CCL24-Eval任务4系统报告：面向中文抽象语义表示解析的大模型评估与增强

陈荣波，裴振武，白雪峰*，陈科海，张民
哈尔滨工业大学（深圳），计算机科学与技术学院
baixuefeng@hit.edu.cn

摘要

本文介绍了我们在第二十三届中文计算语言学大会中文抽象语义表示解析评测任务中提交的参赛系统。中文抽象语义表示（Chinese Abstract Meaning Representation，CAMR）以一个单根可遍历的有向无环图表示中文句子的语义。本系统选择大语言模型作为解决方案。我们首先系统地评估了当下中文大语言模型在AMR解析任务上的性能，在此基础上基于图融合算法整合性能较高的大模型预测结果，最终得到预测的CAMR图。实验结果表明，1）现有大模型已经具备一定的少样本中文AMR解析能力；2）基于微调中文大模型的AMR解析系统能够取得相较以往最优系统更强的性能；3）图融合算法能够进一步增强基于大模型的CAMR解析系统的性能。

关键词： 中文抽象语义表示；大语言模型；图融合

System Report for CCL24-Eval Task 4: Benchmarking and Improving LLMs on Chinese AMR Parsing

Rongbo Chen, Zhenwu Pei, Xuefeng Bai, Kehai Chen, Min Zhang
School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
baixuefeng@hit.edu.cn

Abstract

This paper introduces our submission system for the Chinese Abstract Meaning Representation (CAMR) parsing evaluation task at the 23rd China National Conference on Computational Linguistics. CAMR represents the semantics of Chinese sentences using a single-rooted, traversable, directed acyclic graph. We choose to build CAMR parsing system based on LLMs. We first systematically benchmark the performance of current Chinese large models on the CAMR parsing task. Based on this, we integrate the prediction results of high-performing large models using a graph ensemble algorithm to obtain the final predicted CAMR graph. The experimental results show that: 1) current large models already possess a certain capability in few-shot CAMR parsing; 2) an AMR parsing system based on fine-tuned Chinese large models can achieve superior performance compared to previous systems; 3) the graph ensemble algorithm can further enhance the performance of a large model-based CAMR parsing system.

Keywords: CAMR , Large language model , Graph ensemble

* 通讯作者

1 引言

中文抽象语义表示（Chinese Abstract Meaning Representation，CAMR）将词抽象为概念节点，句子中词与词之间的语义关系抽象为有向弧，从而将整个中文句子的语义结构描述为一个单根有向无环图 (Banarescu et al., 2013)。不同于标准的AMR表示范式，CAMR在保留AMR的语义表示能力的同时，还增加了概念对齐以及关系对齐等信息，可以更好地表达中文句子语义。AMR与CAMR解析旨在预测输入文本中对应的语义图 (Bai et al., 2022a; Bai et al., 2022b; Bevilacqua et al., 2021; Cai and Lam, 2020; Flanigan et al., 2014; Konstas et al., 2017; Lyu and Titov, 2018)。受益于自动解析技术的发展，AMR与CAMR已经被广泛应用于机器翻译 (Nguyen et al., 2021)、对话系统 (Bai et al., 2021)、文本摘要 (Liao et al., 2018) 等自然语言处理下游任务领域中。

现有AMR/CAMR解析的方法可以分为以下三类：基于转移的方法（Transition-based Parsing）、基于图的方法（Graph-based Parsing）以及基于序列到序列的方法（Seq2Seq-based Parsing）。其中，基于转移的方法首先将CAMR图转换成一系列“转移动作”，然后通过预测该转移动作序列来一步步地构造AMR图 (Wang et al., 2018)。基于图的方法一般通过“节点预测-关系生成”两阶段的方式进行AMR图预测，包括自回归方式以及非自回归方式 (Zhou et al., 2022; Chen et al., 2022)。基于序列到序列的方法的核心思想是将CAMR图进行序列化，从而以序列生成的形式进行自回归生成 (Huang et al., 2021)。以上三种方法中，基于序列到序列的方法无需人工设计复杂的中间特征，极大降低了任务的复杂度。

