NLP

Named Entity Recognition

NER做什么

国务院	总理	李克强	调研	上海	外高桥	时	提出	，	支持	上海	积极	探索	新	机制	。
机构		人名		地名						地名					

中国	银行	西藏	自治区	分行	行长	索朗达吉
B-ORG	I-ORG	B-GPE	O	O	O	B-PER

- NER标注一般采用如下标注形式
  - O 非实体
  - S 单独构成实体
  - B 为实体的开始
  - I 为实体的中间
  - E 为实体的结尾

NER的发展

- 基于词典和规则的方法 [略]
- 传统机器学习
  - 无监督学习
  - CRF
- Emb[NN/Attention]-NN-CRF/LSTM [重点]
- 少量标注的训练集的研究 [重点]
  - 半监督学习
  - 主动学习
  - 对抗生成网络
  - 迁移学习
- 主要的问题[研究热点]
  - 标注集少
  - 标注结果准确率与成本的平衡

无监督学习[1][13]

- 不用标注数据，按数据的相似度聚类

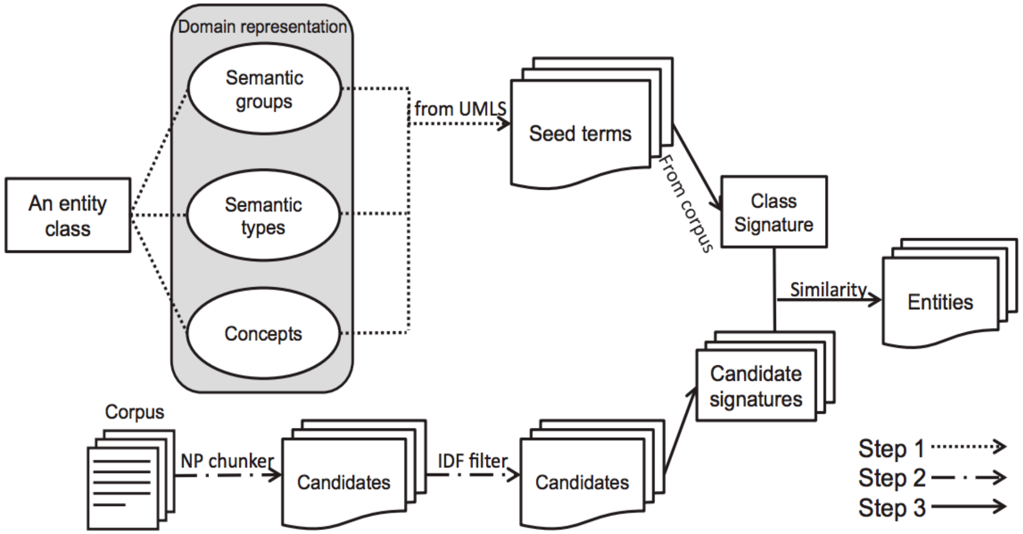


Fig. 1. Overall approach to unsupervised biomedical named entity recognition.

Step 1：Seed term collection

通过对实体类的语义组，类型，概念从UMLS中获取相关类的术语实体作为种子术语。

Step 2：Boundary detection

对Corpus进行分块，获取名词短语，使用IDF过滤

$$IDF(t,D) = log(|D|/|d \in D : t \in d|)$$

Step 3：Entity classification

构建词表大小V（all possible unigrams）

t[step1和step2中的结果]用特征维度为2V的s^t表示：

$$s^t = \langle s^t_1, s^t_2, \ldots, s^t_V, s^t_{V+1}, \ldots, s^t_{2V} \rangle \quad (2)$$

Values in the vector are calculated as follows:

$$s^t_i = w_i * f(v_i, t) * IDF(v_i, D), \quad i = 1 \ldots V \quad (3)$$

$$s^t_i = w_o * f(v_i, context_t) * IDF(v_i, D), \quad i = V + i \ldots 2V \quad (4)$$

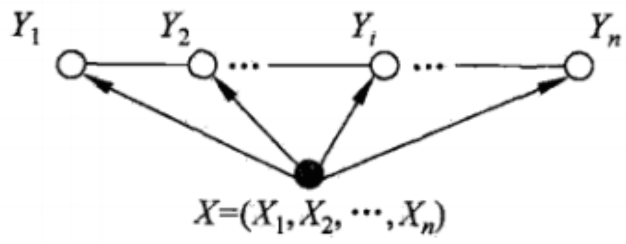
类别特征按类取平均，候选特征和类别特征用余弦相似度聚类

改进：可以使用K-means继续优化

[Shaodian Zhang et al.2013] Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts

CRF

- 观测序列X[未标记序列]和状态序列Y[标记序列]
- 转移概率：$Y_i \rightarrow Y_{i+1}$， 状态概率：$X \rightarrow Y_i$
- CRF：训练集中寻找实体上下文的模板，测试集中将上下文和模板匹配的中间词作为对应实体标注
-



$$P(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

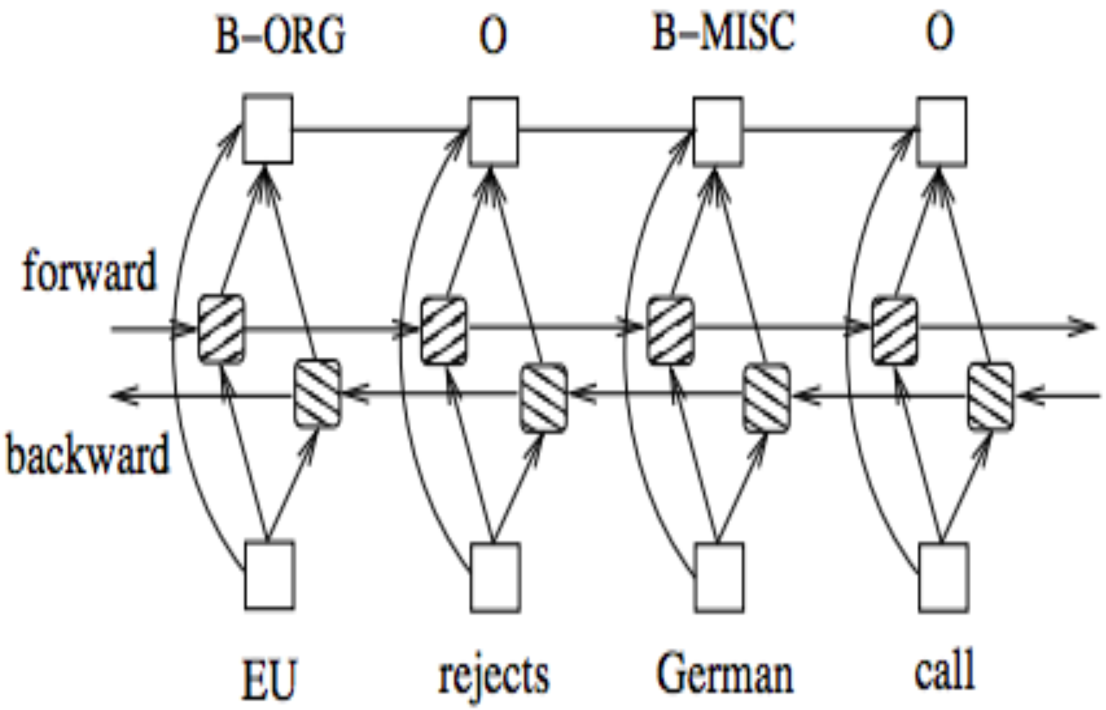
```python
features = [
    'bias',
    'word='+word,
  # 'word.lower=' + word.lower(),
  # 'word[-3:]=' + word[-3:],
   #'word[-2:]=' + word[-2:],
  # 'word.isupper=%s' % word.isupper(),
   #'word.istitle=%s' % word.istitle(),
    'word.isdigit=%s' % word.isdigit(),
    'postag=' + postag,
    'cuttag=' + cuttag,
  # 'postag[:2]=' + postag[:2],
]
if i > 0:
    word1 = sent[i - 1][0]
    postag1 = sent[i - 1][2]
    cuttag1 = sent[i - 1][1]
    features.extend([
        '-1:word='+word1,
        '-1:postag=' + postag1,
        '-1:cuttag=' + cuttag1,
      # '-1:postag[:2]=' + postag1[:2],
    ])
else:
    features.append('BOS')

if i < len(sent) - 1:
    word1 = sent[i + 1][0]
    postag1 = sent[i + 1][2]
    cuttag1 = sent[i + 1][1]
    features.extend([
        '+1:word=' + word1,
        '+1:postag=' + postag1,
        '+1:cuttag=' + cuttag1,
```

