基于上下文的拉丁维文拼写校对的研究①

何晋一[1,2]，陈红英[1]，姜文斌[2]，张海波[2,3]，刘 群[2]

[1](华南师范大学 计算机学院，广州 510631)
[2](中国科学院计算技术研究所 智能信息处理重点实验室，北京 100190)
[3](四川大学 软件学院，成都 610065)

摘 要：根据拉丁维文的特点，分析了拉丁维文常见的拼写错误类型，提出了一种将最小编辑距离、基于有向图模型的词语切分和 trigram 语言模型融合的方法，实现了基于上下文的拉丁维文的自动拼写校对系统，从而大大提高了拉丁维文的校对准确率。在新疆大学提供的维文语料库的测试中，拉丁维文的校对准确率达到了 90.1%。

关键词：拉丁维文；最小编辑距离；有向图模型；词语切分；语言模型；上下文；拼写校对

Latin-Uighur Spelling Check Based on Context

HE Jin-Yi[1,2], CHEN Hong-Ying[1], JIANG Wen-Bin[2], ZHANG Hai-Bo[2,3], LIU Qun[2]

[1](School of Computer, South China Normal University, Guangzhou 510631, China)
[2](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)
[3](School of Software Engineering, Sichuan University, Chengdu 610065, China)

Abstract: According to the characteristics of Latin-Uighur, this paper analyzed the common spelling error types of Latin-Uighur, and then proposed a method which merged the minimum edit distance, directed graph model based lexical segmentation, trigram language model together. Finally, we implemented the automatically spelling check system of Latin-Uighur based on context. It has increased the accuracy of Latin-Uighur spelling check largely. The experiment on the Uighur corpus provided by Xinjiang University reaches an accuracy of 90.1%.

Key words: Latin-Uighur; minimum edit distance; directed graph model; lexical segmentation; language model; context; spelling check

在自然语言处理领域中，对文本的拼写校对已经有了相当深入的研究，并且取得了一定可观的成果。虽然中文的拼写校对技术比较成熟[1,2]，但对维文来说，尤其是拉丁维文的拼写校对的研究还不是很成熟，现有的维文拼写校对系统还有很多缺陷。截止到目前为止，对维文的拼写校对研究基本上还是基于传统维文的，而且没有考虑维文上下文，这大大影响了维文拼写校对的效果。

维吾尔语是黏着性语言，其构词和构形都是在词干后追接不同的成分（称为词缀）来实现的，从理论上来讲维文的词汇量是无限的，词典中不可能包含所有的词或每个词的所有形态变化。因此，在对维文进行拼写校对时，词语切分的工作是十分必要的。

本文在前人提出的基于最小编辑距离的维文拼写校对[3]的基础之上，提出了一种将最小编辑距离、基于有向图的词语切分和 n-gram 语言模型融合的方法，并且实现了一个基于上下文的拉丁维文的拼写校对系统。

1 常见的拼写错误类型分析

1.1 文本错误分析

在输入文本字符时，常见的错误有以下几种[4,5]：

① 基金项目:国家自然科学基金(60736014)
收稿时间:2011-03-29;收到修改稿时间:2011-05-04

非词错误、真词错误、句法错误和语义错误。(1)非词错误，指文本中输入的字符串根本就不在词典中。如下面的：维文 aprle(正确的应该是 aprel，意思是“四月”)。这些错误可以概括为插入、脱落、替代和换位等。(2)真词错误，指由于输入人员的粗心所形成的字符串，虽然拼写正确，但却导致了上下文搭配不当，不是当前语境所需要的单词。(3)句法和语义错误往往是由于真词错误造成的，或由于原稿本身存在语法错误，或输入时丢失了某个单词。通常人们将“非词错误”称为“单词错误”，而将“真词错误”称为上下文相关的文本错误。

1.2 单词错误类型分析

我们在研究中发现，在输入的文本中，60%的错误是由“单个错误”(Single-error Misspelling)引起的。所谓“单个错误”，就是下列错误中的某一个：

