中文错别字检索方案

0.背景

随着统计学习、机器学习、深度学习的发展，中文错别字检索的效率有了极大的提升。本文简单提出一个中文错别字检测的方案。

1. 初步设想

方案的实施顺序：

1）数据收集：正确的语料库，修正字典（训练过程中会继续添加），困惑集;

2）选择语言模型;

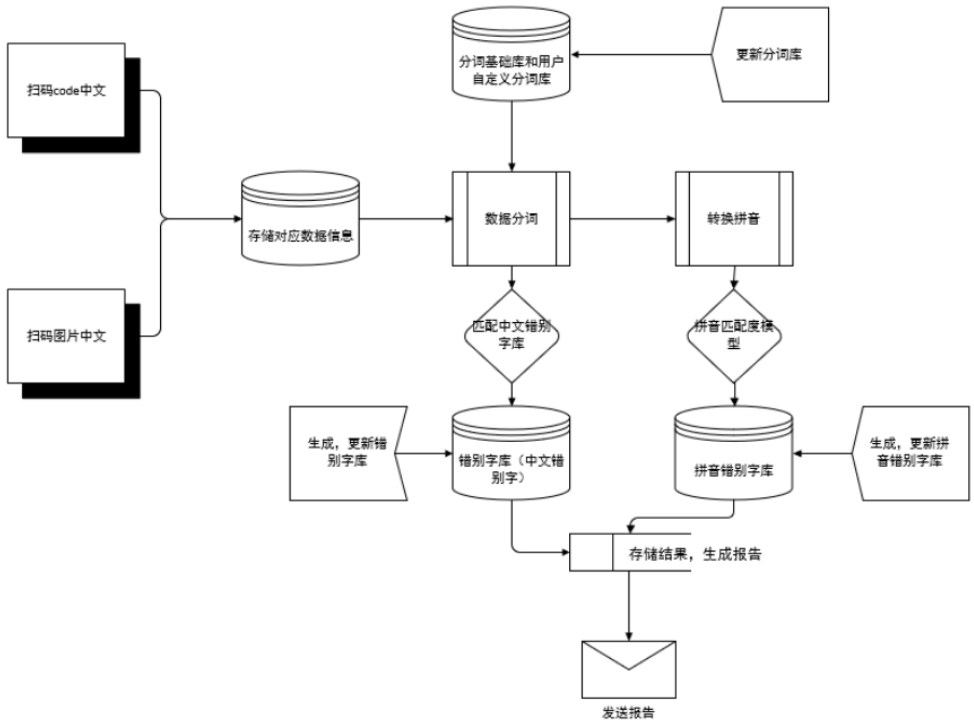
3）使用语言模型在正确的语料库上进行训练，构建高可信度模板;

4）选择分词器对需要检测的文章进行分词;

5）对切词结果进行检测，使用可信度模板过滤一部分正确的序列，对剩余部分进行检测，从字粒度和词粒度两个层次进行检测，整合疑似错误的结果形成错误位置候选集;

6）遍历所有疑似错误位置，并使用音似、形似词典替换错误位置的词，使用语言模型计算句子困惑度，对所有候选集结果比较并排序，得到最优纠正词;

7）将（错误的字词：替换词）加入修正字典和困惑集，提高后续检测效率。



方案流程图

2. 数据收集

正确语料库：1）北京大学语料库（http://icl.pku.edu.cn/)
2）台湾中研院平衡语料
（http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/)
3) Chinese LDC
…

分词器选择：1）哈工大 LTP
2）中科院计算所 NLPIR
3）清华大学 THULAC
4）“结巴”中文分词

修正字典：　　暂无

待检测数据：　　互联网上爬取的文字数据、图片中的文字信息

3. 语言模型选择

语言模型是一种概率模型，它是基于一个语料库创建，得到每个句子出现的概率。语言模型主要分为两类：以 N—gram 为代表的统计语言模型，RNN/LSTM 神经网络语言模型。

统计语言模型通过条件概率公式展开得到，统计待查询单词前所有单词连续出现的情况下，判断语句是否正常。随着技术的发展，N—gram 取代了对所有单词进行概率统计的语言模型，通过统计语料库中词串出现的次数，一次性计算得到词串的概率并将其保存起来，在预测一个句子时，认为一个词出现的概率只与它前面 n—1 个词有关，据此计算得到句子出现的概率，极大减小了统计的可能性、提高了计算效率。