近年来，大规模语言模型（Large Language Models，LLMs）在诸多自然语言处理任务上取得了巨大的成功 (Touvron et al., 2023; Zhao et al., 2023; Wang et al., 2023; Ren et al., 2023; Zeng et al., 2022)。得益于巨大的参数规模和训练数据规模，大模型产生了在小型模型中不存在的涌现能力，例如上下文学习、指令遵循和逐步推理等能力 (Zhao et al., 2023)。目前，将大模型应用于下游任务的通用范式是将任务输入和输出拼接并转换为一个文本序列，从而使得大模型能够以与预训练阶段相同（即序列生成）的形式处理下游任务。对于CAMR解析任务，Gao et al. (2023) 首个将中文大语言模型用于CAMR解析，初步揭示了大语言模型在CAMR解析任务上的潜力。尽管如此，Gao et al. (2023) 的性能相对受限，并且目前尚缺乏大模型在CAMR解析任务上的系统研究。

为了充分挖掘中文大模型在CAMR解析任务上的潜力，本文首先在CAMR解析任务上的两种设置下对现有多个中文大模型进行了系统评估，实验设置包括少样本学习和监督学习，评估对象包括ChatGPT、GPT4两个商用模型以及Baichuan-2、LLaMA-3、LLaMA-3-Chinese三个开源中文大模型。在此基础上，本文选取最优的几个解析系统并使用图融合算法 (Hoang et al., 2021) 整合多个输出得到最终结果。实验结果表明，1）现有的商用模型已经具有一定的少样本CAMR解析能力；2）微调后的开源中文大模型能够取得与以往CAMR解析系统可比的性能，同时具有更好的泛化性；3）图融合算法能够进一步提升系统的性能。最终，本文所提出的系统在三个测试集上分别取得了80.96、74.85和66.91的分数。

2 方法

本文使用基于序列到序列的方法进行CAMR解析。首先对CAMR图进行序列化预处理，以训练集中分词句子和序列化CAMR图作为组合，对中文大模型进行训练，在此基础上实现将文本转换成CAMR序列，然后将从多个模型得到的CAMR序列进行后处理，还原成形式合法的CAMR图，最后将多个模型生成的CAMR图进行融合和二次后处理后，得到目标CAMR图。

2.1 CAMR线性化

本文参照CAMRP2023[1]中SUDA[2]的策略，对虚词关系对齐和概念同指进行简化特殊处理。如图 1 所示，我们将虚词关系对齐的表示“:arg0(x8/的) (x7 / 政党)”处理成“:arg0 (x7 / 政党 :ralign (x8 / 的))”，将概念同指的表示“:quant() (x5 / x1)”处理成“:quant (x5 / 一 :coref x1)”。另外，我们还对CAMR图中不具有重要意义的括号、空格和换行等符号进行去除，从而

[1]https://github.com/GoThereGit/Chinese-AMR/tree/main/CAMRP%202023
[2]https://github.com/EganGu/camr-seq2seq

简化所得到的CAMR图的线性序列。通过以上方法，我们由原始CAMR图得到可用于大模型处理的线性化CAMR序列。

分词文本：这/种/理念/是/一/个/政党/的/执政/基础/。
分词下标文本：x1_这 x2_种 x3_理念 x4_是 x5_一 x6_个 x7_政党 x8_的 x9_执政 x10_基础 x11_。

CAMR图：(x10 / 基础
:mod() (x9 / 执政-01
:arg0(x8/的) (x7 / 政党
:quant() (x5 / x1)
:cunit() (x6 / 个)))
:domain(x4/是) (x3 / 理念
:mod() (x1 / 这)
:unit() (x2 / 种)))

序列化结果：(x10 / 基础 :mod (x9 / 执政-01 :arg0 (x7 / 政党 :ralign (x8 / 的)
:quant (x5 / 一 :coref x1) :cunit (x6 / 个))) :domain (x3 / 理念
:ralign (x4 / 是) :mod (x1 / 这) :unit (x2 / 种)))

Figure 1: CAMR序列化示例（序列化结果为单行）

2.2 基于中文大模型的CAMR解析

为了系统地激发与评估现有中文大模型在CAMR解析的能力，本文选取了现有5种具有代表性的大模型进行实验，并探索了少样本学习和有监督微调两种设置：对于商用闭源模型，本文使用少样本提示学习来激发模型的CAMR解析能力[3]；对于开源模型，本文使用有监督微调来充分发挥模型的性能。

