武漢大學

基于深度学习的法律文书命名实体识别研究与实现

指导老师：彭敏（博导）　　报告人：刁永祥（专硕）

目录

CONTENTS

背景意义

背景意义

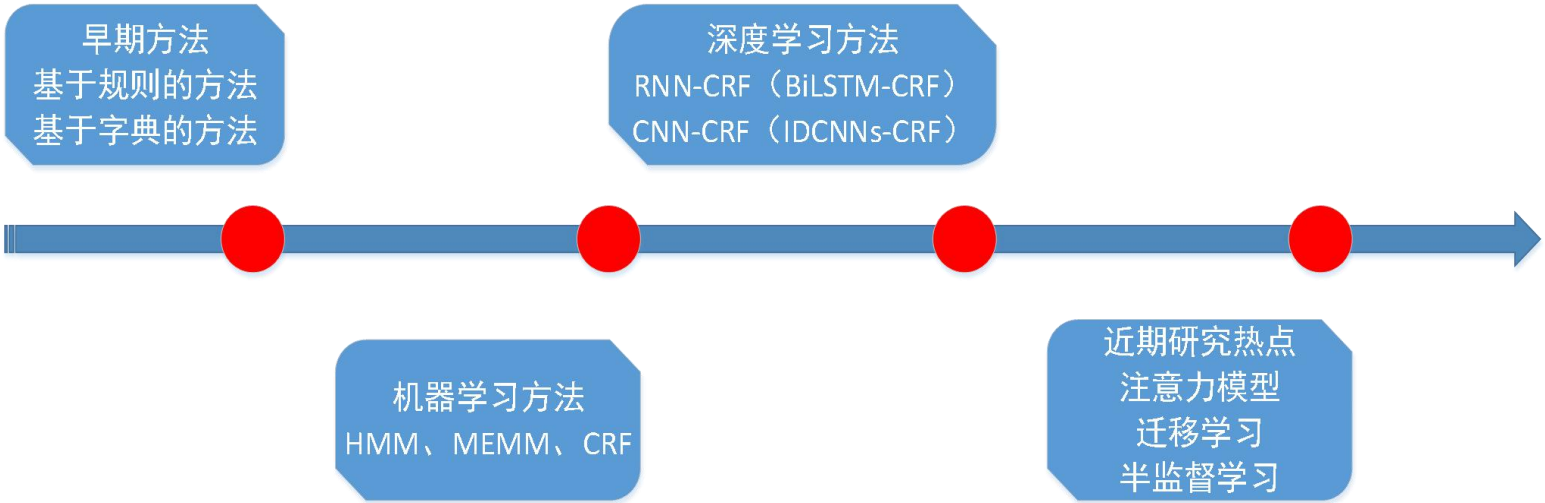
- ✓命名实体识别是自然语言处理的一项基本任务，旨在识别文本中具有特定意义的实体（如人名、地名、组织名等）
- ✓命名实体识别已成为信息抽取、自动文摘、机器翻译、问答系统等任务的重要组成部分
- ✓学术上，有三大类（实体类、时间类、数字类）和七小类（**人名、地名、组织名、时间、日期**、货币、百分比）

背景意义

- ✓互联网+ 的时代，信息量呈几何级数般的增长，政务、商务等对文本信息的处理和分析节奏越发急促、紧密
- ✓国家号召司法公开，增强司法透明度，防止司法权滥用，及时公布法院生效的裁判文书
- ✓调查显示，中国裁判文书网的访问量已超过210亿次，文书总量近5800万篇
- ✓本文研究重点：从海量法律文书中自动、准确、快速识别命名实体，主要是人名、地名、机构名，为后期的研究奠定基础

