# CCL24-Eval任务2总结报告：中文意合图语义解析评测

郭梦溪[1]，李梦[1]，靳泽莹[1]，吴晓靖[1]，饶高琦[2]，唐共波[1]，荀恩东[3*]
[1]北京语言大学 信息科学学院
[2]北京语言大学 国际中文教育研究院
[3]北京语言大学 语言资源高精尖中心
guo_mengxi@foxmail.com

## 摘要

中文意合图是近年提出的中文语义表示方法。本次评测是首次基于意合图理论的语义分析评测，旨在探索面向意合图理论的语义计算方法，评估机器的语义分析能力。本次评测共有14支队伍报名，最终有7支队伍提交结果，其中有5支队伍提交技术报告与模型，均成功复现。在评测截止时间内，表现最好的队伍使用大语言模型LoRA微调方法获得了F1值为72.06%的成绩。在最终提交技术报告的5支队伍中，有4支队伍使用了大语言模型微调方法，在一定程度上表明了目前技术发展的趋势。

**关键词：** 意合图；语义分析；通用语义框架；评测任务

# Overview of CCL24-Eval Task2: Chinese Parataxis Graph Parsing Evaluation

**Mengxi Guo[1],Meng Li[1], Zeying Jin[1],Xiaojing Wu[1],**
Gaoqi Rao[2],Gongbo Tang[1],Endong Xun[3*]
[1]School of Information Science, Beijing Language and Culture University
[2]Research Institute of International Chinese Language Education,
Beijing Language and Culture University
[3]Beijing Advanced innovation Center for language Resources,
Beijing Language and Culture University
guo_mengxi@foxmail.com

## Abstract

The Chinese Parataxis Graph is a recently proposed method for Chinese semantic representation. This evaluation is the first semantic analysis evaluation based on the theory of the Chinese Parataxis Graph. It aims to explore semantic computation methods oriented towards the Chinese Parataxis Graph theory and to evaluate the semantic analysis capabilities of machines.A total of 14 teams registered for this evaluation, with 7 teams submitting results. Among them, 5 teams submitted technical reports and models, all successfully replicated the results. By the evaluation deadline, the best-performing team, which used the LoRA fine-tuning method with a large language model, achieved an F1 score of 72.06%. Among the 5 teams that submitted technical reports, 4 teams used the fine-tuning method with large language models, indicating a current trend in technological development to some extent.

**Keywords:** Chinese Parataxis Graph , Semantic parsing , Evaluation , Universal semantic framework

* 通讯作者

## 1 评测背景

随着自然语言处理技术的不断发展，语义分析作为语言理解的重要组成部分，受到了广泛的关注和研究。意合图是荀恩东近年来提出的一种以事件为中心的语义表示方法，采用单根有向图的形式承载事件、实体、属性及其相互关系。过往的语义表示多是在不同层级单位进行相应的表示，如词、句子、段落、篇章等不同层次的语言单元都有不同的语义表示方法。而意合图力求通过统一的表示框架，对不同层级的语言单元进行一致表示。

意合图理论经历了早期理论架构 (荀恩东, 2023)，以及基于早期理论架构的工程实践 (王诚文, 2021; 王贵荣, 2023)。通过工程实践我们优化并完善了的意合图理论架构，构建了完整的意合图通用语义体系 (郭梦溪等, 2024)，并基于意合图理论与通用语义体系构建了一批意合图语义标注资源 (郭梦溪等, 2024)。为了探索意合图的最佳计算方法，并评估当前机器的语义分析能力，我们组织了本次中文意合图语义解析评测。

## 2 评测任务

### 2.1 相关概念

意合图将通过语言所表征的事件定义为两种，一种是现实世界或可能世界中的事物的动作行为或关系描述，另一种是现实世界或可能世界中的事物间关系，相对应地，意合图将第一种事件称为意合图所表征的一般事件，第二种事件称为意合图所表征的关系事件。意合图将事件词作为事件的核心表达，其中一般事件的事件词常为在句中出现的连续或非连续谓词性语言单元，如汉语的离合词即为非连续事件词；如果事件词在句中省略情况或汉语特殊的名词谓语句，则对事件词进行补全；关系事件的事件词为抽象出的关系概念词，如“因果关系”“同指关系”等。

