命名实体识别

Name Entity Recognition

马新宇

2018/10/23

问题历史

- 命名实体识别任务首先在1996年(R. Grishman & Sundheim)第六届Message Understanding Conference (MUC-6)提出。
- 该会议最初关注的是信息抽取(Information Extract)任务，但是发现识别信息中的一些实体是非常必要的，后将其列为子任务。

问题重要性

- NER是NLP中一项基础性关键任务。
- 从自然语言处理的流程来看，NER可以看作词法分析中未登录词识别的一种，是未登录词中数量最多、识别难度最大、对分词效果影响最大问题。
- 同时NER也是关系抽取、事件抽取、知识图谱、机器翻译、问答系统等诸多NLP任务的基础，通常是这些任务的第一步。
- NER是序列标注任务（chunk tagging, part-of-speech, name entity recognition）的一种。

问题定义

- 命名实体
  - 实体是存在于现实世界中并且可以与其他物体区分开来的物体。我们可以用一系列的属性描述这个实体并把它的差别描述出来。
  - 命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。
  - 3大类（实体类，时间类，数字类）和7小类（人名、地名、组织机构名、时间、日期、货币、百分比）
- 识别（Recognition）
  - 命名实体标识（NE Identification）和命名实体分类（NE Classification）
  - 发现哪个是，是什么类型

评价指标

- 常用的评价指标
  - Precision $= \frac{TP}{TP+FP}$, Re$call = \frac{TP}{TP+FN}$, F1 $= \frac{2PR}{P+R} = \frac{2TP}{2TP+FP+FN}$

	相关(Relevant)，正类	无关(NonRelevant)，负类
被检索到(Retrieved)	True positives (TP, 正类判定为正类)	False positives (FP，负类判定为正类)
未被检索到(Not Retrieved)	False negatives (FN, 正类判定为负类)	True negatives (TN，负类判定为负类)

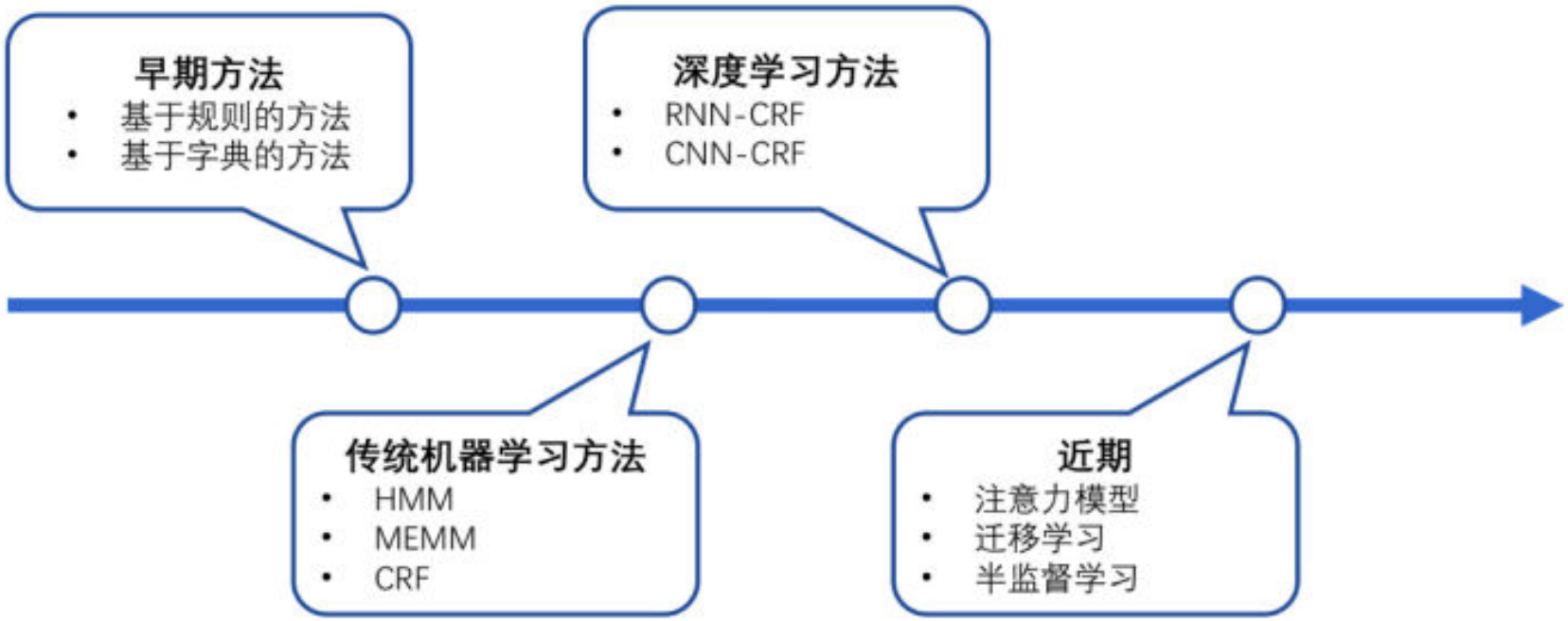
评价指标(CoNLL2003, MUC-6, ACE)

