Overview of CCL23-Eval Task 2: The Third Chinese Abstract Meaning Representation Parsing Evaluation

Zhixing Xu[1], Yixuan Zhang[1], Bin Li[1], Junsheng Zhou[2] and Weiguang Qu[2]
1. School of Chinese Language and Literature, Nanjing Normal University, China
2. 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 emerged as a prominent area of research in sentence-level semantic parsing within the field of natural language processing in recent years. Substantial progress has been made in various NLP subtasks through the application of AMR. This paper presents the third Chinese Abstract Meaning Representation Parsing Evaluation, held as part of the Technical Evaluation Task Workshop at the 22nd Chinese Computational Linguistics Conference. The evaluation was specifically tailored for the Chinese and utilized the Align-smatch metric as the standard evaluation criterion. Building upon high-quality semantic annotation schemes and annotated corpora, this evaluation introduced a new test set comprising interrogative sentences for comprehensive evaluation. The results of the evaluation, as measured by the F-score, indicate notable performance achievements. The top-performing team attained a score of 0.8137 in the closed test and 0.8261 in the open test, respectively, using the Align-smatch metric. Notably, the leading result surpassed the SOTA performance at CoNLL 2020 by 3.64 percentage points when evaluated using the MRP metric. Further analysis revealed that this significant progress primarily stemmed from improved relation prediction between concepts. However, the challenge of effectively utilizing semantic relation alignments remains an area that requires further enhancement.

1 Introduction

With the growing maturity of morphological analysis and syntactic analysis techniques, natural language processing in general has advanced to semantic analysis level. Sentence-level meaning parsing, to be more specific, has already occupied the core position of semantic analysis research. To address the lack of whole-sentence semantic representation and the domain-dependent problem of sentence semantic annotation, Banarescu et al. (2013) proposed a domain-independent whole-sentence semantic representation method called Abstract Meaning Representation (AMR) that can abstract the meaning of a sentence with a single-rooted, acyclic and directed graph and predicts the semantic structure of the targeted sentence. There have been large-scaled corpora constructed for AMR and two international conferences held for AMR semantic parsing evaluation tasks. The latest one was CoNLL 2020, where there have been five languages in cross-lingual track including Chinese. And yet parsing Chinese via AMR was not flawless given that Chinese Mandarin differs a lot from English in terms of syntax and semantics. Li et al. (2016) therefore introduced several major changes into Chinese Abstract Meaning Representation (Chinese AMR, CAMR) so as to better parse Chinese. And similar to AMR, the corpus of CAMR has also begun to take shape and played an important role in the stage of CoNLL 2020.

2 Evaluation Task

Our evaluation task is to parse input sentences and output AMR graphs of the targeted sentences with data from CAMR corpus. It is noteworthy that the alignment of concept and relation are added in CAMR and some extra semantic role labels as well to better distinguish characteristics in Chinese. The evaluation task at CoNLL 2020, however, failed to leverage the alignment of concept and relation. Therefore, in our former CAMRP 2022 evaluation task, we adopted the newly-designed metric named Align-smatch, which contains the alignment of concept and relation, aiming to better evaluate the performance of automatic parsing. CAMRP 2023 is a follow-up and extension of CAMRP 2022, with key difference including the addition of a blind test set with 2,000 interrogative sentences.

