北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY

语言资源高精尖创新中心
Beijing Advanced Innovation Center for Language Resources

基于搜索数据的生僻字知识库构建和应用

作者：刘廷超、王雨、饶高琦、杨兆勇、荀恩东

报告人：刘廷超

2024.11.16

第一届国际多语种智能信息处理大会
(IMLIP2024)

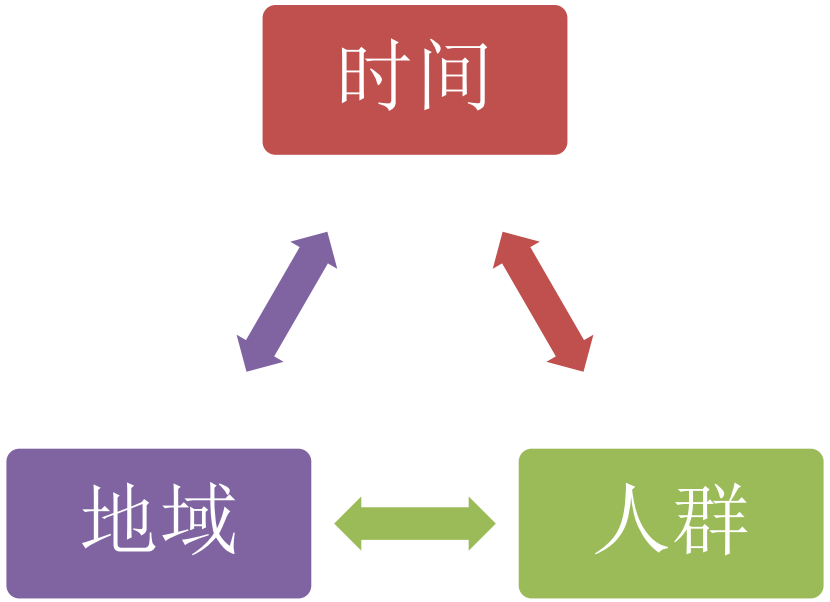
01 生僻字现象

- 生僻字，又称冷僻字，指不常见的或不熟悉的文字
  - 分布广泛，影响交流
  - 值得关注的语用现象

图片来源：https://www.zcool.com.cn/work/ZMzI4MzM4MzY=.html

01 生僻字现象

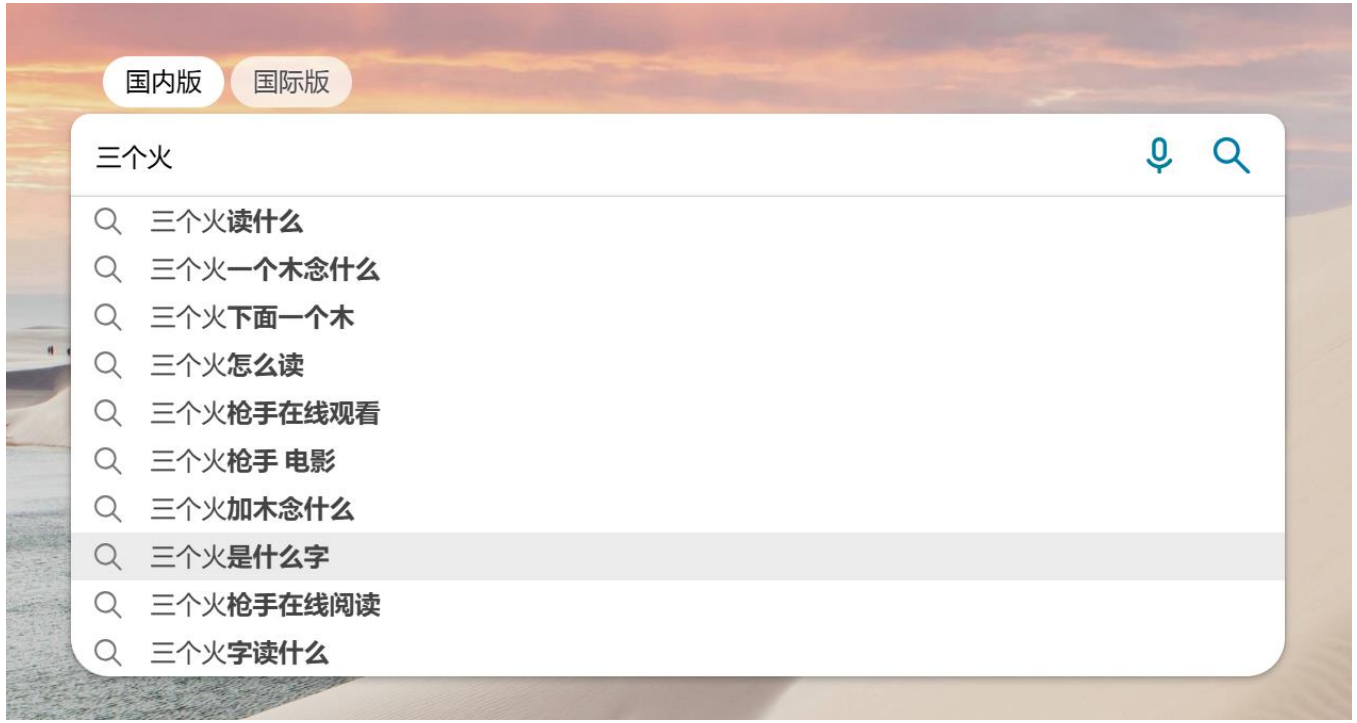
- **生僻字研究困境**
  - **多种环境因素相关（人名、地名、方言、网络用语）**
  - **缺少有效的统计方法**



产生生僻字的主要因素

02 搜索引擎与生僻字

- 当遇到生僻字，用户一般使用搜索引擎查询
- 搜索日志可以反映出一个字的整体认知情况



使用搜索引擎搜索生僻字截图

02 搜索引擎与生僻字

- 使用2015年7月-2016年10月的搜索日志数据
