2024年厦门市大数据创新应用大赛

2024 XIAMEN BIG DATA INNOVATION APPLICATION COMPETITION

基于标准地址体系的标准化地址匹配引擎

团队名称：kelvincjr

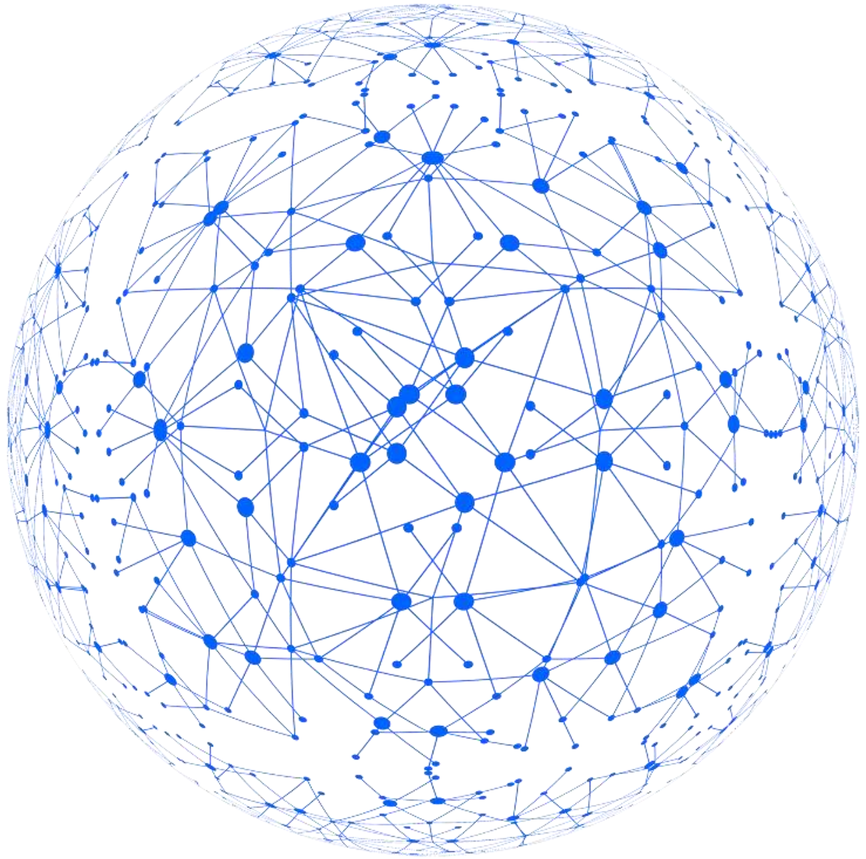
参赛赛道：算法赛

参赛赛题：基于标准地址体系的标准化地址匹配引擎

主办单位：厦门市数据管理局、厦门市公安局、厦门市生态环境局

承办单位：厦门市信息中心

协办单位：厦门大数据有限公司



目录

赛题解读

赛题任务简介

- **赛题背景：**非标准地址转化的意义在于提升社会治理精细化水平、提高社会治理工作效率、推动信息化建设以及服务民生和企业发展等方面。这一工作对于构建智慧城市、提升政府治理能力、促进经济社会发展具有重要意义。当前社会治理地址数据来源多样，如部门共享、基层采集等多种方式相结合，也导致了地址数据质量不一，无法实现与标准地址的关联，对进一步提高治安管理、增强公共服务等方面造成了较大的影响。

- **赛题任务：**主办方提供非标地址测试集的相关数据，选手基于测试集进行模型开发，构建标准地址匹配引擎，实现基于非标地址计算关联得出标准地址。

序号	非标地址	对应标准地址
1	文兴东三里15号前埔南-305室	福建省厦门市思明区文兴东三里15号305室
2	文兴东一里19号(8192)莲前思明前埔南社区\\莲前\\202室	福建省厦门市思明区文兴东一里19号202室

- **评分标准：**

按照标准地址全匹配的方式，选手匹配标准地址的正确率。公式如下：

$$ACC = \frac{\text{选手匹配正确标准地址的数量}}{\text{需匹配的标准地址的总数量}}$$

研究思路

- **任务类型：**本赛题的任务主要是基于批量文本数据的格式转换，主要的研究内容聚焦在如何通过AI人工智能，特别是自然语言处理技术（NLP）实现这种非标文本地址到标准地址的自动转换。

- **解题方案思考：**这种文本自动转换的任务，主要属于NLP的研究范畴，业界主要有两大类解决方案：
  - 把该任务作为一个文本生成的任务来解决，具体可以通过seq2seq，transformer（unilm，bart，T5）和LLM（Llama，chatglm，qwen）等深度神经网络模型来解决。
  - 把问题分成两步：先基于非标地址文本做关键实体信息识别，利用提取到的关键信息，叠加规则和业务处理逻辑，生成最终的标准地址文本。

- **方案评估和选择：**
  - **方案一：**训练数据无需标注，直接使用<非标地址，标准地址>对就可以训练模型，但缺点就是训练和推理效率都比较低，需要配置有较好GPU的环境，同时因为结果是直接生成的，存在错误生成、重复信息、添加额外信息和无中生有的概率较高，很难控制最终结果的准确性。而且生成的效果和模型的大小关系很大。
  - **方案二：**相比而言，方案二需要有标注好的数据来训练模型。但优点就是生成的过程比较可控，可以基于规则和业务策略控制生成过程，基本不会出现重复信息、添加额外信息和无中生有的情况。而且对于关键实体信息抽取，使用基于Bert系列的中小规模的模型已经可以得到比较好的效果，而且训练和推理效率都比较高，生产环境可以直接使用CPU进行推理，成本较低。

