【20210615】长尾query意图识别优化

本页面由王晓敏于2021/12/08迁移自wiki 王晓敏空间

icafe链接： https://console.cloud.baidu-int.com/devops/icafe/issue/map-search-test-1393/show?source=drawer-header

1. 项目背景

目前query意图解析对非属性类的query解析能力空间已经很小（在地图随机query的解析准确率95%+，阿拉丁随机query解析准确率88%），在然而用户属性query的描述是没有边界的，我们目前只解决了其中很小的一部分，然后这部分需求非常影响用户的决策体验，阿拉丁的属性需求主要分布在排行榜、营业范围（菜品、商品、属性），精确属性检索（电话、营业时间、评论、咨询等），要想满足好用户的这类决策需求，需要做好三方面的工作，1、分析用户意图，2、挖掘知识，3、提升召回能力；精确属性的解析已经比较完善了，目前我们重点会放在排行榜和营业范围需求的解析能力提升；

2. 属性检索调研

需求分类	细化需求	query举例	主要行业	影响面	当前效果	理想效果	技术方案
榜单检索	排名topN	1、柳江古镇最出名的美食	景点 美食	10.34%	1、中心点附近检索泛需求词 2、杂质结果	1、准确跳转榜单落地页 2、泛需求列表页结果，插入推荐榜单链接	1、意图解析优化：基于预训练模型的优化，提升榜单以及其他属性需求的解析能力 2、特定属性排序过滤机制建设：优化榜单需求列表页排序排序效果

		2、北京必去景点 3、襄阳最有特色的餐馆					3、列表页插入榜单热词：从召回的poi挂接的榜单进行选取推荐
营业范围	商品 服务 菜品	1、潮阳棉城那里在卖水皮艇 2、浪琴表哪里可以	购物 生活服务 培训 美食	10.53%	1、部分菜品类检索能给出有推荐菜的poi列表 2、部分能命中poi名称的服务或者商品需求展示相关结果 3、杂质结果或者无结果	展示能够提供对应菜品、商品、服务的poi列表	1、意图解析优化：lexparser基于评论的标签同义； 2、tag映射：基于点击等来源的tag标签归一化； 3、普通泛需求模型建设：借助精确需求的召回能力优化属性泛需求召回 4、推荐菜优化：召回优化（更新张冰建设的推荐菜数据），展示优化（店内提供xx菜品（没有推荐数量的菜品），xx人推荐xx菜品（推荐数量较少的菜品），本店topx推荐菜品（推荐数量较多的菜 ）

		维修 3、鸭脖培训 4、炸灌肠					
精确属性	评论	吴家山第四小学怎么样	多数行业	40.40%	展示对应的poi点	1、展示poi的同时，突出展示相应的属性 2、聚合poi的相关咨询链接	1、电话、时间、评论这些数据地图都有，可以突出展示用户描述的属性 2、资讯类数据我们没有，建设成本很高，可以抓取或者和大搜合作，引入相关的链接，挂接在poi上
	电话	于桥派出所电话号码	多数行业	18.76%			
	咨询	1、岭南印象园游	景点 休闲娱乐 教育培训	10.06%			