一般的CRF线性链如左图所示，中间的图是实际中使用CRF进行NER的特征提取代码，比如要预测Y_i，word为X_i，可以看到除了对X_i提取特征外，还涉及到了X_i前面的一个词和后面一个词。

但是并没有将全部的上下文考虑进去，这是CRF问题，后面的NN结构解决了该问题

Emb-BiLSTM-CRF[3][15]



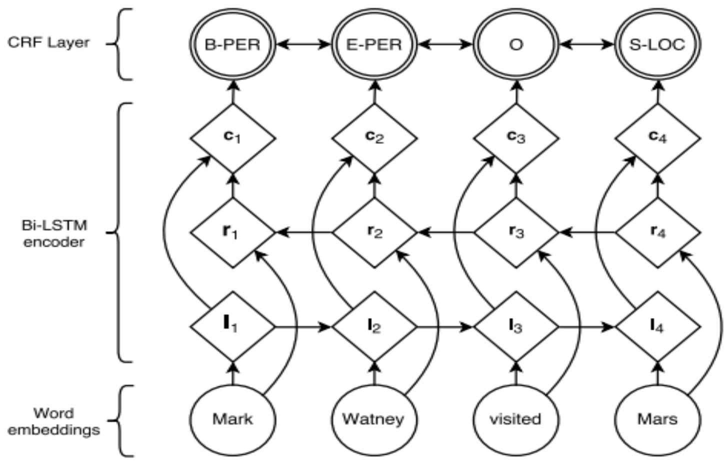
Embedding：使用前置训练，
130K Vocabulary，50 维

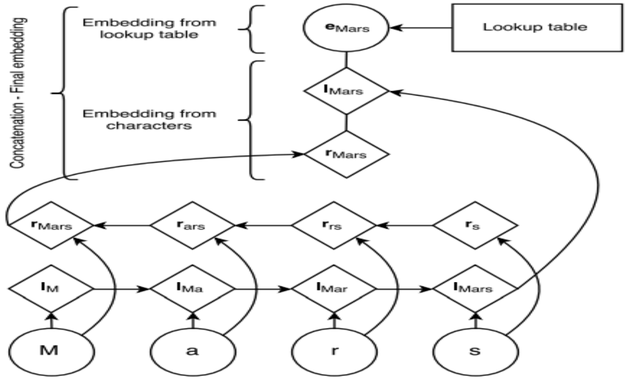
直接将BiLSTM的输出传给CRF层

CRF层的输入：手工添加了很多
特征，比如spelling,context等特征

[Zhiheng Huang et al.2015] Bidirectional LSTM-CRF Models for Sequence Tagging

Emb[LSTM]-BiLSTM-CRF[5][16]





句子的长度为 n，不同的标记个数为 k，将 Bi-LSTM 输出（要经过 softmax 层，变成 k 维的向量）作为打分矩阵 P，

即$P_{i,j}$ 表示句子的第 i 个单词对应的是第 j 个标签的分数

$$\mathrm{X} = (x_1, x_2, \dots, x_n)$$

$$\mathrm{y} = (y_1, y_2, \dots, y_n)$$

$$\mathrm{s(X, y)} = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$$

A 表示的是转移分数，$A_{i,j}$表示从第 i 个标记转到第 j 个标记的分数

softmax:

$$\mathrm{p(y|X)} = \frac{e^{s(X,y)}}{\sum_{y'} e^{s(X,y')}}$$

log 最大似然进行训练:

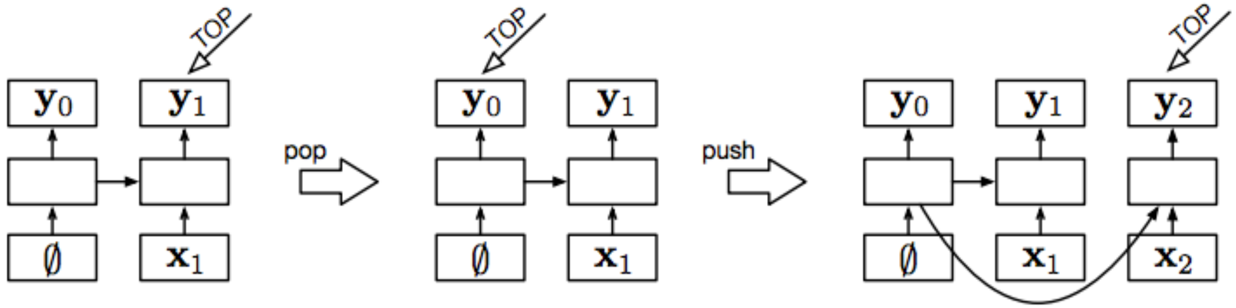
$$\log(\mathrm{p(y|X)}) = \mathrm{s(X, y)} - \log\left(\sum_{y'} e^{s(X,y')}\right)$$

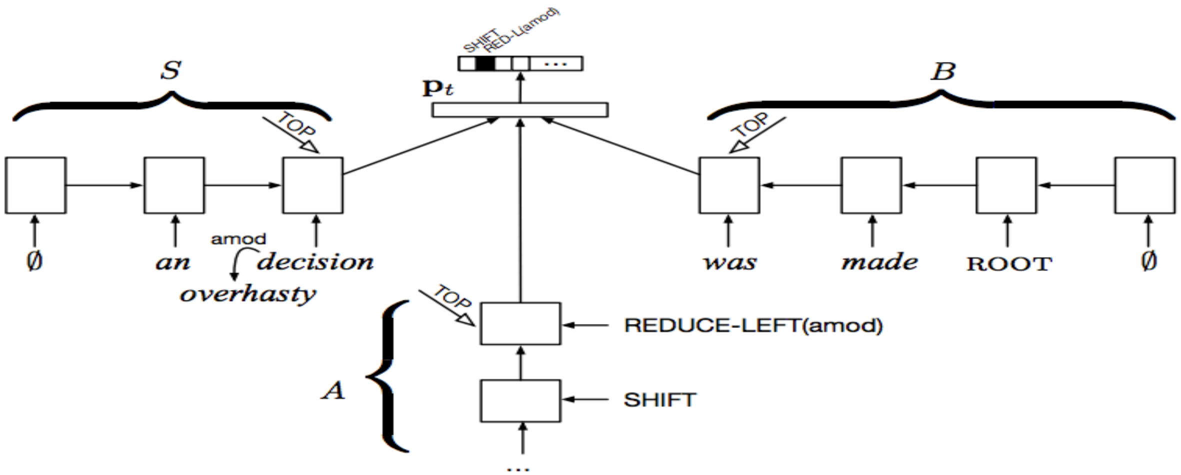
动态规划来预测:

