

SEU 知识抽取-命名实体识别

2019年11月14日 14:37

• 一、实体识别的基本概念

实体识别的任务是识别出文本中的三大类命名实体（实体类、时间类、数字类）

○ 实体类

- 人名
- 组织/结构
- 地理位置

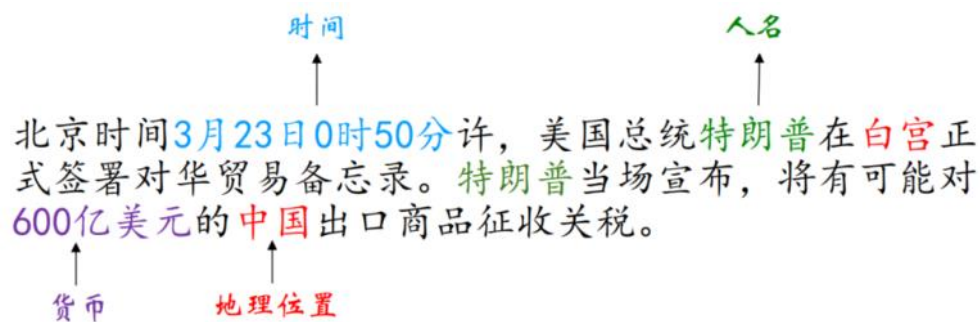
○ 时间类

- 时间
- 日期

○ 数字类

- 货币
- 百分比

北京时间3月23日0时50分许，美国总统特朗普在白宫正式签署对华贸易备忘录。特朗普当场宣布，将有可能对600亿美元的中国出口商品征收关税。



• 二、基于规则和词典的实体识别

基于规则和词典的命名实体识别流程：

○ 预处理

- 划分句子
- 分词+词性标注
- 构建词典

○ 识别实体边界

- 初始化边界：词典匹配、拼写规则、特殊字符、特征词和标点符号等

○ 命名实体分类

- 使用分类规则
- 基于词典的分类

○ 词典主要在三个地方使用

- 在分词时辅助分词
- 实体抽取时根据词典匹配实体
- 基于词典对实体分类

○ 词典的构建

基于统计分析得到候选词典，然后使用人工做筛选，同时人工提取领域中重要的技术和复用领域现有词典。现有的综合中文语义库包括：**CSC、hownet和Chinese open Wordnet**

词典构建统计分析方法：

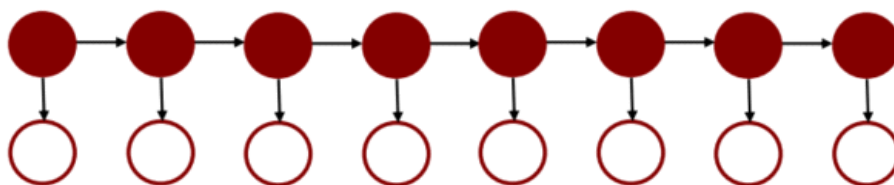
- 去停用词后统计词频，选取一定范围的名词
- 关键词抽取：TF-IDF、TextRank
- 借助维基百科页面的分类系统
- 特征词分词：词共现、特定模式
- 词性分析：从标记为人名（nh）、组织（ni）、日期（nt）等词中抽取
- 依存句法分析

