携程AI助力产品内容化实践

孙玉霞

孙玉霞

携程平台中心AI研发部，目前担任资深算法工程师，负责携程产品内容化中台相关算法的落地实践和优化。

5年+的自然语言处理经验，专注于文本挖掘/匹配/生成方向。2016年加入携程，负责小诗机，智能客服等项目。

目录

目录

内容化痛点

找主题难

选产品难

挖内容难

AI赋能

马可波罗平台

平台功能

特色发现

榜单对标

主题产品挖掘

智能主题配置

主题图片挖掘

主题优质文本抽取

主题文章挖掘

文章自动挂货

文章评级

首图优选

微游记生成

短亮点

长推荐理由

算法层

事件发现

实体链接

敏感内容发现

内容聚合

文本生成

情感分析

主题匹配

摘要抽取

词法分析

句法分析

NER

知识图谱

图文匹配

图片优美度

图片分类

图片去重

数据层

商品信息

用户图片

用户评论

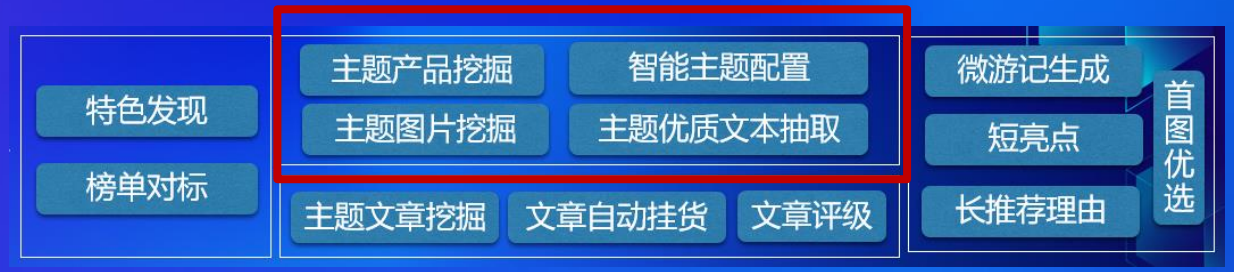
游记

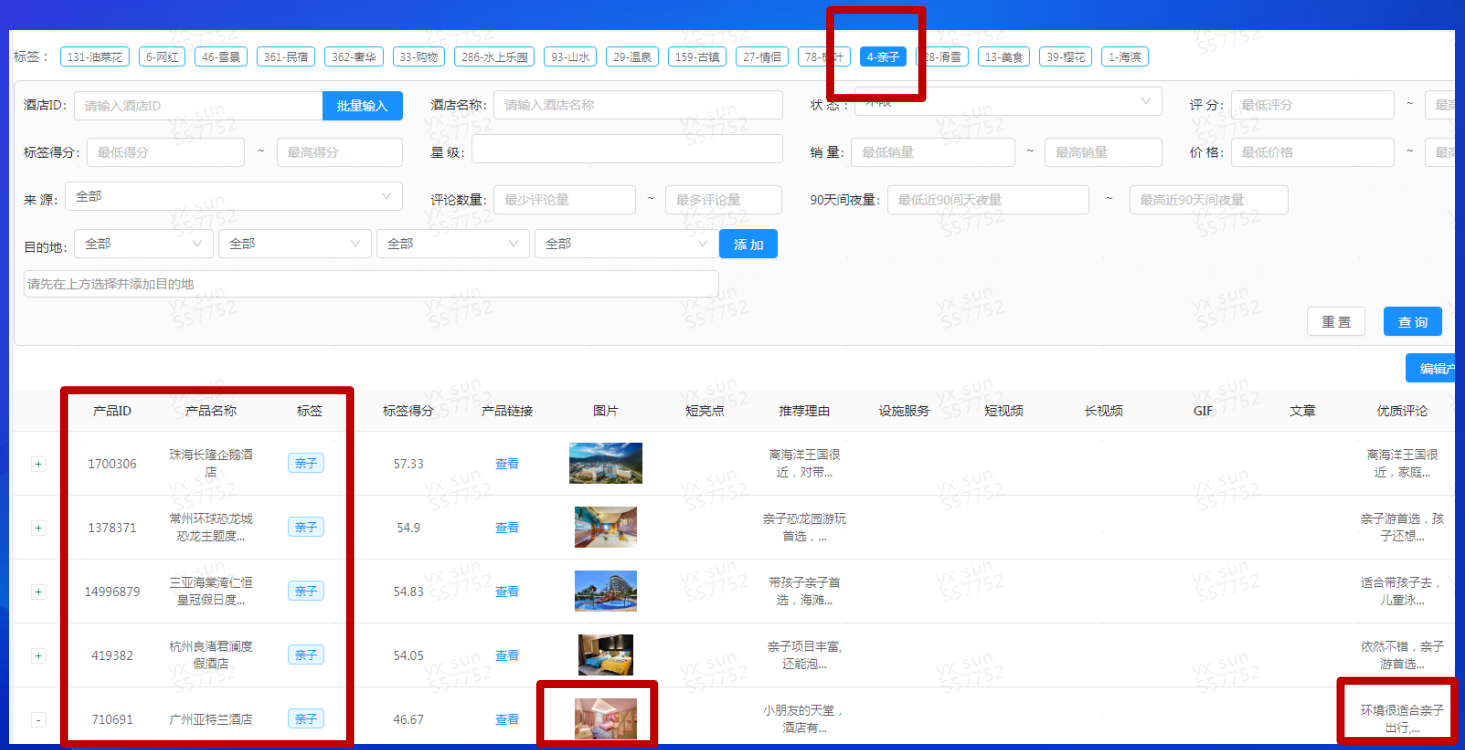
旅拍

问答

马可波罗平台

主题产品挖掘 - 产品主线的内容化





马可波罗平台

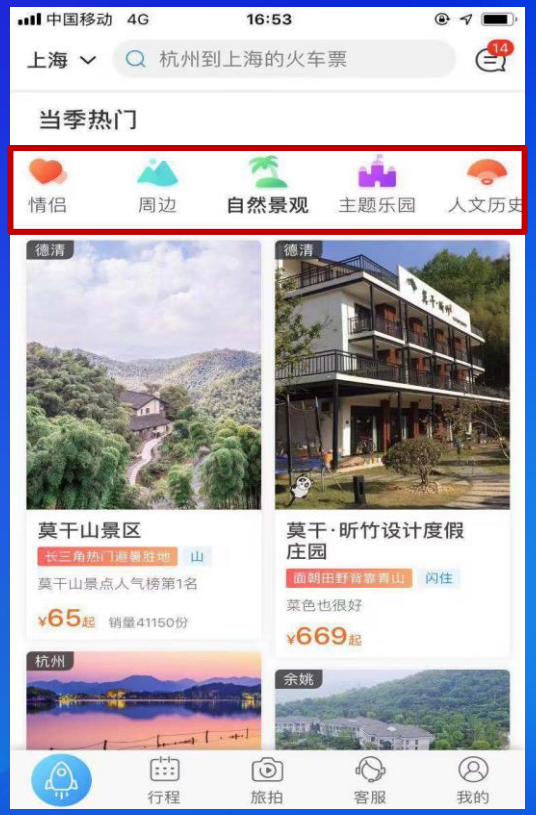
主题产品挖掘 - 产品主线的内容化

亲子

产品ID	产品名称	标签
1700306	珠海长隆企鹅酒店	亲子
1378371	常州环球恐龙城恐龙主题度...	亲子
14996879	三亚海棠湾仁恒皇冠假日度...	亲子
419382	杭州良渚君澜度假酒店	亲子
710691	广州亚特兰酒店	亲子

1. 这是一家性价比很高的亲子酒店，距离广州长隆动物园非常近，打车10分钟左右；
2. 亚特兰亲子酒站是小朋友的天堂，酒店里就有游乐园，能让小朋友乐不思蜀；
3. 房间走亲子风格，可爱卡通，小朋友喜欢；
4. 在长隆边上，去长隆玩比较适合，酒店内部装修风格适合亲子游，卫生服务很好；
5. 很适合亲子旅游，亲子房设计合理，床够大，餐厅也不错；

业务应用



首页瀑布流

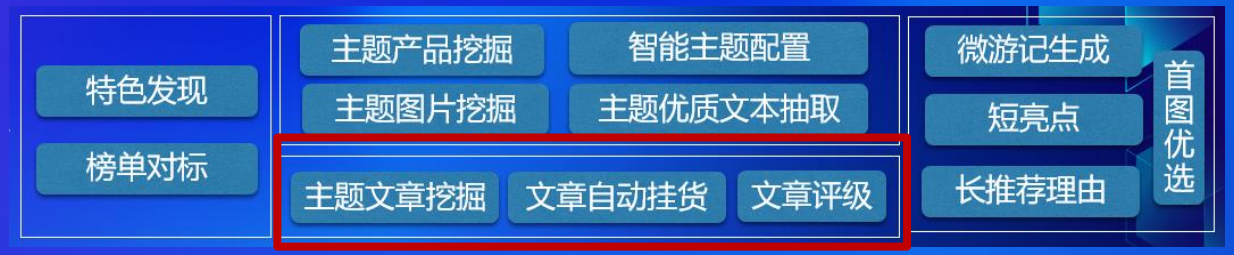
酒店瀑布流

机票瀑布流

主题产品挖掘

马可波罗平台

主题文章挖掘 - 文章主线的内容化



1

- 根据当前话题自动获取最相关的文章。
- 话题文章聚合。

主题文章挖掘

2

- 对图片/文本等内容的量级以及质感进行综合评分，级别粗筛。
- 有效过滤低质文章，缩小候选文章的量。

文章评级

3

- 识别文章中提及到的城市/景点/酒店/餐厅等。
- 提高产品挂载的灵活性。

文章自动挂货

业务应用

中国移动 4G
21:36
#网红打卡地
关注
内容
地点
达人
全部地区
照点——澳门渔人码头
Batumi Botanical Ga...
叶包-sanny
771
和亚美尼亚小姐姐疯狂一夜#向往的生活
doubleliker
23
香港迪士尼乐园
香港迪士尼夜间大巡游光影秀
霏儿
32
花莲
花莲县，原始的
牛奔走天下
参与话题

