58搜索拼写纠错

原创 王成焱，赵国柱 58技术 2019-06-27

在搜索引擎中，用户希望得到和输入查询词相关的并且质量较好的网页或文档。但是往往出于各种原因，用户输入的查询词本身质量不高或是错误的，如果搜索引擎不对这种错误进行修正弥补，会导致召回错误的结果，或者结果数少甚至没有结果。

拼写纠错，主要应用于58集团内部主站搜索中，可以对用户输入的房产，招聘，黄页，二手车等类别搜索质量不佳的查询词进行纠错改写，以能返回更好的搜索结果。拼写纠错能够挽回错误输入导致的流量损失，对于提升用户体验起到重要作用，在58搜索引擎中是非常重要的一个模块。

拼写常见错误

在58的搜索场景下，用户输入查询词较短，较普遍使用拼音输入法输入拼音选择候选词，在部分方言中，一些音节是与普通话发音不同的(例,zh=>z、ang=>an、fei=> hui), 用户发音的地域特色常出现模糊音问题。另外由于移动端屏幕小，用户误选或未选，也很容易出现错误，同时伴随有手写及语音输入等其他输入方式，输入随意，也是错误的来源。

整体上看，常见的查询词拼写错误类型包括:1)纯拼音;2)错字母或缺字母的英文;3)拼音汉字混合。拼音类型的错误包含同音别字、模糊音别字, 汉字方面包含错字、缺字错误。

纠错方案

基于规则的纠错方法

是根据用户输入错误的类型进行设计的，通过离线生成错误词到正确词的索引，线上纠错时用查询词或者处理后的查询词查询索引，如果索引中存在查询词到正确词的索引，则对查询词纠错为正确词，否则不做纠错处理。

基于规则的纠错方法离线使用包含58集团服务领域全部词条的语料文件和搜索日志统计的搜索热词来生成索引文件，主要生成两个索引词典，拼音索引文件和编辑距离索引文件。遍历原始文件的每个词条，生成关键词的全拼，简拼，尾部不完整拼音，拼音的模糊音变换到关键词的索引，保存到拼音索引文件；生成关键词的每个位置的字删掉后余下的字串到关键词的索引，保存到编辑距离索引文件。

拼音索引可以纠错查询词是全拼，简拼，拼音汉字混合，不完整拼音，同音错别字，模糊音错别字的错误，编辑距离索引可以纠错查询词是错字，缺字的错误。
对于很长的全拼音查询词，索引文件中没有对应的索引词条，可以从查询词右侧开始和拼音索引文件匹配最长的拼音，如果找到匹配的拼音，将匹配的拼音截掉后继续进行前述过程，一直到处理完全部查询词。最后将匹配的拼音索引的词组合在一起作为纠错结果。

基于统计语言模型的纠错方法

用于纠错长查询词。离线使用包含58集团服务领域全部词条的语料文件和搜索日志统计的搜索热词生成语言模型。线上处理时，对查询词先进行分词，然后将相邻词组合在一起进行拼音纠错或编辑距离纠错，对纠错得到的词进行分割，这样原始查询词的每个分词会得到若干个候选词，然后使用维特比算法找到每个分词对应的候选词的最优组合，最优组合的标准是每个候选词的3-gram条件概率和最大。

以下是纠错实现的详细描述。

离线词典生成

离线使用语料文件生成拼音索引文件，编辑距离索引文件，用于检查候选词正确性的词频文件，ngram模型文件。

(1)拼音词典：读取语料文件中的每一个词条的关键词，生成四种索引：关键词的所有汉字转换为对应的拼音到原始关键词的索引；关键词的所有汉字转换为对应拼音的首字母到原始关键词的映射；关键词的所有汉字转换为对应的拼音，其中的卷舌音转换为非卷舌音，非卷舌音转换为卷舌音，进行任意组合后形成的词到原始关键词的映射；关键词的所有汉字转换为对应的拼音，最后一个汉字的拼音只保留首字母生成的词到原始关键词的映射。如下例所示：

索引	原始关键词	索 引 类型	可用于纠错的词示例
ershoudiannao	二手电脑	1	ershoudiannao，二手diannao，二手点脑
esdn	二手电脑	2	esdn
ersoudiannao	二手电脑	3	ersoudiannao，二搜电脑
ershoudiann	二手电脑	4	ershoudiann，二手电n