$$y^* = \arg max_{y \in Y_x} s(X, y)$$

[Guillaume Lample et al.2016] Neural Architectures for Named Entity Recognition

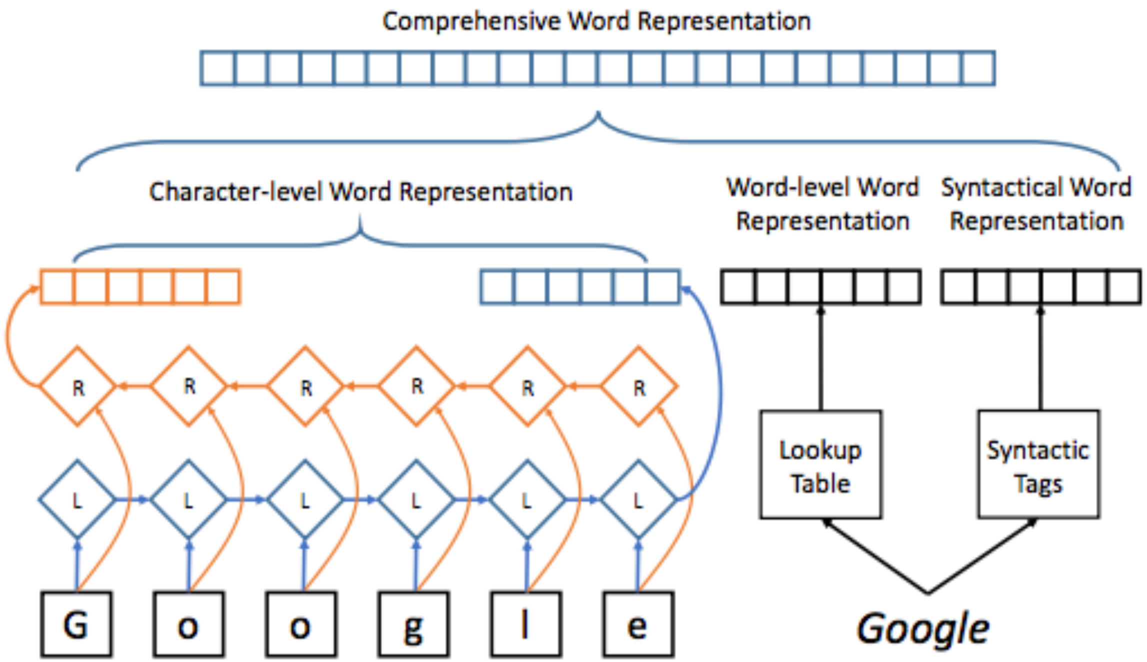
Stack-LSTM[5][16]





[Guillaume Lample et al.2016] Neural Architectures for Named Entity Recognition
[Chris Dyer et al.2015] Transition-Based Dependency Parsing with Stack Long Short-Term Memory

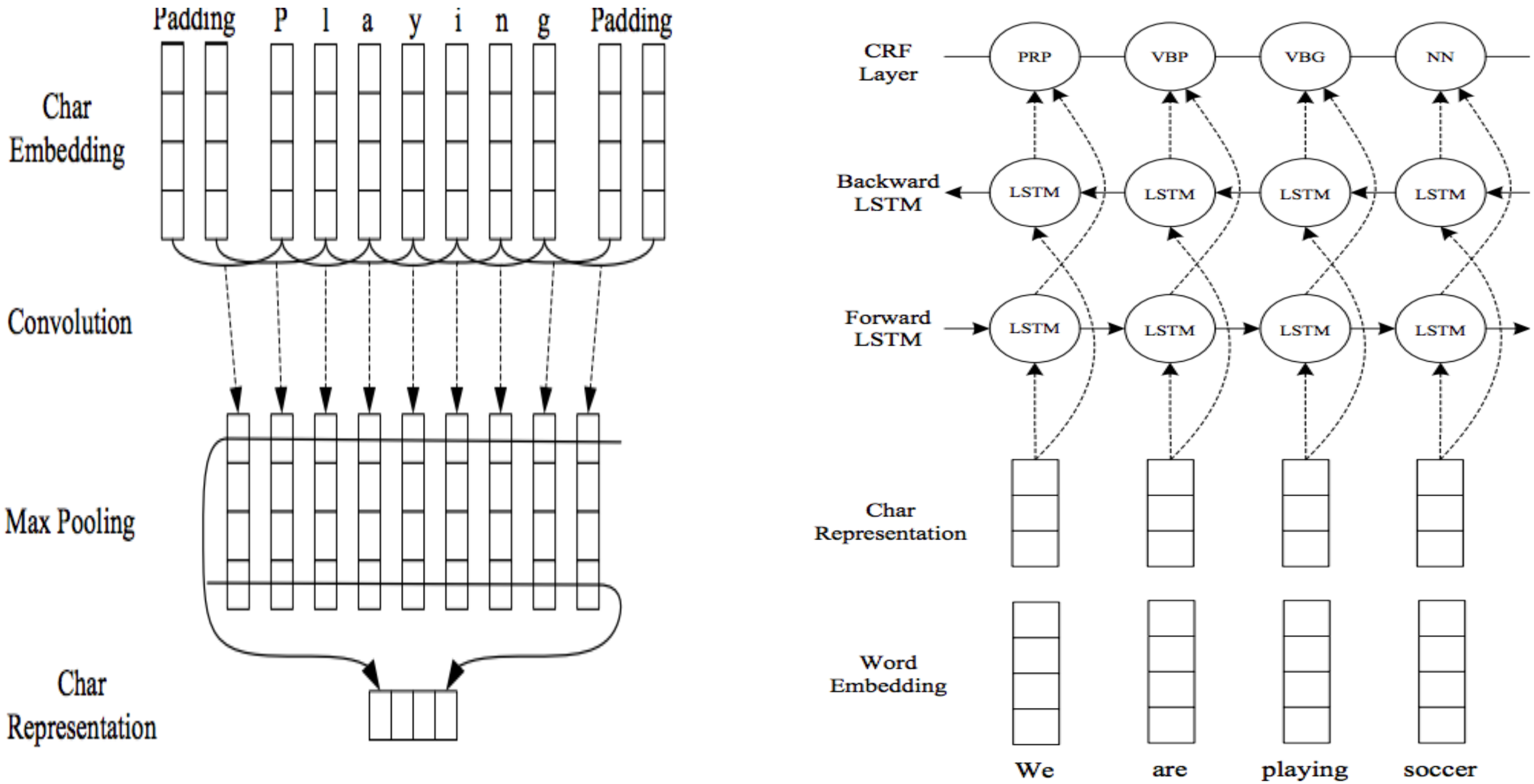
Multi-channel-BiLSTM-CRF[15][17]



主要是嵌入层的改变，如左图，除了加入字符嵌入，词嵌入外，还加入了词性，句法，语义等信息。

[Bill Y.Lin et al.2017] Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media

Emb[CNN]-BiLSTM-CRF[6][16]



[Xuezhe Ma et al.2016] End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF

Emb[Attention]+BiLSTM+CRF[7][16]

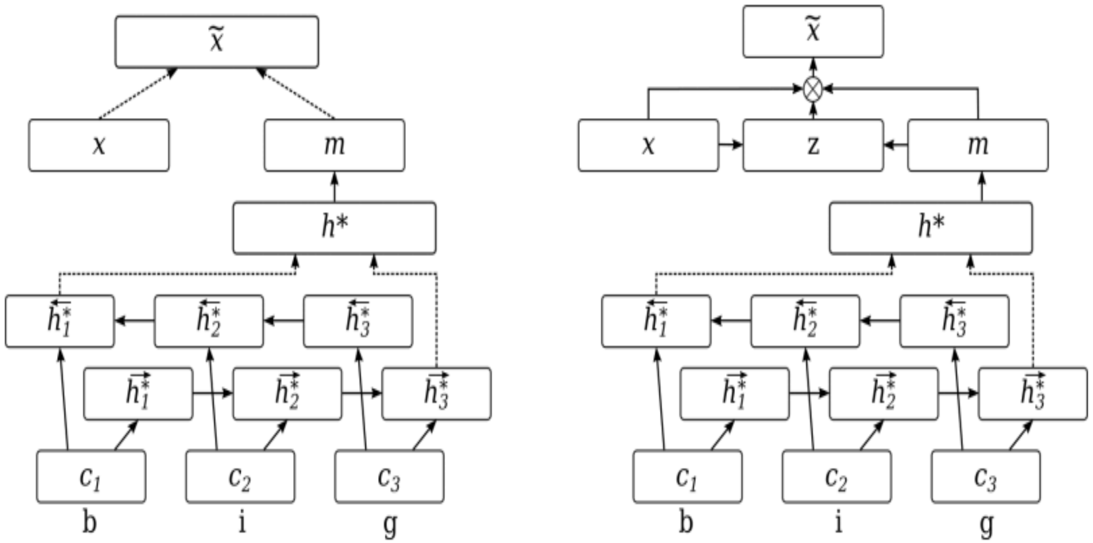


Figure 2: Left: concatenation-based character architecture. Right: attention-based character architecture. The dotted lines indicate vector concatenation.

