

中文通讯地址解析

反欺诈-数据算法 @王国印

Outline

项目背景

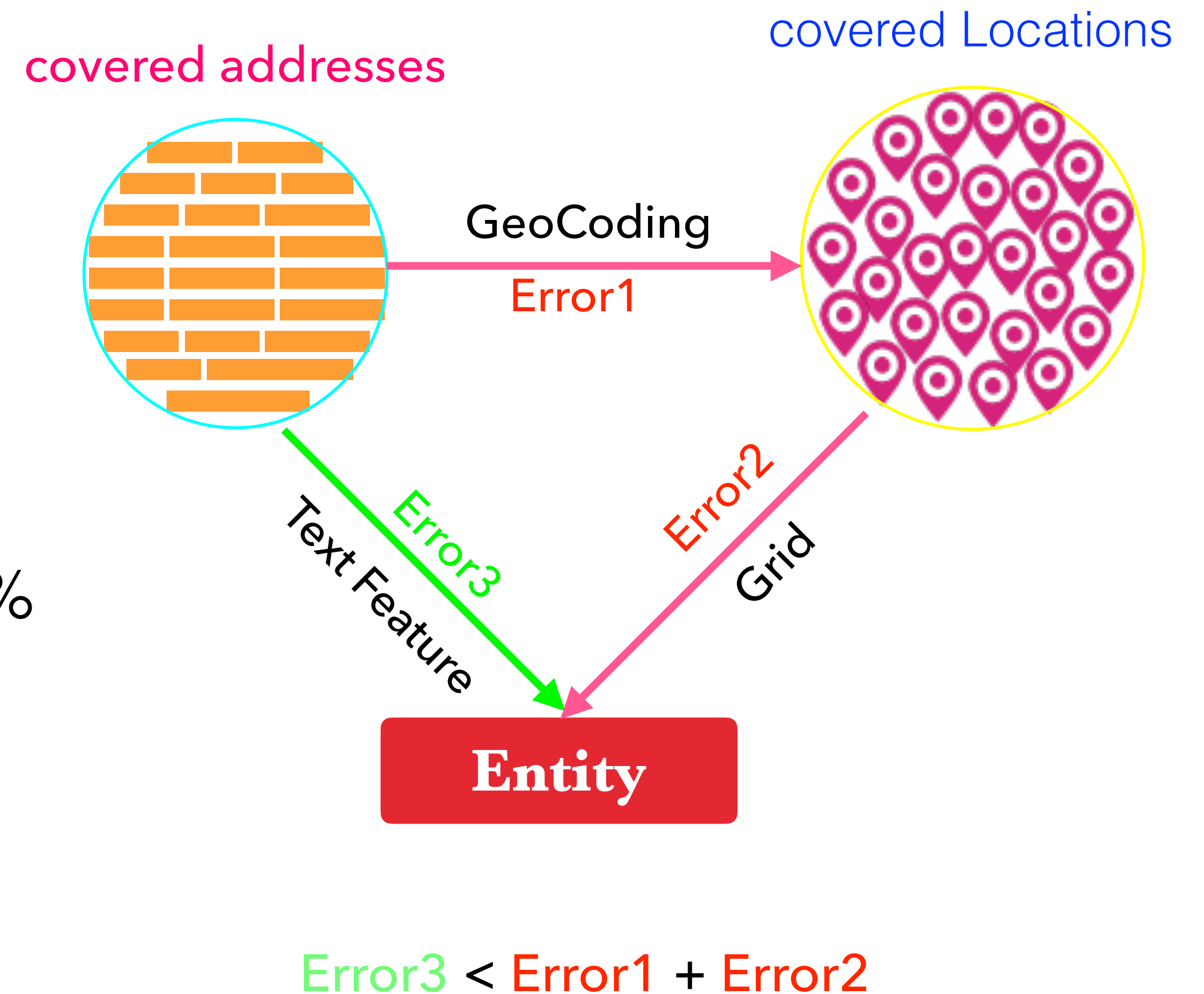
Demo

地址解析keypoint

项目效果

项目背景

- ▶ 用户收寄地址是物流流转的核心要素
- ▶ 业务流转沉淀了大量物流订单历史数据
 - ▶ 网点/快递员派送&揽收数据
 - ▶ 自提柜/驿站收件数据
 - ▶ 仓库发货数据
 - ▶
- ▶ 地址转经纬度误差较大——100米以内不到90%
- ▶ 高德分词不适合物流场景
 - ▶ 性能较差（单机65KB/s, TPS:497, RT 6ms）
 - ▶ 实现代码不透明
 - ▶ 分词粒度单一，比较适合导航场景
 - ▶ 跨团队合作因素



Demo

余杭仓前良睦路999号圣殿公交站对面万利大厦后面乐佳国际2号楼小邮局



prov=浙江省 city=杭州市 district=余杭区 town=仓前街道 road=良睦路
roadNo=999号 poi=乐佳国际 houseNo=2号楼 person=小邮局
otherPoi=[圣殿公交站, 万里大厦] addrInfos=[对面, 后面]
wrdList=[圣殿公交站, 对面, 万利大厦, 后面, 乐佳国际, 2号楼, 小邮局]

结构化的优势

词的语义一目了然，降低使用成本

template-based ngram 取k-skip-n-gram和ngram之长，避其短

避免跨区域badcase 【跨省、市、区】引起的巨大亏损

精确定位用户收货地址上的作弊行为：

- 1. 提供多套行政区划降低邮费
- 2. 多个POI绕过地理围栏框定的业务限制

构建标准地址模型服务于地址验证、地址编码、同址判定等通用场景

编号	字段名称	字段语义
1	prov	省级行政区
2	city	地级行政区
3	district	县级行政区
4	devZone	开发区
5	town	乡级行政区
6	community	社区村委会
7	group	组
8	bizZone	商业圈
9	road	主路
10	subRoad	支路辅路
11	roadNo	主路号
12	subRoadNo	支路号
13	poi	兴趣点
14	subPoi	第二个POI
15	alley	村中路
16	houseNo	楼栋号
17	cellNo	单元号
18	floorNo	楼层号
19	roomNo	房间号
20	person	自然人&法人
21	addrInfos	其他
	otherPois	其他POI
	wrdList	门牌号之后词序

结构化的优势

词的语义一目了然

template-based n

避免跨区域badca

精确定位用户收货

- 1. 提供多套行
- 2. 多个POI绕过

构建标准地址模型
应用场景

```
# templates
city,town,poi
city,devZone,poi
district,town,road,poi
district,town,poi,subPoi
district,town,poi
road,roadNo,poi
district,road,roadNo
```

