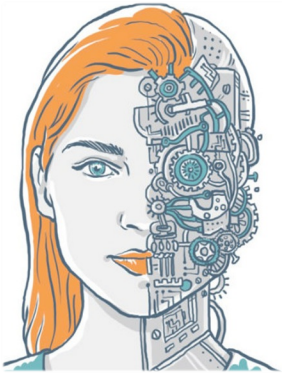
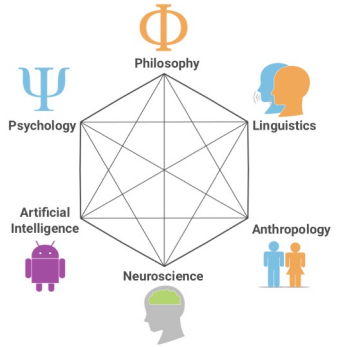
知识抽取：命名实体识别

(Knowledge Extraction: Named Entity Recognition)



Mind or Machines
Cognitive Science Changing
Artificial Intelligence

汪 鹏
pwang@seu.edu.cn
东南大学 KGCODE实验室



东南大学 计算机科学与工程学院/人工智能学院

提纲

- 一、实体识别基本概念
- 二、基于规则和词典的方法
- 三、基于机器学习的方法
- 四、基于深度学习的方法
- 五、基于半监督学习的方法
- 六、基于迁移学习的方法
- 七、基于预训练的方法

语言理解是人工智能领域皇冠上的明珠。

比尔盖茨

如果我有10亿美元，我会建立一个专门研究自然语言处理的项目

Michael Jordan

下一个五年最值得关注的领域是如何理解视频和文字(2014)

Geoff Hinton

深度学习下一个进步点在自然语言理解方面，目标是让机器不仅能理解单个单次，还能理解整个句子和段落(2015)

Yann Lecun

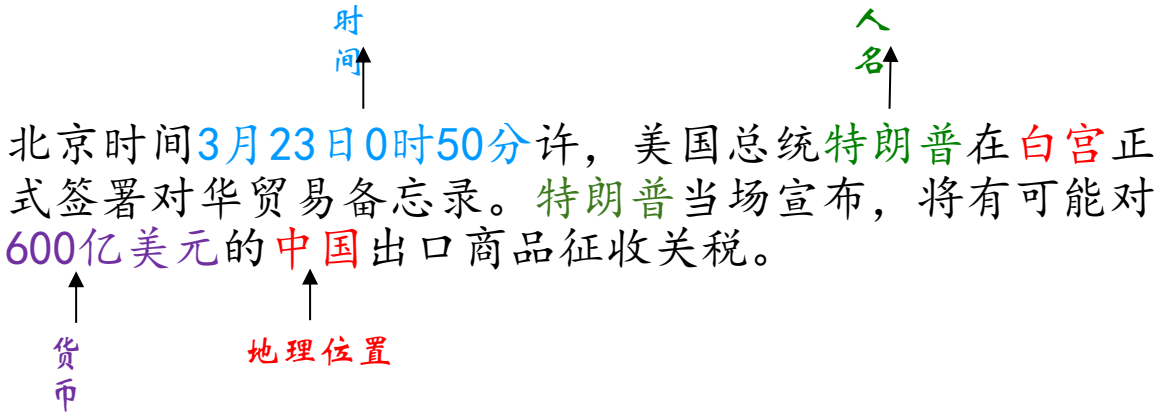
命名实体识别问题

- 实体识别的任务是识别出文本中三大类命名实体（实体类、时间类和数字类），具体如下所示：
  - 人名
  - 组织/机构
  - 地理位置
  - 时间
  - 日期
  - 货币
  - 百分比
  - ……

实体类：人名、组织/机构、地理位置

时间类：时间、日期

数字类：货币、百分比



命名实体识别标注

序列标注体系：

Token	IO	BIO	BIOES	BMEWO
特	I-PER	B-PER	B-PER	B-PER
朗	I-PER	I-PER	I-PER	M-PER
普	I-PER	I-PER	E-PER	E-PER
在	O	O	O	O
白	I-LOC	B-LOC	B-LOC	B-LOC
宫	I-LOC	I-LOC	E-LOC	E-LOC
签	O	O	O	O
署	O	O	O	O

