届中国计算语言学大会论文集，第393页-第404页，太原，中国，2024年7月25日至28日。
卷3：评测报告
(c) 2024 中国中文信息学会计算语言学专业委员会

5 相关工作介绍

手语数字人翻译属于手语生成任务，已有大量研究致力于开发高效、准确的翻译系统，以便更好地服务于听障人士。这些研究主要集中在手语翻译与生成两个方向，旨在通过技术手段实现自然语言到手语间的翻译。

5.1 手语翻译

在手语翻译方面，研究者们提出了多种方法来实现从自然语言到手语的翻译。早期的研究大多依赖于将自然语言转换为Gloss，然后再生成手语动作。例如，Jin(2022)、Zhu(2023)和Kan(2022)采用了这种两步法，通过Gloss作为中间层来实现翻译。这种方法的优点是能够利用大量的Gloss数据进行训练，从而提高翻译的准确性。然而，它也存在一些局限性，例如Gloss数据的获取和标注成本较高，并且Gloss本身不能完全捕捉到自然语言中的复杂语义信息。

为了克服这些局限，近年来有研究者尝试直接从自然语言生成手语，而不依赖于中间的Gloss表示。例如，Lin(2023)提出了一种基于端到端的翻译模型，通过直接将自然语言映射到手语动作序列来提高翻译的流畅性和自然度。类似地，Wong(2024)和Yin(2023)也提出了不同的无Gloss翻译方法，这些方法在处理复杂句子结构和捕捉上下文信息方面表现出色。

随着自然语言处理（NLP）领域的快速发展，手语翻译模型也在不断演进。早期的模型主要基于传统的循环神经网络（Recurrent Neural Network, RNN）和长短期记忆网络（Long Short-Term Memory, LSTM），如Guo(2018)提出的层次化LSTM模型，用于捕捉手语中的时间依赖关系。随着注意力机制和Transformer模型的引入，翻译效果得到了显著提升。例如，Cihan Camgoz(2020)提出的基于Transformer 的手语翻译模型，通过全局注意力机制能够更好地捕捉长距离依赖关系，提高了翻译的准确性和流畅性。

近年来，预训练语言模型（如BERT 和GPT）的成功应用，进一步推动了手语翻译技术的发展。Zhao(2023)利用BERT进行手语翻译，通过预训练模型的强大语义理解能力，显著提升了翻译性能。同时，大语言模型（LLMs）如GPT也开始应用于手语翻译领域。Gong(2024)和Wong(2024)探讨了将LLM应用于手语翻译的可能性，这些模型能够处理更加复杂和多样化的语言输入，使翻译结果更加自然和连贯。

手语翻译技术在不断发展，从早期的基于Gloss的两步法到如今的端到端翻译，从传统的RNN和LSTM模型到现代的Transformer和大语言模型，这些技术的进步极大地提高了手语翻译的准确性和自然性。未来的研究将继续探索如何利用更先进的模型和方法，进一步提高翻译质量，满足实际应用需求。

5.2 手语生成

近年来，手语生成和手语数字人技术逐渐受到广泛关注。手语数字人是通过3D建模和动画技术，实现手语动作逼真展示的虚拟人形形象，这一技术在辅助听障人士的交流中具有重要作用。例如，Lacerda(2023)开发了一种基于Unity 3D引擎的手语数字人系统，该系统能够实时模拟复杂的手语动作，实现了高度的准确性和自然度。

在手语生成领域，姿态估计和关键点推理技术发挥了关键作用。Zuo(2024)、Forte(2023)和Yu(2024)提出了基于姿态估计的方法，通过捕捉人体关键点，并将这些关键点渲染成数字人，达到了逼真的手语展示效果。尤其是Forte(2023)，他们的研究展示了如何通过优化关键点推理算法，提高手语动作的流畅性和自然度，使手语数字人在实际应用中更加实用和可靠。

此外，扩散模型在手语生成中的应用也取得了显著进展。Baltatzis(2024)提出了基于扩散模型的手语生成方法，通过端到端的训练实现了从自然语言到手语的直接翻译。该方法有效地结合了文本信息和视觉信息，使生成的手语更加连贯和自然，克服了传统方法中由于信息割裂导致的动作生硬问题。

另一类重要的方法是基于生成对抗网络（Generative Adversarial Networks, GAN）的手语生成系统。Saunders(2020a)提出了一种利用GAN进行手语生成的系统，通过对抗训练，使生成的手语视频更加逼真和自然。该系统不仅提高了手语动作的视觉效果，还增强了模型对不同手语表达的适应性和鲁棒性。

(c) 2024 中国中文信息学会计算语言学专业委员会

手语数字人技术的进展为手语翻译系统的发展提供了新的可能性。例如，Lakhfif(2020)研究了手语数字人在教育服务中的应用，证明了其在增强用户体验和提高交流效率方面的潜力。通过手语数字人，听障人士可以更直观地理解教育内容，极大地改善了学习效果。

本研究在现有工作的基础上，提出了一种创新性的手语翻译技术，融合了LLM的文本处理能力和3D动画技术的表现力。与以往工作相比，本系统不仅在翻译准确性上有所提升，更在手语动画的流畅性和表现力上实现了质的飞跃。通过深度语义理解，本系统能够深入分析复杂的语言结构和上下文信息，生成更准确和自然的手语翻译。此外，本系统设计的多样化上下文学习样本，确保了翻译结果在语义和语境上的双重精准，满足了多样化的交流场景需求。此外，本研究还特别关注了手语的视觉呈现，通过3D动画技术创造出既流畅又逼真的手语动作，这些动作在视觉上吸引人，并在表达上贴近自然手语的非言语特征，如节奏、力度和表情。系统的实时交互与反馈能力，以及多模态融合技术的应用，使得手语翻译更加生动和直观，极大地提升了用户体验。