长，避其短
员

判定等通

编号	字段名称	字段语义
1	prov	省级行政区
2	city	地级行政区
3	district	县级行政区
4	devZone	开发区
5	town	乡级行政区
6	community	社区村委会
7	group	组
8	bizZone	商业圈
9	road	主路
10	subRoad	支路辅路
11	roadNo	主路号
12	subRoadNo	支路号
13	poi	兴趣点
14	subPoi	第二个POI
15	alley	村中路
16	houseNo	楼栋号
17	cellNo	单元号
18	floorNo	楼层号
19	roomNo	房间号
20	person	自然人&法人
21	addrInfos	其他
	otherPois	其他POI
	wrdList	门牌号之后词序

地址解析key point

- 地名识别
 - 分词
 - 语义标注：地名要素标注
 - 地名同名消歧义
- 未登录地名识别
 - 消除备注
 - overmerge
- 四级地址标准化
 - 前四级补全&纠错
 - 第五级关联验证
- 结构化
 - 地名要素和结构化等级映射
 - 顺序矫正
 - 参照物识别

分词

汉语文本是基于单字的

汉语的书面表达方式也是以汉字作为最小单位

词与词之间没有显性的界限标志

分词是汉语文本分析处理中首要问题

添加合适的、显性的词语边界标志使得所形成的词串反映句子的本意，这个过程就是通常所说的分词

NLP之困

歧义(Ambiguity)

- ▶ 注音歧义：快乐 (le4) , 音乐 (yue4)
- ▶ 分词歧义
 - ▶ 乒乓球/拍卖/完/了
 - ▶ 乒乓球拍/卖/完/了
- ▶ 短语歧义
 - ▶ [咬死猎人]的狗
 - ▶ 咬死[猎人的狗]
- ▶ 词义歧义
 - ▶ [打]乒乓球
 - ▶ [打]毛衣
 - ▶ [打]电话
- ▶ 语用歧义
 - ▶ “你真讨厌!”

病构(Ill-Formedness)

- ▶ 未登录词OOV
- ▶ 已知词的新用法
- ▶ 不合乎语法的句子
 - ▶ 他非常男人
- ▶ 不合乎语义约束的搭配
 - ▶ My car drinks gasoline like water.
- ▶ 由于疏忽造成的错误

分词之困

1.消除歧义

1)交叉歧义： 汉字串AJB被称作交集型切分歧义如果满足AJ、JB同时为词(A、J、B分别为汉字串)。 此时汉字串J被称作交集串

- ▶ [例] “结合成分子”

- ▶ 结合|成分|子|

- ▶ 结合|成|分子|

- ▶ 结|合成|分子|

2)组合歧义： 汉字串AB被称作组合型切分歧义， 如果满足 条件:A、B、AB同时为词

- ▶ 他站|起|身|来

- ▶ 他明天|起身|去北京

2.识别未登录词：

- ▶ 专有名词： 人名、地名、机构名、译名、时间词

- ▶ 领域术语：“互联网+”、“自然语言处理”