		玩攻略 2、鼓浪屿珍奇世界馆门票 3、如东欧尚影城今日影讯					
	营业时间	1、恭王府开放时间 2、1号线运营	景点 政府机构 公交/地铁线路	3.66%			

		时间 3、车管所上班时间2018年					

来源	query	city		
~~阿拉于~~	~~厦门晚上必去的地方~~	~~厦门~~	~~厦门0-B-CENTER(0,2)晚上必去2-B-ATTR_TAG(2,6)的地方14-B-ORAL(6,9)~~	~~厦门0-B-CENTER(0,2)晚上必去2-B-ATTR_TAG(2,6)的地方14-B-ORAL(6,9)~~
~~阿拉于~~	~~三亚篮球培训有哪凡家~~	~~三亚~~	~~三亚0-B-CENTER(0,2)篮球培训1-B-KEYWORDS(2,6)有哪凡家14-B-ORAL(6,10)~~	~~三亚[0]-B-CENTER(0,2)篮球培训[1]-B-KEYWORDS(2,6)有哪凡家[14]-B-ORAL(6,10)~~
阿拉丁	河北温泉度假村排名榜	廊坊	河北温泉度假村排名榜0-B-CENTER(0,10)	河北温泉度假村0-B-CENTER(0,7)排名榜14-B-ORAL(7,10)
无结果	上海地跌10线虹桥到五角场是一条线吗?	北京	上海0-B-CENTER(0,2)地跌10线虹桥到五角场是一条线1-B-KEYWORDS(2,17)吗[UNK]14-B-ORAL(17,19)	上海0-B-CENTER(0,2)地跌10线虹桥1-B-KEYWORDS(2,9)到14-B-ORAL(9,10)五角场1-B-KEYWORDS(10,13)是一条线吗？14-B-ORAL(13,19)
~~无结果~~	~~赤水源镇初级~~	~~金华~~	~~赤水源镇初级中学快递公司0-B-CENTER(0,12)~~	~~赤水源镇初级中学快递公司[0]-B-CENTER(0,12)~~

	~~中学快递公司~~			
中长尾	附近的网红手抓羊内	石家庄市	附近2-B-ATTR_TAG(0,2)的14-B-ORAL(2,3)网红手抓羊内2-B-ATTR_TAG(3,9)	附近2-B-ATTR_TAG(0,2)的14-B-ORAL(2,3)网红手抓羊内1-B-KEYWORDS(3,9)

在附近找顺丰/在附近找美食		在附近找顺丰1-B-KEYWORDS(0,6) 在附近找美食1-B-KEYWORDS(0,6)	在附近找14-B-ORAL(0,4)顺丰1-B-KEYWORDS(4,6) 在附近找14-B-ORAL(0,4)美食1-B-KEYWORDS(4,6)
中国石油(服务区西侧) 北京市海淀区		中国石油(服务区0-B-CENTER(0,8)西侧14-B-ORAL(8,10))[UNK]北京市海淀区1-B-KEYWORDS(10,18)	中国石油(服务区0-B-CENTER(0,8)西侧)[UNK]14-B-ORAL(8,12)北京市海淀区1-B-KEYWORDS(12,18)
人均100的料理		人均1000-B-CENTER(0,5)的14-B-ORAL(5,6)料理0-B-CENTER(6,8)	人均100的料理[0]-B-CENTER(0,8)
山东明至财务咨询有限公司		山东明9-B-ORIGIN(0,3)至14-B-ORAL(3,4)财务咨询有限公司10-B-DESTINATION(4,12)	山东明至财务咨询有限公司0-B-CENTER(0,12)
望京西路53号楼1至2层107室		望京西路53号楼19-B-ORIGIN(0,9)至2层107室0-I-CENTER(9,16)	望京西路53号楼1至2层107室0-B-CENTER(0,16)
莫莉幻想		莫莉幻0-B-CENTER(0,3)想14-B-ORAL(3,4)	莫莉幻想0-B-CENTER(0,4)
城市绿心森林		城市绿心森林公园冬9-B-ORIGIN(0,9)至14-B-ORAL(9,10)数	城市绿心森林公园冬9-B-ORIGIN(0,9)至[illegible]0-I-CENTER(9,12)

公园冬至数九		九10-B-DESTINATION(10,12)	
成都西部中西医结合医院		成都0-B-CENTER(0,2)西部14-B-ORAL(2,4)中西医结合医院0-B-CENTER(4,11)	成都西部中西医结合医院0-B-CENTER(0,11)
中共浙江省工委机关旧址		中共浙江省工委机关旧0-B-CENTER(0,10)址14-B-ORAL(10,11)	中共浙江省工委机关旧址0-B-CENTER(0,11)
内山书店		内14-B-ORAL(0,1)山书店0-B-CENTER(1,2)	内山书店0-B-CENTER(0,4)

3. 项目目标

基于polaris，依存句法分析提升属性query解析能力，重点优化排行榜和营业范围的解析准确率；

4. 设计方案

4.1 样本数据构建

4.1.1 teacher模型样本数据：

1、人工标注样本：（23000+）

bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/map/wangxiaomin/QU/attr_intent/all_samples/manual_sample/**manual_sample（23238）（*50）**