2.3 基于少样本提示学习的CAMR解析

本文选取ChatGPT和GPT-4两种商用模型进行基于少样本提示学习的CAMR解析。首先构造包含任务指令、示例样本、用户输入的提示，随后将提示作为历史上下文信息输入中文大模型中，最终将大模型的输出作为解析后的CAMR图。本文使用的提示模板如下：

任务指令：你是一个中文抽象语义表示解析系统，根据接下来的几个示例进行学习，随后将用户输入的文本转换为中文抽象语义表示图。
示例样本：示例$_1$；示例$_2$；示例$_3$；示例$_4$；示例$_5$。
用户输入：x1_这 x2_种 x3_理念 x4_是 x5_一 x6_个 x7_政党 x8_的 x9_执政 x10_基础 x11_。

Figure 2: 用于CAMR解析的少样本提示学习模板示例

2.4 基于有监督微调的CAMR解析

为了充分激发开源中文大模型的CAMR解析性能，本文采用监督学习的形式对开源大模型进行全量微调，从而获得用于CAMR解析的专用模型。在大模型的微调过程中，本文将任务指令、分词后的源端文本以及线性化的CAMR序列进行拼接，随后以自回归的方式进行预测。以图 1 中“这种理念是一个政党的执政基础。”为例，构造输入数据：

给定如下分词后的中文文本，输出其所对应的中文抽象语义表示图：“x1_这 x2_种 x3_理念 x4_是 x5_一 x6_个 x7_政党 x8_的 x9_执政 x10_基础 x11_。”

为了引导模型专注于生成CAMR结构，本文取消了对于任务指令、原句内容的损失计算。形式化而言，假设以$\mathcal{D}$代表整个训练数据集，I表示任务指令，x表示原句，y表示线性化的CAMR序列，本方法通过优化以下损失函数进行训练：

[3]由于在预实验中观测到零样本性能较差，本文只对少样本学习结果进行评估。

$$\mathcal{L} = -\sum_{\{I,x,y\}\in\mathcal{D}} log\, P(\ y \mid I \oplus x), \tag{1}$$

其中 $\mathcal{L}$ 代表损失函数，$\oplus$ 代表序列拼接操作。本文使用基于梯度下降的方法来进行参数更新。

2.5 后处理

由于模型生成的线性序列并不总是符合CAMR的规范，因此需要对其进行后处理，主要有三个方面，括号补全，节点修正以及特殊关系处理。

- 括号补全：由于模型偶尔会产生括号错误匹配的问题，导致CAMR图中节点和边的关系无法正确表示。如果右括号提前与左括号闭合，则将此右括号置于序列末尾；如果右括号数量不足以匹配左括号，则在序列末尾添加一定数量的右括号使其匹配；如果节点缺少括号，则在最小范围内添加一对括号，使其格式满足规范，例如":op1 x12 / 唱罢"，补全括号后修改为":op1 (x12 / 唱罢)"。
- 节点修正：模型解析预测过程中，可能会产生节点信息错误或者缺失的问题。例如节点的编号与节点的内容不匹配或者缺失，或者节点内容不连续，例如对于连续编号"x1_x2_x3"对应的节点内容中词之间出现多余空格。我们通过规则匹配，补全或重新匹配不正确的节点，使其满足规范。
- 特殊关系处理：包含虚词关系对齐以及概念同指，即":coref"与":ralign"。按照预处理中的对应规则，逆向进行复原。对于虚词关系对齐，搜索到它匹配的父亲节点进行复原；对于概念同指，以同指节点中的一个为核心节点，将标签为编号的节点的标签替换为对应核心节点的标签。