分词粒度

截止到目前没有统一的分词标准，但各自的标准大体相似

大家遵循：结合紧密、使用频繁

分词方法介绍

简单模式匹配

- ▶ 正向最大匹配
 - ▶ 错误切分率1/169
- ▶ 逆向最大匹配
 - ▶ 错误切分率1/245
- ▶ 双向匹配

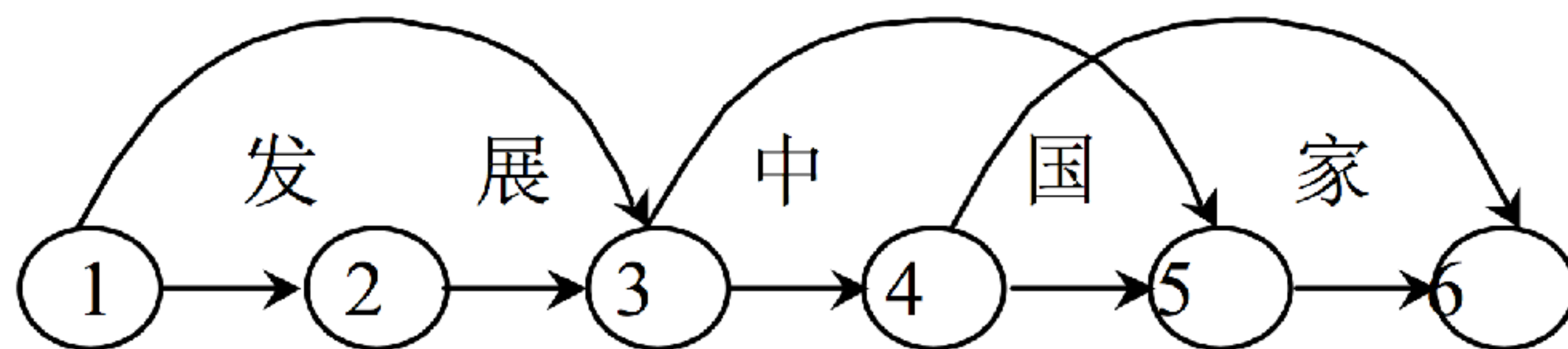
消除交叉歧义和组合歧义效果不明显

识别不了未登录词

基于规则的方法

- ▶ 最小分词方法

- ▶ 分词结果中含有词数最少：等价于在有向图里搜索最短路径问题



统计语言模型

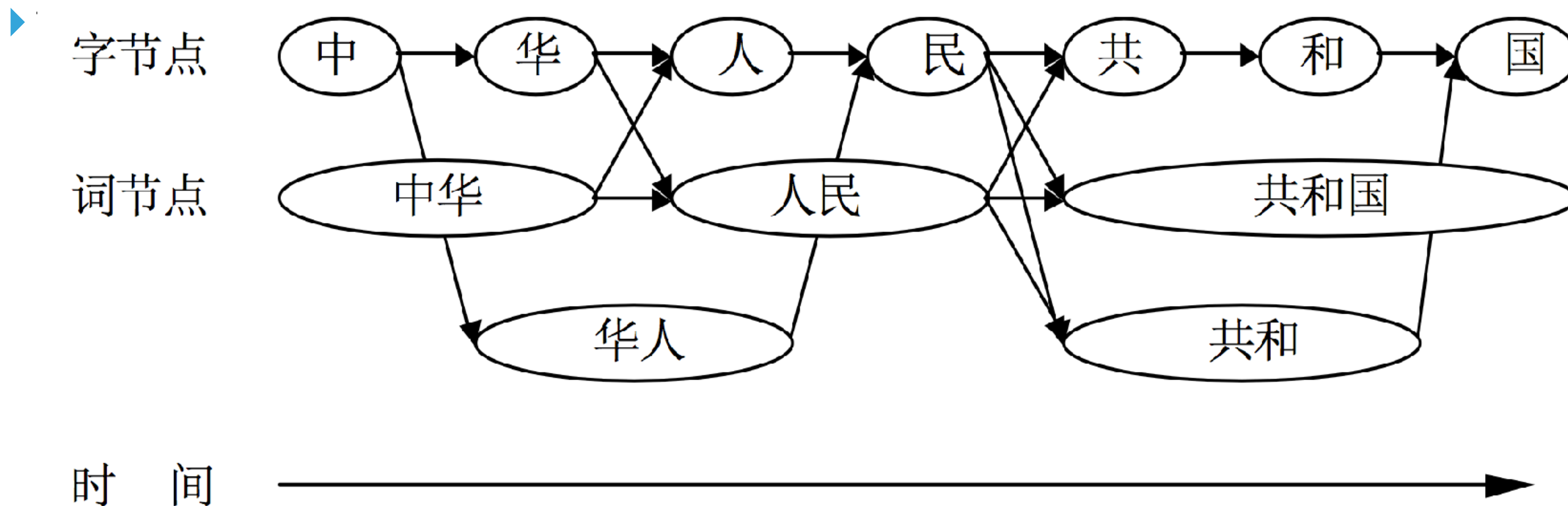
▶ 词网格

- ▶ 第一步：候选词网格构造:利用词典匹配，列举输入句子所有可能的切分词语，并以词网格 形式保存
- ▶ 第二步：计算词网格中的每一条路径的权值，权值通过计算图中每一个节点(每一个词)的一元统计概率和节点之间的二元统计概率的相关信息
- ▶ 根据图搜索算法在图中找到一条权值最大的路径，作为最后的分词结果

可利用不同的统计语言模型计算最优路径，具有较高的分词正确率

消除歧义的能力上升一个level，但识别OOV的能力依然较弱

统计语言模型



可利用不同的统计语言模型计算最优路径，具有较高的分词正确率

消除歧义的能力上升一个level，但识别OOV的能力依然较弱

字粒度的序列标注

原句：我爱北京天安门

标注：S S B E B M E

前提：需要丰富的标注训练语料

优点：同时cover消歧义和OOV的识别问题

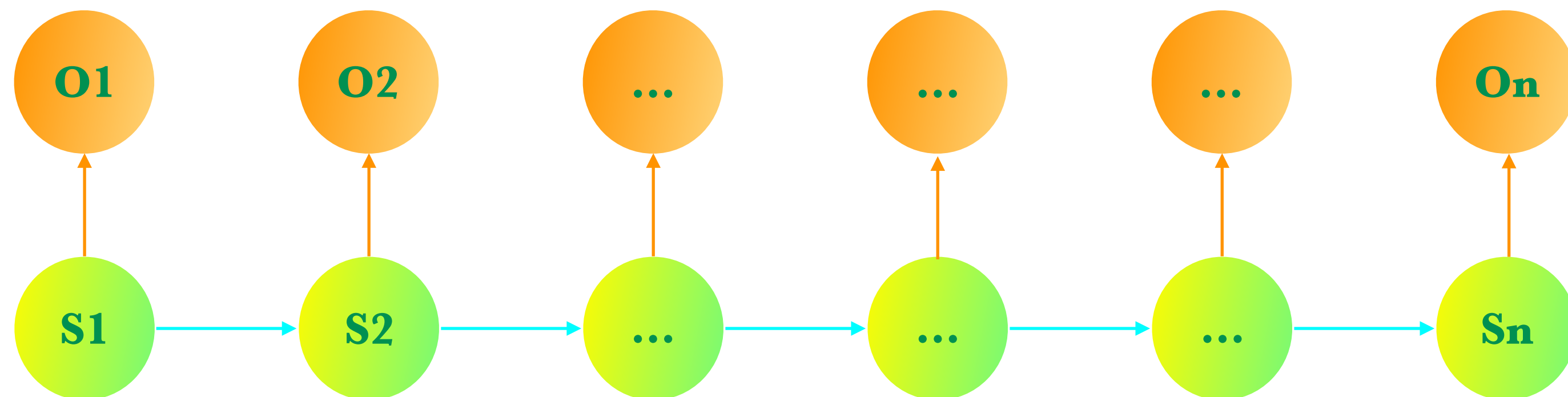
隐马尔科夫模型HMM

最大熵马尔科夫模型MEMM

条件随机场CRF

隐马尔可夫模型HMM

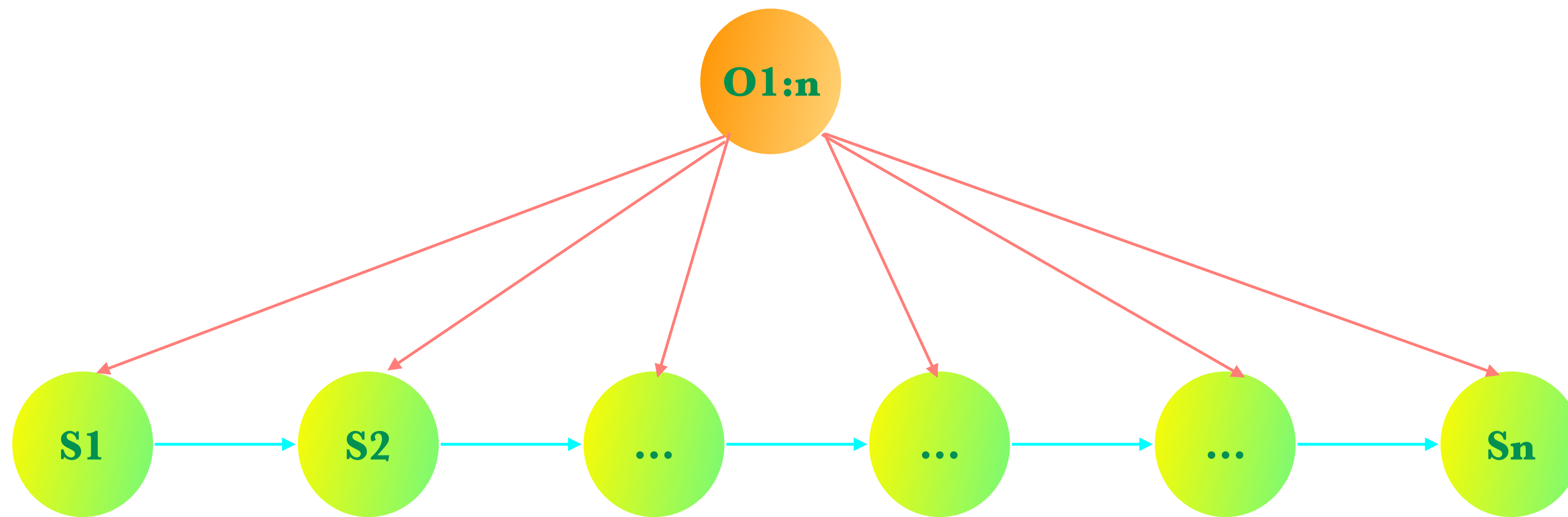
- ▶ 对转移概率和发射概率直接建模，统计共现概率，HMM三要素：
 - ▶ 状态转移矩阵A
 - ▶ 发射矩阵B
 - ▶ 初始向量 P_i
- ▶ 三个假设：
 - ▶ 马尔科夫假设
 - ▶ 稳定性假设：与具体时间无关
 - ▶ 输出独立性假设



