命名实体识别（NER）

杜一凡

什么是NER

- 命名实体识别（Name Entity Recognition,NER）,也称作“专名识别”，是指识别文本中具有特定意义的实体，包括人名、地名、机构名、专有名词等。
- 实体与命名实体的区别：

实体是一个抽象的对象，并不指定描述它的方式。命名实体一般指的是受“Named”限制的对象。

- 命名实体识别(Name Entity Recognition, NER)与实体指代识别(Entity mention Recognition, EMR)的区别：

例子：Sue and her brother studied in University of Toronto.

NER的作用

- NER往往作为natural language applications的基础，比如QA(question answering),text summarization, machine translation等.
- NER是一种序列标注任务,常见的序列标注任务还有分词，词性标注(POS)，关键词抽取，词义角色标注等

NER的关键

- 1. 实体边界的确定(定位)
- 2. 实体类别的判断

NER的评价指标

- 常用的：Acc、Precision、Recall和F1值。
- 绝大多数的NER任务需要识别多种实体类别，需要对所有的实体类别评估NER的效果。基于这个思路，有两类评估指标：

1. 宏平均F-score：分别对每种实体类别分别计算对应类别的F-score，再求整体平均
2. 微平均F-score：对整体数据求F-score(每个实体视为平等)

$$Acc=\frac{TP+TN}{TP+TN+FN+FP} \qquad P=\frac{TP}{TP+FP} \qquad R=\frac{TP}{TP+FN} \qquad F1=\frac{2PR}{P+R}$$

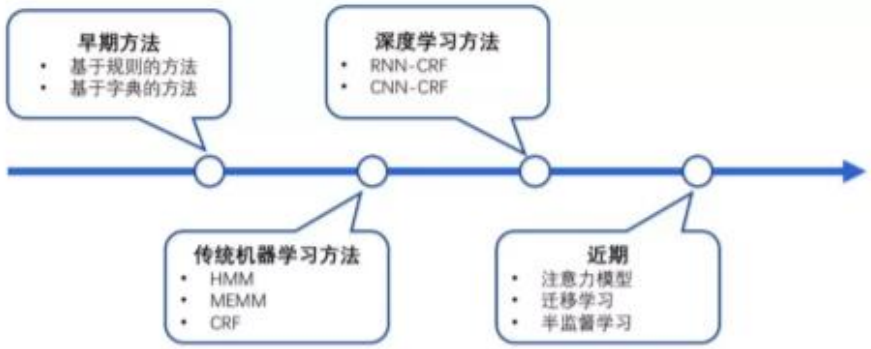


图 2: NER 发展趋势

自1996年NRE提出后的发展趋势

TABLE 1
List of annotated datasets for English NER. Number of tags refer to the number of entity types.

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal texts	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC96T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts, news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine, news, conversation, web	89	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	18	http://noisy-text.github.io
BBN	2005	Wall Street Journal texts	64	https://catalog.ldc.upenn.edu/ldc2005t33
NYT	2008	New York Times texts	5	https://catalog.ldc.upenn.edu/LDC2008T19
WikiGold	2009	Wikipedia	4	https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	113	https://github.com/xiaoling/figer
N^3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
GENIA	2004	Biology and clinical texts	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	790	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

NER的数据集

常用的数据集是CoNLLo3和OneNotes,分别表示粗粒度任务和细粒度任务

NER数据标注体系

- 主要的数据标注体系是:IO,BIO,BMEWO和BMEWO+

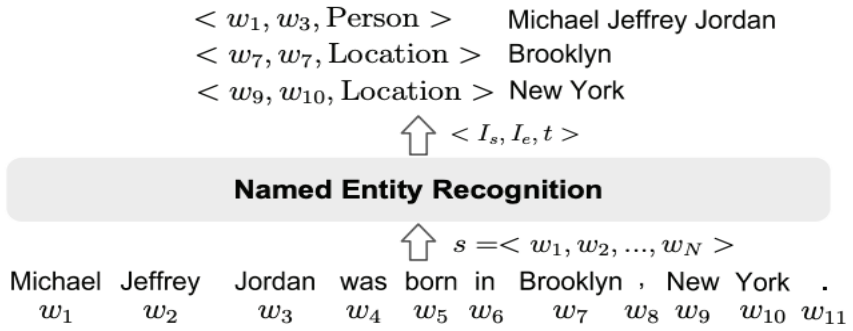
Tokens	IO	BIO	BMEWO	BMEWO+
Yesterday	O	O	O	BOS_O
afternoon	0	O	O	O
,	0	O	O	O_PER
John	I_PER	B_PER	B_PER	B_PER
J	I_PER	I_PER	M_PER	M_PER
.	I_PER	I_PER	M_PER	M_PER
Smith	I_PER	I_PER	E_PER	E_PER
traveled	O	0	O	PER_O
to	O	O	O	O_LOC
Washington	I_LOC	B_LOC	W_LOC	W_LOC
.	O	O	O	O_EOS

NER的工具

TABLE 2
Off-the-shelf NER tools offered by academia and industry/opensource projects.

