【算法分析赛-作品】

厦门大数据安全开放创新应用大赛·食品安全专题

——（算法分析题）分析报告

参赛编号：SP-001-0112

赛题类型：算法分析题

团队人数：1

1. 赛题名称

食品安全信息抽取模型建立

2. 摘要

“民以食为天，食以安为先”，食品安全就是民生。运用**无监督学习**、**半监督学习**、**预训练微调学习**与**预训练提示学习**等范式构建**主动学习系统**，在现有的各类综合信息库中抽取与食品安全相关的信息，以助力相关部门监管高效精准。

3. 问题需求

参赛者需要基于主办方提供的综合信息数据，对信息数据进行分类，通过模型建立、语义分析等方法筛选出食品安全相关的信息，输出属于食品安全相关的信息编号及信息名称，以助力相关部门监管高效精准。

针对食品及食品安全名词定义如下：

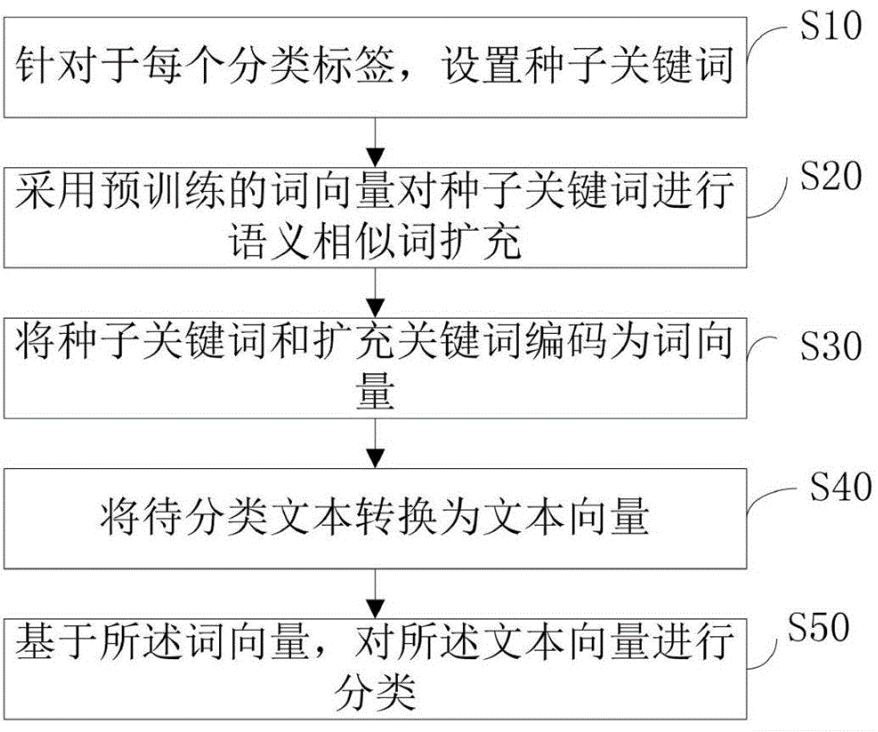
(1) 食品，指各种供人食用或者饮用的成品和原料以及按照传统既是食品又是中药材的物品，但是不包括以治疗为目的的物品。

(2) 食品安全，指食品无毒、无害，符合应当有的营养要求，对人体健康不造成任何急性、亚急性或者慢性危害。食品安全包括食品卫生、食品质量、食品营养等相关内容和食品（食物）种植、养殖、加工、包装、储藏、运输、销售、消费等各个环节。

本赛题是基于**无监督**数据的二元**文本分类**问题，最大**难点**在于没有真实标签的**无监督数据集**如何建立有效的模型。

在工业领域，获取大量标记数据成本往往很大，需要一些无监督或者半监督的方式解决数据标记问题。近年在学术上，无监督方法关注度有所提升，但相对有监督方法，比例还是很小。