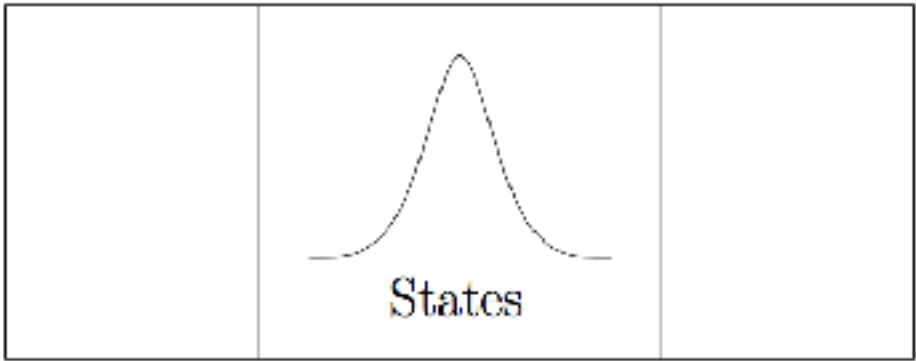
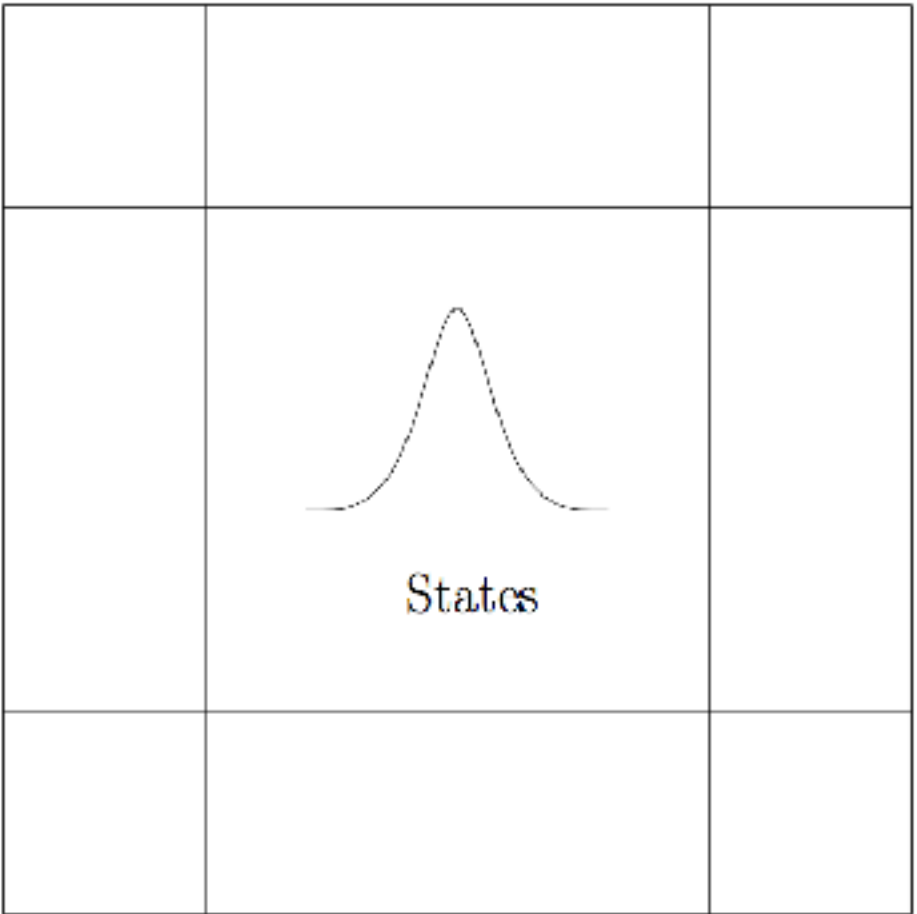
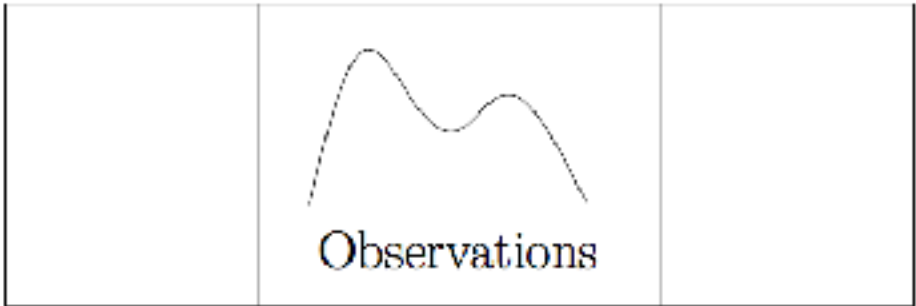
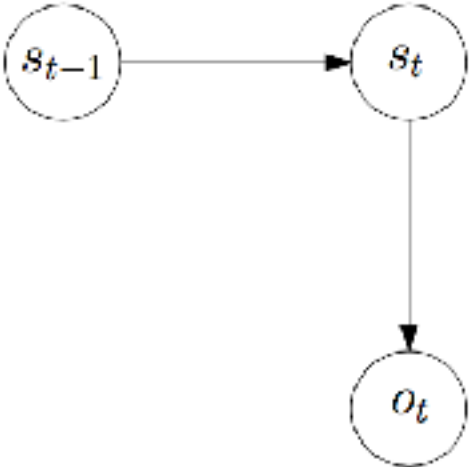
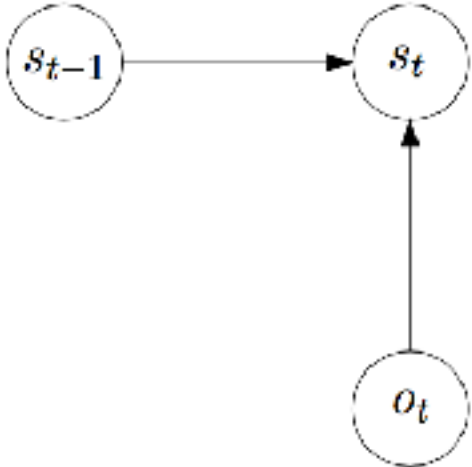
最大熵马尔可夫模型MEMM

