文本纠错的探索与实践

分享人：陈乐清

平安人寿人工智能研发团队

CONTENTS

1-背景介绍：

平安寿险AI

- 互联网新媒体和社交用语->错误率在**2%以上**
- 语音识别->错误率**8%-10%**
- 平安人寿ASKBOBO机器人->错误率达到**9%**

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 → 癜
- 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

垂域客服机器人

特点：领域受限
缺乏点击数据（无监督）

2-研究现状：*pycorrector*

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\国藉\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：国藉 ➔ 国籍
- 候选召回
  - 近音字典替换错误位置　藉\ji ➔ 籍,际,集，…
  - 近形字典替换错误位置　藉➔ 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(韩国国籍投保)　P(韩国国际投保)

    P(韩国国积投保)　P(韩国国藕投保)　…

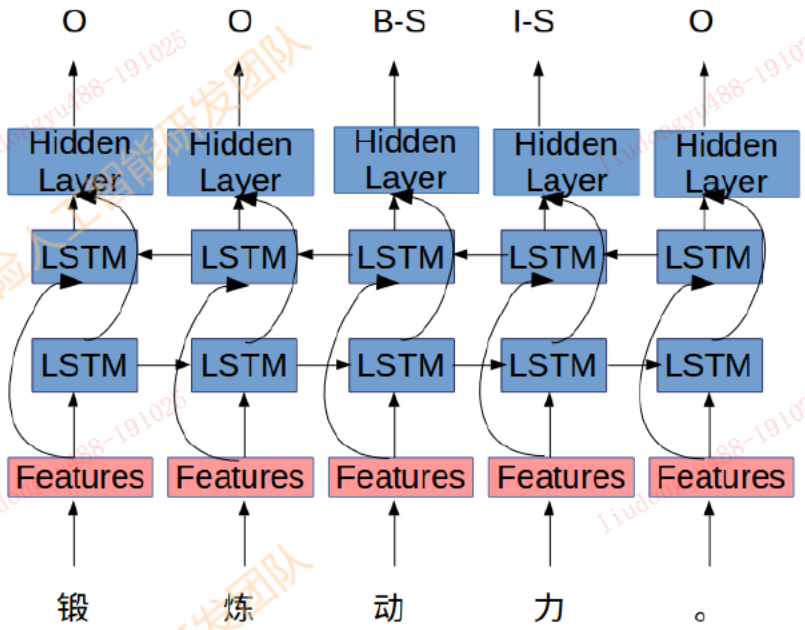
2-研究现状：*学术界进展*

平安寿险AI

基于序列标注的检错方案：

《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

