拼音输入法增量说明文档

Z 输入法, V 2.0

周昊一,张绍文,郑健,张昊,周佳雨

一.简介

Z 输入法(暂定名)是一个基于 NLP 课程内容完成的输入法,涉及到许多自然语言处理方面的内容.目前版本是 2.0, 主要涉及了词法分析,语法分析等方面的内容,当前版本已经加入了一些统计方面的内容,如词频,语言模型等.

Z 输入法是基于 Android 平台的,开发语言为 Java 和 C++(JNI).开发工具为 eclipse with ADT.通过 GitHub 来进行版本管理和协同工作.

选择 Android 平台的原因一是我们由于过去的开发经验, 相比 win32 开发更熟悉 Android 开发;二是在 Android 平台上开发输入法在 UI 的处理方面比较方便,可以把更多的精力放在输入法内核方面.

但是 Android 平台与 PC 相比运算速度要慢很多,并且内存分配上有限制(在我们的测试机上 heap size 最大不能超过 50m),这就对输入法的性能(执行效率以及内存占用方面)有了更高的要求.

二.更新内容

此次更新主要是涉及到语言模型方面.基于 HMM 模型训练并建立了一个语言模型,这个语言模型主要的用途是更好地实现简拼功能.同时,在语言模型的训练过程中,也获取了新的带有权重数据的字典与词典,进一步提升了输入过程中的准确率.

此外,还对原有的一些方面进行了一些优化.例如在音节切分方面,再加入了反向最大匹配,用于处理类似于”xian->xi’an”这样的情况.另外还对运行速度进行了一些优化,由于 HMM 模型的加入,原本性能会有显著下降,通过这些优化尽可能的将性能维持在原有水平.

三.基于语料库的新词库生成

1. 语料库

本次所使用的语料库是 2007~2011 年 5 年的人民日报的数据.总计 196051 篇文章,892M 的文本.

语料库是本次整个工作的基础,新的字库,词库,以及语言模型都是在它的基础上产生的.因此语料库的语言偏向性即代表了最后整个输入法的偏向性.

由于人民日报中的文章用词都比较正式,鲜有口语并且多涉及人名地名,因此最后可以看到生成的整个语言模型对于口语化的语句输入不能够很好的处理,但是对于比较正式的语句能够较为顺畅的输入.

2. 新的字库与词库的挖掘与生成

在上一个版本中,所使用的字库与词库都是人工搜集或生成的.具体的讲,字库是我们在半手工地生成的完整字库,并对其中一些字的权重作了一定的人工标注.而词库则是从网上收集的高频词词库,由于词数太多,做人工标注不现实,并且也无法保证标注结果的有效性,因此没有做标注.

因此,在这一版本中,我们希望能通过语料库来得到一个更为完善,有数据支撑的词库.

我们的做法是,先对文本进行分词,再对分好词的文本进行统计,记录每个出现了的词的词频,同时也记录每一个词的词频.

接下来进行一次筛选,将词频过低的词去除,因为这些词频过低的词显然不是常见的词语.根据观察,通常低词频的词都是人名与地名,这都是我们所不希望出现在我们的核心词库中的.

随后我们根据词(单字也算作词)的词频 f 算出它的权重 w.虽然可以直接令 w=f,但是实际上高词频的词与低词频的词之间词频差距太大,在计算概率时会因为浮点数精度问题导致得到为 0 的概率,因此需要某种手段来缓和高频词和低频词之间的差距.我们所选用的方法是取对数,w=ln(f+1),之所以+1 是为了保证 w 不为 0..

最后还有一个步骤,就是推断词语的拼音.由于语料库中是没有拼音资料的,因此所有词语的拼音都需要从另外的词库,或者是字库中进行推断.

我们的方法是搜集了许多词库,如果在词库中的词,就用词库中的音,否则就从字库推断,如果有多音字,就人工标注.由于多音字比较多(2000 条需要人工标注),因此这一部分虽然简单,但是花去了大家不少时间.

四.基于 HMM 的简拼查找

1. 第一版中的简拼

第一版中的简拼查询是基于简拼还原为全拼的思路来做的.将简拼还原为所有可能的全拼,再对这些全拼进行查找.

这种方法的主要问题是随着简写部分的增多,搜索空间是呈指数级增长的.原因是它将简拼还原为所有可能的全拼,但是这些全拼在词典中很有可能是不存在的.

对于全首字母简拼则是将这些全首字母全拼也全部加入到词典中,当作词语进行直接查询.

第一版中的简拼还有一个很大的问题就是无法对最终结果进行有效的排序,因为没有相关的信息可以参考,只能当作等概率处理.