结合本赛道赛题的要求：需要有较高的准确率，而且模型规模不能太大（2G以内），同时要基于CPU进行推理，推理时间在1个小时以内。我们最终选择了方案二作为最终解决方案的框架。

市场上地址标准化产品竞品分析

- 调研了一下市场上的地址标准化产品，比较有代表性的是几个电商和地图服务巨头公司提供的云服务产品。这主要是因为构建一个地址标准化引擎需要有海量的数据，而电商、物流和地图服务天生就能够收集和存储海量的地址信息数据，为构建地址标准化引擎提供强大的数据基础。下面简单介绍一下几家地址标准化服务：

➢ 阿里云地址标准化服务：

- ✓ 阿里云的地址标准化服务功能很丰富，提供了包括地址信息抽取、地址结构化、地址纠错、地址补全、地址异常检测、地址相似性判断、多源地址归一等功能，但主要还是针对**电商**和**物流**行业的需求和特点提供相关的功能。

- ✓ 并没有直接提供非标地址转标准地址的服务。我们可以通过使用里面的地址信息抽取功能来实现一定的地址标准化，但通过测试发现，针对赛题的这种变化多样的非标地址，阿里云地址服务的抽取效果一般。

市场上地址标准化产品竞品分析

- 市场上几家提供的地址标准化服务，主要还是针对电商、物流、金融等一些通用领域，这些场景中的非标准地址不会像赛题里面提供的用于户籍、人口、城市房屋治理等领域的地址那样变化多样。
- 要解决赛题中这种特点数据的地址标准化问题，还是需要针对这种海量数据建立基于定制化模型的地址标准化引擎。

➢ 百度地图地址标准化服务：

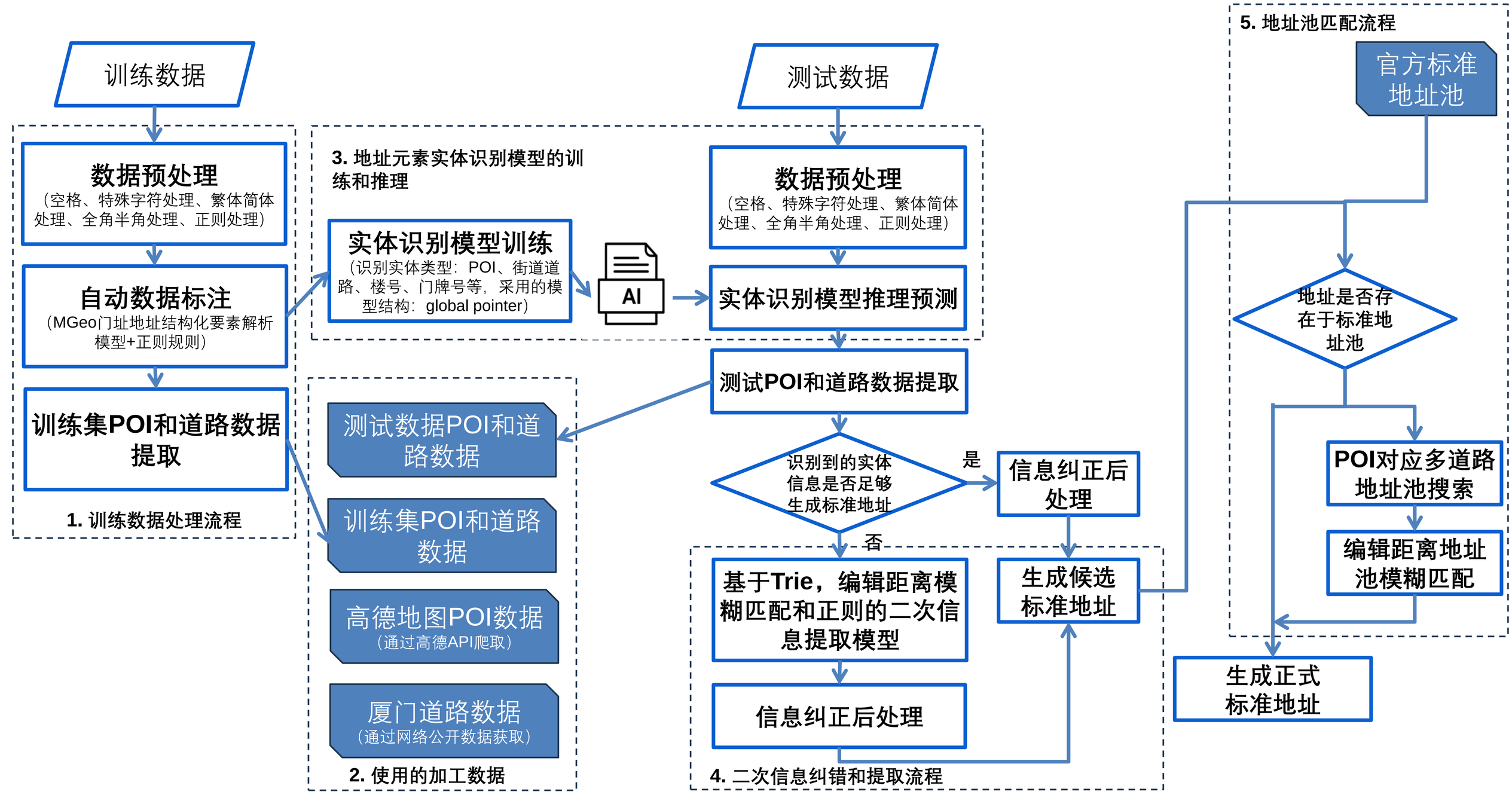
✓ 百度地图的地址标准化服务提供了包括地址结构化解析、地址异常识别和地址归一等功能，同样没有直接提供非标地址转标准地址的服务。其地址结构化解析功能效果比阿里云好，但针对赛题的非标地址，解析的效果同样比较一般。

