

英文单词拼写错误自动检查系统

包利强，雷雪林

2017 年 6 月 22 日

1 课题简介

在英语写作中，单词的拼写错误是一种较为常见的错误，因此英文单词拼写错误自动检查的研究成为了自然语言处理技术的重要子课题。课题11要求我们实现一个英文文本中单词拼写错误自动检查系统，如果在检查的同时还要求进行改错，那么这个系统将包括两个基本的部分，即检错（detecting）和改错（correcting）。在现代自然语言处理的研究中，英文单词拼写改错系统又可以按照不同的情境分别进行研究，其一是上下文无关单词的检错和改错（*context-free word error detection and correction or isolated-word error detection and correction*），在该情境下，每个单词的检查都是与它的上下文环境相独立的，为了检测该单词是否正确，很自然的一个做法是在词典中搜索该单词，如果搜索成功则拼写正确，否则错误，在这种方法中，算法对词典非常敏感。这个方法的一个很大的缺点是无法对那些拼写正确但是不符合上下文的单词进行检错，因此依赖上下文的检错和改错（*context-dependent error detection and correction*）也是一个非常重要的研究课题，但由于这个课题需要语法相关的分析，因此比较复杂。[1]

2 项目目标

目前在自然语言处理研究中，拼写检查和纠错方面的研究已经进行了很长一段时间，有了很多成果和基础，也有了不少开源的拼写检查与纠错的成熟系统，主要以GNU Aspell和Hunspell为代表。我们小组在阅读相关文献和实践的基础上，将动手实现一个简单的类GNU Aspell系统，由于在1中所解释的原因，我们将只实现上个一下文无关的单词检错和改错系统。拼写错误改错意味着要对文本中出现的错误单词提供可能的正确单词建议，一般是一个概率递减的建议列表，但在自动改错时，我们只选择概率最大的一个单词替换错拼单词即可。

参考文献

- [1] S. Deorowicz and M. G. Ciura, “Correcting spelling errors by modelling their causes,” *International journal of applied mathematics and computer science*, vol. 15, no. 2, p. 275, 2005.