2.6 图融合

得到多个模型的后处理CAMR图后，本文采用图融合算法 (Hoang et al., 2021) 来融合多个预测结果，从而得到更加精确、全面的CAMR图。考虑到普通的图和AMR图的不同点在于AMR的结构中存在节点和边的标签， Hoang et al. (2021) 的工作提出一种有效的启发式算法来计算图融合问题的近似最优解，将图融合的预测问题转化成了图挖掘问题，即在图与图之间寻找最大公共子图。具体而言，图融合算法输入m个AMR图，依次选取每个AMR图作为支点图，对于每个支点图来说，以剩下m-1个图中所包含的节点和边进行投票，对m个支点图选出m个聚合图，最终从这m个聚合图中选出一个最佳的聚合图作为输出。

3 实验

3.1 实验设置

数据：本次评测分为封闭测试和开放测试两个赛道，我们选择参加了开放赛道。本次测评训练集包含16576句数据，开发集A、B分别包含1789和500条数据，测试集A、B、C分别包含1713、1999和2000条数据。其中测试集C旨在考察解析系统在古汉语上的自动解析能力。

基线系统：对于闭源商用模型，本文选取了ChatGPT和GPT-4两个模型进行评估。根据预实验的测试结果，本文选用5-ICL作为少样本测试设置，即示例样本由训练集中随机抽选的5个样本组合而来[4]。对于开源中文大模型，基于其在中文评测数据集（C-EVAL）上的性能，本文选取了Baichuan-2[5]，LLaMA-3[6]，LLaMA-3-Chinese[7]三个开源模型进行有监督微调（S.F.T）。此外，为了进行实验对比，本文选取了基于中文BART模型的CAMR解析系统（BARTseq2seq）作为基线。

参数设置：本文所使用的模型在A800 GPU上进行训练，对于本文所提到的开源模型，训练总批次大小为128，训练轮次为5，最大序列长度为1024，使用AdamW (Loshchilov and Hutter, 2019) 优化器进行优化，学习率搜索空间为{1e-5, 3e-5, 5e-5, 8e-5}。

[4]受资源限制，本文只采用了一组示例样本，没有进行多次随机并评估。
[5]https://github.com/baichuan-inc/Baichuan2
[6]https://github.com/meta-llama/llama3
[7]https://github.com/CrazyBoyM/llama3-Chinese-chat

评估指标：本文采用Align_Smatch作为评测指标。曾经作为评测指标的Smatch将每个AMR图转化为三元组的集合，每个集合包含三种数据类型的三元组：表示节点（Instance）的三元组、表示有向弧（Relation）的三元组和表示节点属性（Attribute）的三元组。Align-smatch在Smatch的基础上增添了两种汉语特有的新的数据，概念对齐信息和关系对齐信息 (Xiao et al., 2022)，在评测CAMR图时可以更加客观准确。

3.2 实验结果

不同大模型的CAMR解析性能：表 1 对比了不同模型在测试集TestA上的表现。在少样本测试场景中，ChatGPT模型取得了43.07的F_1分数。而相比之下，GPT-4取得了更高的性能，获得了51.95的F_1分数。以上结果表明现有的商用大模型已经具备一定的CAMR解析能力，为接下来的研究提供了新的思路。此外，通过细粒度分析，我们发现现有的商用大模型在CAMR关系预测上准确率较低（分别是8.62和19.55），原因可能是CAMR的关系标签在大模型的预训练语料中几乎不存在，因此难以预测。在全量微调设置下，现有最强的系统为基于编解码器架构的BARTseq2seq模型。与之相比，基于中文大模型的模型取得了可比或更强的性能。特别地，LLaMA-3-Chinese取得了79.69的F_1分数，这表明基于解码器的大模型仍然具有一定的潜力。另外，通过对比三个基于大模型的系统，可以发现LLaMA-3模型已经具备较强的中文处理能力，继续在中文数据上进行微调未能在CAMR解析任务上带来较大提升。

Setting	Model	Total			Instance	Attribute	Relation
		P	R	F_1	F_1	F_1	F_1
5-ICL	ChatGPT	48.03	39.04	43.07	52.06	77.57	8.62
	GPT-4	55.26	49.01	51.95	61.32	84.41	19.55
S.F.T.	BARTseq2seq	78.63	79.46	79.04	84.94	93.86	64.69
	Baichuan-2	79.64	78.30	78.97	84.77	93.33	64.28
	LLaMA-3	79.79	79.43	79.61	**85.36**	**94.20**	64.72
	LLaMA-3-Chinese	79.69	79.68	**79.69**	**85.36**	94.19	**64.95**

