

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, Yuelel 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 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

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

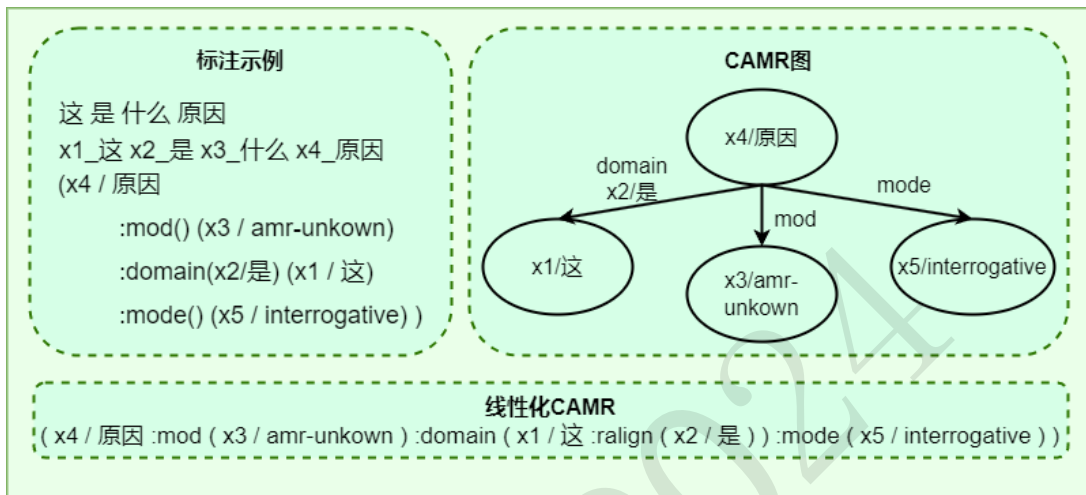


Figure 1: 句子: “这是什么原因?” 的 CAMR 示例

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-HUAWEI¹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

