错别字的定义应是相同读音或相似形态中文字之间的错用、混用等，因为汉字词库是确定的，能输入到电脑系统里的都是正确的字，那么错别字识别的任务就是找出用错地方的汉字，这个任务是很有挑战性的，属于科技前沿，人工智能领域自然语言理解是未解之题。

在 2004 年，语音识别未取得理论上的重大突破时，李开复曾提出了基于统计学的方法，这个思想可以借鉴过来，就是**建立错字库**。搜狗输入法可以一定程度辨别错字，如：

这是基于读音的，可能后台有一个很丰富的词库。

对于一行文本：写完就去次饭。并有错词库：次饭。写了简单的 demo

>> demo.py

当然建立词库是关键和重点，热门的机器学习方法获取有所帮助，个人能力搞不定。

还有另一个设想，自然语言之所以难处理是因为格式随意，没办法范式处理，有一门范式语言**本体**以及相应的 **owl** 语言，高度概念化和结构化，以**主谓宾**的形式表达知识，英语或者汉字都有主谓宾，可以据此分割。