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)

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

```
# ::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

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.

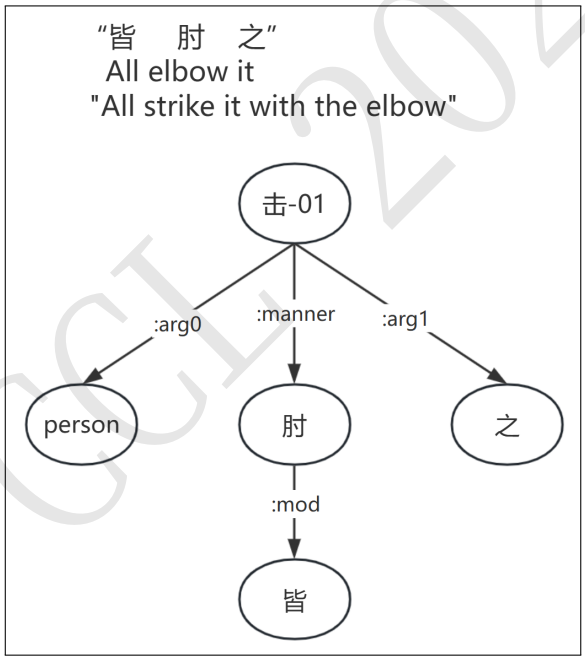

Figure 2: Example of noun used as verb in Ancient Chinese

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

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

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

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.

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,

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-

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.

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.

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