第十一周周记

周一	
完成内容	阅读了 2016 年大连理工大学金留可的硕士论文《基于递归神经网络的生物医学命名实体识别》
内容描述	（１）基于传统递归神经网络的生物命名实体识别 基于 Biocreative II GM 的评测任务，分别米用了两种不同的递归神经网络（ＲＮＮ）进行生物医学命名实体识别。然后通过在隐层和输出层分别增加递归连接来扩展ＲＮＮ，并利用布朗聚类和ＬＤＡ无监督学习算法构建特征层模式化范围更广的上下文信息。 （２）基于长短时记忆递归神经网络的生物命名实体识别为了能够解决传统的ＲＮＮ在处理长句子时出现的问题，更进一步提升命名实体的识别性能，本文采用了 LSTM 替代ＲＮＮ的隐层计算方式。然后在双向 LSTM 的基础上，保留预训练的词向量不变，和微调之后的词向量一起构成双词向量输入到 LSTM 中。最后，通过计算双词向量的差值得到句子向量表示，同时增加新的读入控制口，构建本文的 ST—BLSTM 模型实现命名实体识别。
未解决问题	无

周二	
完成内容	阅读 2016 年计算机工程与应用期刊的论文《基于条件随机场的中文领域分词研究》
内容描述	首先通过条件随机场的基本特征模板和自己定义的特征得到一个初次分词结果，之后利用领域词典对结果进行逆向最大匹配，达到一个校正的效果，对不同领域的分词，只需要增加相应的领域词典，可极大地提高条件随机场模型对未登录词的识别，提高分词的正确率，并且也不需要对不同领域训练新的模型，从而解决了条件随机场模型和单纯的词典分词适应性差的问题。针对分词的歧义问题，提出了固定词串消解、动词消解、词频消解三种方法消除歧义。
未解决问题	无

周三	
完成内容	阅读 2016 年软件导刊期刊的论文《一种基于词典的中文分词改进算法》
内容描述	基于词典的分词系统的分词效率很大一部分取决于词典选择，所以词典能否快速高效的存储和查找数据是关键。目前最常用的词典机制有基于 tire 索引树的词典和基于整词二分法的词典机制。 基于词典的分词算法主要有最大正向匹配算法 MM 和最大逆向匹配算法 DMM 两种，对这两种算法的研究时是这篇文章进行算法改进的基础。
未解决问题	无

周四	
完成内容	阅读四川大学学报(工程科学版)期刊的论文《基于深度学习的中文微博命名实体识别》
内容描述	首先对字典中的每个字特征随机初始化，并基于该字典获取微博句子中每个字的特征，对不同长度的微博句子特征采用周期卷积得到固定长度的特征，

	以此作为微博的句特征; 然后训练微博数据自动编码器，通过栈式自动编码器得到高层微博句子的特征; 最后采用高层句特征与字特征的组合训练字的标注网络模型，并基于该标注模型得到未知字的标注值，获取微博句子中的命名实体。
未解决问题	无

周五	
完成内容	写开题报告
内容描述	
未解决问题	无

周末	
完成内容	写开题报告
内容描述	
未解决问题	无

工程汇总	
完成任务	阅读了几篇论文
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	[1] 《基于递归神经网络的生物医学命名实体识别》 [2] 《基于条件随机场的中文领域分词研究》 [3] 《一种基于词典的中文分词改进算法》 [4] 《基于深度学习的中文微博命名实体识别》
论文摘要	摘 要: 针对条件随机场分词不具有良好的领域自适应性，提出一种条件随机场与领域词典相结合的方法提高领域自适应性，并根据构词规则提出了固定词串消解，动词消解，词概率消解三种方法消除歧义。实验结果表明，该分词流程和方法，提高了分词的准确率和自适应性，在计算机领域和医学领域的分词结果 F 值分别提升了 7.6%和 8.7%。
未解决问题	无

下周任务	
工作	运行 python 程序,继续研究 Co-training 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:2018/03/12 - 2018/03/17