Ngram

当前识别的字为 c_t

前面已识别的字为$prev = \{c_0, \ldots, c_{t-2}, c_{t-1}\}$

c_t的前一个词为w_{t-1}

目前的做法

目前我的思路是将该任务转换成判断c_t前一个词w_{t-1}的语义是否完整，若完整的话就直接输出该单字。

因此我们需要：1 . 得到前一个词w_{t-1} 2. 判断w_{t-1}的语义是否完整

目前的我的解决方法是通过类似计算熵的大小，来判断c_t 和 $prev \in w_{t-1}$ 结合后的词与 w_{t-1} 哪个语义更完整，输出更完整(即概率更大的那个)来同时解决上面两个问题的。

目前的问题

1. 输出的分数的合理性
2. 是否有更好的方法来解决词的界定上的问题，这个任务比较类似分词
3. 中英文混合时，没法纠正英文的拼写错误
4. 英文大小写的问题
5. 标点符号与数字的概率设定
6. 目前"一"经常错判断成~ - 等标点符号
7. 训练ngram语言模型时的一些细节，比如是否该加入标点符号、数字、其他语种的字符，英文的大小写是否该用同一个case