

文章编号: 1004-4353(2020)03-0260-05

基于字典树语言模型的专业课 查询文本校对方法

李丹阳, 赵亚慧*, 罗梦江, 崔荣一

(延边大学 工学院, 吉林 延吉 133002)

摘要: 针对中文文本校对技术中存在的校对准确率较低的问题,提出了一种基于字典树模型的专业课查询文本校对方法.首先,通过计算错误文本与匹配文本间的编辑距离对错误关键词进行模糊匹配;其次,采用字典树语言模型建立搜索树,以提高查询效率.最后,通过对比不同文本相似度阈值下的校对效果选取最佳文本相似度阈值.在最佳阈值下(0.5),将本文模型与传统的拼音模型和 N-gram 模型进行问句校对对比显示,本文方法的准确率(77.91%)、召回率(67%)、F 值(72.04%)比传统的拼音模型校正方法分别提高了 5.69%、23.67% 和 11.57%,比 N-gram 模型校正方法分别提高了 0.64%、10.33% 和 7.89%.因此,本文提出的方法在专业课查询文本校对方面具有很好的应用价值.

关键词: 字典树; 文本校对; 语言模型; 自动纠正

中图分类号: TP391.41

文献标识码: A

Query text proofreading method of professional courses based on trie tree language model

LI Danyang, ZHAO Yahui*, LUO Mengjiang, CUI Rongyi

(College of Engineering, Yanbian University, Yanji 133002, China)

Abstract: Aiming at the problem of low accuracy in Chinese text proofreading technology, a method of text query and proofreading is proposed for professional courses based on trie tree model. Firstly, the error keywords were fuzzy matched by calculating the edit distance between the error text and the matching text. Then, the trie tree language model was used to build the search tree to improve query efficiency. Finally, by comparing the proofreading effect under different text similarity thresholds, the best text similarity threshold was selected. Under the best threshold (0.5), the model was compared with the traditional Pinyin model and N-gram model in question proofreading. The accuracy rate (77.91%), recall rate (67%) and *F* value (72.04%) of the proposed method are 5.69%, 23.67% and 11.57% higher than those of the traditional Pinyin model correction method, and 0.64%, 10.33% and 7.89% higher than that of the N-gram model correction method. Therefore, the method proposed in this paper has good application value in the text query and proofreading of professional courses.

Keywords: trie tree; text proofreading; language model; automatic correction

随着网络信息化的迅速发展,如何在海量的
网络信息中准确、快速地搜索出所需的信息日益

受到人们的关注.但在网络信息搜索时,常会出现
因错误问句而造成的“答非所问”的现象,为此学

收稿日期: 2020-06-22

* 通信作者: 赵亚慧(1974—),女,教授,研究方向为自然语言处理.

基金项目: 国家语委“十三五”科研规划项目(YB135-76);延边大学外国语言文学世界一流学科建设科研项目(18YLPY13,18YLPY14)

者提出了文本校对技术^[1].传统的中文文本校对是借鉴英文文本校对的方法^[2]使用拼音纠错模型进行文本校对^[3],即将错误文本转化为拼音,然后在定义好的词典上进行纠错^[4].传统的中文文本核对方法仅适用于同音字的校正,而且在文字转化成拼音的过程中因增加了字符串的长度使得计算量增大.近年来,一些学者对中文文本校对技术进行了改进,如:文献[5]通过构建语法-词语搭配双层知识库,提出了一种基于互信息和聚合度双重评价条件下的词语搭配校对算法;文献[6]提出了一种将 LSTM 和集成算法结合的中文文本校对算法;文献[7]提出了一种将 N 元语言模型与字典分词相结合的中文文本校对算法.但上述方法仍存在召回率、准确率、纠正率偏低等问题^[8-9].为此,本文利用编辑距离与字典树结合的方法,对计算机专业领域的文本校对问题进行研究,并通过实验验证了本文方法的有效性.

1 专业课查询文本的校对设计

专业课查询其本质是对问句中包含专业课关键词的查询,因此在文本校对前需要对问句进行文本预处理.

1.1 文本预处理

文本预处理主要包含分词和去停用词.经过预处理的文本可以保留对查询有用的词语,剔除对查询无用或干扰查询的词语,进而提高检索的效率和准确率.

1.1.1 分词 词是最小有意义的语言成分.中文分词是将中文语句序列以词为单位切分成词序列,其结果直接影响后续的文本处理结果.经典的分词系统有中科院的 NLPIR 和哈工大的 LTP 语言云等,因哈工大的 LTP 语言云在各测试数据集上有较高的分词准确率,因此本文采用哈工大的 LTP 语言云的 Java 接口进行实验.