NER System	URL
StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP/
OSU Twitter NLP	https://github.com/aritter/twitter_nlp
Illinois NLP	http://cogcomp.org/page/software/
NeuroNER	http://neuroner.com/
NERsuite	http://nersuite.nlplab.org/
Polyglot	https://polyglot.readthedocs.io
Gimli	http://bioinformatics.ua.pt/gimli
spaCy	https://spacy.io/
NLTK	https://www.nltk.org
OpenNLP	https://opennlp.apache.org/
LingPipe	http://alias-i.com/lingpipe-3.9.3/
AllenNLP	https://allennlp.org/models
IBM Watson	https://www.ibm.com/watson/

NER任务的示意图



NER可以视为一种多分类的任务

NER算法分类

- 基于学习方式的分类：

1.监督学习方法：隐马尔可夫模型、语言模型、最大熵模型、支持向量机(分类)、决策树和条件随机场等

2.半监督学习方法：这一类方法利用标注的小数据集（种子数据）自举学习(bootstrap)。

3.无监督学习方法：这一类方法利用词汇资源（如WordNet）等进行上下文聚类。

4.混合方法：几种模型相结合或利用统计方法和人工总结的知识库。

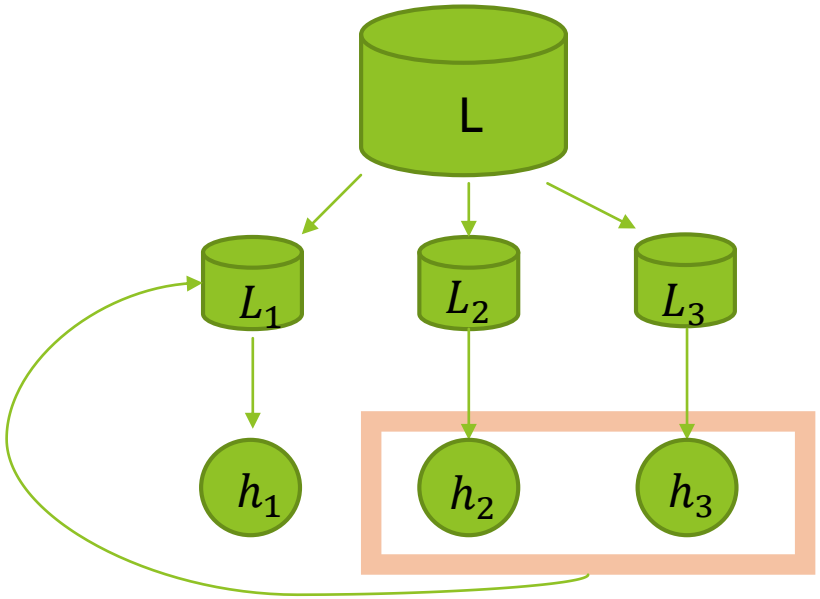
- 基于方式的分类：

1.基于规则和词典的方法

2.基于统计的方法

3.基于深度学习的方法

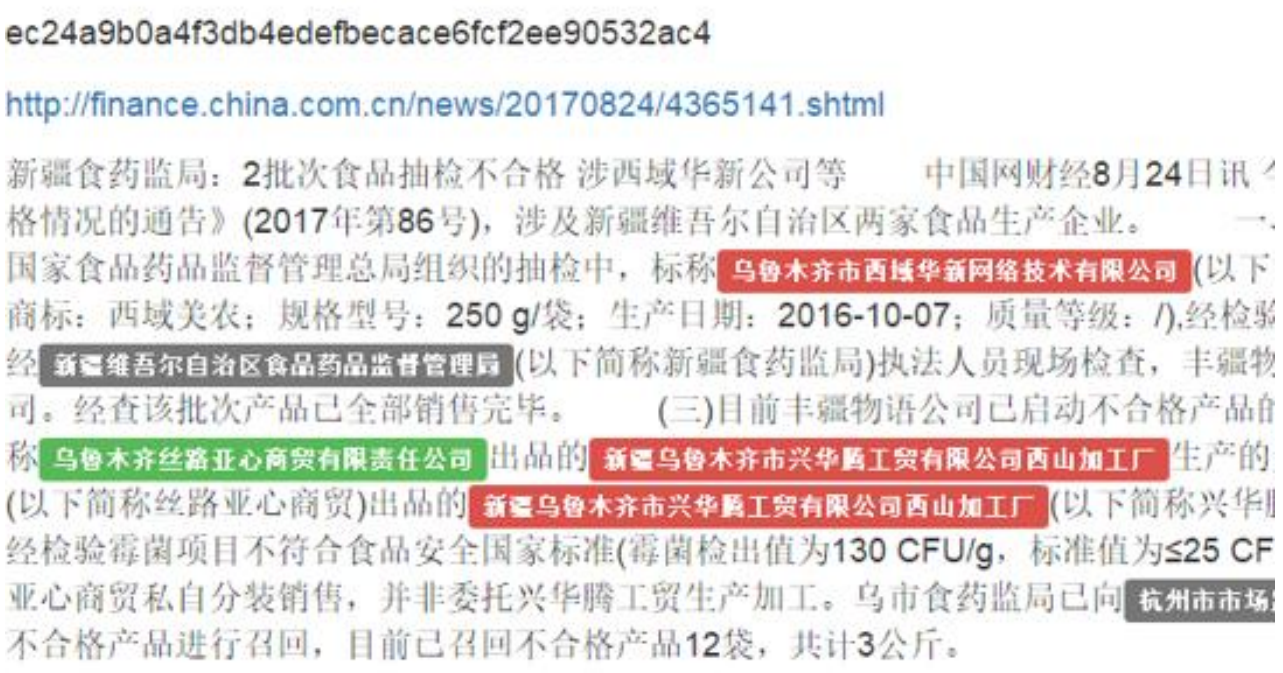
半监督学习：bootstrap



Psudo-label

NER算法：基于规则

- 基于规则的方法往往需要语言学专家手工构造规则模板以及特征词。


ec24a9b0a4f3db4edefbecace6fcf2ee90532ac4

http://finance.china.com.cn/news/20170824/4365141.shtml