①插入(Insertion)：把 urumqi(乌鲁木齐)错误地打成 urumiqi

②脱落(Deletion)：把 urumqi 错误地打成 urumq

③替代(Substitution)把 urumqi 错误地打成 urumwi

④换位(Transposition)把 urumqi 错误地打成 uruqmi

还有一种单词错误是词干和词缀连接错误。这种错误也是占 30%的错误。因为维文有 300 多种词缀，出错的概率也是比较大的。

2 最小编辑距离的计算

最小编辑距离[6]是指由一个字符串变化到另一个字符串所需要的编辑操作的最小数量。编辑操作是指对字符串中某一个位置的字符进行插入、脱落、替代和换位的操作，如 1.2 中所述。每进行这四种当中的一种操作，编辑距离加 1。在计算最小编辑距离时，需要计算字符串之间转换所需要的最小编辑操作次数。由于词典中的词比较多并且维文中第一个字母的写错率比较低，我们采用了 trie 树[7]的结构来组织词典，然后用动态规划的方法来计算最小编辑距离。

设 $S=s_1s_2\cdots s_m$ 为待匹配的字符串，而 $T=t_1t_2\cdots t_n$ 为目标字符串，建立一个 $m\times n$ 的矩阵 D，其中 m 和 n 分别为 S 和 T 的长度。该矩阵中 $D_{ij}(1\le i\le m,1\le j\le n)$ 的值从左到右，从上到下进行计算，D_{ij} 的值表示 $s_1s_2\cdots s_i$ 与 $t_1t_2\cdots t_j$ 之间的编辑距离，计算完之后，矩阵中 D_{mn} 的值就是两个串 S 和 T 的最小编辑距离。D_{ij} 的计算方法如下：

$$D_{ij}=\min\begin{cases}D_{i,j-1}+1(\text{插入})\\ D_{i-1,j}+1(\text{脱落})\\ D_{i-1,j-1}+1(\text{替代})\\ D_{i-2,j-2}+1(\text{换位})\end{cases}$$

我们将允许的最小编辑距离最大值设定为 2，下面给出待匹配串 S 在词典的 trie 树结构中的遍历匹配算法：

(1)从词典的 trie 树结构中从顶层向下搜索，计算当前节点与字符串 S 的编辑距离，对于顶层，由该层的第一节点开始，如果得到的编辑距离：

①小于或等于 2，则重复本步骤以对该节点的每一个子节点进行匹配；

②大于 2，终止以当前结点为根的子树匹配，转而比较当前节点所在层的下一个节点。重复本步骤；

③如果当前节点为末端节点，则找到一个目标字符串。输入该末节点代表的词到候选词当中。比较当前节点所在层的下一个节点。重复本步骤。

(2)当所有子树都被遍历或终止后，全部算法结束。

下面给出 tea 与 tra 以及 ten 与 tra 的最小编辑距离计算的示例，最小编辑距离即为矩阵中最后一行与最后一列交叉处的数：

		t	e	a
		0	1	2
t		1	1	2
r		2	2	1
a				

		t	e	n
		0	1	2
t		1	1	2
r		2	2	2
a				

图 1 最小编辑距离计算的示例

3 基于有向图模型的词语切分

由于维文是黏着性语言，理论上具有无限的词汇量，所以光靠查询词典来得出候选词有时候是行不通的，因此，我们需要将维文单词进行词干和词缀的切分，对其进一步分析。

我们为维文词语切分建立了一种基于有向图的生成式概率统计模型。该模型将维文语句的词语切分结果描述为有向图结构，图中节点表示分析结果中的词干、词缀，而边则表示节点之间的转移或生成关系，

它们刻画了词干、词缀连接成词的规律。生成式概率统计模型为这些转移或生成关系赋以合适的概率形式，词语切分的过程就是寻找其所有概率乘积最大的有向图。

我们把语句中各词的切分结果定义为链状结构：

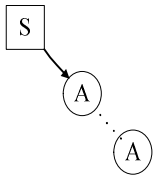

图 2　链状结构

这里，S(stem 表示词干)，A(affix)表示词缀。我们用虚线连接的两个 A 表示 0 个或者多个词缀。在词干到词缀之前以及词缀到后续词缀之间，箭头表示生成或者转移关系。对于整个语句，分析结果则可描述为树状结构：