未来的研究将集中在优化手语数字人的动作自然度和系统的响应速度，以满足实际应用的需求。这包括更精细的3D建模技术、更高效的动画生成算法，以及更智能的自然语言处理模型。随着技术的不断进步，手语数字人将在更多领域中发挥重要作用，如教育、医疗和公共服务，为听障人士提供更好的支持和服务。

6 结论

本研究通过使用大规模语言模型和3D动画技术，开发了一个高效的手语数字人翻译系统。实验结果表明，系统在准确性和实用性方面表现出色，特别是文心4.0在自动评测和人工评测中均表现出显著优势，显示了其在手语翻译中的高可靠性和准确性。本研究开发的手语翻译系统在提高手语翻译准确性和用户体验方面具有巨大潜力，为听障者提供了更加准确和便捷的沟通工具。未来的研究和优化工作将进一步推动手语翻译技术的发展和应用。

致谢

本研究受国家自然科学基金面上项目(No. 62376019, 61976015, 61976016, 61876198, 61370130)资助。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。

参考文献

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context Examples Selection for Machine Translation, December. arXiv:2212.02437 [cs].

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2Pose: Animating Sign Language Notation into Pose Sequences. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21046–21056, Vancouver, BC, Canada, June. IEEE.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report, September. arXiv:2309.16609 [cs].

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural Sign Actors: A diffusion model for 3D sign language production from text, April. arXiv:2312.02702 [cs].

Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. 2021. SIGN: Spatial-information Incorporated Generative Network for Generalized Zero-shot Semantic Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9536–9546, Montreal, QC, Canada, October. IEEE.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *2020 IEEE/CVF*

(c) 2024 中国中文信息学会计算语言学专业委员会

Conference on Computer Vision and Pattern Recognition (CVPR), pages 10020–10030, Seattle, WA, USA, June. IEEE.

Maria-Paola Forte, Peter Kulits, Chun-Hao Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. 2023. Reconstructing Signing Avatars From Video Using Linguistic Priors, April. arXiv:2304.10482 [cs].

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are Good Sign Language Translators, April. arXiv:2404.00925 [cs].

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical LSTM for Sign Language Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling Intensification for Sign Language Generation: A Computational Approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022. Prior Knowledge and Memory Enriched Transformer for Sign Language Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775, Dublin, Ireland. Association for Computational Linguistics.

Navroz Kaur Kahlon and Williamjeet Singh. 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 22(1):1–35, March.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 2022. Sign Language Translation with Hierarchical Spatio-Temporal Graph Neural Network. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2131–2140, Waikoloa, HI, USA, January. IEEE.

Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee, and Jong Park. 2022. Sign Language Production With Avatar Layering: A Critical Use Case over Rare Words. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1519–1528, Marseille, France, June. European Language Resources Association.

Inês Lacerda, Hugo Nicolau, and Luisa Coheur. 2023. Enhancing Portuguese Sign Language Animation with Dynamic Timing and Mouthing, July. arXiv:2307.06124 [cs].

Abdelaziz Lakhfif. 2020. Design and Implementation of a Virtual 3D Educational Environment to improve Deaf Education, May. arXiv:2006.00114 [cs].

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-Free End-to-End Sign Language Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.

Amit Moryossef. 2023. sign.mt: Real-Time Multilingual Sign Language Translation Application, October. arXiv:2310.05064 [cs].

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. 2023. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation, August. arXiv:2303.11089 [cs, eess].

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video, November. arXiv:2011.09846 [cs].

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive Transformers for End-to-End Sign Language Production. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12356, pages 687–705. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

(c) 2024 中国中文信息学会计算语言学专业委员会

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1899–1909, Montreal, QC, Canada, October. IEEE.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation, July. arXiv:2107.02137 [cs].

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance, June. arXiv:2211.09102 [cs].

Liming Wang, Junrui Ni, Heting Gao, Jialu Li, Kai Chieh Chang, Xulin Fan, Junkai Wu, Mark Hasegawa-Johnson, and Chang Yoo. 2023. Listen, Decipher and Sign: Toward Unsupervised Speech-to-Sign Language Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6785–6800, Toronto, Canada. Association for Computational Linguistics.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation, May. arXiv:2405.04164 [cs].

Pan Xie, Taiying Peng, Yao Du, and Qipeng Zhang. 2024. Sign Language Production with Latent Motion Transformer. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3012–3022, Waikoloa, HI, USA, January. IEEE.

Ken Yano and Akira Utsumi. 2021. Pipeline Signed Japanese Translation Focusing on a Post-positional Particle Complement and Conjugation in a Low-resource Setting. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2021–2032, Online. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. 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 7347–7360, Online. Association for Computational Linguistics.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss Attention for Gloss-free Sign Language Translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562, Vancouver, BC, Canada, June. IEEE.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark, April. arXiv:2310.20436 [cs].

Jan Zelinka and Jakub Kanis. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3384–3392, Snowmass Village, CO, USA, March. IEEE.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study, January. arXiv:2301.07069 [cs].