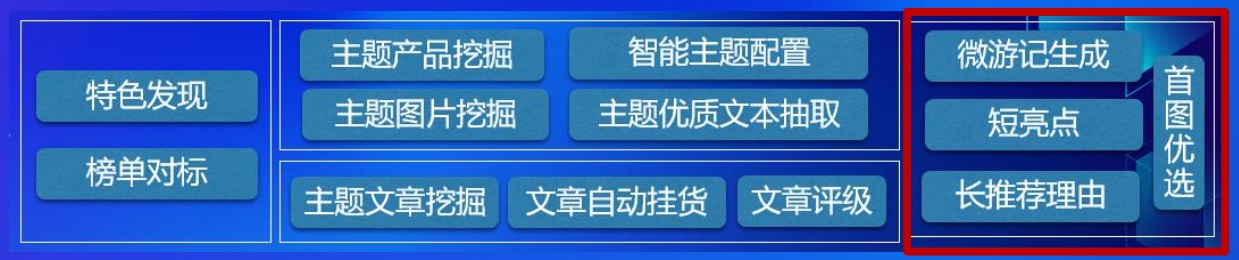
中国移动 4G
18:14
香港迪士尼乐园
"置身奇妙的迪士尼世界"
4.8
46217条点评
香港景点人气榜第1名
香港大屿山 香港迪士尼乐园度假区
需购票 ¥512起
预订
香港迪士尼乐园开闭园时间...
游玩约需1-2天
更多详情
预订
玩法
点评
攻略
推荐

中国移动 4G
迪士尼乐园度假区 迪士尼乐园
地图/周边
4.8分 棒
8405条点评
2017年开业
详情/设施
2019酒店口碑榜
11月3日
11月4日 周一
1间 · 1人
含早餐
双床
立即确认
筛选
每间总价
标准客房
34㎡ 无餐食 双床 2人入住
不可取消
立即确认
¥1876