相关研究

相关研究

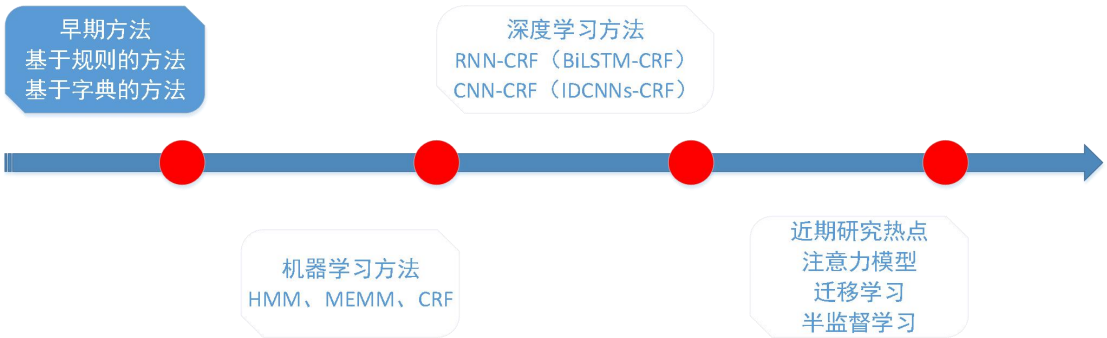

早期方法
基于规则的方法
基于字典的方法
机器学习方法
HMM、MEMM、CRF
深度学习方法
RNN-CRF（BiLSTM-CRF）
CNN-CRF（IDCNNs-CRF）
近期研究热点
注意力模型
迁移学习
半监督学习