2. HMM 模型的训练

首先我们描述一下输入法中所使用的 HMM 模型:

用户的输入I_a表示一系列用户输入状态,例如(xin),(xing)就是两个不同的状态.每个用户输入状态(即拼音)对应一个中文字串C_b.如假设在某次输入时(xin)对应的中文字串是(新),但在其他时候(xin)还可能对于(心),(信)等.

在这里I_a表示的是可观测的外部状态,而C_b表示的是不可观测的内部状态.不同的内部状态有可能对应相同的外部状态,不同的外部状态也可能对应相同的内部状态(比如简拼的情况).而我们的目标就是对一个给定的I_k,要能得到一个内部状态的序列L_c,其中内部状态是按照可能性的大小降序排列的.

对于全拼输入而言,这个内部状态集合是很容易得到的,直接查字典即可.

关键是对于简拼而言,这个内部状态集合是隐含的.因为在字典中无法得到简拼的相关信息.

但是我们可以根据字典中信息得到从简拼到全拼的转移概率.如对于简拼(mingt),对于全拼(mingtian)有一个转移概率 p,对于(mingti)也有一个转移概率 p'.而(mingtian)对于(明天)的概率若是 q,则(mingt)对应到(明天)的概率就是 pq.

根据这个计算方法,我们可以计算出一个简拼到所有可能的全拼的概率,并进一步推算出到内部状态的概率.

在实际的算法中,并不是从简拼到全拼进行推算,而是从所有已知的全拼推算出对应的所有简拼情况的转移概率.

由于我们的输入法并不是要算最可能的情况,而是要算出所有情况,并进行排序,因此只能用朴素算法算出完整的转移矩阵.

3. 转移矩阵的压缩

由于在 Android 平台上,受到系统性能的制约(CPU 速度,以及文件 IO 速度都明显的低于 PC 机,并且内存使用总量受到限制),在设计算法时,还必须要考虑到性能方面的问题.用上面的方法生成的转移矩阵文件相对来说是很大的.一个 90k 项,1.6m 的词典文件,生成的转移矩阵大概有 400k 行,占 20m,这对与我们的情况而言是不可接受的,首先这样一来安装包就太大了,其次 20m 的文件读取的时间太长,并且占内存太多,可能会导致程序崩溃.

因此我们需要采取一定方法来尽可能减小转移矩阵文件的数量.

根据观察,词典中有许多条目,特别是音节较多的词,拼音与词之间是一一对应的.例如 zhong'hua'ren'min'gong'he'guo,对应的词是"中华人民共和国".

首先,这样一个长词,对应的简拼可能是非常多的,有$2^7 - 1 = 127$种.需要占去文件中 127 行.

其次,它的所有简拼的转移目标都只有一个,概率为 1.

因此,对于这些情况,可以把它们合并成一个条目,用特殊的格式标出,在我们的实现中是'z'h'r'm'g'h'g, 开头的'表示这是一个这样子的特殊情况.在查询时,如果简拼对应的转移矩阵查不到,就说明要么是不存在的简拼,要么就是被合并的情况.此时去查找被合并的简拼(取所有首字母,再在开头加上'),如果查到了,再做一次判断,判断这个正在查询的简拼是不是那个被合并的简拼转移到的全拼的简拼.比如对于 zhen'he'ren'ming'g'hen'g 这样一个简拼,它的转移矩阵也是查不到的(因为不存在),此时去查合并后的简拼是能查到的,但是是不匹配的.

实践证明,这样的压缩是非常有效的,可以将文件大小至少减小一半.20m 的文件可以压缩到 8m.

4. HMM 模型的使用

经过训练,我们得到的是一个完整的由简拼到全拼的转移矩阵.在使用它来对简拼进行查找时,只需要做一步的推测,由简拼到所有可能的全拼的转移概率,以及这些全拼对应的中文串及其概率算出这个简拼对应的所有中文串及其概率.

例如,在我们的转移矩阵文件中有这样一行:

j'n'w ji'nian'wu,3;jing'nei'wai,6

表示简拼(j’n’w)可以以 3/9 的概率转移到全拼(ji’nian’wu),以 6/9 的概率转移到全拼(jing’nei’wai).

在词典中可以查到

ji'nian'wu 纪念物,3

jing'nei'wai 境内外,6

因此,(j’n’w)到(纪念物)的概率就是 3/9,到(境内外)的概率就是 6/9.并且,(纪念物)这个候选词的权重是 3*3/9=1,(境内外)这个词的权重是 6*6/9=4.