$$z = \sigma(W_z^{(3)} tanh(W_z^{(1)} x + W_z^{(2)} m)) \qquad \widetilde{x} = z \cdot x + (1 - z) \cdot m$$

[Marek Rei et al.2016] Attending to Characters in Neural Sequence Labeling Models

Emb[LSTM]-BiLSTM-CRF[10][16]

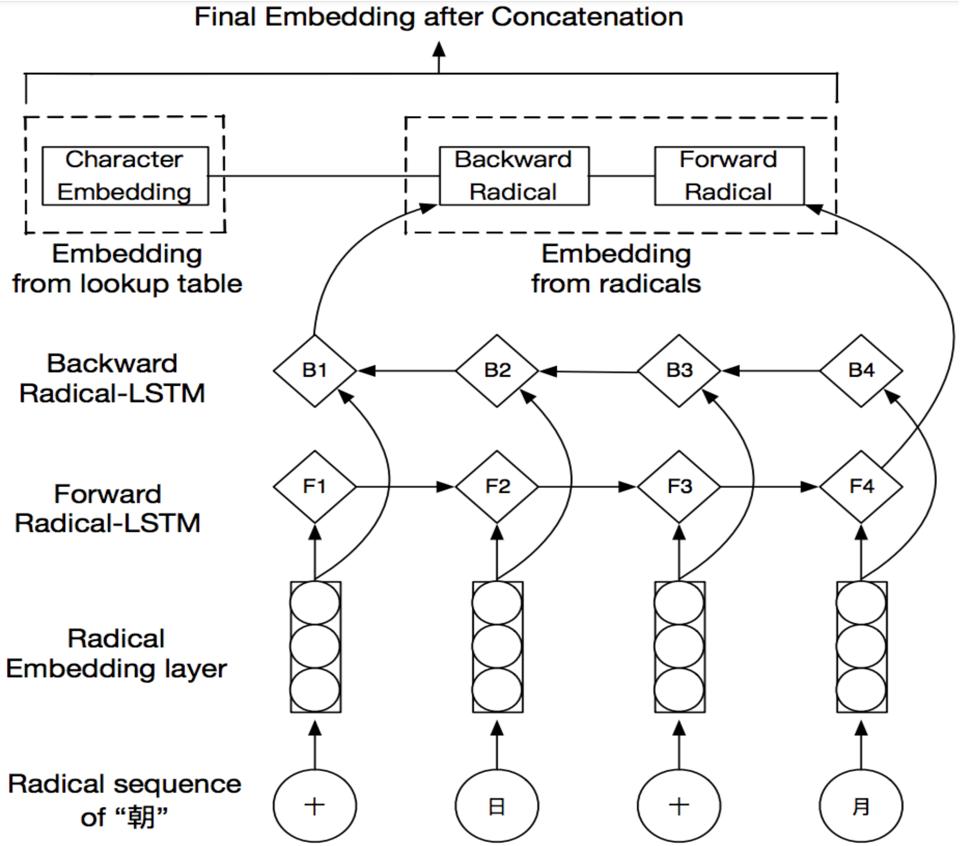


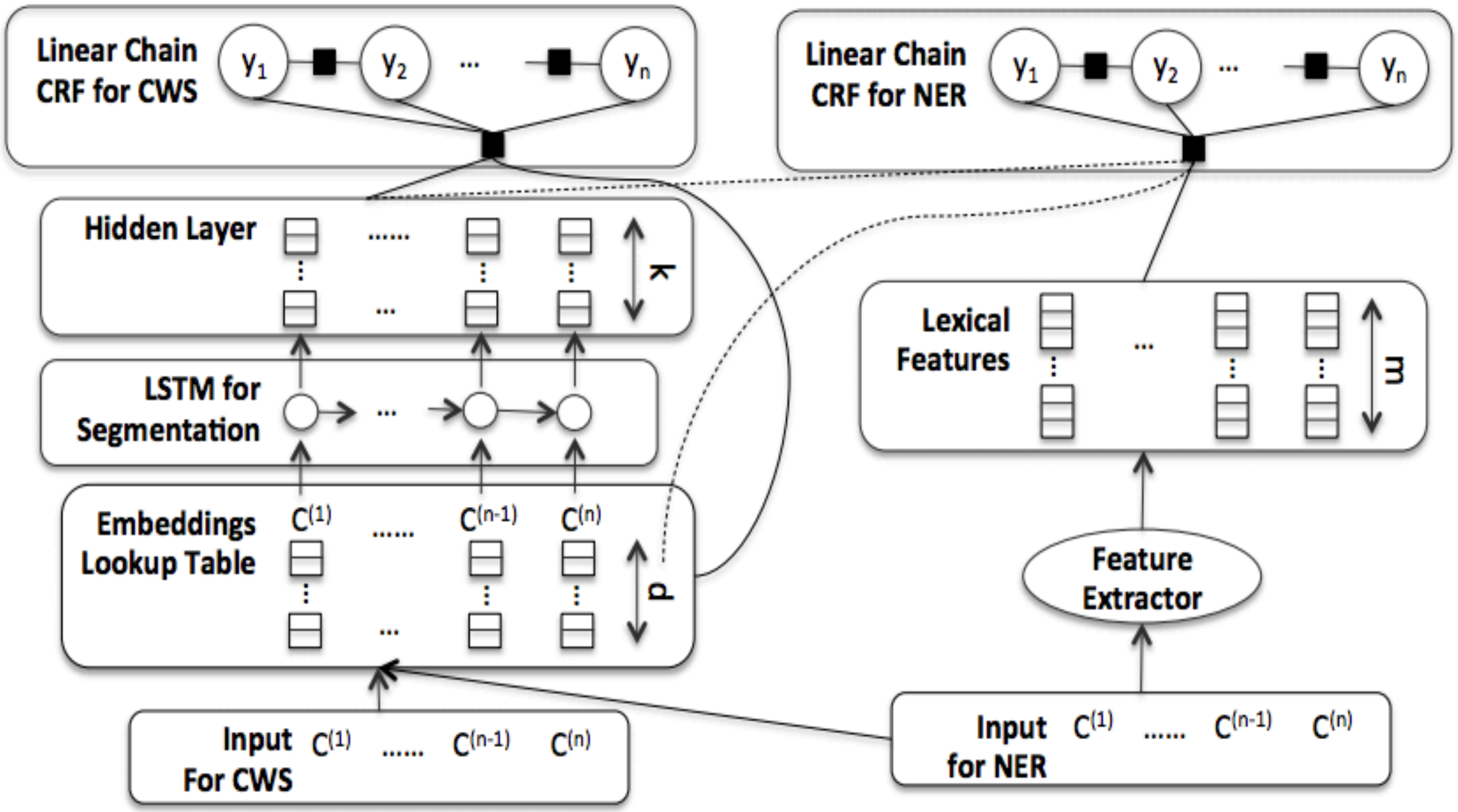
Table 3: Results with different components.

Variant	F1
random + dropout	88.91
random + radical + dropout	89.44
pretrain + dropout	90.75
pretrain	86.87

Model	PER-F	LOC-F	ORG-F	P	R	F
Zhou2006	90.09	85.45	83.10	88.94	84.20	86.51
Chen2006	82.57	90.53	81.96	91.22	81.71	86.20
Zhou2013	90.69	91.90	86.19	91.86	88.75	90.28
Zhang2006*	96.04	90.34	85.90	92.20	90.18	**91.18**
BLSTM-CRF + radical	89.62	91.76	85.79	91.39	88.22	89.78
BLSTM-CRF + pretrain	91.77	**92.10**	**87.30**	91.28	**90.62**	**90.95**

