

知乎

首发于  
平安寿险AI

## AI LIVE | 文本纠错技术探索和实践



最AI的小...

微信公众号：平安寿险PAI；来，一起AI呀~

[关注他](#)

30 人赞同了该文章

### • 小PAI导读 •

「AI LIVE」是平安人寿AI团队打造的**AI专业知识分享和学习专栏**，将通过直播、沙龙等形式，分享平安寿险AI技术及创新成果，推动实现与AI领域同行共成长。

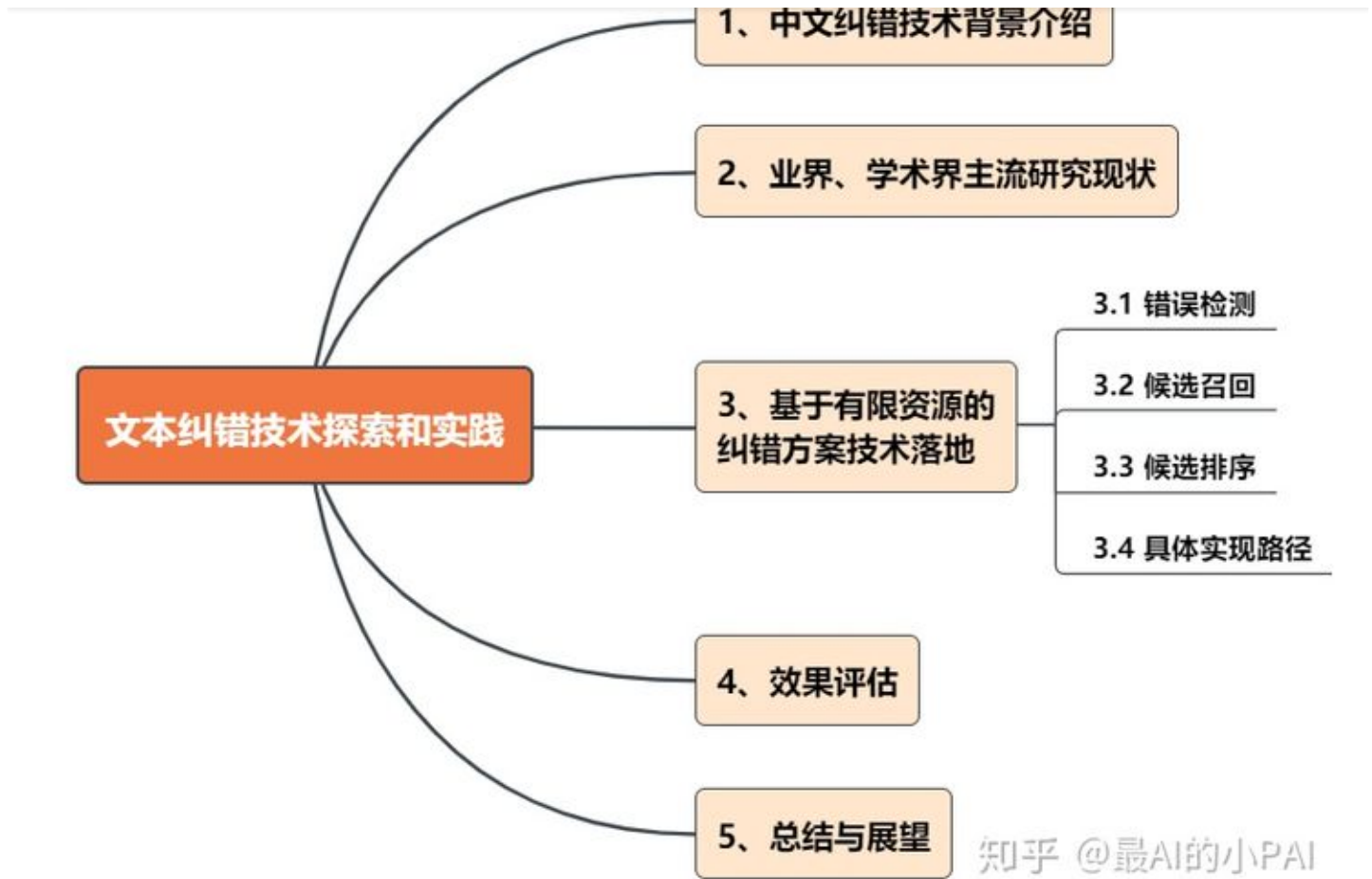
本期「AI LIVE」将回顾我们在“AI研习社”直播间进行的主题为「**文本纠错技术探索和实践**」的技术分享，由平安人寿AI团队高级算法工程师陈乐清老师主讲。

为了让大家能够快速get本期直播干货，小PAI特别整理了这篇直播内容文字稿，一起来复习一下吧~

### 全文框架概览

知乎

首发于  
平安寿险AI



知乎 @最AI的小PAI

## 一、背景与意义

**中文纠错技术**是实现中文语句自动检查、自动纠错的一项重要技术，其目的是提高语言正确性的同时减少人工校验成本。**纠错模块**作为自然语言处理最基础的模块，其重要程度不言而喻。

在日常生活中，我们经常会在微信、微博等社交工具或公众号文章中发现许多错别字。我们在几个方面对文本出错概率进行了统计：在微博等新媒体领域中，文本出错概率在2%左右；在语音识别领域中，出错率最高可达8-10%；而在平安人寿问答领域中，用户提问出错率在去重后仍高达9%。

在平安人寿问答领域的用户问题中，我们发现多种类型错误。其中占比最高的错误是语言转化和发音不标准的错误，占错误总量的50%。比如一款保险产品“少儿平安福”被语言识别转化为“少儿平安符”、“飞机”因方言差异被读成“灰机”、“难受想哭”变成“难受香菇”等。

占比第二高的错误类型是拼写错误，占错误总量的35%。这些错误主要发生在通过拼音、五笔和手写输入文本的场景。比如“眼镜蛇” - “眼睛蛇”、“缺铁性贫血” - “缺铁性盆血”等。剩余的错误我们将其分类为语法和知识错误，语法错误包括多字少字乱序，如“地中海投保” - “投保地中

## 1-背景介绍:

意义: 提升query理解准确性及对话效果, 增强用户体验

### 常见中文错误类型:

#### 发音&音转错误

特点: 音近, 发音不标准用  
原因: 地方发音, 语言转化

- 少儿平安符 → 少儿平安福
- 灰机
- 输暖管手术投保 → 卵

#### 拼写错误

特点: 正确词语错误使用  
原因: 输入法-拼音\五笔\手写

- 眼睛蛇咬了
- 紫癜投保 → 痲
- 缺铁性盆血

#### 语法&知识错误

特点: 多/少字, 乱序, 知识错误  
原因: 知识缺乏, 语言不熟悉

- 投保地中海
- 在南山平安金融中心入职 → 福田

### 常见商用场景:

#### 通用搜索领域

特点: 超大规模的web语料  
用户点击数据

#### 垂直搜索引擎

特点: 用户检索意图明确  
数据规模小、质量差

#### 垂直客服机器人

特点: 领域受限  
缺乏点击数据(无监督)

## 二、研究现状

在通用领域中, 中文文本纠错问题是从互联网起始时就一直在解决的问题。在搜索引擎中, 一个好的纠错系统能够对用户输入的查询词进行纠错提示, 或直接展示正确答案。

在此给大家介绍一个比较受欢迎的纠错项目: **Pycorrector**。该项目由规则纠错和深度学习纠错两部分组成。深度学习纠错项目中提到一些前沿的方法, 比如机器翻译, 但作者未提供直接调用接口; 而规则纠错虽然可以直接调用, 但因其性能和准确率无法满足我们项目需求, 无法直接使用。下面简单介绍一下规则纠错, 主要分为经典三步曲: 第一步通过常用词词典匹配结合统计语言模型的方式进行错误检测; 第二步利用近音字, 近形字和混淆字进行候选召回; 最后一步利用统计语言模型进行打分排序。

▲ 赞同 30 ▼    8 条评论    分享    喜欢    收藏    申请转载    ...