3 Data Set

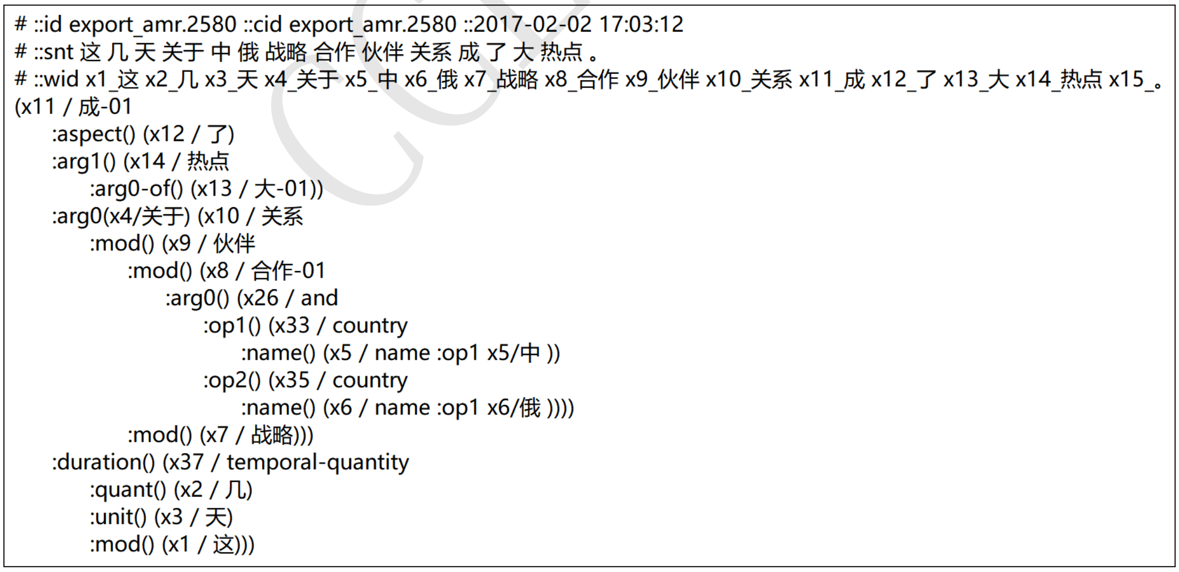
CAMR Corpus has been constructed and co-operated by Nanjing Normal University and Bradeis University since 2015 (Li et al., 2016) (Li et al., 2019). Specifically, the data provided at CAMRP 2023 is the CAMR v2.0 released via Linguisitc Data Consortium (LDC), of which the original data was from Chinese Tree Bank 8.0 including 20,000 Chinese sentences in total. The data sets as usual include training set, dev set and test set, and have been proven with high quality in the evaluation task at CAMRP 2022. We hereby use the exact same data sets in order to see whether there is any progression of CAMR parsing in recent two years. Newly added blind set (Test C) including 2,000 sentences is also provided to measure the generalization performance of parsers. Table 1 shows the distribution of each data set.

Data Set	Sentences	Word Tokens
Train Set	16,576	386,234
Dev Set	1,789	41,822
Test A	1,713	39,228
Test B	1,999	36,940
Test C	2,000	18,909

Table 1: Data set distribution

3.1 Data Format

The data sets we offer are in three different formats, which include the following representations: raw text annotations, dependency analysis results, and tuples.



```
# ::id export_amr.2580 ::cid export_amr.2580 ::2017-02-02 17:03:12
# ::snt 这 几 天 关于 中 俄 战略 合作 伙伴 关系 成 了 大 热点 。
# ::wid x1_这 x2_几 x3_天 x4_关于 x5_中 x6_俄 x7_战略 x8_合作 x9_伙伴 x10_关系 x11_成 x12_了 x13_大 x14_热点 x15_。
(x11 / 成-01
    :aspect() (x12 / 了)
    :arg1() (x14 / 热点
        :arg0-of() (x13 / 大-01))
    :arg0(x4/关于) (x10 / 关系
        :mod() (x9 / 伙伴
            :mod() (x8 / 合作-01
                :arg0() (x26 / and
                    :op1() (x33 / country
                        :name() (x5 / name :op1 x5/中 ))
                    :op2() (x35 / country
                        :name() (x6 / name :op1 x6/俄 ))))
            :mod() (x7 / 战略)))
    :duration() (x37 / temporal-quantity
        :quant() (x2 / 几)
        :unit() (x3 / 天)
        :mod() (x1 / 这)))
```

Figure 1: Sample of CAMR text representation

Figure 1 is a copy of CAMR text representation sample from 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. Translation of the original sentence is “这*/this* 几*/several* 天*/day* 关于*/about* 中*/China* 俄*/Russian* 战略*/strategy* 合作*/cooperation* 伙伴*/companion* 关系*/relationship* 成*/become* 了*/already* 大*/big* 热点*/hot-spot*”, which means “*In the past few days, the strategic partnership between China and Russian has become a hot topic*”.

ID	Token	Part-of-Speech	Head word	Head word ID	Dependency
1	这	DT	成	11	dep
2	几	CD	天	3	nummod
3	天	M	这	1	dep
4	关于	P	成	11	prep
5	中	NR	伙伴	9	nn
6	俄	NR	伙伴	9	nn
7	战略	NN	伙伴	9	nn
8	合作	NN	伙伴	9	nn
9	伙伴	NN	关系	10	nn
10	关系	NN	关于	4	pobj
11	成	VV	root	0	root
12	了	AS	成	11	asp
13	大	JJ	热点	14	amod
14	热点	NN	成	11	dobj

Table 2: Sample of dependency analysis result

Table 2 is a copy of dependency analysis result. Note that in the closed modality, participants are allowed to use dependency analysis results as the external resource for training.