相关研究

✓预定义规则模板（根据词法、句法、语法等语言学知识）

✓善于捕获命名实体特征，规则表达简单，易于理解

✓早期系统，NTU、FACILE、OKI

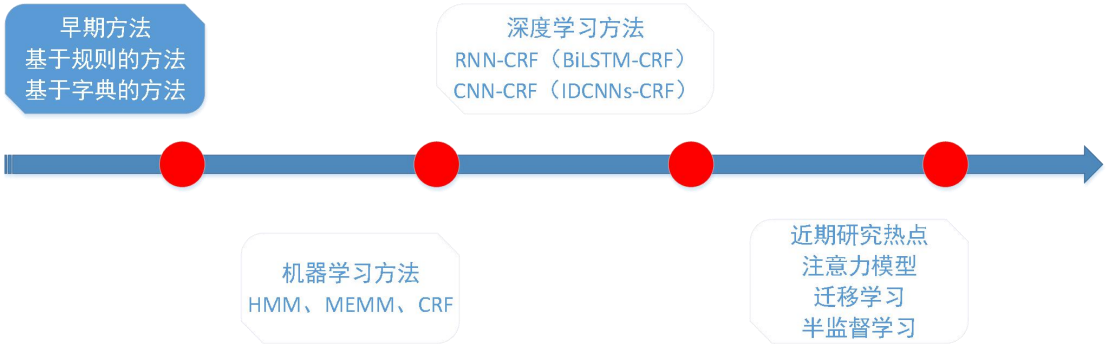


相关研究

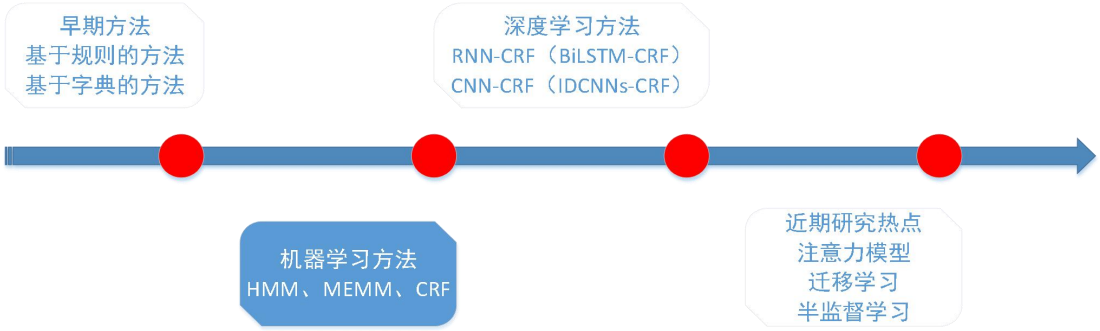
✓词典法意将命名实体收录于词典，匹配识别文本中的实体

✓结合词典与规则，实时有效生成新词典

✓适应性差，难以胜任各种领域、语言的命名实体识别任务

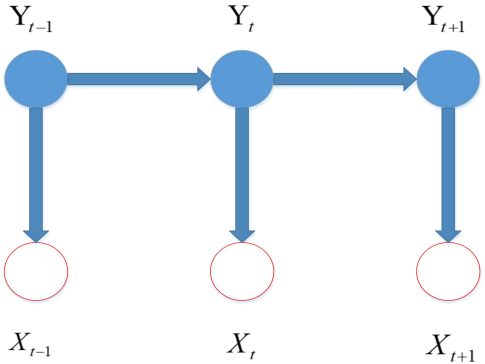


相关研究



- ✓NER视为序列标注问题（利用大规模语料学出标注模型，再对新句子的各个位置标注）
- ✓代表性模型有HMM、 MEMM、 CRF

相关研究



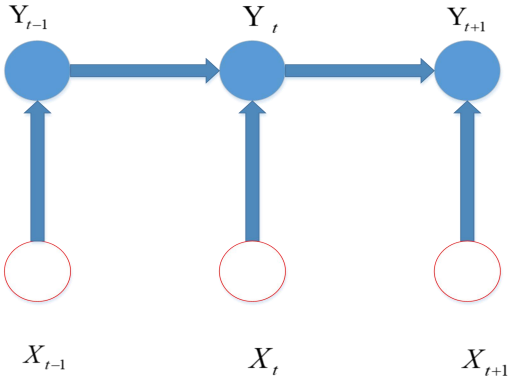
✓HMM结构图示

✓X：观察序列，Y：隐藏状态序列

✓HMM定义序列标注的三个基本问题：概率计算、解码、参数估计问题

✓分别通过前向算法、Viterbi算法、EM算法来求解

相关研究



✓MEMM结构图示

✓X：观察序列，Y：隐藏状态序列

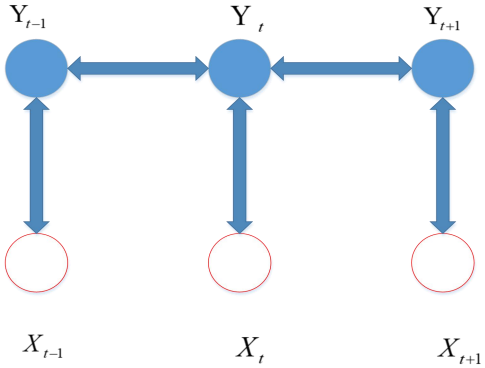
✓McCallum认为HMM存在两个问题：

① 序列标注任务，观察序列需大量特征来刻画（大小写、上下文）

② 理所应当地，观察序列来决定状态序列

✓结合HMM和MEM特点，提出MEMM

相关研究



✓X：观察序列

✓Y：隐藏状态序列

✓标签之间具有约束关系，I-LOC不可能出现在B-PER后，但是MEMM不能识别

✓基于MEMM，加入标签间约束的CRF应运而生

$$P(y_t \mid y_{t-1}, x) = \frac{1}{Z(y_{t-1}, x)} \exp\left(\sum_j \lambda_j t_j(y_{t-1}, y_t, X, t) + \sum_k \mu_k s_k(y_t, X, t)\right)$$