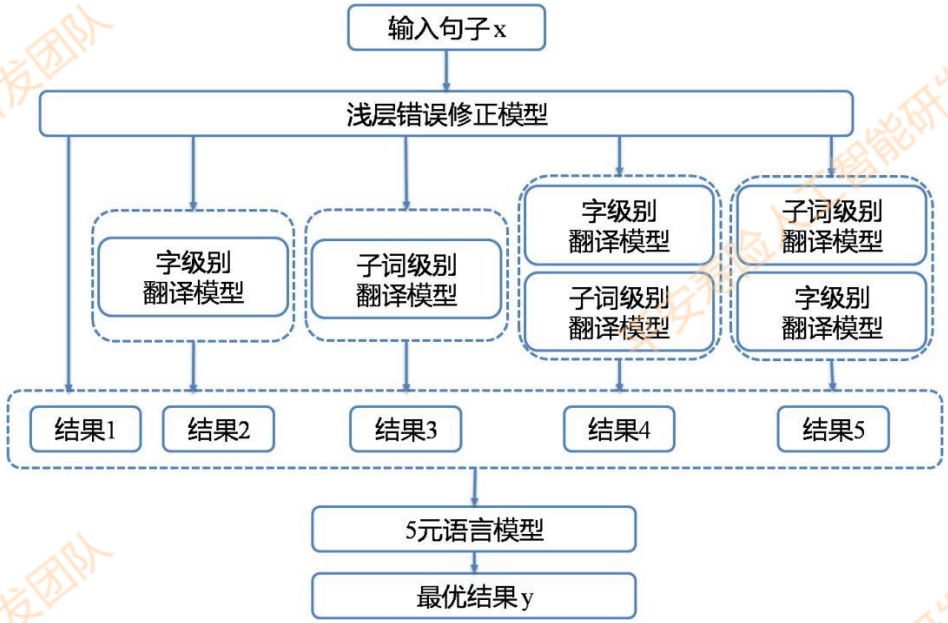
- **Position Level：**

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
0.34	0.18	0.29

2-研究现状：*平安人寿问答纠错项目现状*

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

2-研究现状：*纠错指标参考*

平安寿险AI

- **评价指标：**
  - 过纠率/误报率：

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

  - 召回率：

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

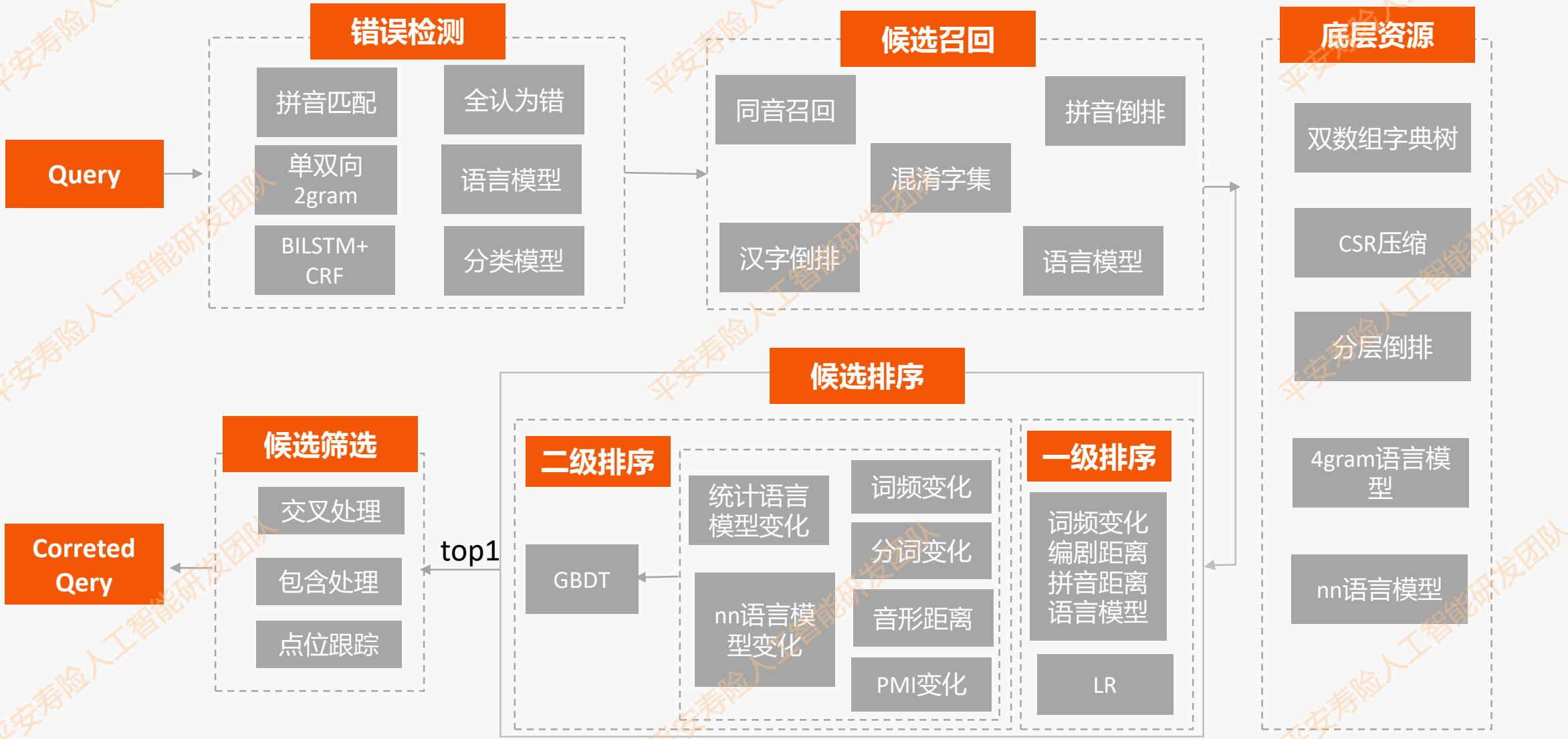
- **纠错目标：** 被改正的句子数 >> 被改错的句子数

$$K * \mathrm{RECALL} \gg (1-K) * \mathrm{FAR}$$

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<0.2%，尽量提高RECALL		

3-技术落地：*平安人寿问答纠错模块架构*

平安寿险AI

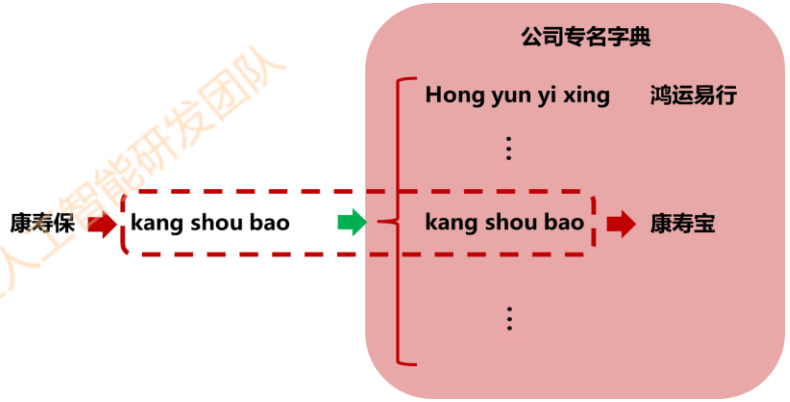


3-技术落地：*错误检测*

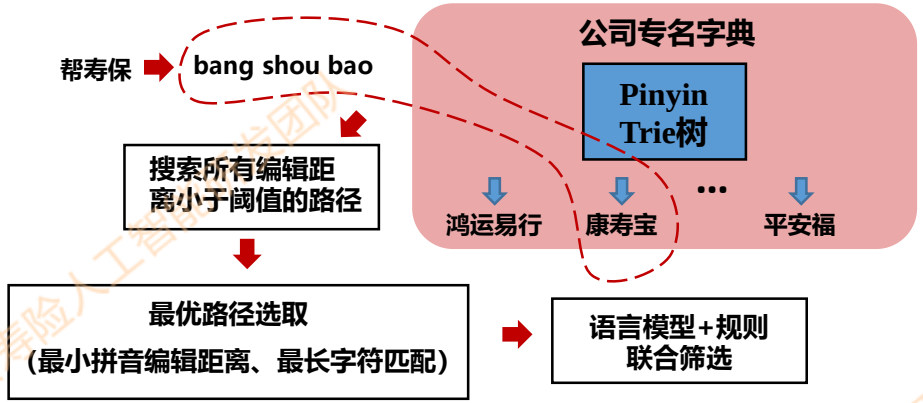
平安寿险AI

➢ 基于规则的错误检测