• 三、基于机器学习的命名实体识别

- 基于机器学习的方法主要包括：
 - 隐马尔科夫模型HMM
 - 最大熵马尔科夫模型MEMM
 - 条件随机场CRF
 - 支持向量机SVM

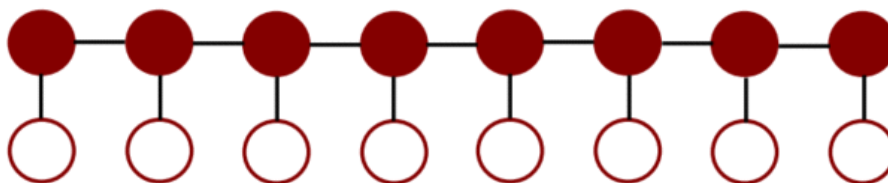
▪ 隐马尔科夫模型

- 有向图模型
- 生成模型
- 特征分布独立假设



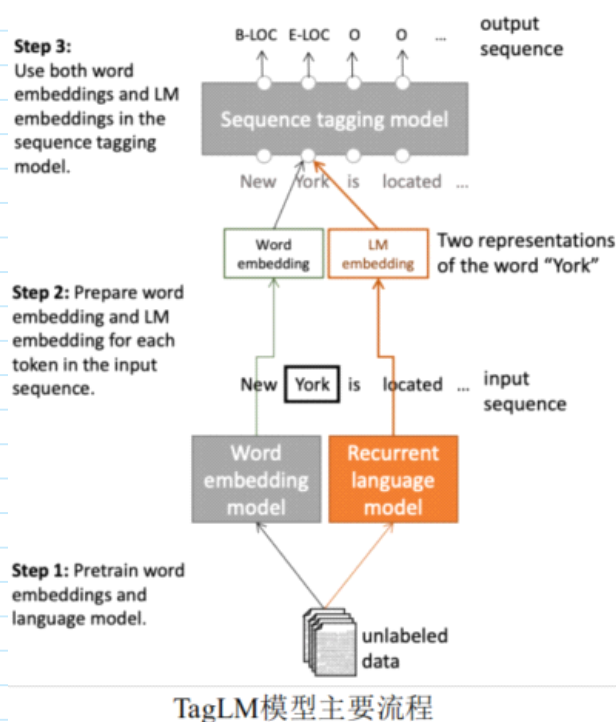
▪ 条件随机场模型

- 无向图模型
- 判别式模型
- 无特征分布独立假设



• 四、基于半监督学习的实体识别

- 使用海量 **无标注** 语料训练Bi-LSTM
- 获取LM embedding和Word embedding
- 将词的向量和语言模型向量混合输入到序列标注模型中进行预测



▪ TagLM模型表现

| Model | $F_1 \pm \text{std}$ |
|-------------------------|------------------------------------|
| Chiu and Nichols (2016) | 90.91 \pm 0.20 |
| Lample et al. (2016) | 90.94 |
| Ma and Hovy (2016) | 91.37 |
| Our baseline without LM | 90.87 \pm 0.13 |
| TagLM | 91.93 \pm 0.19 |

English NER results (CoNLL-2003 test set).

• 五、基于迁移学习的实体识别

迁移学习的核心在于找到新问题和原问题之间的相似性。迁移学习属于机器学习的一个新种类，但是在如下几个方面又有别于传统的机器学习

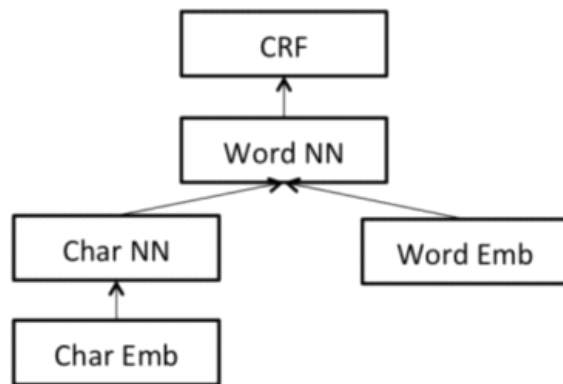
| 比较项目 | 传统机器学习 | 迁移学习 |
|------|----------------|----------------|
| 数据分布 | 训练和测试数据服从相同的分布 | 训练和测试数据服从不同的分布 |
| 数据标注 | 需要足够的数据标注来训练模型 | 不需要足够的数据标注 |
| 模型 | 每个任务分别建模 | 模型可以在不同任务之间迁移 |