意合图在符合人类对语言认知的基础上，充分考虑落地应用的可操作性，使其尽可能地层次化，以便于后续语义分析路径的设计，实现通用性与扩展性兼具的语义表征方案。按照层次可将意合图分为事件结构与实体结构两大部分。事件结构分为事件内事件内结构与事件外结构，事件内结构可进一步分为以事件词为核心的论元结构、情态结构、时空结构，事件外结构为多个事件构成的关系事件结构；实体结构分为实体内结构与实体外结构，实体内结构即实体属性与属性值结构，实体外结构即多个实体构成的实体关系事件结构。本次评测所发布的标注语料的语义标签体系如表1所示，关系论元与关系事件具有对应性，单独于表2展示。其抽象表示如图1所示。

| 层级 | 包含的语义标签类 |
|---|---|
| 论元结构<br>（一般论元） | A0, A1, A2, 工具, 材料, 方式, 依据, 原因, 目的, 范围,<br>数量, 数量源点, 数量终点, 状态, 状态源点, 状态终点<br>（关系事件论元见表2） |
| 情态结构 | Mod（此次评测不对情态内部细分） |
| 时空结构 | 时间, 时间源点, 时间终点,Time（此次评测不对时态时制等时间信息细分）；<br>处所, 处所源点, 处所终点, 趋向 |
| 实体属性 | EntityRel（此次评测不细分实体属性） |
| 特殊标签 | Comp，NER，插入语，Merge，离合，重叠，不宜还原，宜还原 |
| 形式标记 | PN，CompPN，Conj，FW，TM，PF，SF, X |

Table 1: 数据集语义标签（部分）

| 关系事件（词） | 对应的关系论元 |
|---|---|
| 时序关系 | 先行事件，后继事件，伴随事件 |
| 递进关系 | 基本事件，递进事件 |

Table 2: 关系事件及关系论元

Table 2 – 续

| 关系事件（词） | 对应的关系论元 |
|---|---|
| 转折关系 | 让步事件，转折事件 |
| 因果关系 | 原因事件，结果事件 |
| 条件关系 | 条件事件，关系事件 |
| 目的关系 | 目的事件，行动事件 |
| 并列关系（And） | 并列事件，并列实体 |
| 选择关系（Or） | 候选事件，选定事件，候选实体 |
| 同指关系（Ref） | Entity，Event |
| 领属关系 | （此次评测不区分领属关系） |
| 整分关系 | （此次评测不区分整分关系） |