<https://zhuanlan.zhihu.com/p/159101860>

3/17

知乎

首发于  
平安寿险AI**基于规则的通用纠错项目：**<https://github.com/shibing624/pycorrector>

- 错误检测：
  - 常用字典匹配：切词后词不在常用字典中认为有错 韩国\国籍\投保
  - 统计语言模型：某个字的似然概率值低于句子文本平均值
  - 混淆字典匹配：国籍 → 国藉
- 候选召回
  - 近音字典替换错误位置 籍\ji → 籍,际,集, ...
  - 近形字典替换错误位置 籍 → 籍,藕,箱
- 候选排序
  - 利用统计语言模型计算句子概率,取概率超过原句且最大的
    - $P(\text{韩国国籍投保})$   $P(\text{韩国国际投保})$
    - $P(\text{韩国国积投保})$   $P(\text{韩国国藕投保})$  ...

知乎 @最AI的小PAI

而中文文本纠错学术进展主要集中在比赛项目上，如前几年SIGHAN举办的中文拼写纠错比赛以及近几年NLPCC等举办的中文语法检测和纠错的比赛等。在2017年的中文语法错误检测比赛中，其Top1的主要方案利用了序列标注模型结合人工提取特征。在NLPCC2018年举办的中文语法纠错比赛中，冠军团队应用基于Transformer的翻译模型，其主要原理是将错误句子翻译为正确句子。

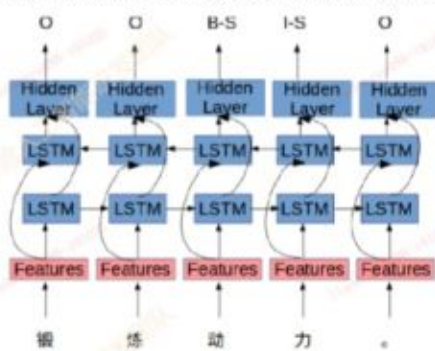
**2-研究现状：学术界进展**

pai 平安寿险AI

**基于序列标注的纠错方案：**

《Alibaba at IJCNLP-2017 Task 1: Embedding Grammatical Features into LSTMs for Chinese Grammatical Error Diagnosis Task》

利用序列标注模型+人工提取特征进行错误位置的标注。



2017年IJCNLP举办的CGED比赛中阿里团队提出的Top1方案

Position Level:			
Precision	Recall	F1	
0.36	0.21	0.27	

**基于NMT的纠错方案：**

《Youdao's Winning Solution to the NLPCC-2018 Task 2 Challenge: A Neural Machine Translation Approach to Chinese Grammatical Error Correction》

利用模型将错误语句翻译为正确语句，利用Transformer模型完成端到端的纠正过程。



2018年NLPCC举办CGEC比赛中有道团队提出的Top1方案

Correction Level:			
Precision	Recall	F0.5	F1
0.34	0.18	0.29	

在学术界的进展中我们可以发现，很多成熟的方案都是基于有监督的深度学习，比如：序列标注模型、翻译模型。但在标注资源受限的情况下，此类深度学习模型很难应用落地。同时纠错技术本身

▲ 赞同 30

● 8 条评论

➤ 分享

♥ 喜欢

★ 收藏

📄 申请转载

...



知乎

首发于  
平安寿险AI

- 没有点击语料，有的只是没有标注过的机器人的问题；送标效率低，而且还会引入很多标注错误；
- 没有标注语料，很难开展基于深度学习的有监督学习；
- 纠错是作为基础模块，对内存和时效要求很高，当前线上纠错要求3ms/句，大规模字典和复杂模型无法在线上使用；
- 线上纠错要求很高准确度，宁愿少召回也要保证高准确度，过纠率0.2%
- 85%以上的错误都是替换错误，比如语言转化错误，拼写错误

知乎 @最AI的小PAI

对于纠错系统的性能评估指标我们将过纠率与召回率看作硬性指标。过纠率代表正确的句子被改错的比率，召回率代表错误的句子被全部纠正的比率，其中较大的过纠率将会对系统和用户体验带来负面效果。我们的目标就是要让纠对句子数量远远大于被改错句子的数量，公式表述如下：如果句子出错概率是K，则 $K * RECALL \gg (1 - K) * FAR$ 。

这里罗列了一些参考值，假设句子出错概率为2%，过纠率为0.5%，那么召回率必须要大于25%；如果我们句子的出错概率为9%，在同样的召回率情况下，我们可以容忍更大的过纠率。所以本系统的目标就是在控制过纠率在0.2%左右，尽量提高召回率。

## 2-研究现状：纠错指标参考

pai 平安寿险AI

### 评价指标：

- 过纠率/误报率：

$$FAR = \frac{\text{正确句子被错纠的个数}}{\text{正确句子个数}}$$

- 召回率：

$$RECALL = \frac{\text{含错误的句子被改正的句子数}}{\text{含错误的句子数}}$$

- 纠错目标：被改正的句子数  $\gg$  被改错的句子数

$$K * RECALL \gg (1 - K) * FAR$$

句子出错概率(K)	过纠率 (FAR)	召回率 (RECALL)
2%	0.5%	24.5%
2%	0.1%	4.9%
9%	2.5%	25.3%
9%	0.5%	5.1%