句子编号	节点编号1	概念1	关系	关系编号	关系对齐词	节点编号2	概念2
sid	nid1	concept1	rel	rid	ralign	nid2	concept2
2580	x0	root	:top	-	-	x11	成-01
2580	x11	成-01	:aspect	-	-	x12	了
2580	x11	成-01	:arg1	-	-	x14	热点
2580	x11	成-01	:arg0	x4	关于	x10	关系
2580	x11	成-01	:duration	-	-	x37	temporal-quantity
2580	x14	热点	:arg0-of	-	-	x13	大-01
2580	x10	关系	:mod	-	-	x9	伙伴
2580	x9	伙伴	:mod	-	-	x8	合作-01
2580	x9	伙伴	:mod	-	-	x7	战略
2580	x8	合作-01	:arg0	-	-	x26	and
2580	x26	and	:op1	-	-	x33	country
2580	x26	and	:op2	-	-	x35	country
2580	x33	country	:name	-	-	x5	中
2580	x35	country	:name	-	-	x6	俄
2580	x37	temporal-quantity	:quant	-	-	x2	几
2580	x37	temporal-quantity	:unit	-	-	x3	天
2580	x37	temporal-quantity	:mod	-	-	x1	这

Table 3: Sample of CAMR tuples

Table 3 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 New Blind Test

As the predecessor of CAMRP 2022, the evaluation task this year includes a brand new blind test comprising 2,000 interrogative sentences, namely Test C. Original data was collected and filtered from Zhihu website, and presented with alignment annotations. We expect to exam the parsing potential for interrogative focus in Chinese with the favor of new blind test.

4 Evaluation Design

In spirit of innovation and comparison, there are three evaluation metrics and two modalties include at CAMRP 2023.

4.1 Evaluation Metrics

4.1.1 Smatch

As the most widely-used evaluation metric, Smatch focuses on the overlapping of two AMR graphs (Cai and Knight, 2013). For two AMR graphs to be matched, Smatch first renames the nodes of AMR graphs and transforms each AMR graph into a set of triples. There are three categories of triples as following:

- Node triple:

  $$\texttt{instance}(\texttt{node\_index}, \texttt{concept})$$

  where `instance` represents the concept nodes. `node_index` is the index of nodes in AMR graph and denoted as a_i. Without loss of generality, we have $i \in 0, 1, \ldots, n$. `concept` is abstracted from the word accordingly. As shown in Table 4, for example, the triple instance(a_0, 希望-01) indicates the instantiation of the word “希望” including its index a_0 and the abstracted concept “希望-01”.

- Arc triple:

  $$\texttt{relation}(\texttt{node\_index1}, \texttt{node\_index2})$$

  where `node_index1` and `node_index2` are indexes of two different concept nodes, and their mappings are a_i and a_j, respectively. As always, $j \in 0, 1, \ldots, n$. `relation` is the semantic role between the index a_i and a_j. For example, the arc triple arg1(a_1, a_4) means that the semantic relation between the mapping words of the index a_1 and a_4 is `arg1` (Object).

- Node property triple:

  $$\texttt{property}(\texttt{node\_index}, \texttt{value})$$

  As shown in Table 4, the property triple root(a_0, top) indicates that the property of the index a_0 is `root`, in which `value` equals `top`, implying that it is the root node in the graph.

4.1.2 Main metric: Align-smatch

With two types of information added, including concept alignment and relation alignment, Align-smatch now transforms Chinese AMR graph into tuples (Xiao et al., 2022).

- New triple for **Concept Alignment**:

  $$\texttt{anchor}(\texttt{node\_index}, \texttt{token\_num})$$

  We name it concept alignment triple and add it into the same category with node property triple. `anchor` stands for it node property. `node_index` remains the same as in Smatch. `token_num` means the number of the word in original sentence (as we mentioned earlier). As shown in Table 5, for example, the property triple anchor(a_7, x3) indicates that the mapping concept node “惨痛-01” of the index a_7 is aligned with the mapping word “惨痛” of the token number x3.

Category	Triple
Node	instance(a0, 希望-01)
	instance(a1, 给-01)
	instance(a2, expressive)
	instance(a3, 经历)
	instance(a4, 大家)
	instance(a5, 教训)
	instance(a6, 我)
	instance(a7, 惨痛-01)
	instance(a8, 1)
	instance(a9, 个)
Arc	mode(a0, a2)
	arg1(a0, a1)
	arg0(a1, a3)
	arg2(a1, a4)
	arg1(a1, a5)
	arg0-of(a3, a7)
	poss(a3, a6)
Node Property	root(a0, top)

Table 4: Triple representation in Smatch

- New tuple for **Relation Alignment**:

$$(\texttt{Word\_on\_Arc}, \texttt{token\_num}, \texttt{node\_index1}, \texttt{node\_index2})$$

Likewise, we name it relation alignment tuple and add it into the same category with arc triple (tuple). Word_on_Arc represents the function word on arc for it actually matters a lot and conveys relations between content words in Chinese. As shown in Table 5, the arc tuple "(的, x4, a_3, a_7)" indicates that the function word "的" is on the arc from the index a_3 pointing to a_7, and assigned with the token number x4 for it is the fourth word in the original sentence (after word segmentation).