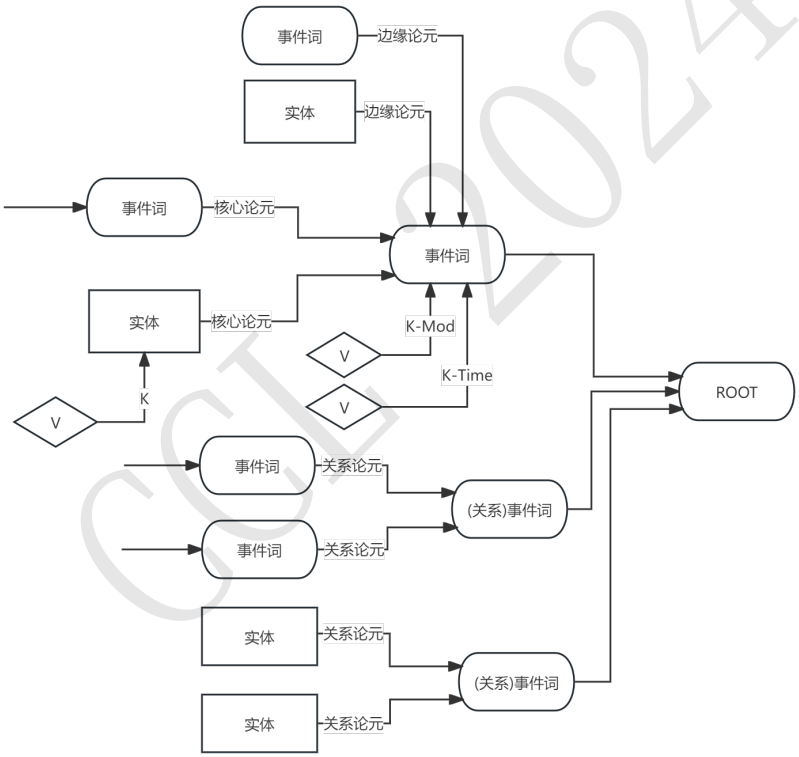


Figure 1: 意合图的抽象表示

## 2.2 任务要求

本次中文意合图语义解析评测任务仅要求生成句子级意合图框架即可，即输入单元为分词后的句子，输出为意合图框架结构，无需生成细化实体结构、情态结构、时空结构等的内部语义分类，仅判断是否属于该结构成分即可，所提供的语料也为粗粒度标签。此外，本次评测任务的参赛队伍可自行借助形式标记辅助个别语义的识别，但最终不要求对形式标记进行识别，不在结果中进行计算。

# 3 评测数据

## 3.1 数据分布

本次评测任务的数据集语料抽取自BCC（Beijing Language and Culture University Corpus Center）(荀恩东等, 2016)国际中文教育阅读语料库与中文宾州树库8.0的新闻语料，共计5000条标注语料。数据分布如表3所示。

| | 国际中文教育阅读语料 | 新闻语料 | 总计 |
|---|---|---|---|
| 训练集 | 701 | 1299 | 2000 |
| 验证集 | 351 | 649 | 1000 |
| 测试集 | 351 | 649 | 1000 |
| 盲测集 | 351 | 649 | 1000 |
| 总计 | 1754 | 3246 | 5000 |

Table 3: 数据分布

需要说明的是，本次评测任务允许参赛队伍根据需求对除盲测集外的数据集分布进行重分配。盲测集仅提供给参赛队伍分词后的输入，由参赛队伍进行结果推理，将预测结果返回。最终根据盲测集结果进行排名。

## 3.2 数据标注

本次评测的数据集由8位具有语言学背景的研究生在标注规范的指导下完成标注。每次任务先由随机两人进行独立标注，然后双方再根据管理者返回的不一致标注结果进行讨论，确定唯一标注结果。无法达成一致的情况，由管理者介入进行确认。最后管理者对标注结果进行全检，再次修正错误。因此，每条语料经过多次确认，以保障标注数据的质量。且数据集中的每条语料的标注结果均经验证，能够生成完整意合图，即标注结果中不存在游离成分或违反意合图原则的情况。

## 3.3 数据格式

除盲测集外，其他发布给参赛队伍的数据集文件均为UTF8编码，Json格式，包括句子的分词信息和标注三元组。如样例（图2）所示，“sent”值域是句子分词结果，“relData”值域是该句子的所有标注信息，其中“word1”和“word2”值域均包含节点内容word和节点编号idx，当word为句中词汇时，idx为该词在句中的编号（从0开始），当word为隐式事件词或实体省略标签时，idx的对应关系如表4所示。需要注意的是，当句中存在不止一个某一种隐式事件时，word对应的内容将在字符串后面增加数字，并向上叠加，idx也将在-13的基础上减1。例如，句中存在三个因果关系，则其对应的word分别为：因果关系、因果关系1、因果关系2，idx分别为：-8、-14、-15。


```
[
  {"sent":["我","对不起","大家",",","我","没有","完成","任务","。"],
"relData":[
{"word1":{"word":"我","idx":0},"word2":{"word":"对不起","idx":1},"relVal":"A0"},
{"word1":{"word":"对不起","idx":1},"word2":{"word":"ROOT","idx":-1},"relVal":"CoreWord"},
{"word1":{"word":"对不起","idx":1},"word2":{"word":"因果关系","idx":-8},"relVal":"结果事件"},
{"word1":{"word":"大家","idx":2},"word2":{"word":"对不起","idx":1},"relVal":"A1"},
{"word1":{"word":"我","idx":4},"word2":{"word":"完成","idx":6},"relVal":"A0"},
{"word1":{"word":"没有","idx":5},"word2":{"word":"完成","idx":6},"relVal":"m否定"},
{"word1":{"word":"完成","idx":6},"word2":{"word":"ROOT","idx":-1},"relVal":"CoreWord"},
{"word1":{"word":"完成","idx":6},"word2":{"word":"因果关系","idx":-8},"relVal":"原因事件"},
{"word1":{"word":"任务","idx":7},"word2":{"word":"完成","idx":6},"relVal":"A1"},]}
]
```