口语化query：manual_all.txt（18180）

随机query：manual_1w.txt（5063）

2、poi名称：（397294）

包含口语词/属性词/起终点的poi名称，整体标注为center，避免poi名称识别错误

a、包含oral/attr/destation的名称标注为center

b、包含keywords的名称使用qu进行需求分析，其中精确需求标注为center

bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/map/wangxiaomin/QU/attr_intent/all_samples/poi_name_sample/**poi_name_sample（1165976）(*5)**

c、poi_type、brand标注为keywords

bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/map/wangxiaomin/QU/attr_intent/poi_type_sample/**poi_type_sample（111483）(*10)**

d、包含destation的名称标注为center作为模型样本，同时放在poi_no_route词表中

bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/disk1/wangxiaomin/QU/attr_intent/all_samples/poi_name_sample/poi_name_res/poi_name_route（21702）

3、属性/口语化query标注样本优化（200w）

a、属性/口语化样本使用意图模型标注后，使用基于依存语法分析的标注优化策略进行修正：128485（6.42%）

map@bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/map/wangxiaomin/QU/attr_intent/all_samples/oral_query_sample/oral_query_sample_modify_200w

b、榜单需求识别能力提升：排序属性识别能力提升

query中识别出的属性词/口语词，调用lexparser，如果匹配到describe的overall_rating，识别为排行榜需求，填充sort字段

4、阿拉丁、地图自动化样本（1000w+1000w），删除其中的poi名称，done

map@bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/map/wangxiaomin/QU/attr_intent/all_samples/

a、地图样本：map_query_sample

非属性/口语化样本：map_query_sample_nooral（9919988）

属性/口语化样本：map_query_sample_oral_modify（117703）其中5340（4.54%）条样本进行了修正

b、阿拉丁样本：aladin_query_sample

非属性/口语化样本：aladin_query_sample_nooral（9125347）

属性/口语化样本：aladin_query_sample_oral_modify（851284）其中81902（9.62%）条样本进行了修正

5、后置处理：针对3、4样本

a、删除既标注了poi成分,又标注了路线成分

b、修正:如果成分标注位ORAL

map@bjyz-mapse-xj-bs013.bjyz.baidu.com:/home/map/wangxiaomin/QU/attr_intent/all_samples/oral_map_aladin_query_sample/**oral_map_aladin_sample_intent_new_0709（21625262）**

badcase：

1、和、与、 -道路 错误识别为ORAL（样本问题）

2、ktv，风景区识别错误

3、标注错误，有I开头的错误标注（可能是网络结构问题）

4、一号线、三号线识别错误，棒棰岛（错误品牌）

5、人工样本中，没有区分连续多个属性tag，进行修正

4.1.2 student模型样本数据：

蒸馏数据:

map@yq01-ns-map26-a3e68.yq01.baidu.com:/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/student/data_ori

原始数据：

map@yq01-ns-map26-a3e68.yq01.baidu.com:/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/auto_manual_0709/data_ori

4.1.3 模型训练数据、配置

机器：yq01-ns-map26-a3e68.yq01.baidu.com

模型	
Polaris+ bigru-crf	
训练数据	/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/auto_manual_0709/data_ori
测试数据	/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/auto_manual_0709/test_data
数据处理	/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/auto_manual_0709/cmd.sh
模型配置	/home/map/rd/wangxiaomin/ernie-poi/paddle-frame/conf/ernie_finetune/intent/ernie_finetune.local.conf
模型数据	/home/map/rd/wangxiaomin/ernie-poi_data/intent/model/auto_manual_0709/job-0bb60ebea454db8c/checkpoint_400000
bigru-crf	
训练数据	/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/student/da ori

数据处理	/home/map/rd/wangxiaomin/ernie-poi_data/intent/data/train_data/student/cmd.sh
模型配置	/home/map/rd/wangxiaomin/paddle_frame/baidu/mapsearch/paddle-frame/conf/query_intent/query_intent_0714.local.conf
模型数据	/home/map/rd/wangxiaomin/paddle_frame/baidu/mapsearch/tmp/model/query_intent/auto_manual_0709_bigru-crf_40w/checkpoint_550000

