第五周周记

周一	
完成内容	阅读了曹勇刚的论文《面向信息检索的自适应中文分词系统》
内容描述	1. 了解了分词的基于二元迭代的切分方法 2. 了解了将搜索引擎索引数据作为词频词典实现逐级递进的分词的方法
未解决问题	无

周二	
完成内容	阅读了西北大学硕士林冬盛的论文《中文分词算法的研究与实现》
内容描述	1. 学习了基于字符串匹配的分词方法 2. 学习了基于统计的分词方法 3. 学习了基于理解的分词方法
未解决问题	无

周三	
完成内容	阅读了电子科技大学的硕士张小欢的论文《中文分词系统的设计和实现》
内容描述	1. 学习了最大匹配算法 2. 学习了最小词切分方法 3. 学习了 N-最短路径法
未解决问题	无

周四	
完成内容	阅读了哈尔滨工业大学薛天竹的硕士论文《面向医疗领域的中文命名实体识别》
内容描述	1. 了解了深度学习的神经网络模型 2. 了解了基于 LSTM 的 NER 实现
未解决问题	无

周五	
完成内容	继续学习 python 的有关知识
内容描述	观看了莫烦 python 的视频《python 基础教程》《Github 代码管理》
未解决问题	无

周末	
完成内容	安装了 theano 框架
内容描述	完成了 ubuntu 16.04+Anaconda+theano+keras **安装**
未解决问题	无

工程汇总	
完成任务	阅读了四篇论文
任务描述	了解了中文分词和命名实体识别的常用方法

代码量	无
未解决问题	无

论文汇总	
论文列表	[1] 《面向信息检索的自适应中文分词系统》 [2] 《中文分词算法的研究与实现》 [3] 《中文分词系统的设计和实现》 [4] 《面向医疗领域的中文命名实体识别》
论文摘要	中文分词是按照特定的规范将汉语中连续的字序列切分为合理的词序列的过程。作为自然语言处理基础性任务,中文分词已经被广泛应用在相关领域中。因此,研究中文分词算法具有重要的理论和现实意义。 为了满足上层应用对分词实用性要求,本文将机械分词和基于统计的分词法有机结合,提出了基于词典和统计规则的中文分词算法。该算法首先使用切分速度快的机械分词法对预处理后的文本进行初步切分,采用改进的双向最大匹配检测法检测出歧义字段,并运用基于二元统计模型的全切分消解歧义。其次,采用基于角色的命名实体识别方法识别出未登录词。最后,引入规则库对分词结果进一步修正。本文的研究工作主要有采用二次索引的词典结构,提升词典查找速度,使用对象序列化技术实现词典文件的加载反序列化和词典对象的序列化。在歧义检测方面,提出了改进的双向最大匹配检测算法,不仅能检测到链长为奇数的歧义字段,而且能检测出所有同时满足链长为偶数且交段长度为的歧义字段。在歧义字段上,采用全切分法消解歧义。在未登录词识别方面,将隐马尔科夫模型中解决编码问题的前向算法用以角色标注,采用角色模式集上的模式串匹配出中文专有名词。使用一个小型校正规则库进行分词碎片的修正。目前中文分词软件包大都以语言开发,而作为主流开发语言之一的,其中文分词组件相对较少。因此,在分词算法的研究基础上,设计并实现了支持 Java 语言的中文自动分词系统。 实验表明,该中文分词算法在,内存的环境下,切分速度约为 21000 字秒,分词准确性指标一值达到了左右,基本能够满足大部分上层应用的要求
未解决问题	无

下周任务	
工作	继续查找相关的论文和继续阅读相关文献。
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:2018/01/29 - 2018/02/03