Figure 2: 数据格式

| 隐式事件词 | ROOT | And | Or | Ref | 时序关系 | 递进关系 | 转折关系 | 因果关系 | 条件关系 | 目的关系 | 重叠 | Is | QS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **idx** | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 | -9 | -10 | -11 | -12 | -13 |

Table 4: 隐式事件词索引

## 4 评价指标

本次评测采用F1值作为模型表现的评价标准与排名的主要依据，计算方式如下：

$$P = \frac{\text{count(Matching Tuples)}}{\text{count(Generated Tuples)}} \tag{1}$$

$$R = \frac{\text{count(Matching Tuples)}}{\text{count(Gold Tuples)}} \tag{2}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

其中“Generated Tuples”为模型预测的三元组集合数，“Gold Tuples”为测试集/盲测集的三元组集合数，“Matching Tuples”为模型预测的三元组集合与测试集/盲测集的三元组集合间的最大匹配个数。

## 5 评测结果

### 5.1 提交结果

本次评测共有14支队伍报名参赛，包括北京大学、北京理工大学、北京语言大学、郑州大学、湘潭大学、南京理工大学、沈阳航空航天大学等高校，中科院计算所等研究所，中国太平洋保险集团、中电信人工智能科技有限公司、北京恺望数据科技有限公司等企业。最终有7支队伍提交了结果，其中有5支队伍提交技术报告，并成功复现其模型。由于本次评测中大多参赛队伍使用了大模型，无法保证每次复现结果完全一致，因此我们接受复现结果与提交结果存在小幅度差异，如复现没有问题，最终排名仍以参赛队伍提交的结果为准。在评测结果提交的规定时间内所提交的盲测集结果如表5所示，来自北京理工大学的队伍1所提交的结果F1值最高，为72.06%。

| 队伍编号 | 队伍单位 | **Precision(%)** | **Recall(%)** | **F1-score(%)** |
|---|---|---|---|---|
| 1 | 北京理工大学 | 70.11 | **74.10** | **72.06** |
| 2 | 北京理工大学 | **70.28** | 73.18 | 71.70 |
| 3 | 北京语言大学 | 62.65 | 66.70 | 64.61 |
| 4 | 北京语言大学 | 64.22 | 64.66 | 64.44 |
| 5 | 中国太平洋保险公司 | 57.80 | 61.95 | 59.80 |
| 6 | 湘潭大学 | 57.89 | 58.53 | 58.21 |
| 7 | 沈阳航空航天大学 | 55.89 | 57.51 | 56.69 |

Table 5: 盲测集结果排名

### 5.2 方法概述

#### 5.2.1 基于大语言模型的微调方法

随着大语言模型的不断发展，其参数量不断增大，模型的泛化能力不断增强，微调方法便捷。提交技术报告的5支队伍中，有4支队伍使用了大语言模型微调方法。所使用的基座模型涉及闭源模型百度文心系列的ERNIE4.0-Speed-8K，开源模型Llama3-Chinese-8B-Instruct、Chinese-Llama2-7B-Chat，且均使用了LoRA微调方法。

来自北京理工大学的队伍1和队伍2均使用了百度文心系列的ERNIE4.0-Speed-8K闭源模型作为基座模型在百度智能云的千帆大模型平台进行微调。二者的微调方法不同，队伍1使用