1.1.2 去停用词 停用词是一些完全不起作用或者没有意义的词,如助词“之”“而已”、语气词“啊”“呀”,副词“很”“一般”、介词“在”“对于”等.这些词不仅会降低查询效率,还会影响搜索结果的准确性;因此,本文在处理搜索请求时将这些词过滤,仅对问句中的课程专业词进行文本校对和搜索.目前,常用的停用词表有哈工大停用词表、

百度停用词表、四川大学机器智能实验室停用词库等.因哈工大停用词表的覆盖内容较为全面(包括标点符号、数字、英文和各种无具体意义的词等),因此本文采用哈工大停用词表.

1.2 文本纠错

中文纠错包括问句查错和错误字符串纠正两个模块.问句查错是以字典树语言模型为基础,采用相关查错方法识别出错误问题,如一个含有错别字的问句经过分词后会出现字符串个数异常增加的现象.错误字符串纠正是根据字符串间的编辑距离进行模糊匹配和相似度计算,然后根据计算结果在词典中查找出几个相近的字符串,并在其中挑选一个最为合适的作为最终纠错建议.

1.3 文本相似度计算

发现错误词语后,首先比较错误文本与字典中其他文本的最小编辑距离^[10],然后再根据最小编辑距离计算出不同词语间的文本相似度,最后通过合适的文本相似度值从字典中挑选出用于纠错的正确词语.

编辑距离是指将中文词串 X 转换为另一个词串 Y 而添加、删除、易位或替换的汉字个数^[10],记为 $dist(X, Y)$.字符串 X 的前 i 个字符转换成字符串 Y 的前 j 个字符所需的最少操作次数即为最小编辑距离,记为 $dist[i][j]$. $dist[i][j]$ 的计算公式为:

$$\begin{aligned} dist[i][j] &= \min\{dist[i-1][j-1], \\ &\quad dist[i-1][j], dist[i][j-1]\}, \\ X[i] &\neq Y[j]; \\ dist[i][j] &= dist[i-1][j-1], \\ X[i] &= Y[j]; \\ dist[i][j] &= dist[i][0] = i, \\ dist[0][j] &= j. \end{aligned} \quad (1)$$

计算出最小编辑距离后即可判断文本是否相似.设 X 和 Y 的较长字符串长度为 L_{\max} ,编辑距离为 LD ,则文本相似度(S)可表示为

$$S = 1 - LD \cdot L_{\max}^{-1}. \quad (2)$$

由式(2)可以看出,字串之间的编辑距离越小,文本相似度越大,即成功纠错词的可能性越大.

2 字典树语言模型

本文采用字典树存储词典,通过字典树查找

关键词并结合中文文本分词技术和相似度计算实现问句文本校对。

2.1 字典树的数据结构

字典树^[11]又称为单词查找树、Trie 树和前缀树,是一种类似于哈希树的变种搜索树,其结构如图 1 所示。字典树具有以下特点:①树的根节点不存储任何数据,其他节点只存储一个字符。②每次检索从根节点检索至叶子节点,将检索路径上的所有字符连接起来即构成一个词。因此,字典树可以共享公共前缀,节省存储空间。当搜索某个字符串时,直接查找以该字符串的首字符开始的链路即可,由此可最大限度地减少字符串的比较,提高检索效率。

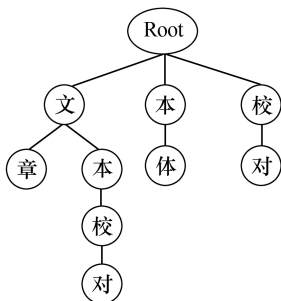


图 1 字典树结构示意图

2.2 字典树的基本操作

1)向字典树中添加字符串,其具体步骤如下:

Step1 输入待插入字符串 key 和字典树根节点 rootNode,将字符检索位置 charIndex 初始化为 0;如果 rootNode 为空,则新建一个字符节点。

Step2 令当前节点 currentNode 从根节点 rootNode 开始。

Step3 记当前树节点与待插入 key 在 char-Index 位置的差值为 compare。

Step4 ① $compare = 0$ 时, $charIndex++$ 。如果已经全部匹配完毕,则添加新字符信息,转至 Step5;如果 currentNode 的下节点为空,则新建字符节点后继续向下遍历,转至 Step3。

② $compare < 0$ 时,如果 currentNode 的左节点为空,则新建字符节点后继续向左遍历,转至 Step3。

③ $compare > 0$ 时,如果 currentNode 的右节点为空,则新建字符节点后继续向右遍历,转至 Step3。

Step5 结束。

2)在字典树中查找字符串,其具体步骤如下:

Step1 输入待查询字符串 key,开始查询位置 offset,将字符检索位置 charIndex 初始化为 offset,字符串 word 初始化为空。

Step2 如果根节点 rootNode 为空或 key 为空,转至 Step6。

Step3 令当前节点 currentNode 从根节点 rootNode 开始;如果 currentNode 为空,则转至 Step6。

Step4 记当前树节点与待查询 key 在 char-Index 位置的差值为 compare。

Step5 ① $compare = 0$ 时, $charIndex++$ 。如果还未匹配完,则候选最长匹配词;如果 char-Index 等于 key 的长度,则转至 Step6;否则继续向下遍历,转至 Step4。

② $compare < 0$ 时,继续向左遍历,转至 Step4。

③ $compare > 0$ 时,继续向右遍历,转至 Step4。

Step6 结束,返回字符串 word。

2.3 基于字典树的专业课问答

本文使用计算机专业词典作为校对标准,采用字典树结构进行存储。词典共包含 10 299 条专业词汇。当用户输入一个课程问句时,若问句中含有错误,则利用词典对问句的错误词进行文本校对,并给出纠正建议。文本校对处理的流程如图 2 所示。

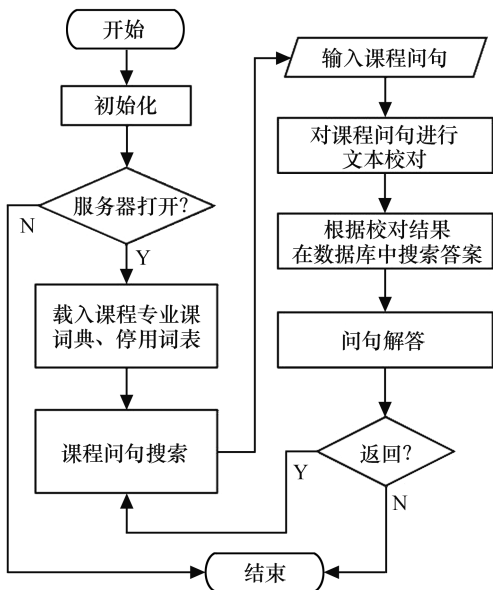


图 2 文本校对处理流程图

3 测试结果与分析

3.1 评测指标

本文采用查准率(P)、查全率(R)和综合评价指标值(F)来度量查错模块,采用纠错准确率(AR)和纠正率(CR)来度量纠错模块。 P 、 R 和 F 的计算公式如下:

$$P = \frac{N}{M} \times 100\%, \quad (3)$$

$$R = \frac{N}{T} \times 100\%, \quad (4)$$

$$F = \frac{2 \times R \times P}{R + P}. \quad (5)$$

其中, N 为正确发现问句的总数, M 为发现问句的总数, T 为文本中错误问句总数。这里规定,只要对错误文本进行改动即认为发现了错误问句,对专业词汇进行修改即认为正确发现了错误词的位置。 AR 和 CR 的计算公式如下:

$$AR = \frac{Q}{N} \times 100\%, \quad (6)$$

$$CR = \frac{Q}{T} \times 100\%. \quad (7)$$

其中, Q 为正确纠错的总数。若修改后的专业词汇与正确文本保持一致即认为纠正正确。

3.2 阈值的选定

在课程领域的专业词典中有许多相近词,能否挑选出最合适的纠错词取决于设定的相似度阈值(f)。由文本相似度定义可知:相似度值越大,纠错结果与输入的课程问句越接近,即越难以实现文本校对的目的;相似度值越小,在词典中检索到的可能的纠错词数目就越多,即纠错效率越低。本文使用0.4、0.5和0.6作为阈值进行测试,并与拼音模型校正方法和N-gram模型校正方法进行对比。实验选取100条问句文本作为输入数据,结果如表1和图3所示。

由图3可看出,本文方法在阈值0.5时,其各项指标均高于传统拼音模型方法、N-gram模型方法和阈值为0.6时的本文方法。本文方法的阈值为0.4和0.5时,其 AR 值和 CR 值虽各有优势,但考虑到阈值为0.5时的 P 、 R 和 F 值均高于0.4时的 P 、 R 和 F 值,因此本文的文本相似度阈值取为0.5。

表1 不同方法纠错率的结果

实验方法	P	R	F	AR	CR
Pinyin model	72.22	43.33	60.47	76.92	33.34
N-gram model	77.27	56.67	64.15	76.47	43.33
本文方法 ($f=0.4$ 时)	62.50	55	58.51	90.91	50
本文方法 ($f=0.5$ 时)	77.91	67	72.04	79.10	53
本文方法 ($f=0.6$ 时)	74.12	63	68.11	22.22	14