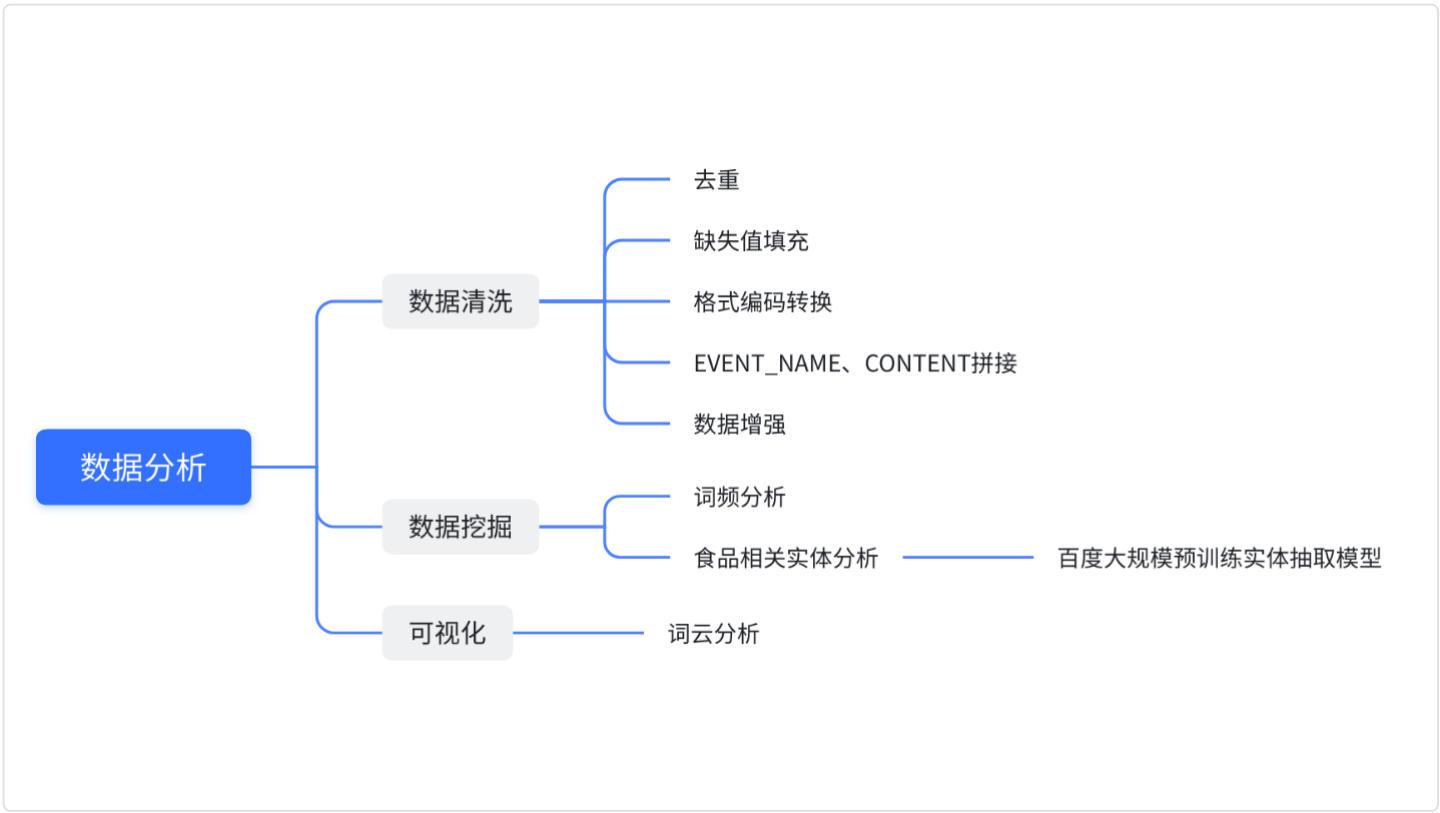
一般方案：该方案的目的在于改善现有技术中所存在的上述不足，提供一种无监督文本分类系统及方法，无需进行人工标注，大大提高文本分类的效率，降低人工成本。