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

基于规则和词典的实体识别

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

基于规则和词典的实体识别

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术语和复用领域现有词典。现有的综合中文语义词库包括：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

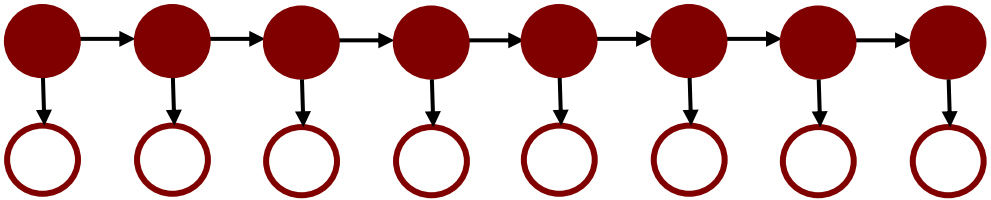
基于机器学习的实体识别

基于机器学习的方法主要包括：

- 隐马尔科夫模型（Hidden Markov Model, HMM）
- 最大熵马尔科夫模型（Maximum Entropy Markov Model, MEMM）
- 条件随机场（Conditional Random Fields, CRF）
- 支持向量机（Support Vector Machine, SVM）

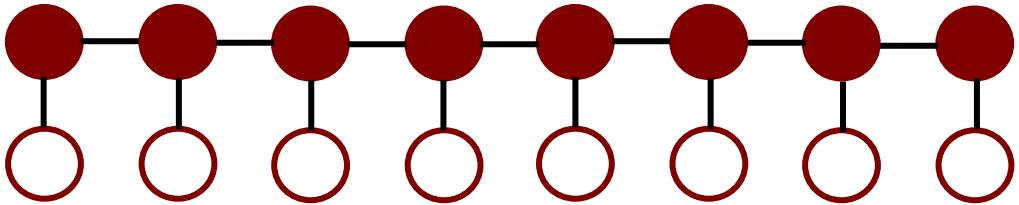
基于机器学习的实体识别

- 隐马尔可夫模型
  - 有向图模型
  - 生成模型
  - 特征分布独立假设



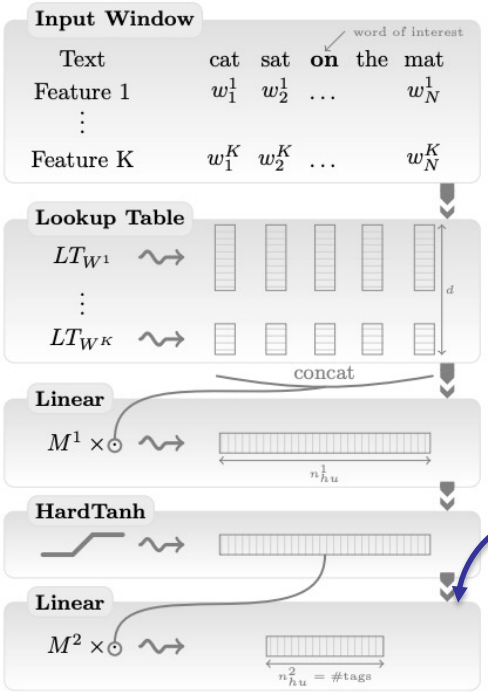
基于机器学习的实体识别

- 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



基于深度学习的实体识别

NN/CNN + CRF

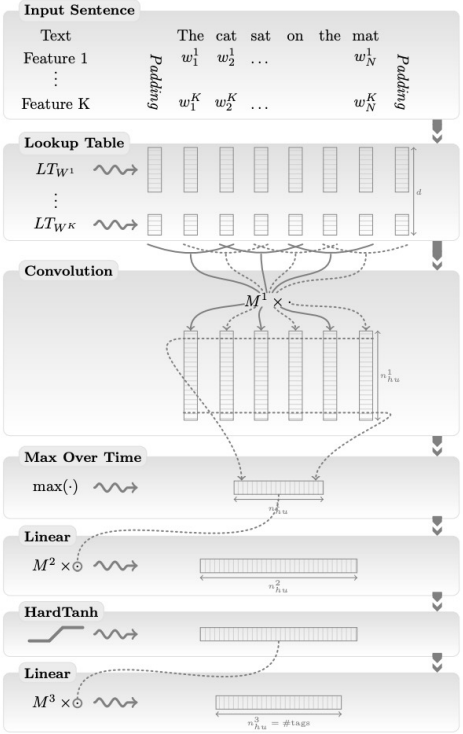


(a) window approach

词级别的对数似然：
softmax

句子级别的对数似然：
CRF

Collobert et al.[2011]



(b) sentence approach

基于深度学习的实体识别

NN/CNN + CRF模型表现

System	F1
Ando and Zhang (2005)	89.31%
Florian et al. (2003)	88.76%
Kudo and Matsumoto (2001)	88.31%

English NER Benchmark model (CoNLL-2003 test set).