了LoRA微调方法，队伍2使用了全参数微调。在Prompt-Response的设计上，二者的Prompt设计类似，输入均为分词后的句子，无意合图概念性描述。Response的设计上，队伍1保留了索引编号，而队伍2未直接输出索引编号。在训练集条数上，队伍1自行分配了2850条语料，队伍2分配了3600句语料。在超参数设置上，两个队伍在迭代轮次、学习率、学习率调整计划等设置上存在差异。最终两个队伍所提交的盲测集预测结果的F1值较为接近，分别为72.06%、71.70%。在结果提交截止后，队伍2又将训练语料条数增加至4000句，在参数上进行少量调整，使得模型盲测集预测结果的F1值提升至74.76%（该结果经过复现核对），并且又进一步探讨了使用小参数规模的模型达到比赛中所用的闭源模型在这一任务上的表现的可行性。该队伍集成了包括Qwen1.5-7B，ChatGLM3-6B，Yi1.5-6B，DeepSeek-7B在内的国内开源模型的轻量级版本，设计了一种循环增强微调训练模块，并采用一种级联监督的方式融合模型的输出结果。这种方法成功地实现结合小规模参数的开源模型达到与百度文心系列的ERNIE-Speed-8K模型相近的效果。

来自北京语言大学的队伍3使用了开源模型Llama3-Chinese-8B-Instruct在本地进行LoRA微调。该参赛队伍对数据集分布进行了统计，根据统计结果进行了样本设计工程（Sample Design Engineering，SDE），将上下文、指令及输出指示放置于输入的任务文本之前以提升模型的任务理解能力，并且根据意合图标签的出现频次，由高到低排序以提高模型对出现次数多的标签的关注程度。该队伍通过实验发现大模型对不同语义标签解析难度不同，并设计了不同的模型训练策略，对解析结果进行了组合分析。最终该队伍所提交的盲测集预测结果的F1值为64.61%。

来自湘潭大学的队伍6使用Chinese-Llama2-7B-Chat在阿里云人工智能平台PAI进行LoRA微调。最终所提交的盲测集预测结果的F1值为58.21%。

### 5.2.2 基于roBERTa的关系抽取方法

来自北京语言大学的队伍4将意合图语义解析任务转换为传统的关系抽取任务。而意合图与传统关系抽取任务相比，其三元组内词的顺序不可变，且存在句外词，即意合图的隐式事件词、根节点“ROOT”、实体省略标签“QS”等。对此，该队伍将三元组内不符合原句语序的“实体对”的关系改为“关系标签_reverse”，以解决词对顺序不可变问题；将句外词添加在原句末尾作为输入，以此解决句外词问题。通过上述处理，将意合图语义解析任务转变为了关系抽取任务。但该处理方式也使得原本就不平衡的标签分布加重，因此该参赛队伍将任务划分为两个子任务，即不包含隐式事件词的关系抽取和包含隐式事件词的关系抽取。该参赛队伍将任务分为关系识别与关系分为两部分，均使用了哈工大版本的chinese-roBERTa-wwm-extlarge模型。最终该参赛队伍取得了F1值为64.44%的成绩。

## 5.3 其他分析

本次评测中有些参赛队伍不仅完成了评测，还在过程中进行了对比实验、结果分析等，对意合图的进一步研究提供了参考。

来自北京理工大学的队伍1对六个开源大模型进行了测试，Yi-1.5-9B在测试集上的F1值为60.21%，其余五个在测试集上的F1值均不足60%，而ERNIE-Speed在同一份测试集上的F1值达69.56%。为进一步探究各种因素对于模型性能影响的程度，该队伍对参数规模和基座模型系列两个因素的影响进行探究。实验结果表明，同属Qwen-1.5系列的四种参数规模的模型在该任务上的表现基本一致，并没有随着参数量的增加而在该任务上表现更优越。其次，参赛队伍选用Baichuan2-7B、Qwen-1.5-7B和Yi-1.5-9B三个参数规模类似的不同系列的模型，采用完全相同的超参数和工具进行微调。实验结果表明，三者表现差异较大，说明了基座模型对于该任务的影响较大。通过该实验，参赛队伍得出其所使用的文心系列ERNIE4.0-Speed-8K模型在该任务中的出色表现与百度的预训练语料、模型的技术细节以及千帆平台的微调实现等因素更为相关。