Table 1: 不同模型在测试集TestA上的Align-Smatch得分

图融合性能：我们选取了评分较高的几个模型进行下一步的图融合，以确保最终图融合后的结果效果最好。具体而言，我们选取了表 1 中所有微调后的系统作为图融合候选，并利用不同的学习率构造出另外两种候选模型，总计6个候选模型。表 2 列出了在测试集TestA上我们以模型的不同组合进行图融合后的结果表现。总的来说，在进行图融合过程中，采用的模型数量越多，最终由这些模型预测的结果融合得到的CAMR图的分数也就越高，但存在边际效益递减的情况，随着模型数量增多，平均每个模型带来的提升逐渐降低。另外我们发现，在通过图融合得到输出CAMR图后，进行二次后处理可以使效果小幅度提升。

Setting	#Models	P	R	F_1
Ensemble_1	4	80.43	80.69	80.56
Ensemble_2	5	80.75	80.94	80.84
Ensemble_3	6	80.80	81.11	**80.96**

Table 2: 测试集TestA不同组合图融合后的CAMR图的Align-Smatch得分

最终系统性能：表 3 列出了我们方法的得分以及与往年系统的对比结果。我们的系统在三个测试集上分别获得了80.96、74.85以及66.92的分数。其中，我们的方法在TestA和TestB两组测试集上，达到了和去年测评任务中SUDA的方法（该方法在开放赛道的TestA 和TestB上分别获得了81.30和74.71的分数）相当的水平，表明大语言模型在复杂句子语义图的解析上也可达到良好效果，但依然存在进步空间。另外，相较于去年评测任务中同样采用微调大模型的WestlakeNLP，我们的方法在测试集TestA和TestB上表现较优。此外，表 4 列出了三个测试集的细粒度评测分数，我们注意到TestC的结果相较于测试集TestB和TestC分数偏低，原因可能是TestC由古汉语句子组成，古汉语相较于现代汉语可能在语法结构和语言形式上存在一定的差异，使得模型从现代汉语到古汉语的迁移学习存在较大误差。

Model	TestA			TestB			TestC		
	P	R	F_1	P	R	F_1	P	R	F_1
SUDA-HUAWEI$_{22}$	82.16	78.20	80.13	75.52	71.79	73.61	-	-	-
ECNU$_{22}$	73.83	66.05	69.72	66.01	57.71	61.58	-	-	-
BUPT$_{22}$	50.41	42.55	46.15	49.95	42.72	46.05	-	-	-
GDUFE$_{23}$	75.53	75.60	75.56	69.71	67.33	68.50	-	-	-
SJTU$_{23}$	47.41	46.45	46.92	46.44	45.68	46.06	-	-	-
SUDA$_{23}$	80.82	81.79	**81.30**	74.39	75.03	74.71	-	-	-
WestlakeNLP$_{23}$	74.40	70.24	72.26	70.42	68.63	69.52	-	-	-
Ours	80.80	81.11	80.96	75.13	74.57	**74.85**	67.05	66.77	**66.92**

Table 3: 评测结果对比

	Total			Instance	Attribute	Relation	Coref	Ralign
	P	R	F_1	F_1	F_1	F_1		
TestA	80.80	81.11	**80.96**	**86.34**	94.71	**66.96**	1.60%	11.63%
TestB	75.13	74.57	74.85	81.09	89.02	59.94	2.48%	15.42%
TestC	67.05	66.77	66.92	68.52	**95.97**	46.79	3.18%	11.55%

Table 4: 三个测试集的Align-Smatch得分和特殊关系概率

3.3 分析

对于现代文测试集的语义解析，大语言模型依然存在一定的进步空间。在长难句的语义解析上，微调后模型的表现仍然不尽如人意，容易出现解析不完整或者解析偏离的情况，可能是由于幻觉现象以及窗口限制的原因。另外，对于关系的预测这一瓶颈，大模型在发现句子中词与词之间逻辑关系的能力上仍需改进。