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. BEST: BERT Pre-training for Sign Language Recognition with Coupling Tokenization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3597–3605, June.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural Machine Translation Methods for Translating Text to Sign Language Glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. 2024. A Simple Baseline for Spoken Language to Sign Language Translation with 3D Avatars, January. arXiv:2401.04730 [cs].

CCL 2024

Overview of CCL24-Eval Task 10: Translation Quality Evaluation of Sign Language Avatar

Yuan Zhao[1*], Ruiquan Zhang[2,3*], Dengfeng Yao[1,4†], Yidong Chen[2,3†]
[1]Beijing Key Laboratory of Information Service Engineering, Beijing Union University
[2]School of Informatics, Xiamen University
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University
[4]Lab of Computational Linguistics, School of Humanities, Tsinghua University
{annzy,tjtdengfeng}@buu.edu.cn, rqzhang@stu.xmu.edu.cn, ydchen@xmu.edu.cn

Abstract

Sign Language Avatar technology aims to create virtual agents capable of communicating with deaf individuals through sign language, similar to the text dialogue agent ChatGPT but focusing on sign language communication. Challenges in sign language production include limited dataset sizes, information loss due to reliance on intermediate representations, and insufficient realism in generated actions. In this event, we particularly focus on the ability of the Sign Language Avatar to translate spoken language text into sign language that is easily understood by deaf individuals. As the first sign language avatar event held by the China National Conference on Computational Linguistics(CCL), this event attracted wide attention from both industry and academia, with 14 teams registering and 10 of them submitting their system interfaces on time. We provided a dataset consisting of 1074 text-video parallel sentence pairs for training, and the evaluation team comprised proficient Chinese sign language users and professional sign language translators. The scoring method employed a comprehensive evaluation based on multiple metrics, focusing primarily on sign language grammar accuracy, naturalness, readability, and cultural adaptability. The final scores were determined by considering performance across these four aspects. The final scores, taking into account these four aspects, showed that four teams demonstrated good readability, with Vivo Mobile Communication Co., Ltd. ranking first with a score of 3.513 (out of a full score of 5), leading the baseline model by 1.394 points. According to the analysis of the results, most teams used the traditional method of converting text into Gloss sequences before generating sign language. Additionally, some teams experimented with emerging methods, including gloss-free end-to-end training and Large Language Model(LLMs) prompt learning, which also achieved promising results. We anticipate that this event will promote the development of sign language avatar technology and provide higher-quality communication tools for the deaf community. For more information on this task, please visit the website of the CCL24-Eval: Translation Quality Evaluation of Sign Language Avatar Task[1].

1 Introduction

Sign language, a rich and complex form of communication with its own unique vocabulary and grammar, is used by over 70 million deaf and hard of hearing people worldwide. Unlike spoken languages, sign language emphasizes body language, incorporating hand shapes, movements, positions, and palm orientations, as well as non-manual elements like body posture and facial expressions (Qiu et al., 2018; Yao et al., 2019). Despite its widespread use and significance, the distinct differences and unique modes of expression in sign language pose challenges to its dissemination and understanding. The purpose of this evaluation competition is to assess sign language avatars that can translate spoken text into sign language, enhancing comprehension for deaf or hard of hearing individuals who use sign language.

[*] Co-First Author
[†] Corresponding Author
[1]Our task website:https://github.com/ann-yuan/QESLAT-2024

©2023 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

Recently, sign language research has become an active area within Computer Vision (CV) and Natural Language Processing (NLP)(Yin et al., 2021; Yu et al., 2023; Zhao et al., 2024), particularly in Sign Language Recognition (SLR)(Chen et al., 2022; Wei et al., 2023; Hu et al., 2023) and Translation (SLT)(Fu et al., 2023; Sun et al., 2024; Hu et al., 2024). However, research publications in Sign Language Production (SLP), which are closely related to this evaluation task, are scarce (Rastgoo et al., 2021; Yao, 2022).

The goal of SLP is to translate spoken or written content into sign language, making it accessible for the deaf. SLP faces challenges such as capturing detailed hand and body movements and dealing with unique visual semantics and grammar. Additionally, limited datasets, difficulties in simulating sign details, and technological constraints in producing contextually accurate signs complicate SLP tasks(Ren et al., 2024).

Early sign language processing (SLP) methods relied on one-to-one gloss-sign correspondences, producing only isolated sign words (Stoll et al., 2018). Subsequent attempts employed machine translation techniques to translate spoken text into continuous gloss sequences, which were then converted to sign language gestures (Saunders et al., 2021; Zhu et al., 2023). (Saunders et al., 2021; Zhu et al., 2023). However, this approach often failed to capture contextual information, leading to semantic losses. A more recent development is the end-to-end SLP method, which directly translates text into sign language videos without intermediate gloss, showing significant improvements with larger data volumes (Baltatzis et al., 2023).

Regarding avatar technologies, initial methods employed 2D or 3D skeletal models extracted from videos (Kapoor et al., 2021; Saunders et al., 2020). Recent advancements have focused on generating sign language from these models through rendering techniques (Saunders et al., 2020; Zelinka and Kanis, 2020; Xiao et al., 2020). The use of pretrained avatar models like SMPL-X has also been explored for enhanced sign language representation (Pavlakos et al., 2019; Stoll et al., 2022).