- New arc triple:

$$\texttt{relation}(\texttt{node\_index1}, \texttt{node\_index2})$$

When processing the word on the root node, we now replace the original property triple with new arc triple. As shown in Table 4, the root node triple in Smatch metric was root(a_0, top), and has been changed into root(a_0,a_0) as we can see in Table 5.

4.1.3 MRP

MRP (Oepen et al., 2020), with its great compatibility, has been used as the only metric in both CoNLL 2019 and CoNLL 2020. And yet when it comes to AMR or CAMR parsing evaluation, MRP normally returns score higher than the other two metrics mentioned above due to its comparatively loose scoring method. For more details, please refer to their Github repository[1].

With concept alignment and relation alignmetn added, Chinese AMR parsing is perfected and completed. Therefore, with full considerations, we take Align-smatch as the main metric at CAMRP 2023. Metrics like MRP and Smatch are for reference only and can mirror if there's any fluctuation or progression in last couple years.

[1] https://github.com/cfmrp/mtool

Category	Tuple

Node	instance(a0, 希望-01)
	instance(a1, 给-01)
	instance(a2, expressive)
	instance(a3, 经历)
	instance(a4, 大家)
	instance(a5, 教训)
	instance(a6, 我)
	instance(a7, 惨痛-01)
	instance(a8, 1)
	instance(a9, 个)

Arc	root(a0, a0)
	mode(a0, a2)
	arg1(a0, a1)
	arg0(a1, a3)
	arg2(a1, a4)
	arg1(a1, a5)
	arg0-of(a3, a7)
	(的, x4, a3, a7)
	poss(a3, a6)

Node Property	anchor(a0, x1)
	anchor(a1, x6)
	anchor(a2, x11)
	anchor(a3, x5)
	anchor(a4, x7)
	anchor(a5, x10)
	anchor(a6, x2)
	anchor(a7, x3)
	anchor(a8, x8)
	anchor(a9, x9)

Table 5: Tuple representation in Align-smatch

4.2 Two Modalities

The evaluation task includes Open Modality and Closed Modality:

- **Closed Modality.** Participants must use the training data, test data and pre-trained model which are all designated in advance. No alternative is allowed. We also offer dependency analysis results of the train set for each team under Closed Modality. HIT_Roberta from Harbin Institue of Technology (Cui et al., 2021) as pre-trained model is highly recommended.

- **Open Modality.** Participants are allowed to use other pre-trained models and external resources such as named entities and dependency analysis results with no limits. Note that all kinds of resources that participants employ should be mentioned and written in detail in the final technical report. Manual correction is forbidden in both modalities. Table 6 shows the requirements of two modalities respectively.

Modalities Resources	***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 6: Requirements of two modalities

5 Evaluation Results

CAMRP 2023 initiates on 1st May, and data set inlucding train set and dev set are authorized and released via LDC. Test sets are provided on 1st June via our GitHub repository[2]. Participants are to submit their technical report by 25th June and Camera-ready by 28th June. The evaluation task will be hosted as part of the 22nd China National Conference on Computational Linguistics (CCL 2023) in Harbin, China.

5.1 Participants

There are 21 teams enrolled and 6 teams stick to the end. 48 results in total are returned as shown in Table 7 along with detailed information. Majority has chosen closed modality and a few has chosen open modality only. Teams like SUDA and WestlakeNLP have overdue submissions which we mark with an asterisk in Table 7. Each team is listed alphabetically here and throughout.

Team	Affiliation	Test A		Test B		Test C	
		closed	*open*	*closed*	*open*	*closed*	*open*
BUPT	Beijing University of Posts and Telecommunications	2	0	2	0	2	0
GDUFE	Guangdong University of Finance and Economics	1	1	1	1	1	1
SJTU	Shanghai Jiao Tong University	0	1	0	1	0	1
SUDA	Soochow University	2+2*	2+2*	2+2*	2+2*	2+2*	2+2*
WHU	Wuhan University	1	0	1	0	1	0
WestlakeNLP	Westlake University	0	1+1*	0	1+1*	0	1+1*
Total	48	8	8	8	8	8	8

Table 7: Participants information overview

5.2 Overall Results

Results from 6 teams encompassing a total of 48 runs exhibit an unexpected level of parsing performance across a broad spectrum. For the sake of better display and clearer comparison, we accordingly drew 6 tables (Table 8-13) to present all results of three test sets, in two modalities 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 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.