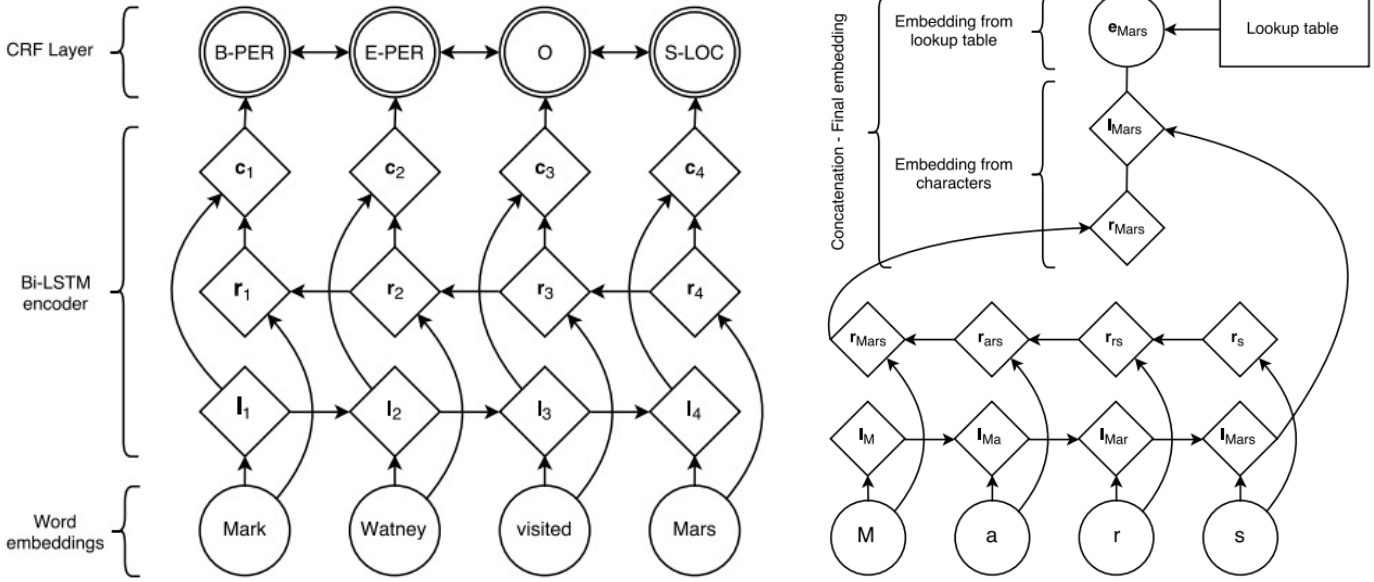
Task		**Benchmark**	**SENNA**
Part of Speech (POS)	(Accuracy)	97.24 %	97.29 %
Chunking (CHUNK)	(F1)	94.29 %	94.32 %
Named Entity Recognition (NER)	(F1)	89.31 %	89.59 %
Parse Tree level 0 (PT0)	(F1)	91.94 %	92.25 %
Semantic Role Labeling (SRL)	(F1)	77.92 %	75.49 %

English NER results (CoNLL-2003 test set).