新疆食药监局：2批次食品抽检不合格 涉西域华新公司等　　中国网财经8月24日讯
格情况的通告》(2017年第86号)，涉及新疆维吾尔自治区两家食品生产企业。　　一、
国家食品药品监督管理总局组织的抽检中，标称乌鲁木齐市西域华新网络技术有限公司(以下
商标：西域美农；规格型号：250 g/袋；生产日期：2016-10-07；质量等级：/),经检验
经新疆维吾尔自治区食品药品监督管理局(以下简称新疆食药监局)执法人员现场检查，丰疆物
司。经查该批次产品已全部销售完毕。　　(三)目前丰疆物语公司已启动不合格产品的
称乌鲁木齐丝路亚心商贸有限责任公司出品的新疆乌鲁木齐市兴华腾工贸有限公司西山加工厂生产的
(以下简称丝路亚心商贸)出品的新疆乌鲁木齐市兴华腾工贸有限公司西山加工厂(以下简称兴华
经检验霉菌项目不符合食品安全国家标准(霉菌检出值为130 CFU/g，标准值为≤25 CF
亚心商贸私自分装销售，并非委托兴华腾工贸生产加工。乌市食药监局已向杭州市市场
不合格产品进行召回，目前已召回不合格产品12袋，共计3公斤。

模板：若干地名+若干其他成分+若干特征词
特征词：公司、有限公司、管理局等

NER算法：基于统计的方法
最大熵

- 基本思想：利用给定训练数据选择适当统计模型，使其满足所有已知事实(约束),而对未知事实不做任何假设(熵最大)——最优化问题。
- 最大熵与其他模型的不同之处：一般说的“特征”都是指输入的特征,而最大熵模型中的“特征”指的是输入和输出共同的特征(共现时对应的特征)——特征函数。
- $$\max H(P) = \sum_{x,y} \bar{P}(x)P(y|x)logP(y|x)$$

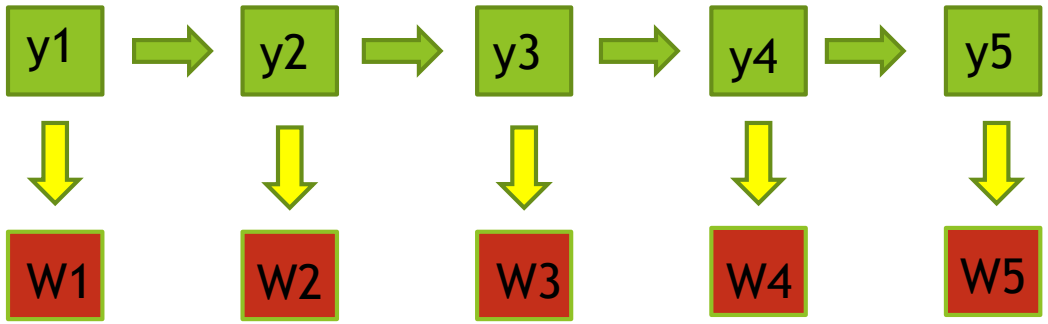
约束:

$$\mathrm{E}_{\bar{P}}(f_i) = E_P(f_i)(i = 1,2,3,..M)$$

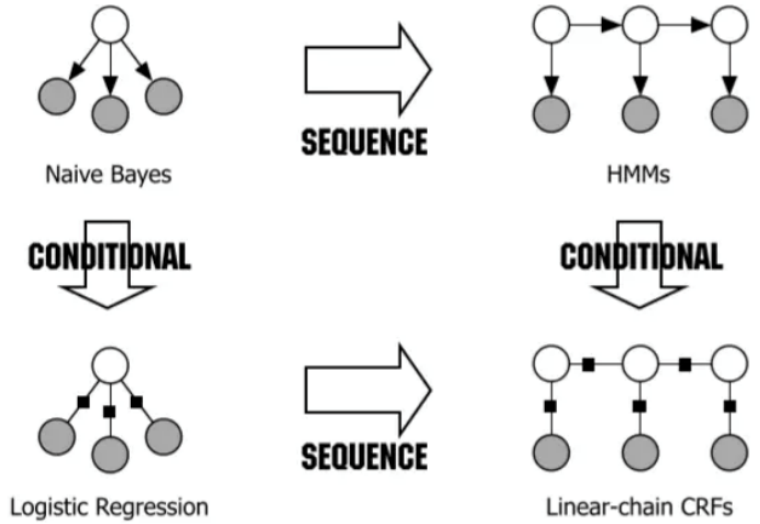
$$\sum_y P(y|x) = 1$$

NER算法：基于统计的方法
HMM（隐马尔可夫模型）

- 目标：求观察序列背后可能的标注序列
- 假设：马尔科夫假设(每个状态只依赖于前一个状态)

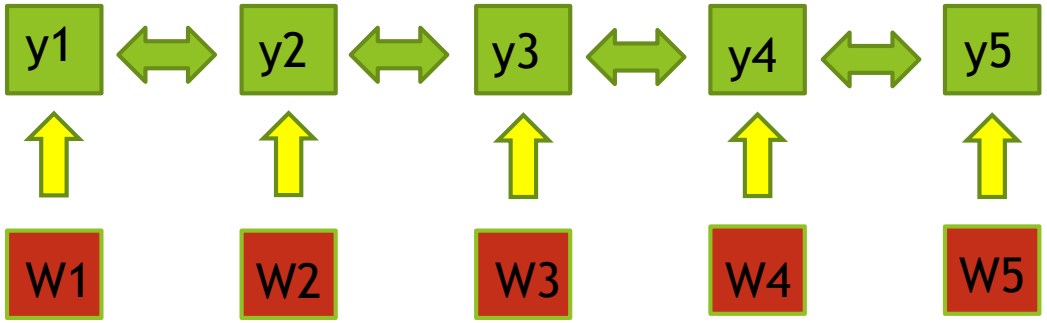


NER算法：基于统计的方法
CRF(条件随机场)