寿险问答机器人目标：FAR&lt;0.2%，尽量提高RECALL

知乎 @最AI的小PAI

▲ 赞同 30 ▼

● 8 条评论

➤ 分享

♥ 喜欢

★ 收藏

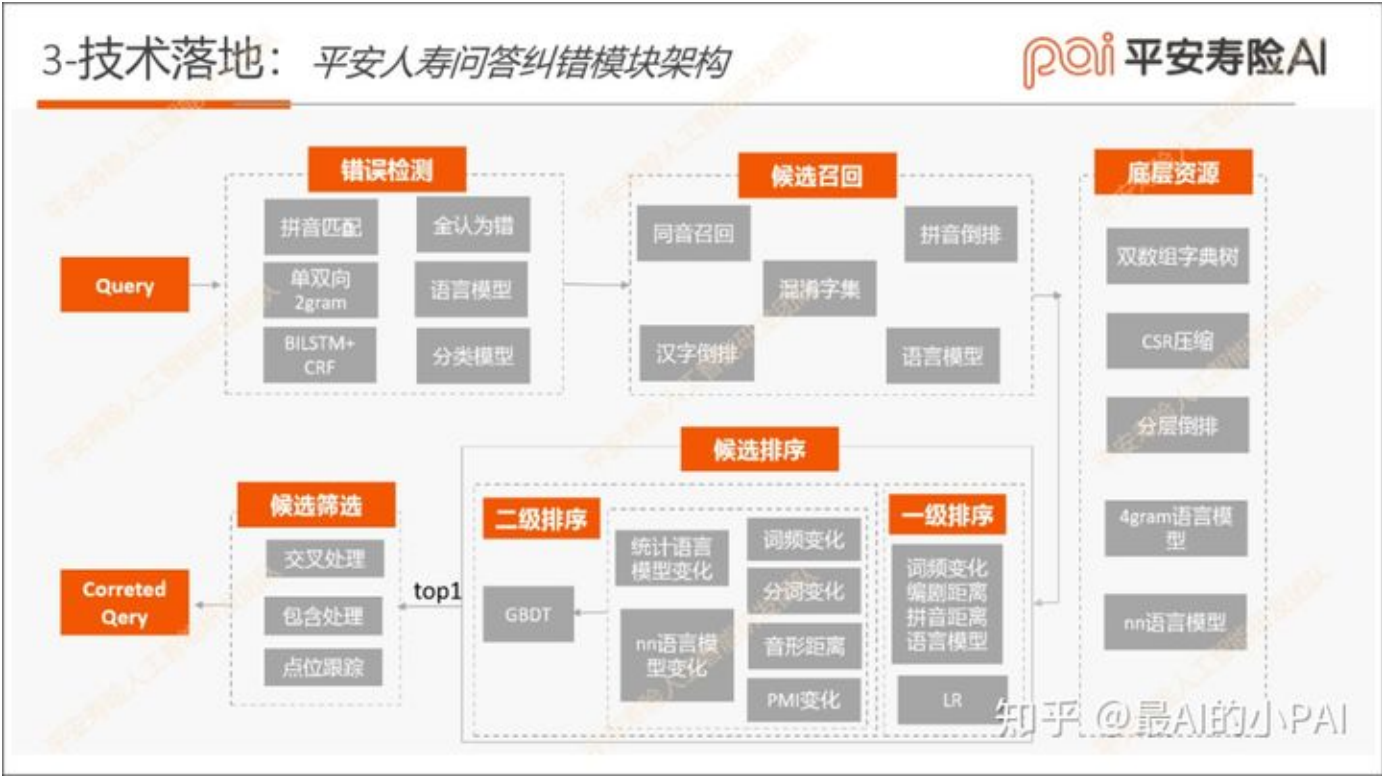
📄 申请转载

...

为了解决上述挑战，团队设计了纠错框架（如下图），在主流程上我们延用了业界经典的纠错系统架构：**错误检测、候选召回、候选排序**。这里为了更好地展示我们纠错系统的细节，我们也将“底层资源”与“候选筛选”作为基础模块添加至主逻辑中。

在系统输入用户问题后，首先进入到错误检测模块找出相应的错误点位，然后针对错误片段进行正确候选词召回，其次经过粗排序、精排序等排序模块，最后通过候选筛选模块处理候选交叉冲突等情况。

其中底层资源用于存放各类底层字典资源、上下文语义信息从而方便各模块调用，其存储的数据结构主要包括：双数组字典树、稀疏矩阵压缩算法CSR和分层倒排索引等；上下文语义信息则以统计语言模型及NN语言模型的形式保存。



在主流程中，“错误检测、候选召回、候选排序” 每一步我们都进行了多种解决方案的探索，下面将进行详细介绍。

### 3.1 错误检测

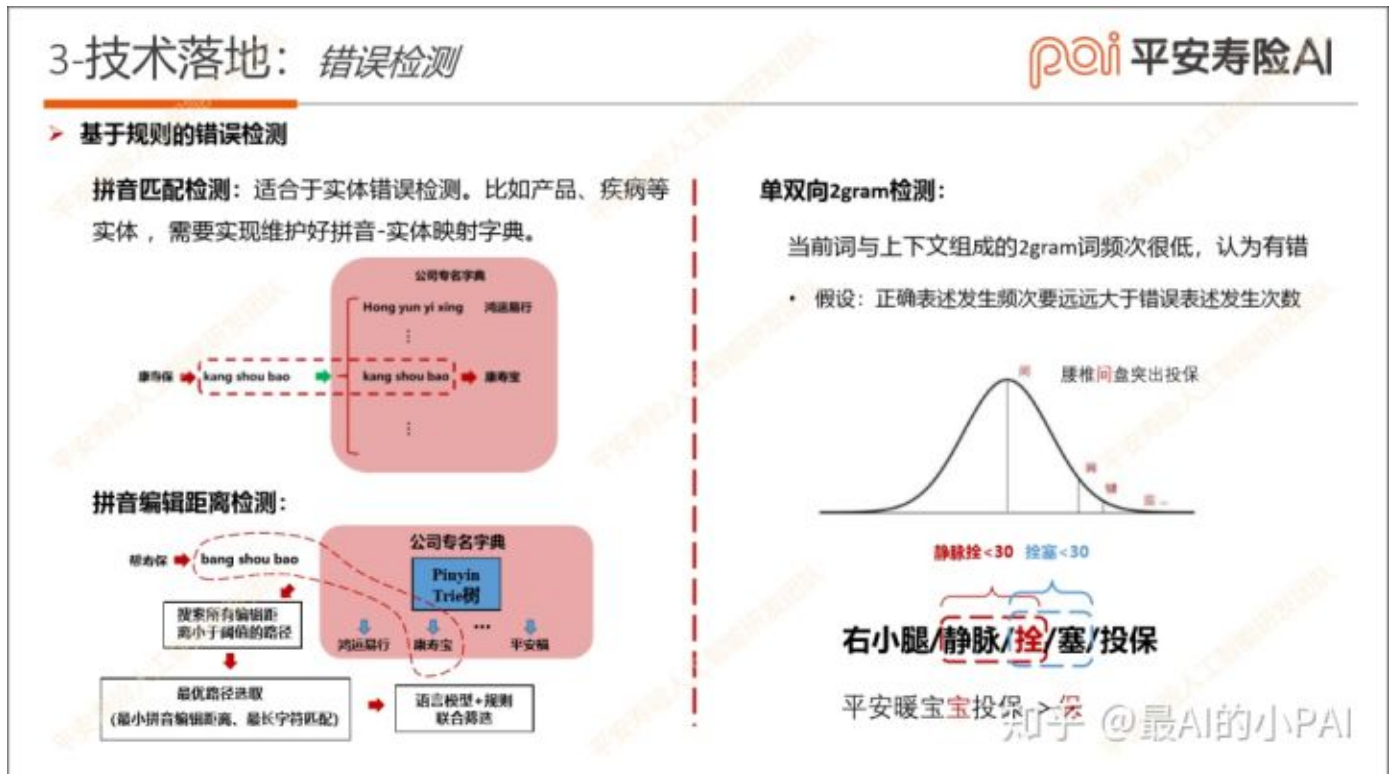
#### 3.1.1 基于规则的错误检测

##### 1) 基于近似拼音匹配的产品专名纠错

离的匹配。

## 2) 双向2gram检测

该方法使用我们基于一个假设：就是正确表述发生频次要比错误的表述发生频次要高很多。我们将语料中所有2gram的联合概率分布拟定为正态分布，正确2gram片段的出现概率将远大于错误出现的概率，从直观上来看这种假设也符合常理，因此可在此假设的基础上将联合概率的比值作为错误判断的依据。



