CCL25-Eval任务12系统报告：基于端到端模型以及指令微调方法的面向中文语音的实体关系三元组抽取研究

李南书
赣西肿瘤医院/ 江西省萍乡市
879075435@qq.com

摘要

传统的关系三元组抽取任务主要集中于书面文本，通过识别实体及其相互关系来构建结构化的知识图谱。然而，语音作为人机交互的主要形式之一，在智能助手、智能客服、语音搜索等诸多应用中发挥着日益重要的作用。因此，如何高效、准确地从语音数据中提取有价值的结构化信息成为研究的热点之一。本研究通过测试模型在数据集上的性能，探究如何增强模型在三元抽取任务中的能力。本研究使用的训练框架是LLamafactory，使用的大模型是两个7B量级的开源模型（qwen2-audio，qwen2.5-omin(Qwen Team, 2025)），首先任取其中的一个模型（本研究选取的为qwen2-audio）设置lora参数为LLamafactory默认参数，修改参数中验证集比例为0.2，epoch为5，进行lora监督微调，获得验证集最佳的epoch。然后，设置lora参数为默认，修改其中的epoch参数为验证集最佳epoch+1，同时对两个模型进行全数据lora监督微调，择其中更优胜者，最后进行进一步的lora调参，以期模型在该任务上达到相对最优性能。最终在B榜获得了end-to-end赛道的第二名，分数为0.5292。

关键词： 中文语音；三元组抽取；监督微调；大语言模型；LLamafactory

System Report for CCL25-Eval Task 12: End-to-End Model and Instruction Fine-Tuning-Based Chinese Speech-Oriented Entity-Relation Triplet Extraction Research

Nanshu LI
Ganxi Cancer Hospital /Pingxiang City, Jiangxi Province
879075435@qq.com

Abstract

Traditional relation triplet extraction tasks have primarily focused on written text, constructing structured knowledge graphs by identifying entities and their mutual relationships. However, speech, as one of the main forms of human-computer interaction, plays an increasingly important role in many applications such as smart assistants, intelligent customer service, and voice search. Therefore, how to efficiently and accurately extract valuable structured information from speech data has become a research hotspot. This study investigates how to enhance model capabilities in triplet extraction tasks by testing model performance on a dataset. The training framework used in this study is LLamafactory, and the large models employed are two open-source models of the 7B scale (qwen2-audio, qwen2.5-omin(Qwen Team, 2025)). First, one of the models (qwen2-audio was selected in this study) was chosen, with LoRA parameters set to LLamafactory's default values. The validation set ratio in the parameters was

第二十四届中国计算语言学大会论文集，第470页-第474页，济南，中国，2025年8月11日至14日。

modified to 0.2, and the number of epochs was set to 5 for LoRA supervised fine-tuning to obtain the optimal epoch on the validation set. Then, with LoRA parameters kept as default, the epoch parameter was adjusted to the optimal validation epoch + 1, and both models underwent LoRA supervised fine-tuning on the full dataset. The superior model was selected, followed by further LoRA parameter tuning to achieve relatively optimal performance on this task. Finally, the model ranked second in the end-to-end track on the B Benchmark with a score of 0.5292.

Keywords: Chinese speech , triplet extraction , supervised fine-tuning , large language models , LLamafactory

1 引言

面向中文语音实体关系三元组抽取任务（Chinese Speech Entity-Relation Triple Extraction Task，CSRTE）旨在从中文语音数据中端到端地自动识别并提取实体及其相互关系，构建结构化的语音关系三元组（头实体、关系、尾实体）。任务的目标在于提升中文语音关系三元组抽取的准确性和效率，增强系统在不同语境和复杂语音环境下的鲁棒性，实现从语音输入到三元组输出的全流程自动化处理。通过本次评测，旨在推动中文语音信息抽取技术的发展，促进语音与自然语言处理技术的深度融合，为智能应用提供更加丰富和精准的基础数据支持。

在end-to-end赛道中，本研究主要运用的是qwen2-audio-7b，qwen2.5-omin-7b两组开源大模型，通过LLamafactory框架，采用框架默认lora参数，比较两组模型一般效果，选取其中优胜者，再采用固定微调策略，选取不同组合的lora-rank,lora-alpha的值，进行多轮测验，最终获取相对最佳的lora-rank,lora-alpha的值为最终取值，在B榜获得了0.5292分，为该赛道第二名。

2 方法

2.1 监督微调

预训的有语音模块的大模型擅长通用任务（语音内容理解，语音转述等），但在具体任务中表现可能不足。通过监督微调，模型可学习任务特定的模式，本研究通过构建提示词模版与数据融合，使模型加强了格式化抽取语音内容中的三元组关系的能力。

2.2 Lora

Lora（低秩适应）是Meta 于2021 年(Hu et al., 2022)提出的一种参数高效微调（PEFT, Parameter-Efficient Fine-Tuning）技术，旨在减少大语言模型（LLMs）微调时的参数量，同时保持接近全量微调的性能。其核心思想是通过低秩分解来近似参数更新，从而大幅降低可训练参数的数量。

3 研究设置

3.1 数据集介绍

本次评测所使用的数据集来源于开源的Common Voice 17中文子集和AISHELL语音识别数据集，经过专业标注人员对其中的实体及其关系进行精确标注，构建了中文语音关系三元组抽取的评测数据集。该数据集总计约40小时的中文语音数据，包含约20,000条语句，每条语句均提供对应的语音文件以及其中包含的实体文本和关系标注结果（不含原始的转录文本）。标注的实体类别涵盖人物、组织、地点、产品、著作等10类，关系类型预定义超过50种，确保涵盖多样化的语义关系。数据内容涵盖新闻报道、日常对话、正式演讲等多种场景，保证了数据的多样性和代表性，旨在支持系统在不同语境和复杂语音环境下的全面评测。