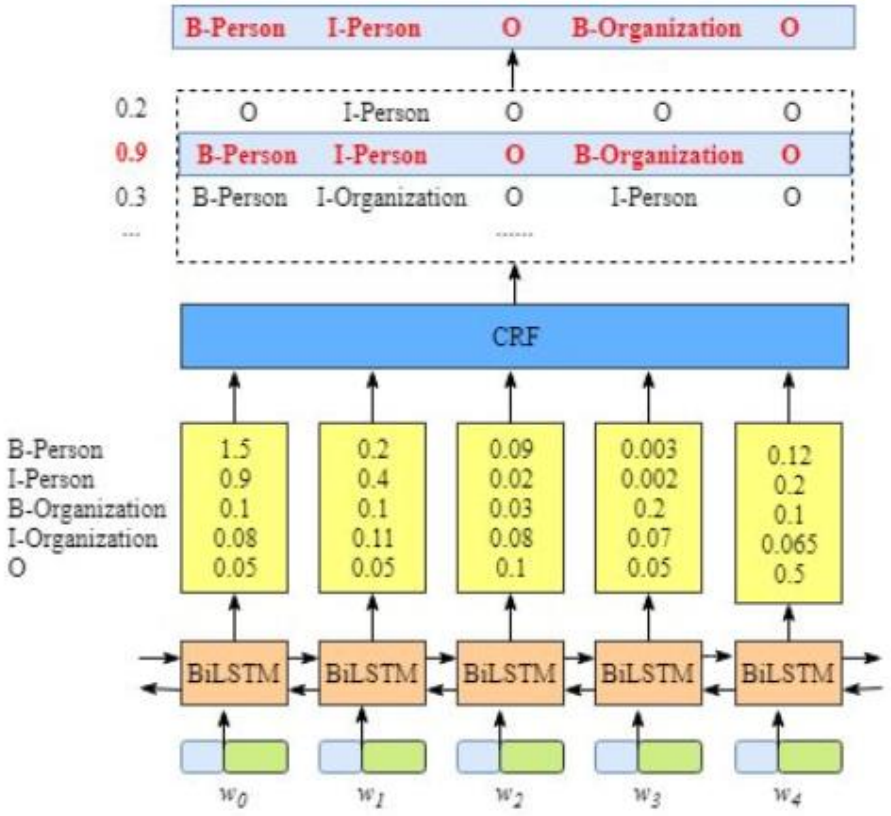


特征函数：$f_j(s, i, l_i, l_{i-1})$
从特征函数到概率：
$\text{Score}(l|s)=\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s, i, l_i, l_{i-1})$

$P(l|s)=\frac{\exp(score(l|s))}{\sum_{l'}\exp(score(l'|s))}$



NER算法：基于深度学习的方法 LSTM+CRF

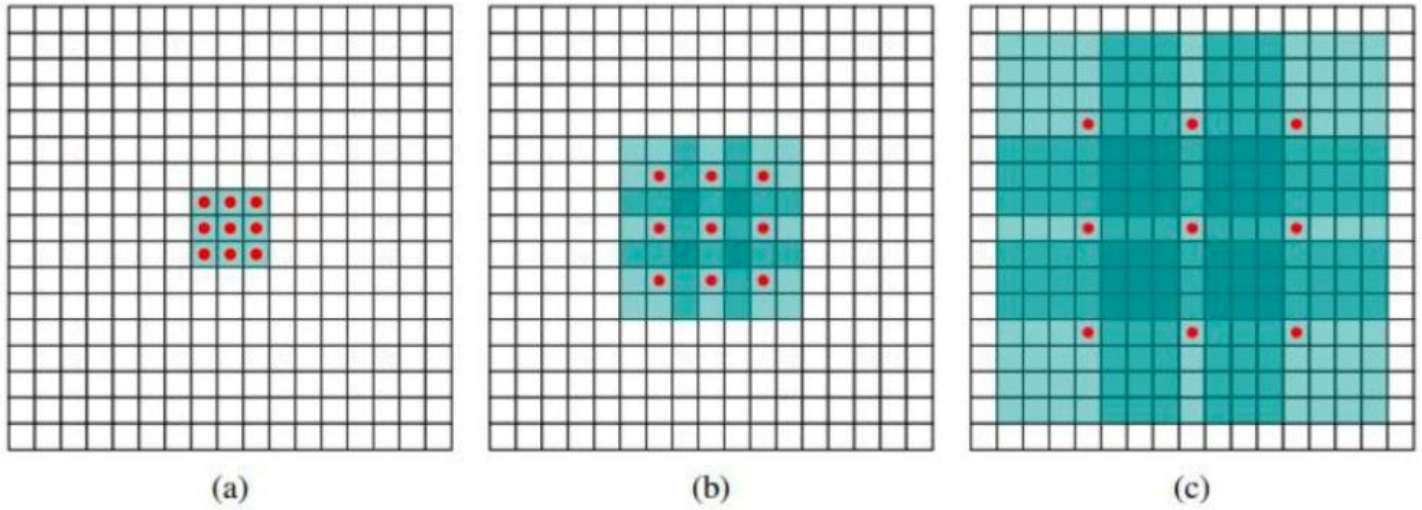


- 传统的CRF的优点就是能对隐含状态建模，学习状态序列的特点，但它的缺点是需要手动提取序列特征.所以一般做法是，在LSTM后加一层CRF，以获得两者的优点
- **如何理解LSTM后接CRF**