拼音匹配检测：适合于实体错误检测。比如产品、疾病等实体，需要实现维护好拼音-实体映射字典。



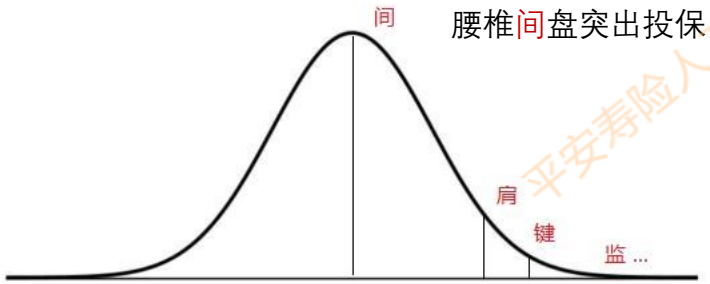
拼音编辑距离检测：



单双向2gram检测：

当前词与上下文组成的2gram词频次很低，认为有错

- 假设：正确表述发生频次要远远大于错误表述发生次数



静脉拴<30　拴塞<30

右小腿/静脉/拴/塞/投保

平安暖宝宝投保 -> 保

3-技术落地：*错误检测*

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

安 心 白 分 百 的附加险

an xin bai fen bai

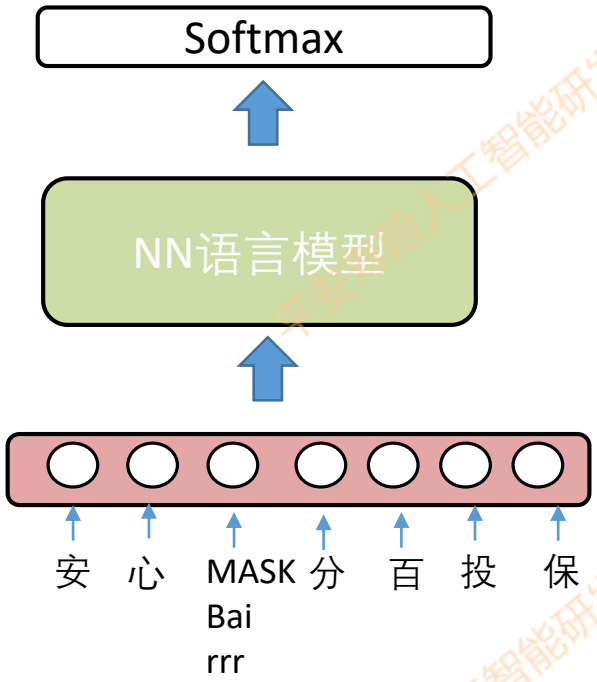
rrr

百-0.57 白-0.32 受限前

百-0.94 白-0.05 受限后

安心白分百投保 → 百

百-0.94;白-0.05;拜-0.0034;....



3-技术落地：*错误检测*

平安寿险AI

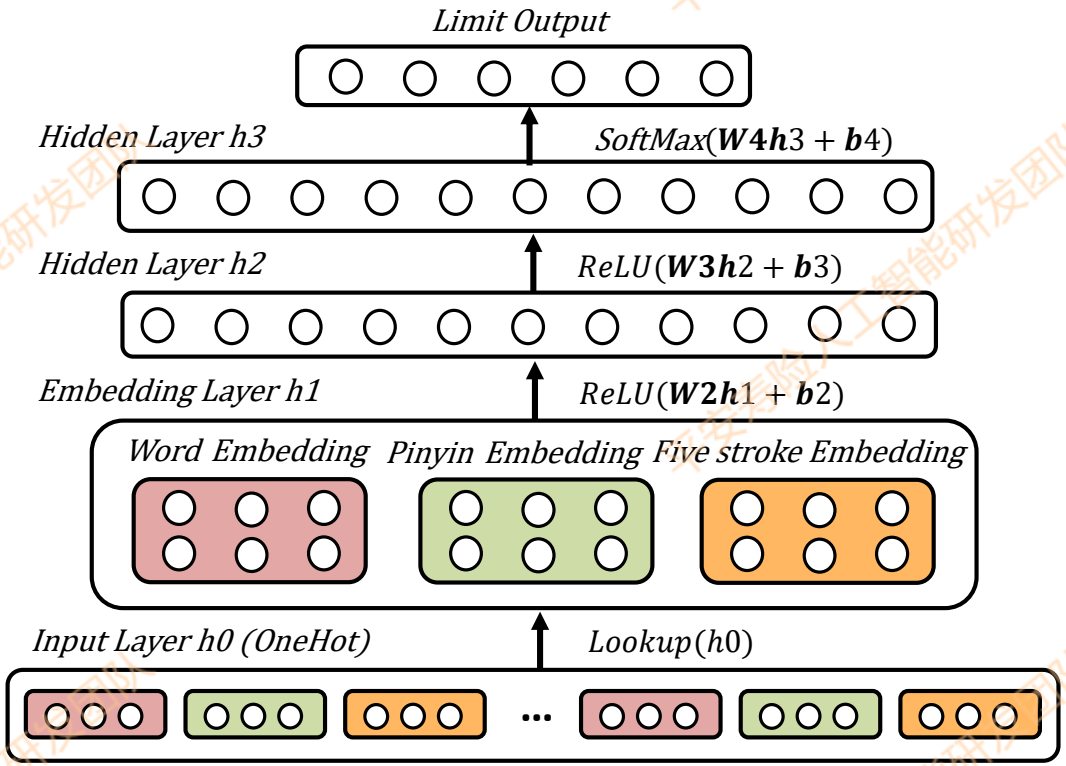
- **基于word2vec-cbow改造的音字混合受限字表语言模型错误检测算法**

  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 -> 玩



3-技术落地：*错误检测*

平安寿险AI

- **基于BiLstm改造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：基于字符级别nn网络的中文拼写错误检测和识别》，2019