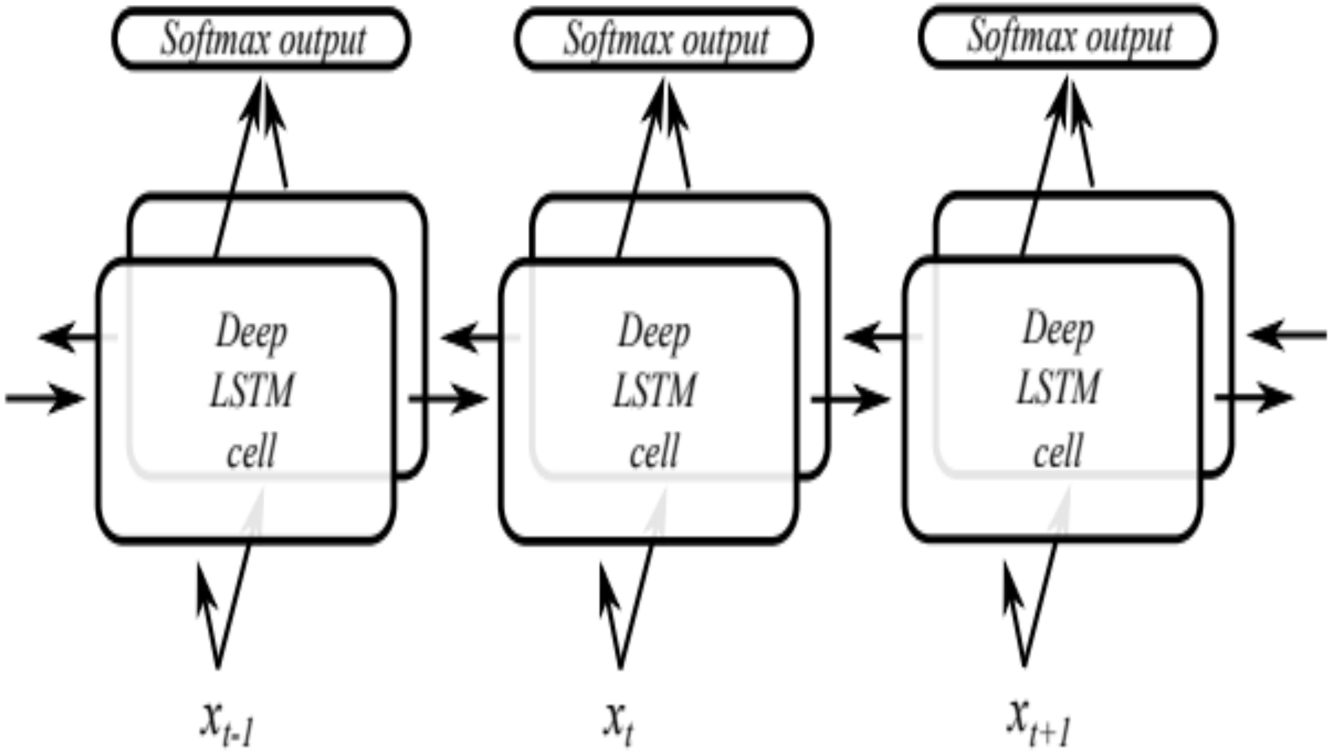
[Chuanhai Dong et al.2016] Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition

With Word Segmentation[8][16]



[Nanyun Peng et al.2016] Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning

Emb[Character]-DBi-LSTM-Softmax[9][16]

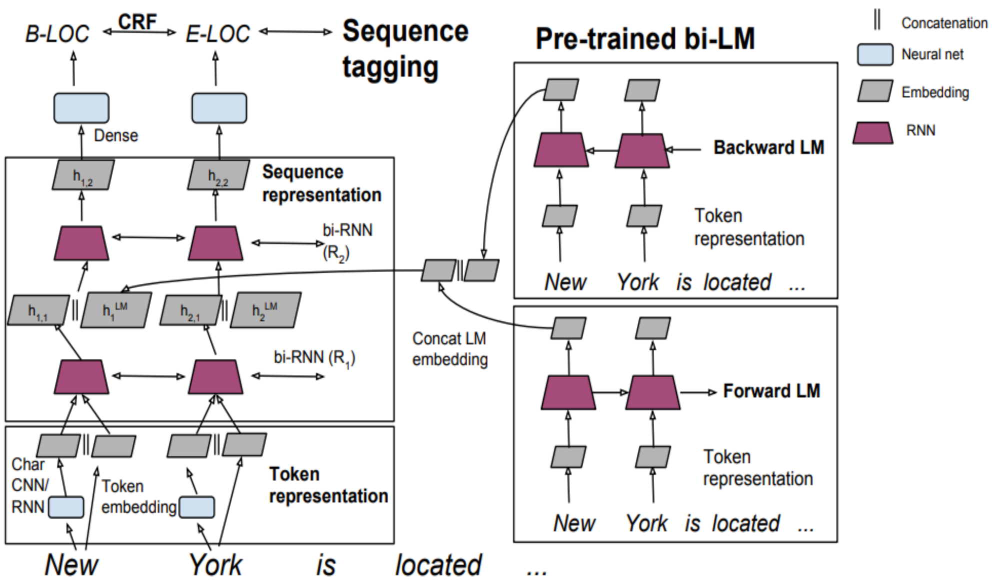


这篇paper中发现以下三种实体：
disorders and findings[疾病 and …]
pharmaceutical drugs[药物]
body structure[身体结果]

[Simon Almgren et al.2016] Named Entity Recognition in Swedish Health Records with Character-Based Deep Bidirectional LSTMs

Semi-supervised sequence tagging[11][17]

- 利用无标签数据来优化有标签数据训练的模型



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	**91.93 ± 0.19**

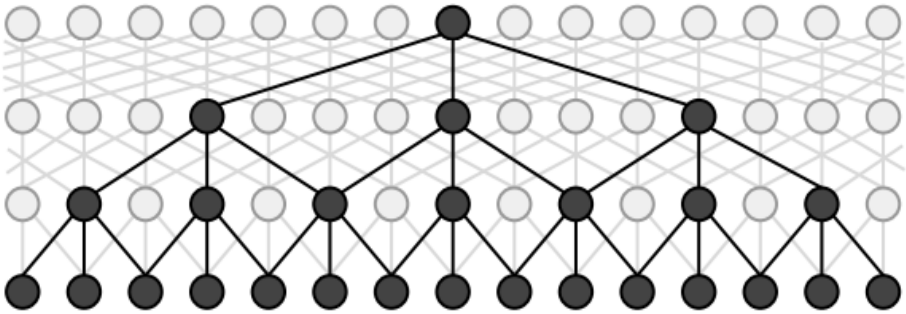
Table 1: Test set F_1 comparison on CoNLL 2003 NER task, using only CoNLL 2003 data and unlabeled text.

左边是NER的三层结构，中间层使用了两层BiLSTM
右边使用语言模型(前面词预测下一个词)训练词的上下文向量[语料是未标记的数据]

[Peters et al.2017] Semi-supervised sequence tagging with bidirectional language models

Emb-IDCNN-CRF[12][17]

- 从CNN和RNN对句子特征提取方面出发，使用迭代扩张卷积替换BiLSTM层



窗口大小为3，4层扩张CNN
本文将block定义为上面的一个结构，将这些block进行堆叠意为迭代

GPU资源匮乏，CNN在并行上的优势[特征独立]
CNN固定窗口大小无法获取更多上下文信息，不断增加CNN虽然可以扩大窗口，但参数增多，分辨表现差
ID-CNN不会损失分辨率，限定参数

Bi-LSTM-CRF (re-impl)	90.43 ± 0.12
ID-CNN-CRF	**90.54 ± 0.18**

Model	Speed
Bi-LSTM-CRF	1×
Bi-LSTM	9.92×
ID-CNN-CRF	1.28×
5-layer CNN	12.38×
ID-CNN	14.10×

[Emma Strubell et al.2017] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions

Active Learning - CRF in clinical text[2][15]

- 主动学习分为以下几个部分：
  - 1.初始模型生成：使用少量带标签的样本来构建模型，样本的采样可以使用下面两种方法：一种是随机采样，一种是使用最长句子采样
  - 2.查询：在pool中没有标注的句子使用查询算法被排序，一些算法是使用CRF模型来进行排序，一些不是，选择前N个句子进行注释，然后放到注释过的集合中，每次迭代的batch size按照8，16，32，64的规律选择。
  - 3.训练：CRF模型在更新的注释集中重新训练
  - 4.迭代：重复2，3过程，直到达到某个标准停止。
- 查询算法：
  - 基于不确定的查询算法，最不确定的句是最有信息量的句子[高度依赖于模型的质量，比如通过CRF模型来排序句子] [论文中比较的结果是该类算法整体优于下面的算法]
  - 基于多样性的查询算法，通过单词，语义，句法来构建向量，通过各个句子之间的相似度来排序

