R&S[23] | 搜索系统中的纠错问题

原创 机智的叉烧 CS的陋室 2020-03-08

国内疫情基本稳定下来，我们看到曙光了，大家继续坚持，一切都会好起来。

Children Of A Miracle

Don Diablo;Marnik - Shut Up Lets Dance(Vol. II)

往期回顾：

- NLP.TM[28] | 浅谈NLP算法工程师的核心竞争力
- ML&DEV[13] | bad case分析
- R&S[22] | 搜索系统中的召回
- R&S[18] | SIGIR2018：深度学习匹配在搜索与推荐中的应用
- R&S[17] | 手把手搞推荐[6]：回顾整体建模过程

纠错是搜索引擎中一个非常有特色的模块，对用户输入的内容进行改写从而让用户得到正确的结果，有的时候也会带有一些惊喜度，所以纠错技术是一个搜索体验的加分项，近期突然对这块有兴趣，所以就了解了一下。**（学习周报本周停，学习内容都在这了）**

纠错技术的背景

人非圣贤，孰能无过，别说是搜索的时候，哪怕是我们打字、写作文的时候，都会出现错字，一般的错别字不会对最终目标带来很大影响，且出现频率很低，不拘小节的我们常常会忽略这样的小问题，但是，在搜索场景下，错别字意味着可能就搜不到内容了，对于用户而言，就是需求无法满足，造成了很差的体验，因此在搜索场景中，就很有必要去纠错。

错误是如何产生的

要去纠错，先要去看看错误是怎么产生的。

首先是误操作类型，这种类型可以从输入法角度去看。

- 拼音输入法。常会出现同音异形字，例如周节伦等。
- 笔画输入法或者手写输入法。常会出现形似字，例如博和傅。

然后是用户的主观理解，有的时候用户只是听说过而没见过，或者就是理解问题，导致主动地出入了错误的内容，例如飞扬拔（跋）扈，然后有一些名词，例如小说、音乐、电影等，写错字是非常容易的。

当然，也有用户图方便，或者输入问题，导致直接输入拼音或者拼音前缀，或者就是因为记忆的原因，输错了。

当然这里也要补充一些常见的问题举例：

- 谐音。深圳-森圳。
- 别字。师傅-师博。
- 中英文。Taylor swift-泰勒斯威夫特。
- 近义词。爱情呼叫转移-恋爱呼叫转移。
- 形近字。高粱-高梁。
- 全拼。深圳-shenzhen。
- 拼音前缀。北京-bj。
- 内容不完整。唐人街探案-唐人。

总之错误千奇百怪。理解错误产生的机理，我们就可以尝试去处理这些问题。

词典与规则方法

词典是搜索系统中非常常用的方法，词典具有高速、高准的优点，如果词典的覆盖度高，甚至可以达到高召回的效果，因此词典基本是搜索系统中的核心存在，我们不应该小看他，而是尽可能挖掘他的潜能。

词典方法，说白了就是对query找对应词典里有没有，如果有就改写过去，这种方法的优点在于速度快，而难点在于怎么去挖掘这个词典。

至于怎么挖掘这个词典，方法有很多底层数据库抽取，用户日志等，都有很多构建起这样的词典，能够大大降低耗时，复杂度至于query和单词长度有关。那么一般都有什么词典呢，我们来一个一个看看。

- 拼音和拼音前缀词典。先将query或者单词转为拼音，然后通过通过拼音召回对应的结果，完成纠错。
- 别字词典，记录一些常见的错别字，例如百度的形近词表就很不错（就在百度百科里面）。
- 其他改写字典。一般基于具体业务来改写，例如用户输入唐人街探案，其实唐人街探案有3部，我们应该给那个，需要基于热度等方面去改写到具体最合适的一部。

词典只是能够匹配到合适的结果，但是我们需要知道的是，改写的内容不能和原来差距太远，否则会出现很多意料之外的结果，因此改写不能大改，只能改微调，否则出来的结果会让用户感到很懵逼。控制的方法主要是**编辑距离**。

所谓的编辑距离，就是改写前到改写后，需要经过的操作多少，说人话就是两句话的不同点有几个，精确到字级别。深圳-森圳的编辑距离就是1。通过编辑距离的约束，一般能够让两者的差距不是很大。

我知道很多人热衷于用语义相似度之类的操作，不管别的什么方法，编辑距离一定要约束，用户强调的是直观感受，语义相近与否不是他们第一个关心的，只有当字相近的结果不好的时候考虑语义相近才是用户的实际反映，且错别字带来的语义变化非常大，此处用予以相似度其实不完全合适。

模型类方法

说是词典和规则好处很多，但是在泛化能力上，模型还是很强的。那么在模型视角下，其实会分为下面3个步骤进行分析处理。

- 错误诊断。即判断有没有错。
- 修正召回。召回可能的修改项。保证召回率
- 修正确认。判断最终需要的修改项。保证准确率。

当然，如果模型足够强力，召回和确认两个步骤也可以合并，具体看准招和耗时了。

其实这个思路最广泛的应用就是推荐系统，召回和排序分离，这个我在大概是去年很早的一篇文章里谈到在这个，这是推荐系统里面非常重要的思想，这个思想其实在很多地方可以迁移：

技术向：推荐学习推荐系统（深度思考，不是广告）

至于模型层面，有下面的思路。

- kenlm统计语言工具。运用统计学方法进行语言建模从而检测和修正错误。
- rnn_attention。RNN加上attention还是一个非常有意思的方法。
- rnn_crf模型：说起来你们可能不信，这个思路来自阿里2016参赛中文语法纠错比赛的第一名的方法。
- seq2seq_attention模型：比RNN强一些，长文本效果不错，但是容易过拟合。
- transformer：线性优秀的序列表征模型，大家懂的。
- bert：中文微调，最妙的是mask可协助纠正错别字。
- conv_seq2seq模型：基于Facebook出品的fairseq，在NLPCC-2018的中文语法纠错比赛中，是唯一使用单模型并取得第三名的成绩。

小结

怎么说呢，目前我还只是在探索，深度不是很够，后面有所补充，再和大家交流，参考文献放这里吧：

- 中文文本纠错算法--错别字纠正的二三事：https://zhuanlan.zhihu.com/p/40806718
- pycorrector：https://github.com/shibing624/pycorrector
- 中文文本纠错算法走到多远了？：https://blog.csdn.net/sinat_26917383/article/details/86737361