Collobert et al.[2011]

基于深度学习的实体识别

Bi-LSTM+CRF



(a) word-level (b) character-level

Lample et al.[2016]

基于深度学习的实体识别

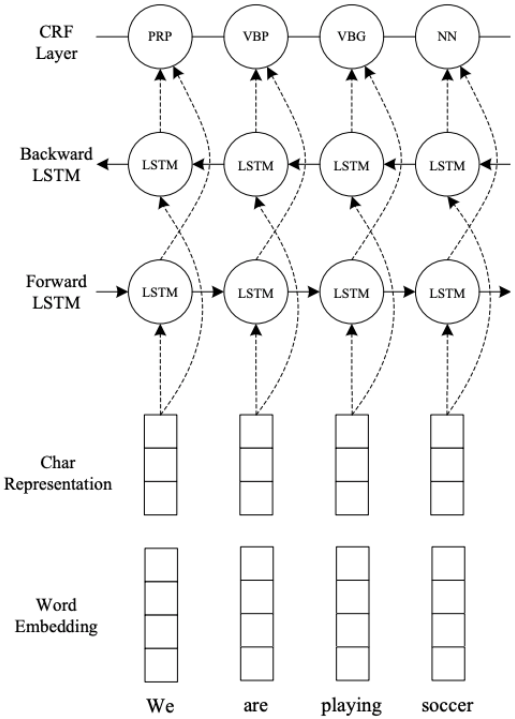
Bi-LSTM+CRF模型表现

Model	F_1
Collobert et al. (2011)*	89.59
Lin and Wu (2009)	83.78
Lin and Wu (2009)*	90.90
Huang et al. (2015)*	90.10
Passos et al. (2014)	90.05
Passos et al. (2014)*	90.90
Luo et al. (2015)* + gaz	89.9
Luo et al. (2015)* + gaz + linking	**91.2**
Chiu and Nichols (2015)	90.69
Chiu and Nichols (2015)*	90.77
LSTM-CRF (no char)	90.20
LSTM-CRF	**90.94**

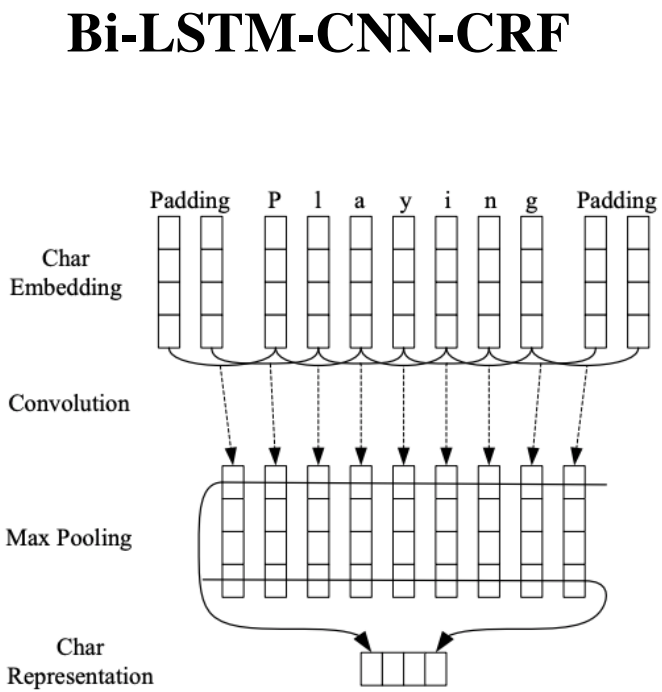
English NER results (CoNLL-2003 test set).

Lample et al.[2016]

基于深度学习的实体识别



(a) Bi-LSTM-CNN-CRF



(b) CNN获得字符级别表示 Ma and Hovy.[2016]