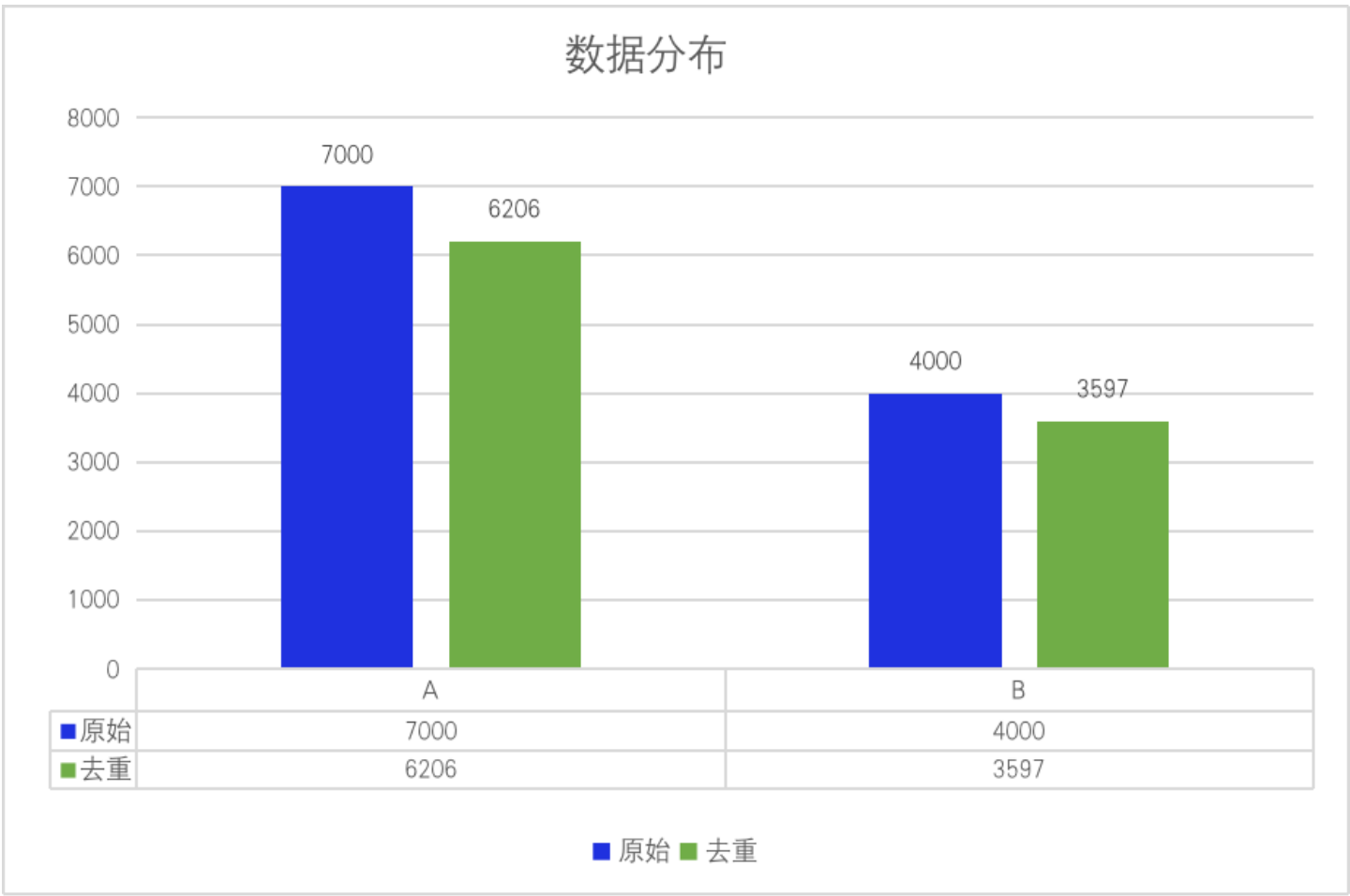


本报告方案：可以借鉴上述方案，将**无监督学习**转化为**半监督学习**，甚至**有监督学习**。

4. 数据应用

- 数据清单
  - 食品安全-算法分析题初赛A榜-综合信息数据.xls，关键字段：
    - EVENT_NAME
    - CONTENT
  - 食品安全-算法分析题初赛B榜-综合信息数据.xls，关键字段：
    - EVENT_NAME
    - CONTENT
- 数据分析细节


数据分析
数据清洗
去重
缺失值填充
格式编码转换
EVENT_NAME、CONTENT拼接
数据增强
数据挖掘
词频分析
食品相关实体分析
百度大规模预训练实体抽取模型
可视化
词云分析


数据分布
8000
7000
6000
5000
4000
3000
2000
1000
0
7000
6206
4000
3597
A
B
原始
7000
4000
去重
6206
3597
原始
去重

