第十四周周记

周一	
完成内容	阅读了论文《基于 CRF 的互联网文本命名实体识别研究》
内容描述	本文提出使用条件随机场（ＣＲＦ）并结合匹配规则的方法对互联网文本进行命名实体识别。通过分析互联网文本特点，对文本进行规范化，利用统计和规则相结合的方法进行识别。实验获得了良好效果，但仍然存在缺陷，识别效果有待提高。下一步要进行的工作包括扩大训练语料规模、获取更加简单有效的规则以及对上下文信息的处理等。
未解决问题	无

周二	
完成内容	阅读论文《基于神经网络的中文词法分析系统的研究与实现》
内容描述	在命名实体识别上，我们在输入层将随机初始化的词嵌入、预训练词嵌入和词性标签的向量表示以非线性方式组合，再利用双向 LSTM 编码输入，将双向输出结果采用非线性方式组合，最后采用全局最优的标签预测方法，在测试集上取得的 F1 值为 94.71%，比基准线高 0.57 个百分点。
未解决问题	无

周三	
完成内容	阅读论文《基于条件随机场的命名实体识别及实体关系识别的研究与应用》
内容描述	命名实体识别是将文本中的元素分成预先定义的类，如人名、地名、组织机构名、时间、货币等等。作为自然语言的承载信息单位，命名实体识别属于文本倍息处理基础的研究领域，是信息抽机构取、信息检索、机器翻译、问答系统等多种自然语言处理技术中必不可少的组成部分。在实体识别领域，国外科研机构针对英文实体的识别已取得了突出的成绩，识别准确度达到以上。由于中文在分词及语义方面存在着众多的困难，内针对该问题还处于研究和探索阶段。所以针对中文实体及关系的识别的研究有养重大的怠义。
未解决问题	

周四	
完成内容	阅读论文《基于条件随机场的学术期刊中理论的自动识别方法》
内容描述	本文认为命名实体识别的方法可以分为依据词典的方法和使用规则的方法两大类。基于统计的方法实质是以概率分布的思想利用规则，因此归入使用规则的方法。关于如何获取词典和规则，则有人工专家法和机器学习的方法两类。人工专家法在早期基于规则的算法中较为常用，专家总结规则并提交给算法使用。机器学习的方法按照是否需要标注语料，一般分为有监督的学习 (Supervised learning) 、弱监督的学习 (semi -supervised) 和无监督 (unsupervised) 的学习。
未解决问题	无

周五	
完成内容	阅读论文《一种中文人名识别的训练架构》

内容描述	在自然语言处理（NLP）任务中，许多高度工程化的 NLP 系统应用，大都采取基于特定任务特征的线性统计模型，这些模型由应用背景激发、受领域知识的限制，通过面向工程的专用特征发现数据表示。这些特征通常由一些特征提取的辅助工具预处理而得到，是一种监督学习的训练方法。但是这种方法不仅会导致复杂的运行时依赖关系，而且要求研发人员必须拥有大量的语言学知识。 为了减少这种依赖，必须捕捉关于自然语言的更多的一般性，分析语言的元信息如词性、实体、语法、句法等，以期获取一种更一般的描述方法，减少甚至忽略先验领域知识对模型的影响，用无监督学习的方式，尽量避免工程化特征，在大规模未标记数据上学习产生模型。 本文描述一种基于深度神经网络的字词训练模型，通过发现其内部表示，尽量避免了工程特征对于模型的限制，采取一种无监督学习的方式对中文人名识别进行了研究，最后通过实验验证该模型的合理性。
未解决问题	无

周末	
完成内容	看了下 github 里的 python 代码
内容描述	了解了神经网络 LSTM-CRF 模型
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 的一些知识
代码量	无
未解决问题	无

论文汇总	
论文列表	[1] 《基于 CRF 的互联网文本命名实体识别研究》 [2] 《基于神经网络的中文词法分析系统的研究与实现》 [3] 《基于条件随机场的命名实体识别及实体关系识别的研究与应用》 [4] 《基于条件随机场的学术期刊中理论的自动识别方法》 [5] 《一种中文人名识别的训练架构》
论文摘要	
未解决问题	无

下周任务	
工作	运行 python 代码
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:2018/04/02 - 2018/04/07