简单说就是条件随机场可以把label的上下文学出来。lstm加softmax分类的时候只能把特征的上下文关系学出来，label的没学出来。

NER算法：基于深度学习的方法
IDCNN-CRF

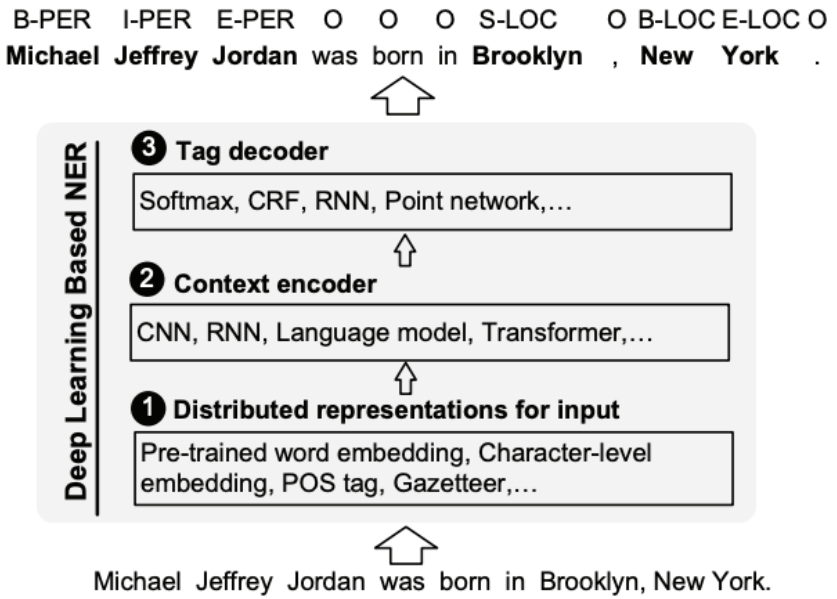
- **关键：对NER来讲，整个输入句子中每个字都有可能对当前位置的标注产生影响，即所谓的长距离依赖问题。**
- 2015 提出了dilated CNN模型，意思是“膨胀的”CNN。



NER算法：基于深度学习的方法创新思路

- 基于深度学习的NER模型的模块：

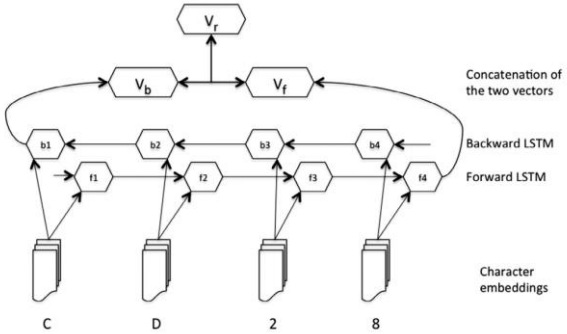
1. Distributed representations for input
2. Context encoder
3. Tag decoder



参考文献： Li J , Sun A , Han J , et al. A Survey on Deep Learning for Named Entity Recognition[J]. 2018.

NER算法：基于深度学习的方法
Distributed representations for input方面的创新

- 1. Gridach, Mourad. Character-Level Neural Network for Biomedical Named Entity Recognition[J]. Journal of Biomedical Informatics, 2017:S1532046417300977.



- 2. Kuru, O., Can, O.A., Yuret, D.: Charner: Character-level named entity recognition. In: Proceedings of The 26th International Conference on ComputationalLinguistics. pp. 911–921 (2016)

John	works	for	Globex	Corp.	.
B-PER	O	O	B-ORG	I-ORG	O

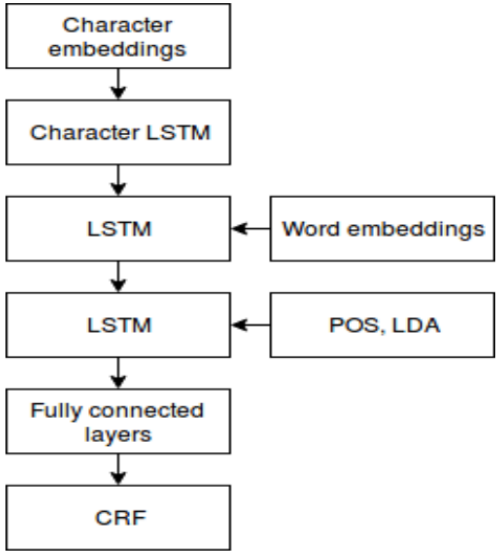
J	o	h	n		w	o	r	k	s		f	o	r		G	l	o	b	e	x		C	o	r	p	.	.
P	P	P	P	O	O	O	O	O	O	O	O	O	O	O	G	G	G	G	G	G	G	G	G	G	G	O	O

Figure 1: An example sentence with word level and character level NER tags.

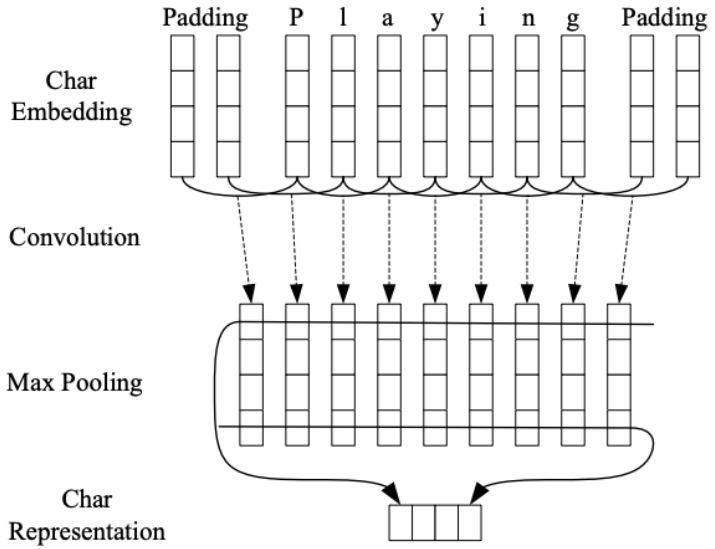
	PER	ORG	O	PER	ORG	O	PER	ORG	O
PER	1	0	0	1	0	1	1	0	0
ORG	0	1	0	0	1	1	0	1	0
O	0	0	1	0	0	1	1	1	1
	$c \rightarrow c$			$c \rightarrow s$			$s \rightarrow c$		