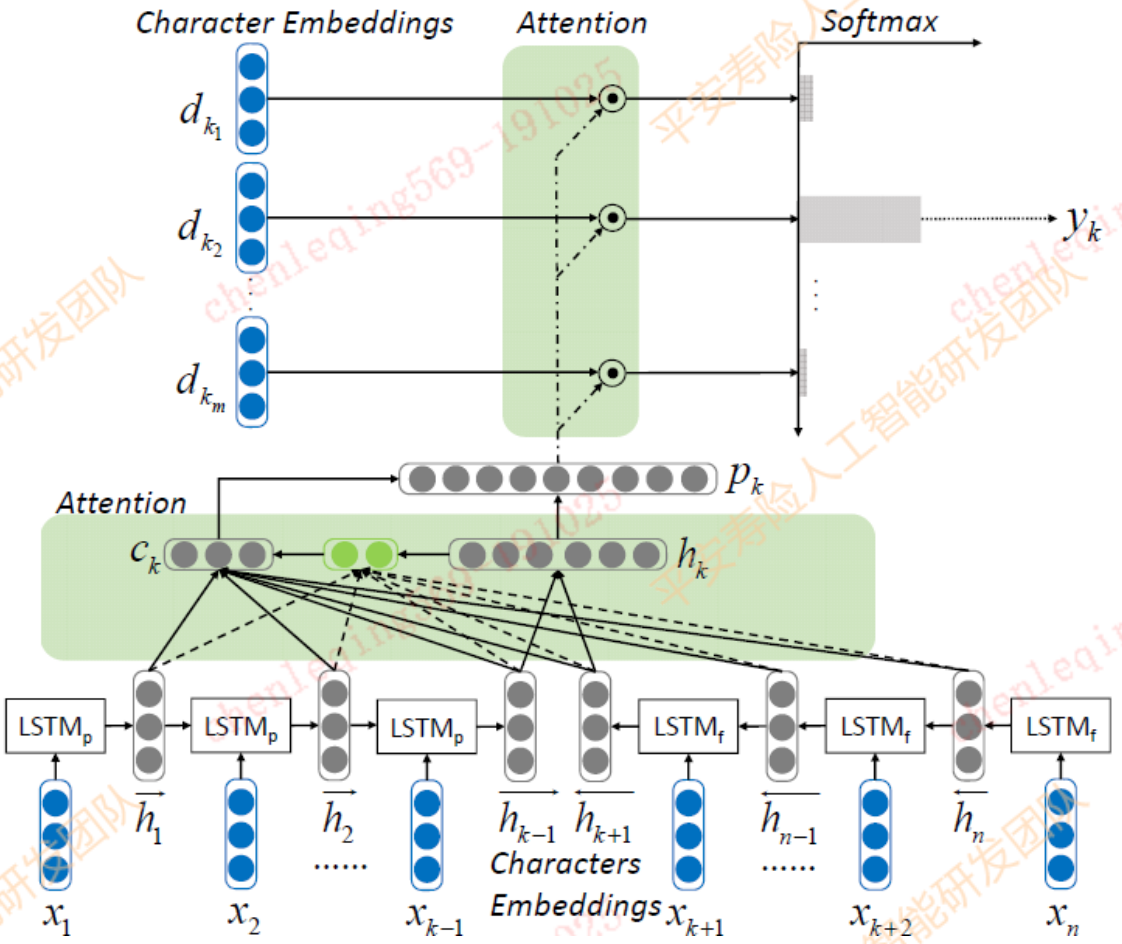
Automatic Proofreading in Chinese : Detect and Correct Spelling Errors in Character-Level with Deep Neural Networks，2019

二层注意力机制

- 第一层：hk与其他字的输出向量hs做attention，如果候选词领近的字也是错别字，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ck与hk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