➢ 京东系的与图地址标准化服务：

✓ 京东系的与图提供了包括电商下单环节的地址识别、地址自动补全，用于物流的地址排重和用于金融风控的地址诊断功能。同样没有直接提供非标地址转标准地址的服务。

建模思路

总体解决方案概览


训练数据
数据预处理
（空格、特殊字符处理、繁体简体处理、全角半角处理、正则处理）
自动数据标注
（MGeo门址地址结构化要素解析模型+正则规则）
训练集POI和道路数据提取
1. 训练数据处理流程
3. 地址元素实体识别模型的训练和推理
实体识别模型训练
（识别实体类型：POI、街道道路、楼号、门牌号等，采用的模型结构：global pointer）
AI
测试数据
数据预处理
（空格、特殊字符处理、繁体简体处理、全角半角处理、正则处理）
实体识别模型推理预测
测试POI和道路数据提取
测试数据POI和道路数据
训练集POI和道路数据
高德地图POI数据
（通过高德API爬取）
厦门道路数据
（通过网络公开数据获取）
2. 使用的加工数据
识别到的实体信息是否足够生成标准地址
是
否
信息纠正后处理
基于Trie，编辑距离模糊匹配和正则的二次信息提取模型
信息纠正后处理
生成候选标准地址
4. 二次信息纠错和提取流程
5. 地址池匹配流程
官方标准地址池
地址是否存在于标准地址池
POI对应多道路地址池搜索
编辑距离地址池模糊匹配
生成正式标准地址

1. 训练数据处理流程 – 数据预处理

■ 数据预处理主要包括特殊字符处理，繁体字处理，全角半角处理等流程

特殊字符处理

{"id": 11, "o_u_address": "益辉花园36号3o2室", "u_address": "益辉花园36号302室",
{"id": 169, "o_u_address": "83洪山柄北区147号102之1室", "u_address": "83洪山柄北区147号102-1室",
{"id": 49, "o_u_address": "42前埔北一 里 160号304", "u_address": "42前埔北一里160号304",

繁体字处理

{"id": 714, "o_u_address": "蓮薇-蓮前東路736之1", "u_address": "莲薇-莲前东路736-1",
{"id": 715, "o_u_address": "萬景//洪蓮北路1001", "u_address": "万景//洪莲北路1001",

全角半角处理

{"id": 632, "o_u_address": "华林绿景花园１７０号１００１室", "u_address": "华林绿景花园170号1001室",
{"id": 644, "o_u_address": "57富山公寓１４６号５０４室", "u_address": "57富山公寓146号504室",

1. 训练数据处理流程 – 训练数据自动标注

■ 训练数据自动标注分成三个主要处理流程：

- 1）街道道路、楼号、门牌号的自动标注，这几类实体直接在训练数据中的标准地址里面，使用正则表达式从训练数据中提取相关实体信息，并在非标文本中进行标注即可；
- 2）POI数据需要采用通用的POI实体识别模型先提取再标注，我们采用阿里开源的MGeo门址地址结构化要素解析模型来初步提取POI进行初始标注，后面再通过步骤三进行规则纠正；
- 3）通过一些规则对1，2步初步标注的数据进行纠错和补全，形成最终的训练标注数据；

赛题提供的训练数据

```
序号,非标地址,对应标准地址
1,68庐山公寓25号704室,福建省厦门市思明区东浦三里25号704室
```

街道道路，楼号，门牌号自动标注

```
def process_train_data(data, road_dict):
    pattern = "^福建省厦门市(.*?)区([^\\d\\-]*)?([\\d\\-]*)?号(([\\w\\-]*)?室)?$"
    p = re.compile(pattern)
```

POI数据自动提取和标注

```
inputs:  68庐山公寓25号704室
2024-09-15 16:57:57,115 - modelscope - WARNING - Model revision not specified, use revision: v1.1.1
2024-09-15 16:57:58,377 - modelscope - WARNING - Model revision not specified, use revision: v1.1.1
2024-09-15 16:57:58,581 - modelscope - INFO - initiate model from /root/.cache/modelscope/hub/iic/mgeo_geographic_elements_tagging_chinese_base
2024-09-15 16:57:58,581 - modelscope - INFO - initiate model from location /root/.cache/modelscope/hub/iic/mgeo_geographic_elements_tagging_chinese_base.
2024-09-15 16:57:58,582 - modelscope - INFO - initialize model from /root/.cache/modelscope/hub/iic/mgeo_geographic_elements_tagging_chinese_base
2024-09-15 16:58:02,157 - modelscope - INFO - cuda is not available, using cpu instead.
2024-09-15 16:58:02,162 - modelscope - WARNING - task token-classification input definition is missing
2024-09-15 16:58:02,438 - modelscope - WARNING - task token-classification output keys are missing
{'output': [{'type': 'poi', 'start': 0, 'end': 2, 'prob': 0.41024846, 'span': '68'}, {'type': 'poi', 'start': 2, 'end': 6, 'prob': 0.37556648, 'span': '庐山公寓'}, {'t
ype': 'houseno', 'start': 6, 'end': 9, 'prob': 0.00042913, 'span': '25号'}]}
```