攻略 – 话题圈子

马可波罗平台

内容多样化



素材库
- 优质长文本/短亮点抽取
- 首图优选
- 内容改写和生成

微游记生成
- 语义去重
- 标签化
- 图文匹配

文章框架
- 产品-多维度模式
- 主题模式
- 行程模式

业务应用

酒店 – 首页二屏

酒店 - 推荐理由

餐厅 - 推荐理由

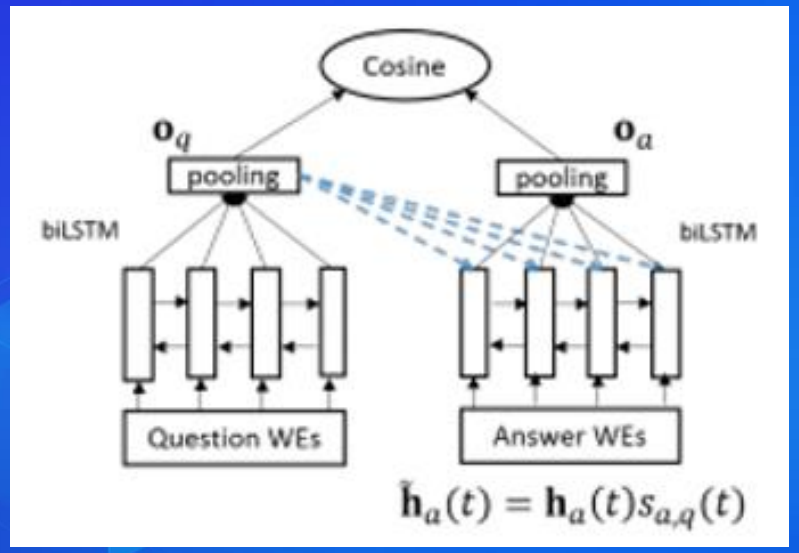
IM+ 酒店推荐理由

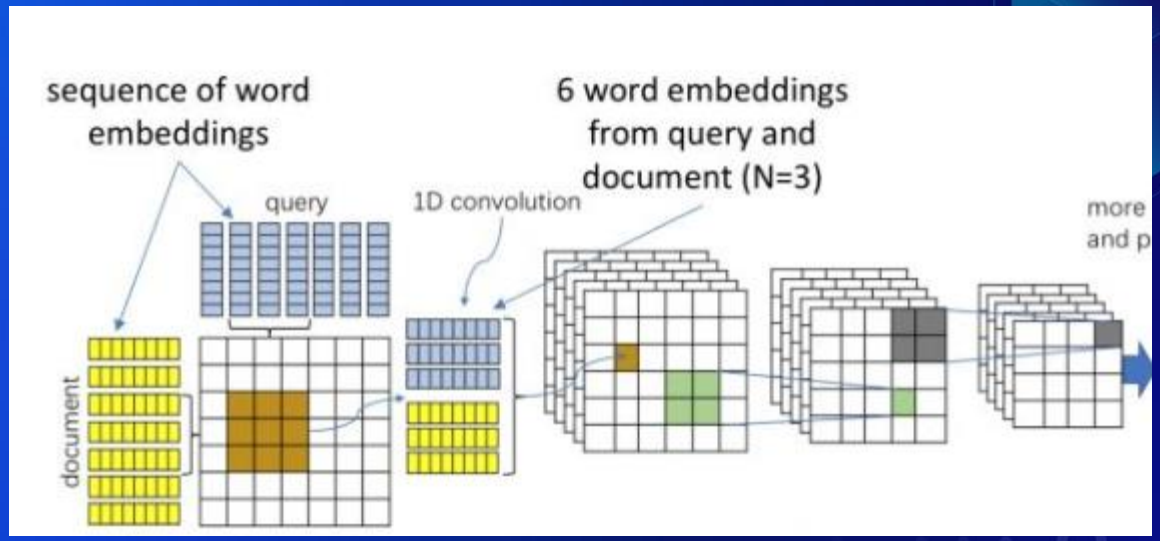
目录

2.1 主题产品/文章挖掘

匹配模型简单分类

- 交互式和非交互式
- Loss Pointwise（0，1分类）和pairwise max{0,margin-(S(Q,D+)-S(Q,D-))}





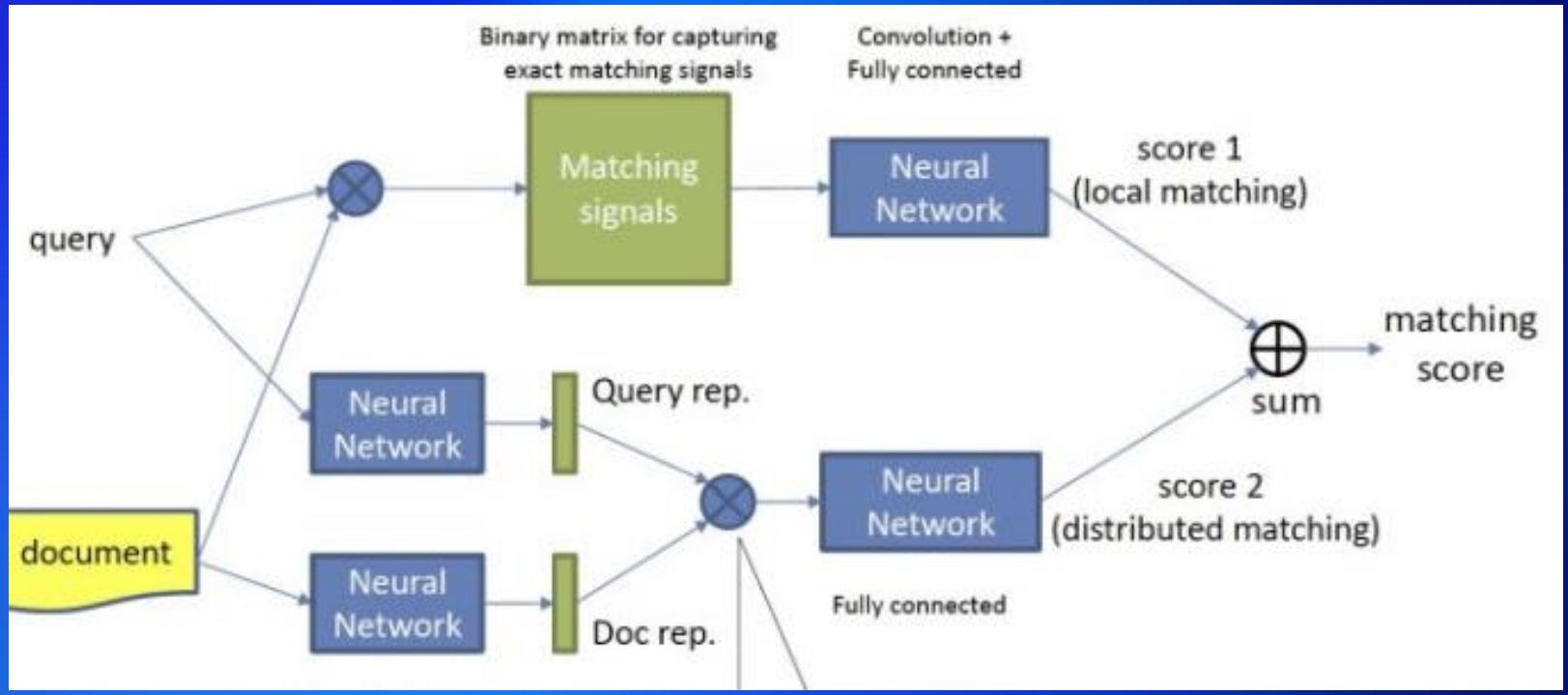
Part 1

- q,a使用同样的网络抽取特征
- 计算相似度前有交互，如pooling、attention。

Part 2

- q,a对应的embedding后的word计算点积 matching signals
- CNN方法多次卷积
- 2分类模型

2.1 主题产品/文章挖掘

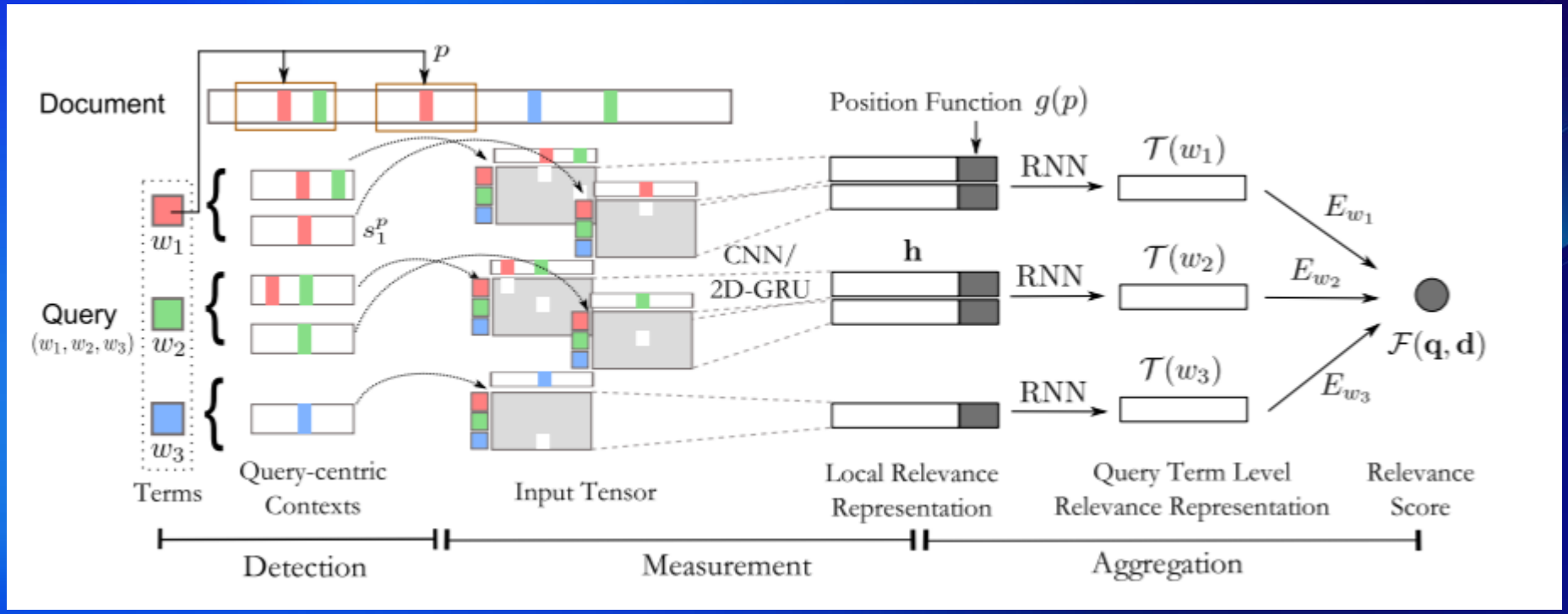


- 短文本之间的匹配
- 结合上文的两种抽取方法进行网络组合。
- 采用pairwise的训练方法，输入为（q, a+, a-）,a+表示跟q匹配，a-不匹配。
  pointwise 每个输入为q,a,label，label表示是否匹配。
- Loss = max{0,margin-(S(Q,D+)-S(Q,D-))}+loss（分类模型交叉熵损失）

Learning to Match using Local and Distributed Representations of Text for Web Search

2.1 主题产品/文章挖掘

- 短文本-长文本主题相关性
- 长文本主题不聚焦
- 长文本主题多样性
- 类似于query-document问题



DeepRank: A New Deep Architecture for Relevance Ranking in Information Retrieval

2.1 主题产品/文章挖掘

1 短文本偏好

2 主题产品优质度不一

3 关键词定义时随意性

“避暑胜地”，弱关键词“暑假 清凉”

4 关键词歧义

“樱花季” – 前台樱花服务热情，xx樱花酒店；主题“江南水乡” – 首尔江南

5 千万级候选

2.2 文章自动挂货

基础实体识别和链接

附近优质酒店

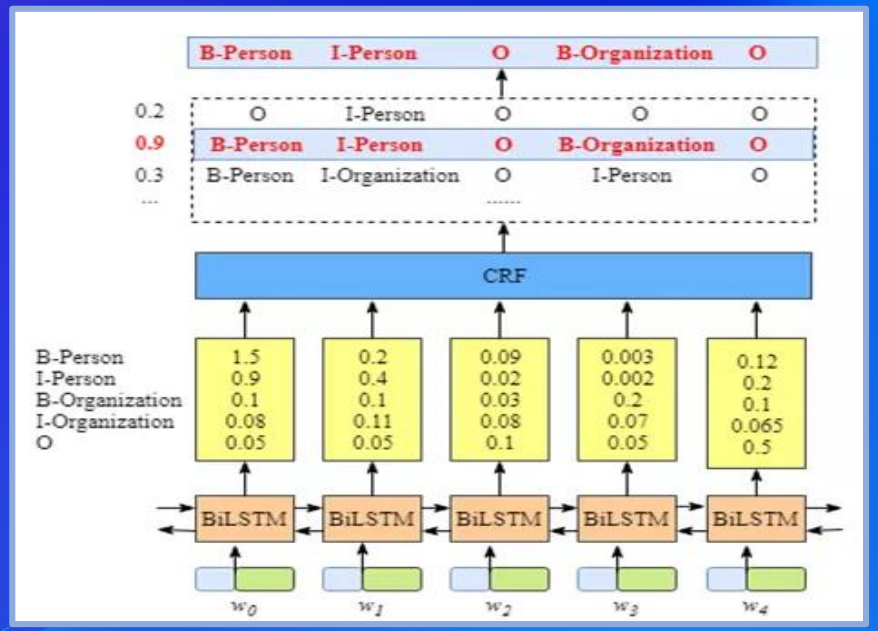
城市热门景点

相关门票推荐

.....

多场景灵活的挂货策略

2.2 文章自动挂货



NER → 人名/地名/机构名

类别细分 → 城市/景点/酒店

链接校验 → 反向校验

1 识别不全

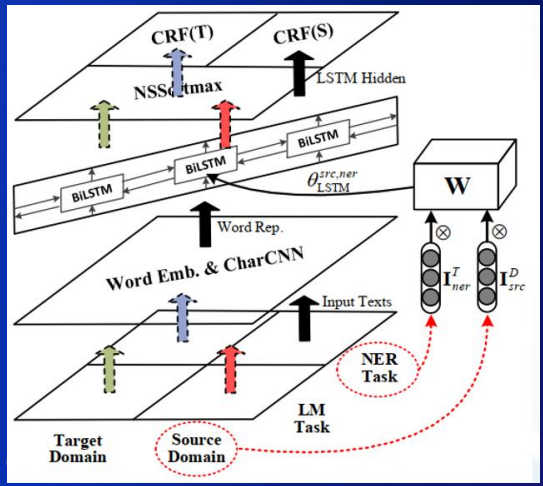
瘦***西湖景区***

实体补全+反查

2 识别有误

另外酒店还非常人性化地为我们把退房时间延迟.

