第四周周记

周一	
完成内容	阅读了论文《领域自适应中文分词系统的研究与实现》
内容描述	1. 了解了领域适应性分词的难点 2. 了解了基于条件随机场的中文分词方法 3. 了解了基于神经网络的中文分词方法
未解决问题	无

周二	
完成内容	阅读论文《面向手机短信的命名实体识别研究》
内容描述	1. 手机短信的语言特点和手机短信的专家知识的构建 2. 手机短信的命名实体识别 3. 手机短信的人名实体的识别
未解决问题	无

周三	
完成内容	无
内容描述	
未解决问题	

周四	
完成内容	阅读论文《领域自适应中文分词系统的研究与实现》
内容描述	1. 了解了歧义切分和未登录词识别 2. 了解了神经网络概述和神经网络分词实现 3. 了解了多模型的中文分词方法
未解决问题	无

周五	
完成内容	下载了《head first python》的 pdf
内容描述	学习了创建了简单的 python 列表和向列表增加数据
未解决问题	无

周末	
完成内容	看了莫烦 Python 的视频
内容描述	了解了神经网络和动态神经网络的基本知识
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	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] 《领域自适应中文分词系统的研究与实现》 [2] 《面向手机短信的命名实体识别研究》 [3] 《领域自适应中文分词系统的研究与实现》
论文摘要	中文分词是指将连续的字序列依照特定的规范切分为合理的词序列的过程。作为自然语言处理最基本的一个步骤，是信息检索、知识获取以及机器翻等应用必须处理的关键环节。因此，研究中文分词具有重要的理论和现实意义。 本文提出了一种基于字的多模型分词方法。该方法采用神经网络模型结构针对每个字单独建立模型。由于中文汉字本身带有语义信息，不同的字在不同语境中其含义和作用不同，造成每个字的构词规律存在差异。与现有字标注分词方法不同的是，该方法能够有效区分每个特征对不同待切分字的影响，从而学习出字构词的特殊性规律。通过与单模型方法、CRF 方法以及前人的工作进行对比，本文提出的基于字的多模型方法取得了更好的分词效果。并在 SIGHAN Backoff2005 提供的中文简体语料 PKU 和 MSR 上，取得的 F 值分别为 93.4%和 95.5%。 根据上述方法，面向领域自适应分词任务，本文提出了一种基于字的领域自适应分词方法。由于字模型相互独立，模型更新时，保留迁移性能强的字模型，对迁移性能弱的字模型进行更新训练。解决了大规模切分数据难与共享，源领域与目标领域数据混合需要重新训练等问题。对目标领域进行分词时，通过模型的自适应能力实现领域自适应。特征嵌入的表示方法能够有效地解决特征稀疏问题，本文采用特征嵌入来表示输入特征。实验结果表明，本文提出的分词方法能够明显提高领域适应性能力。 最后，设计并实现了领域自适应中文分词系统。该系统可以实现利用已有的基础模型对输入的句子或文本进行分词，并且支持添加相关领域词典，还可根据待分词领域训练数据对基础模型进行更新，从而获得相关领域较好的分词结果。
未解决问题	无

下周任务	
工作	用 python3 读取文件里中文并分词
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:2017/01/22 - 2017/01/27