## 实验设计

实验1：原始prompt+静态fewshot（常规做法）

实验2：原始prompt+动态fewshot（主题聚类法）（其他论文提出）

实验3：原始prompt+静态fewshot+动态规则约束（本文论文提出）

实验4：原始prompt+动态fewshot+动态规则约束（本文论文提出）

原始prompt=【T：任务描述】+【D：实体与关系定义】+【E：任务强调与思维链引导】+【E：审核指令与校验规则】+【S：Few-Shot】+【C：输入模板定义】

## 目前结论

1. 5篇文献测试

实验2优于实验1（效果评估方法：人工看）

实验3优于实验1（实验1到实验3阶段，初步人工增加规则约束）

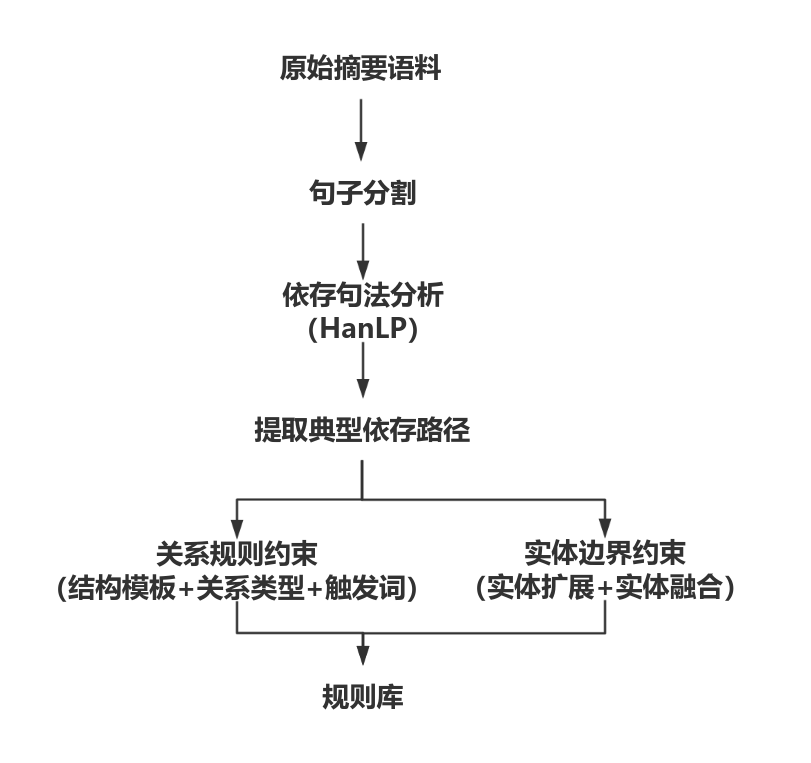
1. 规则的增加会提高模型提取的效果
   1. 在没有集中生成规则前我们用大模型提取实体和关系，为了更好的提取也会在prompt中描述一些规则与边界。例如：着重看基于XX的方法或者是提出XX的结构句子；常见的算法、研究方法名称有XX。本研究提出的方法可批量生成句法以及实体边界的规则并加入prompt，提高提取的效率

## 任务进度

1. 动态规则约束-句法依存分析

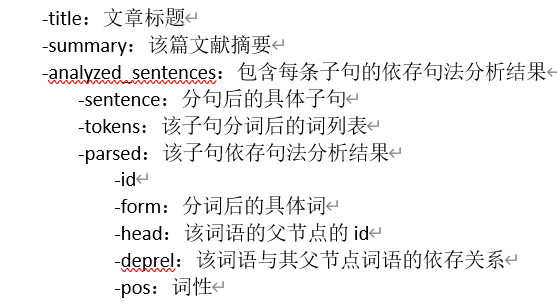
整体思路：利用依存句法分析自动识别句法结构，构建规则库以抽象常见实体关系模式，再将规则信息嵌入到Prompt中，指导大语言模型（如ChatGPT、DeepSeek等）在抽取过程中进行结构感知与语义对齐