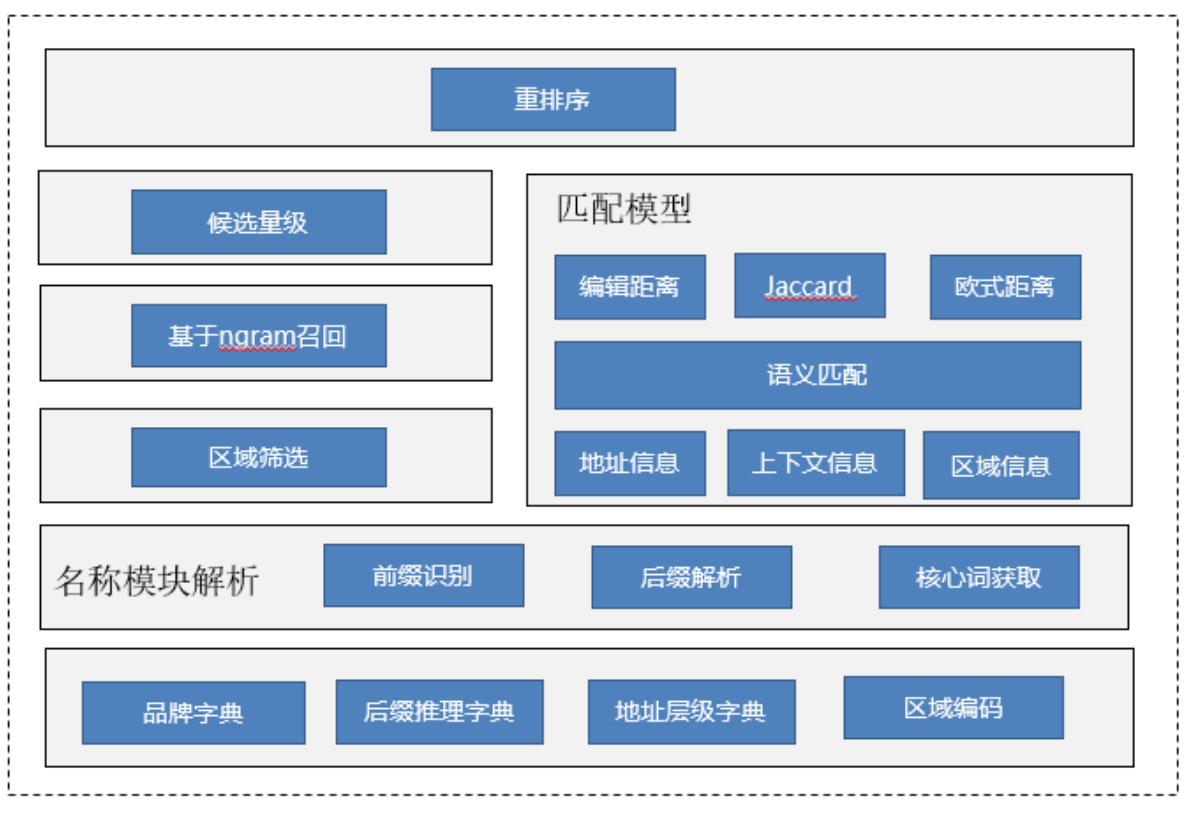
句法分析+纠错

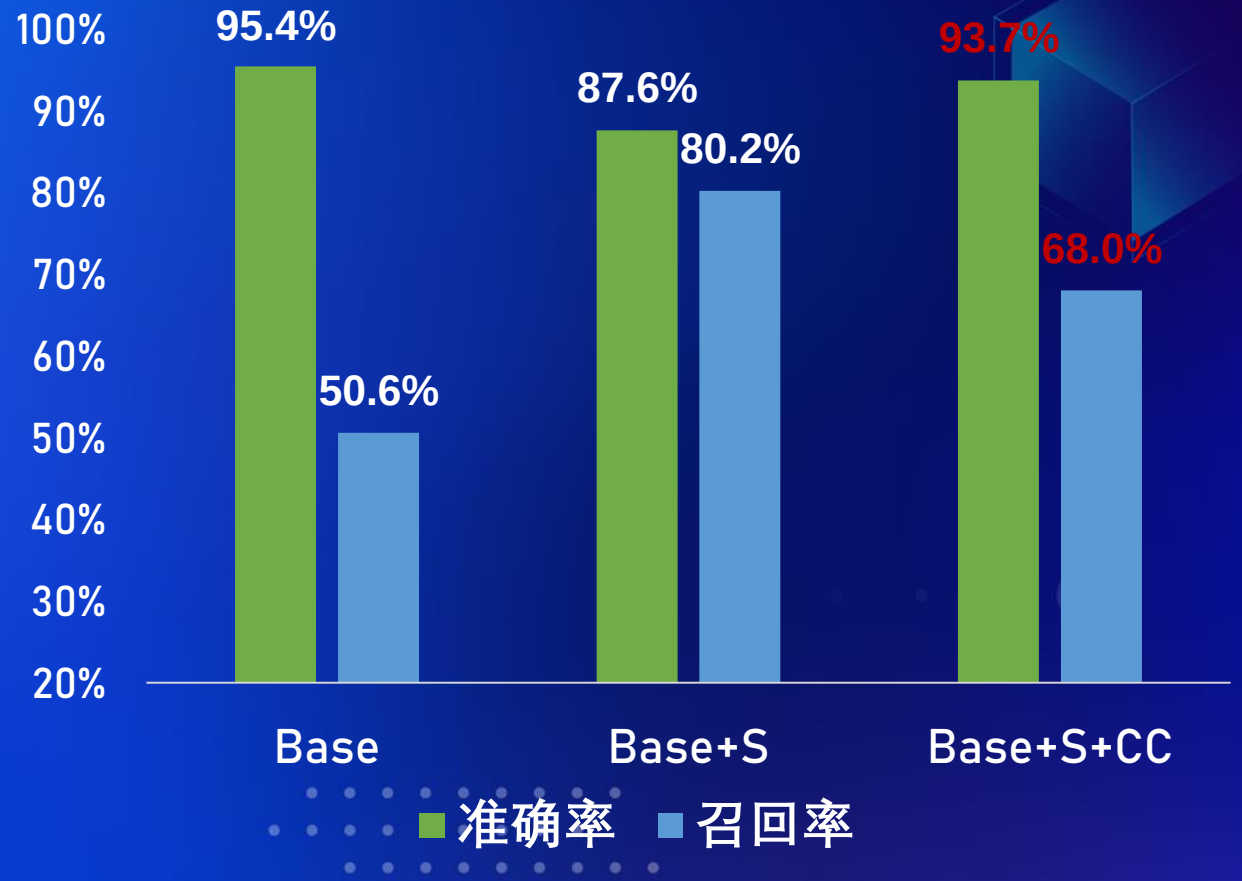


Cross-Domain NER using Cross-Domain Language Modeling

2.2 文章自动挂货

实体链接算法框架

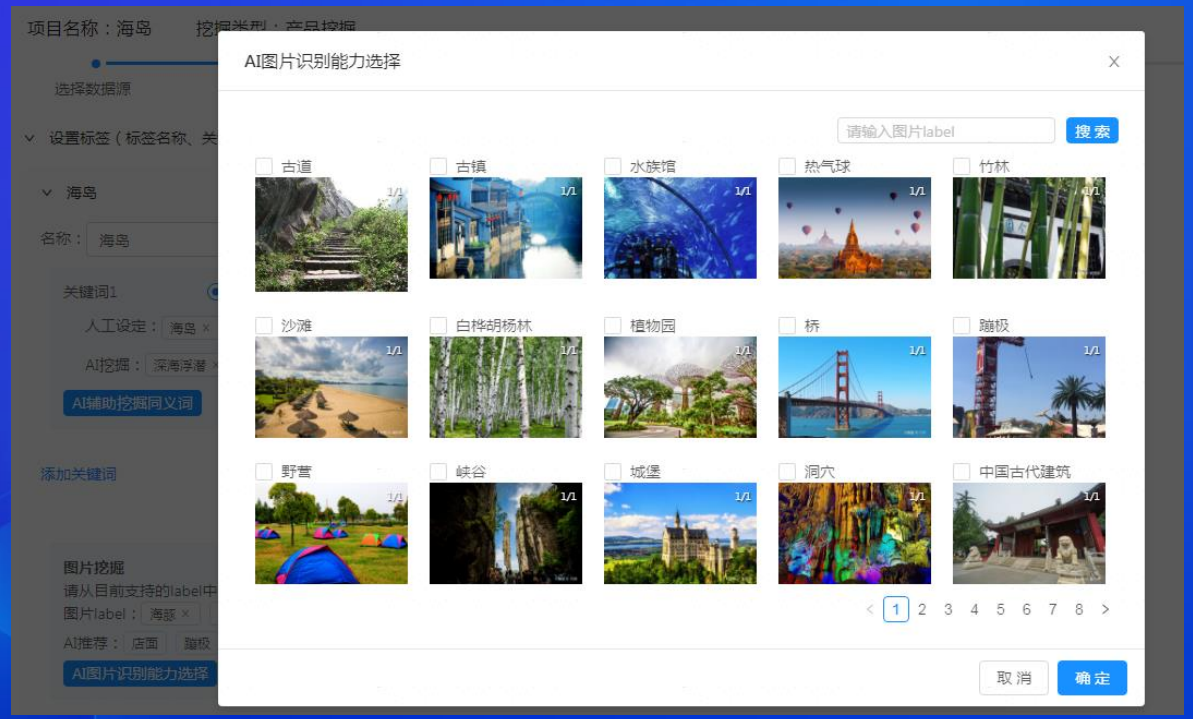


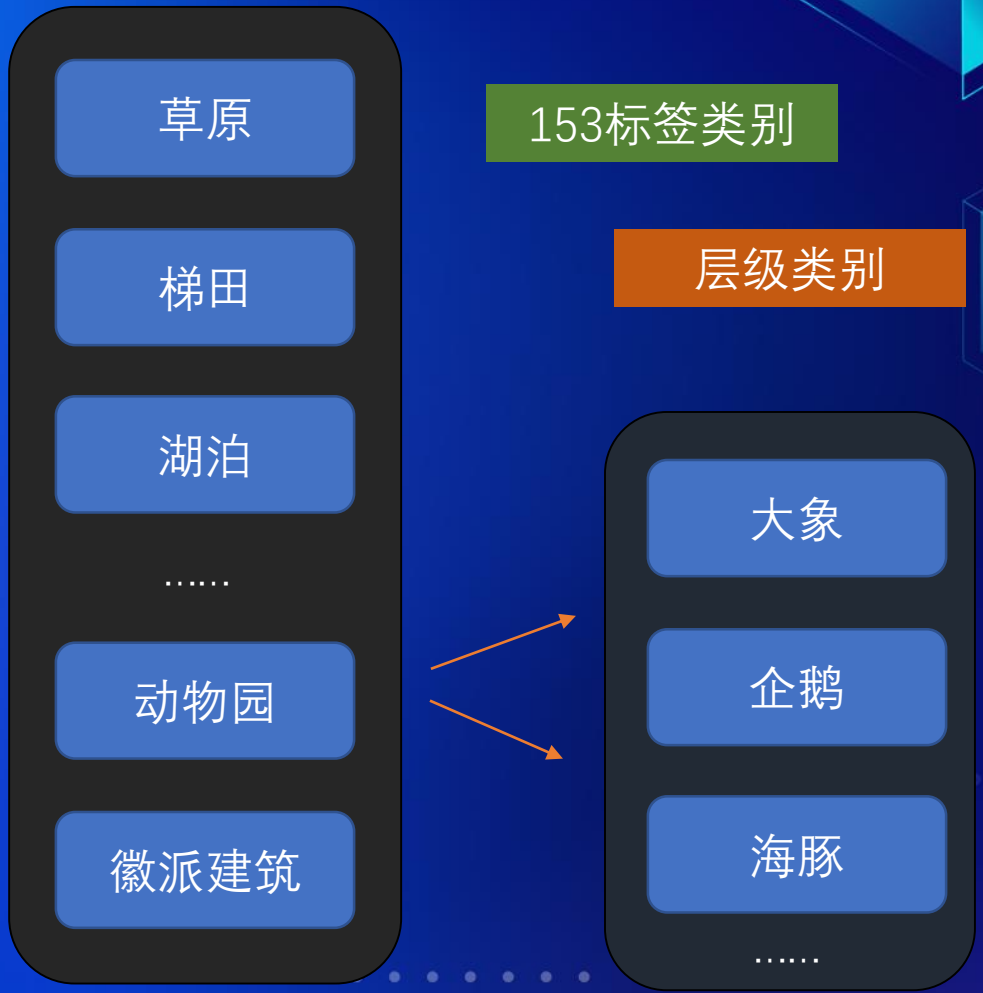


2.3 主题图片挖掘

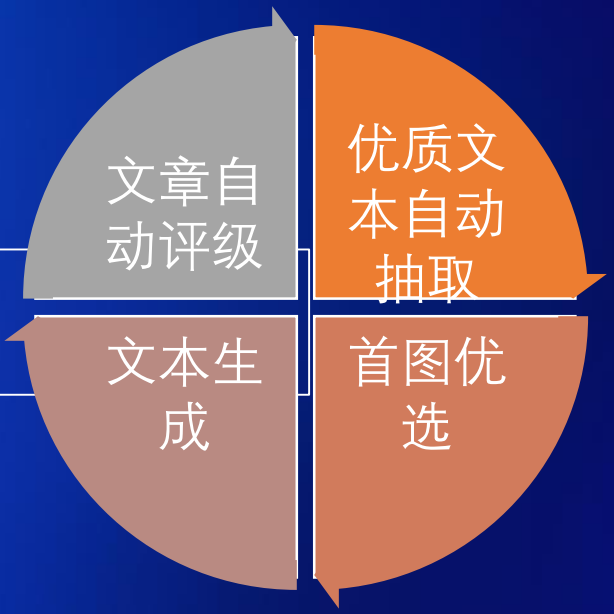
用户自定义主题

手动设置关联





目录



3.1 文章自动评级

选优

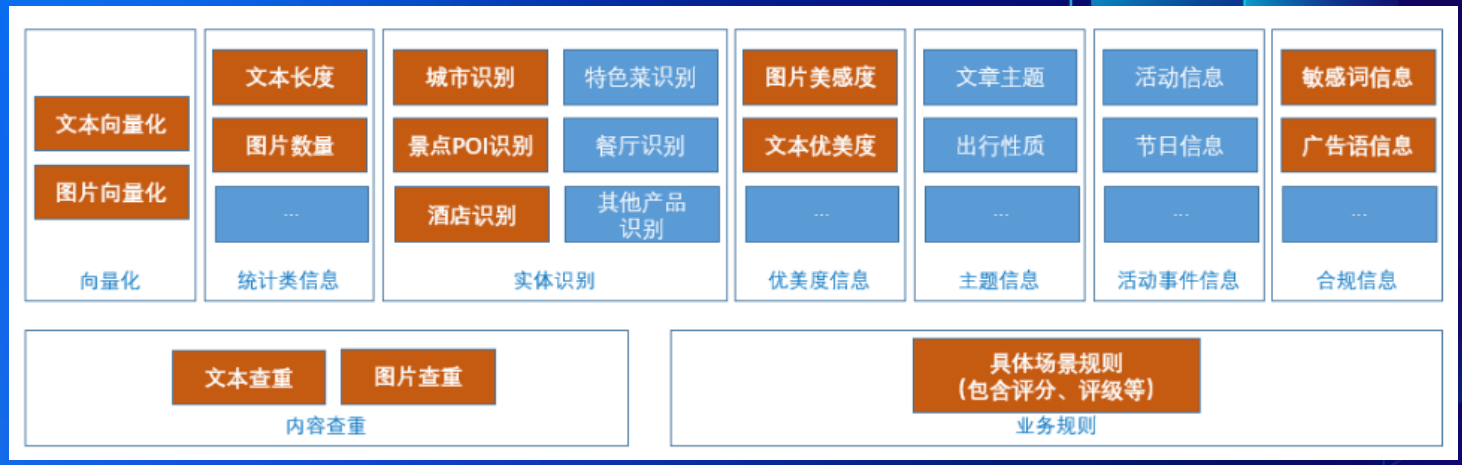
文章内容复杂

标准不一，难以定义

低质筛除

关键点的检测

标准较为清晰，单一