[Yukun Chen et al.2015] A study of active learning methods for named entity recognition in clinical text

Proactive Learning[14][17]

- Active learning
  - 假设标记的句子没有错误[乏味/困难导致出错在所难免]
  - 句子只拿给专家标注[成本高]
- Proactive learning
  - 对于困难的标注问题出现错误不可避免，允许专家出错
  - 使用两类注释器(者)，reliable expert 和 fallible expert[节约成本]
- Algorithm

1.在未标记数据集中的所有句子使用active learning 准则排序，最有信息的N个句子作为批次采样的输入，这这一步中，批次的句子分布到两个集合分别给reliable 和fallible。分给fallible的句子是fallible有一个很高概率标记正确的句子，同时，只有那些fallible很难标注的句子才会被送给reliable。这样标记的成本就会减少。

2. 第一步：设定阈值α，当fallible expert判别该句子的概率大于这个阈值时，将这个句子分配给fallible expert，否则进到第二步；

第二步：计算句子对reliable和fallible的不同：

$$
\begin{aligned}
diff(reliable, fallible, \boldsymbol{x}) \\
&= |p(CorrectLabels|reliable, \boldsymbol{x}) \\
&\quad - p(CorrectLabels|fallible, \boldsymbol{x})| \qquad (4)
\end{aligned}
$$

设定一个阈值β，当diff大于该阈值时，将句子交给reliable，否则交给fallible

Algorithm 1: Proactive Learning for NER

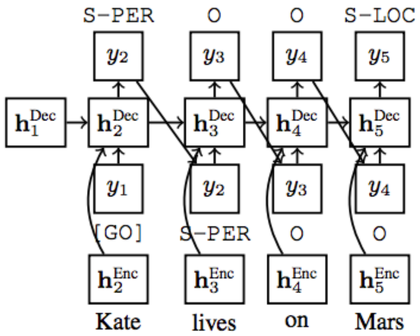
```
Input: a labelled dataset L, an unlabelled dataset UL, a
       test dataset T, a budget B, a reliable expert e_r
       with cost C_r for each sentence, a fallible expert
       e_f with cost C_f, the current cost C
Output: a labelled dataset L
Estimate the performance of each expert as described in
  Section 2.1;
while C < B do
    Train a named entity recognition model M on L;
    Sort all sentences in the unlabelled dataset
      according to an active learning criterion;
    Select the top N sentences;
    UL_r, UL_f =
      BatchSampling(M, top N sentences);
    L_r, L_f ← e_r and e_f annotate UL_r and UL_f
      respectively;
    L = L ∪ L_r ∪ L_f;
    UL = UL − UL_r − UL_f;
    C = C + C_r * |L_r| + C_f * |L_f|;
end
```

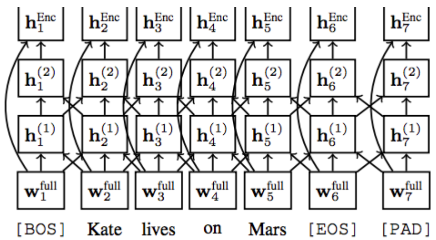
Algorithm 2: Batch Sampling

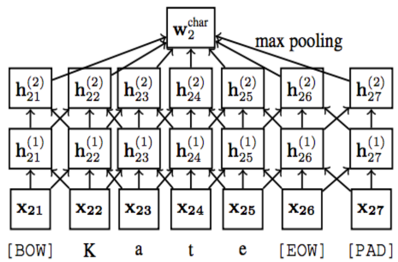
```
Input: a named entity recognition model M, top-N
       sentences selected according to an active
       learning criterion
Output: UL_r, UL_f
UL_r = ∅;
UL_f = ∅;
while Batch Size do
    // Stage 1
    foreach sentence x do
        if p(CorrectLabels|fallible, x) > α then
            UL_f = UL_f ∪ {x};
            BatchSize = BatchSize - 1
        end
    end
    // Stage 2
    if Batch Size ≠ 0 then
        Sort the remaining sentences according to a
          re-ranking criterion;
        Calculate threshold β;
        foreach sentence x do
            if Batch Size ≠ 0 then
                if diff(reliable, fallible, x) < β
                  then
                    UL_f = UL_f ∪ {x};
                else
                    UL_r = UL_r ∪ {x};
                end
                BatchSize = BatchSize - 1;
            end
        end
    end
end
```

[Maolin Li et al.2017] Proactive Learning for Named Entity Recognition

Deep Active Learning-CNN-CNN-LSTM[16][18]







模型实现加速：使用了CNN-CNN-LSTM模型，这个模型在标准数据集上实现了近乎最先进的效果，同时在计算上比最佳性能的模型更有效[时间和表现综合来说模型最优]

模型在小数据集上表现：在训练过程中执行增量的主动学习，仅仅使用原有训练集的25%，模型就能达到原有的最好表现[节约成本]

嵌入层使用character-level和word-level嵌入级联：

$$\mathbf{w}_i^{\text{full}} := \left(\mathbf{w}_i^{\text{char}}, \mathbf{w}_i^{\text{emb}}\right).$$

character-level的嵌入和信息提取层都使用的是CNN层，两层一维卷积+一层最大池化，窗口大小为3

$$\mathbf{h}_i^{\text{Enc}} = \left(\mathbf{h}_i^{(l)}, \mathbf{w}_i^{\text{full}}\right)$$

解码层使用CRF模块时间复杂度为O(nT^2)，而使用LSTM时间复杂度为O(nT)，且能达到同样的效果。

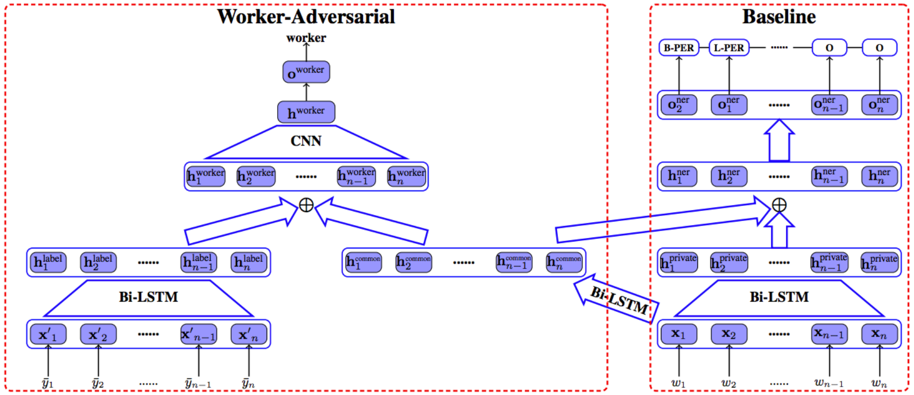
Char	Word	Tag	Reference	F1	Sec/Epoch
None	CNN	CRF	Collobert et al. (2011)	88.67	-
None	LSTM	CRF	Huang et al. (2015)	90.10	-
LSTM	LSTM	CRF	Lample et al. (2016)	90.94	-
CNN	LSTM	CRF	Chiu & Nichols (2016)	90.91 ± 0.20	-
GRU	GRU	CRF	Yang et al. (2016)	90.94	-
None	Dilated CNN	CRF	Strubell et al. (2017)	90.54 ± 0.18	-
LSTM	LSTM	LSTM		90.89 ± 0.19	49
CNN	LSTM	LSTM		90.58 ± 0.28	11
CNN	CNN	LSTM		90.69 ± 0.19	11
CNN	CNN	CRF		90.35 ± 0.24	12

Table 3: Evaluations on the test set of CoNLL-2003 English

[Yanyao Shen et al.2018] DEEP ACTIVE LEARNING FOR NAMED ENTITY RECOGNITION

Adversarial Learning for Crowd[17][18]