### 3.1.2 基于模型的错误检测

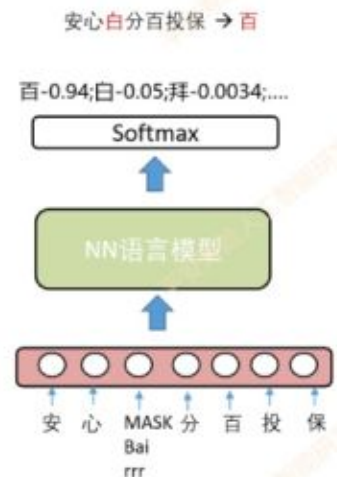
受到Bert的Mask Language Model的启发，我们可以把错误检测的过程转化成字级别的完形填空问题；在进行错误检测时对每个字所在位置进行逐一预测，从而根据预测的概率分布判断当前位置是否有错。

知乎

首发于  
平安寿险AI

## ➤ 基于nn语言模型错误检测

- 通过完形填空的方式来预测候选字的概率分布
- 如果原字的概率不在topk里或者与top1比值超过阈值认为有错
- 改进措施:
  - 传统语言模型从左到右预测, 只利用上文, 改进成利用上下文;
  - 传统语言模型直接把预测字MASK, 不会带入预测字的信息, 通过引入当前字的去除后鼻音和翘舌音的拼音和五笔等信息;
  - 传统语言模型会直接预测这个字表, 比如字表大小是3800, 会直接得到3800个字的概率分布, 通过将预测字约束在近音、近形和混淆字表里, 提高正确字与错误字的区分度



知乎 @最AI的小PAI

第一种语言模型的方法是基于word2vec的cbow改造。在word2vec训练过程中我可以通过指定窗口大小的上下文来预测当前字的概率。同时在此基础上, 我们又进行两个方面的改造:

- 第一、传统语言在预测当前位置时候是不会带入当前字的先验信息, 但是在我们的场景中, 因为正确字可能为错误字的近音、近形词, 所以我们加入了待预测字的拼音和五笔特征;
- 第二、我们将字典压缩为领域内的高频字, 同时对NN的输出进行受限, 在预测时输出层只计算近音、近形和混淆字所对应的神经元, 在提高正区分度的同时, 增加计算效率。

## 3-技术落地: 错误检测

pai 平安寿险AI

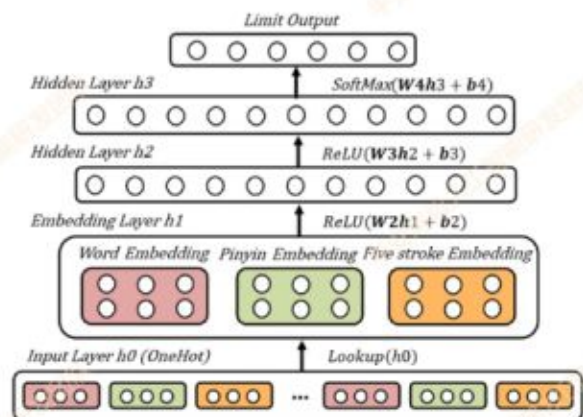
## ➤ 基于word2vec-cbow改造的音字混合受限字表语言模型错误检测算法

基于CSLM的中文拼写错误检测, 2015

Chinese Spelling Errors Detection Based on  
CSLM, 2015

- 带入预测字及上下文拼音、五笔特征;
- 去掉前后鼻音和翘舌音, 并利用混淆音集映射的方式来提高模型对谐音错误的识别性能;
- 预测字表受限于近音字、近形字与混淆字表中;

badcase: 哎, 我好困哪, 好晚了 -&gt; 玩



知乎 @最AI的小PAI

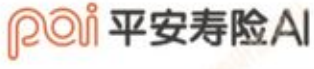


知乎

首发于  
平安寿险AI

年的一篇论文进行改造的，模型利用了BiLstm，前向Lstm从左到右学习 $h_{k-1}$ ，后向Lstm从后到左学习 $h_{k+1}$ ，然后合并两个得到 $h_k$ ，得到 $h_k$ 先与输入的字向量 $x_i$ 做Attention得到 $C_k$ ，然后 $C_k$ 与 $h_k$ 拼接得到 $P_k$ ；再用 $P_k$ 与候选字向量做Attention，用Attention后的分数作为预测概率分布。使用该模型可以缓解邻近字也是错别字的情形；比如“腰椎键潘突出投保”，正确是“间盘”，我们在纠正“键”的时候，因为可以利用更长距离的“腰椎、突出、投保”等字符信息，可以减少右边错别字“潘”的影响。

### 3-技术落地：错误检测

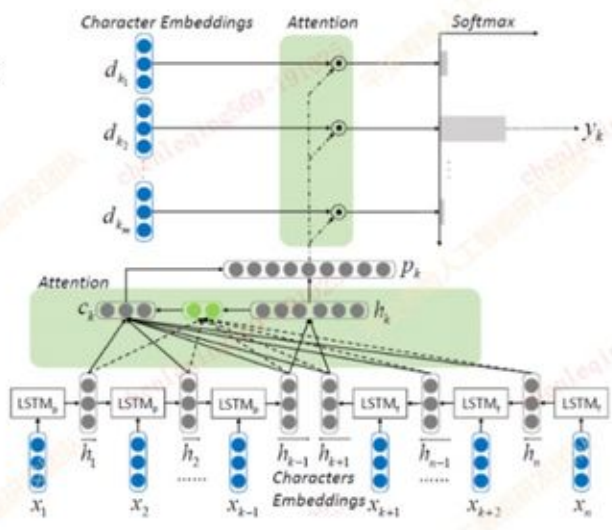


➤ 基于BiLstm改进的音字混合受限字表语言模型错误检测算法

《中文自动校对：基于字符级nn网络的中文拼写错误检测和识别》，2019  
Automatic Proofreading in Chinese : Detect and Correct Spelling Errors  
in Character-Level with Deep Neural Networks , 2019

二层注意力机制

- 第一层：hk与其他字的输出向量hs做attention，如果候选词邻近的字也是错别字，可以利用全文的信息  
腰椎**监潘**突出投保
- 第二层：第一层attention后的 $c_k$ 与 $h_k$ 拼接后与每个候选词做attention，直接利用attention的分数作为最后候选排序分数