迁移学习的三种模式：

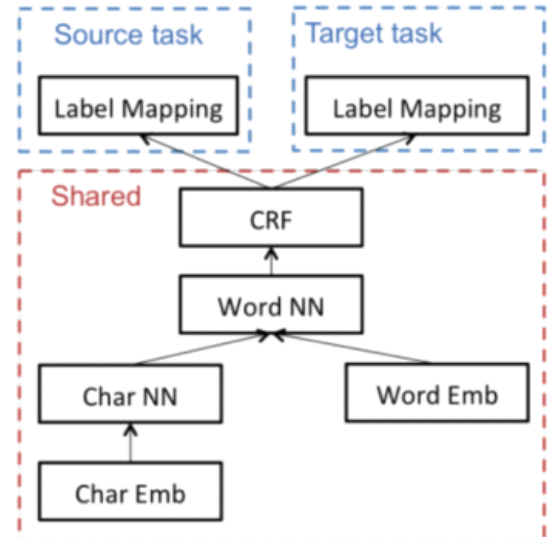
跨领域、跨应用、跨语言

迁移学习的二种模式：

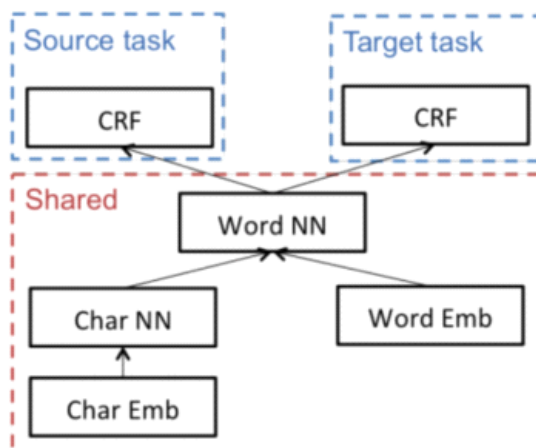
跨领域、跨应用、跨语言



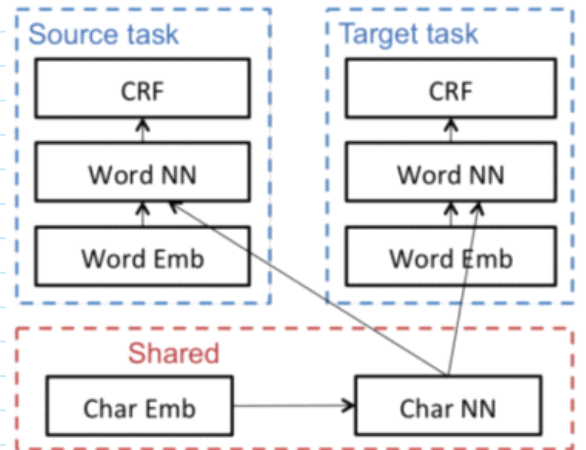
(a) Base model: both of Char NN and Word NN can be implemented as CNNs or RNNs.



(b) Transfer model T-A: used for cross-domain transfer where label mapping is possible.



(c) Transfer model T-B: used for cross-domain transfer with disparate label sets, and cross-application transfer.

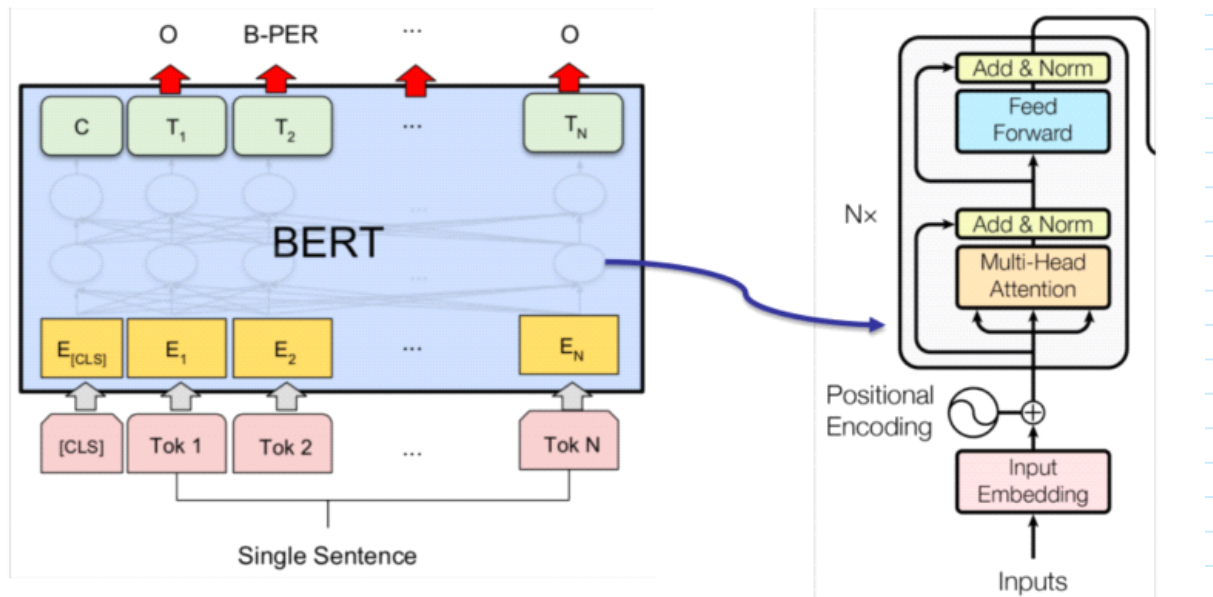


(d) Transfer model T-C: used for cross-lingual transfer.

- a. 基本模型：字符神经网络 单词神经网络可以实现为卷积神经网络 or 循环神经网络
- b. 迁移模型T-A：被使用于标签映射可能时的跨领域转换
- c. 迁移模型T-B：用于具有不同标签集的跨领域转换 和 跨应用转换
- d. 迁移模型T-C：用于跨领域转换

• 六、基于预训练的实体识别

BERT模型



- **BERT模型**重新设计了语言模型预训练阶段的目标任务，提出了**遮挡语言模型** (Masked LM) 和下一个**句子预测** (NSP)
- **Masked LM**是在输入的词序列中随机选15%的词进行[MASK],然后在这15%的词中，有80%的词被真正打上[MASK]标签，10%的词被随机替换成任意词汇，10%的词不做任何处理。相比于传统的语言模型，**Masked LM**可以从前后两个方向预测这些带有[MASK]标签的词。
- **NSP**实质上是一个二分类任务，以50%概率输入一个句子和下一个句子的拼接，标签属于正例；另外50%的概率输入一个句子和非下一个随机句子的拼接，对应标签为负例。