基于深度学习的实体识别

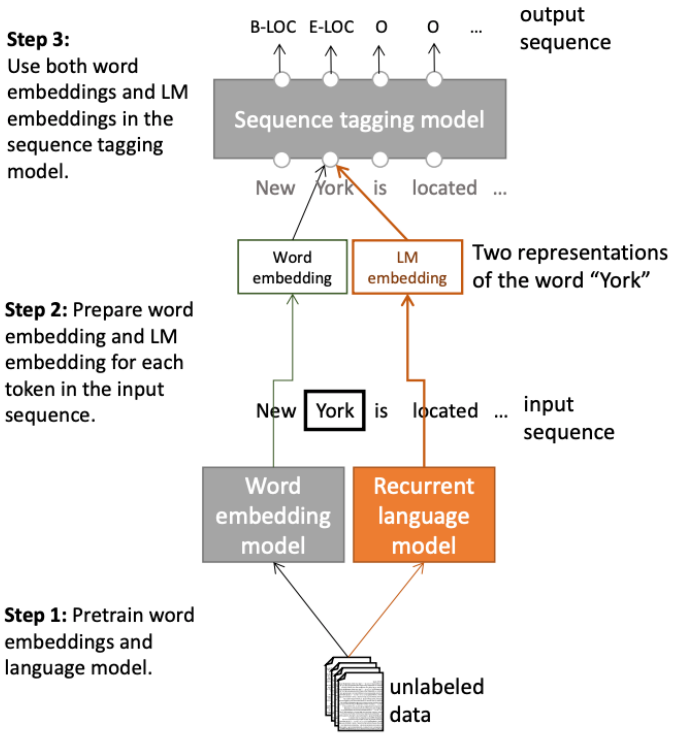
Bi-LSTM-CNN-CRF 模型表现

Model	F1
Chieu and Ng (2002)	88.31
Florian et al. (2003)	88.76
Ando and Zhang (2005)	89.31
Collobert et al. (2011)‡	89.59
Huang et al. (2015)‡	90.10
Chiu and Nichols (2015)‡	90.77
Ratinov and Roth (2009)	90.80
Lin and Wu (2009)	90.90
Passos et al. (2014)	90.90
Lample et al. (2016)‡	90.94
Luo et al. (2015)	91.20
This paper	**91.21**

English NER results (CoNLL-2003 test set).

Ma and Hovy.[2016]

基于半监督学习的实体识别



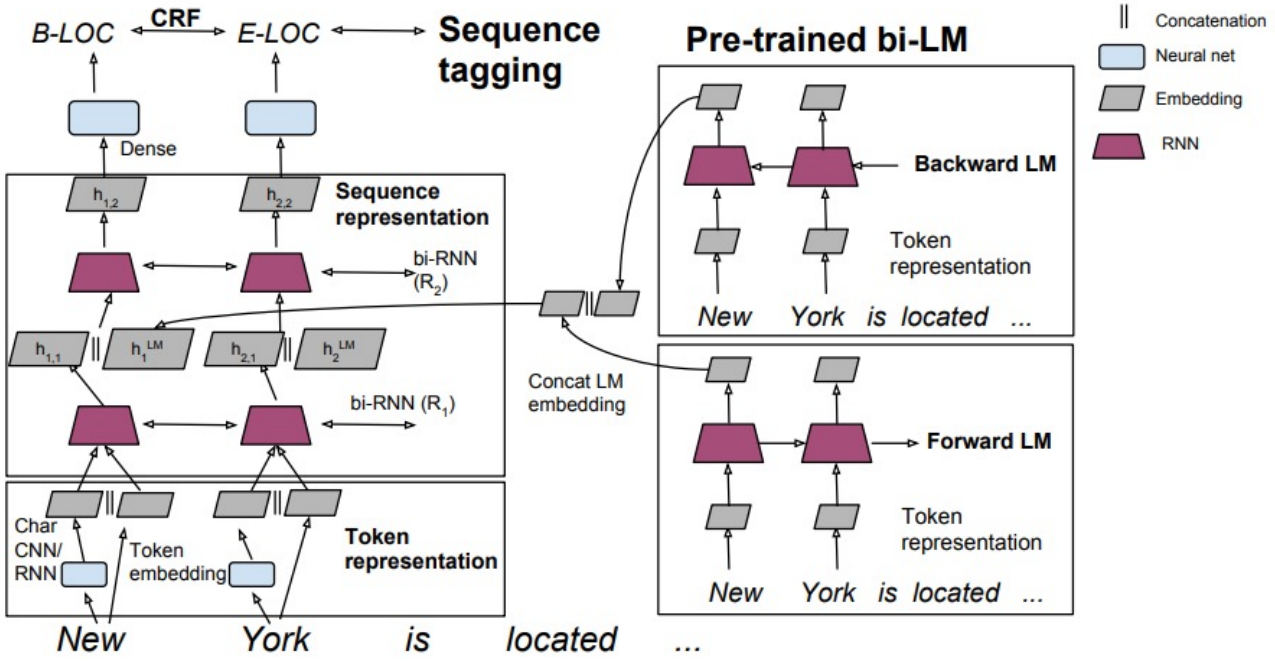
TagLM模型主要流程

Language Model Augmented Sequence Taggers(TagLM)

a. 使用海量无标注语料训练Bi-LSTM
b. 获取LM embedding和Word embedding
c. 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

Peters et al.[2017]

基于半监督学习的实体识别



TagLM模型结构

Peters et al.[2017]

基于半监督学习的实体识别

TagLM 模型表现

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

English NER results (CoNLL-2003 test set).