(2)编辑距离词典：读取语料文件中的每一个词条的关键词，生成两种索引：省略关键词中的每个字生成的字串到原始关键词的索引，并记录省略掉的字在原始关键词中的位置；将词条中的所有汉字

转换为拼音，省略掉拼音中的每个字母生成的字符串到原始关键词的索引，并记录省略掉的字母在词条中的位置。如下表的例子所示：

索引	省略字的位置	原始关键词	索引类型	可用于纠错的词示例
二手电脑	-1	二手电脑	1	无纠错
手电脑	1	二手电脑	1	手电脑，无手电脑
二电脑	2	二手电脑	1	二电脑，二无电脑
二手脑	3	二手电脑	1	二手脑，二手无脑
二手电	4	二手电脑	1	二手电，二手电无
Ershoudiannao	-1	二手电脑	2	无纠错
Rshoudiannao	1	二手电脑	2	rshoudiannao
Eshoudiannao	2	二手电脑	2	eshoudiannao
Erhoudiannao	3	二手电脑	2	erhoudiannao
…	…	…	。	…

(3)ngram模型：统计语言模型纠错方法使用3-gram模型。使用开源工具Srilm训练ngram模型。首先对语料文件的每个词条的关键词分词，将分词用空格分割后保存；用Srilm先统计相邻3个分词组合的数目，生成ngram计数文件，然后统计每个组合的概率，生成语言模型。Srilm在训练模型时可以对数据进行平滑。模型可以用来评估一个词组的合理性，或者说可以用来评估两个字符串之间的差异程度。ngram模型的基本原理是一个句子中的每个词出现的概率只与其前面出现的词有关。可以计算每个词和其前面词组合在一起的条件概率，用以计算一个词序列组成词组的整体概率，如下公式所示：

$$P(\omega_1, \omega_2, \dots, \omega_m) = P(\omega_1)P(\omega_2|\omega_1)P(\omega_3|\omega_1\omega_2) \dots P(\omega_m|\omega_1, \dots, \omega_{m-1})$$

利用马尔科夫链假设，当前词仅仅跟前面几个有限的词相关，因此就不必追溯到最开始的那个词，这样可以大幅缩减上诉算式的长度。3-gram模型中当前词仅仅跟前面2个词相关，如下公式所示：

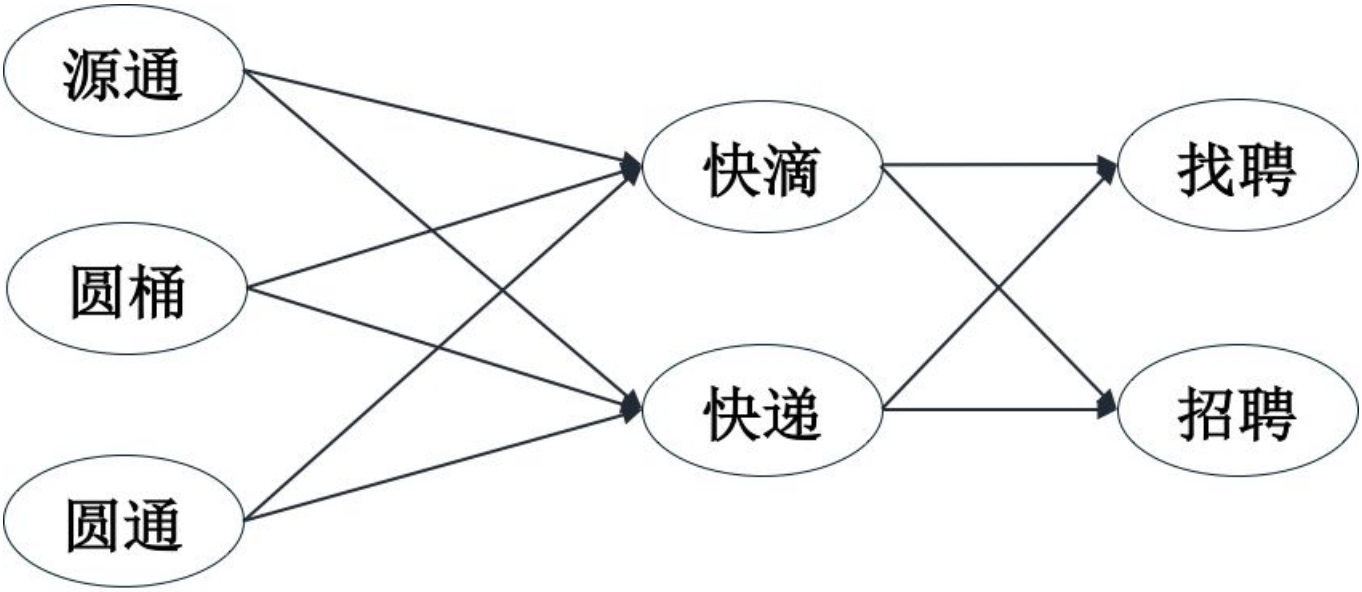
$$P(\omega_1, \omega_2, \dots, \omega_m) = \prod_{i=1}^{m} P(\omega_i|\omega_{i-2}\omega_{i-1})$$