The best record is 0.8000 in closed Test A, 0.7264 in closed Test B, and 0.8137 in closed Test C. Open modality, on other hand, axiomatically enable participants to reach their limits even more. The highest score is 0.8130 in open Test A, 0.7471 in open Test B, and 0.8261 in open Test C, which is around two percentage points higher than that of in closed modality respectively. MRP metric, given its relatively not that strict scoring method, yields better results than other two metrics. What is worth mentioning is

[2] https://github. com/GoThereGit/Chinese-AMR

Team	Run	*Align-smatch*			*Smatch*			*MRP*		
		P	R	F_1	P	R	F_1	P	R	F_1
BUPT	1	0.7774	0.7682	0.7728	0.7598	0.7539	0.7569	0.8086	0.8057	0.8071
	2	0.7840	0.7644	0.7741	0.7529	0.7350	0.7438	0.8035	0.7947	0.7991
GDUFE	1	0.8080	0.7287	0.7663	0.7905	0.7121	0.7492	0.8308	0.7631	0.7955
	2	-	-	-	-	-	-	-	-	-
SUDA	1	0.8183	0.7824	**0.8000**	0.8104	0.7696	0.7895	0.8463	0.8142	0.8299
	2	0.8185	0.7654	0.7911	0.7515	0.8104	0.7798	0.8460	0.7963	0.8204
WHU	1	0.7894	0.7490	0.7687	0.7528	0.7326	0.7426	0.8036	0.7941	0.7988
	2	-	-	-	-	-	-	-	-	-

Table 8: Results of Test A in closed modality

Team	Run	*Align-smatch*			*Smatch*			*MRP*		
		P	R	F_1	P	R	F_1	P	R	F_1
BUPT	1	0.6959	0.7103	0.7030	0.6999	0.7045	0.7022	0.7564	0.7527	0.7545
	2	0.7209	0.6968	0.7087	0.7041	0.6872	0.6956	0.7484	0.7509	0.7497
GDUFE	1	0.7575	0.6118	0.6769	0.7515	0.6111	0.6741	0.7921	0.6617	0.7210
	2	-	-	-	-	-	-	-	-	-
SUDA	1	0.7516	0.7028	**0.7264**	0.7569	0.7119	0.7337	0.7964	0.7529	0.7740
	2	0.7535	0.6968	0.7240	0.7622	0.7058	0.7329	0.8008	0.7452	0.7720
WHU	1	0.7241	0.6783	0.7004	0.7028	0.6823	0.6924	0.7489	0.7488	0.7488
	2	-	-	-	-	-	-	-	-	-

Table 9: Results of Test B in closed modality

that team SUDA has scored a 0.8416 in MRP, which literally outperforms the SOTA at CoNLL 2020 by 3.64 percentage points[3].

In nutshell, results vary according to different modalities, metrics and test sets. Parsing performance on open modality inevitably exceeds that of on closed modality:

$$F_1{}^{open} \gg F_1{}^{closed}$$

And three test sets, with distinct language flavor and characteristics, are too revealing a degree of complexity. Test C comprising of all short simple sentences is the easiest, without a shadow of doubt:

$$F_1{}^{testC} > F_1{}^{testB} > F_1{}^{testA}$$

Lastly, the variability in scores arises when there is a change in the chosen metrics. Counter-intuitive as it may appear, Align-smatch is not the metric with lowest scores:

$$F_1{}^{mrp} > F_1{}^{align-smatch} \geq F_1{}^{smatch}$$

We are to further discuss more technical details in the subsections below.

5.3 Models and Analysis

Given the significant advancements in natural language processing and the increased recognition of the potential of large language models (LLMs), participants at CAMRP 2023 have been influenced by the success of models such as ChatGPT (Ouyang et al., 2022). These models have demonstrated their effectiveness in various tasks, showcasing their ability to generate human-like responses and comprehend

[3] The test set used at CoNLL 2020 is exactly the same with the Test A at CAMRP 2023.