- 不同会议，指标不同
- CoNLL2003
  - Exact matching完全匹配，包括边界和类型
- MUC-6
  - Partial matching部分匹配，边界或者类型
- ACE
  - 更复杂的方案：部分匹配，并为每种类型分配权重

评价指标

- 由于属于多分类问题，综合考察在不同类别下分类器的优劣，这时候就需要引入宏平均（Macro-averaging）、微平均（Micro-averaging）
- 宏平均（Macro-averaging）
  - 所有**类别**的每一个统计指标值的算数平均值
  - Precision = $\frac{1}{n}\sum_{i=1}^{n} P_i$
- 微平均（Micro-averaging）
  - 所有**样本**的每一个统计指标值的算数平均值
  - Precision = $\frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}$
  - 受样本多的类别影响很大

常用方法


早期方法
基于规则的方法
基于字典的方法
深度学习方法
RNN-CRF
CNN-CRF
传统机器学习方法
HMM
MEMM
CRF
近期
注意力模型
迁移学习
半监督学习

早期方法

基于规则+词典

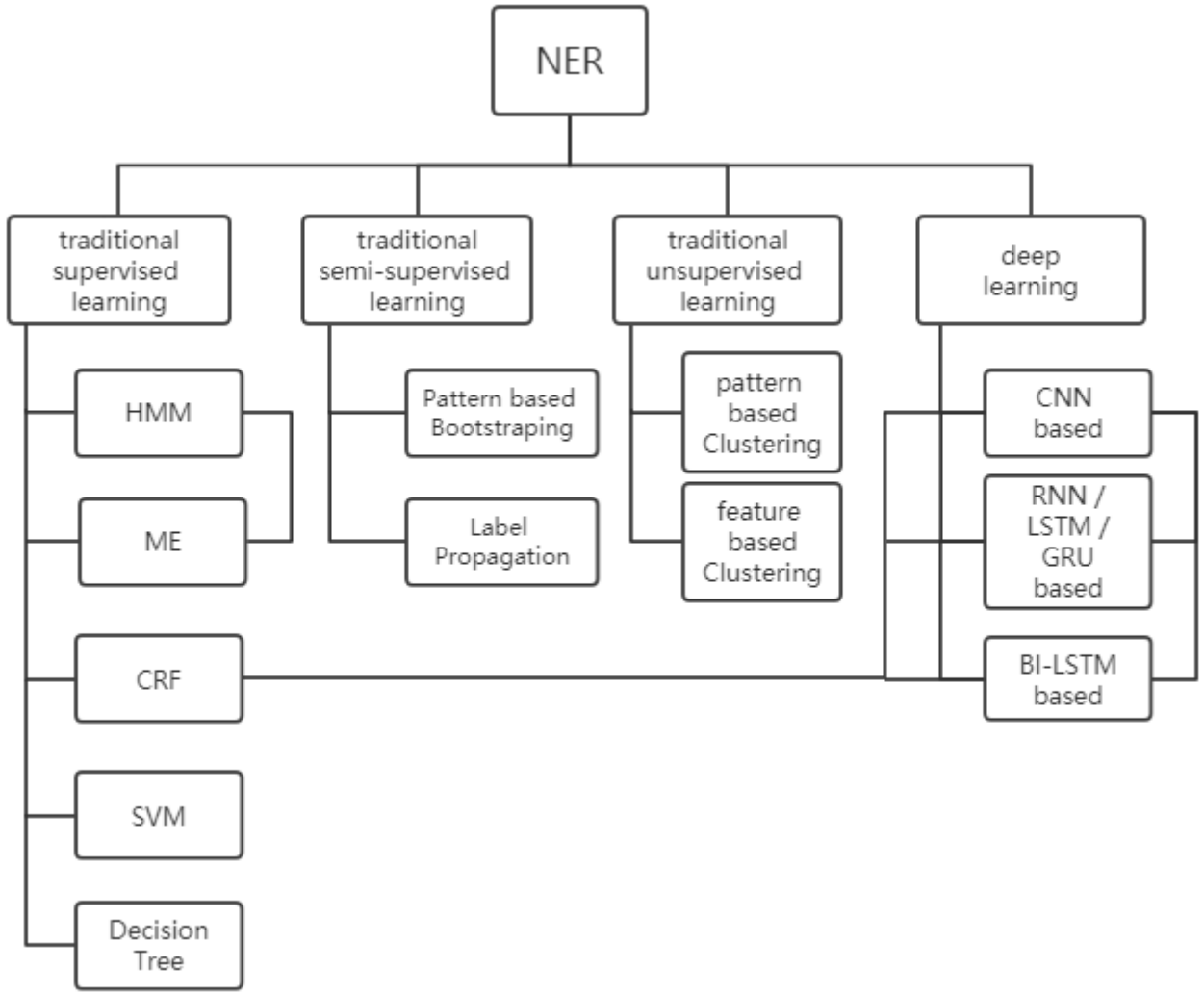
- 例句：我来到北京天安门。
- 分词：我 来到 北京 天安门。依赖词典
- 词性标注：O O S E
- 规则：S+O*E+
- 在文本中进行匹配，找出类似的实体