Peters et al.[2017]

基于迁移学习的实体识别

在教育和心里学上，迁移学习是基于人类已有的经验来研究人类的行为、学习或表现。探讨人类如何从一个环境中迁移到具有相似特性的另一个环境中。任何一种学习都要受到学习者已有知识经验、技能和态度的影响。只要有学习，就有迁移。

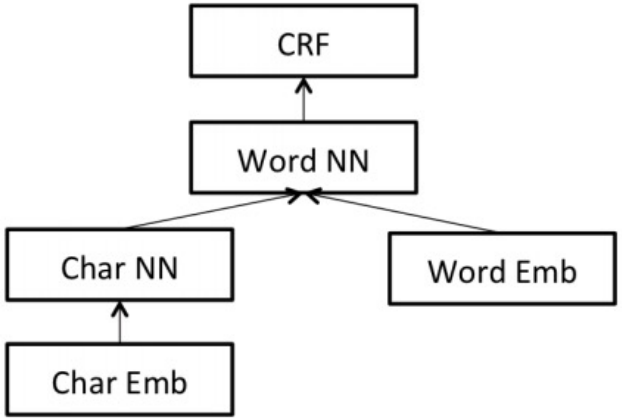
迁移学习的核心在于找到新问题和原问题之间的相似性。迁移学习属于机器学习的一个种类，但在如下几个方面又有别于传统的机器学习。

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

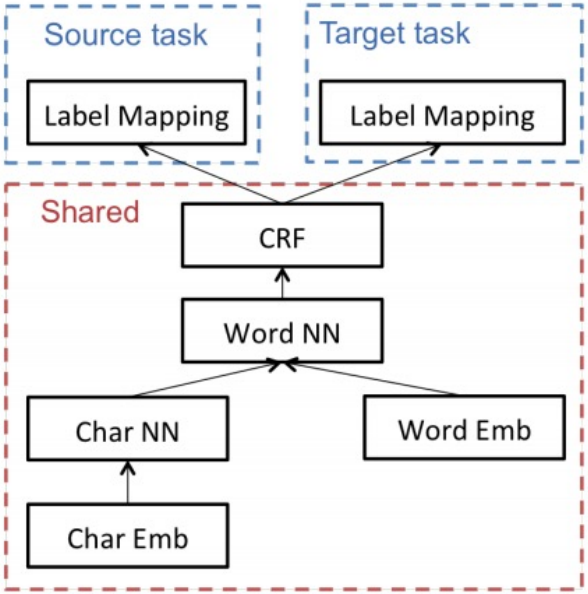
基于迁移学习的实体识别

迁移学习的三种模式：

跨域、跨应用、跨语言



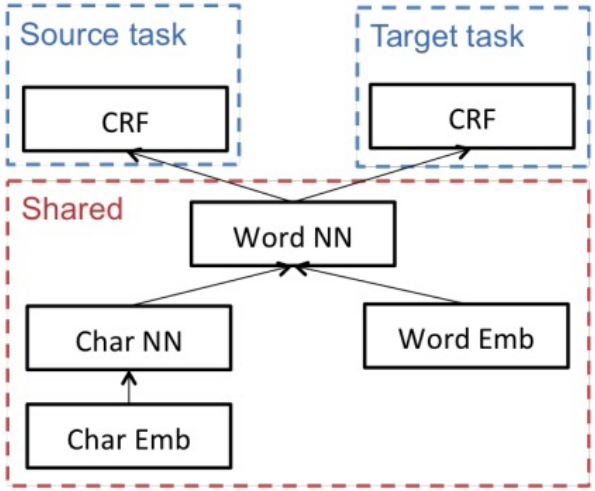
(a) Base model: both of Char NN and Word NN can be implemented as CNNs or RNNs.