Double Attentive Checker

知乎 @最AI的小PAI

第三种语言模型是基于Bert模型进行改造的，加入了拼音和字形特征，然后将训练的超参数进行调整，比如层数变成3层，embedding\_size从750降到150，使用5个头，字典从原来2w变成3.8k。

基于bert改造的音字混合受限字表语言模型错误检测算法

BERT: Pre\_training of Deep Bidirectional Transformers for Language Understanding, 2018

改造huggingface的代码，输入层加入拼音，字形特征

开源项目地址：  
<https://github.com/huggingface/transformers>

参数细节：  
Num\_hidden\_layers: 3  
Num\_attention\_head:5  
Hidden\_size:150  
Vocab\_size:3800  
Max\_position\_embedding:86

知乎 @最AI的小PAI

以上三个语言模型，考虑到效率问题，其中基于word2vec改造的模型用于线上纠错中；其他两个模型用于对效率要求较低，但准确度要求很高的离线版本的纠错中。

### 3.2 候选召回

上面介绍的语言模型除了用于错误检测，还可以用于候选召回，即正确的候选字从语言模型预测输出的topk中获取。

除了通过语言模型召回候选以外还需要通过混淆词典来召回正确候选词。相关的混淆词典包括：基于近音、近型、编辑距离的1、2gram混淆词典。为了提高字典的索引效率及搜索时间，我们将1gram词及词频和1gram近音词词典使用双数组字典树存储，而2gram词典采用CSR数据结构存储，同时2gram的近音混淆词可以从以上词典里恢复出来。而为了进行编辑距离召回候选词，我们建立了分层倒排索引词典从而提高搜索效率。

近音候选词召回:

大规模的基础字典的依赖,使其在存储空间与读取速度方面受到极大挑战

底层字典整体存储架构

双数组字典

↑存储

1gram-词频字典

CSR

↑存储

1gram-拼音-同音词字典

↑存储

2gram-词频字典

信息

恢复出

2gram-拼音-同音词字典

腰椎间盘突出

↓

腰椎肩盘

↓

yao zui jian pan

↓

2gram

↓

腰椎间盘

- 降低存储空间: a、利用Trie树降低信息冗余  
b、利用经典结CSR压缩稀疏矩阵  
c、使用词典间的关联信息恢复2gram同音词典
- 提高读取速度: Trie树、CSR技术的高效索引

字、音编辑距离召回:

分层倒排索引: 对于每个字的倒排词集按照词中字的数量按照分层的方式存储

字1

↓

字2

↓

字3

↓

字4

↓

字5

↓

字6

↓

字7

↓

字8

↓

字9

↓

字10

↓

字11

↓

字12

↓

字13

↓

字14

↓

字15

↓

字16

↓

字17

↓

字18

↓

字19

↓

字20

↓

字21

↓

字22

↓

字23

↓

字24

↓

字25

↓

字26

↓

字27

↓

字28

↓

字29

↓

字30

↓

字31

↓

字32

↓

字33

↓

字34

↓

字35

↓

字36

↓

字37

↓

字38

↓

字39

↓

字40

↓

字41

↓

字42

↓

字43

↓

字44

↓

字45

↓

字46

↓

字47

↓

字48

↓

字49

↓

字50

↓

字51

↓

字52

↓

字53

↓

字54

↓

字55

↓

字56

↓

字57

↓

字58

↓

字59

↓

字60

↓

字61

↓

字62

↓

字63

↓

字64

↓

字65

↓

字66

↓

字67

↓

字68

↓

字69

↓

字70

↓

字71

↓

字72

↓

字73

↓

字74

↓

字75

↓

字76

↓

字77

↓

字78

↓

字79

↓

字80

↓

字81

↓

字82

↓

字83

↓

字84

↓

字85

↓

字86

↓

字87

↓

字88

↓

字89

↓

字90

↓

字91

↓

字92

↓

字93

↓

字94

↓

字95

↓

字96

↓

字97

↓

字98

↓

字99

↓

字100

↓

字101

↓

字102

↓

字103

↓

字104

↓

字105

↓

字106

↓

字107

↓

字108

↓

字109

↓

字110

↓

字111

↓

字112

↓

字113

↓

字114

↓

字115

↓

字116

↓

字117

↓

字118

↓

字119

↓

字120

↓

字121

↓

字122

↓

字123

↓

字124

↓

字125

↓

字126

↓

字127

↓

字128

↓

字129

↓

字130

↓

字131

↓

字132

↓

字133

↓

字134

↓

字135

↓

字136

↓

字137

↓

字138

↓

字139

↓

字140

↓

字141

↓

字142

↓

字143

↓

字144

↓

字145

↓

字146

↓

字147

↓

字148

↓

字149

↓

字150

↓

字151

↓

字152

↓

字153

↓

字154

↓

字155

↓

字156

↓

字157

↓

字158

↓

字159

↓

字160

↓

字161

↓

字162

↓

字163

↓

字164

↓

字165

↓

字166

↓

字167

↓

字168

↓

字169

↓

字170

↓

字171

↓

字172

↓

字173

↓

字174

↓

字175

↓

字176

↓

字177

↓

字178

↓

字179

↓

字180

↓

字181

↓

字182

↓

字183

↓

字184

↓

字185

↓

字186

↓

字187

↓

字188

↓

字189

↓

字190

↓

字191

↓

字192

↓

字193

↓

字194

↓

字195

↓

字196

↓

字197

↓

字198

↓

字199

↓

字200

↓

字201

↓

字202

↓

字203

↓

字204

↓

字205

↓

字206

↓

字207

↓

字208

↓

字209

↓

字210

↓

字211

↓

字212

↓

字213

↓

字214

↓

字215

↓

字216

↓

字217

↓

字218

↓

字219

↓

字220

↓

字221

↓

字222

↓

字223

↓

字224

↓

字225

↓

字226

↓

字227

↓

字228

↓

字229

↓

字230

↓

字231

↓

字232

↓

字233

↓

字234

↓

字235

↓

字236

↓

字237

↓

字238

↓

字239

↓

字240

↓

字241

↓

字242

↓

字243

↓

字244

↓

字245

↓

字246

↓

字247

↓

字248

↓

字249

↓

字250

↓

字251

↓

字252

↓

字253

↓

字254

↓

字255

↓

字256

↓

字257

↓

字258

↓

字259

↓

字260

↓

字261

↓

字262

↓

字263

↓

字264

↓

字265

↓

字266

↓

字267

↓

字268

↓

字269

↓

字270

↓

字271

↓

字272

↓

字273

↓

字274

↓

字275

↓

字276

↓

字277

↓

字278

↓

字279

↓

字280

↓

字281

↓

字282

↓

字283

↓

字284

↓

字285

↓

字286

↓

字287

↓

字288

↓

字289

↓

字290

↓

字291

↓

字292

↓

字293

↓

字294

↓

字295

↓

字296

↓

字297

↓

字298

↓

字299

↓

字300

↓

字301

↓

字302

↓

字303

↓

字304

↓

字305

↓

字306

↓

字307

↓

字308

↓

字309

↓

字310

↓

字311

↓

字312

↓

字313

↓

字314

↓

字315

↓

字316

↓

字317

↓

字318

↓

字319

↓

字320

↓

字321

↓

字322

↓

字323

↓

字324

↓

字325

↓

字326

↓

字327

↓

字328

↓

字329

↓

字330

↓

字331

↓

字332

↓

字333

↓

字334

↓

字335

↓

字336

↓

字337

↓

字338

↓

字339

↓

字340

↓

字341

↓

字342

↓

字343

↓

字344

↓

字345

↓

字346

↓

字347

↓

字348

↓

字349

↓

字350

↓

字351

↓

字352

↓

字353

↓

字354

↓

字355

↓

字356

↓

字357

↓

字358

↓

字359

↓

字360

↓

字361

↓

字362

↓

字363

↓

字364

↓

字365

↓

字366

↓

字367

↓

字368

↓

字369

↓

字370

↓

字371

↓

字372

↓

字373

↓

字374

↓

字375

↓

字376

↓

字377

↓

字378

↓

字379

↓

字380

↓

字381

↓

字382

↓

字383

↓

字384

↓

字385

↓

字386

↓

字387

↓

字388

↓

字389

↓

字390

↓

字391

↓

字392

↓

字393

↓

字394

↓

字395

↓

字396

↓

字397

↓

字398

↓

字399

↓

字400

↓

字401

↓

字402

↓

字403

↓

字404

↓

字405

↓

字406

↓

字407

↓

字408

↓

字409

↓

字410

↓

字411

↓

字412

↓

字413

↓

字414

↓

字415

↓

字416

↓

字417

↓

字418

↓

字419

↓

字420

↓

字421

↓

字422

↓

字423

↓

字424

↓

字425

↓

字426

↓

字427

↓

字428

↓

字429

↓

字430

↓

字431

↓

字432

↓

字433

↓

字434

↓

字435

↓

字436

↓

字437

↓

字438

↓

字439

↓

字440

↓

字441

↓

字442

↓

字443

↓

字444

↓

字445

↓

字446

↓

字447

↓

字448

↓

字449

↓

字450

↓

字451

↓

字452

↓

字453

↓

字454

↓

字455

↓

字456

↓

字457

↓