在本纠错服务中当索引文件中不存在与查询词对应的词条时，可以先对输入词进行分词，然后相邻词组合在一起进行分段纠错，这样的将相邻词组合在一起可以利用每个词的上下文信息。分段纠错后会产生很多候选的纠错结果，最后从每段选出一个候选词组合成最优的结果。

在本纠错方法中计算出各组合中3-gram概率和最大的一个组合作为最优的组合。假设查询词是“源通快滴找聘”，经过分段纠错得到如下图所示的候选词，候选词组合“源通快滴 找聘”的3-gram概率和为：

P（源通|S,S）+ P（快滴|源通,S）+ P（找聘|源通,快滴）

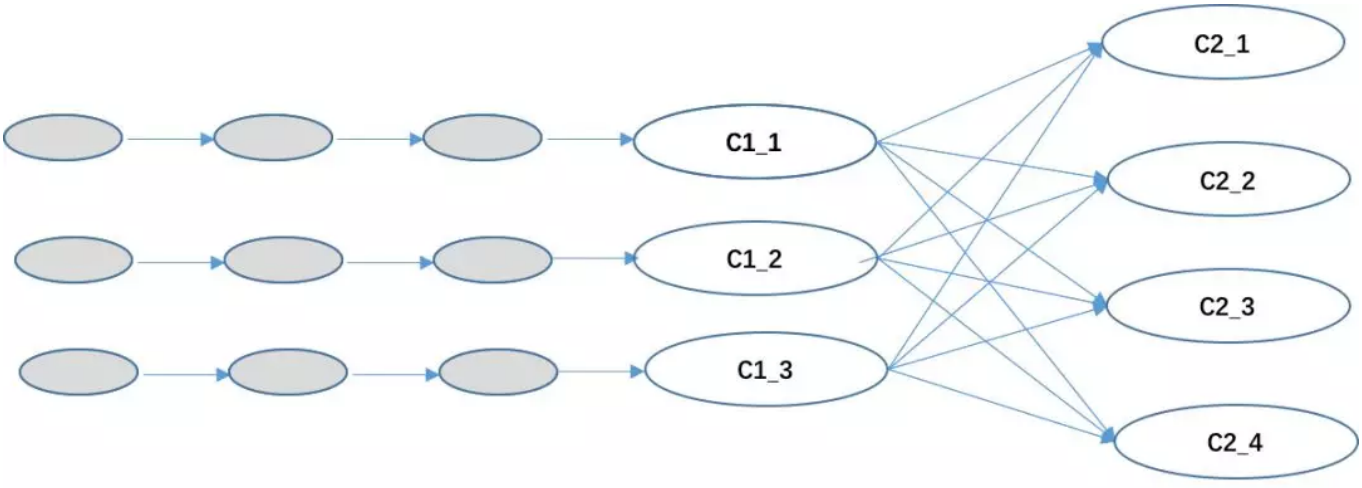
其中P（找聘|源通,快滴）类似的是“找聘”在“源通快滴”后的条件概率。



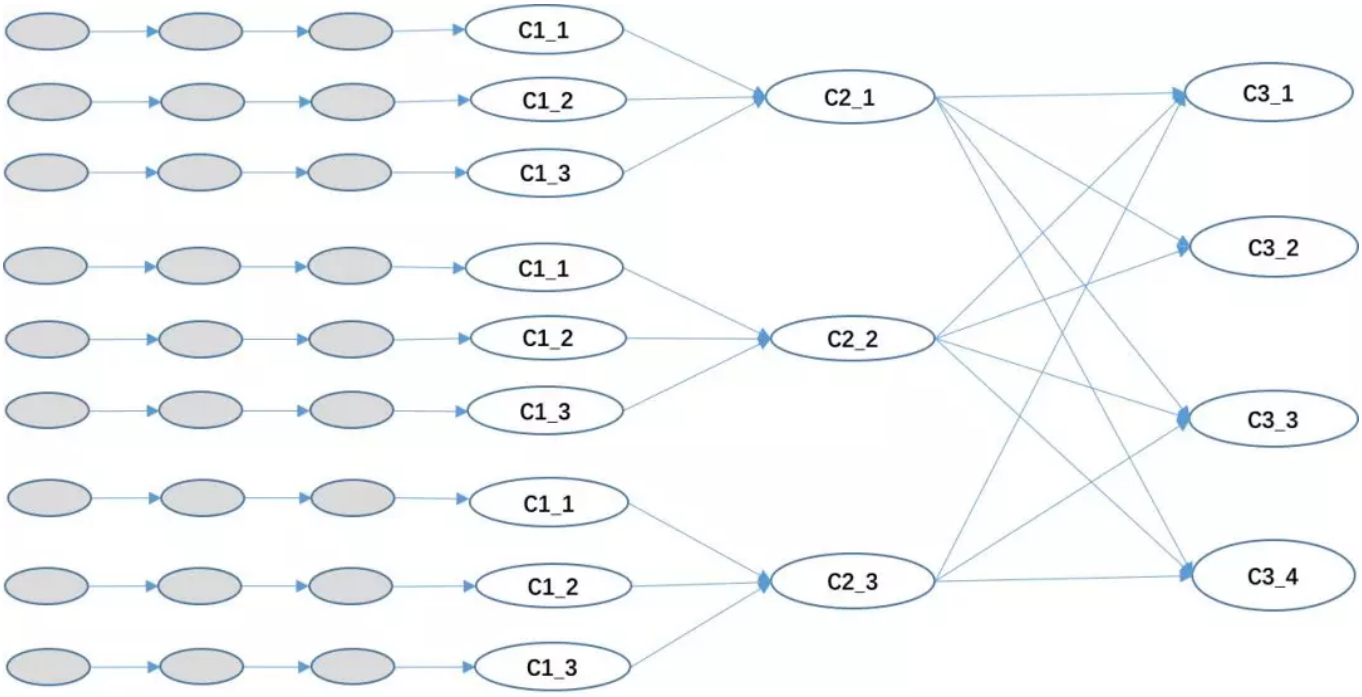
以上候选词组成一个栅格图，可以使用维特比算法求出这个栅格图的最大概率路径。

维特比算法使用动态规划的思想，每个分词位置的最大概率路径必定由前一个位置的某个候选词的最大概率路径组成。

在本纠错方法中，由于使用3-gram，对维特比算法进行一些修改，如下图所示：



2-gram



3-gram

在图中，C_x_x这样的标记是候选词。在2-gram中，因为每个候选词的条件概率只与前面一个位置的候选词计算出来，所以使用前面一个位置每个候选词的最大概率路径来计算当前位置每个候选词的最大概率路径。如下式所示：

$$PSUM_{\max}(C2\_1) = MAX(PSUM_{\max}(C1\_1) + P(C2\_1 | C1\_1),$$
$$PSUM_{\max}(C1\_2) + P(C2\_1 | C1\_2),$$
$$PSUM_{\max}(C1\_3) + P(C2\_1 | C1\_3))$$

在3-gram中，因为每个候选词的条件概率是由前面两个位置的候选词计算出来的，所以每个候选词对应的最大概率路径不一定由前面一个位置的每个候选词的最大概率路径组成。前一个位置的每个候选词需要保存多条路径及对应的最大概率值，每条路径是经过前两个位置候选词的路径中的最大概率路径。如下式所示：