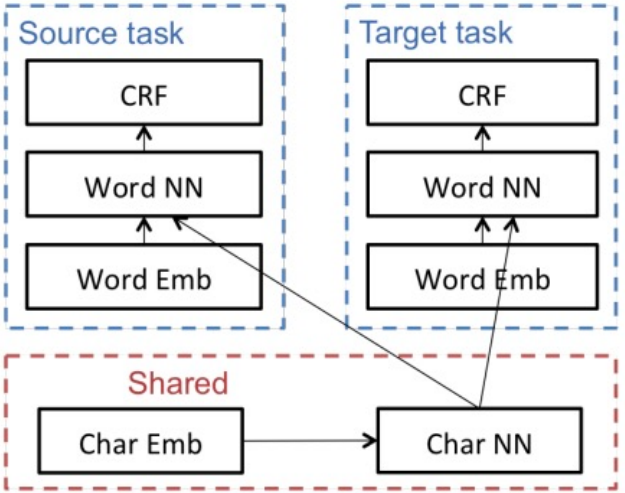
(b) Transfer model T-A: used for cross-domain transfer where label mapping is possible.

Yang et al.[2017]

基于迁移学习的实体识别



(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.

Yang et al.[2017]

基于迁移学习的实体识别

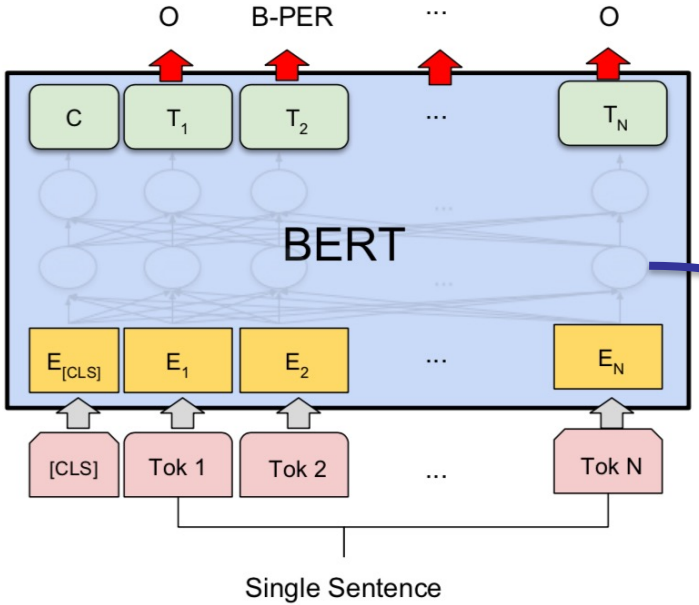
迁移学习模型表现

Source	Target	Model	Setting	Transfer	No Transfer	Delta
PTB	Twitter/0.1	T-A	dom	83.65	74.80	8.85
CoNLL03	Twitter/0.1	T-A	dom	43.24	34.65	8.59
PTB	CoNLL03/0.01	T-B	app	74.92	68.64	6.28
PTB	CoNLL00/0.01	T-B	app	86.73	83.49	3.24
CoNLL03	PTB/0.001	T-B	app	87.47	84.16	3.31
Spanish	CoNLL03/0.01	T-C	ling	72.61	68.64	3.97
CoNLL03	Spanish/0.01	T-C	ling	60.43	59.84	0.59
PTB	Genia/0.001	T-A	dom	92.62	83.26	9.36
CoNLL03	Genia/0.001	T-B	dom&app	87.47	83.26	4.21
Spanish	Genia/0.001	T-C	dom&app&ling	84.39	83.26	1.13
PTB	Genia/0.001	T-B	dom	89.77	83.26	6.51
PTB	Genia/0.001	T-C	dom	84.65	83.26	1.39

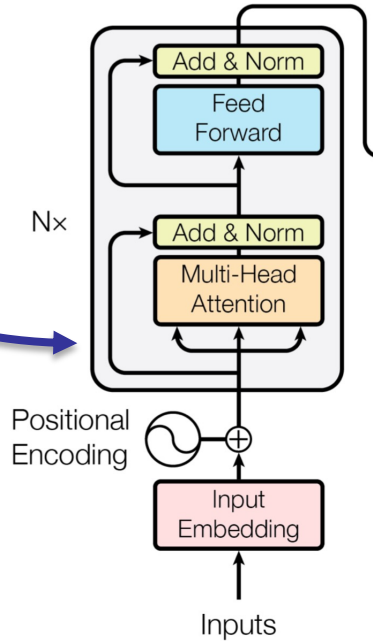
Yang et al.[2017]

基于预训练的实体识别

BERT模型



Transformer--Encoder



Devlin et al.[2018]

基于预训练的实体识别

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，有10%的词被随机替换成任意词汇，10%的词不做任何处理。模型的任务是去正确预测带有[MASK]标签的词。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%的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。

基于预训练的实体识别

BERT模型表现

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

English NER results (CoNLL-2003).