流程图如下：

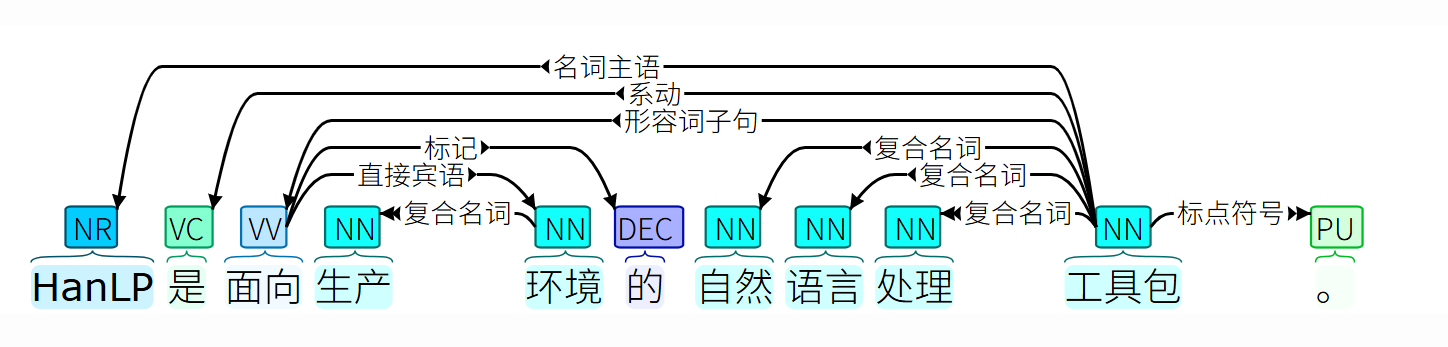


依存句法分析已经调HanLP处理完成待进行依存路径提取

具体得到结果包括：



句法分析示意图如下：



**依存路径提取步骤：**

定义：依存路径是指，在依存句法树中，连接两个实体（词元）之间的路径关系（包含路径上的依存关系和词性）

研究方法 ——[nsubj]→ 提出 ——[dobj]→ 模型

1. 确定实体对

从句子中找出所有可能的实体对组合（实体1，实体2）

1. 依存图中找“最短路径”

·在 HanLP 的依存分析结果中构建依存图（树）

·使用 BFS / DFS 寻找 实体1 到 实体2 的最短依存路径

·保留路径中每个节点的：词、词性、依存关系、方向

1. 路径格式化

·实体1 -[nsubj]→ 动词（谓词） -[dobj]→ 实体2

或[

{"word": "研究方法", "dep": "nsubj", "upos": "NN"},

{"word": "提出", "dep": "root", "upos": "VV"},

{"word": "模型", "dep": "dobj", "upos": "NN"}

]