早期方法

- 总结：采用语言学专家手工构造规则模板，选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词(如尾字)、中心词等方法，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。

- 缺点：依赖于具体语言、领域和文本风格，编制过程耗时且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则

机器学习方法


NER
traditional supervised learning
traditional semi-supervised learning
traditional unsupervised learning
deep learning
HMM
ME
CRF
SVM
Decision Tree
Pattern based Bootstraping
Label Propagation
pattern based Clustering
feature based Clustering
CNN based
RNN / LSTM / GRU based
BI-LSTM based

传统机器学习方法-MEMM

特征工程+模型

MEMM（最大熵马尔科夫模型）

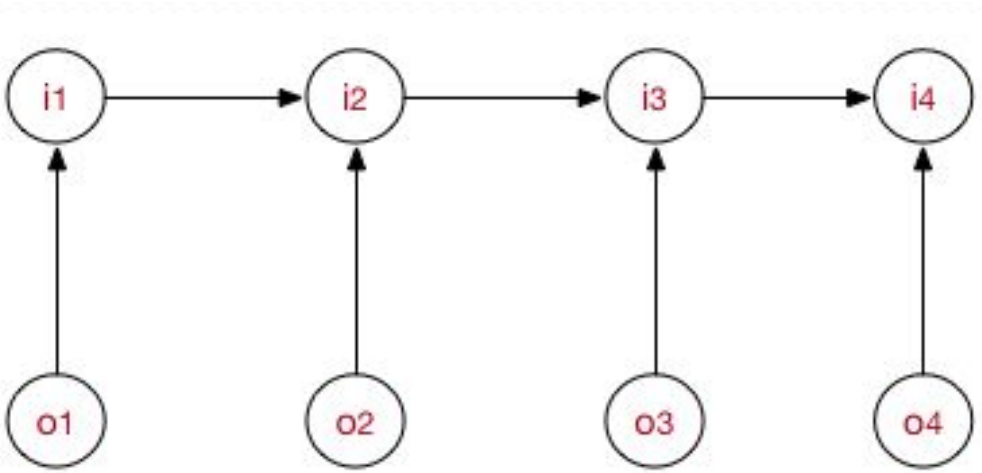
- 概率有向图模型
- 判别式模型，直接对条件概率建模

目标函数：

$$P(I|O)=\prod_{t=1}^{n}P(i_i|i_{i-1},o_i),i=1,\cdots,n$$

$$P(i|i^{'},o)=\frac{1}{Z(o,i^{'})}exp(\sum_{a})\lambda_a f_a(o,i)$$

并且，$P(i|i',o)$概率通过最大熵分类器建模



传统机器学习方法-MEMM

- 公式展开后：

$$P(I|O)=\prod_{t=1}^{n}\frac{exp(\sum_a)\lambda_a f_a(o,i)}{Z(o,i_{i-1})},i=1,\cdots,n$$

$Z(\mathrm{o},i')$是归一化因子；$f_a(o,i)$是特征函数，这个函数是需要去定义的；λ是特征函数的权重，是未知参数，需要从训练阶段学习而得。

