|  |  |
| --- | --- |
| **自然语言处理及应用**  **实验报告** | |
|  | |
|  |  |
| **名称** | 基于IBM模型的机翻文本对齐 |
| **姓名** | 黄隆宁 |
| **班级** | 硕1224班 |
| **学号** | 3121155049 |
| Email | hln18773372567@stu.xjtu.edu.cn |
| **日期** | 2020-10-23 |

# 实验目的

分别实现IBMModel1和IBMModel2两个机器翻译模型，并使用这两个模型实现英/中翻译的单词对齐。

# 实验环境

Ubuntu 18.04

Python3.8

调用build-in库：

time：用于算法运行时间统计

math：用于数学函数的计算

pip依赖：

numpy：绝大部分的数据表示和数学计算

nltk：用于自然语言数据的预处理

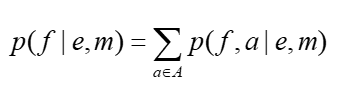
# 实验方法

所有的实验代码都更新在我的Github，可在Github在线查看：<https://github.com/HLNN/NLP>

IBMModel是于上世纪九十年代开始，IBM公司推出的一系列统计机器翻译模型，本次实验要求实现的是其中最基础的IBMModel1和IBMmodel2。

IBMModel1不考虑单词在句子中的位置，将句子所有排列方式同等看待，仅注意单词与单词之间的对应概率。不适用于较为复杂的句子结构。

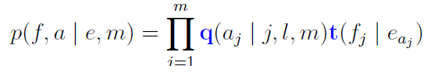
IBMModel2在IBMModel1的基础上，额外增加了单词在句子中位置的概率。统计源句子和目标句子长度给定时，目标句子第i个词与源句子第j个词对应的概率，这个概率在IBMModel1中被认为时固定的。



翻译模型计算的时所有的对齐情况下概率之和的最大值，而如果我们考虑给定翻译的情况下的最佳翻译，求解对齐概率的最大值。

C:\Users\HLN\AppData\Local\Temp\1634975465(1).png

该概率可以进一步推导为：



其中对于IBMModel1，认为对齐概率q为恒定值。

具体IBMModel和IBMModel2模型的实现，参考了NLTK库的源代码，在其基础上完成。

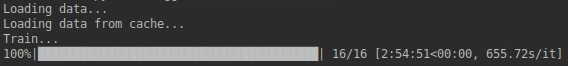
为了提高对齐效果，首先对平行语料库进行预处理，使用nltk的word\_tokenize将分词后的语料句子转为token列表。再使用nktk中的PorterStemmer、LancasterStemmer和SnowballStemmer等词干化工具进行词干化。

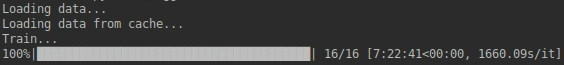
模型的概率表使用EM算法迭代求解。在E过程中，使用概率表计算count值，在M过程中再用count值迭代概率表

1. **实验结果**

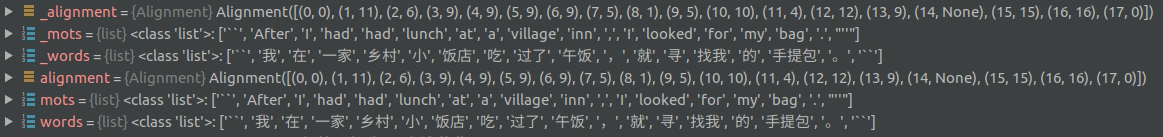
因为文本对齐没有Ground Ture作为评估依据，因此随机从对齐结果中选取展示对齐效果。

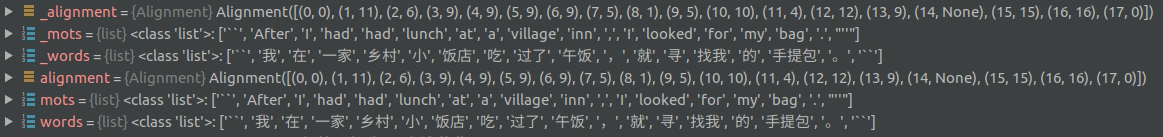
训练集包括100000条的平行语料，经过16轮的EM算法迭代。其中IBMModel1模型的训练时间约3h，IBMModel2模型的训练时间约7.5h，该训练时间在每次训练时稍有差异。

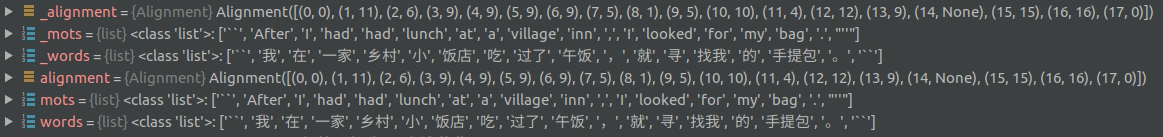




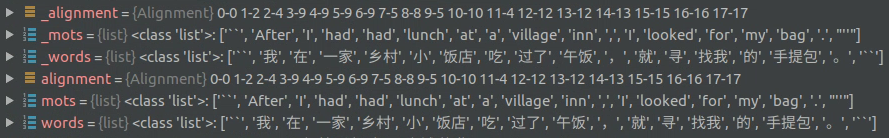
下图是IBMModel1得到的其中一对平行语料的对齐结果，数据以处理为nltk中的AlignedSent类型：

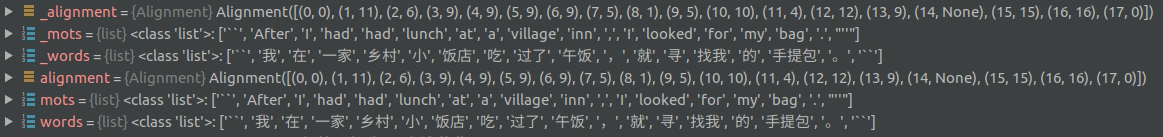


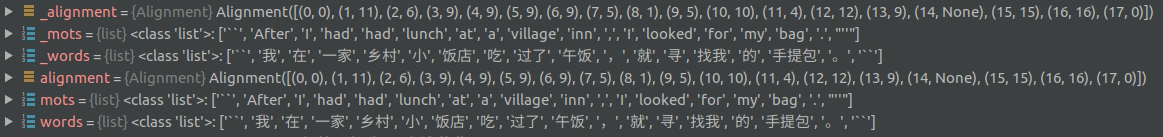




下图是IBMModel2得到的其中一对平行语料的对齐结果，数据以处理为nltk中的AlignedSent类型：







可以看出IBMModel2在一些地方仍对齐效果不佳，但相较于IBMModel1，IBMModel2在解决词与词之间的对应关系上有明显改进。当一个词或标点符号在句子中重复出现时，IBMModel2能将先出现的源语言单词与先出现的目标语言单词对齐，后出现的与后出现的对齐。

综合其他语料的对齐结果，可以看出IBMModel1模型对名词单词能实现非常好的对齐功能，对于一些有不同含义的动词，或者一些表示状态时态的单词对齐效果较差。同时如果相同单词在一个句子中多次出现，模型无法判断与平行句子中的哪一个对应。比如上图中的目标句子出现两个“我”，模型无法判断两个“我”分别和哪一个“I”对应。

1. **遇到问题及解决思路**

相同单词多次出现，目标句子的多个相同单词分别和源句子中的多个相同单词中的哪一个对应，这个问题IBMModel1模型很难解决。

另外模型要求有足够大且预处理良好的数据集。在小范围测试时，比如只取平行语料库的100或1000条时。概率表中大量的翻译概率和对齐概率均为默认值，此时模型难以有效对齐。