Team	Run	*Align-smatch*			*Smatch*			*MRP*		
		P	R	F_1	P	R	F_1	P	R	F_1
BUPT	1	0.8096	0.7862	0.7977	0.7925	0.7980	0.7952	0.8354	0.8386	0.8370
	2	0.8060	0.7777	0.7916	0.7780	0.7562	0.7669	0.8300	0.8072	0.8185
GDUFE	1	0.8238	0.7308	0.7745	0.8048	0.7161	0.7578	0.8489	0.7653	0.8049
	2	-	-	-	-	-	-	-	-	-
SUDA	1	0.8331	0.7951	**0.8137**	0.8265	0.7870	0.8063	0.8652	0.8282	0.8463
	2	0.8111	0.8050	0.8081	0.8126	0.8102	0.8114	0.8563	0.8445	0.8504
WHU	1	0.8098	0.7635	0.7859	0.7798	0.7548	0.7671	0.8313	0.8069	0.8189
	2	-	-	-	-	-	-	-	-	-

Table 10: Results of Test C in closed modality

Team	Run	*Align-smatch*			*Smatch*			*MRP*		
		P	R	F_1	P	R	F_1	P	R	F_1
GDUFE	1	0.7553	0.7560	0.7556	0.7333	0.7403	0.7368	0.7832	0.7944	0.7887
	2	-	-	-	-	-	-	-	-	-
SJTU	1	0.4741	0.4645	0.4692	0.6173	0.6094	0.6133	0.5131	0.5022	0.5076
	2	-	-	-	-	-	-	-	-	-
SUDA	1	0.8081	0.8174	0.8128	0.7960	0.8060	0.8010	0.8375	0.8456	0.8415
	2	0.8082	0.8179	**0.8130**	0.7955	0.8054	0.8004	0.8376	0.8457	0.8416
Westlake-NLP	1	0.7440	0.7024	0.7226	0.7300	0.6936	0.7114	0.7816	0.7322	0.7561
	2	-	-	-	-	-	-	-	-	-

Table 11: Results of Test A in open modality

complex language patterns. Some choose to utilize the great power of LLMs, while some refer to prior parsing systems which have proven to come in handy still. Five out of six teams have completed their technical report, and we are to analyse their pros and cons.

BUPT and GDUFE, following the same path, both have reproduced the SOTA system of SUDA-HUAWEI[4] in last year's CAMRP 2022, which uses RoBERTa-BiLSTM as encoder and a Biaffine classifier as decoder. Both results and performance have been promising, achieving decent scores of 0.7728 and 0.7663 in closed Test A, respectively. Similiarly, WHU reproduced the CAMR parsing model of PKU (Chen et al., 2022), which has won the second prize last year at CAMRP 2022. And yet for some reasons, WHU failed to implement relation alignment while parsing, leading to the decrease of their final score.

SJTU and WestlakeNLP choose to explore novel approaches with LLMs. SJTU follows two primary ideas including (1) Predict and infer with ChatGPT in zero-shot and few-shot, (2) Fine-tune ChatGLM-6B (Du et al., 2021) with LoRA (Hu et al., 2021). In the stage of zero-shot and few-shot modeling, they tend to some certain prompt engineering after pre-processing so as to convert Chinese AMR parsing into Seq2Seq task. The outcome, however, was not promising. It appears that when faced with complex prediction tasks such as Chinese AMR parsing, the performance of ChatGPT in zero-shot and few-shot scenarios did not meet the expectations. So is fine-tuning ChatGLM-6B, even though SJTU has tries different strategies, the best record is 0.6052 in open Test C.

WestlakeNLP shared the same inspiration with SJTU, fine-tuning LLMs. Instead of relying on ChatGPT, they choose to utilize baichuan-7B[5] for the complex task of Chinese AMR parsing. This model is renowned for its large size and impressive performance compared to alternative models. WestlakeNLP follows the step of pre-propcessing and prompt engineering as well. They also add post-propocessing

[4] https://github.com/zsLin177/camr

[5] https://github.com/baichuan-inc/baichuan-7B

Team	Run	*Align-smatch*			*Smatch*			*MRP*		
		P	R	F_1	P	R	F_1	P	R	F_1
GDUFE	1	0.6971	0.6733	0.6850	0.6882	0.6778	0.6830	0.7394	0.7306	0.7350
	2	-	-	-	-	-	-	-	-	-
SJTU	1	0.4644	0.4568	0.4606	0.6037	0.6001	0.6019	0.5137	0.5099	0.5118
	2	-	-	-	-	-	-	-	-	-
SUDA	1	0.7433	0.7485	0.7459	0.7505	0.7635	0.7570	0.7899	0.7963	0.7931
	2	0.7439	0.7503	**0.7471**	0.7521	0.7669	0.7595	0.7916	0.7975	0.7945
Westlake-NLP	1	0.7042	0.6863	0.6952	0.7021	0.6930	0.6975	0.7501	0.7301	0.7400
	2	-	-	-	-	-	-	-	-	-