- 统计包含后缀的搜索字符串：“字读什么”“字念什么”“是什么字”“念什么字”“念什么”“读什么”

搜索字符串	搜索次数
爱奇艺	327951
盗墓笔记	285674
花千骨	233485
奔跑吧兄弟第二季	194550
…	…
胥怎么读	123
…	…

搜索日志数据示例

02 搜索引擎与生僻字

- **生僻字搜索方式分类**
  - **分为直接生僻字搜索和拆字生僻字搜索两类**

搜索字符串	次数
灏读什么	153
睢怎么读	112
闰字读什么	99
阚怎么读	68
臻读什么	53
靳字读什么	46
斐读什么	44
窦字怎么读	39

直接生僻字搜索

搜索字符串	次数
三个土读什么	160
四个火字读什么	156
三个牛念什么	124
两个方一个土念什么	95
三个水读什么	82
三个火读什么	76
三个火加一个木是什么字	68
或加两撇念什么	65

拆字生僻字搜索

03 拆字生僻字搜索处理

- **构建拆字搜索的规则**
  - **将拆字搜索字符串转换为汉字部件列表**

规则	搜索字符串	转换结果
数量词	一个立一个羽 三个火加一个木	立羽 火火火木
部首	三点水一个女 一个单人旁一个吉	氵 女 亻 吉
结构	斌下面一个贝 左边一个革右边一个斤	上下结构 左右结构
加	三个火加一个木 上面一个加下面一个可	删除 保留
修饰语	男女男合起来 更生组成的	删除 删除
去除	埠去掉土 演去掉三点水	埠|土 演|氵