▶ MEMM

- ▶ 对转移概率和发射概率建立联合概率，统计的是条件概率，但它容易陷入局部最优，是因为它只在局部做归一化

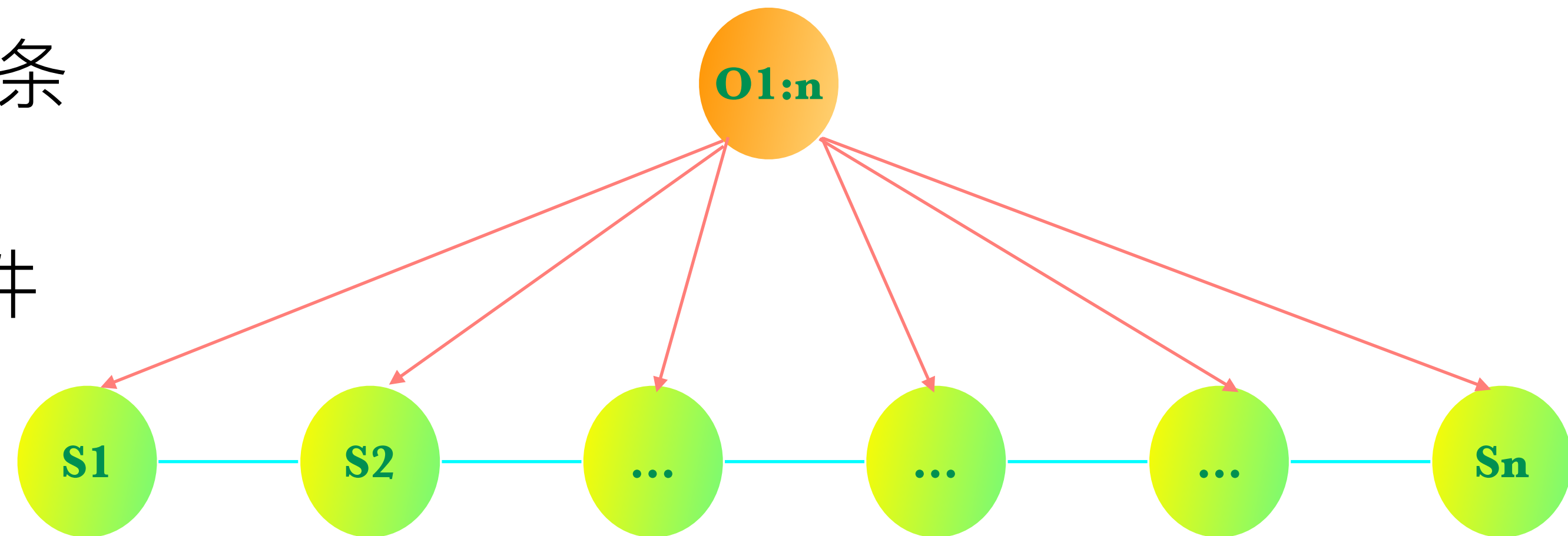


HMM VS. MEMM

HMMs	MEMMs
$P(s s'), P(o s)$ State_i	$P(s s', o) \Leftrightarrow S $ distributions: $P_{s'}(s o)$ State_j
 <p>States</p>	 <p>States</p>
State_i  <p>Observations</p>	
	

条件随机场CRF

- 统计了全局概率，在做归一化时，考虑了数据在全局的分布，这样就解决了MEMM中的标记偏置的问题，CRF的优点：
 - CRF没有HMM那样严格的独立性假设条件，因而可以容纳任意的上下文信息。
 - 由于CRF计算全局最优输出节点的条件概率，它克服了MEMM标记偏置的缺点。
 - CRF是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率分布，而不是在给定当前状态条件下，定义下一个状态的状态分布。



序列标注模型选型

- ▶ 前提：一定规模的训练语料
- ▶ 硬件有约束：HMM首选
- ▶ 硬件充裕：CRF

工业界主流分词：HMM+词典/CRF+词典

地址解析架构



地名识别

余杭仓前良睦路999号圣殿公交站对面万利大厦后面乐佳国际2号楼小邮局

分词&
语义标注

余杭/district_abbrev,town_abbrev,community_abbrev 仓前/town_abbrev,community_abbrev 良睦路/road 999/num 号/hao 圣殿/first 公交站/poi_feature 对面/position 万利/first 大厦/building_feature 后面/position 乐佳/first 国际/biztype 2/num 号楼/house 小邮局/poi

语义标签

- ▶ 词性(parts of speech): 语言学家将词按照相似的语法结构行为和典型的语义类型聚成不同的类。又称为句法类或语法类
 - ▶ 现代汉语的词可以分为两类12种词性。一类是实词：名词、动词、形容词、数词、量词和代词。一类是虚词：副词、介词、连词、助词、叹词和拟声词。
- ▶ 语义标签与词性很类似都是词的一种分类。
- ▶ 语义标签是在词性基础上的细分，不仅能体现出词的语用，更能体现词更具体的语义。
- ▶ 词关联上语义标签之后，语义标签相同的词，其处理方式一致的。因此可将处理规则或文法可定义在语义标签上，从而大大降低规则的复杂度及规模
- ▶ 语义标签定义标准
 - ▶ 行政区划
 - ▶ 关联词的字面特征&结构
 - ▶ 使用场景
 - ▶ 词的语义角色
 - ▶ 同类词的规模

目前为止定义了150多个语义标签

行政区划

- ▶ 国家级
 - COUNTRY # 国家
- ▶ 省级行政区：省、自治区、直辖市简称
 - PROVINCE # 省级行政区_标准名字
 - PROVINCE_ALIAS # 省级行政区_别名
 - PROVINCE_ABBR # 省级行政区_简称
 - MUNICIPALITY_ABBR # 直辖市简称
- ▶ 地级行政区：地级市、直辖市全称、自治州、地区、盟
 - CITY # 地级行政区
 - CITY_ALIAS # 地址行政区_别名
 - CITY_ABBR # 地级行政区_简称
 - MUNICIPALITY # 直辖市_全称
- ▶ 县级行政区：区、县、县级市、自治县等
 - COUNTY # 县级行政区_县/县级市
 - COUNTY_ALIAS # 县级行政区_县市别名
 - COUNTY_ABBR # 县级行政区_县市简称
 - DISTRICT # 县级行政区_区
 - DISTRICT_ALIAS # 县级行政区_区的别名
 - DISTRICT_ABBR # 县级行政区_区的简称
- ▶ 开发区、高新区等
 - DEVZONE
 - DEVZONE_ALIAS
 - DEVZONE_SUFFIX
 - DEVZONE_ALIAS_FEATURE
 - DEVZONE_FEATURE
- ▶ 乡级行政区：乡、镇、街道、苏木、民族乡、区公所、乡级单位[农场、牧场、林场]
 - TOWN # 国家
 - TOWN_ABBR
 - TOWN_ALIAS
 - TOWNSHIP
- ▶ 社区级：社区、居委会、村委会、行政村等
 - COMMUNITY # 省级行政区_标准名字
 - COMMUNITY_ALIAS # 省级行政区_别名
 - COMMUNITY_ABBR # 省级行政区_简称
- ▶ 行政区划后缀词汇：省、市、区/县、自治州、乡/镇/街道、开发区、社区/村委会
 - SHENG # 省
 - PROVICNE_FEATURE # 自治区
 - SHI # 市
 - CITY_FEATURE # 自治州
 - QU # 区
 - XIAN # 县
 - COUNTY_FEATURE # 自治县、旗
 - TOWN_FEATURE # 乡/镇/街道
 - TOWN_ALIAS_FETURE # 街道办，街道办事处，办事处
 - TOWNSHIP_FEATURE #
 - COMMUNITY_FEATURE # 社区、村委会
 - COMMUNITY_ALIAS_FEATURE # 居委会、村民委员会、社区委员会