NER算法：基于深度学习的方法
Distributed representations for input方面的创新

- 3. Distributed Representation, LDA Topic Modelling and Deep Learningfor Emerging Named Entity Recognition from Social Media



4. Ma X , Hovy E . End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.



NER算法：基于深度学习的方法
Context encoder方面的创新

- 1. Tran Q , Mackinlay A , Yepes A J . Named Entity Recognition with stack residual LSTM and trainable bias decoding[J]. 2017.

a. 为了增加网络的表达能力采用layer stacking,为了避免网络layer stacking过程中的degradation 问题，引入residual connection.

b. 直接在评估指标F-measure上训练。但问题是难以训练。作者采用了一种混合的解决方案，先用传统的log-likelihood训练，然后用一个更简单的自适应模型来使模型的输出更适合F-measure.方法是在解码器中加入一个可训练的噪声。

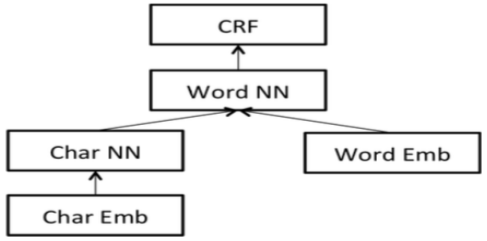
- 2. Named Entity Recognition With Parallel Recurrent Neural Networks

a. 将单个LSTM分割为多个维度更小的的LSTM

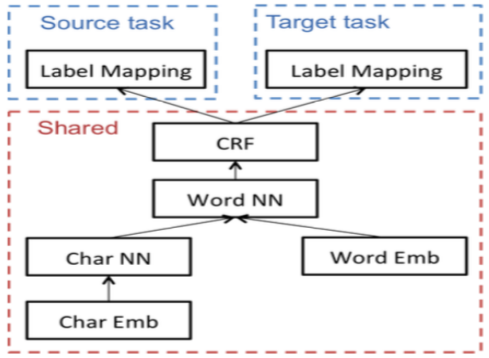
b. 为了提升多样性：引入了一个正则化项(约束)

$$\Phi = \begin{pmatrix} \mathrm{vec}(W_c^{(1)}) \\ \mathrm{vec}(W_c^{(2)}) \\ \vdots \\ \mathrm{vec}(W_c^{(N)}) \end{pmatrix} \qquad \lambda \sum_i \|\Phi\Phi^\top - I\|_F^2$$

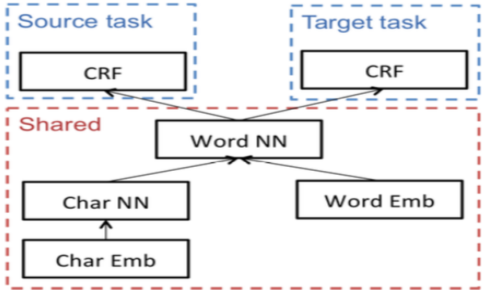
小规模数据集的NER算法：迁移学习



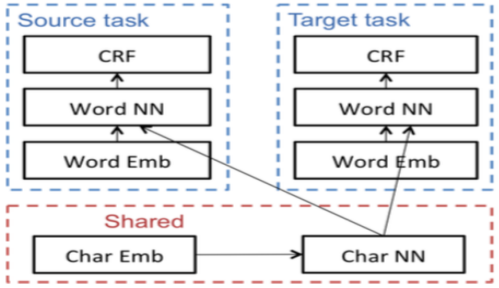
(a) Base model: both of Char NN and Word NN can be implemented as CNNs or RNNs.



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



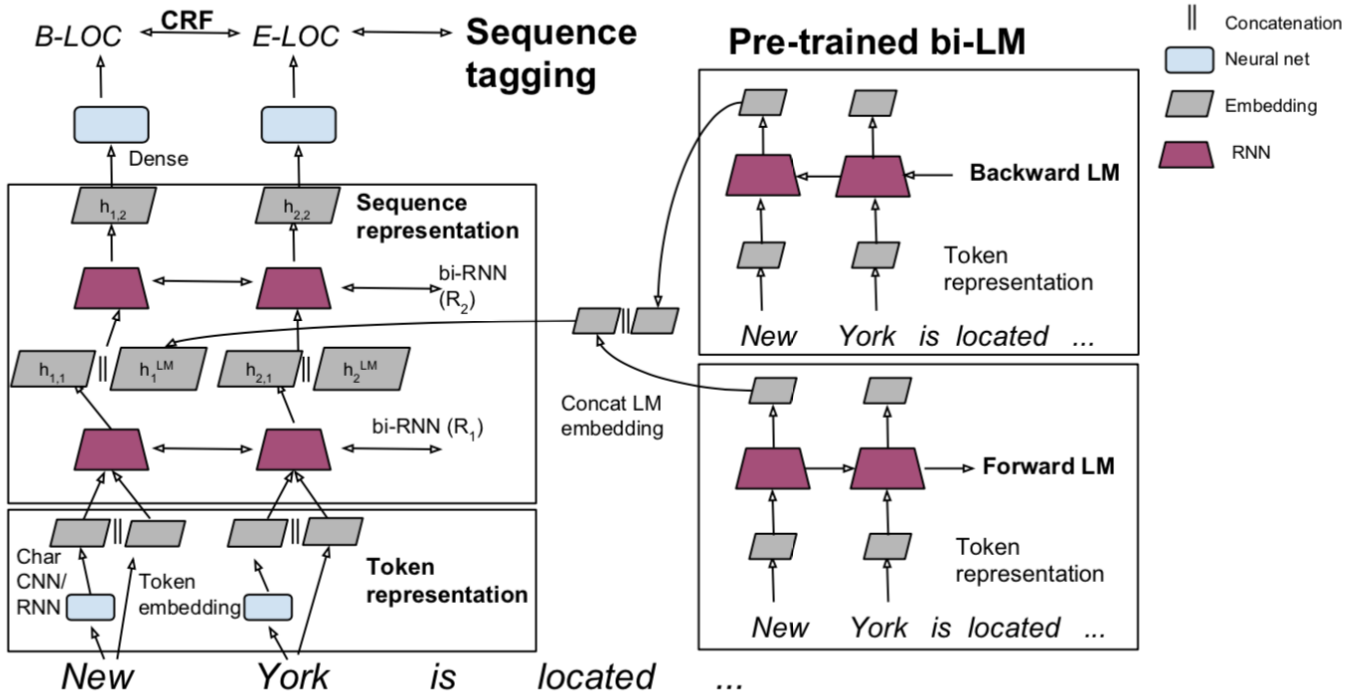
(c) Transfer model T-B: used for cross-domain transfer with disparate label sets, and cross-application transfer.