字458

↓

字459

↓

字460

↓

字461

↓

字462

↓

字463

↓

字464

↓

字465

↓

字466

↓

字467

↓

字468

↓

字469

↓

字470

↓

字471

↓

字472

↓

字473

↓

字474

↓

字475

↓

字476

↓

字477

↓

字478

↓

字479

↓

字480

↓

字481

↓

字482

↓

字483

↓

字484

↓

字485

↓

字486

↓

字487

↓

字488

↓

字489

↓

字490

↓

字491

↓

字492

↓

字493

↓

字494

↓

字495

↓

字496

↓

字497

↓

字498

↓

字499

↓

字500

↓

字501

↓

字502

↓

字503

↓

字504

↓

字505

↓

字506

↓

字507

↓

字508

↓

字509

↓

字510

↓

字511

↓

字512

↓

字513

↓

字514

↓

字515

↓

字516

↓

字517

↓

字518

↓

字519

↓

字520

↓

字521

↓

字522

↓

字523

↓

字524

↓

字525

↓

字526

↓

字527

↓

字528

↓

字529

↓

字530

↓

字531

↓

字532

↓

字533

↓

字534

↓

字535

↓

字536

↓

字537

↓

字538

↓

字539

↓

字540

↓

字541

↓

字542

↓

字543

↓

字544

↓

字545

↓

字546

↓

字547

↓

字548

↓

字549

↓

字550

↓

字551

↓

字552

↓

字553

↓

字554

↓

字555

↓

字556

↓

字557

↓

字558

↓

字559

↓

字560

↓

字561

↓

字562

↓

字563

↓

字564

↓

字565

↓

字566

↓

字567

↓

字568

↓

字569

↓

字570

↓

字571

↓

字572

↓

字573

↓

字574

↓

字575

↓

字576

↓

字577

↓

字578

↓

字579

↓

字580

↓

字581

↓

字582

↓

字583

↓

字584

↓

字585

↓

字586

↓

字587

↓

字588

↓

字589

↓

字590

↓

字591

↓

字592

↓

字593

↓

字594

↓

字595

↓

字596

↓

字597

↓

字598

↓

字599

↓

字600

↓

字601

↓

字602

↓

字603

↓

字604

↓

字605

↓

字606

↓

字607

↓

字608

↓

字609

↓

字610

↓

字611

↓

字612

↓

字613

↓

字614

↓

字615

↓

字616

↓

字617

↓

字618

↓

字619

↓

字620

↓

字621

↓

字622

↓

字623

↓

字624

↓

字625

↓

字626

↓

字627

↓

字628

↓

字629

↓

字630

↓

字631

↓

字632

↓

字633

↓

字634

↓

字635

↓

字636

↓

字637

↓

字638

↓

字639

↓

字640

↓

字641

↓

字642

↓

字643

↓

字644

↓

字645

↓

字646

↓

字647

↓

字648

↓

字649

↓

字650

↓

字651

↓

字652

↓

字653

↓

字654

↓

字655

↓

字656

↓

字657

↓

字658

↓

字659

↓

字660

↓

字661

↓

字662

↓

字663

↓

字664

↓

字665

↓

字666

↓

字667

↓

字668

↓

字669

↓

字670

↓

字671

↓

字672

↓

字673

↓

字674

↓

字675

↓

字676

↓

字677

↓

字678

↓

字679

↓

字680

↓

字681

↓

字682

↓

字683

↓

字684

↓

字685

↓

字686

↓

字687

↓

字688

↓

字689

↓

字690

↓

字691

↓

字692

↓

字693

↓

字694

↓

字695

↓

字696

↓

字697

↓

字698

↓

字699

↓

字700

↓

字701

↓

字702

↓

字703

↓

字704

↓

字705

↓

字706

↓

字707

↓

字708

↓

字709

↓

字710

↓

字711

↓

字712

↓

字713

↓

字714

↓

字715

↓

字716

↓

字717

↓

字718

↓

字719

↓

字720

↓

字721

↓

字722

↓

字723

↓

字724

↓

字725

↓

字726

↓

字727

↓

字728

↓

字729

↓

字730

↓

字731

↓

字732

↓

字733

↓

字734

↓

字735

↓

字736

↓

字737

↓

字738

↓

字739

↓

字740

↓

字741

↓

字742

↓

字743

↓

字744

↓

字745

↓

字746

↓

字747

↓

字748

↓

字749

↓

字750

↓

字751

↓

字752

↓

字753

↓

字754

↓

字755

↓

字756

↓

字757

↓

字758

↓

字759

↓

字760

↓

字761

↓

字762

↓

字763

↓

字764

↓

字765

↓

字766

↓

字767

↓

字768

↓

字769

↓

字770

↓

字771

↓

字772

↓

字773

↓

字774

↓

字775

↓

字776

↓

字777

↓

字778

↓

字779

↓

字780

↓

字781

↓

字782

↓

字783

↓

字784

↓

字785

↓

字786

↓

字787

↓

字788

↓

字789

↓

字790

↓

字791

↓

字792

↓

字793

↓

字794

↓

字795

↓

字796

↓

字797

↓

字798

↓

字799

↓

字800

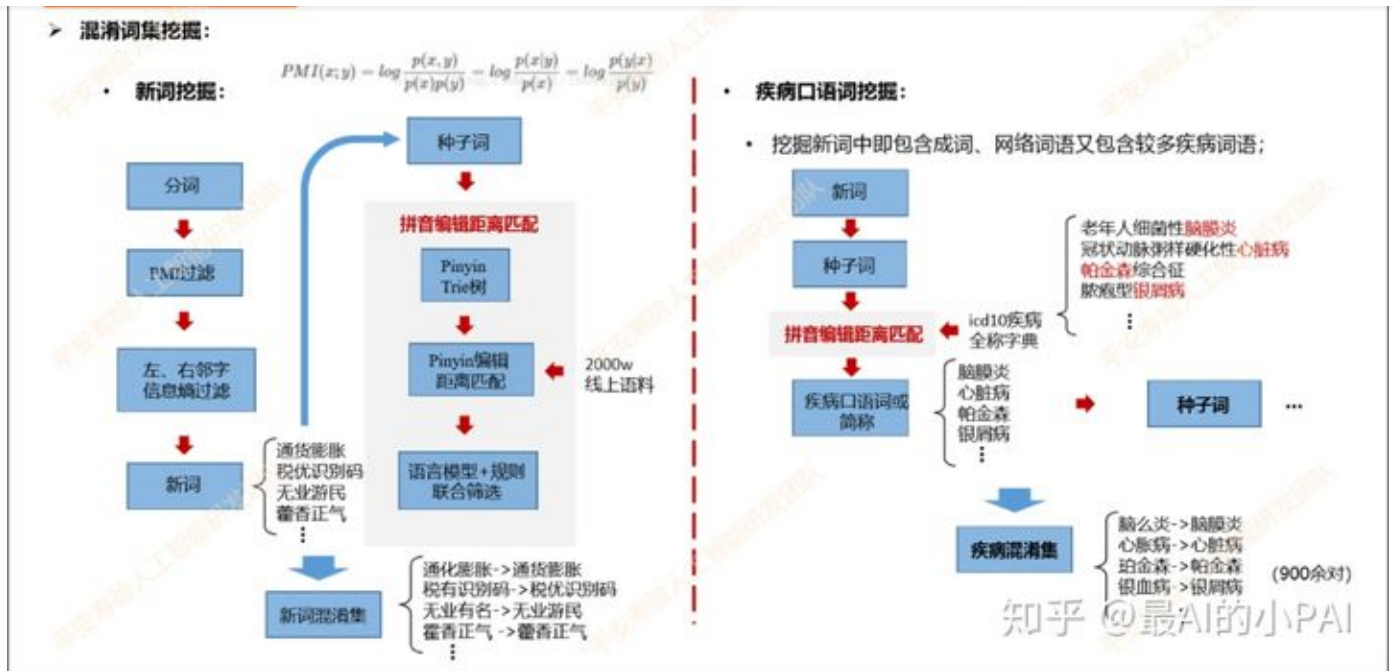
↓

字801

↓

字802

知乎

首发于  
平安寿险AI

### 3.3 候选排序

对于一项纠错任务来说，其正确结果是具有唯一性的，那么就需要通过排序算法对召回的候选词进行打分排序，从而选择分数最高的一项进行替换。然而，候选召回模块得到的候选词数量庞大，逐一通过复杂模型计算替换概率将引入较大的时间损耗，因此在精排序前需要进行一定粗排序，从而通过简单的算法来过滤掉部分明显错误的答案，其中我们采用的是逻辑回归模型。人工抽取的特征主要包括：频率比值、编辑距离、拼音jaccard距离、Ngram统计语言模型分数差值等。





变化、形音变化、PMI互信息变化、Ngram语言模型分数变化以及一些其他的基础特征。

### 3-技术落地：候选排序

二级排序：

模型：xgboost

作用：分数超过设定阈值且是Top1的作为最终候选

要求：正类（接受候选）准确度要很高

Query: 红痕狼仑算重疾吗？ 红痕狼疮

频次: 20 -> 1688

切词: 红\痕\狼\疮 (1\1\2) -> 红痕狼疮 (4)

nn语言模型: 痕 (<0.001) -> 斑 (0.979)

4gram语言模型: -19.2 -> -10.6

PMI:红痕(0.33) ->红斑(9.7)

Query: 暖圆孔未闭可以投保吗？ 卵圆孔未闭

拼音: 暖 (nuan) -> 卵 (luan) 声母不一致

拼音Jaccard距离: 0.25

Query: 旺财信息可以册除吗 删除

五笔: 册 (MMGD) -> 删 (MMGI)

Query: 油菜和普才计划的区别？ 优才

	词频	油菜	优才
寿险领域		5	428
通用领域		1190	87

局部特征

切词变化  
(短语个数, 单字个数, 含错字片段长度)

PMI变化  
(最小值, 最大值)

频次变化  
(本身频次变化, 与上下文组成2gram频次变化)

4gram语言模型变化  
(句子概率, 片段概率)

形音变化  
(全\简拼音韵母变化, Jccard距离, 五笔变化)

其他变化  
(停用词\错字位置, 候选来源等)