近年也流行起神经网络语言模型，从机器学习的角度来看，一开始不全部计算这些词串的概率值，而是通过一个模型对词串的概率进行建模，然后构造一个目标函数，不断优化这个目标，得到一组优化的参数，当需要哪个词串概率时，利用这组优化的参数直接计算得到对应词串概率。在神经学习领域，是将一个样本对象抽象为一个向量，神经网络语言模型中是将词或短语表示为向量，通常叫做 word2vec。

主要的语言模型：

kenlm: kenlm 统计语言模型工具

rnn_lm: TensorFlow、PaddlePaddle 均有实现栈式双向 LSTM 的语言模型

rnn_attention 模型: 参考 Stanford University 的 nlc 模型，该模型是参加 2014 英文文本纠错比赛并取得第一名的方法

rnn_crf 模型: 参考阿里巴巴 2016 参赛中文语法纠错比赛并取得第一名的方法

seq2seq 模型: 使用序列模型解决文本纠错任务，文本语法纠错任务中常用模型之一

seq2seq_attention 模型: 在 seq2seq 模型加上 attention 机制，对于长文本效果更好，模型更容易收敛，但容易过拟合

kenlm 的特点：

快速，节省内存，最重要的是，允许在开源许可下使用多核处理器；比 SRILM 和 IRSTLM 更快，更低的内存；使用用户指定的 RAM 进行磁盘估计；用于时空权衡的两种数据结构；mmap 的二进制格式。或直接加载 ARPA 文件；如果安装了相应的库，它还可以读取使用 gzip，bzip2 或 xz 压缩的文本和 ARPA 文件，线程安全的；更多假设重组的机会，如果模型退避，则 State 仅存储匹配的单词；FullScore 函数还返回模型匹配的 n-gram 长度。

完全基于 attention 的 seq2seq 的特点：

传统的基于 CNN 或 RNN 的 Seq2Seq 模型都存在一定的不足：CNN 不能直接用于处理变长的序列样本，RNN 不能并行计算，效率低。虽然完全基于 CNN 的 Seq2Seq 模型可以并行实现，但非常占内存，很多的 trick，大数据量上参数调整并不容易。传统的 Seq2Seq 模型不适用于长的句子，Seq2Seq+attention 虽然提升了处理长句子的能力，但 encoder 解码得到隐变量 Z 时，任然对观测序列 X 的计算添加了约束。

基于 Attention Mechanism + LSTM 的 Seq2Seq 模型的优点：自适应地计算一个权值矩阵 W，权重矩阵 W 长度与 X 的的词数目一致，每个权重衡量输入序列 X 中每个词对输入序列 Y 的重要程度，不需要考虑输入序列 X 与输出序列 Y 中，词与词之间的距离关系。缺点：attention mechanism 通常是和 RNN 结合使用，但 RNN 依赖 t-1 的历史信息来计算 t 时刻的信息，因此不能并行实现，计算效率比较低，特别是训练样本量非常大的时候。

基于 CNN 的 Seq2Seq+attention 的优点： 基于 CNN 的 Seq2Seq 模型具有基于 RNN 的 Seq2Seq 模型捕捉 long distance dependency 的能力，此外，最大的优点是可以并行化实现，效率比基于 RNN 的 Seq2Seq 模型高。缺点：计算量与观测序列 X 和输出序列 Y 的长度成正比。

完全基于 Attention 的 Seq2Seq 模型，在基于 CNN 和 RNN 的 Seq2Seq 模型存在的不足之处表现良好。基于 Attention 来构造 encoder 和 decoder。

泛化能力强，也因此容易导致过拟合，从而带来错别字误判。

4.具体实施

1）收集数据

使用采自人民日报的北京大学语料库作为语言模型训练的标准语料库,在开放社交平台上收集大量文本内容，作为语言模型初期检测及训练的样本，在此检测过程中，增加修正词典以及困惑集的深度，从而提高后期错别字的检测速度。

2）选择语言模型

使用快速的统计语言模型 **kenlm 结合短距离**的上下文进行检查词粒度、字粒度层面是否存在错别字，对句子里的字符进行打分，得分低的地方视为待纠错位置。将待纠错位置与上下文组合进行字典查词，当所有组合在词典中都查找不到，则将其视为错字。

使用神经网络中的 **seq2seq_attention 语言模型进行长距离**的上下文错别字检测，可以充分利用神经网络在检测长距离词串上的优势.