机构实体可视化

地区实体可视化

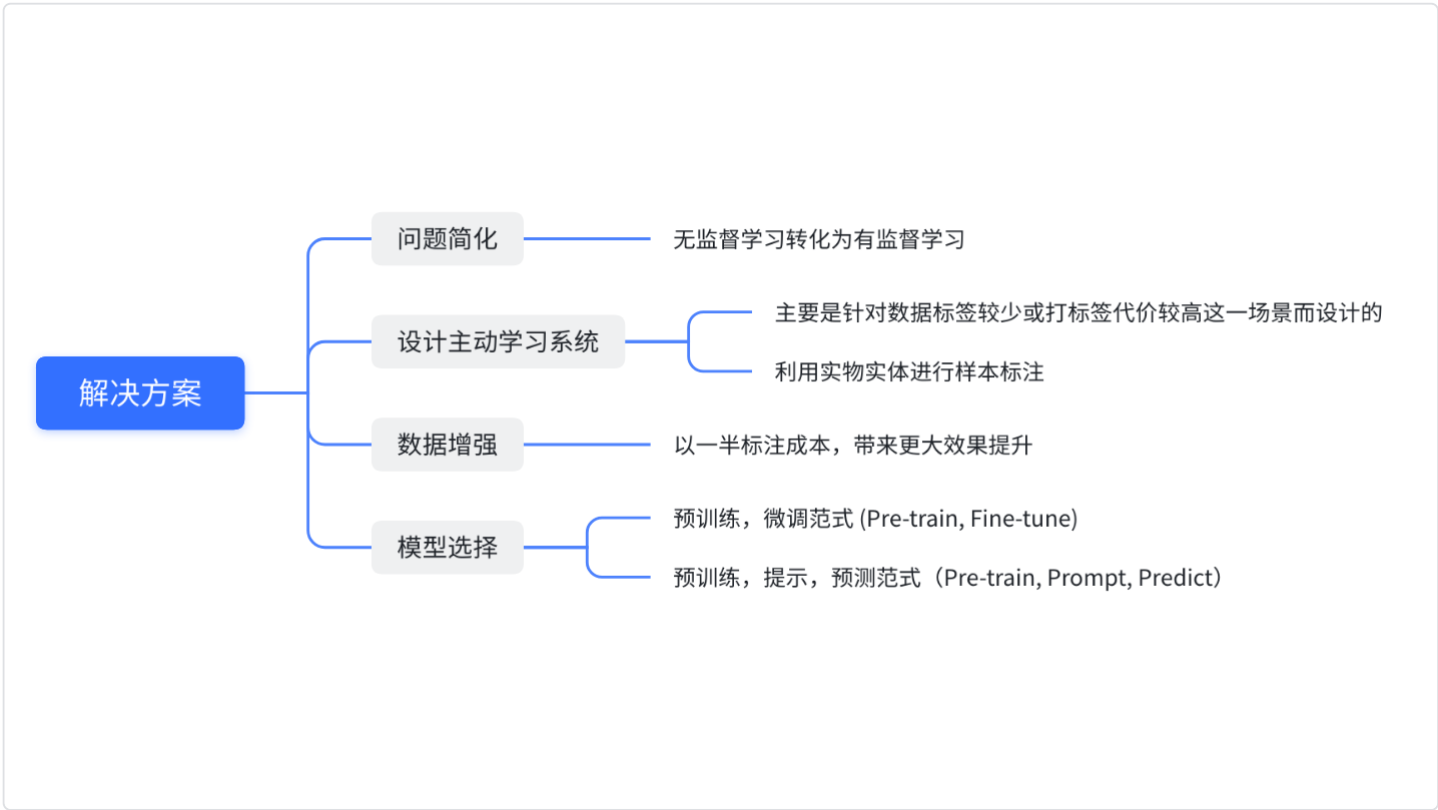
名词实体可视化

食品实体可视化

- 数据安全保障
  - 加强安全意识培训
