基于搜索数据的生僻字知识库构建和应用

刘廷超[1] 王雨[1] 饶高琦[2] 杨兆勇[1] 荀恩东[3]

1. 北京语言大学 信息科学学院
2. 北京语言大学 国际中文教育研究院
3. 北京语言大学 高精尖语言资源中心

摘要

- 探讨了生僻字的识别与应用问题。首先通过对搜索引擎日志数据的分析，将用户搜索生僻字的行为划分成直接搜索和拆字搜索两个类别。然后制定生僻字拆分规则，生成生僻字表，结合相关信息构建生僻字知识库。最后讨论了知识库的应用，包括构词领域分布、历时统计分析以及汉字拆分搜索等方面。尽管面临样本数据有限和拆分歧义性等挑战，但本文的方法依然为生僻字的研究提供了新的视角和工具。这些成果为相关领域的进一步研究奠定了基础，并将为汉字文化的传承与发展起到了积极的推动作用。

生僻字搜索

- 当用户遇到生僻字时，通常会使用搜索引擎来查找其含义和用法。用户的搜索行为可以分为直接生僻字搜索和拆字生僻字搜索。
- 直接生僻字搜索示例

搜索字符串	次数
灏读什么	53
闰字读什么	6
斐读什么	6
…	…

- 拆字生僻字搜索示例

搜索字符串	次数
三个火字读什么	160
三个火念什么	156
三个火加一个木是什么字	53
…	…

拆字搜索部件提取

- 生僻字组成部件提取规则

规则	示例	结果
数量词	一个立一个羽 两个方一个土 三个火加一个木	立羽 方方土 火火火木
部首	三点水一个女 一个单人旁一个吉 草字头下面一个长	氵女 亻吉 艹长
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

- 生僻字组成部件提取结果

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
4	日字下面立	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

生僻字的统计结果

- 直接搜索**TOP50**字表

阜	3483	阚	3312	丵	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				

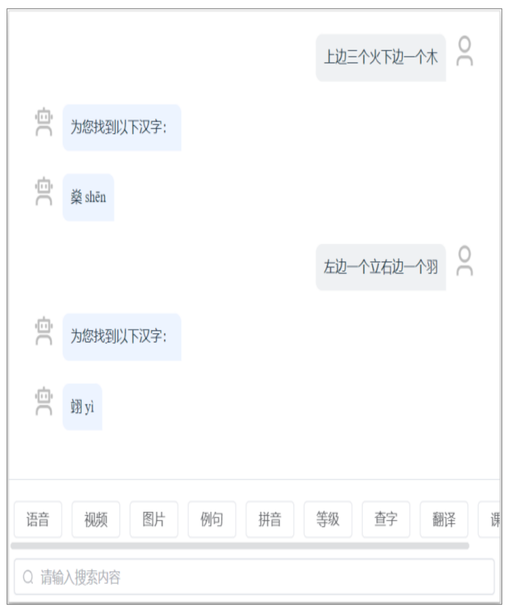
- 拆字搜索**TOP50**字表

垚	130038	燚	71547	堃	60613	焱	56376
昶	53550	昱	45034	犇	43931	赟	42286
翊	37987	燊	36919	靳	35119	晟	32497
鼛	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	狲	17948	頔	17582	騙	17039
昝	16962	圩	16894	艮	16408	泵	15813
彧	14935	泯	14761	斛	14485	岑	14196
黔	14009	粿	13218	昕	13148	趸	13018
墨	12839	柘	12565	灏	12485	龘	11708
阚	11553	怼	11448	颢	11349	聿	10995
珩	10863	梓	10768				

生僻字知识库及应用效果

- 知识库包含汉字的基本信息以及常用词语、搜索字符串等信息;
- 生僻字知识库目前已经在汉语教学产品中得到了应用。

汉字	昱
拼音	yù
通用度	46479
汉字等级	2
组成部件	日立
常用词语，词语通用度	王昱珩，52143 朱相昱，47549 瑞昱，14756 …,…
搜索字符串	日立 上日下立 …



结论

- 总结了生僻字搜索方法，包括直接查询和拆字查询，并通过实验验证其有效性;
- 制定了一套生僻字拆分规则，并通过部件还原和容错处理提高了识别准确性;
- 构建了一个全面的生僻字知识库，为生僻字的领域分布、历时统计和拆分搜索等应用提供了有力支持。

https://github.com/paineliu/RCKB

Email: liutingchao@hotmail.com