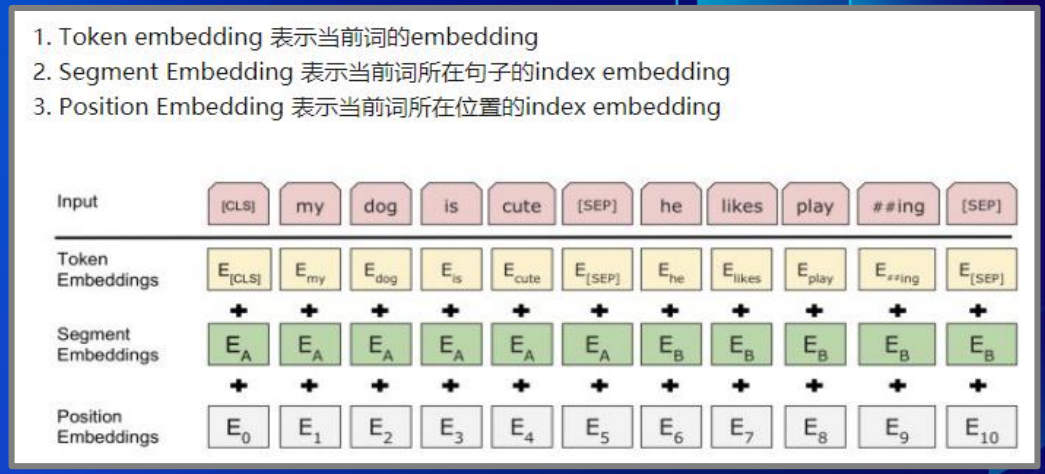
- 敏感内容识别
- 广告识别
- 负面情感识别
- 旅游领域识别

……

3.1 文章自动评级

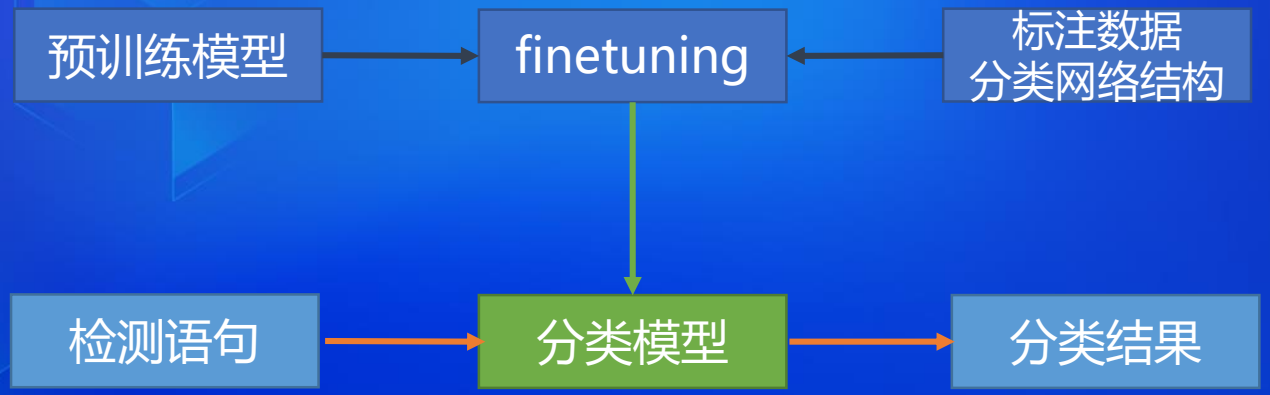
论文：《 **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding** 》

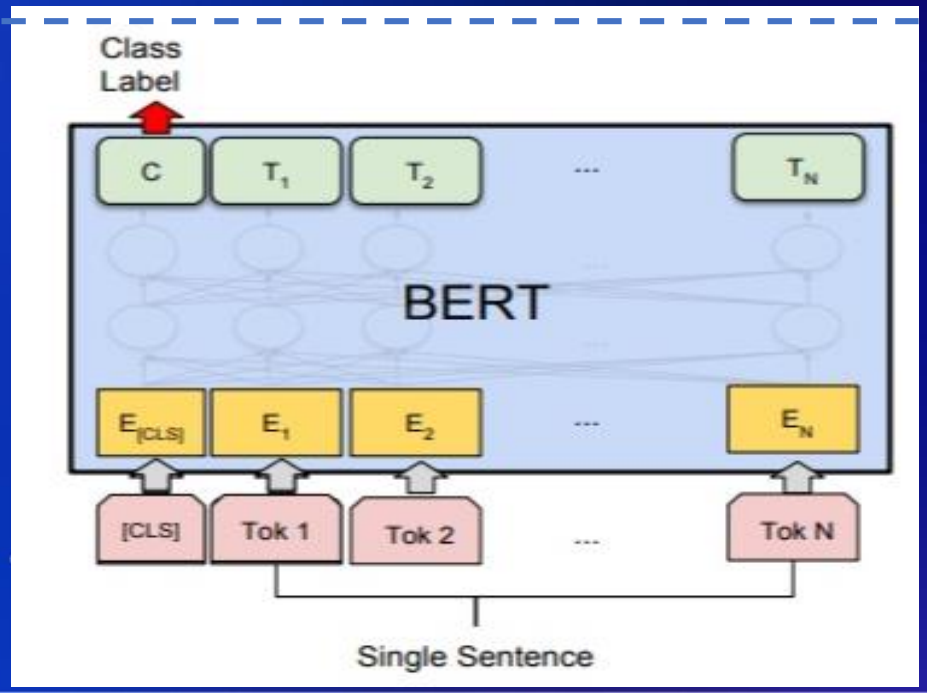
- 用Transformer方法进行特征提取，可同时利用左侧和右侧的词语信息
- Mask（遮挡%15）技术在语言模型上应用
- 能够输出word level 和sentence level
- 下游任务对接方便



使用方法：

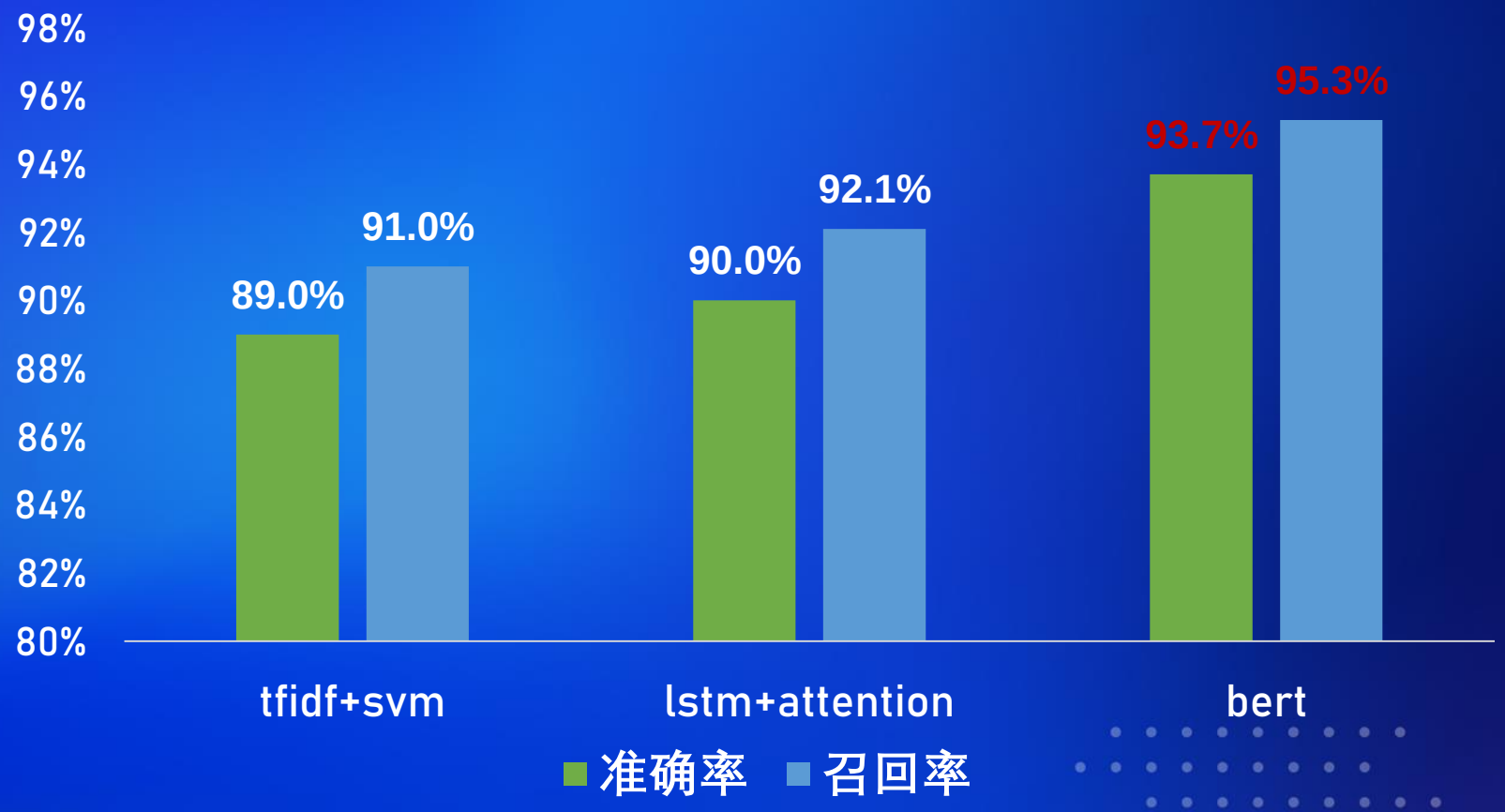
BERTBASE (L=12, H=768, A=12, Total Parameters =110M)