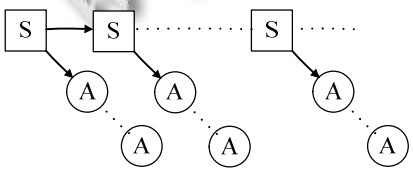

图 3　树状结构

与单个词的切分结果相比，整句分析结构中增加了相邻词的词干之间的生成或转移关系，从而在所有词干和词缀之间形成一个拓扑有序的树状结构。树中结点表示词干或者词缀，而节点之间的边表示词干到词干、词干到词缀以及词缀到词缀的生成或转移关系。我们为树中不同的边设计相应的权重，这些权重的度量反映了节点之间生成或转移规律的强弱，那么求整句词语切分结果的过程，即为在所有可能的候选树中寻找权重之和最高的树的过程。本模型中，我们用类似于隐马模型中的转移概率来描述树中边的权重。根据边指向对象的不同，我们设计了如下两种转移概率：

(1)P(S|S n-gram)：词干到词干的转移概率，类似于 n-gram 语言模型；

(2)P(A|S/A n-gram)：其它词缀的生成概率。S/A n-gram 指当前词缀之前词干或词缀组成的 n-gram 历史。

给定一个候选树 T，我们用这些概率的乘积表示该候选的整体生成概率：

$$P(T)=\prod_{S\in T}P(S|...)\ \times\prod_{A\in T}P(A|...\ )$$

为了简洁起见，公式中隐藏了两个条件概率的历史条件。容易看出，这可以理解为传统的 n-gram 语言模型向树结构的拓展。

4　基于上下文的候选词的选取算法

输入一段带有拼写错误的拉丁维文文本，对它的自动校对算法如下：

(1)先在词典的 trie 树结构中查找与该单词最小编辑距离小于或等于 2 的单词，若查找不到，则跳转到步骤(2)；若查到了，则直接跳转到步骤(3)；

(2)对当前词进行切分，将该词切分为词干、词缀部分，然后分别在词干表、词缀表的 trie 树结构中查找与当前词的词干、词缀最小编辑距离小于或等于 2 的词干、词缀，然后利用词缀的生成概率模型来挑选最佳的词干、词缀组合成的 5 个拉丁维文词；

(3)利用词到词之间的 n-gram 语言模型（上下文信息）来选出困惑度最低的最佳候选词，困惑度的计算公式如下：

$$P(S)=2^{-\frac{1}{l}\log_2(P(s_1)P(s_2|s_1)\cdots P(s_l|s_1s_2\cdots s_{l-1}))}$$

其中,S 表示输入的句子,l 代表 S 中单词的个数。

计算困惑度时，我们给每个句子加上头标记<s>和尾标记</s>，然后在前 n 个词已经进行拼写校对完以后，.根据前 n 个词的信息来计算当前词所有候选词的困惑度，从而挑选出困惑度最小的当前词。这个可以借助成熟的语言模型工具包 SRILM[8]来完成。

5　实验

我们在新疆大学提供的 11 万句子语料库上进行实验，该语料库共包括 108943 个完全正确的句子（作为词到词之间的 3 元语言模型的训练语料）、1057 个有拼写错误的句子以及对应的正确句子（共有 11537 个拼写错误的单词，其中 6956 个词是单词错误类型，而 4581 个词是真词错误类型）。我们使用的是包含 324896 个词条的词典、32589 个词干的词干表以及 319 个词缀的词缀表。其中，进行词语切分时，词干到词干的转移概率、词缀到词缀的转移概率、词干到词缀的生成概率，我们直接借助成熟的语言模型工具包 SRILM，以 WB 平滑方式训练 3 元语言模型。同样，