In China, sign language avatars are gaining traction. Initiatives like ZHIPU's "AI Sign Language Classmate" and vivo's "Sign Language Translator" highlight a deep understanding and proactive response to the deaf community's communication needs. These avatars are now visible at events such as sports games and news programs. With ongoing technological progress, sign language avatars are expected to play a vital role, enhancing communication for the deaf community.

This evaluation assesses the effectiveness of sign language avatars by using recorded videos and multiple-choice questions for feedback. We expect this round to enhance the avatar team's understanding of the deaf community's needs and pinpoint improvement areas.

2 Task Description

The purpose of this evaluation is to assess the naturalness and accuracy of sign language avatars translating Chinese into Chinese sign language(CSL), ensuring that the translations adhere to sign language grammatical rules and are understandable and acceptable to the deaf community. During the evaluation, we collect and construct a rich corpus covering various common scenarios for teams to train on. Additionally, we design a series of test sentences, four in total, each accompanied by the participating team's sign language avatar videos, along with four multiple-choice questions and optional evaluations. Each multiple-choice question provides four options, allowing evaluators to choose during the evaluation process to determine which option best matches the performance of the sign language avatar.

3 Evaluation Method

In the fields of SLR and SLT, common metrics such as Accuracy, Recall, Word Error Rate (WER), and translation-specific indicators like BLEU, CIDEr, ROUGE, METEOR often require comparison with ground truth videos, which are difficult to obtain(Rastgoo et al., 2021; Rastgoo et al., 2022). Therefore, automated machine evaluation is not used. Additionally, the diversity in styles of sign language avatars complicates the use of a standardized automated evaluation method. Instead, human evaluation is preferred due to its flexibility, accuracy, and thorough analytical capabilities, making it the method of choice for assessing the naturalness and fluency of sign language translations.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Multiple-choice: A Specific Example	
Digital Avatar Video Example	已经为您办理完成，这是您的机票。请问需要其他帮助吗? *I have completed the process for you. Here is your plane ticket. Do you need any other help?*
Accuracy Assessment Question	Please assess whether the digital avatar accurately performed the key signs for "办理(process)", "机票(plane ticket)", and "帮助(help)".
A. All included (5 points)	The signs are complete and precise, clearly including all the signs for "办理(process)", "机票(plane ticket)", and "帮助(help)".
B. Partially included (3 points)	The signs are partially correct, including at least 1 or 2 of the specified signs.
C. Signs semantically incorrect (1 point)	There are sign changes, but these signs do not accurately reflect the specific semantics of "办理(process)", "机票(plane ticket)", and "帮助(help)".
D. Signs unclear and unreadable (0 points)	The sign language does not correspond to the provided content, or the signs are too unclear to be recognizable.

Table 1: A specific example in multiple-choice questions

Participating teams are required to provide a sign language avatar interface for recording videos corresponding to the competition's test scenarios. These translation results will be scored by an evaluation team, consisting of 20 deaf individuals and professional translators certified by the Committee for Sign Language Research and Promotion of the Chinese Deaf Association, with each evaluation metric having a maximum score of 5 points. The final scores for each team will be calculated based on the feedback and scores from the evaluators, ensuring a comprehensive and fair assessment.

Next, we will introduce the evaluation metrics and scoring methods for this competition.

3.1 Metrics

Sign Language Grammar Accuracy Sign language grammar accuracy refers to the translation adhering to the semantics and grammatical rules of the target language. The translation must follow the word order and structural rules of CSL, as the subject-verb-object structure of Mandarin may need to be adjusted to a sequence more customary in sign language. The correctness of gesture forms is also crucial, ensuring the appropriate use of finger positions, palm directions, and hand movements. Additionally, grammatical markers in sign language, such as tense, negation, and questions, must also be accurately expressed in the translation, which is vital for conveying the complete meaning of sentences.

Naturalness Naturalness emphasizes the fluency and naturalness of the translation, making the sign language close to the natural communication methods of the deaf community. The gestures in the translation must be coherent and fluid, mimicking the fluency of natural sign language, avoiding stiff and disjointed movements. It is also crucial to assess whether the translation conforms to the everyday expressive habits of the deaf community, including the accurate application of common sign language phrases and idiomatic expressions. Non-verbal elements, such as facial expressions, body postures, and

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

spatial arrangement, should also be naturally integrated to achieve more natural and expressive communication.

Readability Readability ensures that the expression of sign language is clear and easy to understand, promoting effective communication. The clarity of gestures is crucial; they must be clear enough for observers to understand easily while avoiding any ambiguous movements that could cause confusion. Consistency is also key, ensuring that the same concept or vocabulary is expressed with the same gesture in different contexts, which helps observers better understand and remember the meaning of the sign language. Adaptability is also critical; translations need to adjust gestures and expressions according to different contexts to ensure effective communication.

Cultural Adaptability Cultural adaptability focuses on the suitability and accuracy of the translation across different cultural backgrounds, avoiding cultural misunderstandings. The translation must carefully consider cultural differences to avoid direct translations that might lead to cultural misunderstandings or inappropriate expressions. The translation of sign language avatars must not only be accurate on a literal level but also appropriate culturally, ensuring that the message is correctly understood across different cultural contexts. It also needs to adapt to specific social contexts, such as the appropriate use of polite expressions and industry-specific terminology. The accurate conveyance of emotional tones is also an important aspect of cultural adaptability; translations need to capture and convey the emotional aspects of the original text, such as sarcasm or humor, which is crucial for ensuring comprehensive message delivery and emotional resonance with the receiver.