POI

▶ 小区、村庄

- GARDEN # 居民小区、城市住宅区
- VILLAGE # 自然村

▶ 工业区、工业园

- INDUSTRIAL_PARK # 工业区、工业园

▶ 写字楼

- BUILDING # 写字楼、大厦、建筑物名

▶ 全国普通高校、中小学、幼儿园

- UNIV # 大学
- UNIV_ALIAS # 大学_别名
- UNIV_ABBR # 大学_简称
- UNIV_CAMPUS # 含有校区信息的大学
- CAMPUS # 校区
- SCHL # 大学下属的学院
- UNIV_SCHL # 大学学院
- EDU # 中小学, 幼儿园, 中专等初等教育机构
- EDU_ALIAS # 初等教育机构的别名
- EDU_ABBR # 初等教育机构的简称
- EDU_CAMPUS # 含有校区信息的初等教育机构

▶ 公司

- CORP # 公司企业
- CORP_CORE # 公司企业名核心部分
- CORP_BRANCH # 含有分公司信息

▶ 工厂

- FACTORY # 工厂
- FACTORY_BRANCH # 工厂分厂

▶ 政府机构

- GOV # 政府机构
- GOV_BRANCH

▶ 医疗机构

- HOSPITAL
- HOSPITAL_BRANCH

▶ 其他

- POI
- POI_BRANCH
- AOI

▶ POI组成词

- NATION # 全国各民族
- BRAND # 品牌词汇
- FRIST # POI核心词
- BIZTYPE # POI类型词

同名消歧义

余杭/district_abbrev, town_abbrev, community_abbrev 仓前/town_abbrev, community_abbrev 良睦路/road 999/num 号/hao 圣殿/first 公交站/poi_feature 对面/position 万利/first 大厦/building_feature 后面/position 乐佳/first 国际/biztype 2/num 号楼/house 小邮局/poi

消歧义

余杭/district_abbrev 仓前/town_abbrev 良睦路/road 999/num 号/hao 圣殿/first 公交站/poi_feature 对面/position 万利/first 大厦/building_feature 后面/position 乐佳/first 国际/biztype 2/num 号楼/house 小邮局/poi

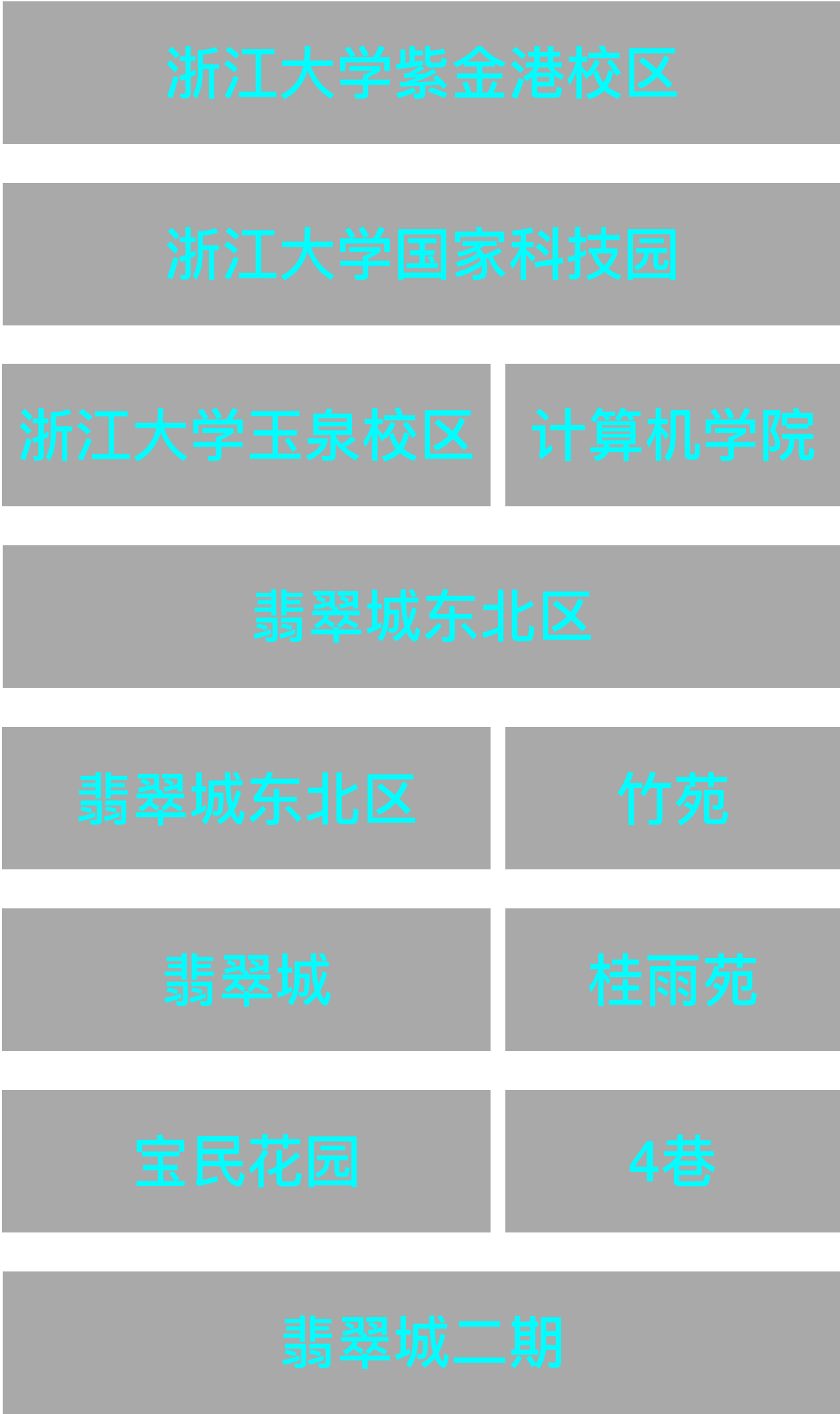
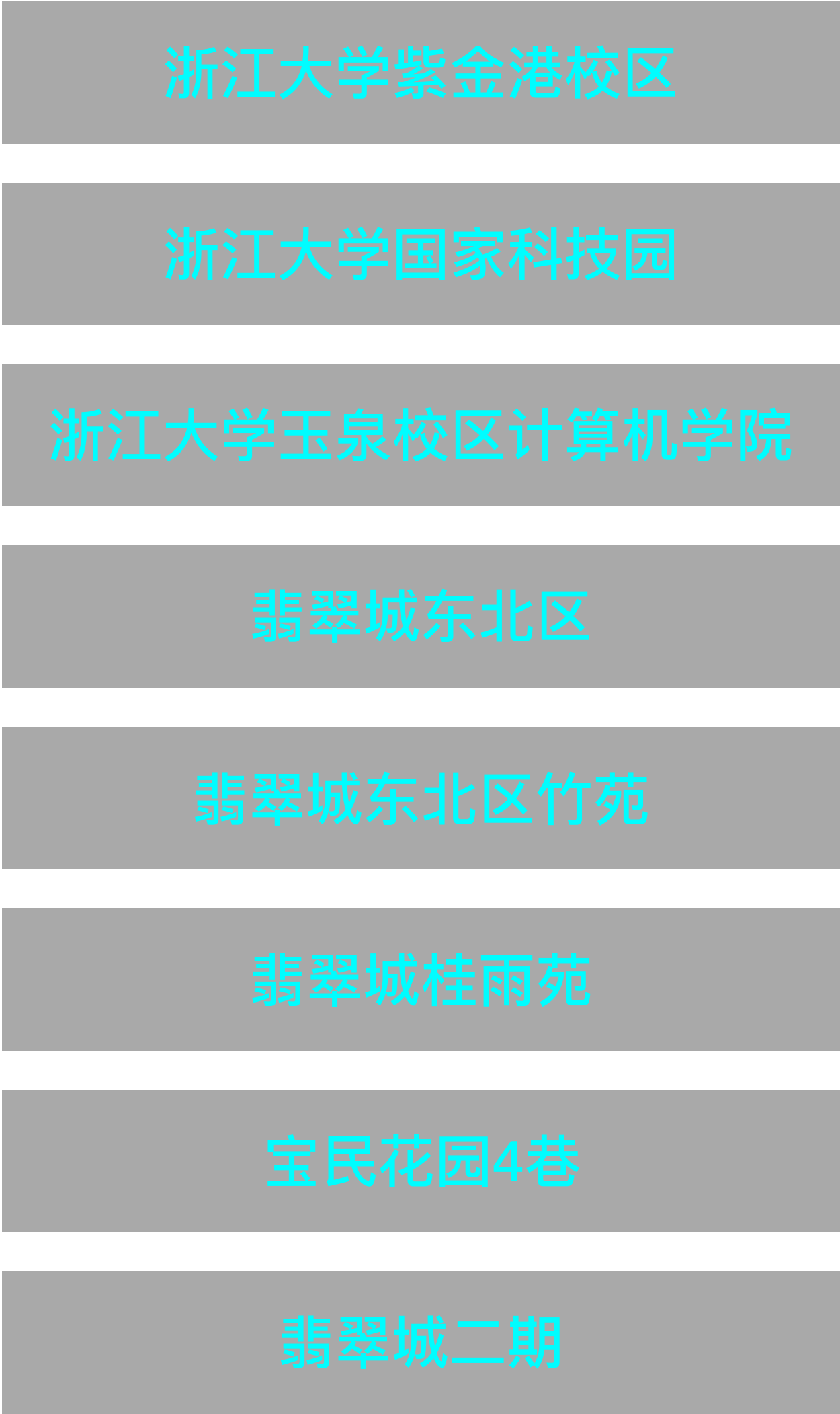
OOV地名识别