3.1 文章自动评级

情感模型-不同模型效果对比图



3.2 优质文本内容抽取

4.重排召回	模型融合 需求规则	重排、召回
3.结果优化	优美度模型 语义匹配模型	语义匹配 相似度去重 优美度判断
2.内容丰富度衡量	评价维度模型 实体识别 知识图谱 亮点模型	场景维度评价 产品特征识别 语句基础信息
1.基础处理	情感模型 基础处理逻辑	长度筛选 拼写错误 敏感词过滤 情感控制 繁体转简

3.2 优质文本内容抽取

内容丰富度

语句基础信息

- 信息熵
- 词性占比
- 依存句法分析

语句层面

产品特征识别

- 实体识别
- 知识图谱

产品层面

场景维度评价

- 规定各场景（酒店，美食，景点）评价维度
- 建立维度的分类模型

场景层面

3.2 优质文本内容抽取

内容丰富度

词法句法

- **信息熵：**

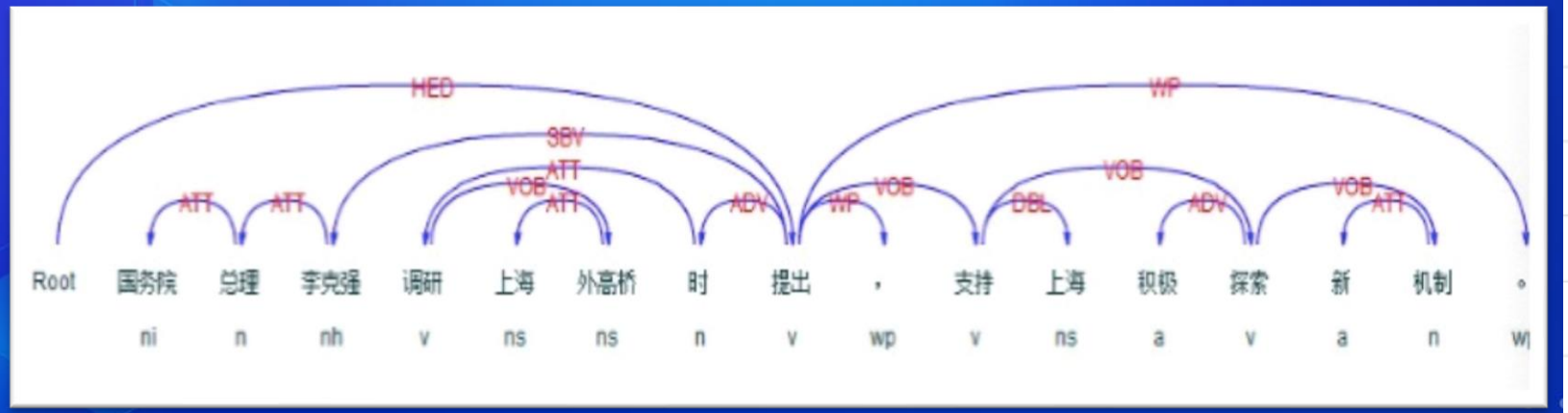
$$H_s = \sum_{i=1}^{n} p_i I_e = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- **词性分析：**

  考虑到推荐理由的结果大概率的同时包含名词和形容词等词性占比。

- **依存句法分析：**

  重点分析某些关系类型的分布，SBV，ATT，VOB等等。



关系类型	Tag	Example
主谓关系	SBV	我送她一束花 (我 <-- 送)
动宾关系	VOB	我送她一束花 (送 --> 花)
间宾关系	IOB	我送她一束花 (送 --> 她)
定中关系	ATT	红苹果 (红 <-- 苹果)
状中结构	ADV	非常美丽 (非常 <-- 美丽)