3-技术落地：*检索和深度语义匹配-BERT for QA*

平安寿险AI

- **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

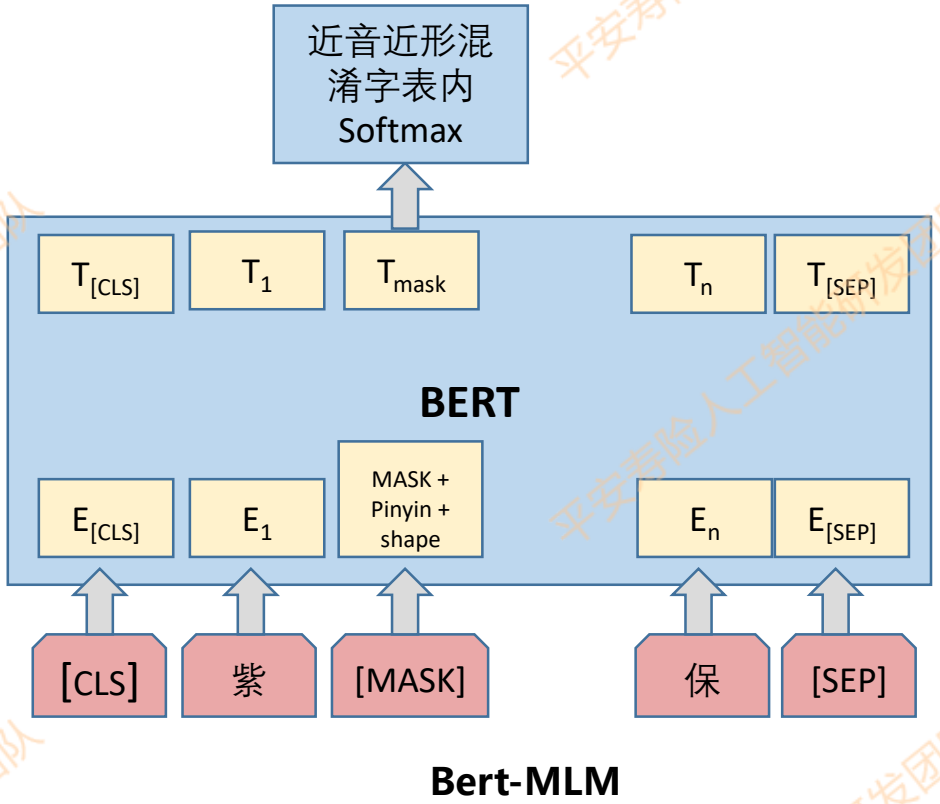
改造huggingface的代码，输入层加入拼音，字形特征

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



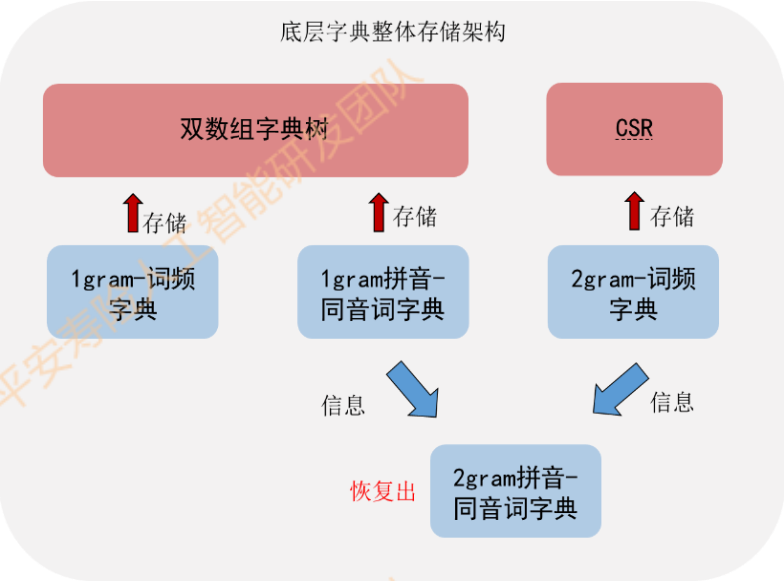
Bert-MLM

3-技术落地：*候选召回*

平安寿险AI

➢ 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



腰椎/肩/盘/突出

↓

腰椎肩盘

↓

yao zui jian pan

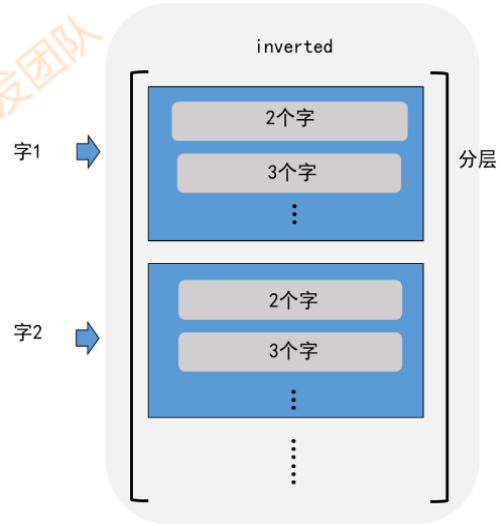
↓ 2gram

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



费　交费,保费,缴费,...
费　交费单, 缴费期,交通费,...
期　准备期,缴费期, 合同期,...

缴会期

↓

(交费单, 缴费期,...)

×

(准备期,缴费期, ...)