权重的作用是进行全局排序.因为输入其实是(jnw),而不是(j’n’w).对于像(xian)这样的输入,可能是(xian)->(先),也可能是(xi’an)->(西安),由于前者是直接通过字典查到的,而字典中只能查到权重,因此这里需要用权重进行一个排序.

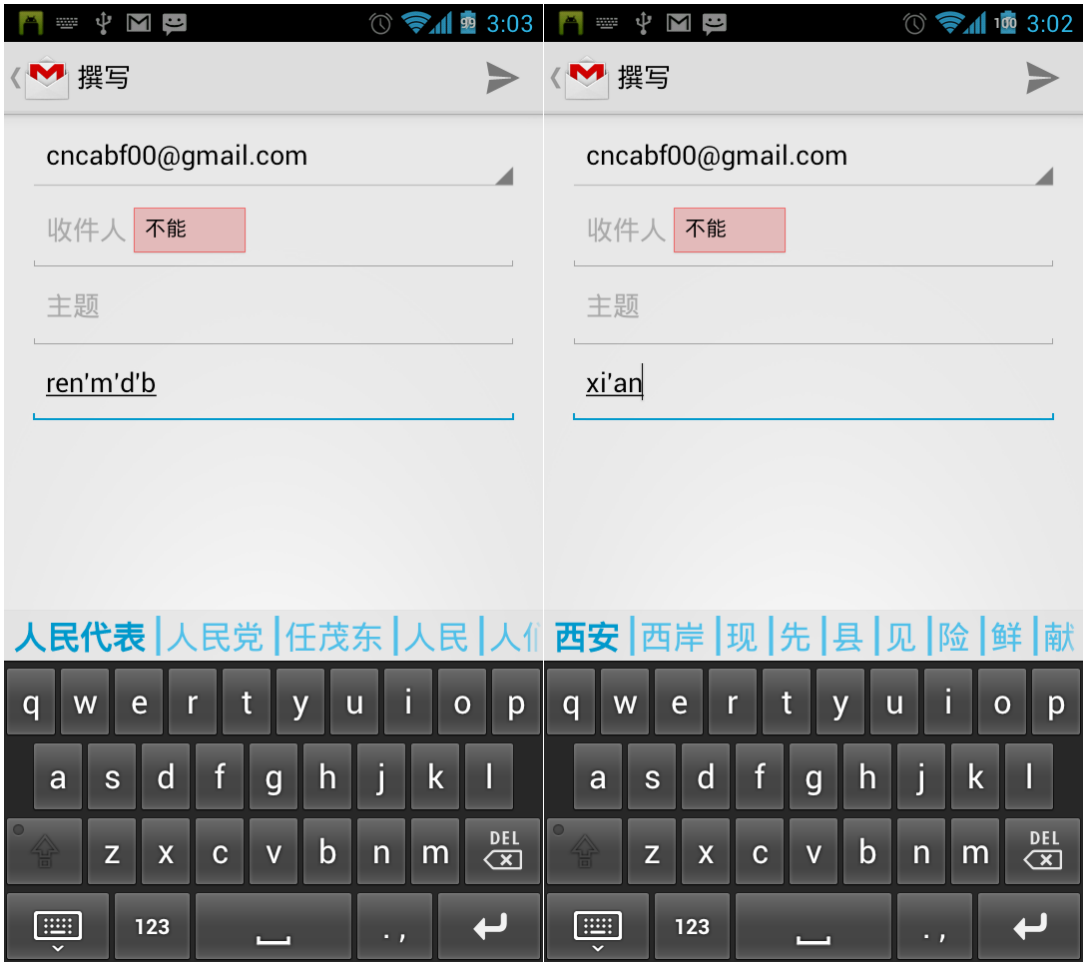
五.已知问题

现在的存在的问题主要有两个:

一是语料库偏向性比较严重,不适合日常输入(这是手机输入法的主要应用场景),需要收集一些类似于微博这样的语料库

二是性能方面还不是太理想,初始化耗时需要 2s,能明显感到需要等待,另外在一些候选词比较多的情况下会有明显卡顿,需要进一步做优化处理.