3.2 Scoring Method

This assessment primarily utilizes manual evaluation to comprehensively assess the performance of digital sign language avatars across various indicators. Each indicator is scored from 1 to 5, where 5 is the best and 1 is the worst. We have designed multiple multiple-choice questions to facilitate the judges in scoring. A specific example in Table 1.

The total score is the arithmetic mean of all scores, excluding one highest and one lowest score. Assuming a set of scores $X = x_1, x_2, ..., x_n$, the specific calculation formula is

$$R = \frac{\sum_{x_i \subset X'} x_i}{n-2} \tag{1}$$

where n is the number of evaluators, and X' is the set of scores excluding the highest score $max(X)$ and the lowest score $min(X)$.

In the overall evaluation framework, the distribution of weights for individual indicators is as shown in Table 2. The composite score is calculated as the weighted sum of the scores for each indicator and their respective weight coefficients. The evaluation focuses on the sign language avatars' ability to accurately express sign language, emphasizing naturalness and readability while also considering cultural adaptability.

Metric Type	Metric Name	Weight Proportion(%)
Human Evaluation of Translation Quality	Sign Language Grammar Accuracy	30%
	Naturalness	25%
	Readability	25%
	Cultural Adaptability	20%

Table 2: Metrics corresponding to the human evaluation of translation quality

4 Dataset

Sign language, as a minority language, typically has a relatively small corpus, which is a common challenge in the field of SLR and SLT. Currently, among publicly available papers, the University of Science and Technology of China(USTC) holds two datasets: the USTC-CCSL dataset (Huang et al., 2018), which contains 25,000 sentences recorded by 50 sign language demonstrators; and the CSL-Daily dataset

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Corpus	Count	Format	Owner	Scenario
XMU-CSL	500	Word-level	Xiamen University	Hospital, Services
BUU-CSL	500	Word-level, Gesture-level	Beijing Union University (College of Special Education)	Shopping, Dining, Accommodation, Tourism, Finance, Hospital, Security, Transportation, Legal, Employment, Public, Government Services, etc.
ZZSZY-CSL	74	Gesture-level	Zhuzhou Voice of Hand Information Technology Co., Ltd.	Legal, Services, Hospital, Services

Table 3: Datasets for the current evaluation task

(Zhou et al., 2021), containing 20,654 sentences recorded by 10 sign language demonstrators. The University of the Chinese Academy of Sciences(UCAS) also offers the RCSD dataset (Wang et al., 2019), though the exact number of videos has not been publicly disclosed, also recorded by 10 sign language demonstrators. Notably, the aforementioned three datasets are proprietary, and usage requires applying under the name of a university or research institute, which poses challenges for the evaluation task's progress.

To address evaluation challenges, Table 3 outlines a corpus of 1,074 sentences with corresponding videos, provided by Xiamen University, Beijing Union University (College of Special Education), and Zhuzhou Voice of Hand Information Technology Co., Ltd. Initially, a text-based annotated corpus was supplied to meet the computational language processing needs of the teams. As training progressed, a composite dataset incorporating both text and video was introduced to support more comprehensive development in sign language avatar technology.

The training dataset used for this evaluation comprises 1,074 sentences from daily life scenarios relevant to the deaf community, such as transportation and medical services. It includes inputs from deaf individuals for whom sign language is the primary mode of communication, ensuring the data's authenticity and representativeness. The dataset incorporates written language, corresponding videos, and sign language glosses, with annotations in both word-level and gesture-level formats to provide expressive richness. These formats comply with the T/CADHOH0004-2023 standard specifications for Intelligent Sign Language Translation System Test. The XMU-CSL dataset features word-level annotations and is focused on hospital services. As shown in Table 3, the BUU-CSL dataset includes both annotation types, covering a broad range of practical scenarios, from shopping to government services. Meanwhile, the ZZSZY-CSL dataset, which uses gesture-level annotations, is tailored for legal and hospital services.

In the evaluation corpus for this event, specific scenarios were selected, and corresponding written language texts along with their sign language glosses are provided as reference examples. Table 4 includes part-of-speech tagging and gesture-level annotations for each reference example[2]. Additionally, sign language video demonstrations of the reference examples are provided[3] [4], with particular attention

[2]Note: Marks [1] denote words that are the same but have different meanings; marks ①② denote words that are the same in both term and meaning, but differ in sign language actions. Other reference materials for sign language include the 'National Common Sign Language Common Words List' and the 'National Common Sign Language Dictionary' APP, among others.

[3]Ex.1 sign language video: https://github.com/ann-yuan/QESLAT-2024/blob/main/video1.gif

[4]Ex.2 sign language video: https://github.com/ann-yuan/QESLAT-2024/blob/main/video2.gif

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	Written Language Texts	Sign Language Glosses	
Example 1 [3]	女儿可能生病了，快带她去医院。 My daughter may be sick. Take her to the hospital.	**Word-level**	女儿②/病/可能②，带[1]/医院/速度 daughter②/sick/might②, take[1]/hospital/quickly
		Gesture-level	女-矮/病/可能②【疑问】，带[1]/医生-家/速度 female-short/sick/might②【doubtful】,take[1]/doctor-home/quickly
Example 2 [4]	为了赢得比赛，儿子一直在专心训练。 In order to win the competition, my son has been concentrating on his training.	**Word-level**	为/比赛/赢②，儿子②/专心/训练/一直 in order to/competition/win②, son/dedicated/train/always
		Gesture-level	为/比较/胜利②，男-矮/认真【眼睛同时向下看】-心/练习/一直 in order to/compare/win②, son-short/careful[eyes looking down at the sametime]-heart/practice/all the time

Table 4: Reference examples of written language texts and sign language glosses

to the changes in facial expressions.