Table 12: Results of Test B in open modality

Team	Run	*Align-smatch*			*Smatch*			*MRP*		
		P	R	F_1	P	R	F_1	P	R	F_1
GDUFE	1	0.773	0.7814	0.7772	0.7521	0.7691	0.7605	0.8020	0.8181	0.8100
	2	-	-	-	-	-	-	-	-	-
SJTU	1	0.6282	0.5839	0.6052	0.7262	0.6840	0.7045	0.6697	0.6228	0.6454
	2	-	-	-	-	-	-	-	-	-
SUDA	1	0.8206	0.8212	0.8209	0.8164	0.8195	0.8179	0.8575	0.8566	0.8571
	2	0.8211	0.8213	0.8212	0.8163	0.8186	0.8175	0.8576	0.8563	0.8569
Westlake-NLP	1	0.8273	0.8249	**0.8261**	0.8143	0.8118	0.8130	0.8561	0.8549	0.8555
	2	-	-	-	-	-	-	-	-	-

Table 13: Results of Test C in open modality

in order to better complement any missing information like parenthesis and nodes. The highest score of WestlakeNLP is 0.8261 in open Test C.

SUDA[6] has taken the unique features in Chinese AMR parsing, information of alignment and co-reference, for example, into consideration, therefore they use multiple auto-regressive and non auot-regressive models and fuses their outputs based on graph ensemble method. In open modality, their whole parsing system is on the base of BART model (Lewis et al., 2019), and fuse dependency results and POS (Part-of-Speech) information layered with a BiLSTM. RoBERTa is the only pre-trained model allowed in closed modality, so they inherit their prior work, finally reaching a 0.8000 in closed Test A.

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., 2016), CAMRP 2023 proposes 8 fine-grained metrics for Chinese AMR parsing, including **CA** (Concept Alignment) and **RA** (Relation Alignment), and **Interr.** (Interrogation) especially for Test C this year.

Table 14 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. **Interr.** is proposed this year at CAMRP 2023, mainly computing on the *amr-unknown* concepts so as to further explore the parsing systems' ability and potential of finding interrogative focus and multiple

[6] https://github.com/EganGu/camr-seq2seq

interrogations in one single sentence.

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
Interr.	Interrogations	*amr-unknown* concepts

Table 14: Eight fine-grained metrics

Metric Team	**Neg.**	**Con.**	**NSF**	**Imp.**	**Reent.**	**CA**	**RA**
closed							
BUPT	0.7219	0.8507	0.8671	0.8264	0.5060	0.9036	0.4669
GDUFE	0.7187	0.8425	0.8602	0.8196	0.4994	0.8738	0.4910
SUDA	**0.7640**	0.8627	**0.8800**	0.8347	0.5865	0.8957	**0.5651**
WHU	0.6416	0.8397	0.8608	0.8041	0.5063	0.9035	-
open							
GDUFE	0.6825	0.8431	0.8638	0.8052	0.4695	0.8786	0.4736
SJTU	0.5719	0.7615	0.7892	0.7142	0.4165	0.3000	0.3265
SUDA	0.7537	**0.8695**	0.8759	**0.8381**	**0.6404**	**0.9079**	0.5515
WestlakeNLP	0.6800	0.8149	0.8168	0.7852	0.5029	0.8348	0.4678

Table 15: Fine-grained metrics and subscores in Test A

Table 15-17 shows participants' performance in each track, including two modalities and three test sets. Fine-grained metric **Interr.** is only set active when scoring in Test C (for interrogative sentences only).

Generally, subscores in metrics like **NSF** or **Con.** are apparently higher than the rest. **Neg.** shifts its difficulty according to different test set. And nearly all subscores in **Reent.** failed to reach 0.6, indicating that the complexity of AMR or CAMR topology structure and the exceptionally challenging nature of the parsing task. It is evident that the utilization of concept alignment annotation in Chinese AMR has had a noticeable impact, leading to higher subscores in metrics related to concepts, **CA**, for example, are to break 0.9 almost. **RA**, however, still remains the lowest results among all (same at CAMRP 2022).

Noticeably, SUDA has achieved the highest subscore in the **Reent.**, thanks to their special pre/post-process of co-reference in Chinese AMR. Their unique treatment of co-reference resolution has allowed for more accurate identification and representation of reentrancies within the AMR graphs. SJTU with modeling via ChatGPT and ChatGLM-6B, yet ends up with around 0.3 in both fine-grained metrics **CA** and **RA**. It is reasonable to argue that when it comes to the task of structural prediction and inference, relying solely on LLMs may not be sufficient.