Devlin et al.[2018]

实体识别的应用示例

PER a ORG s LOC d DATE f

郑帅帅×，男。因涉嫌犯运输毒品罪，于2016年6月6日×被昆明市公安局官渡分局×刑事拘留，2016年7月11日×被执行逮捕。现羁押于昆明市官渡区看守所×。辩护人胡纯蛟×、姚芳×，云南颐高律师事务所律师。昆明市官渡区人民检察院×以官检公一科刑诉1038号起诉书指控被告人郑帅帅×犯运输毒品罪，于2016年10月21日×向本院提起公诉，并以官检公一科量建820号量刑建议书提出"建议判处被告人郑帅帅×十三年以上十五年以下有期徒刑，并处罚金"。

PER a ORG s LOC d DATE f

本院受理后，适用普通程序依法组成合议庭，于2016年12月22日×、2017年1月11日×在本院法庭公开开庭审理了本案。昆明市官渡区人民检察院×指派检察员袁松×、代理检察员詹晶×、书记员何秋虹×出席法庭支持公诉，被告人郑帅帅×及其辩护人胡纯蛟×到庭参加诉讼，现已审理终结。经审理查明：2016年6月6日11时许×，被告人郑帅帅×以体内藏匿的方式携带毒品可疑物欲乘飞机前往郑州×，后在云南省景洪市西双版纳国际机场×安检时，被民警查获。从其体内排出毒品可疑物净重142克。经鉴定，排出的毒品可疑物检出海洛因成分。

东大理解实体识别优化测试结果										
实体类型			优化前(KnowledgeGraph-V1.0)				KnowledgeGraph-V1.1优化结果（红色：下降、绿色：提升）			
			测试实体数量	识别成功数量	实体识别正确率	综合	测试实体数量	识别成功数量	实体识别正确率	综合
姓名NAME	两字姓名_常见姓氏		2136	1949	91.25%	90.59%	2128	1973	92.72%	94.32%
	两字姓名_罕见姓氏		618	528	85.44%		585	539	92.14%	
	三字姓名_常见姓氏		2129	2055	96.52%		2122	2054	96.80%	
	三字姓名_罕见姓氏		571	530	92.82%		604	579	95.86%	
	复姓姓名		606	428	70.63%		621	571	91.95%	
机构COMPANY	机构全称（机构后缀有一定规则的）		4077	3923	96.22%	88.24%	4029	3653	90.67%	81.01%
	机构简称		983	542	55.14%		1031	446	43.26%	
地址ADDR	nplace籍贯地址		1484	1196	80.59%	68.42%	1434	1393	97.14%	80.23%
	hplace（用户户口地址包）包含固定格式内容的地址		1408	945	67.12%		1435	1043	72.68%	
	无固定格式内容的其他地址		1868	1116	59.74%		1891	1383	73.14%	
账号ACCOUNT	强规则账号	Tel	1440	1440	100.00%		1438	1438	100.00%	
		email	1400	1400	100.00%		1397	1397	100.00%	
		idcard	1280	1280	100.00%		1276	1276	100.00%	
		wxid开头的微信	599	599	100.00%		578	578	100.00%	
		车牌号	/				11	10	90.91%	
	非强规则账号（需指明账号类型）	qq	1480	1479	99.93%		在1.0版本的基础上新增了：QQ群、微信群、微博、京东、淘宝、支付宝、抖音共7种非强规则账号识别，测试情况见下表			
		非wxid开头的微信	601	601	100.00%					


धन्यवाद
Hindi
Hindi
多謝
Traditional Chinese
ขอบคุณ
Thai
Спасибо
Russian
Thank You
English
Gracias
Spanish
شكرا
Arabic
Grazie
Italian
Danke
German
Obrigado
Brazilian Portuguese
多谢
Simplified Chinese
Merci
French
நன்றி
Tamil
Tamil
ありがとうございました
Japanese
감사합니다
Korean