余杭/district_abbrev 仓前/town_abbrev 良睦路/road 999/num 号/hao 圣殿/first 公交站/
poi_feature 对面/position 万利/first 大厦/building_feature 后面/position 乐佳/first 国
际/biztype 2/num 号楼/house 小邮局/poi



余杭/district_abbrev 仓前/town_abbrev 良睦路/road 999号/road_no 圣殿公交站/poi 对面
/position 万利大厦/building 后面/position 乐佳国际/aoi 2号楼/house_no 小邮局/poi

地名粒度定义



OOV识别之困

POI构词结构复杂多样

1. POI里含有路：光明路小学
2. POI含有小区：求是（和家园）小学
3. 名址合一：上海市浦东新校区白玉兰小学
4. 括号内容歧义：可以是POI的一部分，也可以是备注

- 1) 恒泰药店（崇明西路）
- 2) 老山蜂蜜园（镇江店）
- 3) 蓝天网报（西园新村西南）
- 4) 真维斯（东方大厦西）

5. POI词很长

- 1) POI:7
- 2) 公司名:13.2
- 3) 行政区划：3.8
- 4) 最长POI：45

OOV POI识别核心

1. 识别构词POI的小粒度词
2. 预测最优的语义标签
3. 基于语义标签构建OOV POI 识别文法，实现字面特征的大规模降维

未登录行政区划识别

```
county_abbrev qu    credible;county_alias
county_abbrev shi   credible;county_alias
district_abbrev shi credible;district_alias
district_abbrev xian credible;district_alias
city_abbrev shi    credible;city
```