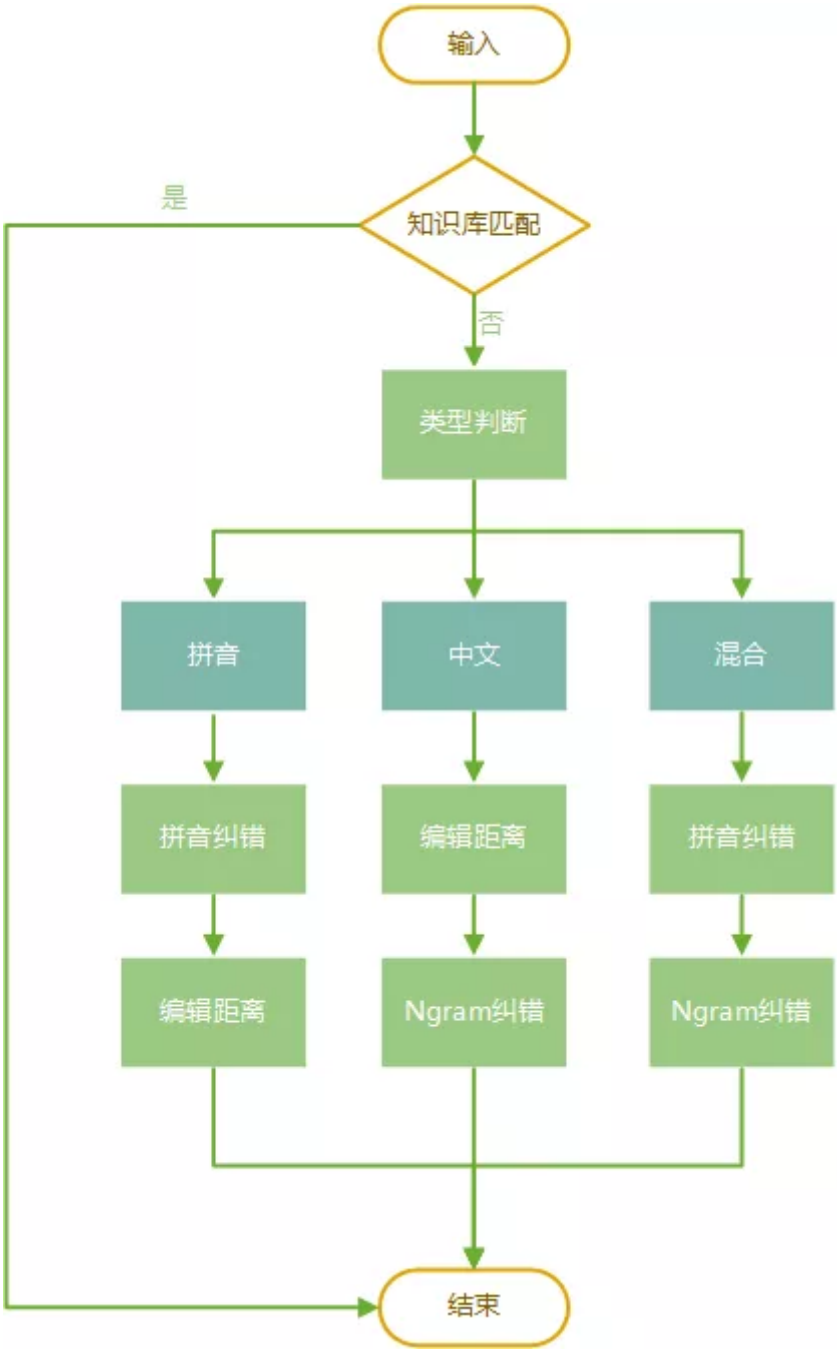
$$PSUM_{\max}(C3\_1,C2\_1)=MAX(PSUM_{\max}(C2\_1,C1\_1)+P(C3\_1|C2\_1,C1\_1),$$
$$PSUM_{\max}(C2\_1,C1\_2)+P(C3\_1|C2\_1,C1\_2),$$
$$PSUM_{\max}(C2\_1,C1\_3)+P(C3\_1|C2\_1,C1\_3))$$

$$PSUM_{\max}(C3\_1,C2\_2)=MAX(PSUM_{\max}(C2\_2,C1\_1)+P(C3\_1|C2\_2,C1\_1),$$
$$PSUM_{\max}(C2\_2,C1\_2)+P(C3\_1|C2\_2,C1\_2),$$
$$PSUM_{\max}(C2\_2,C1\_3)+P(C3\_1|C2\_2,C1\_3))$$
$$PSUM_{\max}(C3\_1,C2\_3)=MAX(PSUM_{\max}(C2\_3,C1\_1)+P(C3\_1|C2\_3,C1\_1),$$
$$PSUM_{\max}(C2\_3,C1\_2)+P(C3\_1|C2\_3,C1\_2),$$
$$PSUM_{\max}(C2\_3,C1\_3)+P(C3\_1|C2\_3,C1\_3))$$

