第八周周记

周一	
完成内容	观看了视频《什么是神经网络》和《神经网络在做什么》
内容描述	1. 了解神经网络的基本概念 2. 了解机器学习与数据点的一些概念
未解决问题	无

周二	
完成内容	观看了视频《什么是激励函数》
内容描述	1. 了解了公式 y=AF(Wx)的相关概念
未解决问题	无

周三	
完成内容	阅读吉林大学秦赞的硕士论文《中文分词算法的研究与实现》
内容描述	1. 了解了正向最大匹配算法 2. 学习使用双向匹配算法进行歧义判断 3. 学习了常见的人名识别方法
未解决问题	无

周四	
完成内容	学习了 python 的一些知识
内容描述	学习了一些文件与异常的知识
未解决问题	无

周五	
完成内容	学习了 python 的一些知识
内容描述	学习了数据保存的一些知识
未解决问题	无

周末	
完成内容	继续阅读论文《Co-training an Improved Recurrent Neural Network with Probability Statistic Models for Named Entity Recognition》
内容描述	更深入地了解了 Co-training 算法
未解决问题	无

工程汇总	
完成任务	阅读了几篇论文，继续学习 python 知识
任务描述	初步学习了 python 的一些基本知识
代码量	无
未解决问题	无

论文汇总

论文列表	《中文分词算法的研究与实现》
论文摘要	《中文分词算法的研究与实现》 在本文中，对自然语言处理的基础性问题中文分词进行了研究。在常见的基于词典的分词算法和基于统计的分词算法的基础之上，提出了一种基于词典与基于统计相结合的分词方法，充分利用了基于词典分词的高效性及基于统计的分词的较强的歧义处理的能力。首先使用改进的双向匹配方法对待切分句子是否包含歧义进行判断，如果判断没有歧义，将分词结果直接作为输入传递给中文人名识别模块；如果判断包含歧义，该句子需要基于统计的方法进行切分，首先，使用正向全切分算法对待切分句子进行处理，得到所有的可能的切分情况，然后，根据训练得到的 bin-gram 语言模型对各种切分情况进行可能性的计算，选出概率最大的三种结果加入到备选集，下一步使用基于隐马尔可夫(HMM)的评价算法对备选集中的三种切分进行出现的可能性评估，选取概率最大的一种作为切分后的结果，最后将该结果作为中文人名识别模块的输入，进行中文人名的识别操作，对于中文人名的识别，本文采取了一种规则与统计相结合的识别算法，人名识别模块的输出便是最终的处理结果。在实际中，只有少部分的中文句子包含歧义，这就意味这大部分的句子使用双向匹配算法就可以得到解决，少部分的句子使用基于统计的分词方法进行歧义的消除，这样就最大程度地兼顾了效率与准确性。实验结果表现出了较好的分词效果。 本文的创新之处在于：使用了词典与统计相结合的分词方法；对基于词典的分词方法进行了改进，并对传统的整词二分法词典及双字哈希词典均进行了优化，引入了词长数组，对于词典正文部分按照长度分开存储，并进行排序，提高了词典的匹配效率并减少了空间占用，引入了结尾词长数组从而使逆向匹配算法可以和正向匹配算法使用同一个词典，实现了词典的复用；使用了一种三层的存储结构存储 bin-gram 语言模型，提高了运算速度；采用了规则与统计相结合的中文人名识别方法，表现出了较好的人名识别率。 最终实现了一个中文分词的系统，提供了便捷的操作界面，系统集成了各种词典结构及分词方法，并支持词典的添加删除等维护操作，方便操作及对比研究。
未解决问题	无

下周任务	
工作	继续阅读论文并学习算法
论文	继续寻找与中文分词和命名实体识别相关的论文
其他	无
汇总	了解更多与我的课题相关的知识

日期:2018/02/19 - 2017/02/24