  - 建立文件保密制度

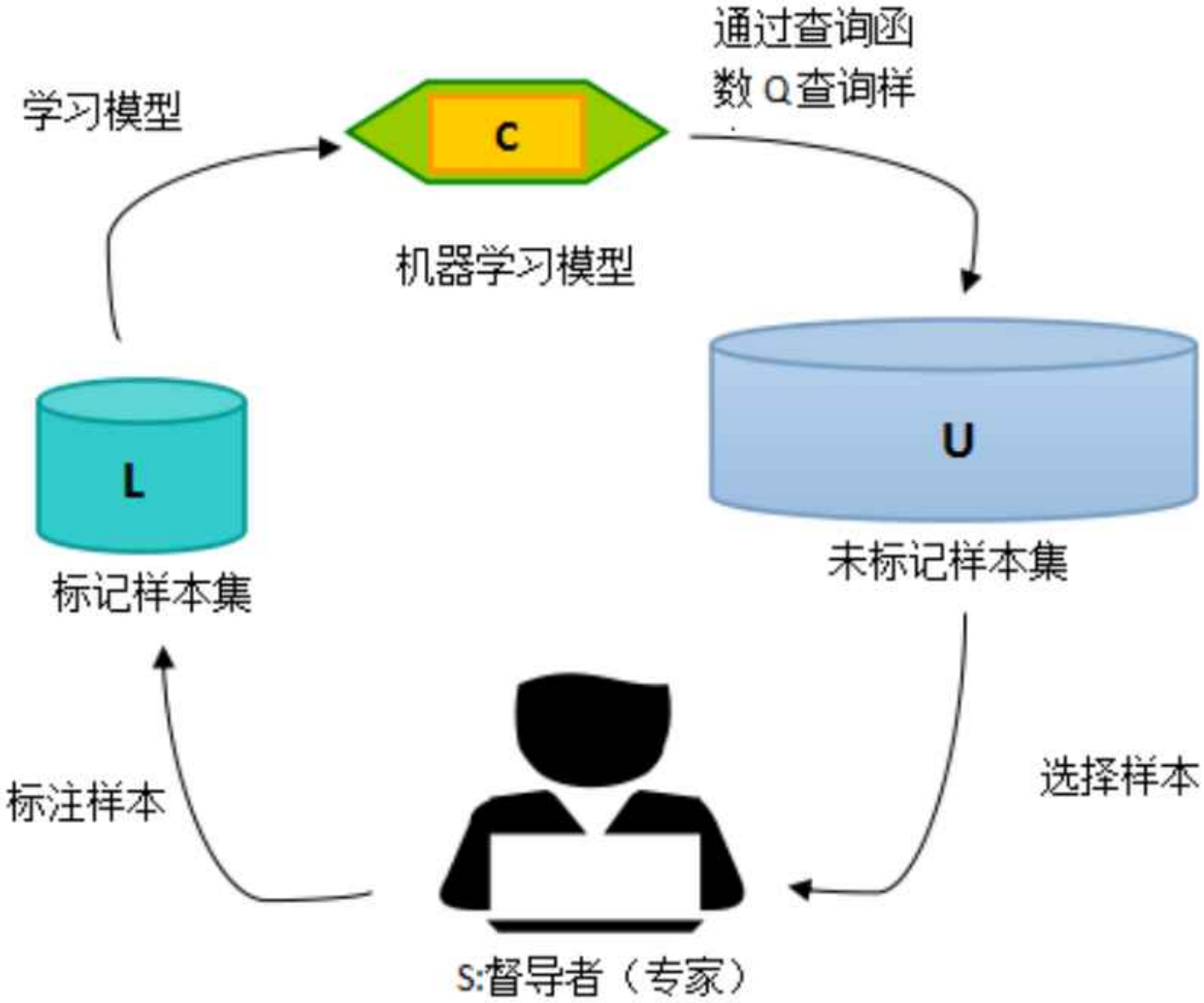
- 弥补系统漏洞
- 密切监管重点岗位的核心数据
- 部署文档安全管理系统等
- 隐私计算，联邦学习