- 从众包标记数据质量低角度出发，通过对抗学习提取公有特征，减轻注释噪声



对抗学习的目的：优化公有模块的学习质量，使之收敛于真实的数据

原有的结构不变(如右部分)，只是加入了对抗网络(左部分)，使用众包标签作为输入，经过Bi-LSTM，和common级联，一维卷积，窗口大小是5，最大池化，最后经过softmax，维度是标注者的个数：

$$\mathbf{h}_t^{\text{worker}} = \mathbf{h}_t^{\text{common}} \oplus \mathbf{h}_t^{\text{label}}$$

$$\tilde{\mathbf{h}}_t^{\text{worker}} = \tanh(\mathbf{W}^{\text{cnn}}[\mathbf{h}_{t-2}^{\text{worker}}, \mathbf{h}_{t-1}^{\text{worker}}, \cdots, \mathbf{h}_{t+2}^{\text{worker}}])$$

$$\mathbf{h}^{\text{worker}} = \text{max-pooling}(\tilde{\mathbf{h}}_1^{\text{worker}}\tilde{\mathbf{h}}_2^{\text{worker}}\cdots\tilde{\mathbf{h}}_n^{\text{worker}})$$

$$\mathbf{o}^{\text{worker}} = \mathbf{W}^{\text{worker}}\mathbf{h}^{\text{worker}},$$

$$p(\bar{z}|\mathbf{X}, \bar{\mathbf{y}}) = \frac{\exp(\mathbf{o}_{\bar{z}}^{\text{worker}})}{\sum_z \exp(\mathbf{o}_z^{\text{worker}})},$$

左右联合：

$$\begin{aligned} \mathrm{R}(\Theta, \Theta', \mathbf{X}, \bar{\mathbf{y}}, \bar{z}) &= \text{loss}(\Theta, \mathbf{X}, \bar{\mathbf{y}}) - \text{loss}(\Theta, \Theta', \mathbf{X}) \\ &= -\log p(\bar{\mathbf{y}}|\mathbf{X}) + \log p(\bar{z}|\mathbf{X}, \bar{\mathbf{y}}), \end{aligned}$$

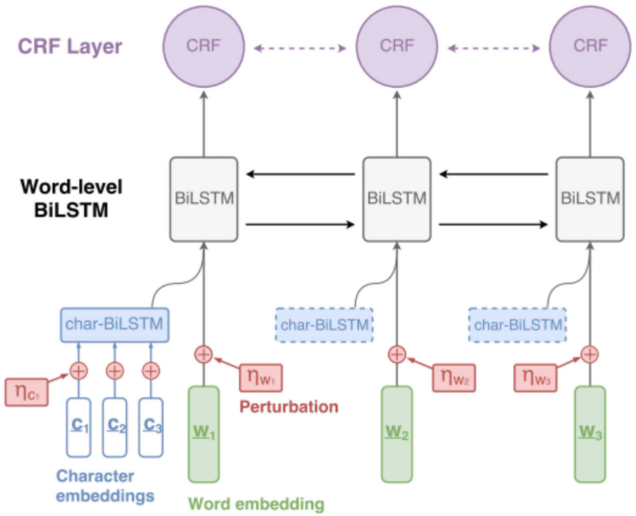
其中Θ表示整个模型中和NER有关的参数，Θ′表示仅仅和worker discriminator有关的参数，其中common Bi-LSTM对应的参数属于Θ中，优化下面式子：

$$\hat{\Theta} = \arg\min_{\Theta} \mathrm{R}(\Theta, \Theta', \mathbf{X}, \bar{\mathbf{y}}, \bar{z})$$

$$\hat{\Theta}' = \arg\max_{\Theta'} \mathrm{R}(\hat{\Theta}, \Theta', \mathbf{X}, \bar{\mathbf{y}}, \bar{z})$$

[Yaosheng Yang et al.2018] Adversarial Learning for Chinese NER from Crowd Annotations

Adversarial Training for Robust[19][18]



	Our Models		Plank et al. (2016)			Berend (2017)	Nguyen et al. (2017)
	Baseline	Adversarial	BiLSTM	TNT	CRF		
bg	98.34	**98.53**	97.97	96.84	96.36	95.63	97.4
cs	98.70	**98.81**	98.24	96.82	96.56	95.83	–
da	96.63	**96.74**	96.35	94.29	93.83	93.32	95.8
de•	94.29	**94.35**	93.38	92.64	91.38	90.73	92.7
en	95.72	**95.82**	95.16	94.55	93.35	93.47	94.7
es	96.26	**96.44**	95.74	94.55	94.23	94.69	95.9
eu•	94.55	94.71	**95.51**	93.35	91.63	90.63	93.7
fa	97.38	**97.51**	97.49	95.98	95.65	96.11	96.8
fi•	94.54	95.40	**95.85**	93.59	90.32	89.19	94.6
fr	96.48	**96.63**	96.11	94.51	95.14	94.96	96.0
he	97.34	**97.43**	96.96	93.71	93.63	95.28	–
hi	97.12	**97.21**	97.10	94.53	96.00	96.09	96.4
hr•	96.12	96.32	**96.82**	94.06	93.16	93.53	–
id	93.95	**94.03**	93.41	93.16	92.96	92.02	93.1
it	98.04	**98.08**	97.95	96.16	96.43	96.28	97.5
nl	92.64	93.09	**93.30**	88.54	90.03	85.10	91.4
no	97.88	**98.08**	98.03	96.31	96.21	95.67	97.4
pl•	97.34	97.57	**97.62**	95.57	93.96	93.95	96.3
pt	97.94	**98.07**	97.90	96.27	96.32	95.50	97.5
sl•	97.81	**98.11**	96.84	94.92	94.77	92.70	97.1
sv	96.39	**96.70**	96.69	95.19	94.45	94.62	–
Avg	96.45	**96.65**	96.40	94.55	94.11	93.59	95.55
el	98.18	**98.24**	–	–	–	97.12	–
et•	90.79	**91.32**	–	–	–	86.30	–
ga	90.66	**91.11**	–	–	–	88.82	–
hu•	93.39	**94.02**	–	–	–	89.47	–
ro	91.24	**91.46**	–	–	–	88.99	–
ta	82.91	**83.16**	–	–	–	81.80	–
Avg	91.20	**91.55**	–	–	–	88.41	–

在训练词嵌入和字符嵌入的时候加入噪声信息，对于加入噪声的输入，模型仍然需要正确标注出来，提高模型的鲁棒性

输入句子所有的单词/字符嵌入用s表示，模型的参数用θ表示，y为目标的词性标记序列，训练过程中，模型要最小化负对数似然函数：

$$L(\boldsymbol{\theta}; \boldsymbol{s}, \boldsymbol{y}) = -\log p(\boldsymbol{y} \mid \boldsymbol{s}; \boldsymbol{\theta})$$

在s上加一个连续的扰动向量，这个扰动向量为使得我们的模型的损失函数最大：

$$\boldsymbol{\eta} = \underset{\boldsymbol{\eta}': \|\boldsymbol{\eta}'\|_2 \leq \epsilon}{\arg\max} \; L(\hat{\boldsymbol{\theta}}; \boldsymbol{s} + \boldsymbol{\eta}', \boldsymbol{y})$$

θ^为模型的参数（在上述求解过程中为常量）对抗训练的样本为：

$$\boldsymbol{s}_{\text{adv}} = \boldsymbol{s} + \boldsymbol{\eta}$$

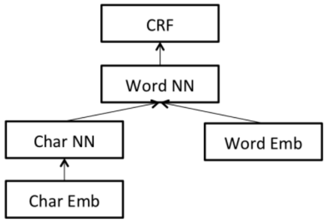
定义如下对抗训练的损失函数(本文γ = 0.5)：

$$\tilde{L} = \gamma L(\boldsymbol{\theta}; \boldsymbol{s}, \boldsymbol{y}) + (1 - \gamma) L(\boldsymbol{\theta}; \boldsymbol{s}_{\text{adv}}, \boldsymbol{y})$$

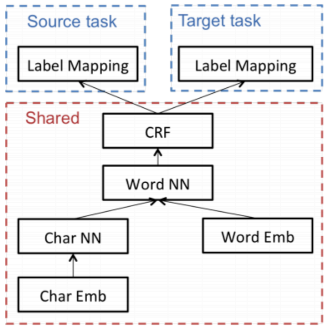
[Michihiro Yasunage et al.2018] Robust Multilingual Part-of-Speech Tagging via Adversarial Training

Transfer Learning for sequence tagging[13][17]

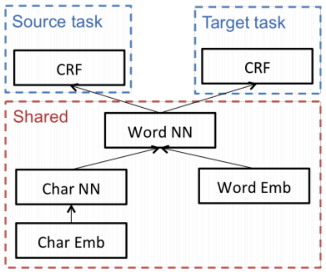
- 迁移学习通过学习相关领域(source)的知识来提高当前领域(target)模型的性能



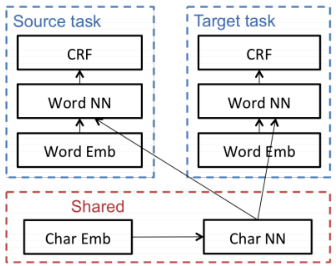
(a) Base model: both of Char NN and Word NN can be implemented as CNNs or RNNs.