5 Registration and Evaluation Results

In this evaluation event, 14 teams, including 9 from academia and 5 from industry, registered and applied for the corpus, highlighting widespread interest in sign language avatar technology. By the submission deadline, 10 teams had submitted their interfaces. The event was supported by 18 professional evaluators from the Chinese Association of the Deaf's Committee for Sign Language Research and Promotion, split into an expert group and a collection group. The expert group, proficient in sign language, established benchmarks and facilitated comparisons across different groups. Meanwhile, the collection group, composed of members from 12 different regions, was responsible for identifying regional variations in sign language. The event also included 9 general evaluators—comprising university students, working adults, and retirees. A notable focus was placed on 24 experienced sign language users over the age of 35, 18 of whom are native users. Their extensive use and deep understanding of sign language provided crucial insights into its natural fluency, accuracy, and cultural nuances, thereby ensuring the evaluation's reliability and enhancing the assessment of the avatars' capability to capture nuanced and emotional expressions in sign language.

After preliminary screening by the evaluation team, we found that the system interfaces submitted by 4 of the teams listed in Table 5 not only function effectively but also well reflect the core characteristics of sign language avatars. Table 6 provides a more detailed breakdown of the scores. This outcome indicates that, despite many challenges, the participating teams have made significant progress in the development and application of sign language avatars. However, the performances of the other 6 teams were not satisfactory. Two of these teams' systems could produce sign language sequences, but upon preliminary review, these sequences did not comply with sign language grammar and were unsuitable for scoring by evaluators. Given that these two teams could produce sign language sequences, we awarded them 1 point

CCL 2024

Rank	Competing Team	Weighted Total Score
1	VIVO: Vivo Mobile Communication Co., Ltd.	3.513
2	GBAT: GoBetter AccessTech (SuZhou) Co., Ltd.	2.447
3	BJTT: Beijing Tian Tang Technology Co., Ltd. **(Baseline)**	2.119
4	QDHDT: Qingdao Heshi Digital Technology Co., Ltd.	1.806

Table 5: Effective Team Rankings

Expert Group Scores (9 people)				
Sign Language Avatar	Accuracy	Naturalness	Readability	Cultural Adaptability
VIVO	3.50	3.75	3.43	2.36
GBAT	2.21	2.36	2.00	1.79
BJTT	2.61	2.25	2.21	2.07
QDHDT	1.04	2.14	1.68	1.75
Collection Group Scores (9 people)				
Sign Language Avatar	Accuracy	Naturalness	Readability	Cultural Adaptability
VIVO	3.39	3.43	2.86	2.43
GBAT	1.11	2.07	1.86	1.29
BJTT	2.14	2.07	1.96	1.11
QDHDT	0.25	1.79	1.54	0.82
General Group Scores (9 people)				
Sign Language Avatar	Accuracy	Naturalness	Readability	Cultural Adaptability
VIVO	4.14	4.14	3.75	3.11
GBAT	1.75	2.89	2.43	2.50
BJTT	2.93	2.89	2.89	2.57
QDHDT	1.79	2.86	2.54	2.04

Table 6: Individual scores to each team given by three groups of evaluators

for encouragement in naturalness and ranked them jointly in fifth place. The other four teams, due to repetitive content that did not meet the requirements of our evaluation task, were given zero points and ranked at the bottom.

In a sign language avatar evaluation, Beijing Tian Tang Technology Co., Ltd. (BJTT) was used as the baseline team, using a classical algorithm with two Transformers for text-to-gloss and gloss-to-3D motion translation. While BJTT showed moderate success in the collection group, it rated higher in naturalness in the general group.

Vivo Mobile Communication Co., Ltd. (VIVO) topped the competition with a score of 3.513, outperforming BJTT by 1.394 points. VIVO's success was credited to its pre-trained multilingual model fine-tuning, back-translation for data augmentation, and strategies to smooth animation, leading to high marks in accuracy, naturalness, and readability from general users.GoBetter AccessTech (SuZhou) Co., Ltd. (GBAT) placed second with a score of 2.447, performing slightly above the baseline and noted for its accuracy and naturalness by general users. Qingdao Heshi Digital Technology Co., Ltd. (QDHDT) ranked fourth, facing difficulties in accuracy and readability in the collection group.

In the evaluation, VIVO excelled, while GBAT, BJTT, and QDHDT showed potential for improvement in accuracy and cultural adaptability. Future efforts should aim at enhancing sign language avatar performance and user satisfaction.

To gather specific feedback, optional evaluations were added to the questionnaire for each avatar. VIVO's avatar was praised for its clear, natural movements and coordinated expressions, though it could improve in expression richness and contextual adaptability. GBAT's avatar stood out for clear hand-

CCL 2024

shapes but needs better movement naturalness, expression richness, and sign language fluency. BJTT's avatar was noted for clear hand movements and expressions but required improvements in accuracy and naturalness. QDHDT's avatar was recognized for good visual design and movement coordination but needs better vocabulary accuracy, expression fluency, and non-manual element depiction.

In summary, each team has made certain progress in the development of sign language avatars, but they also face challenges in accuracy, naturalness, and cultural adaptability. Future research and development should focus on these challenges to continuously optimize the technology, allowing sign language avatar to naturally display sign language translations.