六.运行截图



七.测试

1. 测试 A

1) 按键功能测试

本中文输入法采用常规的全键盘布局。除了 26 个字母输入键外，shift 键实现大小写切换功能，enter 键实现直接选择功能，空格键输入空格，标点符号键选择输入逗号、句号，del 键实现删除功能。

2) 单字输入测试

单字输入是中文输入法基本功能，通过输入单个汉字的全拼音或开头部分，来选择输入单字。本版本输入法测试结果如下：

字母输入	期望结果	测试结果

a	啊、阿、吖……	阿、呵、腌……
ai	哎、爱、唉……	埃、爱、碍、……
d	以 d 开头的所有汉字	带、吊、大、党……
di	地、第、低……	提、地、底、低……
dia	嗲	嗲
diao	掉、叼、调……	调、鸟、雕、掉……
c	以 c 开头的所有汉字	蔡、曾、册……
ci	次、此、词、刺……	次、此、差……
chi	吃、迟、持、池……	持、吃、斥……
i	null	null
lo	咯	咯
ever	null	null

3) 全拼功能测试

词语的全拼输入是中文输入法的重要功能，通过输入整个词语的全拼音，来选择输入词语。本版本输入法测试结果如下：

字母输入	期望结果	测试结果
ayi	阿姨	阿姨
anquan	安全	安全
baba	爸爸、巴巴、粑粑……	爸爸、八八、巴巴
bimian	避免、壁面……	避免
chahua	茶花、插画、插话、插花……	常怀爱、插话、插画、插花……
chuchao	初潮；不应该有：粗糙	出超
cucao	粗糙；不应该有：初潮	粗糙
didi	弟弟、滴滴、的的……	弟弟

eyu	鳄鱼、俄语、阿谀……	俄语、鳄鱼、恶语……
enqing	恩情；不应该有：儿女情	恩情
fukua	浮夸	浮夸
guojia	国家、郭家……	国家、郭家……
haha	哈哈	哈哈
iyi	null	null
jiayuan	家园、佳缘……	家园、加元
kuohao	括号	括号
lvse	绿色、滤色	绿色
muji	母鸡、募集、目击、木屐……	募集、目击、母鸡、木屐……
nver	女儿	女儿
oa	null	oa
piaoliang	漂亮、票量……	漂亮
putong	普通、扑通……	普通、扑通
qishi	其实、骑士、气势……	其实、启示、歧视、气势……
qisi	气死……不应该有：骑士	其四、气死、奇思
riqi	日期	日期
ruizhi	睿智、锐志	睿智
siqi	私企、死期……	私企、死棋、四起
tamen	他们、她们、它们……	他们、她们、它们
uta	null	null
vqi	null	null
wajiao	挖角、蛙叫	瓦济奥
xiamen	厦门、下门、夏门	厦门

yushui	雨水	雨水、鱼水
ziji	自己、字迹……不应有：知己	自己、自给、字迹、子集
zhiji	知己、至极……不应有：自己	之际、制剂、职级……

4) 简拼功能测试

词语的简拼输入也是中文输入法的重要功能，通过输入词语中每个汉字全拼音或者拼音开头部分的组合来联想出词语进行选择输入。本版本输入法测试结果如下：

字母输入	期望结果	测试结果
ar	爱人	爱人
air	爱人	爱人
aire	爱热	埃（单字）
aqua	阿；不应有：安全	阿（单字）
bb	拜拜、爸爸、版本……	辩驳、北边、保本……
baib	拜拜、白白、百倍……	白白、败北、百倍、百般……
bbian	不变、不便……	不便、褒贬、北边、病变……
cc	粗糙、出差、初潮……	草丛、次次、匆促、草草……
chc	出差、初潮……	曹杭村、储存、楚辞……
cuc	粗糙、促成……	粗糙、粗粗、簇簇……
dda	单打	抵达、到达、单打、大大……
dad	大道、大胆……	大道、答道、大德、大典……
ey	鳄鱼、俄语、阿谀……	而言、扼要、耳语……
eu	null；不应有：鳄鱼	额（单字）
diaozi	调子；不应有：迪奥紫	调子
dale	打了、大了；不应有：到了	大（单字）
fusho	扶手、副手……	扶手、俯首、福寿

renm	人民、人名、人们……	人民、人们、人马、任免……
gole	够了、供了……	null
guj	估计、顾及、股价……	股金、古建、痼疾……
hq	回去、花钱……	毁弃、环球、黑钱……
iy	null	null
jiaw	价位、家务……	家务、价位、甲烷、加温……
kuoh	括号、括弧……	括号
lanqi	蓝旗；不应有：篮球	兰（单字）
mus	牧师、暮色、目送……	穆萨、暮色……
nve	女儿	女儿
ner	女儿	那（单字）
ouy	欧阳	欧元、欧亚……
pl	漂亮、票量……	蓬乱、普勒、乒联
qip	欺骗、期盼、奇葩……	棋盘、期盼
rzi	日子	融资、日子
sez	色泽、生儿子	色泽；没有：生儿子
tqi	天气……	透气、跳棋……
uy	null	null
vb	null	null
waq	瓦器……	挖潜
xx	学校、学习、谢谢……	行星、乡下……
yusu	语速；不应有：玉碎	育（单字）
zit	字体、姿态……	姿态、紫藤、字头
zuoti	做题；不应有：昨天	作（单字）