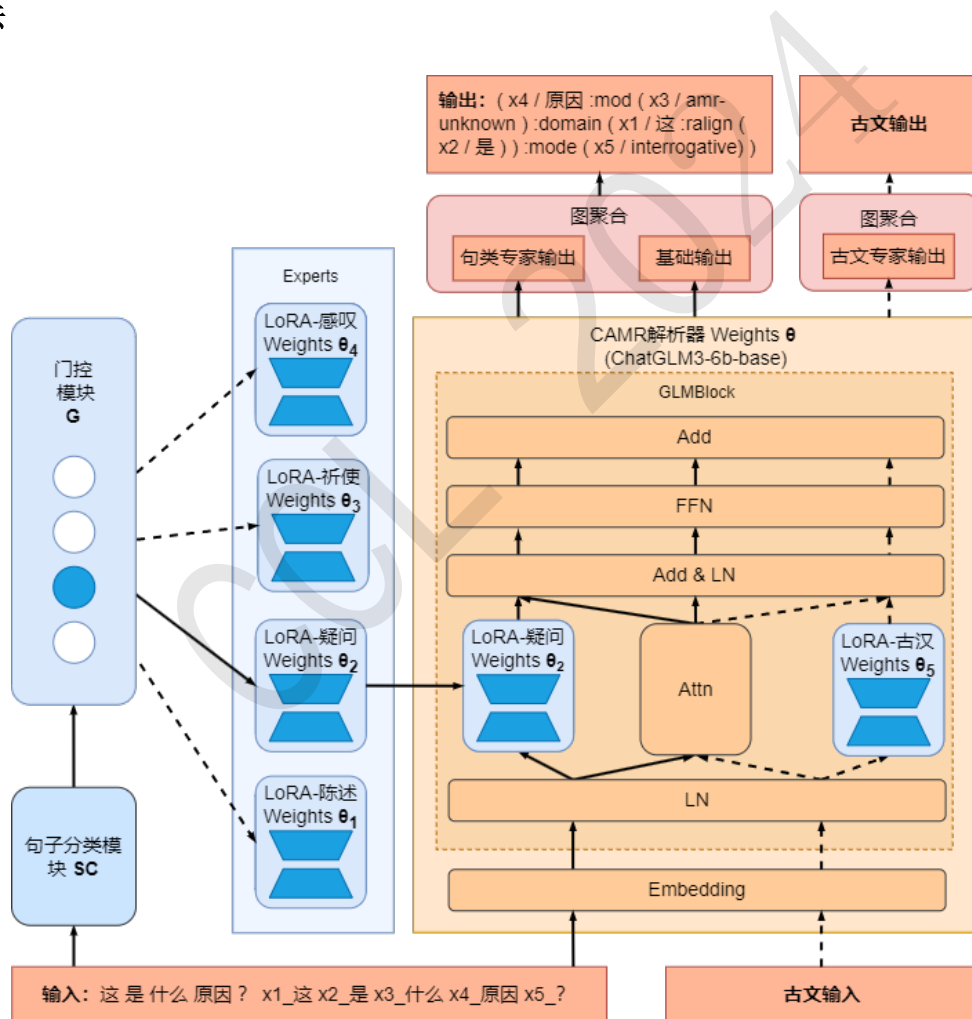


Figure 2: 框架结构示意图

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 $S = \{s_1, s_2, \dots, s_n\}$ 及其带有概念节点标签的句子 $W = \{w_1, w_2, \dots, w_n\}$ ，让模型输出 CAMR 图的 DFS 线性化序列 $Y = \{y_1, y_2, \dots, y_m\}$ ，如图 1 所示。线性化的步骤借鉴了 Gu et al. (2023) 的处

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。ChatGLM3-6B³ 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：

$$X_{\theta} = X + f^{Attn}(LN(X); \theta) \quad (1)$$

$$Basic_{\theta}(X) = X_{\theta} + f^{FFN}(LN(X_{\theta}); \theta) \quad (2)$$

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/>

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:

$$X_{\theta_i} = X + f^{Attn}(LN(X); \theta_i) \quad (3)$$

$$LE_{\theta_i}(X) = X_{\theta_i} + f^{FFN}(LN(X_{\theta_i}); \theta) \quad (4)$$

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 $S = \{s_1, s_2, \dots, s_n\}$, 模型输出为

$$SC_i(S) = OneHot(type_i) \quad (5)$$

其中 $i \in \{\text{陈述句, 感叹句, 祈使句, 疑问句}\}$, $OneHot(\cdot)$ 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:

$$Gate_i(SC_i(S)) = \begin{cases} 1, & \text{if } i = \text{argmax}(SC_i(S)) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

其中 $\text{argmax}(\cdot)$ 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

行结合，以期能得到更好的结果。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:

$$Accuracy(SC) = \frac{1}{n} \sum_{k=1}^n \delta(SC_i(S_k) \in A_k) \quad (8)$$

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

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 | | | |

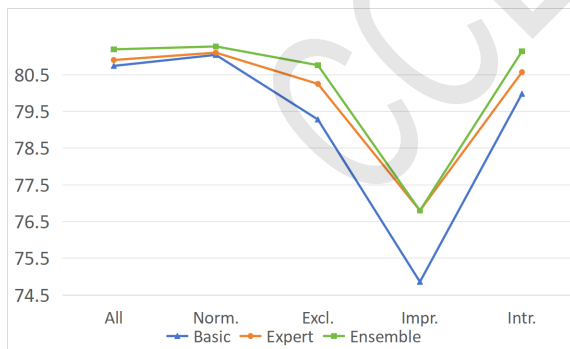


Figure 5: TestA 各得分的折线图

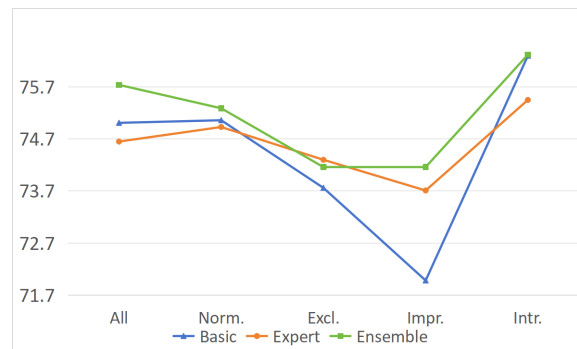


Figure 6: TestB 各得分的折线图

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

- semlanking. 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: 基于图融合的非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.

- 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. 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*, pages 5938–5945, 2022.

- 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, Xinchun 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.