6 Overview of Methods Used by Participating Teams

Currently, the participating teams generally adopt the text-gloss-video technology route. This method first uses a translation model to convert spoken text into gloss. Then, it retrieves corresponding videos from a gloss-video database and uses smoothing techniques to produce fluid, continuous sign language videos. Figure 1 shows the basic process of generating Avatar for sign language translation Below, we will discuss the methods from two aspects: Text2Gloss translation and avatar synthesis.

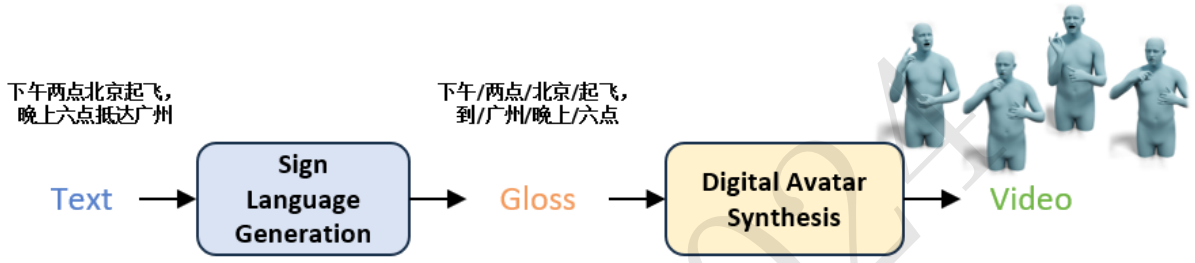


Figure 1: Flowchart of Sign Language Avatar Translation System

6.1 Text2Gloss Translation

Text2Gloss involves translating text into gloss sequences for sign language videos, traditionally using rule-based systems or models trained on text-gloss datasets.However, current methods face several limitations, including dynamic context handling, information loss, data scarcity, and generalization of large models. To address these issues, the participating teams have developed various strategies:

Dynamic Context Handling: Traditional text-to-gloss methods often falter with dynamic contexts, where the meaning of the text depends heavily on its surrounding context. VIVO has fine-tuned a pre-trained multilingual model (mRASP) with one million text-gloss parallel sentences, exploiting its multilingual capabilities for superior context comprehension. This method enhances translation accuracy by considering the context across multiple languages. In a similar vein, BJTT utilized a Transformer model trained on 300,000 parallel sentences, incorporating syntactic tree-based methods to improve contextual understanding, thus preserving the contextual coherence in the gloss translation.

Information Loss: The process of converting text to gloss often leads to information being lost, especially when direct equivalents in sign language are absent. VIVO countered this by implementing a back-translation technique for data enrichment, creating numerous gloss pseudo labels to ensure data variety. This strategy boosts the model's resilience by introducing a broader range of data, minimizing information loss. Conversely, QDHDT adopted an end-to-end approach by integrating text encoding with a pre-trained BERT model and a decoding process via an LSTM sequence model to produce 3D animations directly. This method effectively preserves more information by omitting the intermediate gloss translation.

Data Scarcity: The scarcity of sign language datasets poses a significant challenge for model training and generalization. VIVO addressed this by using a back-translation strategy for data augmentation, significantly expanding the training dataset. BJTT employed data augmentation techniques, including using ChatGPT to generate additional training data. These approaches are effective as they increase the

volume and diversity of training data, which is crucial for training robust models. GBAT prepared high-quality datasets and utilized few-shot learning techniques, leveraging the strong performance of LLMs on small datasets. This method allows the model to learn effectively even with limited data, enhancing its generalization capability.

Generalization of Large Models: While large models perform well on many tasks, their generalization ability in specific domains is still limited. GBAT enhanced generalization by meticulously selecting and preparing parallel sentence pairs, using large language models for few-shot learning to build an agent specifically for text-gloss conversion. This approach leverages the adaptability of large language models to learn specific domain tasks effectively. QDHDT explored end-to-end methods to directly generate 3D sign language animations from text, aiming to overcome generalization issues by avoiding intermediate gloss steps, thus creating a more direct and potentially more accurate translation process.

By addressing these challenges, the participating teams have made significant advancements in Text2Gloss translation and avatar synthesis, contributing to more accurate and fluid sign language translation systems.

6.2 Avatar Synthesis

Avatar synthesis includes action synthesis and smoothing, as well as action rendering. Competing teams mapped gloss to 3D skeleton videos, then used rendering technologies to bind the skeletons to specific avatars to produce the final avatar videos.

VIVO used motion capture technology to acquire common gloss gestures, joint rotation angles, and other skeletal information. They applied a general method proposed by Slot (Lin et al., 2020) for action fusion, achieving seamless transitions between two actions. This method extracts several frames from both the preceding and following skeleton videos, calculates the spatial distances between skeletal joints frame by frame, constructs a cost matrix, and finds the skeleton synthesis plan by calculating the path of minimum total cost.

GBAT employed a more refined approach to constructing individual gloss skeleton videos with 3D keyframe animation. They used keyframe technology to capture critical states and transitions in motion, achieving continuous and natural transitions while solving the challenge of smooth transitions between different sign language videos. In rendering, the team used ThreeJS to implement the sign language avatar, supporting WebGL1 API, suitable for both PC and mobile platforms.