4.2 模型训练

step1、大规模数据基于polaris finetune

1、地图大框query：1000w

2、阿拉丁原始query：1000w

3、口语化专项query：200w

step2、高精度样本finetune

2w人工标注的属性/口语化样本

构建3w地图大框随机样本

step3、模型蒸馏

4.3 在线应用

lexparser模型和nn标注模型融合：

现状：优先选择lexparser模型的结果，无结果的情况下选择nn标注模型

优化：

路线类优先使用nn模型结果

poi类型在lexparser覆盖不足0.7的情况下，使用nn模型结果

5. 模型效果：

	样本	jobid	step	map随机500	阿拉丁随机500	口语化随机500	all

线上 bigru-crf				P: 0.887960 R: 0.876238 F1: 0.882060	P: 0.874576 R: 0.836305 F1: 0.855012	P: 0.851628 R: 0.799006 F1: 0.824478	P: 0.865684 R: 0.825542 F1: 0.845136
polaris-crf	auto_manual_0709	job-0bb60e862f0884ea	45w	~~P: 0.924370 R: 0.938567 F1: 0.931414~~	~~P: 0.868552 R: 0.848780 F1: 0.858553~~	~~P: 0.853362 R: 0.848506 F1: 0.850927~~	
			65w	~~P: 0.932546 R: 0.943686 F1: 0.938083~~	~~P: 0.866667 R: 0.845528 F1: 0.855967~~	~~P: 0.851003 R: 0.844950 F1: 0.847966~~	
			105w	P: 0.952462 R: 0.9573	P: 0.880000 R: 0.858537 F1: 0.869136	P: 0.860650 R: 0.847795 F1: 0.854174	P: 0.886169 R: 0.874052

				38 F1: 0.954894			F1: 0.880525
			115w	~~P: 0.944257~~ ~~R: 0.953925~~ ~~F1: 0.949066~~	~~P: 0.864548~~ ~~R: 0.840650~~ ~~F1: 0.852432~~	~~P: 0.852011~~ ~~R: 0.843528~~ ~~F1: 0.847748~~	
			135w	~~P: 0.954315~~ ~~R: 0.962457~~ ~~F1: 0.958369~~	~~P: 0.876254~~ ~~R: 0.852033~~ ~~F1: 0.863974~~	~~P: 0.853448~~ ~~R: 0.844950~~ ~~F1: 0.849178~~	~~P: 0.881829~~ ~~R: 0.873034~~ ~~F1: 0.877409~~
polaris-bigru-crf	auto_manual_0709	job-0bb60ebea454db8c	40w	P: 0.935593 R: 0.941980 F1: 0.938776	P: 0.898164 R: 0.874797 F1: 0.886326	P: 0.868156 R: 0.857041 F1: 0.862563	P: 0.890570 R: 0.880322 F1: 0.885417

			60w	P: 0.935484 R: 0.940273 F1: 0.937872	P: 0.890000 R: 0.868293 F1: 0.879012	P: 0.861771 R: 0.851351 F1: 0.856530	
			90w	P: 0.944162 R: 0.952218 F1: 0.948173	P: 0.889447 R: 0.863415 F1: 0.876238	P: 0.863211 R: 0.852774 F1: 0.857961	P: 0.887854 R: 0.877637 F1: 0.882716
			100w	P: 0.944257 R: 0.953925 F1: 0.949066	P: 0.888147 R: 0.865041 F1: 0.876442	P: 0.861871 R: 0.852063 F1: 0.856938	P: 0.886866 R: 0.878021 F1: 0.882421
bigru-crf			55w	P(0.961667) R(0.952145)	P(0.885470) R(0.839546)	P(0.871832) R(0.855114) F1(0.863392)	P(0.895947) R(0.[illegible]31 2)

				F1(0.956882)	F1(0.861897)		F1(0.884741)

6. 参考资料

邻近词扩展：邻近指的是语义上的临近，即越是经常共同出现的词之间越相似，就为成为邻近词

邻近词扩展--nlpc-wordemb-neighbor

命名实体识别

nlpc-wordner

依存句法分析

依存句法分析--nlpc-depparser

词性分析

分词+词性标注--nlpc-lextag_new

词性标签定义

标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间