5) 连续输入测试

连续输入是对短语或句子中每个词语或单字进行连续全拼或简拼后，选择实现直接输入整个短语或句子的功能，是中文输入一个重要的扩展功能，可以极大的提高中文输入法的用户体验。本版本输入法测试结果如下：

字母输入	期望结果	测试结果
zhhrenmgongheg	中华人民共和国	中华人民共和国
wodysheng	我的一生	我的一生
tashpant	他是叛徒	他是叛徒
nisnlren	你是哪里人	你是哪里人
woxhchangge	我喜欢唱歌	我喜欢唱歌
njdxjsjx	南京大学计算机系	南京大学计算机系
tingghexz	听歌和写字	听歌和写字
shibadlongzhzhaok	十八大隆重召开	十八大隆重召开
niswdy	你是我的眼	你是我的眼
Wangfeishggesh	王菲是个歌手	王菲是个歌手

2. 测试 B

在这次测试中，我们主要是测试了歧义，简拼，混拼等情况，具体如下：

1) 测试有歧义的词

xian, 前五的是：西安，西岸，现，仙，县

tian， 前五的是：提案，天，填，添，田

dian， 前五的是：堤岸，迪安，点，电，典

gang， 前五的是：港，刚，岗，钢，纲，尴尬排在第九。

dang， 前五的是，当，党，荡，档，档。而蛋糕，单个，单杠等词排在第九之后。

tang,　前五的是，贪官，探戈，唐，堂，塘。

2) 简拼测试

gx,　前五的是：关系，各项，贡献，高效，广西

dt，　前五的是：独特，动态，带头，当天，大厅

hz,　前五的是：合作，孩子，汉族，回族，火灾

mg,　前五的是：美国，每个，目光，曼谷，敏感

nmg，　前五的是：农民工，内蒙古，农民，那么，你们

xz,　前五的是：现在，西藏，小组，选择，协作

wy，　前五的是：委员，无疑，唯一，网友，位于

zc，　前五的是：再次，总裁，自从，早餐，杂草

cg，　前五的是：采购，参观，村官，草根，才干

dy，　前五的是：党员，对于，第一，调研，电影

gwy，　前五的是：国务院，公务员，岗位，国外，更为

3) 混拼测试

haoy,　前五的是：好运，好友，好意，好样，耗油

hyun，　前五的是：航运，海运，好运，货运，怀孕

ziw,　前五的是：自我，滋味，自卫，咨文，自问

zwei，　前五的是：　作为，最为，滋味，座位，自卫

cain,　前五的是：才能，菜农，采纳，采暖，财年

cnian，　前五的是：　次年，财年，采，财，材

hunh，　前五的是：混合，婚后，浑厚，浑河，昏黄

hhun,　前五的是：黄昏，含混，会，后，好

kaix，　前五的是：开心，开学，开销，凯旋，开县

kxin，　前五的是：开心，可信，空心，苦心，口信

mam，　前五的是：妈妈，麻木，马某，骂名，马

mma，　前五的是：妈妈，密码，木马，明码，谩骂

xiaom，前五的是，小麦，消灭，小米，校门，消弭

xming，前五的是，鲜明，姓名，写明，性命，虚名

yingx，　前五的是：影响，营销，英雄，影像，硬性

yxing，　前五的是：英雄，元凶，杨雄，有，员，亚

banlang，前五的是：板蓝根，斑斓，办，班，版

blgen，　前五的是：板蓝根，办理，比例，便利，不良

上面列出了部分的测试结果。从上面的数据可以看出，总体还是不错的。给出的前五的单词都是日常比较常见的词语。这样用户会有较好的体验。但由于资源的问题，我们的输入法暂时还无法显示那些网络上用的较多的词语，这还有待将来的改进。

八.小组情况

本小组共有 5 个人.

Leader 为周昊一(MF1233055),负责核心代码,以及用户界面的编写.

其余 4 个人及本次的工作为:

张绍文(MF123304),负责词典的统计,训练,以及测试.

郑健(MF1233054),拼音表,字典的纠错,优化,以及测试.

张昊(MG1233046),负责词典中拼音的生成以及人工标注,以及测试.

周佳雨(MG1233091),负责搜集语料库,并进行分词以及整理.