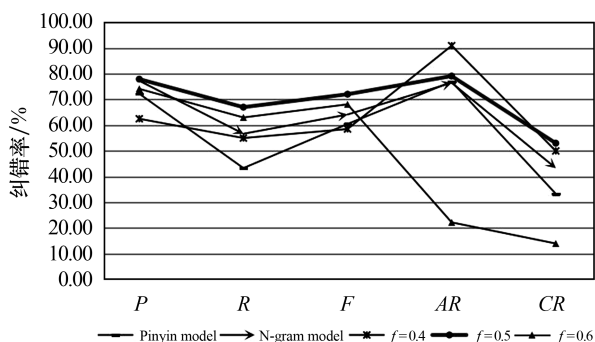


图3 不同方法纠错率的拆线图

3.3 实验数据分析

由表1可知,在文本相似度阈值为0.5时,总样本的纠错准确率为79.10%。部分问句纠错结果的实验数据如表2所示,其中下划线部分为错误字词和纠正后字词。在表2的11个样例中:样例1、3、4、6、7、9、10、11中的文本错误被成功发现。样例2、5由于查错失败导致没有进入纠错模块,如第2句,其正确文本应为“怎样采样”。该样例未被发现的原因是分词过程中“怎样才养”被切分为“怎样/才/养”,因分词部分均为停用词,因此未能纠错。样例8、10是由于纠错误判而导致纠正失败,如第10句,其正确问句应为“汉字编码是什么”。该样例分词结果为“汉/编码/是什么”,而“编码”在专业词典中的纠错建议是“熵编码”,因而导致将“汉编码”纠错为“汉熵编码”。其余样例均被成功纠正。由以上结果可以看出,本文方法在计算机专业领域对错误问句进行文本校正具有较好的准确率。

表 2 错误问句文本及纠错样例

序号	错误问句文本	是否发现	是否发现准确	纠错建议文本	纠错是否正确
1	什么是数据哭	是	是	什么是数据库	是
2	怎样才养	否	—	—	—
3	欧式距离怎么求	是	是	欧氏距离怎么求	是
4	次品统计怎么做	是	是	词频统计怎么做	是
5	蜜月长度怎么求	是	否	蜜月段长度怎么求	否
6	迷药长度怎么求	是	是	密钥长度怎么求	是
7	笛卡尔集怎么求	是	是	笛卡尔积怎么求	是
8	不嘛怎么求	否	—	—	—
9	传书延迟会引起什么	是	是	传输延迟会引起什么	是
10	汉编码是什么	是	是	汉嫡编码是什么	否
11	机器反义应用于什么领域	是	是	机器翻译应用于什么领域	是

4 结论

研究表明,在相似度为 0.5 的条件下,本文提出的基于字典树模型与编辑距离的纠错方法其准确率(77.91%)、召回率(67%)、F 值(72.04%)比传统的拼音纠错模型分别提高了 5.69%、23.67%和 11.57%,比 N-gram 纠错模型分别提高了 0.64%、10.33%和 7.89%,由此表明本文方法适用于专业课程领域中中文非真词错误的校对,尤其是通过拼音输入法造成的替换错误.本文提出的纠错方法仅是对词进行纠错,没有考虑到上下文语义信息;因此,在今后的研究中我们将采用深度学习与传统纠错技术结合的方法,根据上下文语义信息对错误文本进行校正.

参考文献:

[1] 张仰森,丁冰青. 中文文本自动校对技术现状及展望[J]. 中文信息学报,1998(3):3-5.
[2] KAREN KUKICH. Techniques for automatically correcting words in text[J]. ACM Computing Sur-

veys (CSUR), 1992,24(4):377-439.
[3] LIU B Q, WANG X L, WANG Y Y. Incorporating linguistic rules in statistical Chinese language model for pinyin-to-character conversion[J]. High Technology Letters, 2001,7(2):8-13.
[4] 纪兴光. 基于神经网络的带有拼写纠错功能的音字转换模型[D]. 北京:北京邮电大学,2019.
[5] 陶永才,海朝阳,石磊,等. 中文词语搭配特征提取及文本校对研究[J]. 小型微型计算机系统,2018,39(11):2485-2490.
[6] 陶永才,吴文乐,海朝阳,等. 一种结合 LSTM 和集成算法的文本校对模型[J]. 小型微型计算机系统, 2020,41(5):967-971.
[7] 吴淙. 中文文本校对关键技术研究与应用[D]. 成都:电子科技大学,2019.
[8] 曲强. MOOC 环境下课程智能问答系统的设计与实现[D]. 延吉:延边大学,2018.
[9] 王璐. 中文文本真词错误自动校对算法研究[D]. 杭州:浙江工商大学,2018.
[10] 欧晓聪. 基于自动纠错的最小编辑距离优化算法[J]. 网络安全技术与应用,2019(12):44-48.
[11] 孙芳媛. 基于倒排索引和字典树的站内搜索引擎的设计与实现[D]. 哈尔滨:哈尔滨工业大学, 2016.