- 特征函数定义举例：

$$f_a(o,i)=\begin{cases}1,\text{满足特定条件}\\0,\ \text{otherwise}\end{cases}$$

- 训练完成后通过维特比算法进行解码

特征工程

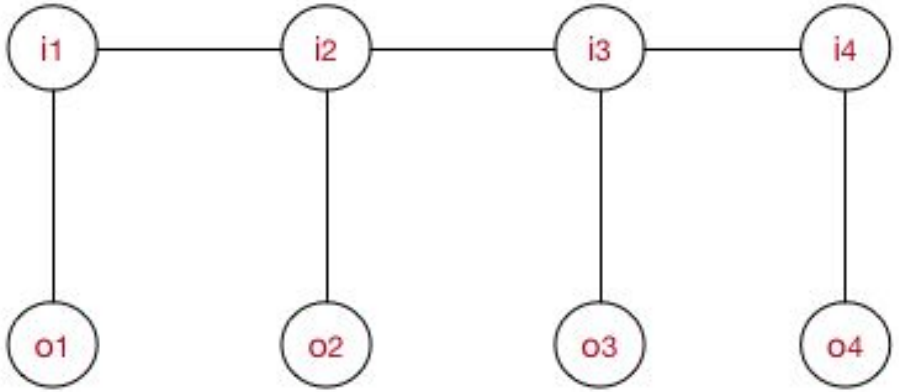
特征类别	特征
Morphological	N-gram character, N-gram word, suffixes, prefixed
Orthographic	Capitalization, symbols
Linguistic	Lemmatization, stemming, POS, chunking, syntactic parsing
Context	Windows, conjunctions
Domain knowledge	Lexicons, gazetteer, existing NLP tools

```
features = {
        'bias': 1.0,
        'word[-3:]': word[-3:],
        'word[-2:]': word[-2:],
        'word[:2]': word[:2],
        'word[:3]': word[:3],
        'word.lower()': word.lower(),
        'word.isupper()': word.isupper(),
        'word.istitle()': word.istitle(),
        'word.isdigit()': word.isdigit(),
        'word.isdigit_char()': self.isDigitAndChar(word),
        'word.stemmer()': porter_stemmer.stem(word),
        'word.lemmatizer()': wordnet_lemmatizer.lemmatize(word),
        'word.len()': len(word),
        'postag': postag,
        'postag[:2]': postag[:2],
    }
```

传统机器学习方法-CRF

- CRF(条件随机场)
  - 给定随机变量X(观测变量o)下，随机变量Y(隐变量i)的马尔科夫随机场
  - 概率无向图模型、判别式模型
- 广义CRF

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$$



- 概率无向图联合概率分布(因子分解)

$$P(Y|X) = \frac{1}{Z(x)} \prod_c \psi_c(Y_c|X)$$

- 线性链CRF满足$\mathrm{P}(I_i|\mathrm{O}, I_1, \ldots, I_n) = \mathrm{P}(I_i|\mathrm{O}, I_{i-1}, I_{i+1})$

传统机器学习方法-CRF

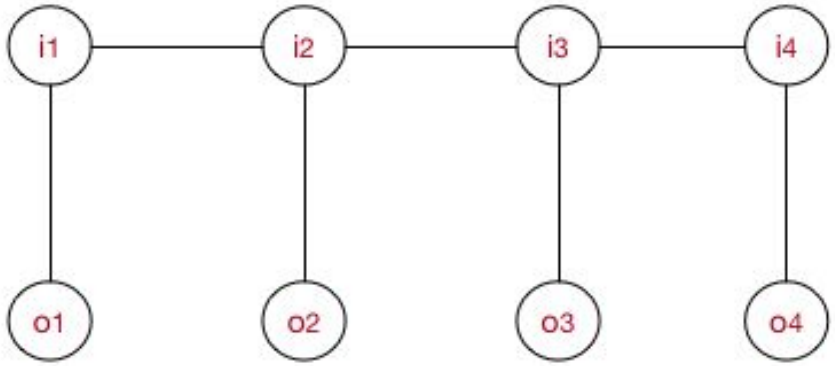
- 目标函数

$$P(I|O)=\frac{1}{Z(O)}\prod_i \psi_i(I_i|O)=\frac{1}{Z(O)}\prod_i e^{\sum_k \lambda_k f_k(O,I_{i-1},I_i,i)}=\frac{1}{Z(O)}e^{\sum_i\sum_k \lambda_k f_k(O,I_{i-1},I_i,i)}$$

- $Z(O)$是归一化因子;$f_k(o,I_{i-1},I_{i+1},i)$是特征函数，这个函数是需要去定义的;λ是特征函数的权重
- 特征函数分解为转移特征和状态特征

$$\sum_i^T\sum_j^J \lambda_j t_j(O,I_{i-1},I_i,i)+\sum_i^T\sum_l^L \mu_l s_l(O,I_i,i)$$