拆字搜索规则应用示例

03 拆字生僻字搜索处理

- 使用搜索规则，将拆字搜索字符串转换为部件列表

序号	搜索字符串	部件	结构
1	日下面加个立	日立	上下结构
2	上边日下边立	日立	上下结构
3	日下面立	日立	上下结构
…	…	…	…
26	一个日字一个立	日立	
27	日下面加立	日立	上下结构
28	一日一立	日立	
29	上面一个日字下面一个立	日立	上下结构

搜索字符串转换部件列表示例

03 拆字生僻字搜索处理

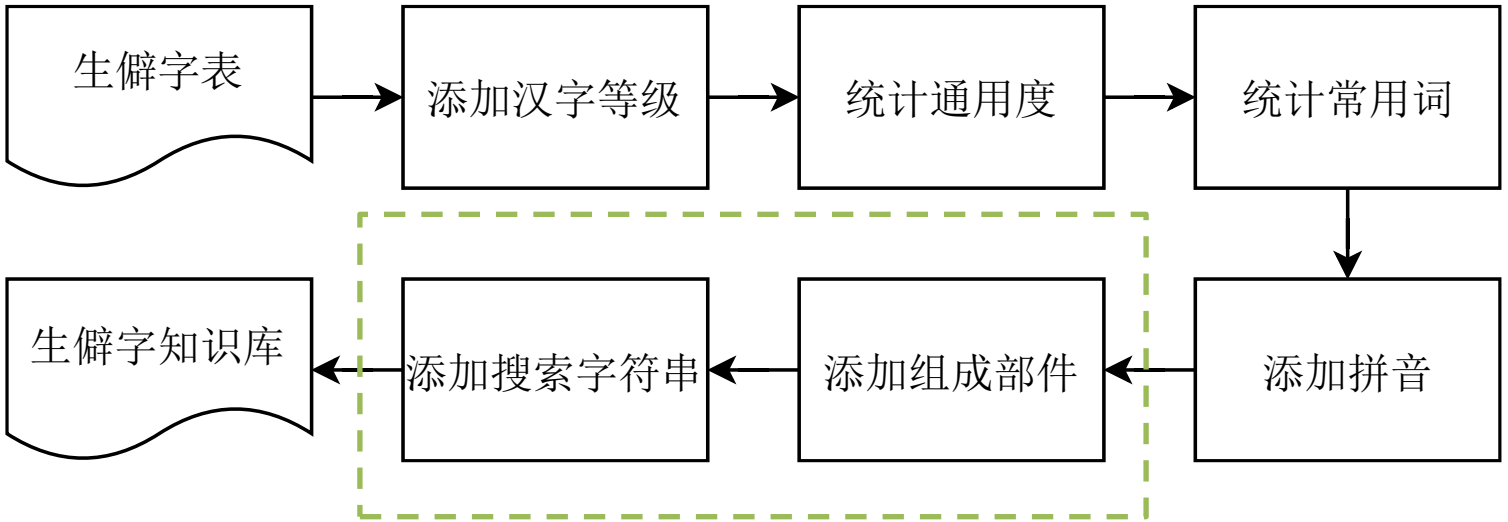
- 将汉字部件列表还原为对应汉字
  - 使用《汉字拆字字典》数据
    - https://github.com/kfcd/chaizi