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

Transfer learning for sequence tagging with hierarchical current networks

小规模数据集的NER算法：引入LM的预训练模型



Semi-supervised sequence tagging with bidirectional language models

小规模数据集的NER算法：ELMo(Embedding from Language Models)

Deep contextualized word representations

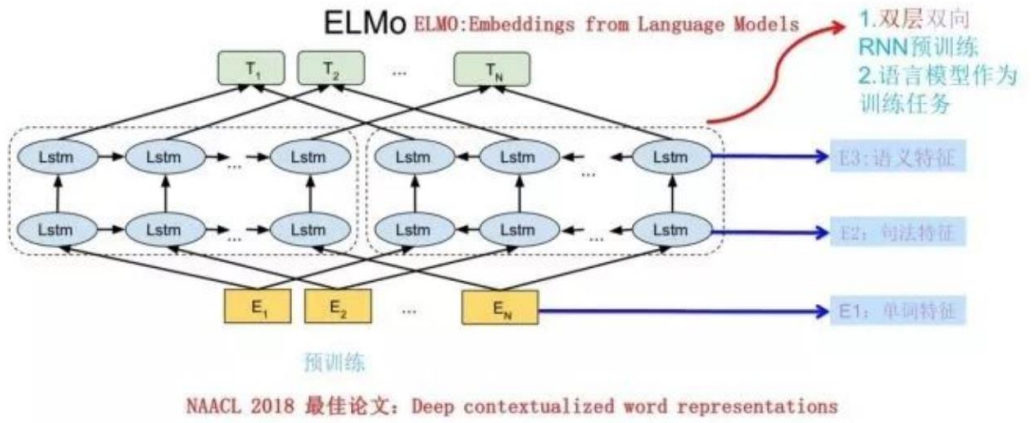
- 这里用到了双向的language model，

$$P(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \ldots, t_{k-1})$$

$$P(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, \ldots, t_N)$$

目标是最大上面两个的对数和

注意：这里的对token进行的上下文无关的编码,即用CNN对字符级进行编码



Token的新表达：

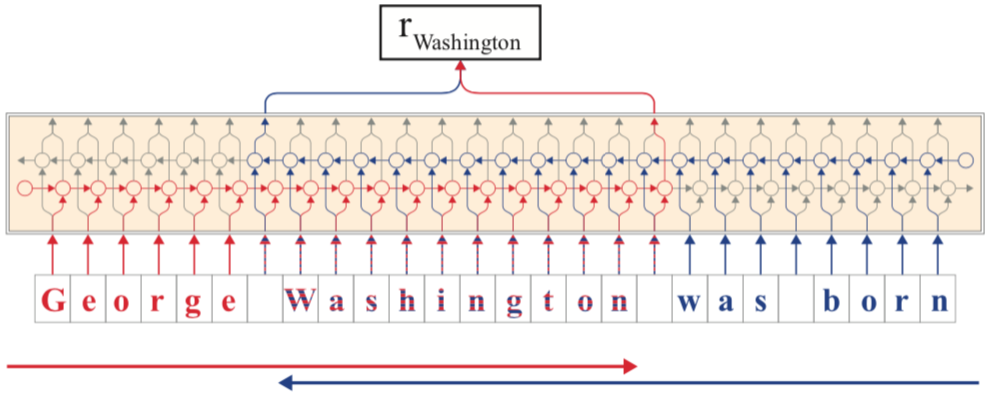
$$h_{k,j}^{LM} = [\overrightarrow{h}_{k,j}^{LM}; \overleftarrow{h}_{k,j}^{LM}]$$