Unlike the first two teams, BJTT used an end-to-end Transformer model to directly generate 3D skeleton videos from text, creating semantically consistent and smoothly acted videos without needing retrieval or smoothing. This approach robustly handles complex sign expressions but is constrained by data and computational resources. Meanwhile, QDHDT optimized animation with state machines and montage techniques, boosting efficiency through graphical state management and flexible code control.

7 Conclusion and Future Prospects

This evaluation involved 14 teams—9 from academia and 5 from industry—highlighting significant interest in sign language avatar technology. By the submission deadline, 10 teams had successfully submitted their interfaces, with 4 advancing past preliminary screening due to their effectiveness and readability.

VIVO stood out by ranking first, excelling across expert, collection, and general groups. Despite all teams following a text-gloss-video route, varied approaches were seen in Text2Gloss translation and avatar synthesis, including fine-tuning pre-trained models, utilizing LLM prompts, ensuring smooth transitions, and implementing end-to-end synthesis.

Evaluators consistently found that while some avatars showcased fluent, clear, and accurate sign language expressions, many still had issues like missing information, inaccurate gestures, and stiff movements. These shortcomings impact the accuracy, readability, and overall viewing experience. Future work should concentrate on refining these technologies to better serve sign language users, aiming for continuous improvement and enhanced user satisfaction.

- **Improving accuracy and readability:** Further optimize Text2Gloss translation and avatar synthesis methods to ensure accurate conveyance of sign language information.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

- **Enhancing naturalness and cultural adaptability:** Enhance the simulation of sign language expressions, rhythms, and intonations, taking into account the unique sign language habits and cultural characteristics of various regions and groups to better cater to users' communication needs.
- **Exploring more methods:** Explore advanced avatar production methods, including direct end-to-end models, generation strategies like diffusion, and the integration of reinforcement learning or transfer learning to enhance the efficiency and quality of sign language avatar production.
- **Continual improvement and optimization:** Continuously adjust and improve system interfaces to adapt to changing user needs and evolving technologies.

We look forward to sign language avatars serving the deaf community better in the future and providing them with a more convenient and friendly communication experience.

8 Acknowledgements

This research was supported by the National Natural Science Foundation of China [62036001; 62076211]; National Social Science Foundation of China [21BYY106]; General Project of the National Language Committee [YB145-25]; and the Support Plan for Beijing Municipal University Faculty Construction - High-Level Scientific Research and Innovation Team Project [BPHR20220121].

References

Yunfeng Qiu, Dengfeng Yao, Rong Li, and Chunda Liu. 2018. *Introduction to Chinese Sign Language Linguistics*. China International Broadcasting Press.

Dengfeng Yao, Minghu Jiang, Hong Bao, Hanjing Li, and others. 2019. Thirty Years Beyond Sign Language Computing: Retrospect and Prospect. *Chinese Journal of Computers*, volume 42, number 1, pages 111–135.

Tianyu Ren, Dengfeng Yao, Chaoran Yang, and Xinchen Kang. 2024. The Influence of Chinese Characters on Chinese Sign Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, volume 23, number 1, pages 1–31.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

Pei Yu, Liang Zhang, Biao Fu, and Yidong Chen. 2023. Efficient sign language translation with a curriculum-based non-autoregressive decoder. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5260–5268.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. *Conditional variational autoencoder for sign language translation with cross-modal alignment*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, number 17, pages 19643–19651.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. *Continuous sign language recognition with correlation network*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539.

Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2023. A token-level contrastive framework for sign language translation. In *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tong Sun, Biao Fu, Cong Hu, Liang Zhang, Ruiquan Zhang, Xiaodong Shi, Jinsong Su, and Yidong Chen. 2024. Adaptive Simultaneous Sign Language Translation with Confident Translation Length Estimation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 372–384.

Cong Hu, Biao Fu, Pei Yu, Liang Zhang, Xiaodong Shi, and Yidong Chen. 2024. An Explicit Multi-Modal Fusion Method for Sign Language Translation. In *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3860–3864.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3451–3461.

Dengfeng Yao. 2022. *A Guide to Sign Language Computing*. Science Press.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. *Sign language production using neural machine translation and generative adversarial networks*. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. *Mixed signals: Sign language production via a mixture of motion primitives*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. *Neural Machine Translation Methods for Translating Text to Sign Language Glosses*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2023. *Neural Sign Actors: A diffusion model for 3D sign language production from text*. *arXiv preprint arXiv:2312.02702*.

Parul Kapoor, Rudrabha Mukhopadhyay, Sindhu B. Hegde, Vinay Namboodiri, and C. V. Jawahar. 2021. *Towards automatic speech to sign language generation*. *arXiv preprint arXiv:2106.12790*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. *Progressive transformers for end-to-end sign language production*. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer.

Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403.

Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, volume 125, pages 41–55. Elsevier.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. *Expressive body capture: 3d hands, face, and body from a single image*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985.

Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. 2022. *There and back again: 3d sign language generation from text using back-translation*. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, volume 164, pages 113794. Elsevier.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. 2022. All You Need In Sign Language Production. *arXiv preprint arXiv:2201.01609*.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, number 1.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving Sign Language Translation With Monolingual Data by Sign Back-Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, June.

Hanjie Wang, Xiujuan Chai, and Xilin Chen. 2019. A novel sign language recognition framework using hierarchical grassmann covariance matrix. *IEEE Transactions on Multimedia*, volume 21, number 11, pages 2806–2814.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. *Pretraining multilingual neural machine translation by leveraging alignment information*. arXiv preprint arXiv:2010.03142.

CCL 2024