**第十二周周记**

|  |  |
| --- | --- |
| 周一 | |
| 完成内容 | 阅读了2017年山东大学武威的硕士学位论文《法官解释法律的方法与路径》 |
| 内容描述 | 由于立法、执法、司法权力的分立，法官只有忠于法律文本才能在法律适用中获得正当性，从而实现解释结果的妥当性。法官是法律的实际使用者而不是法律的创造者，法官忠于立法本意，并将法律视为案件审判的权威性文件，因此法官的法律解释权能够有助于解决个案出现的问题，将立法者对某些问题的价值判断运用到化解矛盾中，这也有助于解决立法本身所存在的滞后性问题。 |
| 未解决问题 | 无 |

|  |  |
| --- | --- |
| 周二 | |
| 完成内容 | 阅读2016年现代计算机(专业版)期刊的论文《命名实体识别综述》 |
| 内容描述 | 有监督学习方法将命名实体识别看做序列标注问题 。序列标注模型包括：隐 马 尔 科 夫 模 型 Hidden Markov Models (HMM)，最大熵马尔科夫模型 Maxi-mum Entropy Markov Models (MEMM) 和 条件随机场Conditional Random Fields (CRF)等。这些模型都是基于大量的标注语料，定义一系列实体，通过学习得到基于特征的判别规则。 |
| 未解决问题 | 无 |

|  |  |
| --- | --- |
| 周三 | |
| 完成内容 | 阅读2017年小型微型计算机系统期刊的论文《双向循环网络中文分词模型》 |
| 内容描述 | 双向循环网络中文分词模型，能有效获取待分类字符的上下文特征，避免局部窗口大小的限制，使用长短时记忆网络( Long Short-Term Memory Neural Network，LSTM) 作为神经网络隐藏层，同时增加一层反向LSTM 抽取字符的将来信息特征． 提出一种语言模型预训练的网络权值初始化方法，该模型同时得到中文字符embeddings 分布式向量特征． 在标准分词数据集上测试表明该模型取得比以往统计标注方法更好的效果． 通过对比实验结果发现深层神经网络能提取出不逊于人工总结的分词特征 |
| 未解决问题 | 无 |

|  |  |
| --- | --- |
| 周四 | |
| 完成内容 | 修改并上交毕业设计开题报告 |
| 内容描述 |  |
| 未解决问题 | 无 |

|  |  |
| --- | --- |
| 周五 | |
| 完成内容 | 阅读2015年华中师范大学程志刚的硕士论文《基于规则和条件随机场的中文命名实体识别方法研究》 |
| 内容描述 | 在命名实体识别过程中，由于数词和时间词的表达结构较为规范，所以采用基于规则的方式来进行识别；同时考虑到人名、地名和机构名的不规则性，而采用条件随机场模型，大量的命名实体识别研究表明，基于单个汉字提取上下文特征能够取得更为丰富的特征值，所Ｗ在本文中Ｗ字为单位进行文本切割，获取每个字的上下文特征，然后采用条件随机场进行机器学习，最终实现对句子中包含的组织机构名、人名和地名的自动识别。在模型的训练过程中，采用多种不同的模板进行训练，选择效果最隹的模板，这在一定程度上提高了实体识别的效果。 |
| 未解决问题 | 无 |

|  |  |
| --- | --- |
| 周末 | |
| 完成内容 | 阅读2017年计算机技术与发展期刊论文《基于循环神经网络序列标注的中文分词研究》 |
| 内容描述 | 循环神经网络 ( RNN)广泛应用于机器翻译、语音识别、图像描述生成等领域。相比于传统前馈神经网络，其特点是可以存在有向环，将上一次的输出作为本次的输入。而与前馈神经网络的最大区别是: 前馈神经网络要求输入的上下文是固定长度的，也就是说n－gram 中的 n 要求是个固定值，而在 LSTM基础上扩展的循环神经网络不限制上下文的长度，可以充分利用所有上文提供的信息来预测下一个词，本次预测的中间隐层信息可以在下一次预测里循环使用。为此，文中试图将循环神经网络模型应用在中文分词方面。 |
| 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 |  |
| 代码量 |  |
| 未解决问题 |  |

|  |  |
| --- | --- |
| 论文汇总 | |
| 论文列表 | [1] 《法官解程法律的方法与路径》  [2] 《命名实体识别综述》  [3] 《双向循环网络中文分词模型》  [4] 《基于规则和条件随机场的中文命名实体识别方法研究》  [5] 《基于循环神经网络序列标注的中文分词研究》 |
| 论文摘要 | 摘要: 分词是中文自然语言处理中的关键技术。在自然语言处理中，序列标注在中文分词中有着极其重要的应用。当前主流的中文分词方法是基于监督学习，从中文文本中提取特征信息。这些方法未能充分地利用上下文信息对中文进行分割，缺乏长距离信息约束能力。针对上述问题进行研究，提出在序列标注的前提下利用双向循环神经网络模型进行中文分词，避免了窗口对上下文大小的限制，可以获得一个词的前面和后面的上下文信息，通过增加上下文能够有效地解决梯度爆炸和爆的问题，然后再在输入层加入训练好的上下文词向量，取得相对较好的分词效果。实验结果表明，该算法的使用可以达到97.3% 的中文分词准确率，与传统机器学习分词算法相比，效果较为显著。 |
| 未解决问题 | 无 |

|  |  |
| --- | --- |
| 下周任务 | |
| 工作 | 用Co-training算法设计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:2018/03/19 – 2018/03/24