未登录POI识别

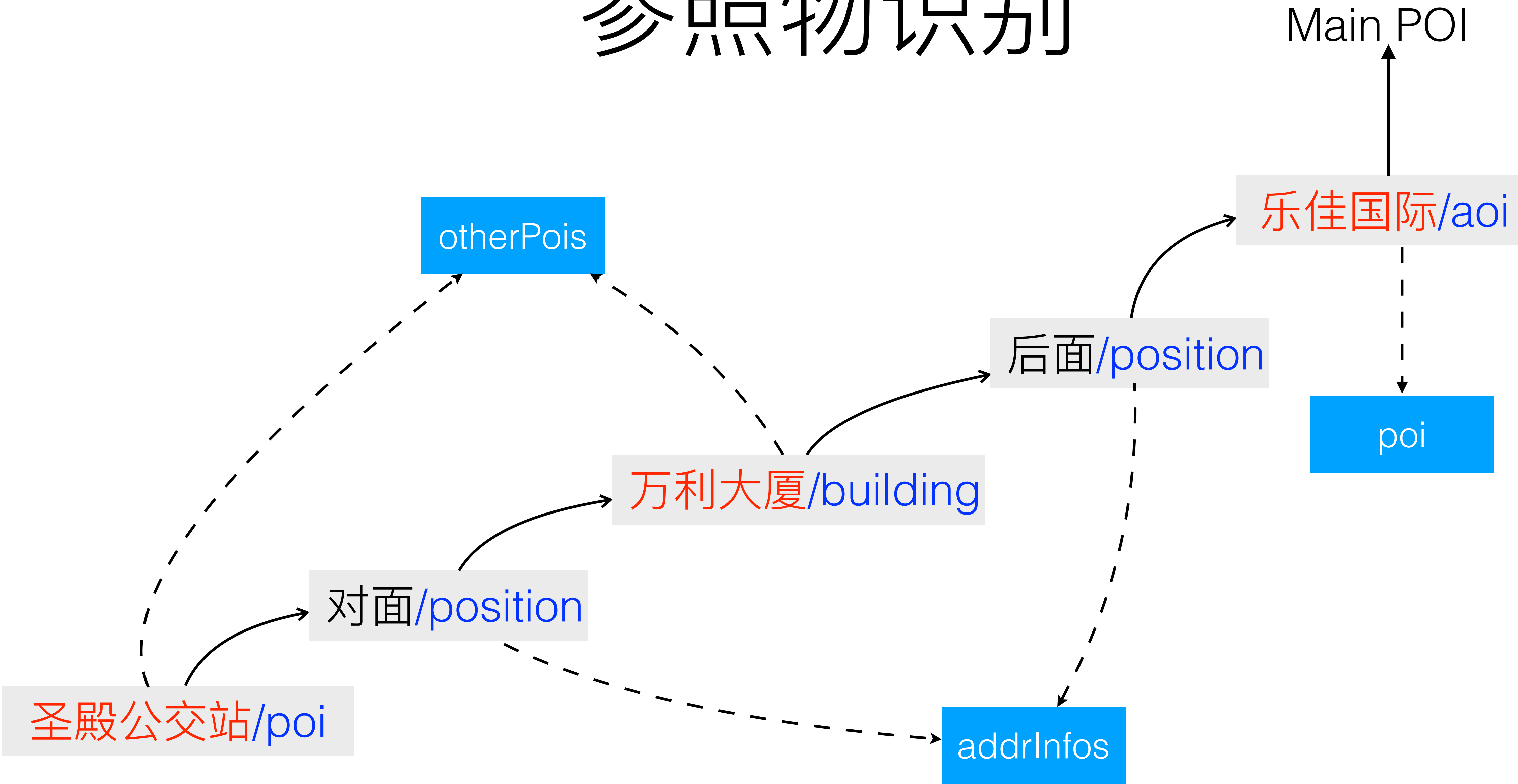
```
univ city_abbrev campus_feature    credible;univ_campus
province_abbrev first biztype biztype corp_feature    credible;corp
city_abbrev first biztype biztype corp_feature    credible;corp
city first biztype biztype corp_feature    credible;corp
city_abbrev biztype univ_feature    credible;univ
```

基于active learning方式挖掘小粒度词和标注标签

参照物识别

圣殿公交站/poi 对面/position 万利大厦/building 后面/position 乐佳国际/aoi

参照物识别



结构化

余杭/district_abbrev 仓前/town_abbrev 良睦路/road 999号/road_no 圣殿公交站/poi 对面
/position 万利大厦/building 后面/position 乐佳国际/aoi 2号楼/house_no 小邮局/poi

Structuration parser



district=余杭 town=仓前 road=良睦路 roadNo=999号

poi=乐佳国际 houseNo=2号楼 person=小邮局

otherPoi=[圣殿公交站, 万里大厦] addrInfos=[对面, 后面]

wrdList=[圣殿公交站, 对面, 万利大厦, 后面, 乐佳国际, 2号楼, 小邮局]

标准化

district=余杭 town=仓前 road=良睦路 roadNo=999号

poi=乐佳国际 houseNo=2号楼 person=小邮局

otherPoi=[圣殿公交站, 万里大厦] addrInfos=[对面, 后面]

wrdList=[圣殿公交站, 对面, 万利大厦, 后面, 乐佳国际, 2号楼, 小邮局]



prov=浙江省 city=杭州市 district=余杭区 town=仓前街道 road=良睦路

roadNo=999号 poi=乐佳国际 houseNo=2号楼 person=小邮局

otherPoi=[圣殿公交站, 万里大厦] addrInfos=[对面, 后面]

wrdList=[圣殿公交站, 对面, 万利大厦, 后面, 乐佳国际, 2号楼, 小邮局]

标准化的作用

信息
检索

行政区划地名归一化

地址补全&纠错

行政区划关联验证

同名地名二次消歧义

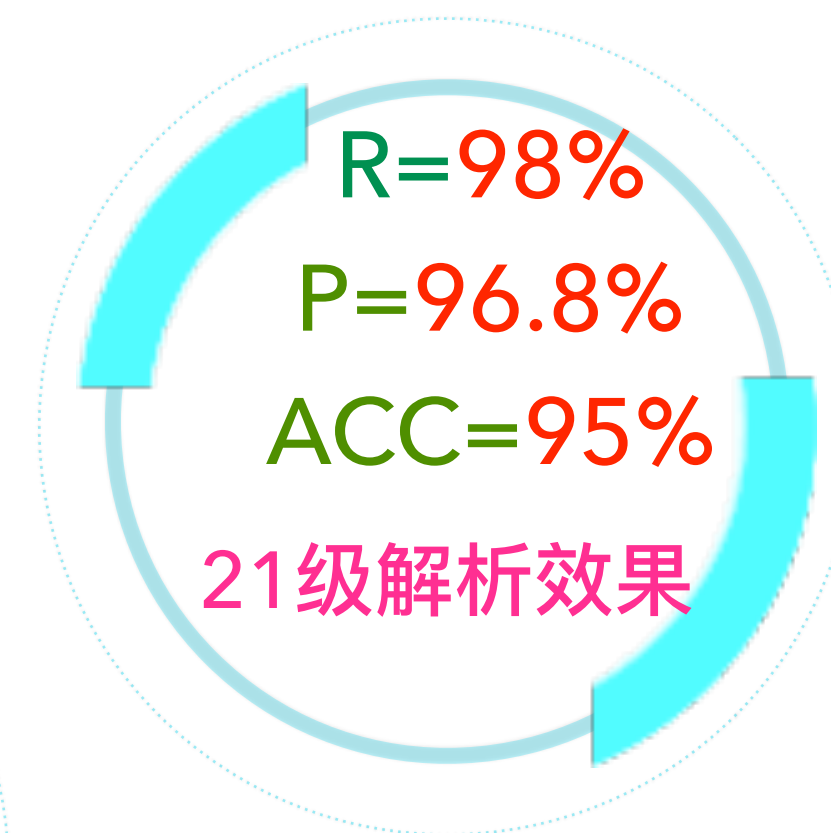
消除overmerge

安徽省安庆市岳西县中医院

安徽省 安庆市 岳西县中医院

安徽省 安庆市 岳西县 岳西县中医院

项目效果



Q/A, Thx!

-王国印