↓

缴费期

- 降低每层词集的空间从而降低索引时间
- 灵活地搜索任意编辑距离的候选词集

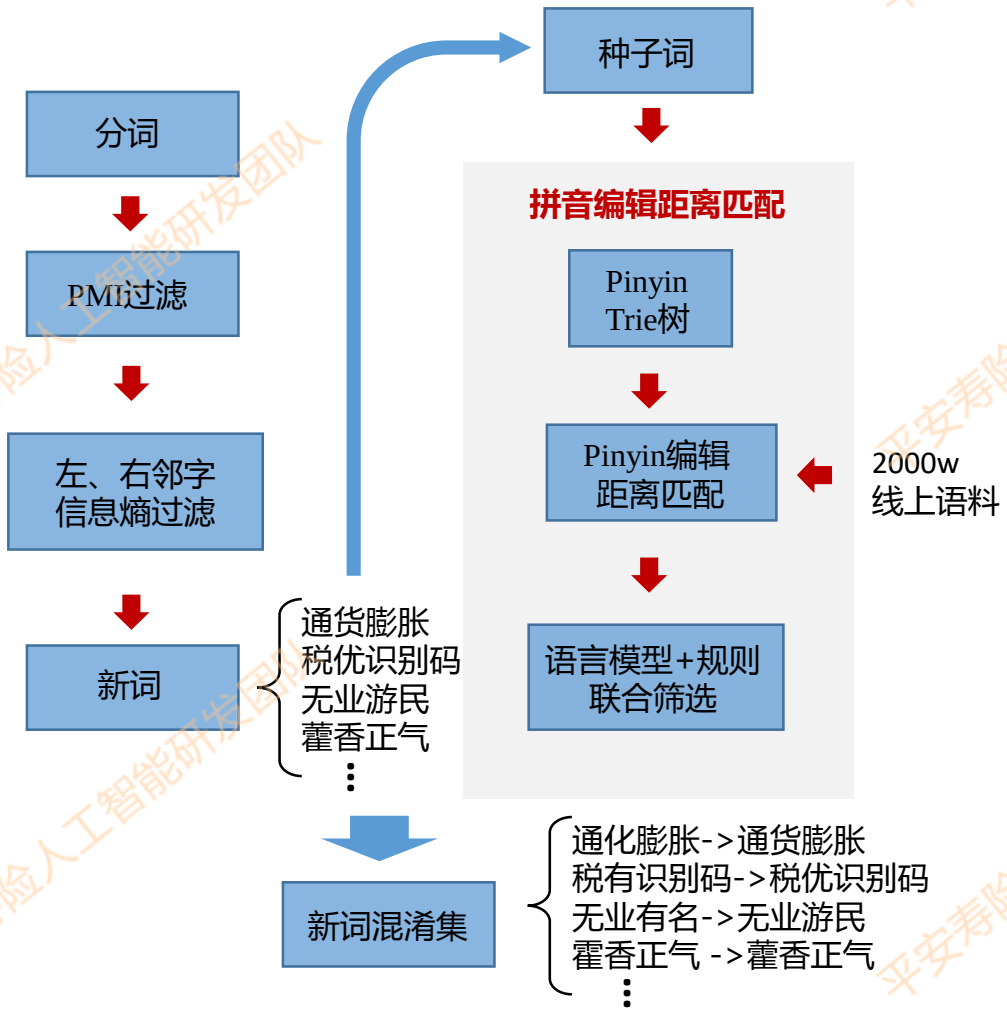
3-技术落地：*候选召回*

平安寿险AI

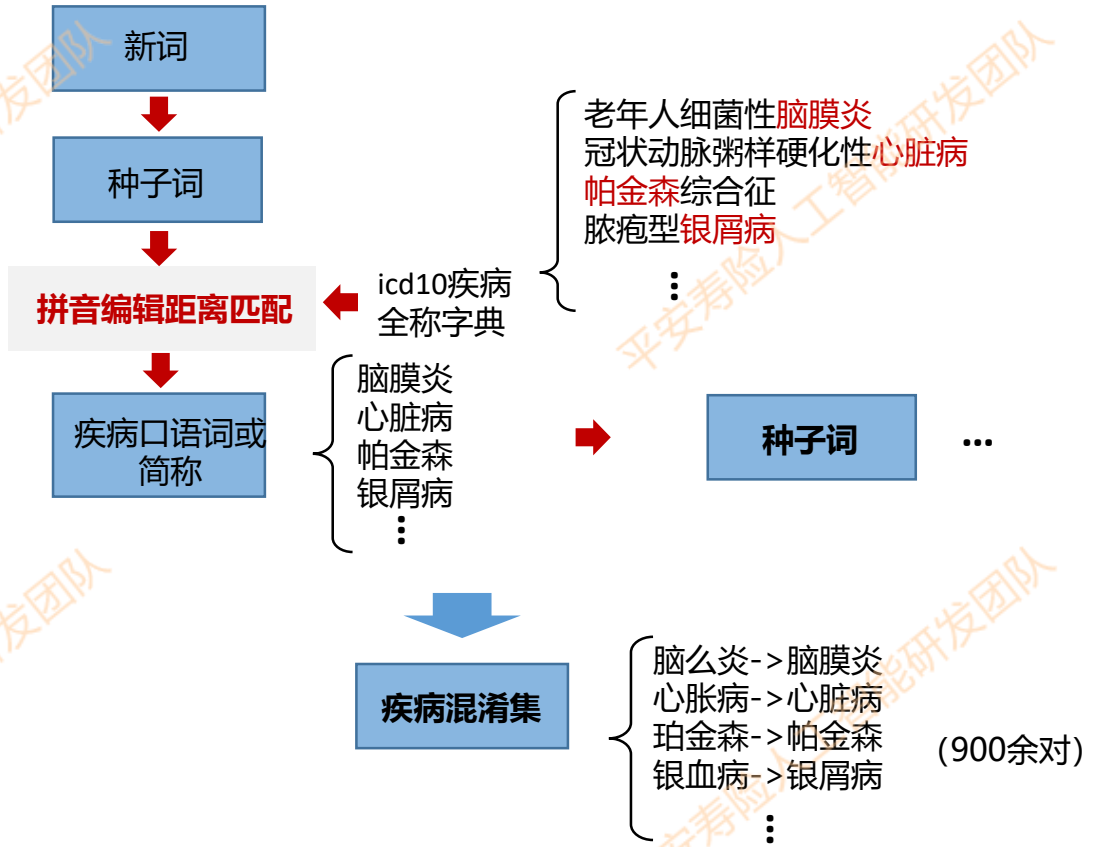
➢ **混淆词集挖掘：**

$$PMI(x;y) = log\frac{p(x,y)}{p(x)p(y)} = log\frac{p(x|y)}{p(x)} = log\frac{p(y|x)}{p(y)}$$