规则对标注数据进行纠错和补全

```
{"u_address": "68庐山公寓25号704室", "s_address": "福建省厦门市思明区东浦三里25号704室",
"u_label": {"house_no": ["25号", 6, 9], "room_no": ["704室", 9, 13], "poi": ["庐山公寓", 2, 6, "", ""]},
"s_label": ["思明", "东浦三里", "25", "704室"]}
```

2. 训练和推理过程使用到的数据

■ 本次比赛的赛题使用到的数据如下：

1. 初赛训练集（1000 条）

序号	非标地址	对应标准地址
1	文兴东三里15号前埔南-305室	福建省厦门市思明区文兴东三里15号305室
2	文兴东一里19号(8192)莲前思明前埔南社区\\莲前\\202室	福建省厦门市思明区文兴东一里19号202室
3	圣华佗中医理疗（85号合并）	福建省厦门市思明区莲前东路87号
4	洪莲里104号侨福社区|602室	福建省厦门市思明区洪莲里104号602室
5	厦禾社区湖滨南路2号？501	福建省厦门市思明区湖滨南路2号501室
6	东坪山社109-3号	福建省厦门市思明区东坪山社109-3号

2. 初赛测试集（2000 条）

id	N_standard_address
1	东浦路236号(6926)莲前龙山桥社区\\莲前\\101室
2	前埔南小区店上东里95号405
3	莲前西路348-4号\3506\莲前莲怡社区|莲前|102室
4	西林东里119号金鸡亭社区|206室
5	洪莲路13号(8862)莲前莲丰社区\\莲前\\902室
6	洪文六里30号莲翔社区601室

3. 决赛测试集（3800 条）

字段与初赛测试集一致

4. 决赛标准地址池（20 万条）

标准地址池
福建省厦门市思明区洪文泥窟社西片区6-3号408室
福建省厦门市思明区岭兜北片区63号101室

5. 公开的厦门思明区街道数据

```
福建省:
    厦门市:
        思明区:
            361001:
                - 斗西路156-158(双号)
                - 北门外街
                - 人和路
                - 故宫路71-99(单号)
                - 大埕头巷
                - 打铁街
                - 开禾路
                - 后岸巷
                - 泰山路
                - 水仙路
                - 第七市场
                - 水流横巷
                - 南荞巷
                - 通奉第巷
                - 西堤南里
                - 厦禾路407-551(单号)
                - 厦禾路395-397(单号)
                - 厦禾路11-393(单号)
                - 定安路
                - 角滨路(1号)
                - 道平路
                - 武当分镇横巷
                - 惠通巷
                - 厦禾路399-401(单号)
                - 豆仔尾路2-168(双号)
```

6. 公开的基于高德地图的厦门 POI 数据

数据通过高德地图开放 API 获取，包括 POI 名称，POI 地址和 POI 类别等字段。

3. 地址元素实体识别模型的训练和推理

■ 本次比赛使用到的地址要素实体识别模型采用了基于Bert的GlobalPointer深度神经网络结构。

- 经典实体识别任务一般采用Bert + CRF的模型结构，但在处理效率和对嵌套实体处理存在不足；
- GlobalPointer的设计，它利用全局归一化的思路来进行命名实体识别（NER），可以无差别地识别嵌套实体和非嵌套实体，在非嵌套（Flat NER）的情形下它能取得媲美CRF的效果，而在嵌套（Nested NER）情形它也有不错的效果。还有，在理论上，GlobalPointer的设计思想就比CRF更合理，处理效率更高。

	北	京	大	学	在	深	圳	有	分	校	吗	？
北	0	0	0	0	0	0	0	0	0	0	0	0
京		0	0	0	0	0	0	0	0	0	0	0
大			0	0	0	0	0	0	0	0	0	0
学				0	0	0	0	0	0	0	0	0
在					0	0	0	0	0	0	0	0
深						0	0	0	0	0	0	0
圳							0	0	0	0	0	0
有								0	0	0	0	0
分									0	0	0	0
校										0	0	0
吗											0	0
？												0

Head-1：人名

	北	京	大	学	在	深	圳	有	分	校	吗	？
北	0	1	0	0	0	0	0	0	0	0	0	0
京		0	0	0	0	0	0	0	0	0	0	0
大			0	0	0	0	0	0	0	0	0	0
学				0	0	0	0	0	0	0	0	0
在					0	0	0	0	0	0	0	0
深						0	1	0	0	0	0	0
圳							0	0	0	0	0	0
有								0	0	0	0	0
分									0	0	0	0
校										0	0	0
吗											0	0
？												0

Head-2：地名

	北	京	大	学	在	深	圳	有	分	校	吗	？
北	0	0	0	1	0	0	0	0	0	0	0	0
京		0	0	0	0	0	0	0	0	0	0	0
大			0	0	0	0	0	0	0	0	0	0
学				0	0	0	0	0	0	0	0	0
在					0	0	0	0	0	0	0	0
深						0	0	0	0	0	0	0
圳							0	0	0	0	0	0
有								0	0	0	0	0
分									0	0	0	0
校										0	0	0
吗											0	0
？												0

Head-3：机构名

3.地址元素实体识别模型的训练和推理

序号,非标地址,对应标准地址
1,68庐山公寓25号704室,福建省厦门市思明区东浦三里25号704室

```
{
    "u_address": "68庐山公寓25号704室",
    "s_address": "福建省厦门市思明区东浦三里25号704室",
    "u_label": {
        "house_no": [
            "25号",
            6,
            9
        ],
        "room_no": [
            "704室",
            9,
            13
        ],
        "poi": [
            "庐山公寓",
            2,
            6,
            "",
            ""
        ]
    },
    "s_label": [
        "思明",
        "东浦三里",
        "25",
        "704室"
    ]
}
```