相关研究

✓CRF、 MEMM应用时的明显劣势

① [xt=ws, yt=ln]会在词集合和标签集合中排列组合，所以特征数量随着上述集合的数量增加而呈现指数级增长

② 特征通常具有领域特性，不同任务必须定义不同的特征函数

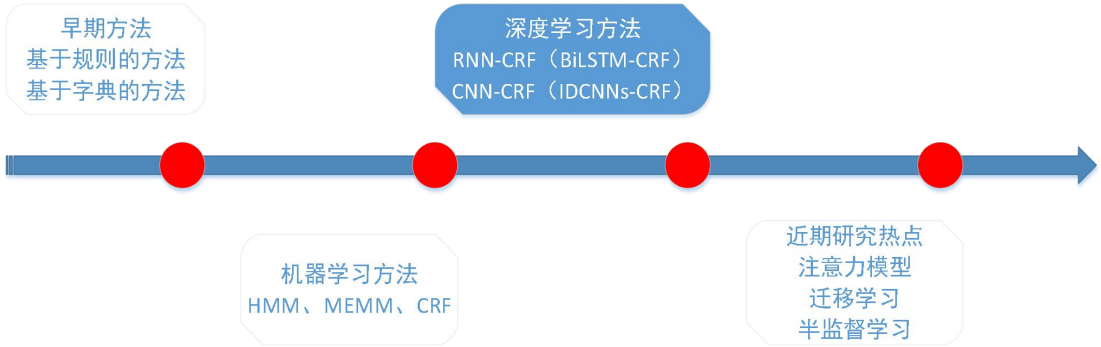
相关研究

✓NER中，深度学习方法的应用主要有两种

① 获得词向量，作为额外特征加到输入中，利用统计学的方法完成NER

② 建立基于神经网络的模型完成NER

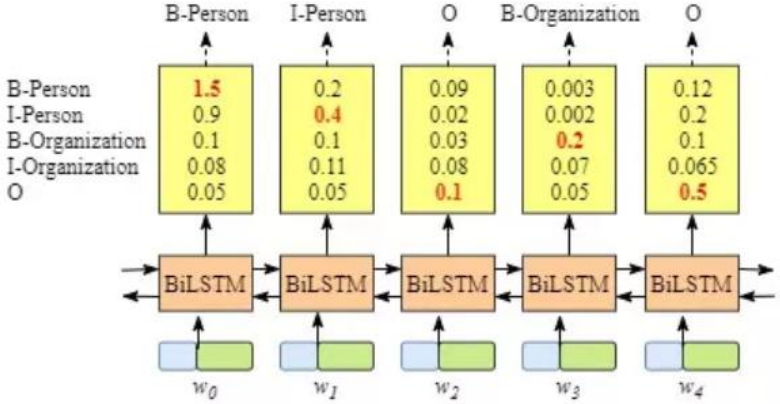
✓主流模型有BiLSTM-CRF、IDCNNs-CRF

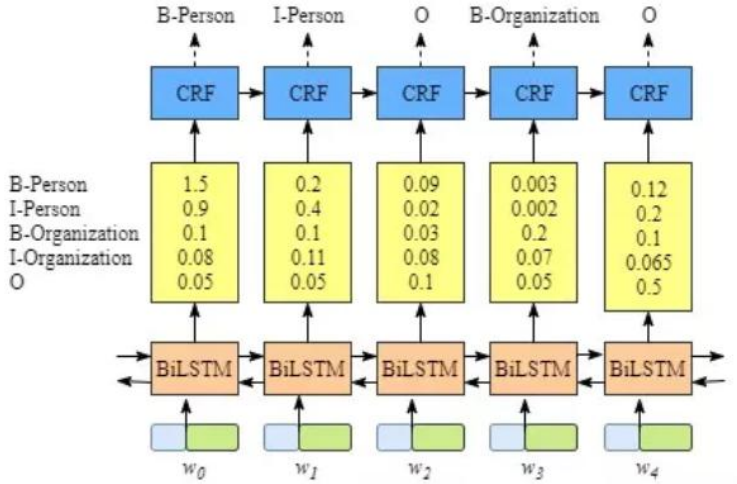


相关研究

✓Baidu AILab proposed BiLSTM-CRF models for sequence tagging in 2015

✓CMU raised nerual architectures for named entity recognition in 2016



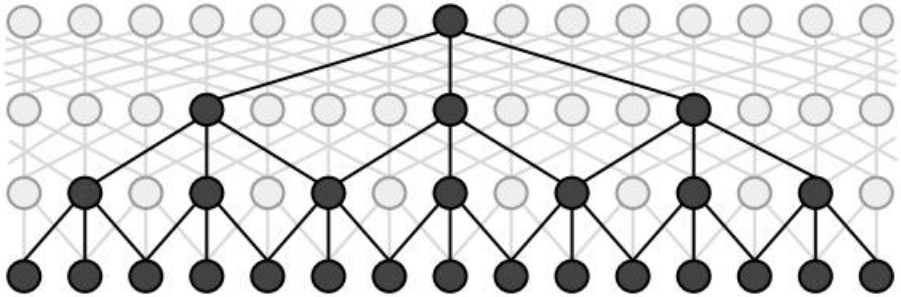


✓无CRF层

✓有CRF层

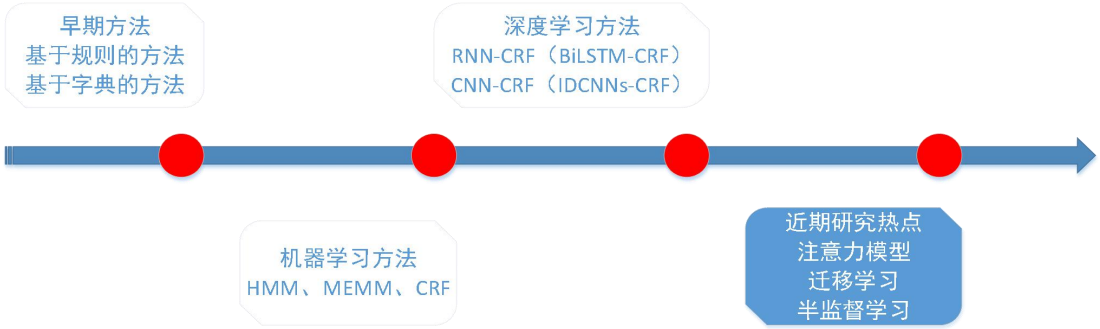
相关研究

✓Fast and Accurate Entity Recognition with Iterated Dilated Convolutions(2017)