3）训练及构建高可信度模板

使用选定的两种语言模型在标准语料库上进行训练，构建高可信度模板，将出现频率较高的词串加入长词词表。同时在含有错词的文本中进行训练，构建困惑集，并将高频次的错误词对加入常用错误词表，检测到该错误出现时可以直接替换。需要长期维护模板、长词词表、常用错误词表。

4）待测中文文本分词

使用 LTP 分词工具对待检测中文文本进行分词。（也可通过神经网络构建自己的分词工具）。长期使用，可添加用户自定义的个性分词库，以提高某专业领域内的检测效率。

5）对分词结果进行检测。

通过高可信度模板进行初步筛选，对剩下的文本在两个方面进行检测：字粒度方面，使用语言模型、困惑度检测某字的似然概率值低于句子文本平均值，则判定该字是疑似错别字的概率大;词粒度方面，切词后不在词典中的词是疑似错词的概率大。

主要的检测方式有两种：在之前构建的困惑集、词表中进行查找，找到后使用语言模型进行纠正评价决定是否更改; 使用语言模型生成纠错候选，利用本身的评价函数，在生成纠错候选过程中进行预筛选。

6）评价纠正结果

遍历所有所有疑似错误的位置，使用语言模型对得到的纠错候选进行评分排序，最终排序最高（如没有错误识别阶段，则仍需比原句评分更高或评分比值高过阈值，否则认为不需要纠错）的纠正候选作为最终纠错结果。

7）扩充困惑集、维护词表

困惑集，是中文文本纠错任务中较为关键的数据之一，用于存储每个字词可能被混淆的错别字词的可能。困惑集的数据格式是 key-value 格式，key 为中文中常用的字词，value 为该字词可能的错误方式。key 可以基于字符，也可以包含词语。通常一个 key 对应多个 value。

错误形式，主要分为两大类：发音混淆，形状混淆。形状混淆，通常是五笔输入笔画输入手写收入带来的错误。发音混淆最为常见，可分为相同读音、相同音节不同音调、相似音节相同音调、相似音节不同音调。
对困惑集进行扩充，并对每一个拼写错误构建倒排索引，拼写错误为索引词，潜在正确结果为检索内容，对于每个潜在正确内容，利用汉字出现的频率进行排名。预测同时，在监测阶段维护一个错词修正表，每次替换之后不在词表的词均加入错词表，最终找到正确结果的词加入正确词表，每次结束之后构建错词修正表。如果下次预测到的时候直接利用错词修正表进行调整。

5.测评方法

评价指标：
该纠正的，即有错文本记为 P，不该纠正的，即无错文本记为 N
对于该纠正的，纠对了，记为 TP，纠错了或未纠正，记为 FP
对于不该纠正的，未纠正，记为 TN，纠正了，记为 FN。
通常情况下，查准比查全更重要，FN 更难接受，可构造下述评价指标
$\frac{1}{F_\beta} = \frac{2}{P} + \frac{1}{R}$，其中$P = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$,$\mathrm{R} = \frac{TP}{TP+2FN}$.
测评过程：
选择含有主要的中文错别字类型进行错别字检测，观察模型的查准率和查全率。

6.优化过程

影响纠错效果的主要因素有如下几点：
困惑集：主要影响召回率，纠错首先需要的就是构建一个好的困惑集，使其尽可能小但是包括绝大多数情况。
语言模型：在纠错任务中，常常使用两种语言模型，一种是基于字符级别的，主要用于错 n-gram 语言模型误的发现，一般字符级别的阶数在 1 到 5 之间。还有一种是词级别的，主要用于排序阶段。
词表：词表主要用于判断字符之后是否可以成词，词表最好是比较大的常用词表加上需要应用的领域词表。
语料：可以利用大规模的互联网语料估计错误拼写，而且语料也应用于语言模型的生成。
后续优化：
添加不同专业领域内的正确语料，丰富语料库；
针对不同领域，完善分词库。