最后一个位置时求出所有候选词的最大概率路径的组合就是最后的纠错结果。
具体的工程实现中，对相邻词组合纠错后得到的纠错结果进行分割来得到每个分词位置的候选词。分割过程中需要注意缺字情况下纠错结果的长度和原始组合词长度不一样情况下分割情况的完整性，将缺字的所有位置都要进行考虑。查询词的分词结果中可能有一些单字，因为单字与相邻词组合后的词大部分比较短，比较短的词不进行拼音纠错和编辑距离纠错。所以单字不作为单独的分词位置，而是将单字组合到周围的分词中。组合时要考虑所有的组合情况。纠错中间结果的分割也需要考虑单字的影响。

在线纠错流程

整个线上纠错流程如下图所示：



输入查询词后

1. 首先解析查询词，判断查询词是全拼音或还是包含中文。
2. 如果查询词是全拼音，则进行拼音纠错。
3. 如果被纠错则退出纠错流程，返回纠错结果。
4. 如果没有被拼音纠错，则进行编辑距离纠错。
5. 如果被编辑距离纠错则退出纠错流程，返回纠错结果。
6. 如果没有被编辑距离纠错，则进行长拼音纠错，处理完后退出纠错流程，如果有纠错结果，则将纠错结果返回。
7. 如果查询词是部分中文或全中文的，则先进行拼音纠错。具体方法为将所有汉字转换为对应的拼音，然后查询拼音索引，查询有结果则纠错。如果没有被拼音纠错，则对查询词进行分词，然后使用基于语言模型的方法纠错。

各种错误query的纠错流程示例：

错误词	纠错后的词	纠错方法	纠错级别
Shuianhuating	水岸华庭	不包含汉字，拼音纠错后退出	强制纠错不提示
ATLS	奥特莱斯	不包含汉字，拼音纠错后退出	强制纠错提示
linshiG	时工	不包含汉字，拼音纠错后退出	强制纠错提示
iphoni4	iphone4	不包含汉字，拼音纠错未能纠错，编辑距离纠错正确纠错，退出	强制纠错提示
05crv	05款crv	不包含汉字，拼音纠错未能纠错，编辑距离纠错正确纠错，退出	强制纠错提示
shijingshanxiaochaoshi	石 景 山 小 超 市	不包含汉字，拼音纠错未能纠错，编辑距离纠错未能纠错，长拼音纠错正确纠错，退出	强制纠错提示
途an	途安	包含汉字，拼音纠错正确纠错，退出	强制纠错提示
保山l	宝山路	包含汉字，拼音纠错正确纠错，退出	强制纠错提示
昂克威	昂科威	全部汉字，拼音纠错正确纠错，退出	强制纠错提示
天津医科大学总医院zufang	天 津 医 科 大 学 总 医 院 租 房	包含汉字，整体拼音纠错未纠错，分词为天津医科大学、总医院、zufang。使用ngram纠错为天津医科大学总医院租房	强制纠错不提示
复试办公公寓	复 式 办 公 公 寓	包含汉字，整体拼音纠错未纠错，分词为复试、办公、公寓。使用ngram纠错为复式办公公寓	建议纠错及提示

总结

本纠错模块就规则、统计模型的传统纠错方法实践进行了介绍，在58场景下可以对大部分的用户错误输入进行纠错，但就技术主要步骤:纠错词表、候选词生成及评价等方面还有不少空间，针对不同领域和业务情况还可以实施独立的策略算法，进行更精细的优化。

目前业界前沿纠错技术不仅可以纠正用词的错误，还可以纠正文法错误，句法错误，知识错误等。例如百度的纠错技术，首先使用CRF技术检测错误，然后召回候选，最后使用deep和wide的混合模型对候选词排序。腾讯的纠错技术基于语义关联，特别对垂直搜索的效果很好，具体的实现和知识图谱的挖掘类似。另外有些开源项目大部分使用了避免人工特征提取的基于深度模型的纠错方法，各有优缺点。需要我们追赶前沿，根据业务场景不断的进行探索实践优化。