6 Conclusion and Future Work

This paper introduced the overview of the Third Chinese Abstract Meaning Representation Parsing Evaluation in CCL 2023. CAMRP 2023 uses the novel metric Align-smatch to better evaluate the parsing performance of each participating parsing system. There have been six teams in total submitted their

Metric Team	Neg.	Con.	NSF	Imp.	Reent.	CA	RA
closed							
BUPT	0.5784	0.7880	0.8003	0.7098	0.5366	**0.8460**	0.4125
GDUFE	0.5562	0.7622	0.7699	0.6521	0.4609	0.7894	0.4138
SUDA	0.6002	0.7999	**0.8161**	0.7154	0.5734	0.8351	**0.5031**
WHU	0.4863	0.7752	0.7781	0.7018	0.5288	0.8455	-
open							
GDUFE	0.5309	0.7799	0.7892	0.6508	0.4728	0.8164	0.4120
SJTU	0.4771	0.7261	0.7408	0.5703	0.4531	0.3131	0.3174
SUDA	**0.6285**	**0.8116**	0.8071	**0.7393**	**0.6334**	0.8447	0.5025
WestlakeNLP	0.5538	0.7831	0.7775	0.6696	0.5612	0.7967	0.4516

Table 16: Fine-grained metrics and subscores in Test B

Metric Team	Neg.	Con.	NSF	Imp.	Reent.	CA	RA	Interr.
closed								
BUPT	0.6364	0.8414	0.8284	0.7004	0.4719	0.8551	0.4904	0.9242
GDUFE	0.6116	0.8173	0.8039	0.6517	0.4019	0.8098	0.4564	0.8839
SUDA	**0.6621**	0.8479	0.8361	0.6959	0.5165	0.8391	**0.5023**	0.9379
WHU	0.6230	0.8181	0.8153	0.7054	0.4604	**0.8556**	-	0.9127
open								
GDUFE	0.6054	0.8352	0.8183	0.6651	0.3616	0.8428	0.4578	0.8775
SJTU	0.5156	0.7873	0.8159	0.5922	0.4054	0.4620	0.3609	0.5833
SUDA	0.6481	0.8493	0.8256	**0.7347**	**0.5637**	0.8321	0.4614	**0.9562**
WestlakeNLP	0.6402	**0.8589**	**0.8399**	0.7265	0.5588	0.8405	0.4680	0.9527

Table 17: Fine-grained metrics and subscores in Test C

results, which are inspiring and motivating. Some has advanced prior works and found creative orientation. Some has probed into LLMs thoroughly. In MRP metric, SUDA has scored a 0.8416, surpassing the best record at CoNLL 2020 by 3.64 percentage points. Semantic parsing for interrogative focus in Chinese seems fairly promising. Significant achievements and continuous progress have been made in Chinese AMR parsing, accompanied by notable advancements and innovative approaches. However, it is important to acknowledge that relation prediction and its alignment continue to pose challenges, acting as bottlenecks in the development of Chinese AMR parsing. Despite the remarkable breakthroughs in various aspects of Chinese AMR parsing, accurately predicting and aligning relations remains a critical area that requires further improvement. The complex nature of relation identification and alignment within AMR structures demands focused attention and innovative techniques.

In our future endeavors, we are committed to dedicating extensive efforts to advance Chinese AMR parsing. This includes hosting evaluation tasks to facilitate the assessment and benchmarking of parsing models. Additionally, we aim to construct and refine parsing models that are specifically tailored to the intricacies of Chinese AMR, ultimately driving forward the field of semantic analysis. By focusing on relation prediction and alignment, we aim to overcome the current challenges and enhance the overall performance and understanding of Chinese AMR parsing. Through continuous research, collaboration, and innovation, we aspire to contribute to the development of 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, Liming Xiao, Jin Chen, Jingya Lu. And particularly Yuan Wen, Yihuan Liu and Peiyi Yan, we thank them for annotating the corpus of Chinese AMR. Lastly, we extend our appreciation to all the anonymous reviewers who provided thoughtful comments and feedback, helping to refine and strengthen this paper. This work is the staged achievement of the projects supported by National Social Science Foundation of China (18BYY127) and National Natural Science Foundation of China (61772278).

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. 2016. An incremental parser for abstract meaning representation. *arXiv preprint arXiv:1608.06111*.

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.

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. Ufal at mrp 2020: Permutation-invariant semantic parsing in perin. *arXiv preprint arXiv:2011.00758*.

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.