第十三周周记

周一	
完成内容	阅读了论文《Multi-domain evaluation framework for named entity recognition tools》
内容描述	从非结构化文本中提取结构化信息对定性数据分析非常重要。利用 NLP 技术进行定性数据分析将有效加速注释过程，实现大规模分析并提供对文本的更多见解以提高性能。获得文本见解的第一步是命名实体识别（NER）。直接影响 NER 过程表现的一个重大挑战是定性数据的领域多样性。代表性文本根据其领域在许多方面有所不同，包括分类，长度，形式和格式。在本文中，我们讨论并分析跨领域的最先进工具的性能，以阐述其稳健性和可靠性。为了做到这一点，我们开发了一个标准的，可扩展的和灵活的框架，以分析和测试工具的性能，使用表示不同领域文本的语料库。我们从各个角度和各个角度对工具进行了广泛的分析和比较。由此产生的比较和分析对于提供最先进工具的整体说明非常重要。
未解决问题	无

周二	
完成内容	阅读论文《Recognizing irregular entities in biomedical text via deep neural networks》
内容描述	命名实体识别（NER）是生物医学文本挖掘的一项重要任务。大多数先前的工作集中于识别由连续的单词序列组成并且彼此不重叠的规则实体。在本文中，我们提出了一个称为 Bi-LSTM-CRF 的神经网络模型，该模型由双向（Bi）长期短期记忆（LSTM）和条件随机场（CRF）组成，以识别正规实体和不规则实体的组成部分。然后根据手动设计的规则将组件组合起来建立最终的不规则实体。此外，我们提出了一个叫做 NerOne 的新模型，它由 Bi-LSTM-CRF 网络和另一个 Bi-LSTM 网络组成。 Bi-LSTM-CRF 网络执行与上述模型相同的任务，并且 Bi-LSTM 网络确定是否应该组合两个组件。因此，NerOne 会自动组合这些组件，而不是使用手动设计的规则。我们在两个数据集上评估我们的模型，以识别规则和不规则的生物医学实体。实验结果表明，在特征工程较少的情况下，我们模型的性能与最先进的系统性能相当。我们表明，自动组合的方法与手动设计规则的方法一样有效。我们的工作可以促进生物医学文本挖掘的研究。
未解决问题	无

周三	
完成内容	阅读论文《Character-level neural network for biomedical named entity recognition》
内容描述	生物医学命名实体识别（BNER）提取重要的命名实体，如基因和蛋白质，这是自动化系统中的一项具有挑战性的任务，它可以挖掘生物医学文本中的知识。先前的最先进的系统需要大量的特征工程，词典和数据预处理形式的特定任务的知识以实现高性能。在本文中，我们引入了一种新型的神经网络架构，通过使用双向长期短期记忆（LSTM）和条件随机场（CRF）的组合，自动获得词和字符级表示，从而不需要大多数特征工程任务。我们在两个数据

	集上评估我们的系统：JNLPBA 语料库和 BioCreAtIvE II 基因提及（GM）语料库。我们通过超越以前的系统获得了最先进的性能。就我们所知，我们是第一个研究深层神经网络，CRF，词嵌入和特征级表示在识别生物医学命名实体。
未解决问题	无

周四	
完成内容	阅读 2015 年广东外语贸易大学赵文慧的硕士学位论文《基于语料库的法律文本翻译中显化现象的研究》
内容描述	研究结果表明显化现象广泛地存在于法律文本翻译中，显化的内容涉及语法、时态、词汇、篇章结构和背景文化等方面，其中主要显化手段主要包括人称代词的添加和时态的变形，情态动词和连接词的添加，文化背景信息的解释和情境说明；研究发现法律系统差异，语言差异，社会文化之间的差异和译者主体因素等是显化的主要成因。
未解决问题	无

周五	
完成内容	阅读 2016 年物联网技术期刊的论文《中文分词算法研究与分析》
内容描述	当前，中文分词算法主要有三大类，即基于字符串匹配的分词算法、基于统计的分词算法以及基于理解的分词算法。其中，基于字符串匹配的分词算法是根据某种分词策略将要分词的字符串和一个“足够大”的词典进行匹配，从而切分出中文单词 ；基于统计的分词算法则是通过统计相邻字与字之间的联合出现概率来判断是否是一个单词 ；基于理解的分词算法是在中文分词时进行句法、语义分析，并利用句法信息和语义信息来处理歧义现象。而这三类中文分词算法即代表着中文分词算法的研究三大方向。
未解决问题	无

周末	
完成内容	阅读了 2018 年计算机科学期刊的论文《基于 BLSTM 的命名实体识别方法》
内容描述	命名实体识别是典型的序列标注问题，而循环神经网络是一种很有效地解决序列标注问题的神经网络模型，能够有效地利用数据的序列信息，具有一定的记忆功能。但 RNN 无法很好地处理长距离依赖问题，并且训练算法存在梯度消失或爆炸问题。文献中提出了一种利用门限机制对历史信息进行过滤的长短时记忆（LSTM）模 型，有 效 地 解 决 了 RNN 中存在的问题。针对本文要解决的问题和 LSTM 模型的优点，提出一种中文命名实体识别方法。
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] 《Multi-domain evaluation framework for named entity recognition tools》 [2] 《Recognizing irregular entities in biomedical text via deep neural networks》 [3] 《Character-level neural network for biomedical named entity recognition》 [4] 《基于语料库的法律文本翻译中显化现象的研究》 [5] 《中文分词算法研究与分析》 [6] 《基于 BLSTM 的命名实体识别方法》
论文摘要	中文分词作为搜索引擎以及自然语言处理的重要组成部分，是当前这一领域的研究热点和难点之一。文中首先分析了中文分词的特点，包含基于字符串匹配分词算法、基于统计分词算法、基于理解分词算法这三大类的各种中文分词算法。并通过分析和对比，对各种中文分词算法进行了总结与展望。
未解决问题	无

下周任务	
工作	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:2018/03/26 - 2018/03/31