# 实验结果数据分析认识与结论

训练集：

pku：19056句 来自报纸（含有：贺词，训话，诗文，新闻，人名）

msr：86924句 文学作品，新闻，影视介绍评价，政治，条款

平均困惑度：

pku 数据 bigram模型：4132.187

pku 数据 trigram模型：9080.823

msr 数据 bigram模型：2148.493

msr 数据 trigram模型：7077.380

可以看到在采用同种训练数据和测试数据时，trigram 模型的困惑度要比bigram 模型的困惑度大的多，这是因为trigram 考虑的历史更完整，相应的句子概率较小。采用不同数据时，可以看到msr数据上的困惑度更小，效果更好，我认为这是因为msr数据中的句子更多，语言模型学习的更好。

pku 训练 msr 测试 bigram模型：3048.676

pku 训练 msr 测试 trigram模型：8105.584

msr 训练 pku 测试 bigram模型：7632.463

msr 训练 pku 测试 trigram模型：14975.679

可以看到在pku训练，msr测试的效果要比msr训练，pku测试的效果要好，msr数据涵盖的范围要比pku数据广，所以用pku来作msr的测试集效果显然不会好。

混合数据训练 bigram 模型 pku 测试：4977.8733

混合数据训练 trigram 模型 pku 测试：11774.780

混合数据训练 bigram 模型 msr 测试：2084.420

混合数据训练 trigram 模型 msr 测试：7496.348

在混合数据后，msr测试的结果要比对应的pku测试的结果要好，这是因为，在混合数据中，来自msr的数据量要比来自pku的数据量大的多，所以相应的在msr测试集上的效果更好。

结论：

1. 在ngram确定时，句子概率越大，语言模型越好，困惑度越小。
2. 训练数据越多，效果越好。
3. 考虑的历史越多困惑度越大。
4. 如果想建立某领域的语言模型，用相应领域的数据效果会更好，所以要处理好数据。
5. 针对不同任务可以选择对标点符号
6. ngram存在OOV问题，需要平滑，可以根据开发集数据选取超参数，本次实验没有开发集，我选取了训练集的1/10作为开发集得到超参数。
7. 困惑度计算时，多个概率连乘带来浮点数下溢问题，需要转化为对数和形式。
8. ngram模型优点：计算简单，可解释性强，适用于固定搭配，诗词，歇后语等。
9. ngram模型缺点：(1) 前序依赖不灵活，缺乏长期依赖，只能建模到前 n-1 个词；(2) 随着 n 的增大，参数空间呈指数增长；(3) 数据稀疏，难免会出现OOV的问题；(4) 单纯的基于统计频次，相同语义和语法功能词不能替代泛化能力差。