地址要素实体识别模型训练

id,N_standard_address
1,洪文六里37号502室

地址要素实体识别模型推理

Globalpointer

	6	8	庐	山	公	寓	2	5	号	7	0	4	室
6	0	0	0	0	0	0	0	0	0	0	0	0	0
8		0	0	0	0	0	0	0	0	0	0	0	0
庐			0	0	0	1	0	0	0	0	0	0	0
山				0	0	0	0	0	0	0	0	0	0
公					0	0	0	0	0	0	0	0	0
寓						0	0	0	0	0	0	0	0
2							0	0	1	0	0	0	0
5								0	0	0	0	0	0
号									0	0	0	0	0
7										0	0	0	1
0											0	0	0
4												0	0
室													0

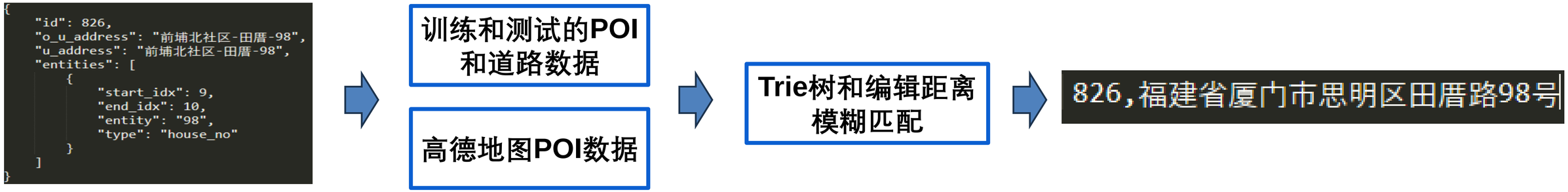
预训练模型：Roberta-Base

```
{
    "id": 1,
    "o_u_address": "洪文六里37号502室",
    "u_address": "洪文六里37号502室",
    "entities": [
        {
            "start_idx": 0,
            "end_idx": 3,
            "entity": "洪文六里",
            "type": "rd_st"
        },
        {
            "start_idx": 4,
            "end_idx": 6,
            "entity": "37号",
            "type": "house_no"
        },
        {
            "start_idx": 7,
            "end_idx": 10,
            "entity": "502室",
            "type": "room_no"
        }
    ]
}
```

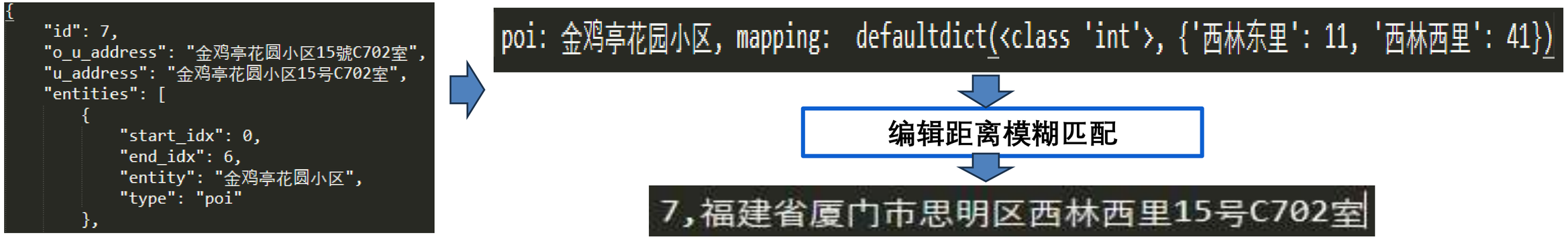
4. 二次信息纠错和提取模型

■ 二次信息纠错和提取模型主要完成POI、街道信息的二次识别（NLP实体链接）和实体信息纠错等任务：

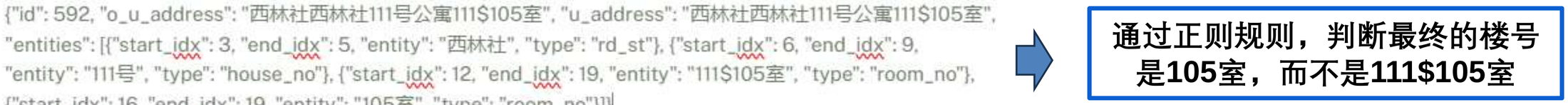
➢ 对于地址元素实体识别模型没有能识别到的街道和POI，通过Trie树和编辑距离模糊匹配等方法进行二次查找和匹配；



