

ngram 实验报告

olhwdb2

实验结果

方法	WER
baseline	6.34
stupid-backoff	5.50
ngram-max	6.58
ngram-mean	6.58
我们的方法	4.87

参数设定

这里使用的都是5-gram模型。输出分数与原识别置信度通过一定比例结合，有 $\alpha * P_{rec} + \beta * P_{lm}$ 。其中 P_{rec} 为识别模型输出的置信度； P_{lm} 为语言模型输出的分数， $\alpha = 0.68$ 。

baseline

不加入语言模型，即识别结果直接输出

stupid-backoff

ngram经典匹配方法，即从长到短寻找匹配项，输出最长的可匹配项的分数。

ngram-max

与stupid-backoff相似，但是输出的不是最长可匹配项的分数，而是最大的分数。

ngram-mean

输出从最长到最短所有分数的平均值。

我们的方法

通过判断当前字与前向字是否能组合成一个词，来找到当前字的最佳分割点，输出组合后的词的分数。

bad case

BASELINE	我们的方法	GROUNDTRUTH
改变3	改变3	改变了
官光经济	官光经济	宏观经济
拒绝杀访	拒绝杀访	拒绝采访
降水过程	降水过程	降雨过程
Harmony	Hororony	Harmony

可以看出：

1. 存在数字改错的问题
2. 词首难以改正，较为依赖识别模型
3. 中间能组合成词的很难改对
4. 英文上还有些问题

除此之外，标点符号上也存在一些问题，目前难以改对。