The Spelling Problem 解题报告

石家庄市第二中学 张若天

1. 题目来源

Codechef Nov 14

2. 题目大意

给一个字典和一个可能有 4%单词错误的小写英文句子，输出更正后的句子。

错误由以下四种构成：

1.交换了两个字母

2.漏掉了一个字母

3.多打了一个字母

4.打错了一个字母

最多只会在一个单词里出现一种错误。

大部分单词出现在了字典中。

数据范围：文本大小约是 10MB。

3. 计分方式

首先我们将你输出的文本以空白字符为依据分成许多单词。

然后我们将每个单词与正确答案相应位置上的单词进行一一比较。

若当前单词在输入文本中是正确的，在输出文本中是不正确的，你将被扣 1 分。

若当前单词在输入文本中是错误的，在输出文本中是正确的，你将得到 3 分。

如果你的分数被扣到 0 分以下，将被记为 0 分。

你初始分数为 100，所以你输出原文就可以得到 100 分。

为了适合 tsinsen 上的测评，你每个测试点的得分为 上述得分/该测试点标准分。标准分为一个参考解得到的分数。

4. 算法讨论

4.1. 算法一

预处理出字典中所有单词的错误方式。

- 交换两个字母
- 漏掉一个字母
- 多大一个字母
- 打错一个字母

我们可以枚举每一种错误方式的每一种情况，生成所有可能的错误字符串，再存在一个 Trie 树中或哈希表中。

然后去扫描错误的句子，若一个单词在错误字典中出现，则纠正。

假设英文单词最长为 26，则时间复杂度$O(\text{句子总长度} * 26)$，空间复杂度$O(\textit{字典大小} * 26 * 26)$。

4.2. 算法二

在上一个算法的基础上，我们发现预处理整个字典的复杂度较高，而整个句子只有 4%出错并且句子中大部分单词都在字典中。可以

考虑对于每个不在字典中出现的单词生成所有可能的正确单词，然后在字典中查询。

假设英文单词最长为 26，则时间复杂度O(句子总长度 $*$ 26 + 4% $*$ 句子总长度 $*$ 26 $*$ 26) ，空间复杂度O(*字典大小*)。

4.3. 算法三

以上一个算法为基础。我们注意到句子中单词不全出现在字典中。而上一个算法是一旦发现单词不在字典中就去枚举所有的情况。我们考虑在代码中扩充一下字典，来减少正确的单词由于不在给定字典中而枚举的情况。需要注意代码长度限制。

4.4. 算法四

以上一个算法为基础。我们注意到一个拼错的单词可能对应着许多正确的单词，比如单词”thes”可能是”this”也可能是”thus”，但前者概率高很多。

我们考虑给字典中每个单词添加权值，在有多个对应的正确单词时选择权值最高的那个。

权值的设置，可以考虑打一个单词常用度表放进代码里，或者也可以学习读入的字符串，毕竟有 96%的成分是正确的，也有可能更符合该字符串的语境。

4.5. 算法五

以上一个算法为基础。我们尝试对错误单词生成所有情况进行记忆化。比如单词 A 没有在字典中出现，最后的结果是用了 B 单词进行替代。我们就把二元组(A,B)记录下来，下次再遇到 A 时，直接返回 B。有两个好处，一是可以减少重复错误的单词查询复杂度，二是一些没有在字典中但是在文本中多次出现的正确词可以减少枚举量。

假设英文单词最长为 26，则时间复杂度O(句子总长度 $*$ 26 + 4% $*$ 句子总长度 $*$ 26 $*$ 26) ，空间复杂度O(*字典大小*)。

4.6. 其他高级的机器学习的算法

通过学习一些词组组合搭配以及英语构词法来提升正确性，如 N-gram 模型。

5. 考察内容

字符串处理，哈希，Trie 树