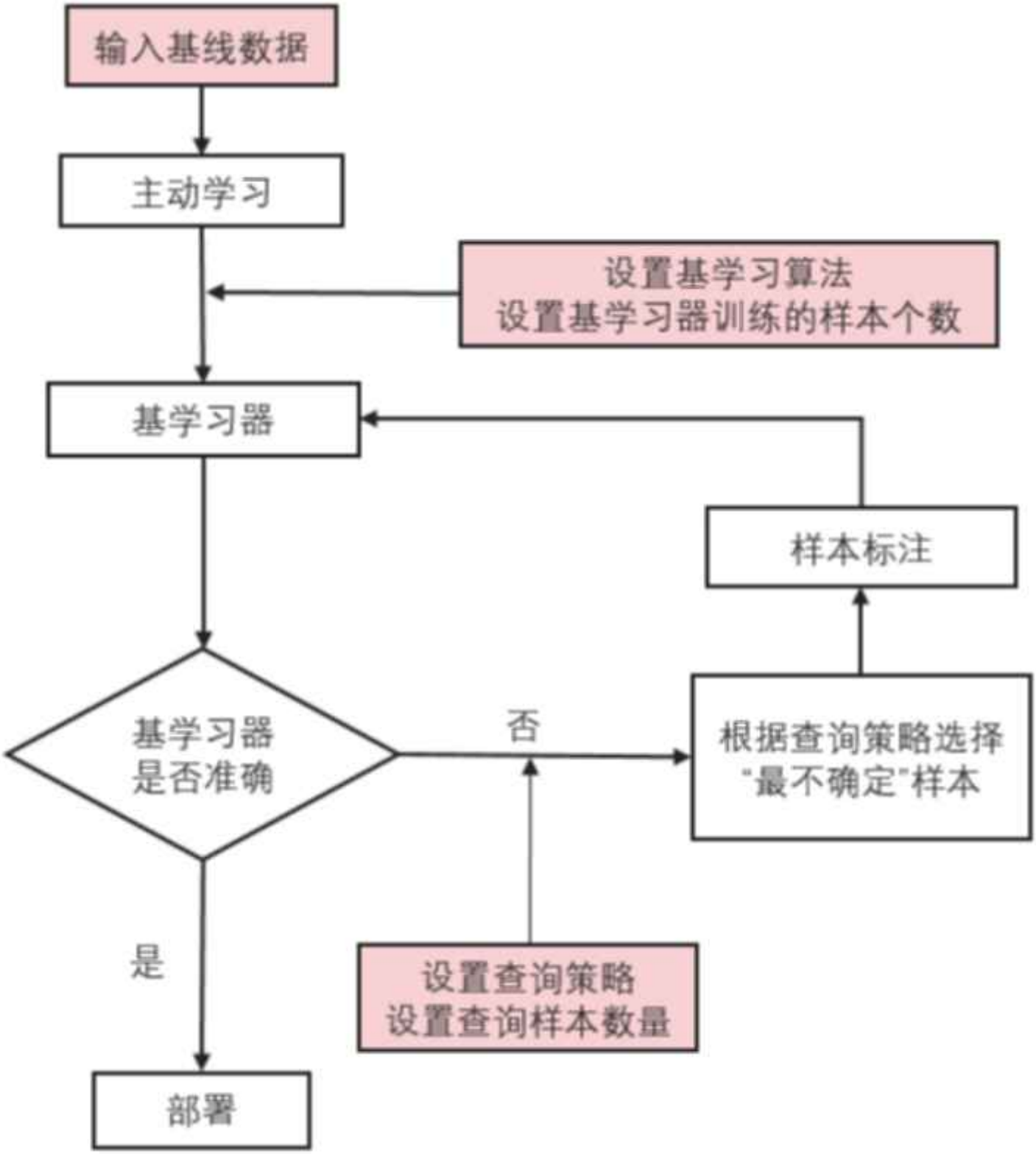
5. 算法分析



设计主动学习系统：

在人类的学习过程中，通常利用已有的经验来学习新的知识，又依靠获得的知识来总结和积累经验，经验与知识不断交互。同样，机器学习模拟人类学习的过程，利用已有的知识训练出模型去获取新的知识，并通过不断积累的信息去修正模型，以得到更加准确有用的新模型。不同于被动学习被动的接受知识，主动学习能够选择性地获取知识


通过查询函
数 Q 查询样
学习模型
C
机器学习模型
U
未标记样本集
L
标记样本集
选择样本
标注样本
S:督导者（专家）



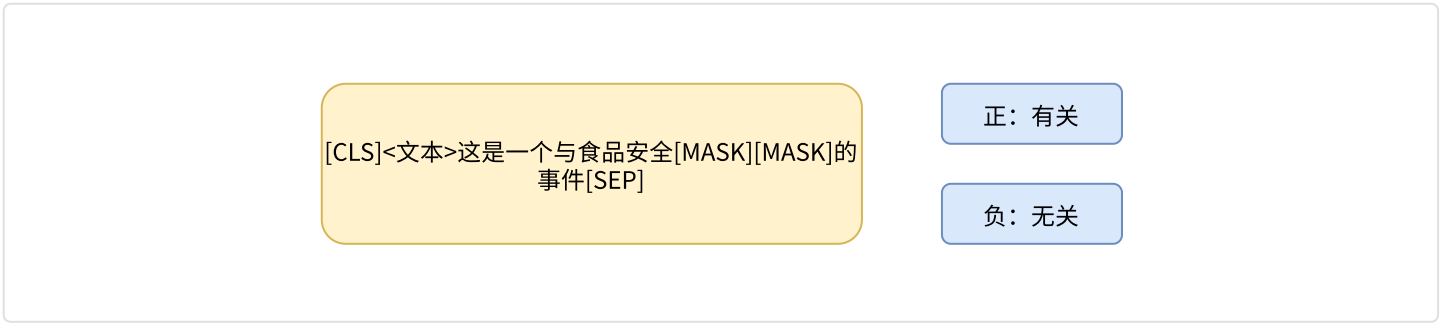
实物实体标注

```
from paddlenlp import Taskflow
ner = Taskflow('ner')
```