3.2 评价指标

评测任务将采用以下指标对参赛系统的三元组抽取性能进行评估，各指标分段描述如下：

- **准确率（Precision, P）**：正确抽取的三元组数Z与系统抽取的总三元组数N之比。

$$P = \frac{Z}{N} \tag{1}$$

- **召回率（Recall, R）**：正确抽取的三元组数Z与数据集中真实三元组总数Nz之比。

$$R = \frac{Z}{Nz} \tag{2}$$

- **F1值（F1-score）**：准确率和召回率的调和平均数，作为综合性能指标。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

3.3 数据预处理

原数据只有id和data-annotation两个关键字，数据简洁明了；出于对提示词是否会对最终结果有较大影响的好奇，对于该任务构建了两组不同的prompt，一组详细复杂即prompt1，另一组粗略简单即prompt2。

prompt1：你是一个专业的三元组抽取工具，请从以下中文语音¡audio¿中提取所有可能的(头实体, 关系, 尾实体)三元组.要求：1.输出格式为列表，例如：[[”实体1”, ”关系”, ”实体2”], ...],[[”实体1”, ”关系”, ”实体3”]],...]2.如果你认为该中文句子中不能提取到三元组关系请返回[],示例：提问：我们需要有能力为这些非美国业务的运转提供资金，回答：[]3.如果你认为该中文句子中能提取到三元组关系请按前述既定的的格式返回,示例：提问：中新社北京七月七日电记者刘舒凌和尚都有钱；回答：[[[”刘舒凌”,”持有称号/职业”,”记者”]],[[”中新社”,”位于/业务地点”,”北京”]],[[”北京”,”位于”,”中国”]]]”。

prompt2：”你是一个专业的三元组抽取工具，请从以下中文语音<audio>中提取所有可能的(头实体, 关系, 尾实体)三元组。”，直接将prompt设置为input。

output设置为数据集data-annotation中的内容如：[[[”刘舒凌”,”持有称号/职业”,”记者”]],[[”中新社”,”位于/业务地点”,”北京”]],[[”北京”,”位于”,”中国”]]]。然后将两组数据构造成符合LLamafactory要求的格式分别加入其data数据库并在数据库中注册。

4 研究流程

4.1 基本参数设置

cutoff_len: 1024
learning_rate: 0.0001
num_train_epochs: 5
lr_scheduler_type: cosine
warmup_ratio: 0.1
bf16: true
optim: adamw_torch

Table 1: 模型训练基本参数设置

4.2 获取相对最佳epcoh

选择qwen2-audio模型，采用默认的lora参数设置以及设置验证集比例为0.2即： 最终结果推荐val最佳位置在epcoh为3附近，因此最终选择之后的全数据训练epoch为4。

lora_rank: 8
lora_alpha: 16
lora_dropout: 0
lora_target: all
val_size: 0.2

Table 2: LoRA参数配置表

模型	分数
qwen2-audio	0.432
qwen2.5-omin	0.503

Table 3: 模型性能对比

4.3 模型选择

将基本设置的epcoh修改为4，注释掉val_size参数，其余参数不变，分别对qwen2-audio，qwen2.5-omin两个模型进行lora微调训练，结果如下：

根据结果确定qwen-2.5-omin为最后选择模型。（因两模型效果差距较大，未对中间保存的不同步数的lora模型进行对比）

4.4 最佳lora参数确定

选择不同的lora参数值以及不同步数保存的lora模型，确定最终的提交模型：

	checkpoint-steps			
	2500	3000	3500	4000
lora_rank: 8，lora_alpha: 16	0.5125	0.5210	0.5211	0.5175
lora_rank: 16，lora_alpha: 32	0.5185	0.5235	0.5205	0.5203
lora_rank: 32，lora_alpha: 64	0.5271	0.5263	0.5268	0.5260
lora_rank: 64，lora_alpha: 128	0.5275	0.5292	0.5280	0.5252
lora_rank: 128，lora_alpha: 256	0.5270	0.5278	0.5281	0.5273

Table 4: 不同LoRA配置的模型性能

5 结论

1.本次比赛后期由于时间和资源的问题，采用的研究方法比较粗糙，因此qwen2.5-omin在更加精细的参数设置下，应该可以取得更好的成绩。

2.qwen2.5-omin 与qwen2-audio 两者均为7b量级的模型，但qwen2.5-omin实为多模态模型其亦具有视频模块功能，因此其音频模块参数应显著少于qwen2-audio，但根据实验结果，qwen2.5-omin在当前任务，显著强于qwen2-audio，鉴于目前大模型基本基于transformer架构，因此推测应是训练方法以及训练数据有显著优化。

3.在4.2步骤时用qwen-audio模型分别对之前生成的两组具有不同的prompt的数据集进行了微调训练，对比结果发现并无明显差别，虽然试验的次数不多，但是否依然可以推测提示词的详尽与否是否对监督微调大模型类任务无太大影响。

4.在4.2步骤时，额外测试了把rag（即将梅尔频谱编码后的语音数据以及其对应的data-annotation生成rag库）加入prompt的效果，对比结果发现并无明显差别。

参考文献

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and others. 2022. *Lora: Low-rank adaptation of large language models. ICLR*, 1(2):3.

Qwen Team. 2025. *Qwen2.5-Omni Technical Report*. https://github.com/QwenLM/Qwen2.5-Omni/blob/main/assets/Qwen2.5_Omni.pdf.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. *Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems*, 36:10088–10115.