➢ 对于测试数据中只有POI而没有道路信息的记录，通过使用训练和测试数据中POI和道路的映射表、高德POI数据，匹配到相应的道路信息；

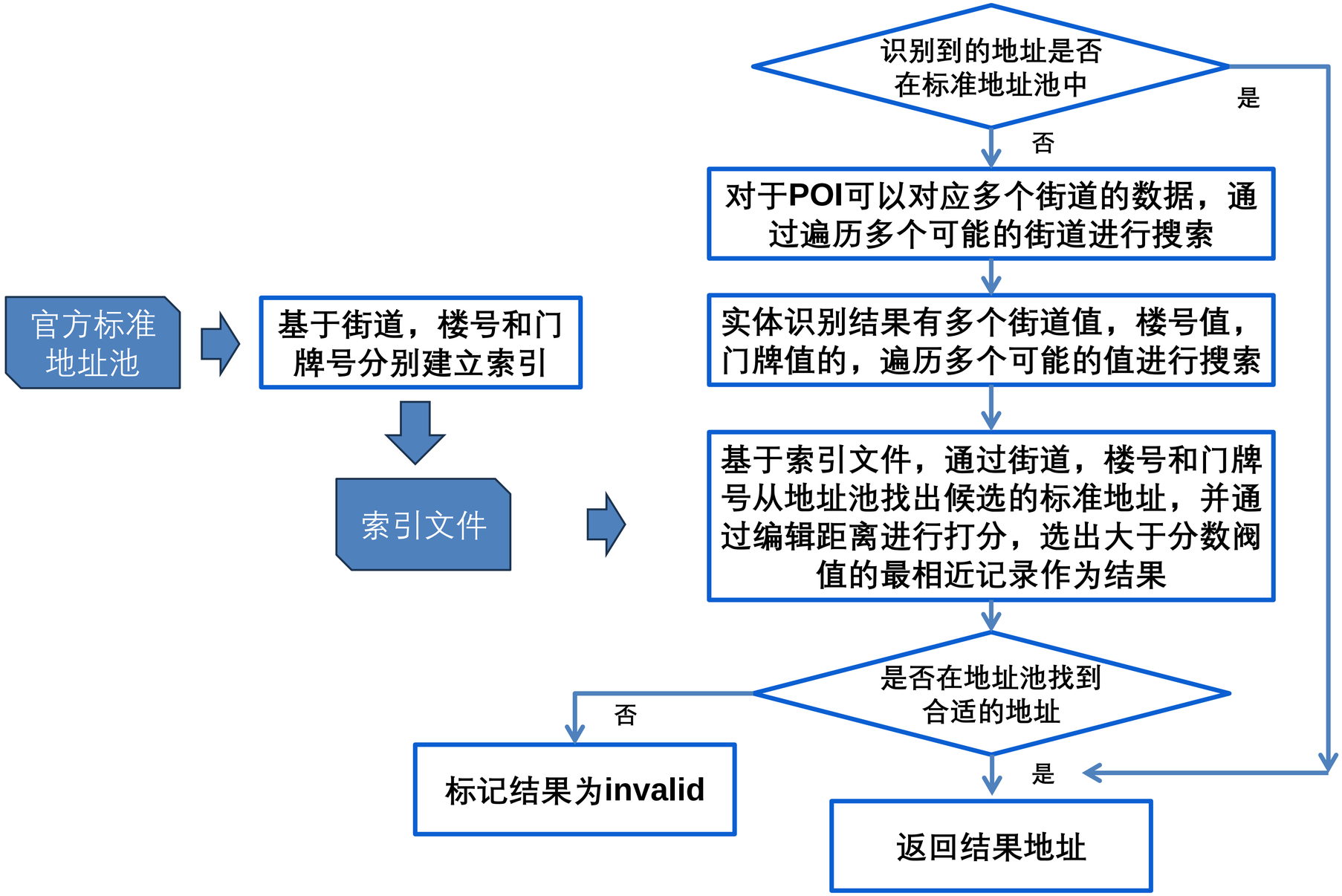


➢ 对于地址元素实体识别模型识别有误的信息，通过正则表达式等方式进行纠正；



5. 地址池匹配流程

■ 首先判断之前识别到的地址是否在标准地址池中，如果不在，则执行地址池搜索任务：



算法效率和参数量

- 地址元素识别模型基于Roberta Base，12层transformer，参数量为102M，之前也用过RBT3，参数量只有38M 。模型处理的时间如下，采用基于Roberta Base的模型处理效率在180-190秒之间（决赛3800条记录）。

Kelvin9903	model_0917.zip	2024-09-17 12:20	1	191.12224507331848	下载 查看
Kelvin9903	model_0916.zip	2024-09-16 14:54	2	186.58047032356262	下载 查看
Kelvin9903	model_0915_...	2024-09-15 20:00	--	185.32470321655273	下载 删除 查看
Kelvin9903	model_0915.zip	2024-09-15 19:05	--	185.445782661438	下载 删除 查看
Kelvin9903	model_0914.zip	2024-09-14 15:51	3	187.73184895515442	下载 查看
Kelvin9903	model_0913.zip	2024-09-13 11:02	4	186.02802777290344	下载 查看
Kelvin9903	model_0912.zip	2024-09-12 11:05	5	182.64008545875	下载 查看
Kelvin9903	model_0911.zip	2024-09-11 10:58	6	80.6223497390747	下载 查看
Kelvin9903	model_0910.zip	2024-09-10 15:47	7	79.5391047000885	下载 查看

基于Roberta-Base

基于RBT3

算法效果

- 截至9.18日，目前决赛成绩排名第一（团队：kelvincjr）。决赛过程中一些重要优化策略包括：标准地址池匹配策略优化，地址元素识别模型基座从RBT3换成Roberta-Base、二次搜索策略优化和数据信息的完善。

排名	团队名称	最优成绩	提交次数	最优成绩提交时间
1	kelvincjr	0.977105	9	2024-09-17 12:20
2	鹭漫漫	0.974474	14	2024-09-18 15:09
3	鹭芊芊	0.971053	13	2024-09-18 03:50
4 ▲11	飞飞加油呀	0.959737	3	2024-09-17 23:44
5 ▼1	小白要努力	0.955263	8	2024-09-17 08:13
6 ▼1	LLMs	0.942632	8	2024-09-18 11:47
7 ▼1	mlf	0.938421	13	2024-09-18 01:05
8 ▼1	长沙小飞侠	0.921316	5	2024-09-17 23:29
9 ▼1	一个好记的名字	0.915526	3	2024-09-17 20:55