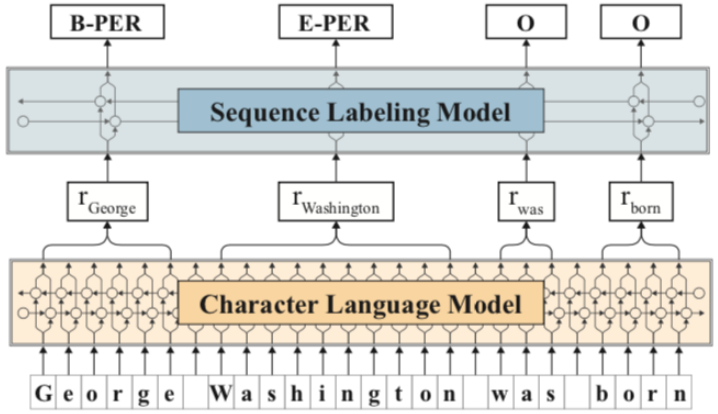
$$ELMo_k^{task} = E(R_k; \theta^{\text{task}})$$

$$= \gamma^{\text{task}} \sum_{j=1}^{L} s_j^{task} h_{k,j}^{LM}$$

小规模数据集的NER算法：引入预训练模型

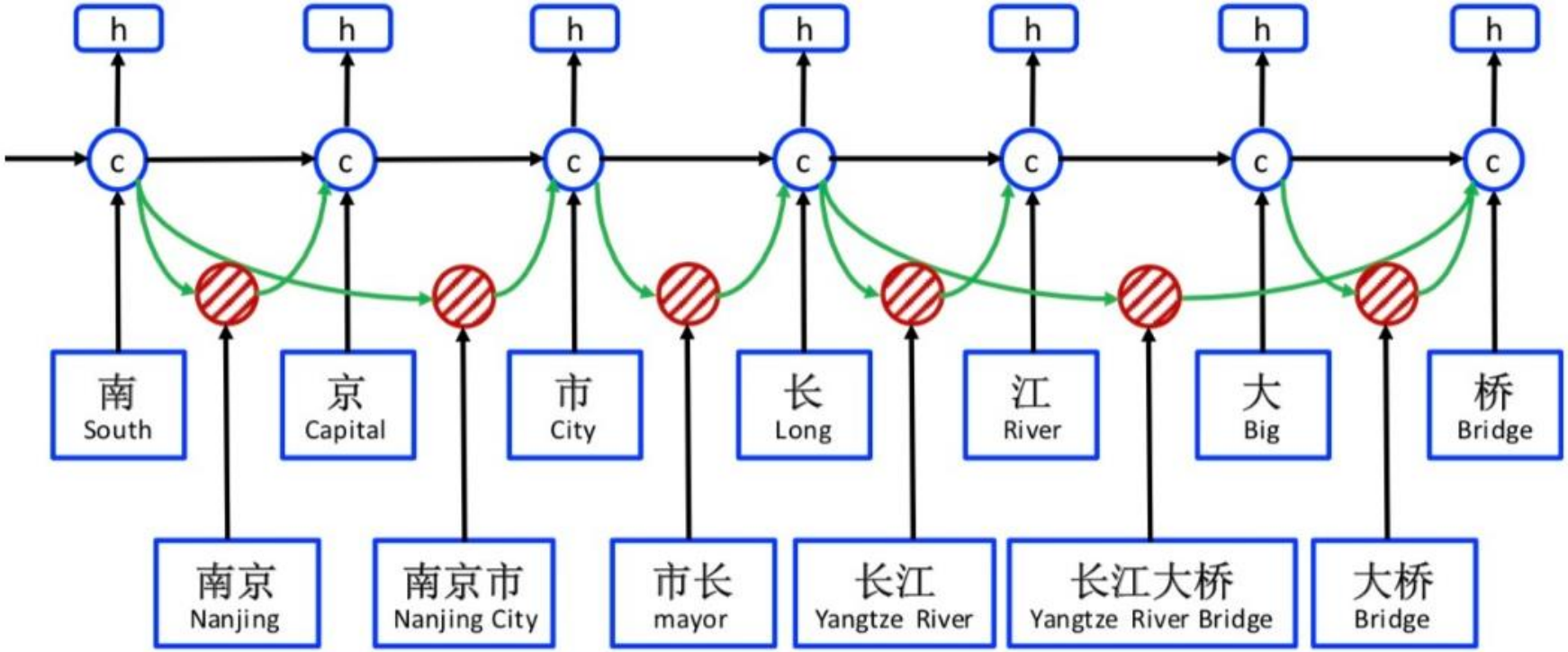
Contextual String Embeddings for Sequence Labeling





$$w_i = [\begin{matrix} w_i^{CharLM} \\ w_i^{Glove} \end{matrix}]$$

最优中文命名实体识别：《Chinese NER Using Lattice LSTM》



NER榜单(CoNLLo3数据集)

Model	F1	Paper / Source
Flair embeddings (Akbik et al., 2018)	93.09	Contextual String Embeddings for Sequence Labeling
BiLSTM-CRF+ELMo (Peters et al., 2018)	92.22	Deep contextualized word representations
Peters et al. (2017)	91.93	Semi-supervised sequence tagging with bidirectional language models
LM-LSTM-CRF (Liu et al., 2018)	91.71	Empowering Character-aware Sequence Labeling with Task-Aware Neural Language Model
Yang et al. (2017)	91.26	Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks
Ma and Hovy (2016)	91.21	End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF
LSTM-CRF (Lample et al., 2016)	90.94	Neural Architectures for Named Entity Recognition

汉语中NER的难点

- 1.汉语文本没有类似英文文本中空格之类的显式标示词的边界标示符。分词会对NER产生影响。
- 2.现代汉语文本，尤其是网络汉语文本，常出现中英文交替使用，这时汉语命名实体识别的任务还包括识别其中的英文命名实体；
- 3.在不同领域、场景下,命名实体的外延有差异,存在分类模糊的问题。不同命名实体之间界限不清晰,人名也经常出现在地名和组织名称中,存在大量的交叉和互相包含现象,而且部分命名实体常常容易与普通词混淆,影响识别效率。
- 4.命名实体构成结构比较复杂,并且某些类型的命名实体词的长度没有一定的限制,不同的实体有不同的结构,比如组织名存在大量的嵌套、别名、缩略词等问题,没有严格的规律可以遵循;人名中也存在比较长的少数民族人名或翻译过来的外国人名,没有统一的构词规范。因此,对这类命名实体识别的召回率相对偏低。