汉字	拆分1	拆分2	拆分3
奤	大明	大日月	
奮	大周		
奫	大朋	大月月	
燚	炎炎	火火火火	
蕊	艹惢	艹心心心	
灪	氵 顯	氵 景页	氵 日京页

汉字拆字字典数据示例

04 构建生僻字知识库

- 生僻字构建流程
  - 生僻字知识库的GIT地址
    - https://github.com/paineliu/RCKB.git



生僻字知识库构建流程

04 构建生僻字知识库

- 词语通用度
  - 兼顾词语的分布率和频率两个方面，把两者有机地结合起来

$$T = \frac{(\sqrt{n_1} + \sqrt{n_2} + \cdots + \sqrt{n_k})^2}{k}$$

词语通用度计算公式

04 构建生僻字知识库

- **直接搜索生僻字通用度**

阜	3483	阚	3312	丰	2451	靳	2310
衢	1995	羽	1783	翟	1703	鄞	1658
赣	1597	邹	1558	晟	1496	祁	1436
华	1417	虞	1417	闫	1391	濮	1390
沂	1374	睢	1254	绥	1233	力	1194
青	1190	肇	1185	斤	1183	忻	1162
页	1065	郴	1006	堃	1004	煜	986
暨	980	黔	966	日	958	缪	934
毓	921	筱	910	婺	897	綦	890
岑	874	文	846	湛	842	兖	829
昱	821	于	820	茜	811	行	809
芮	806	樊	802	覃	789	龚	787
卜	777	荥	775				

直接搜索生僻字TOP50列表

04 构建生僻字知识库

- **拆字搜索生僻字通用度**

垚	130038	燚	71547	堃	60613	焱	56376
昶	53550	昱	45034	犇	43931	赟	42286
翊	37987	燊	36919	靳	35119	晟	32497
鸟	28081	仝	28033	淼	26516	旻	25866
囡	23602	婧	22678	煜	22353	幢	21515
砼	21167	臻	20346	喆	20150	歆	19575
夯	18582	翀	17948	頔	17582	骘	17039
砻	16962	圩	16894	艮	16408	泵	15813
彧	14935	淼	14761	斛	14485	岑	14196
黔	14009	粿	13218	昕	13148	垦	13018
嬰	12839	柘	12565	灏	12485	龘	11708
阚	11553	怼	11448	颢	11349	聿	10995
珩	10863	梓	10768				

拆字搜索生僻字TOP50列表

04 构建生僻字知识库

• 生僻字在《通用规范汉字表》中的等级分布

级别	方式	数量	汉字列表
一级	直接	28	钦欠赣华卜辛斤行于矢文页石乐柏羽殷玉今童尧力青日沁黔中肇
	拆字	12	矗尧兢滇邑泵汝寅幢黔酉夯
二级	直接	66	蠡筱阜臻靳芮翟裴睢荥昱殇缪鄄晟邵兖龚懋鄞詹岑忻瞿菁牟阚邹暹蔺谌赟倪睿砼煜衢钰虞茜郴覃廖毓樊蓟綦胥褚郝雒邳尹麓暨绥祁昕昶灏濮婺湛栾沂斐
	拆字	62	囡阜鋆臻靳曈趸舸岫颉昱罘珩柘辊晟厝訾岑尕泺怼阚赟罡珏梓歆彧砼煜毳翊晁姝乜旻喀骅廿覃劼熠淦弋尹轶婧昝槎暨烨颢昕圩昶灏斛艮沂焱聿
三级	直接	5	仝堃芈闫昇
	拆字	15	犇喆燚森堃仝甦骉淼翀曌崟昇晔垚
级外	直接	1	尐
	拆字	11	叒騳咲犇灬龘峬鱻朤槑尛

TOP100生僻字在《通用规范汉字表》中的等级分布

04 构建生僻字知识库

- **生僻字知识库数据示例**

汉字	豯
拼音	nǔ,nǚ
通用度	1
汉字等级	2
常用词语，词语通用度	翼状豯肉，2282 妓豯，139 豯肉，40 豯肉攀睛，4 将豯肉，0 …

直接搜索生僻字知识库

汉字	昱
拼音	yù
通用度	45034
汉字等级	2
常用词语，词语通用度	王昱珩，52143 朱相昱，47549 瑞昱，14756 …,…
组成部件	日立
搜索字符串	日立 上日下立 …