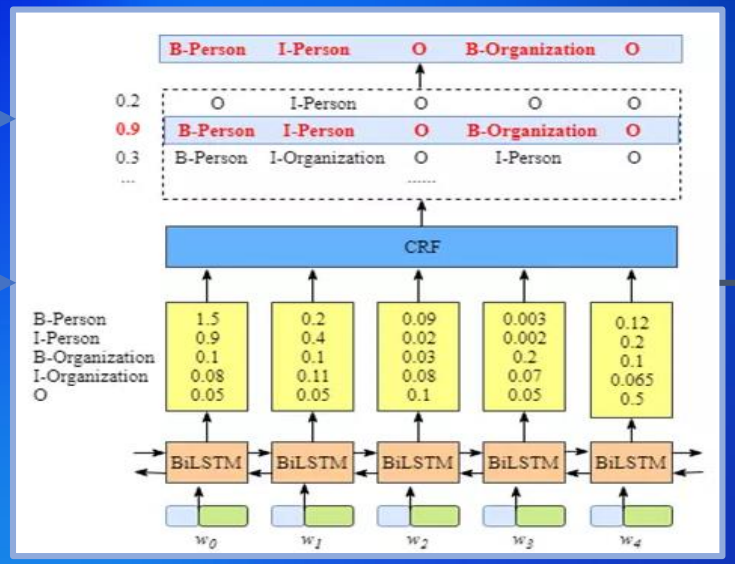
3.2 优质文本内容抽取

信息点
内容丰富度

美食 （菜品，口味等）
酒店 （设施，交通等）
景点 （名称，等）

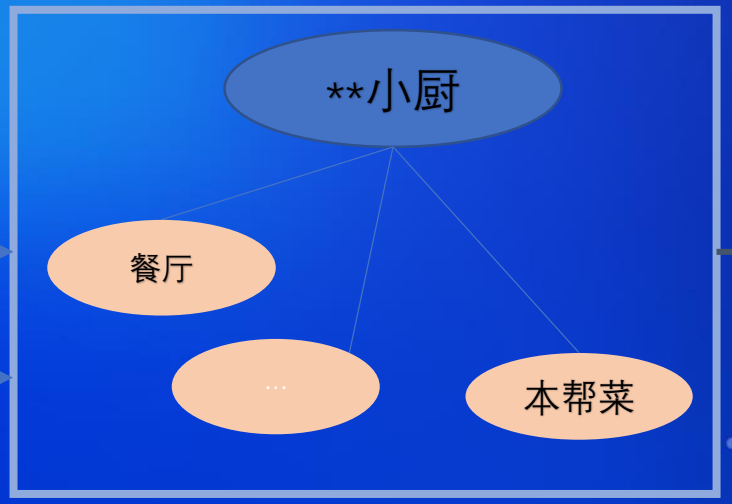
动态提取
实体识别模型

BiLSTM-CRF



识别结果：
干锅花菜，枣泥发糕

**小厨：
枣泥发糕，糕有浓浓的荷叶香。商户详情上说是本帮菜，干锅花菜强烈推荐。



识别结果：
本帮菜

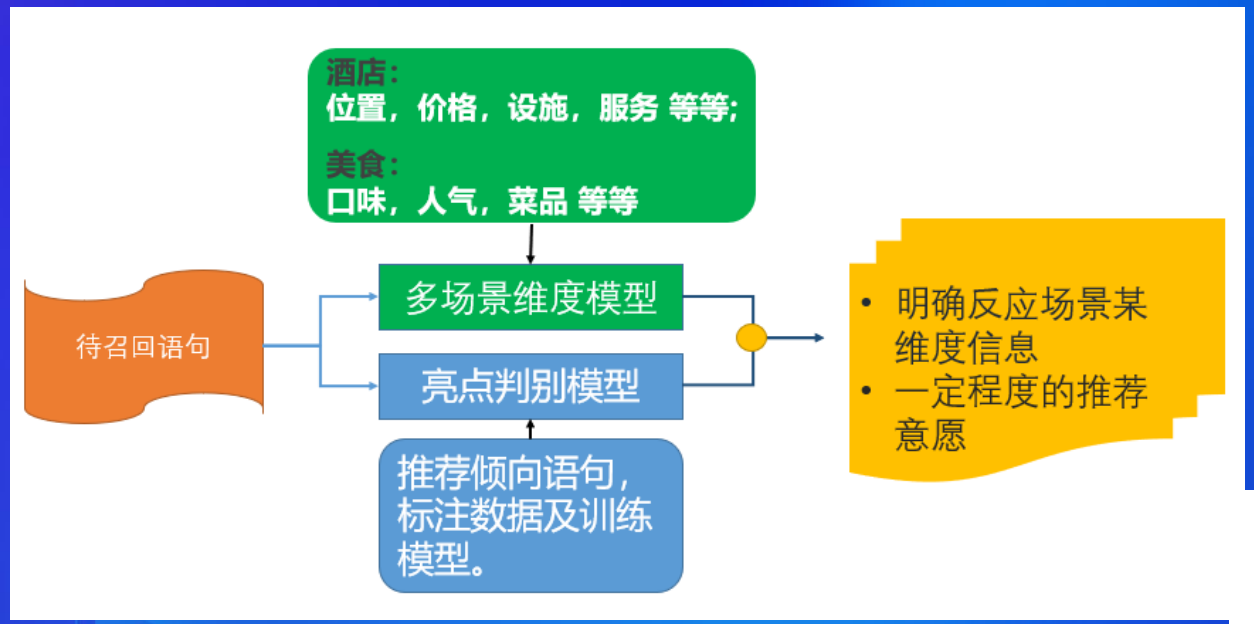
固态提取
知识图谱

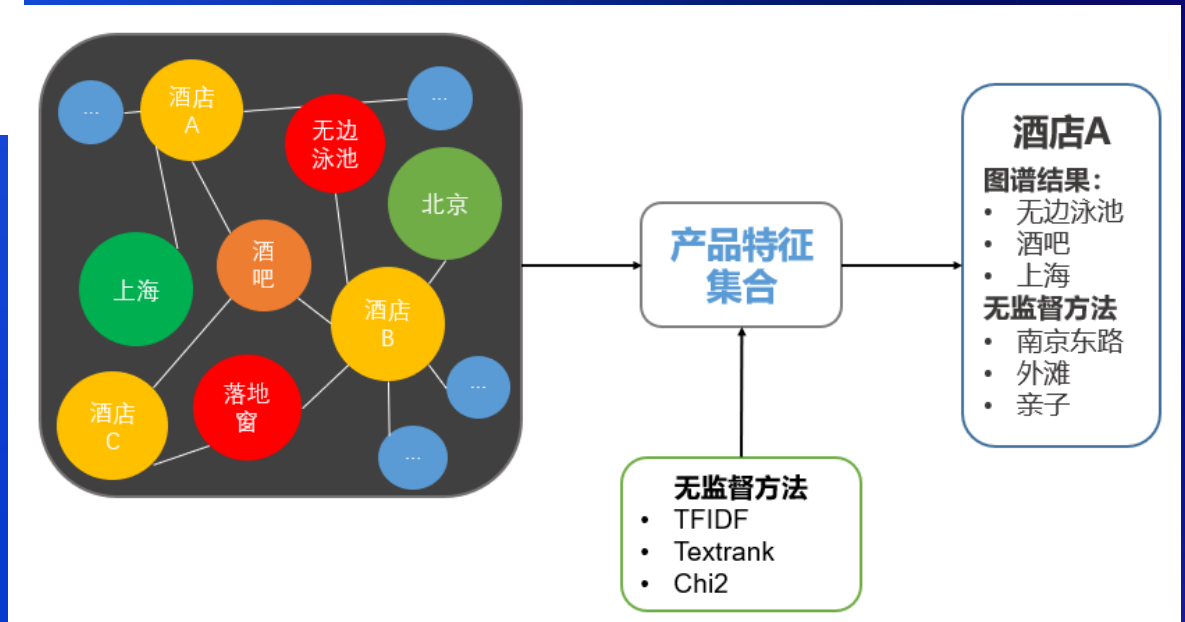
识别结果：
干锅花菜，
枣泥发糕，
本帮菜

3.2 优质文本内容抽取

场景及特色

内容丰富度





3.2 优质文本内容抽取

早餐很好，菜品很好
房间隔音，喜欢沐浴用品
环境优美充满中西方底蕴
早餐丰富，位置很好
早餐品类丰富
地点很好历史悠久

9楼的行政酒廊安静，餐点精致
阳台酒吧是观赏外滩风景的绝佳位置
环境优美充满中西方文化底蕴
下午茶很好，爵士酒吧很喜欢
早餐不错，品类丰富，尤其是西点做得好
一楼的茉莉酒廊值得去

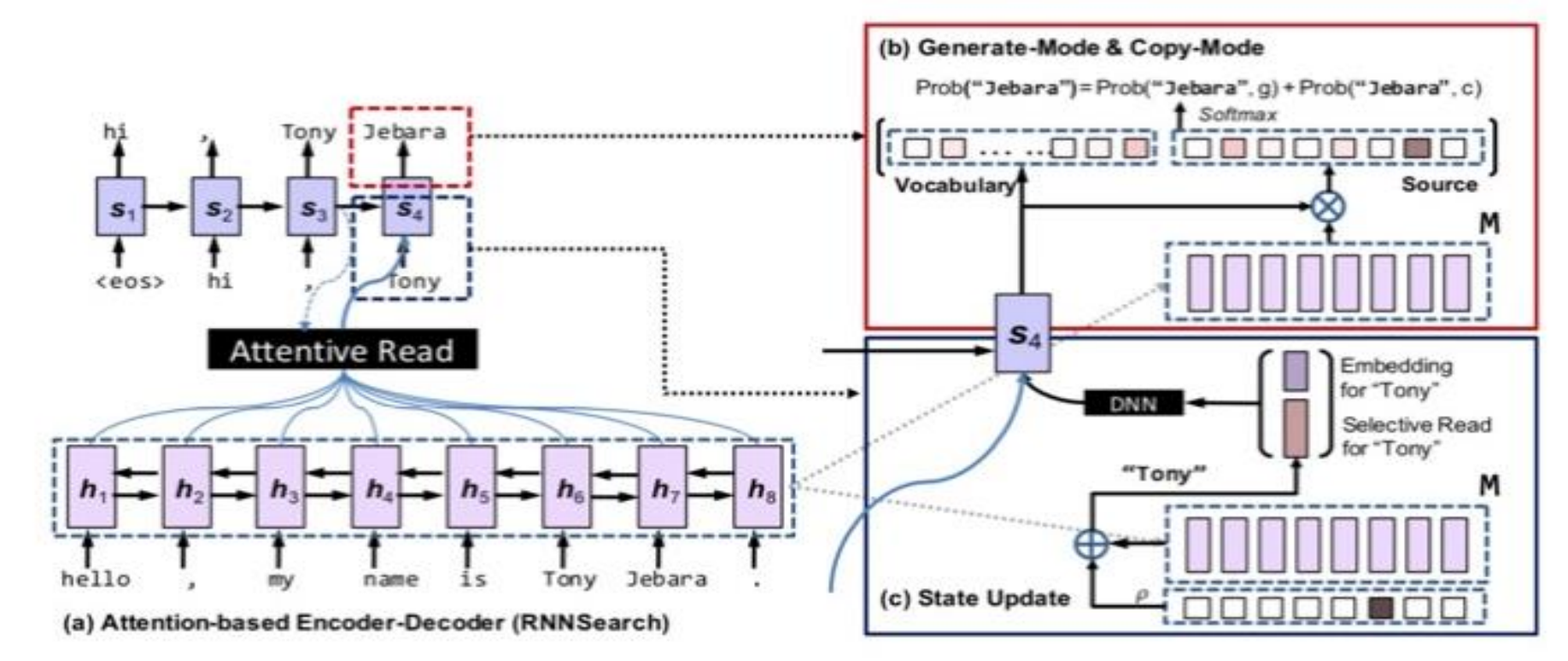
重排前

内容丰富度重排

重排后

**饭店-效果展示

3.3 内容生成



优点：

- Copy机制，较好处理 OOV 问题。
- 场景模式单一的化，能够较好学习到。

缺点：

- 容易出现以偏概全的问题

3.3 内容生成

标题生成

独行在孔庙和国子监博物馆
面试结束后顺便去孔庙、国子监转了一圈放松一下心情走在古香古色的院子里感觉一切忧虑都云消雾散了#孔庙和国子监博物馆[地点]#

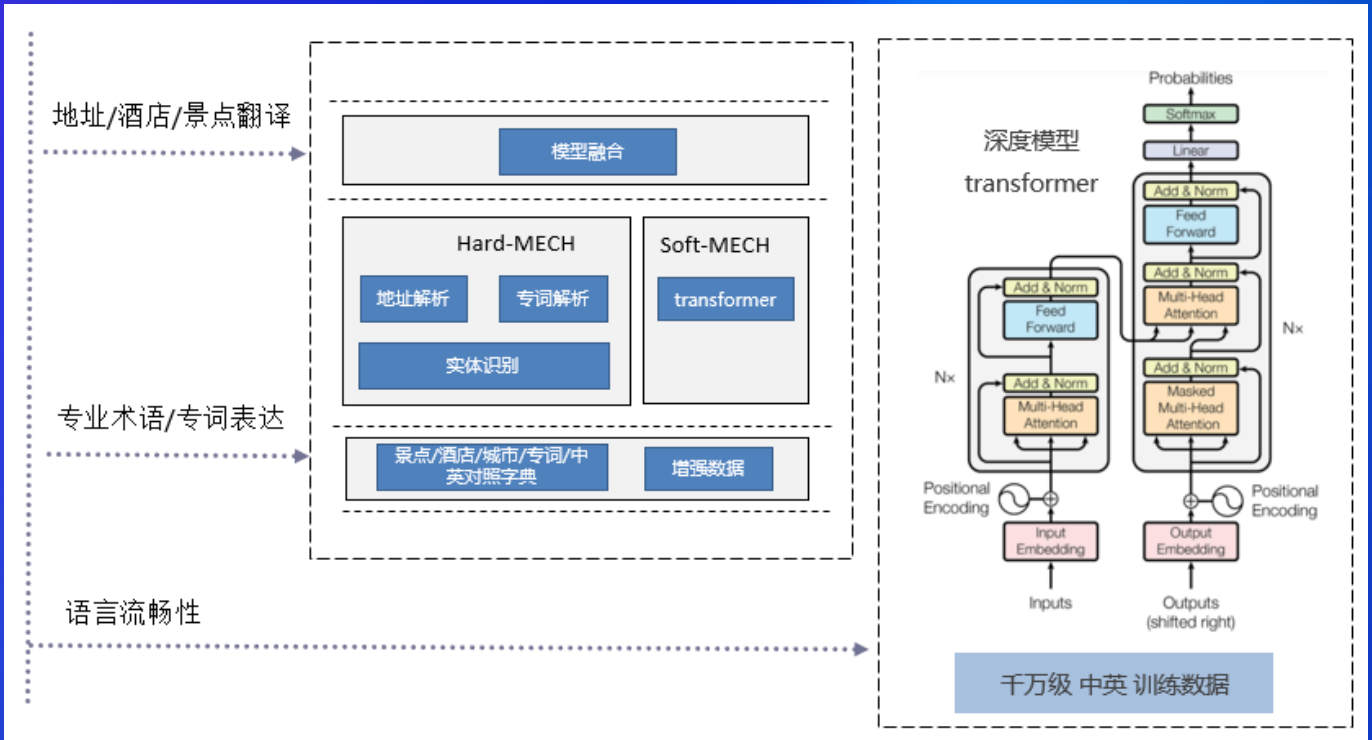
网红地洪崖洞
网红地打卡～洪崖洞洪崖洞重庆美食昨晚去了一趟洪崖洞，人真的超级多，不过景真的很美，还是很值得去的，一边是江景一边是建筑，在灯光的照耀下格外好看。

舌尖上吃日料
又来吃日料啦，这家寿司很棒呢，料足又好吃，那个薯你醉鳗是招牌哦

西安昭陵，你不可错过的地方
西安必打卡的地方：碑林+城墙。碑林博物馆馆藏文物十分丰富，可以说是中国书法艺术的百科全书，我最感兴趣的还是【昭陵六骏】，这是任何一部《中国美术史》必敲黑板的重点。