测评成绩排行榜（截至9.18）

文件名称	提交时间	团队内排名	运行时长	成绩
model_0917.zip	2024-09-17 12:20	1	191.12224507331848	0.977105
model_0916.zip	2024-09-16 14:54	2	186.58047032356262	0.968947
model_0915_...	2024-09-15 20:00	--	185.32470321655273	--
model_0915.zip	2024-09-15 19:05	--	185.445782661438	--
model_0914.zip	2024-09-14 15:51	3	187.73184895515442	0.968684
model_0913.zip	2024-09-13 11:02	4	186.02802777290344	0.967895
model_0912.zip	2024-09-12 11:05	5	182.6400854587555	0.966842
model_0911.zip	2024-09-11 10:58	6	80.6223497390747	0.949474
model_0910.zip	2024-09-10 15:47	7	79.5391047000885	0.947105
model_0909.zip	2024-09-09 14:45	8	3437.4245343208313	0.937105
model_0907.zip	2024-09-08 10:11	9	3424.903464317322	0.919737

提交成绩记录

创新探索

创新点总结

高效AI系统解决问题（深度模型+工程化模型+规则）

- **清晰完善的AI系统架构：**解决问题导向，分别打造基于Bert的核心地址元素识别模块，基于Trie树和编辑距离二次识别模块和基于分级索引的地址池搜索模块。并通过清晰的AI系统架构设计，把多个模块有机整合，形成一个完善的AI系统。
- **高效的处理效率：**基于效率和准确性考虑，没有采用基于生成式大模型的方案，整个系统处理性能高，并发性能好。
- **准确性高：**最终的系统在训练数据，初赛测试数据和决赛测试数据都取得了接近98%的准确率。

高质量的数据体系（数据为王）

- **赛事数据的充分加工利用：**充分利用赛事提供的训练数据，测试数据和决赛地址池数据，基于此加工出了：1）地址元素识别模型的训练实体标注数据；2）训练和测试的街道数据；3）训练和测试的POI数据；
- **公开数据的有效补充：**使用一些公开的数据作为有效补充，增强识别效果，包括：公开的厦门思明区街道数据，基于高德地图API搜集的厦门POI数据。
- **高效的数据搜索和匹配：**通过建立分级索引、Trie树等数据结构，使用编辑距离匹配的方法，高效匹配街道、POI数据和标准地址池数据。

整体流程自动化，方案可复用（节约人力，降本增效）

- **训练标注数据自动生成：**通过自动化的程序处理生成高质量的地址元素识别模型训练实体标注数据，只需要少量的人工复核，大幅减少了人工标注量，节约人力；
- **提供工具自动化搜集公开补充数据：**提供了自动化小工具，从公网上搜集一些公开的补充数据，如街道和POI数据，无需人工参与，节约人力；
- **方案的可复用性：**方案的体系架构清晰，可复用于其它地址转换、搜索和匹配场景。也适用于其它基于实体识别和匹配的应用场景。

跨场景应用的探索和规划

- 地址作为定位到某个位置的关键信息，广泛应用于我们生活工作的方方面面。而地址标准化在地址信息准确定位、地址信息共享和精细化管理等方面发挥着重要作用。我们提出的地址标准化转换方案可复用性强，除了赛题描述的智慧城市户籍、人口、房屋精细化管理场景外，在政府管理的其它方面，还有其它行业都有很多潜在的高频使用场景：

  - **政府管理：**地址信息在政府管理方面主要存在地址使用标准不统一、地址信息无法准确定位和多部门信息共享的痛点。这造成的典型问题包括：
    - ✓ 公安部门在接到群众报警电话时，往往由于地址不详，没有精准的地址信息，无法精准定位，造成警察多跑路、跑错路；
    - ✓ 在商业主体注册登记方面，由于缺乏地址认证，存在大量虚假注册地址，给事后监管带来重重困难；
    - ✓ 在户籍、人口管理方面，同样由于缺乏完整的标准地址库认证，因录入地址不规范、不统一的情况，以房管人、人房纳管难以落地；

    赛题主要聚焦在户籍、人口、智慧房屋管理的应用，但如上述例子，地址标准化在公安、工商等多个其它部门都有众多应用场景。如赛题描述，地址标准化工作包括标准地址库建设、非标-标准地址转换两大块。如上述场景，在公安接到举报的时候，通过我们的程序能实时把群众提供的地址快速转换成标准地址库中的地址，就能够极大提升公安部门的处置效率，在紧急情况下快速赶到现场，解决问题。

  - **电商、物流行业：**现代物流行业飞速发展，但是随着物流行业的进步，该行业也面临着许许多多的挑战，例如常见的物品送错，产生的原因有可能是：1. 原始地址在输入时就有误。2. 地址本身存在一地多名的情况，导致地址目标不明确。3. 行政区划信息变更，未及时更新。地址标准化功能就能有针对性解决这一问题。

跨场景应用的探索和规划

- **移动互联网：**移动APP时代对于地址的需求不降反增，比如平时最常见的APP挪车，外卖APP地址输入，导航地址查询，甚至智能汽车地址寻路等等对于地址的精度要求很高，地址标准化在这些场景也可以直接通过API调用的形式直接集成在APP中，和移动APP融合一体为上层的应用提供各种各样的地址相关数据支持。地址标准化可解决：1.对原始业务地址进行纠错、补全、结构化等标准化处理，返回最新的标准地址；2.将多源业务系统的地址进行标准化处理，并将多库合一，在此基础上实现地址匹配功能；3.在标准地址库的基础上提供地址搜索、地址联想、POI预测等功能。

- **新零售场景：**多源业务地址匹配，建立标准数据库，分析城市POI画像，赋能零售企业的营销推广。1.将用户地址纠错、补全、结构化成标准地址；2.实现标准地址与业务地址1:N的关联关系建立；3.评估产品社区渗透率，指导营销、完善售后；4.门店选址推荐、线下营销推荐、区域人群画像分析等。