- **新词挖掘：**



- **疾病口语词挖掘：**
  - 挖掘新词中即包含成词、网络词语又包含较多疾病词语；



3-技术落地：*候选排序*

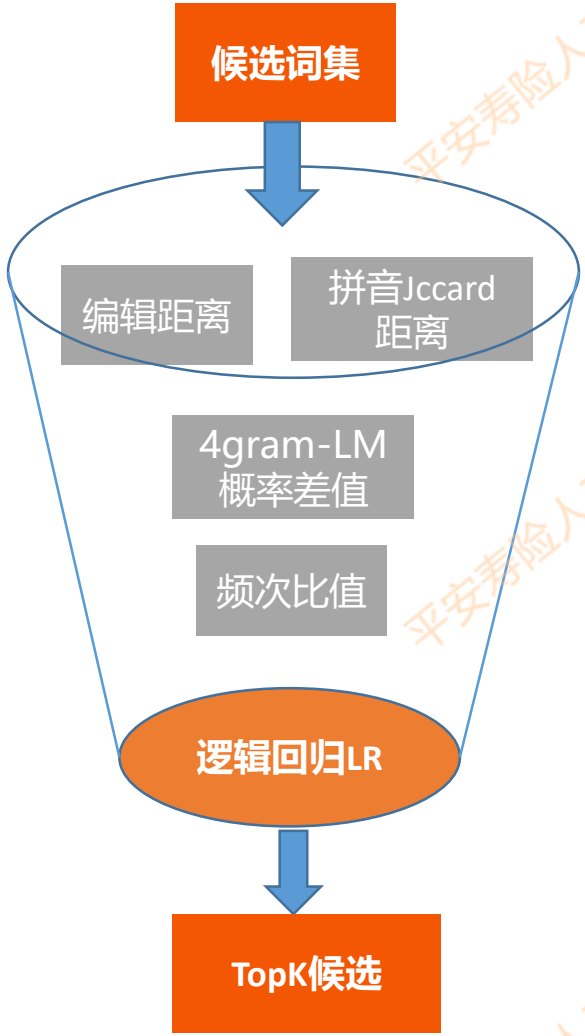
平安寿险AI

- **一级排序：**
  - 模型：逻辑回归LR
  - 作用：二级精排较为耗时，需要一级初筛选Topk进入二级
  - 要求：正类（接受候选）召回率很高，运行速度快
  - 特征：
    - 频次比值：候选频次越高分数越高
    - 编辑距离：编辑距离越小分数越高
    - 拼音Jaccard距离：拼音越相近分数越高
    - 4gram-LM概率差值: 候选替换后句子越通顺分数越高

Example

Query：得了甲状腺姐姐可以投保吗？ ->结节 0.99

- 频次： 55 -> 94
- 编辑距离： 2
- 拼音Jaccard距离： 0
- 语言模型概率： -21.9 -> -8.6



3-技术落地：*候选排序*

平安寿险AI

- **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

Example

Query： 红癍狼仓算重疾吗？ 红斑狼疮

- 频次：20 → 1688
- 切词：红\癍\狼疮 (1\1\2) -> 红斑狼疮 (4)
- nn语言模型：癍 (<0.001) -> 斑 (0.979)
- 4gram语言模型：-19.2 -> -10.6
- PMI:红癍(0.33) ->红斑(9.7)

Query: 暖圆孔未闭可以投保吗？ 卵圆孔未闭

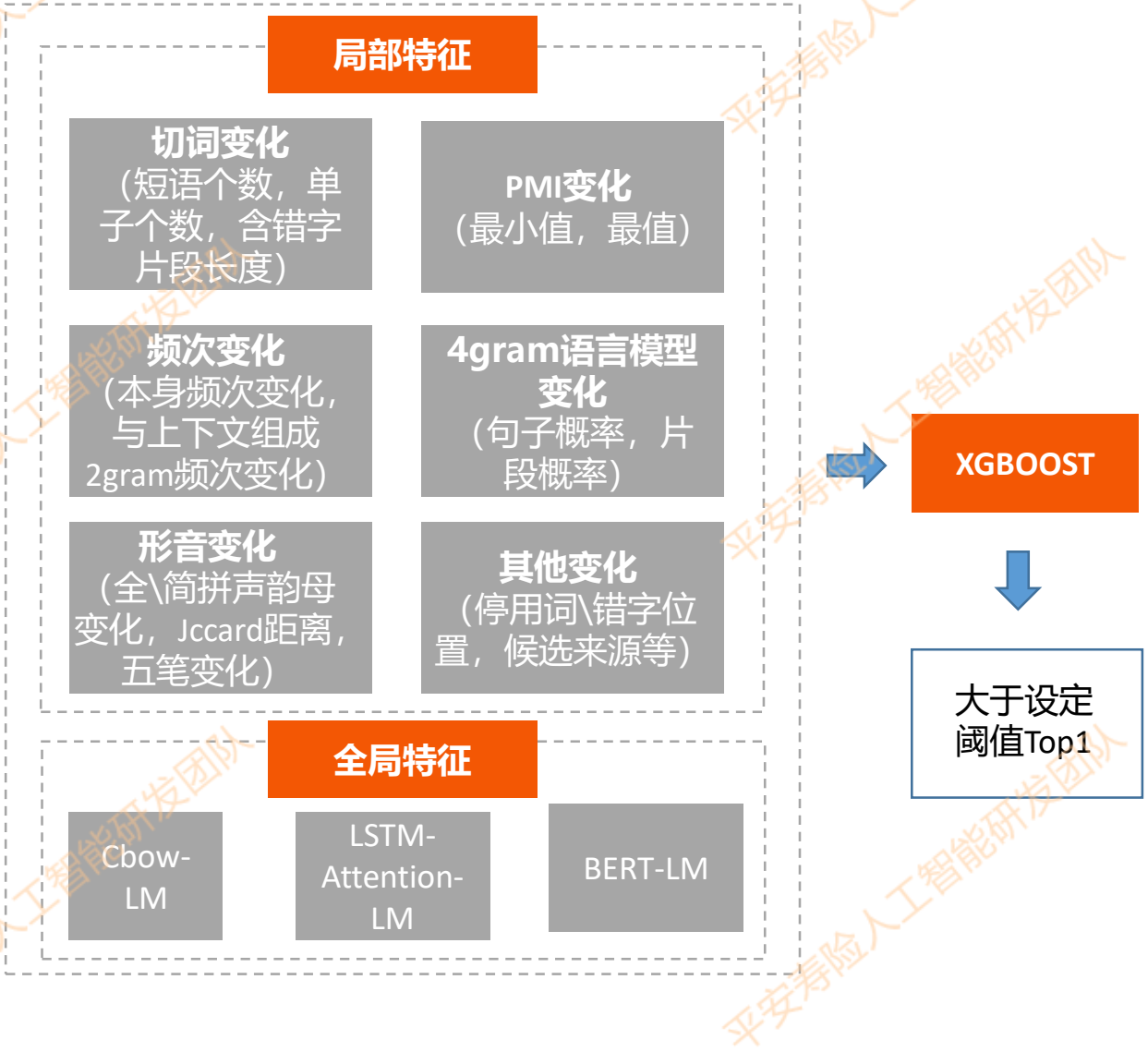
- 拼音： 暖 (nuan) -> 卵 (luan) 声母不一致
- 拼音Jaccard距离：0.25