杭州|象山艺术公社
打卡杭州|象山艺术公社时间匆忙，很多线条，人景融合都没拍，建筑物简洁干练，线条框架感实足，关键是人好少鸭#象山艺术公社[地点]#

3.3 内容生成



优点：
- 借助千万级平行数据，模型更加稳健
- 较为丰富的单向文本数据

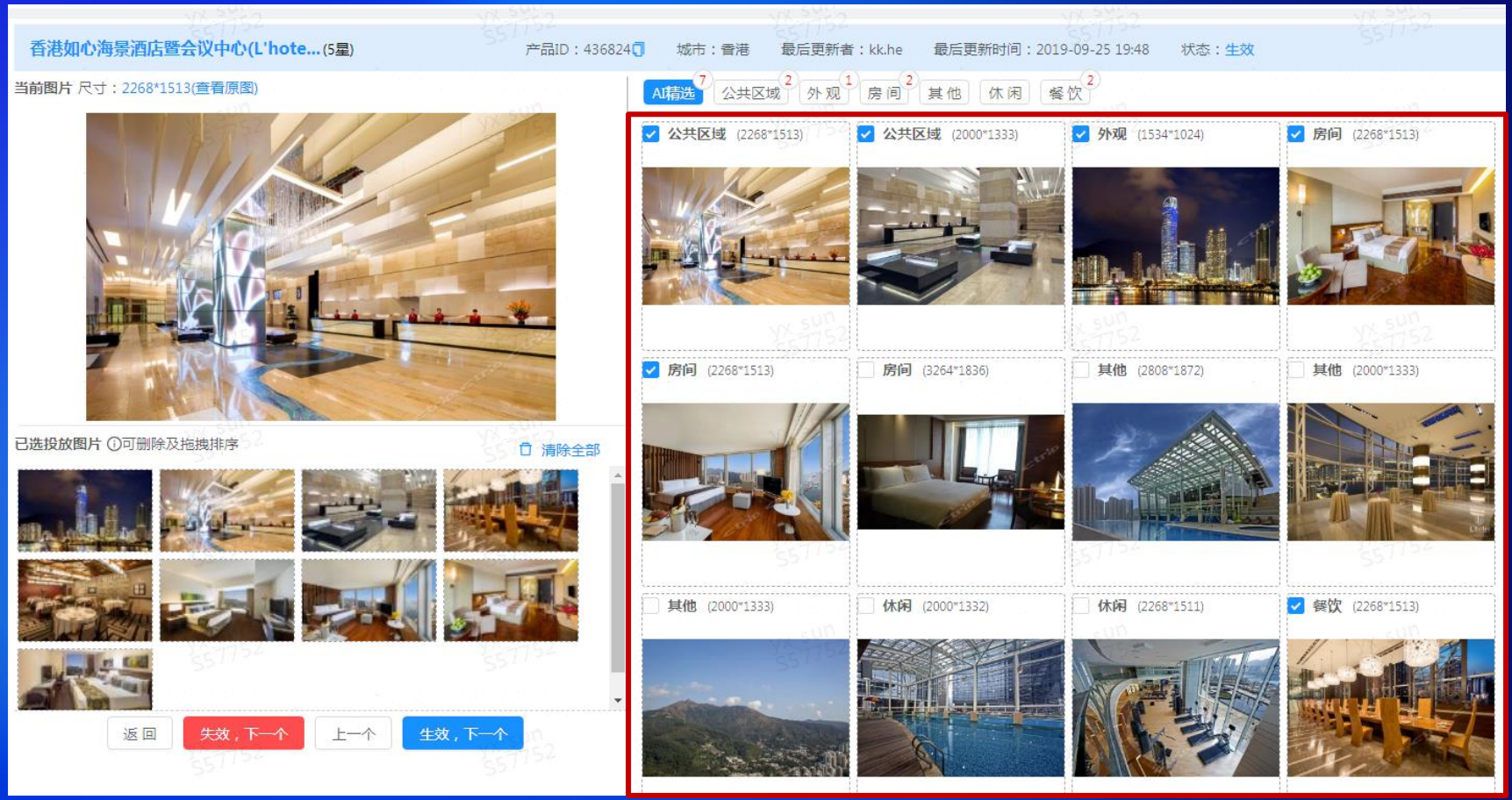
缺点：
- 经过两次翻译，信息双重丢失
- 对实体类型，长尾词语的表达问题比较严重

去平江路和狮子林都很近，走路三分钟就能到，都很方便。
园林景观很美，闹中取静，有苏州特色
很安静，离海信广场很近几分钟走路就可以到。
洗漱用品超级可爱，精心配备了小朋友的拖鞋、牙刷。

改写

靠近平江路和狮子林，步行3分钟，很方便。
园景漂亮，闹中取静，有苏州特色。
非常安静，距离海信广场仅几步之遥。
洗漱用品超级可爱，配有儿童拖鞋和牙刷。

3.4 首图优选


香港如心海景酒店暨会议中心(L'hote...(5星)
产品ID：436824
城市：香港
最后更新者：kk.he
最后更新时间：2019-09-25 19:48
状态：生效
当前图片 尺寸：2268*1513(查看原图)
已选投放图片 可删除及拖拽排序
清除全部
返回
失效，下一个
上一个
生效，下一个
AI精选 7
公共区域 2
外观 1
房间 2
其他
休闲
餐饮 2
公共区域 (2268*1513)
公共区域 (2000*1333)
外观 (1534*1024)
房间 (2268*1513)
房间 (2268*1513)
房间 (3264*1836)
其他 (2808*1872)
其他 (2000*1333)
其他 (2000*1333)
休闲 (2000*1332)
休闲 (2268*1511)
餐饮 (2268*1513)

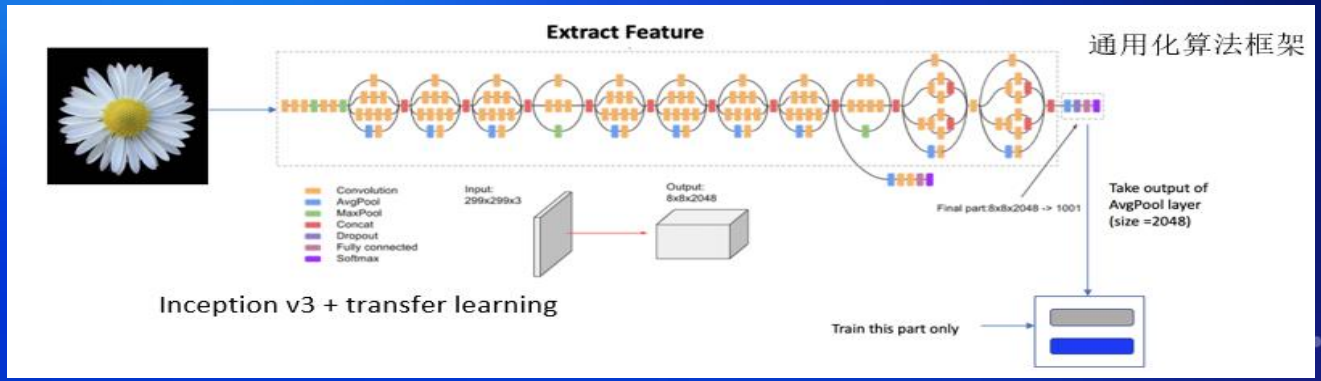
3.4 首图优选

- 多分类：外观 游泳池 大床房 其他 等类别
- 首图场景化类别
- 特定视角，优质度优于覆盖度

1 首图分类

2 美感度

- 二分类 美/丑
- 概率值-图片美感度分值
- 场景样本歧视
- 训练样本的高区分度



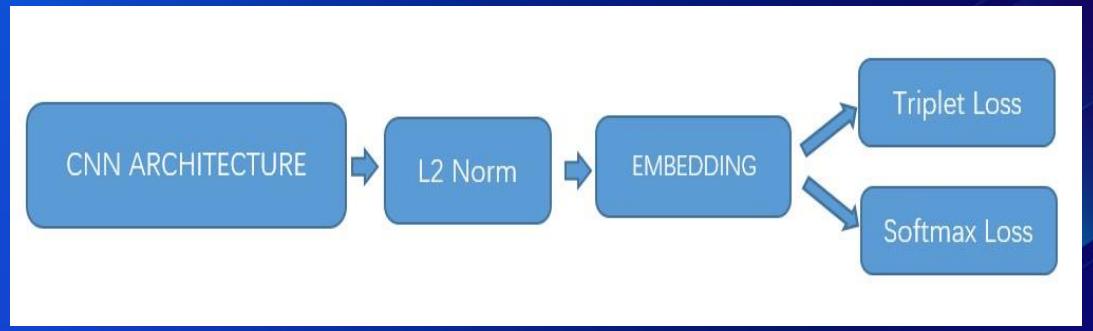
3.4 首图优选

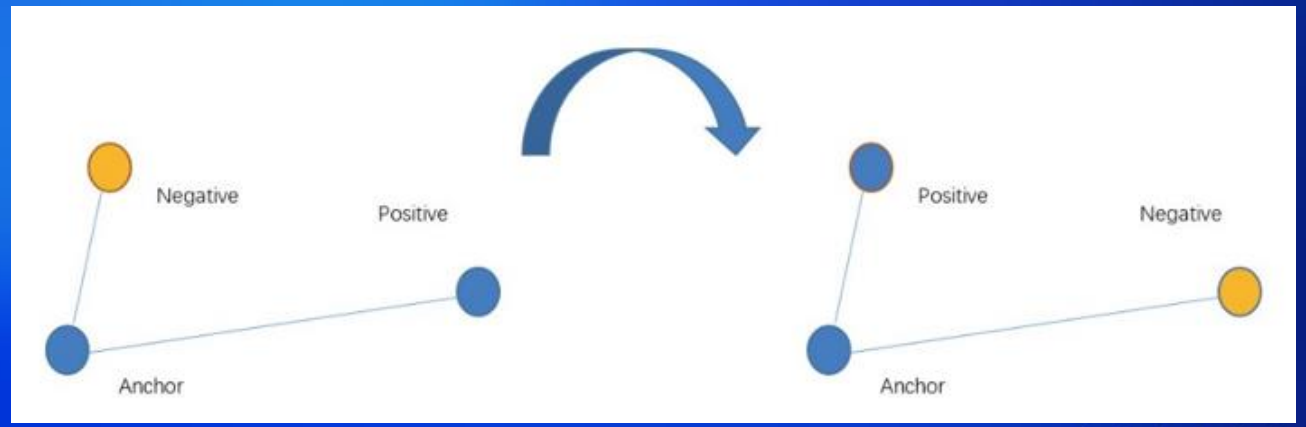
Triplet Loss：最大化异类和同类间的距离差

$$L_t = MAX((|D_A - D_P| + Margin - |D_A - D_N|), 0)$$

3 图片去重

1. 特征更为稳定
2. 特征维数可控，特征冗余较小
3. 鲁棒性强





目录

4 总结

1 结合用户反馈

通过用户反馈有效提高内容化的价值。

2 生成探索

图片自动增强/裁剪/美化文案设计，文本生成进一步探索。

3 内容聚合

内容多样性聚合，更加灵活。

Thanks For Watching