(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.

该论文中 word-level和character-level部分都使用的是GRU
其中CRF部分添加了max-margin principle，如下：

$$f(\mathbf{h}, \mathbf{y}) - \log \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{h})} \exp(f(\mathbf{h}, \mathbf{y}') + \mathrm{cost}(\mathbf{y}, \mathbf{y}')),$$

h 为从 word-level 和 character-level 部分学习到的特征表示，$h = (h_1, .., h_T)$，y 为标记序列，$y = (y_1, ..., y_T)$，其中Y(h)表示 h 对应的标记空间，cost 函数表示当y'和 y 的差距越大，惩罚越大。f 函数表示的是 CRF 中的打分函数，前面提到的s(X, y)。

Table 3: Comparison with state-of-the-art results (%).

Model	CoNLL 2000	CoNLL 2003	Spanish	Dutch	PTB 2003
Collobert et al. (2011)	94.32	89.59	–	–	97.29
Passos et al. (2014)	–	90.90	–	–	–
Luo et al. (2015)	–	91.2	–	–	–
Huang et al. (2015)	94.46	90.10	–	–	97.55
Gillick et al. (2015)	–	86.50	82.95	82.84	–
Ling et al. (2015)	–	–	–	–	**97.78**
Lample et al. (2016)	–	90.94	85.75	81.74	–
Ma & Hovy (2016)	–	91.21	–	–	97.55
Ours w/o transfer	94.66	91.20	84.69	85.00	97.55
Ours w/ transfer	**95.41**	**91.26**	**85.77**	**85.19**	97.55

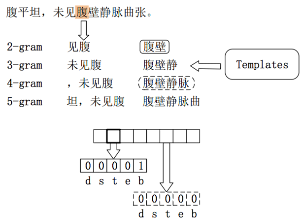
[Zhilin Yang et al.2017] TRANSFER LEARNING FOR SEQUENCE TAGGING WITH HIERARCHICAL RECURRENT NETWORKS
Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition 28 Apr 2018

Incorporating dictionaries for NER[20][18]

- 数据驱动的方法典型缺乏处理稀有或没有出现的实体，本文在深层网络中加入词典

CCKS-2017 Task2：临床命名实体识别
本文使用character-level作为嵌入层，同时构造了一个特征向量di
特征构造分三种：
1.N-gram feature
总共有8个temple，每个temple映射到5维上，d，s，t，e，b，分别表示disease，symptom，treatment，exam，body-part，即每个字符在这个特征类别上有40维度。

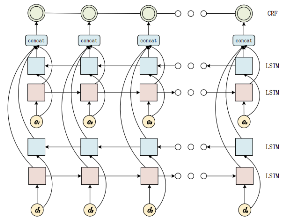
Type	template
2-gram	$x_{i-1}x_i$, x_ix_{i+1}
3-gram	$x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}$
4-gram	$x_{i-3}x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}x_{i+3}$
5-gram	$x_{i-4}x_{i-3}x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}x_{i+3}x_{i+4}$

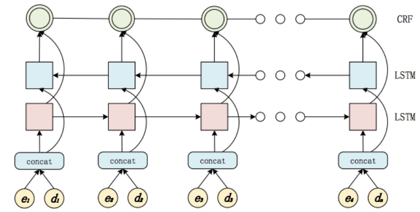


2.Position-Independent Entity Type feature
对句子X基于词典D进行前向最大匹配，并进行归类，然后将每个字符进行映射，这里其特征表示可以使用one-hot或者特征嵌入矩阵，如下第三行所示
3. Position-Dependent Entity Type feature
这个是有2中切词后的结果来获取的，即该字符是否是实体的开始，中间，结束，或者是单个字符实体，分别用B，I，E，S，同时可以用one-hot或者特征嵌入矩阵

Character sequence	腹	平	坦	，	未	见	腹	壁	静	脉	曲	张	。
tag sequence	S-b	O	O	O	O	O	B-b	E-b	B-s	I-s	I-s	E-s	O
PIET features	b	None	None	None	None	None	b	b	s	s	s	s	None
PDET features	S-b	None	None	None	None	None	B-b	E-b	B-s	I-s	I-s	E-s	None

这些特征都受词典或上下文的影响，而不受其他句子或统计信息的影响，因此是有别于流行的数据驱动的方法。





提出两种模型如上图所示，3种特征的5种表现形式和2种模型进行组合，结果如下：

		Model-I			Model-II		
		Precision	Recall	F_1-Measure	Precision	Recall	F_1-Measure
N-gram feature		88.39	88.46	88.43	88.72	88.71	88.71
PIET feature	one-hot encoding	89.53	90.58	90.05	89.38	90.49	89.93
	feature embedding	90.11	90.01	90.56	90.00	90.60	90.30
PDET feature	one-hot encoding	90.51	91.04	90.77	90.22	90.64	90.43
	feature embedding	**90.83**	**91.64**	**91.24**	90.36	91.35	90.85

[Qi Wang et al.2018] Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition

Conclusion

- 经典的三层模型是不会变的
  - Emb representation-info extract-CRF transfor
- 通过各种手段来提高模型的表现
  - 对抗生成网络
  - Attention
  - 加词表，加特征
- 少量标签的学习/加速
  - 迁移学习
  - 主动学习
  - CNN替换LSTM，LSTM替换CRF

paper:

- [1].[Shaodian Zhang et al.2013] Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts
- [2].[Yukun Chen et al.2015] A study of active learning methods for named entity recognition in clinical text
- [3]. [Zhiheng Huang et al.2015] Bidirectional LSTM-CRF Models for Sequence Tagging
- [4].[Chris Dyer et al.2015] Transition-Based Dependency Parsing with Stack Long Short-Term Memory
- [5].[Guillaume Lample et al.2016] Neural Architectures for Named Entity Recognition
- [6].[Xuezhe Ma et al.2016] End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF
- [7].[Marek Rei et al.2016] Attending to Characters in Neural Sequence Labeling Models
- [8].[Nanyun Peng et al.2016] Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning
- [9].[Simon Almgren et al.2016] Named Entity Recognition in Swedish Health Records with Character-Based Deep Bidirectional LSTMs
- [10].[Chuanhai Dong et al.2016] Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition
- [11].[Peters et al.2017] Semi-supervised sequence tagging with bidirectional language models
- [12].[Emma Strubell et al.2017] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions
- [13].[Zhilin Yang et al.2017] TRANSFER LEARNING FOR SEQUENCE TAGGING WITH HIERARCHICAL RECURRENT NETWORKS
- [14].[Maolin Li et al.2017] Proactive Learning for Named Entity Recognition
- [15].[Bill Y.Lin et al.2017] Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media
- [16].[Yanyao Shen et al.2018] DEEP ACTIVE LEARNING FOR NAMED ENTITY RECOGNITION
- [17].[Yaosheng Yang et al.2018] Adversarial Learning for Chinese NER from Crowd Annotations
- [18].[Zhenghui Wan et al.2018] Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition
- [19].[Michihiro Yasunage et al.2018] Robust Multilingual Part-of-Speech Tagging via Adversarial Training
- [20].[Qi Wang et al.2018] Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition

NLP