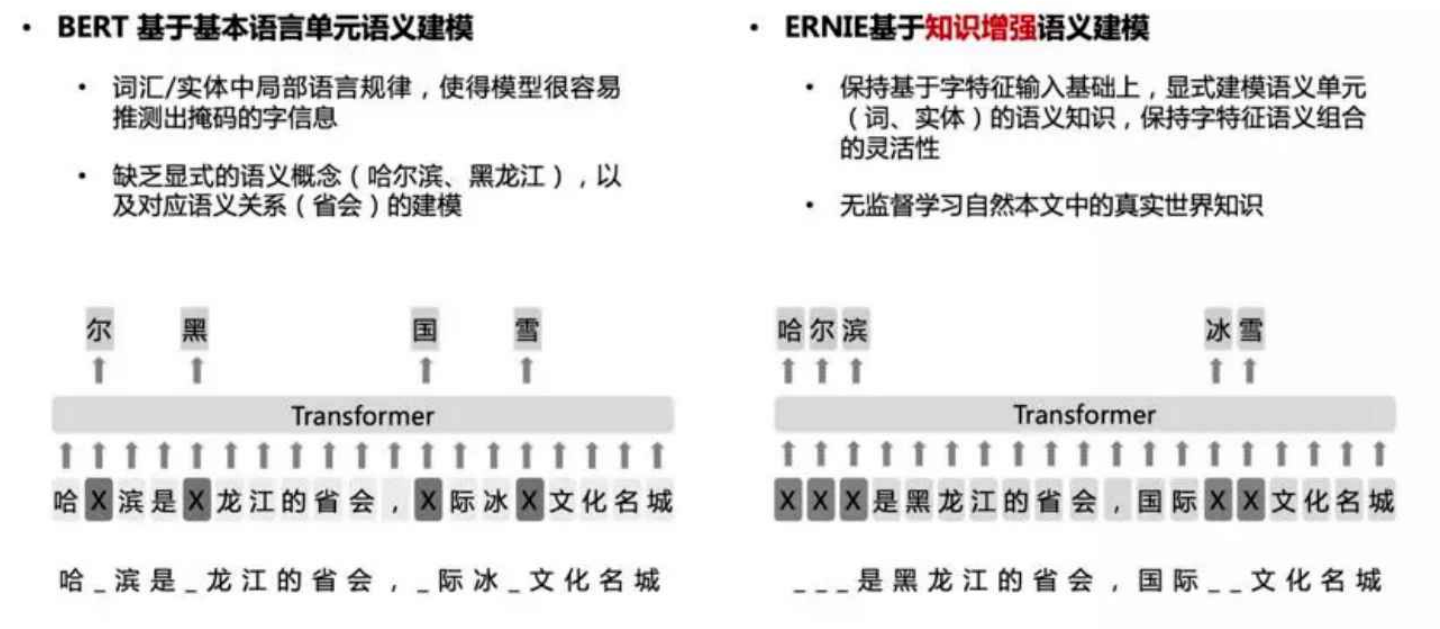
算法设计

预训练，微调范式 (Pre-train, Fine-tune)： 预训练和微调(pre-train and fine-tune) 范式将预训练好的语言模型迁移到具体的下游任务中，并在下游任务上进行微调。这种范式减小了对标记数据集的依赖。但是微调可能会导致语言模型丧失原本的预测能力，这种现象被称为灾难性遗忘(Catastrophic Forgetting)。

预训练，提示，预测范式（Pre-train, Prompt, Predict）： 这种范式通过提示(prompt)来对下游任务进行改造，使得预训练模型可以直接被用于预测下游任务的结果。这种方式可以在不破坏语言模型的同时，最大程度地挖掘语言模型中包含的知识。