- **能源精细化管理场景：**建立标准的业务地址数据库，结合地址围栏、地址坐标等实现网格化精细管理，多维数据上图，便于整体调控。1.将业务地址标准化清洗并建立标准地址数据库；2.地址围栏实现用户的网格化管理；3.地址经纬度转换查询；4.多源业务地址关联匹配；5.各类热力图分析。

- **金融风控场景：**在我从事的金融领域中，个人和企业开卡、开户、登记、填写信息的过程中，地址信息存在行政区划缺、漏、错、假的现象，对于地址不全、错误、虚假等非标准化数据与风险，智能地址解析，可以快速识别此类型数据风险，有效提升数据质量，降低业务风险。1.实现21层地址标签，将用户的地址数据标准化处理；2.针对风险地址、虚假地址进行预警；3.智能化搜索与内容返填；4.基于知识图谱实现地址画像，沉淀地址标签属性体系；5.实现地理位置经纬度与文本之间的互相转换；6.提供地址围栏等服务。

团队介绍

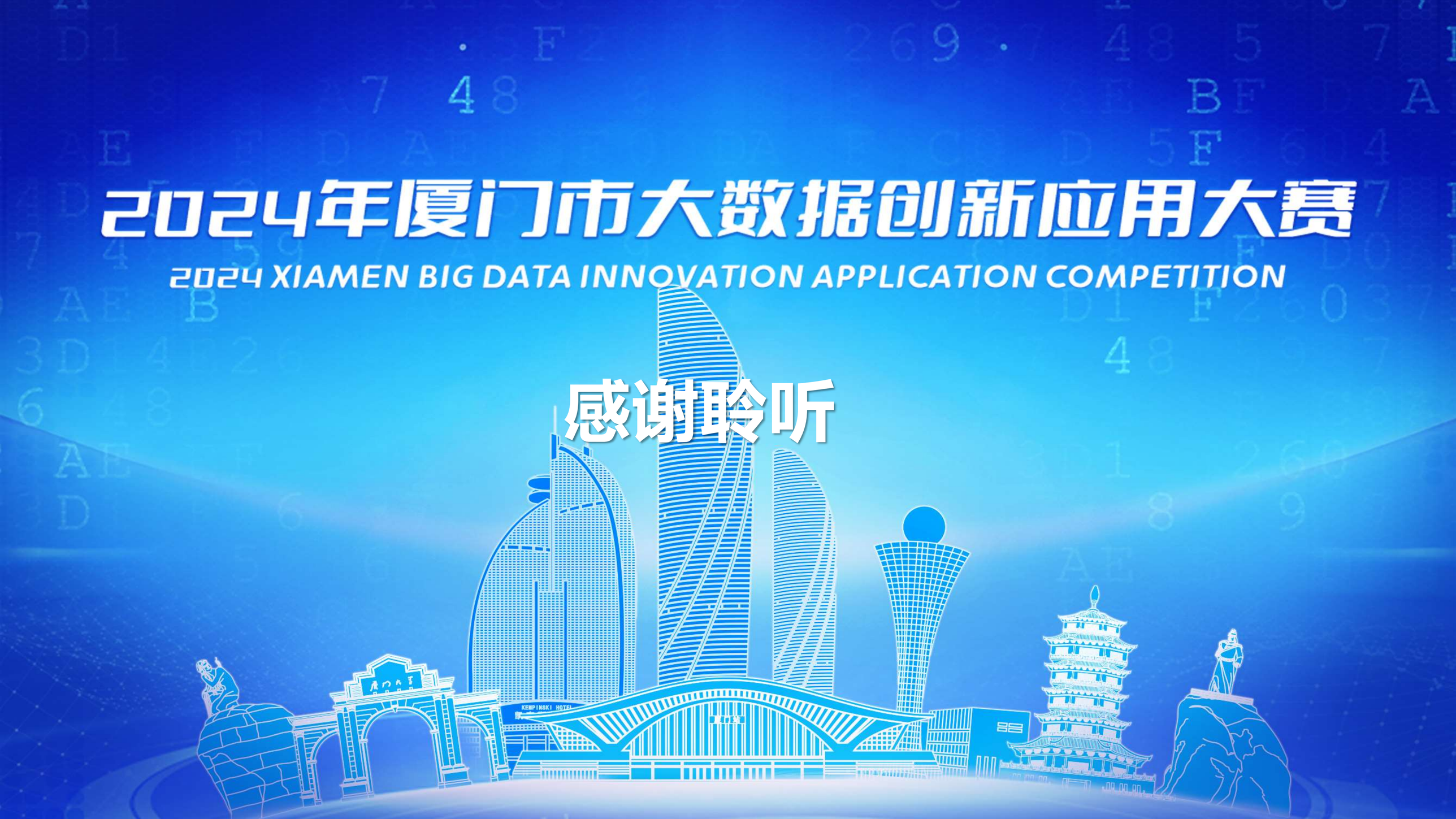
四、团队介绍

蔡嘉荣（队长）- kelvincjr

- 目前任职某银行信用卡中心，数据科学专家，负责数据科学团队的管理。
- 多年的软件开发、架构经验，近几年专注于数据科学、机器学习、NLP和大模型等领域的研究和开发。
- 近年多次参加数据挖掘和算法类比赛，获得一定的名次和奖项。

2024年厦门市大数据创新应用大赛

2024 XIAMEN BIG DATA INNOVATION APPLICATION COMPETITION



感谢聆听