读书报告 9.7-9.13

61518426 周之遥

·读书内容

学习第8章Part-of-Speech Tagging部分内容

Part-of-Speech Tagging就是词性标注，给每个文本中的word预测词性。

词性有两种，closed class和open class。

open class主要包括noun、verb、adj、adv，会随时代不断更新

closed set包括各种常用的function word。

Penn Treebank tagset是常用的词性标准，Brown、WSJ、Switchboard都是标注数据集。

一般都是先分词再进行词性标注。

对一个有多重含义的word，词性标注可以进行disambiguate，词性标注的一个基本的baseline就是认为一个word的词性是最常出现的词性。

·下周计划

学习HMM和MEMM