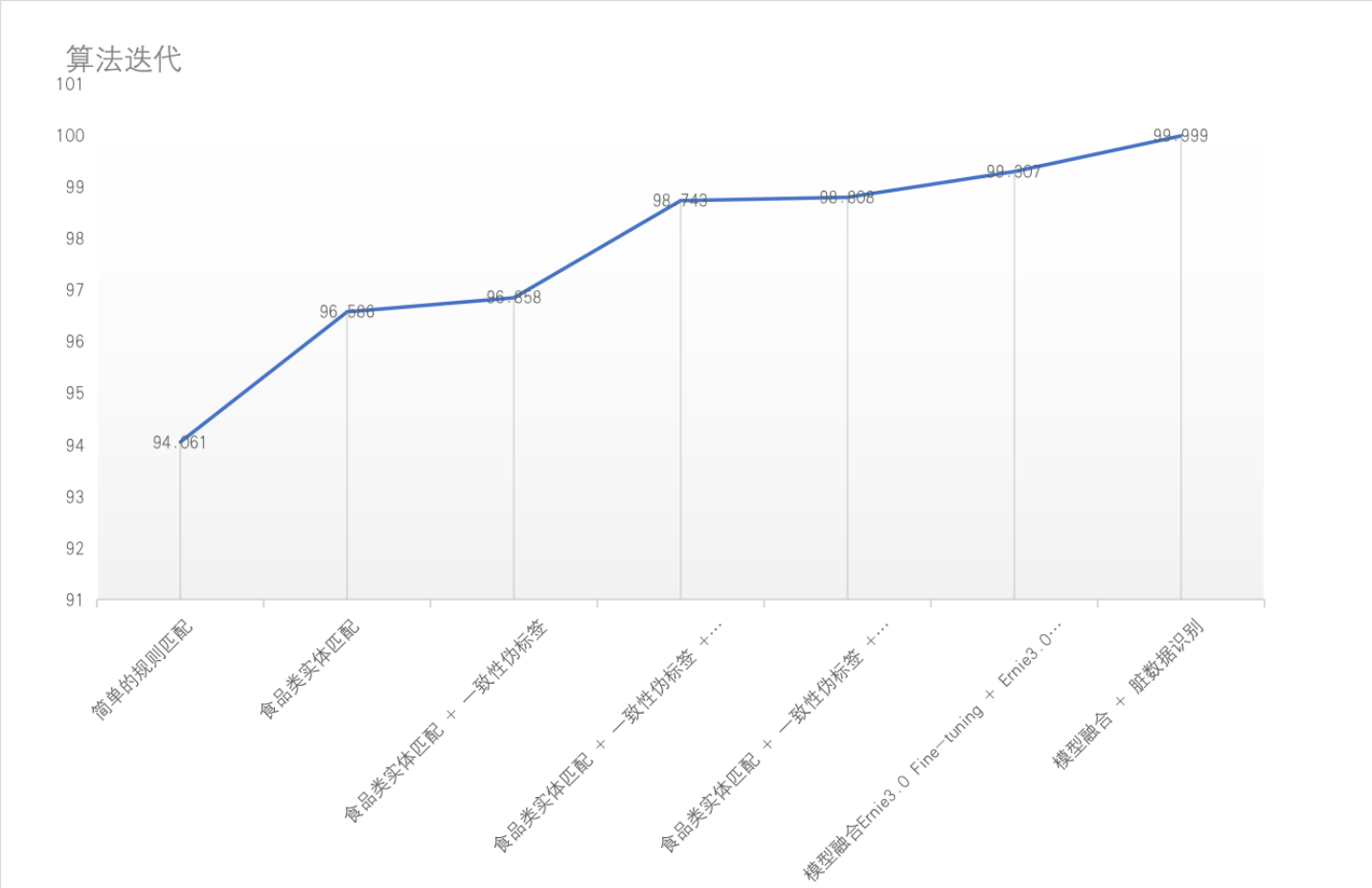


选择更好的预训练模型



6. 结果及结果分析

	A	B
1	算法	得分
2	简单的规则匹配	94.061
3	食品类实体匹配	96.586
4	食品类实体匹配 + 一致性伪标签	96.858
5	食品类实体匹配 + 一致性伪标签 + Ernie3.0 Fine-tuning	98.743
6	食品类实体匹配 + 一致性伪标签 + Ernie3.0 Propmt	98.808
7	模型融合Ernie3.0 Fine-tuning + Ernie3.0 Propmt	99.307
8	模型融合 + 脏数据识别	99.999



> 由于看不到初赛成绩了，所以只有复赛的得分（复赛提交了一次，不是全方案，未得到充分严重）

算法分析题

食品安全信息抽取模型建立

参赛编号：SP-001-0112

提交作品　算法提交记录

序号	提交文件名称	提交时间	得分
1	result.txt	2022-10-24 21:22:33	98.8080
2	result.txt	2022-10-13 18:26:49	0.0000

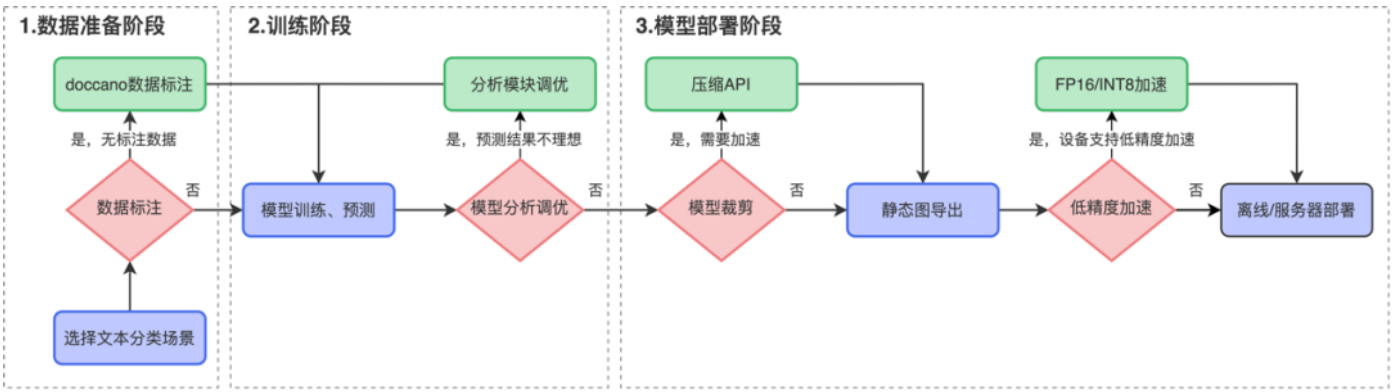
7. 应用成果

- 鲁棒性强：可以提广到新数据集上
- 降低人工成本：以一半标注成本，带来更大效果提升
- 有效识别食品安全分类，助力相关部门监管高效精准

8. 作品价值

- 提出了主动学习系统，提高标注效率，节省10倍人力成本，并带来更大效果提升
- Prompt-learning 优于 Fintune，解释性更强
- 增加**证据识别**及基于证据的预测赛道，这样能更好的识别出食品安全的证据，支撑监管部门更好的判断

9. 实施建议



10. 参考文献（可选）

- Paddle
- Paddlenlp
- TrusAI
- https://github.com/PaddlePaddle