第九周周记

周一	
完成内容	阅读 2012 华南理工大学戴思明的硕士论文《互联网文本热点信息实体识别研究与应用》
内容描述	1. 了解互联网新闻人名实体的识别 2. 了解互联网地名机构名实体的识别 3. 了解了产品名称、属性、属性值、评论命名实体识别
未解决问题	无

周二	
完成内容	阅读 2017 年江苏大学曹菲的硕士论文《基于 Hash 和 CRF 的中文分词算法研究》
内容描述	1. 了解基于 HASH 的正向回溯算法 2. 了解基于 CRF 的命名实体识别
未解决问题	无

周三	
完成内容	学习了一些 python 的知识
内容描述	了解了 API 和 BIF 的一些问题
未解决问题	无

周四	
完成内容	继续学习 python 知识
内容描述	了解了 python 的异常处理机制和特定指定异常的一些知识
未解决问题	无

周五	
完成内容	阅读 2017 年南京师范大学王蕾的硕士论文《基于神经网络的中文命名实体识别研究》
内容描述	1. 基于神经网络的字符级中文命名实体识别知识 2. 基于神经网络的片段级中文命名实体识别知识
未解决问题	无

周末	
完成内容	看了 Python 的视频
内容描述	了解了卷积神经网络 CNN 的一些知识
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	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] 《互联网文本热点信息实体识别研究与应用》 [2] 《基于 Hash 和 CRF 的中文分词算法研究》 [3] 《基于神经网络的中文命名实体识别研究》
论文摘要	本文针对现有主要的中文分词算法进行了分析和研究。主要的创新点如下： 1）提出了基于 Hash 的正向回溯算法，改进了由于利用回溯方法解决歧义带来的复杂度高的问题； 2）针对 CRF 模型识别命名实体过程中的因观察窗口小带来的内嵌以及一些外国译名、网络新词问题，提出了一种组合方法来改进命名实体识别效果。 论文主要内容如下： （1）提出基于 Hash 的正向回溯算法。该算法在回溯机制的基础上，在查询词语时采用新的扫描方式，并结合 hash 词典解决了最长匹配字的问题。此外，针对采用回溯机制发现和处理歧义带来的匹配次数翻倍，导致的时间复杂度高的问题，通过加入结束标识位判断，减少时间复杂度。该方法相对于其他回溯方法，减少了时间复杂度。 （2）提出利用组合进行命名实体识别的算法，该算法将 CRF 识别命名实体和正向最大匹配相结合，改进了命名实体的识别效果。该组合算法基于 CRF 模型，利用基本特征、实体列表特征、边界特征和组合特征构造相应的模板，然后根据实验好坏，决定采用何种模板；针对英文名识别不准确、网络新名词和观察窗口小导致的机构名内嵌的问题，提出建立常见外文名字字典、网络新词以及 5 字以上的机构名字典将匹配法与 CRF 模型结合，利用规则修正分词结果，从而提高准确率和召回率。 （3）为了验证所提出算法的可行性，在 Eclipse 开发平台上，利用 Java 和面向对象的程序设计思想来开发了一套中文分词原型系统，实验结果表明分词效果理想。
未解决问题	无

下周任务	
工作	继续阅读论文并理解 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:2018/02/26 - 2018/03/03