✓filter width = 3, dilation width = 1, 2, 4

相关研究



- ✓ 应对缺乏标记训练数据的窘境

① NITE: A Neural Inductive Teaching Framework for Domain-Specific NER(EMNLP, 2017, ZJU)

② Semi-supervised sequence tagging with bidirectional language models （ACL, 2017）

主要研究内容

主要研究内容

- ✓ 法律文书的特点总结
- ✓ 语料搜集、标注（BIEOS or BIO标签集）

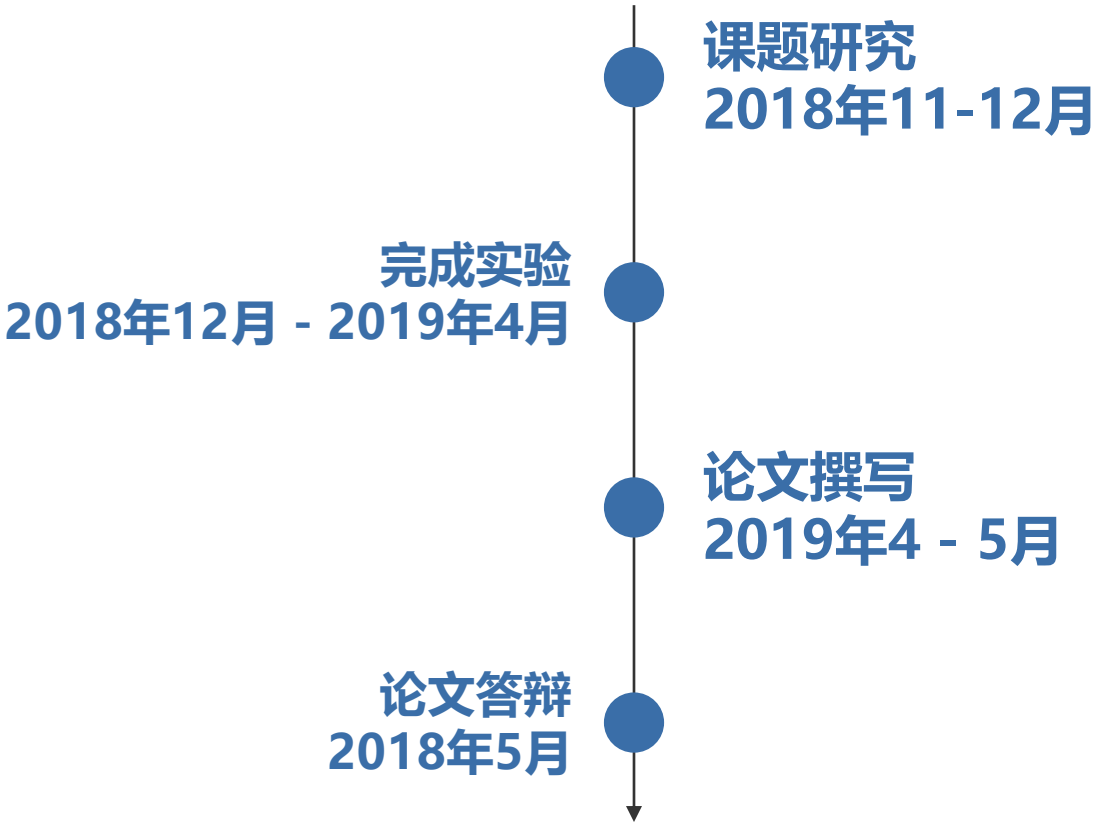
① 截至目前，还未找到公共的语料标注资源

② 暂时考虑，主动学习，或者人工标注

- ✓ 篇章级NER的难点（OOV）与应对策略（暂定：整数线性规划）
- ✓ 对比实验，BERT or Word2Vec， IDCNNs-CRF or BiLSTM-CRF

论文进度安排

论文进度安排


课题研究
2018年11-12月
完成实验
2018年12月 - 2019年4月
论文撰写
2019年4 - 5月
论文答辩
2018年5月

谢谢大家
水平有限，请多指教

汇报人：刁永祥　汇报时间：2018年12月12日