1. 统计并归纳典型路径模式

·统计最频繁出现的依存路径结构

·提取结构上类似的抽象模式

如：方法 –[nsubj]→ 动词 –[dobj]→ 对象

==========理想规则库==========

主要生成实体与关系的约束规则：

* 关系规则约束（依存句法分析）：帮助LLM判断属于哪种关系（例如：
* 实体约束边界规则（词典/LLM语义理解/边界约束规则）：用来明确哪些词可作为实体被抽取

一条关系规则示例：

[

{

"relation\_type": "所属关系", // 映射到定义的目标关系之一

"dependency\_pattern": "nsubj→VERB→nmod:of", // 依存路径结构

"trigger\_words": ["属于", "隶属", "是...的"], // 典型触发词，用于判定关系

"entity\_1\_type": "组织", // 限定抽取的实体类型

"entity\_2\_type": "上级机构",

"example\_sentence": "该学院属于南京大学。",

"matched\_span\_template": ["实体1", "关系", "实体2"], // 用于 Prompt 构造的结构模板

"notes": "用于表达从属关系的主谓介宾结构" // 可选说明

}

{

"id": "R-101",

"type": "boundary\_rule",

"entity\_type": "研究方法",

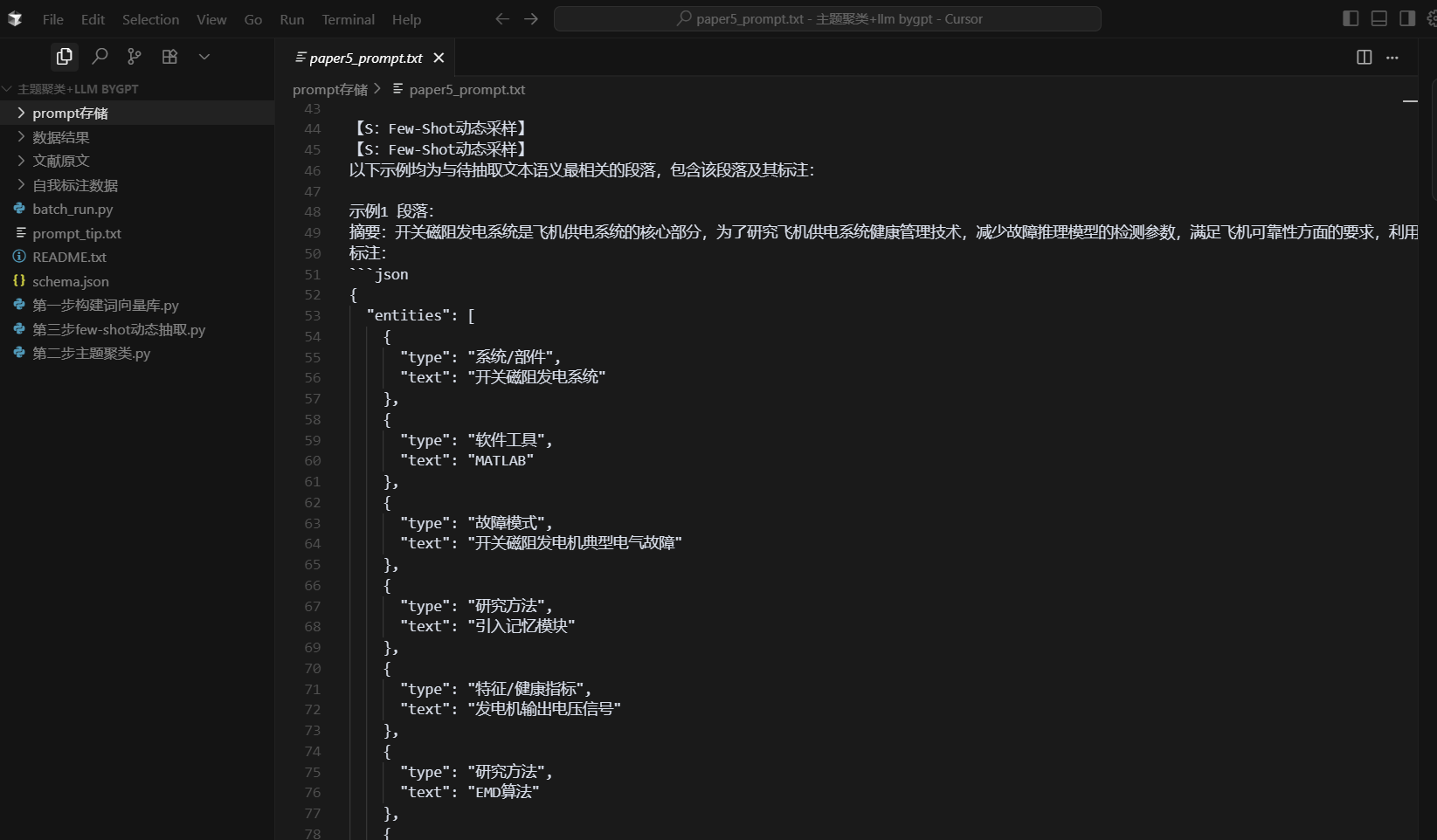
"prefix": ["一种基于", "新提出的"],

"suffix": ["方法", "策略"]

}

]

1. **Prompt中动态fewshot模块更新-主题聚类（针对每进行动态的更新示例）**
   1. 通过语义向量表示与相似度计算，从人工标注数据中选取与当前待抽取文献语义最相关的示例，引导大模型更好理解目标文本语境，提高抽取质量。
   2. **实现步骤**
      1. **构建语义嵌入库**（使用GTE模型）
      2. **主题聚类**（UMAP 算法将高维向量压缩至五维空间，以保留语义邻域结构并去除冗余特征；训练完成后，UMAP 模型被持久化保存，以便后续快速加载与降维。接着，基于降维后的向量执行 HDBSCAN 聚类，自动识别出语义主题簇并标记噪声。）
      3. **构建prompt中动态fewshot**（通过计算论文向量与示例库中各聚类段落向量之间的余弦相似度，确定最匹配的聚类簇。再从该簇中选取若干最相似的示例段落。）



## 进一步工作

1. 衡量效果指标设计
2. 句法依存分析分析已完成，待完成提取典型依存路径
3. 优化效果
   1. Prompt大模型侧：调下模型的上下文记忆token
   2. 规则生成效果：依存句法分析是以“句子”为单位做的，即构建出的依存句法树也是针对一个特定的子句，所以一篇文献的摘要可能被分成若干字句，就有若干依存句法树，如果实体对在一棵树上即在某一子句内，或许可以找到其依存路径，但是对于“跨句”的实体对，可能就找不到。后续如果效果不好可优化为整段。

## 其余工作

1. 软著申请材料已完成（已同步给郭老师，中介那边初步返修了两次）
2. 肖书宇项目那边要重新更新一下数据库数据
   1. 之前的一些数据有的有错误
   2. 增加数据字段：可量化的数据（之前都是字符串类的）