拆字搜索生僻字知识库

05 生僻字知识库的应用

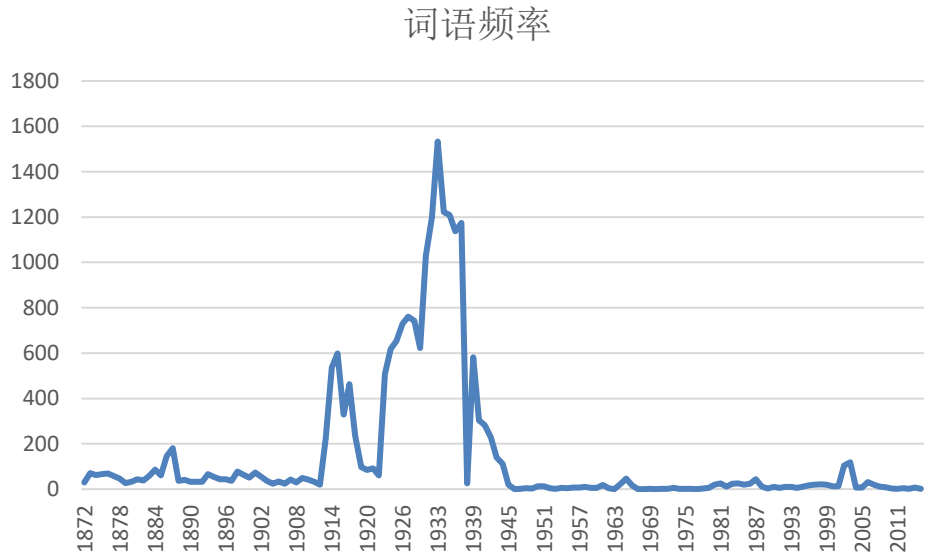
- **生僻字构词的领域分布**

领域	例词
姓名	阚昕、王堃、陆垚、刘垚昕
组织机构	鑫鸟大厦、犇腾牛排、安徽垚森
地名	临沂大学、五邑大学
网络	槑痘痘、泏槑孖、上線發槑
方言	乜嘢、冇乜
其他	石斛、铁皮石斛

生僻字构词的领域分布

05 生僻字知识库的应用

- **生僻字构词的历时分布**
  - **“昶（chǎng）”字构词的历时变化**



“昶”字历时词频和词种分布

05 生僻字知识库的应用

- **生僻字构词的历时分布**
  - **“昶（chǎng）”字构词词频分布**

词语	词频
益昶	1896
大昶	1300
昶庄	564
瑞昶	488
成昶	276
毛昶熙	273
余昶	267
恒昶	266
郑昶	237
巨诚昶	235
…	…

“昶”字构词词频分布

05 生僻字知识库的应用

- **生僻字构词的历时分布**
  - **“鑫”字构词的历时变化**

“鑫”字历时词频和词种分布

05 生僻字知识库的应用

- **生僻字构词的历时分布**
  - **“鑫”字构词词频分布**

词语	词频
路亿鑫	2646
李鑫培	2174
叶开鑫	2144
三鑫	1136
李鑫甫	1076
鑫记大	926
鑫培	826
谭鑫培	762
陈鑫斋	687
苗鑫茹	621
…	…

“鑫”字构词词频分布

05 生僻字知识库的应用

- **拆字汉字检索服务**
  - **通过微调qwen-7b模型，构建拆字汉字检索模型**

拆字检索服务在北京语言大学《国际中文智慧教学系统》中应用情况

总结

1. 使用搜索日志数据对生僻字进行统计分析
2. 总结拆字搜索规则
3. 将搜索字符串转换为汉字部件列表
4. 构造生僻字知识库
5. 生僻字知识库应用
   - 统计生僻字构词的领域分布
   - 结合BCC语料库统计生僻字构词的历时分布
   - 结合大模型实现拆字检索服务

展望

1. 收集更多的搜索数据，对生僻字进行统计
2. 完善生僻字发现方法（例如：羊）
3. 区分部件相同的生僻字（例如：邑、吧）
4. 进一步挖掘生僻字知识库的应用场景

北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY

谢谢！

Email：liutingchao@hotmail.com