来自北京语言大学的队伍3对模型在训练集上各语义标签的错误率进行排序，其中错误率最低的语义标签为“CoreWord”，即核心事件词的判断；错误率最高的语义标签为“选定事件”。对易预测错误的语义标签进行分析，发现一些语义标签错误率高可能是由于训练集中该类数据过少，也存在一些语义标签在训练集中数据不少，但大模型仍难掌握。此外，该队伍还在相同实验条件下微调了六个同等规模的开源大模型，尤其关注了中文大模型与中文语料增量预训练的英文大模型。实验结果显示，在该任务下中文大模型展现出更高的适应性与优越性。

## 6 总结

本次评测是意合图正式提出后首次参与评测活动，共吸引了来自高校、研究院以及企业的14支队伍报名，最终有7支队伍提交了结果，其中来自北京理工大学的队伍以F1值达72.06%的成绩取得了本次评测的第一名。在提交技术报告的5支队伍中，有4支队伍使用了大模型微调的方法。该情况充分展现了大模型微调已成为当前技术研究与应用中的主流选择，更多团队倾向于采用更为高效的微调方法来解决各种任务。

参赛队伍的实验结果为我们的研究提供了宝贵的参考经验。基于本次评测所反映出的情况，我们将在未来的研究中，探索基于更优秀的基座模型和更精细的微调工程，获得更优异的分析效果。为典型的语义关系挖掘提供更多表达方式作为备选资源，使得能够更大程度地提高标签精度，为更精准的语义分析提供支持。在资源建设方面，评测中所呈现的方法将助力我们构建更高质量的意合图语义资源，更高效地推动意合图在各类应用场景中的实践。我们将不断优化和改进意合图的构建和应用方法，期待在语义分析领域取得更大的突破，推动语义分析的发展。

## 致谢

本研究得到国家自然科学基金“中文意合图的表征与生成方法研究”（62076038）与中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）的支持。本次评测所使用的数据集由北京语言大学李梦、何晴、胡星雨、王静怡、吴晓靖、张可芯、周书帆、朱奕瑾（按姓氏排）八位研究生完成标注，感谢各位标注员所作出的贡献。数据集标注所使用的在线标注平台最初由张梦圆构建，于钟洋、宋玉良完成了新功能的实现，感谢三位研究生对数据集构建所作出的贡献。

## 参考文献

Liyang Pang, Chengwen Wang, Guirong Wang, Gaoqi Rao, and Endong Xun. 2021. *Prepositional Frame Extraction and Semantic Classification Based on Chinese ChunkBank.* CLSW, Nanjing.

Shufan Zhou , Chengwen Wang, Endong Xun. 2023. *Recognition of Disyllabic Intransitive Verbs and Study on Disyllabic Intransitive Verbs Taking Objects Based on Structure Retrieval.* Springer Nature Switzerland, 2023: 265–282.

郭梦溪, 荀恩东, 李梦, 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

郭梦溪, 李梦, 荀恩东, 饶高琦, 于钟洋. 2024. 基于意合图语义理论的结构标注体系与资源建设. 第二十三届中国计算语言学大会.

邵田, 翟世权, 饶高琦, 荀恩东. 2023. 基于结构树库的状位动词语义分类及搭配库构建. 中文信息学报,37(06):44-51+66.

田思雨, 邵田, 荀恩东, 饶高琦. 2023. 基于结构树库的补语位形容词语义分析及搭配构建. 第二十二届中国计算语言学大会论文集,第420页-第432页,哈尔滨.

王诚文, 钱青青, 荀恩东, 邢丹, 李梦, 饶高琦. 2020. 三元搭配视角下的汉语动词语义角色知识库构建. 中文信息学报,34(09):19-27.

王诚文. 2021. 面向意合图的汉语动词论知识构建研究. 北京语言大学博士论文.

王贵荣. 2023. 意合图事件结构标注及分析研究. 北京语言大学博士论文.

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

荀恩东. 2023. 自然语言结构计算：*BCC*语料库. 人民邮电出版社, 北京.

荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制. 语料库语言学.