Query: 旺财信息可以册除吗 删除

- 五笔：册 (MMGD) -> 删 (MMGJ)

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



3-技术落地：*实现路径*

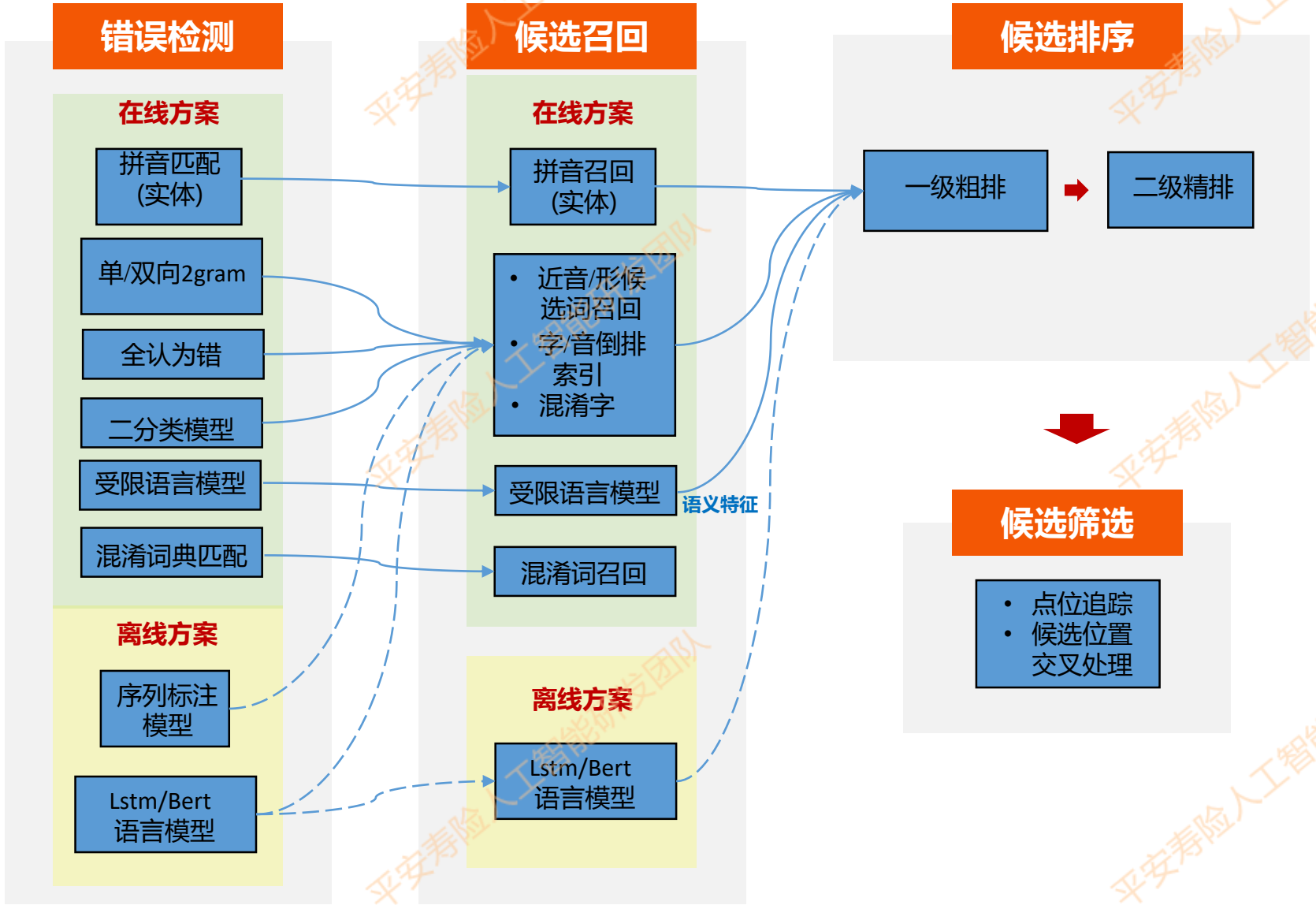
平安寿险AI

➢ **纠错整体架构：**

特性
- pipline 串联，热拔插
- 子模块均遵循检测-召回-排序流程
- 规则+模型混合
- 离线+在线

在线方案
- 低时延、低复杂度模型+规则
- 高速1ms~3ms/句
- **用于**线上实时预测

离线方案
- 大规模、复杂模型
- 低速200ms~500ms/句
- **用于**构造在线模型训练数据



4-效果评估：

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

5-总结与改进方向：

平安寿险AI

- 优点：
  - 无监督，方便将该方法迁移到其他垂域，只需重新无监督挖掘数据；
  - 系统架构很方便拔插特殊编写纠错子模块
- 缺点：
  - 很难迁移到通用领域中
  - Pipeline导致错误逐级传递
  - Pipeline链越长耗时越大
- 改进方向：
  - 强化上下文/全局的语义理解
  - 训练语料去燥处理
  - 探索端到端的算法，如NMT(神经机器翻译)
  - 探索语法错误（多字少字乱序）相关算法和工程实现
  - 知识关联

开放，共赢

平安寿险AI

《Understanding Error Correction and its Role as Part of the Communication Channel in Environments composed of Self-Integrating Systems》

Error correction for high-level languages can demand a large amount of a prior knowledge which is usually too large to be provided as additional information by the sender and hence is provide directly to the receiver`s poll

加入AI技术交流群

扫左方二维码，添加小PAI助手号，备注“直播”即可加入社群

技术干货、产品应用独家分享，招聘信息最新速递...

期待更多交流与碰撞，一起AI~

Thank You for Listening

谢谢您的聆听