全局特征

Cbow-LM

LSTM-Attention-LM

BERT-LM

XGBOOST

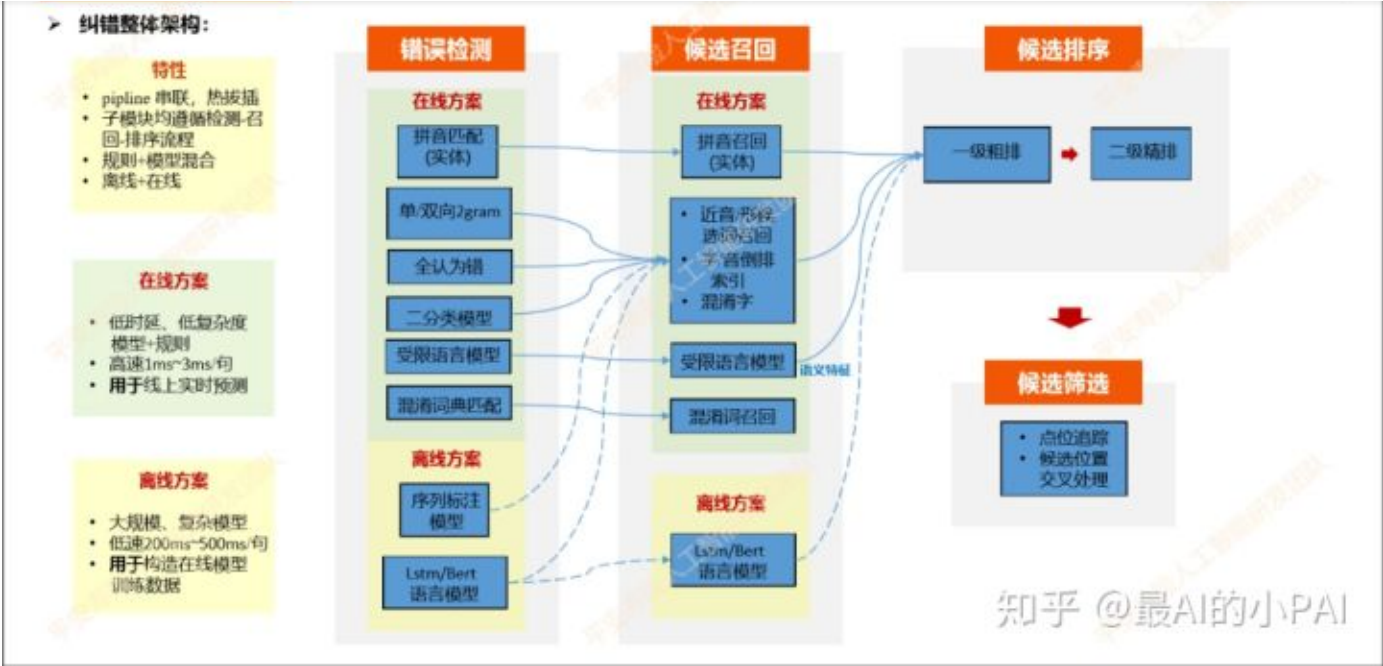
大于设定阈值Top1

知乎@最AI的小PAI

### 3.4 具体实现路径

下面我们对以上工作进行总结。整体的纠错框架可以分为“错误检测”、“候选召回”、“候选排序”三个主要流程，其中的每个流程我们都尝试了不同的方法，而不同方法间进行组合又可衍生出多种纠错方法。

在实际应用中，我们针对垂域中不同类型的错误采取了不同的纠错方法，方法间采用Pipeline串联的方式，使得不同的模块可以相互补充同时也方便模块线上的热拔插。根据算法的时效性我们将各类算法分为在线方案、离线方案两部分，其中离线方案具有较高的准确性但耗时较高，因此可通过离线方案进行原始数据去噪处理，来提升线上模型效果；而相对离线方案来说在线方案具有极高的时效性，线上平均每个Query的处理时间是1-3ms。



四、效果评估

线上我们分别对比了所提方案与Pycorrector的纠错性能。相较于Pycorrector，本系统的误报率为0.1%，召回率为70%，单句平均耗时在1.5ms左右，远高于开源项目。其中，系统仍存在一些badcase，如“好晚了”被误纠为“好玩了”等，此类badcase大都因为模型缺陷导致，其无法处理上下文语句、长依赖问题以及知识关联问题。

4-效果评估:

pai 平安寿险AI

效果	Far(过纠率)	Recall(召回率)	耗时/句
pycorrector	>50%	11%	none
在线纠错v0.5	0.1%	70%	1.5ms

类型	原句子	纠正信息	方法/原因分析
正例	平安福与 <b>大复兴</b> 有什么区别	大福星	产品名称-近音召回
	平安福承保多少种 <b>重疾线</b>	重疾险	保险名词-近音召回
	<b>小山羊</b> 可以投保安心百分百吗	小三阳	疾病简称-近音召回
	<b>省份</b> 正变更	身份证	保险名词-近音召回
	你说 <b>鑫祥</b> 和 <b>诛仙</b> 是平等的吗	重疾 主险	保险名词-近音召回
	<b>关心</b> 并可以投保吗	冠心病	疾病名称-近音召回
	少儿平安福可以加 <b>脱保</b> 人豁免吗	投保人	保险名词-倒排召回
Badcase	<b>安心保</b> 的功能是什么	安鑫保/安心宝	缺乏上下文语境, 无法判断哪个是正确的
	腰椎间盘突出 <b>出头</b>	未纠成 投保	字级别语义破坏
	哎, 我好困哪, 好 <b>晚了</b>	过纠成 玩	缺少相关语料, cbow语言模型无法学习长距离依赖
	帮我查一查 <b>何可意</b>	过纠成 何可以	人名识别模块无法识别
	在 <b>南山</b> 平安金融中心入职	未纠成 福田	缺少知识关联

知乎 @最AI的小PAI

知乎

首发于  
平安寿险AI

本系统的优点可以概括为以下几个方面：

1. 方便扩展与领域迁移，对于新领域只需重新挖掘无监督数据即可；
2. 系统架构方便拔插特殊编写的纠错子模块，方便后续的优化与开发。

同时也存在如下缺点：

1. 难以应用于通用领域纠错；
2. pipeline的机制导致错误逐级放大；
3. pipeline同样导致串联链越长则耗时越长。

未来，当具有充足的标注语料时，需要强化上下文/全局语义理解，可训练深度学习模型提高系统召回性能；同时，基于神经机器翻译模型端到端纠错方法也是一种新的纠错思路。后续如保险垂域扩充了足够大的标注样本时，可尝试应用机器翻译算法优化整体纠错系统。另外，针对知识关联错误后续可尝试利用知识图谱完善整个纠错任务。

更多干货内容欢迎关注「平安寿险PAI」（公众号ID：PAL-AI）。

编辑于 2020-07-13

自然语言处理

## 文章被以下专栏收录



平安寿险AI

公众号：平安寿险PAI；促进交流碰撞，与AI共成长

关注专栏

## 推荐阅读

▲ 赞同 30 ▼    8 条评论    分享    喜欢    收藏    申请转载    ...

知乎

首发于  
平安寿险AI

## 【年终总结】2019年有三AI NLP做了什么，明年要做什么？

小Dream哥

Jianfeng Gao  
Microsoft Research  
jgao@microsoft.comMichel Galley  
Microsoft Research  
mgalley@microsoft.comLihong Li  
Google Brain  
lihong@google.com

## 对话AI综述

forPr...

发表于NLP

BI  
Tr  
数

### 8 条评论

⇌ 切换为时间排序

写下你的评论...



SimpleZqb

2020-07-23

请教BERT加入拼音和五笔特征，待预测的字只有token被MASK，而拼音和五笔特征没有被MASK吗？

👍 赞



clq 回复 SimpleZqb

2020-07-24

是的

👍 赞



SimpleZqb

2020-10-13

拼音特征就是拼音挺好理解的，挺好奇想知道这个字型特征是什么，有开源项目吗？谢谢

👍 赞



clq 回复 SimpleZqb

2020-12-24

字型特征可以看爱奇艺faspell

👍 1



想不出来名字

2020-11-26

请问近音、近形和混淆字表有开源项目可以分享吗？

👍 赞



clq 回复 想不出来名字

2020-12-24

▲ 赞同 30 ▼

💬 8 条评论

➦ 分享

❤️ 喜欢

★ 收藏

📄 申请转载

...



Irie仔储研首，怎么就能解决编辑距离的问题

 赞



clq 回复 Lee-90

2020-12-24

参考elasticsearch的模糊查询算法

 赞