如 3.2 所述，在表 4 中可以看出古汉语测试集与现代汉语测试集分数差异较大，差异主要体现在节点（Instance）和关系（Relation）的预测上。在节点属性（Attribute）的预测上，三个测试集均效果不错；在节点预测上相较于现代汉语测试集 A 与 B 上 86.34 和 81.09 的分数，古汉语测试集 C 的分数仅有 68.52；另外在关系预测上，古汉语测试集的分数仅有 46.79，明显低于两个现代汉语测试集。

综合以上，我们推测 TestC 预测效果较差的原因是古汉语和现代汉语句法结构上的差异。例如，古汉语中存在大量虚词，而现代汉语中虚词被简化甚至省略；古汉语的语序存在较多倒装，句法结构较为复杂，而现代汉语结构简洁明了；古汉语中存在一些文字在现代汉语中出现频率较低，仅依靠分词结果很难直接解析其真实词性。以上分析表明，在不充分掌握古汉语语义表示规律的前提下，模型仅依靠分词句子进行的推测较难达到期望结果。

4 结论

在本次 CAMR 解析评测任务中，我们系统地评估了现有中文大模型在CAMR解析任务上的能力，在此基础上使用图融合策略进一步增强基于大模型的CAMR解析系统的性能。实验结果表明，现有的中文大模型已经具备较强的少样本和全样本训练后CAMR解析能力，并且结合图融合技术能够进一步提升系统的性能。此外，相较于现代汉语，古汉语解析仍然是一个比较难的挑战。今后，对于古汉语的 CAMR 解析问题，我们认为一种潜在的解决方案是尝试将古汉语和现代汉语的对齐信息加入训练过程，另外一种是通过自动构造大量古汉语和对应 CAMR 图进行训练来提升解析性能。

致谢

感谢所有审稿人对本文提出的宝贵建议，使本文的内容更加完善和系统。本工作由深圳市高等院校稳定支持计划项目（GXWD20231130104007001）资助。

参考文献

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022a. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Xuefeng Bai, Sen Yang, Leyang Cui, Linfeng Song, and Yue Zhang. 2022b. Cross-domain generalization for AMR parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Liang Chen, Bofei Gao, and Baobao Chang. 2022. A two-stage method for Chinese amr parsing. *ArXiv*, abs/2209.14512.

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

Wenyang Gao, Xuefeng Bai, and Yue Zhang. 2023. System report for CCL23-eval task 2: WestlakeNLP, investigating generative large language models for Chinese AMR parsing. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*.

Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam M. Nguyen, Dzung T. Phan, Vanessa López, and Ramón Fernandez Astudillo. 2021. Ensembling graph predictions for AMR parsing. In *Annual Conferenceon Neural Information Processing Systems*.

Ziyi Huang, Junhui Li, and Zhengxian Gong. 2021. Chinese AMR parsing based on sequence-to-sequence modeling. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*.

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

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chunchuan Lyu and Ivan Titov. 2018. AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Long H. B. Nguyen, Viet H. Pham, and Dien Dinh. 2021. Improving neural machine translation with AMR semantic graphs. *Mathematical Problems in Engineering*.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, A. V. Podolskiy, Grigory Arshinov, A. Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. 2023. Pangu-σ: Towards trillion parameter language model with sparse heterogeneous computing. *ArXiv preprint*, abs/2303.10845.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Chuan Wang, Bin Li, and Nianwen Xue. 2018. Transition-based Chinese AMR parsing. In *North American Chapter of the Association for Computational Linguistics*.

Xiao Wang, Wei Zhou, Can Zu, Han Xia, Tianze Chen, Yuan Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, J. Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *ArXiv preprint*, abs/2304.08085.

Liming Xiao, Bin Li, Zhixing Xu, Kairui Huo, Minxuan Feng, Junsheng Zhou, and Weiguang Qu. 2022. Align-smatch: A novel evaluation method for chinese abstract meaning representation parsing based on alignment of concept and relation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *ArXiv preprint*, abs/2210.02414.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

Shilin Zhou, Qingrong Xia, Yang Li, Zhefeng Wang, and Zhenghua Li. 2022. Suda-huawei camr2022 比赛技术评测报告. In *Proceedings of the 21nd Chinese National Conference on Computational Linguistics*.