词到词之间 3 元语言模型我们也是借助 SRILM 得来。

测试结果如下：

表 1 不考虑上下文的实验结果

错误类型	错误单词数	纠正单词数	纠错率
单词错误	6956	6470	93.0%
真词错误	4581	0	0%
两者都有	11537	6470	56.1%

表 2 考虑上下文的实验结果

错误类型	错误单词数	纠正单词数	纠错率
单词错误	6956	6524	93.8%
真词错误	4581	3871	84.5%
两者都有	11537	10395	90.1%

从实验结果我们可以看出,考虑上下文时，即采用词到词之间的 3 元语言模型时，能纠正一些真词错误类型的单词；而不考虑上下文时，不能纠正真词错误类型的单词。不论考不考虑上下文，单词错误类型的单词纠错率比真词错误类型的单词纠错率高出很多。

6 结语

基于上下文的拉丁维文的拼写校对是一个难度比较大的研究课题。本系统实现了基于上下文的拉丁维文纠错功能，但是对于真词错误类型的单词的纠错率还是比较低的，光靠统计学的方法还是不能完全解决句法和语义上的某些真词错误的。我们应该针对拉丁维文的文本进行分析，加强句法和语义层次的校对策略研究，与目前的统计方法相结合，从而更进一步提高拉丁维文的校对准确率。语义问题是语言学与自然语言处理研究中的薄弱环节，语义错误的校对在拉丁维文的文本校对中仍未实现，需待进一步研究。

参考文献

1 龚小谨,罗振声,骆卫华.中文文本自动校对中的语法错误检查.计算机工程与应用,2003,8:98–100,127.

2 骆卫华,罗振声,龚小谨.中文文本自动校对的语义级查错研究.计算机工程与应用,2003,12:115–118.

3 玛依热·依布拉音,米吉提·阿不里米提,艾斯卡尔·艾木都拉.基于最小编辑距离的维语词语检错与纠错研究.中文信息学报,2008,22(3):110–114.

4 张仰森,俞士汶.文本自动校对技术研究综述.计算机应用研究,2006,6:8–12.

5 古丽拉·阿东别克,艾尔肯·伊米尔.维吾尔文校对中常见错误分析.计算机工程与应用,2005,27:181–183.

6 Kukich K. Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, 1992,24(4): 377–438.

7 Merrett TH, Shang HP. Trie Methods for representing text. Proceedings of the International Conference on Foundations of Data Organization and Algorithms. Lecture Notes in Computer Science vol. 730, Springer Verlag,1993:130–145.

8 Stolcke, Andreas. Srilm-an extensible language modeling tookit. Proceedings of the International Conference on Spoken Language Processing,2002:311–318.

（上接第 74 页）

虑到节点的不均匀性，提出簇内节点以单元格为单位进行地址复用，簇头节点以复用区域进行地址复用，以通信标志位来避免消息混淆。这样能抵御节点的不均匀性带来的压缩算法的弱化。仿真试验的结果表明：在相同的地址长度下，本文算法能容纳更多的节点，在节点密集分布的情况下，其地址压缩性能较为突出。

参考文献

1 Kan BQ, Cai L, Zhu HS, et a1. Accurate energy model for WSN node and its optimal design.Journal of Systems Engineering and Electronics,2008,19(3):427–433.

2 Jacobson V. Compressing TCP/IP Headers for Low-Speed Serial Links. Request for Comments 1144, February 1990.

3 Bormann C, Burmeister C, Degermark M, et al. RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed. Request for Comments 3095. July 2001.

4 Montenegro G, Kushalnagar N. Transmission of IPv6 Packets over IEEE 802.15.4 Networks. 2007.

5 Chin KW, Lowe D, Sanchez RG. A new technique for reducing MAC address overheads in sensor networks.IEEE Communications Letters, 2006,10(5):338–340.

6 田野,盛敏,李建东.一种新的传感器网络 MAC 地址分配算法.西安电子科技大学学报,2006,33(5):716–720.