- 转移特征可以理解为最大团中对边的分离
- 状态特征可以理解为最大团中对节点的分离



传统机器学习方法–CRF

- 目前我们的结果, pure CRF

	precision	recall	f1	support
I-LOC	0.860	0.871	0.866	1919
I-MISC	0.765	0.762	0.764	909
I-ORG	0.809	0.774	0.791	2491
I-PER	0.864	0.903	0.883	2773
avg/total	0.833	0.838	0.835	8112

神经网络方法(1)

- Collober et.al(2011)
- NN/CNN+hand-drafted features
- 两种方法
  - Window approach network
  - Sentence approach network
- Window approach
- 动机：某个单词的标记主要取决于它的邻居单词
- k_{sz} 窗口大小，$d_{wrd} * k_{sz}$ 输入维度

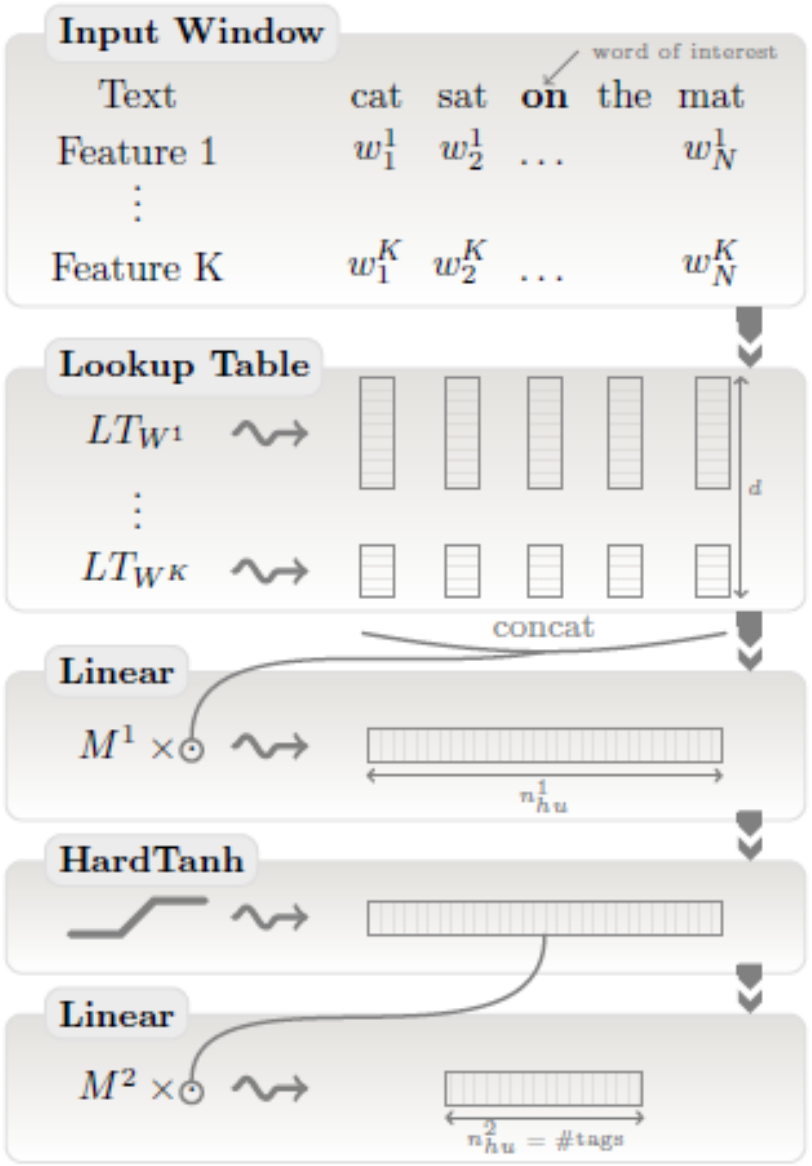


Figure 1: Window approach network.

神经网络方法(1)

- 动机：window方法虽然在大多数NLP任务中表现较好，但受限于窗口大小，在Semantic Role Labeling任务中失败。遂将整个句子作为输入，用卷积提取local feature
- 训练

$$s([x]_1^T, [i]_1^T, \tilde{\theta}) = \sum_{t=1}^{T} \left([A]_{[i]_{t-1}, [i]_t} + [f_\theta]_{[i]_t, t} \right) .$$

$$\log p([y]_1^T \mid [x]_1^T, \tilde{\theta}) = s([x]_1^T, [y]_1^T, \tilde{\theta}) - \underset{\forall [j]_1^T}{\text{logadd}} s([x]_1^T, [j]_1^T, \tilde{\theta}) .$$

- A是标签转移矩阵，f代表NN的输出得分

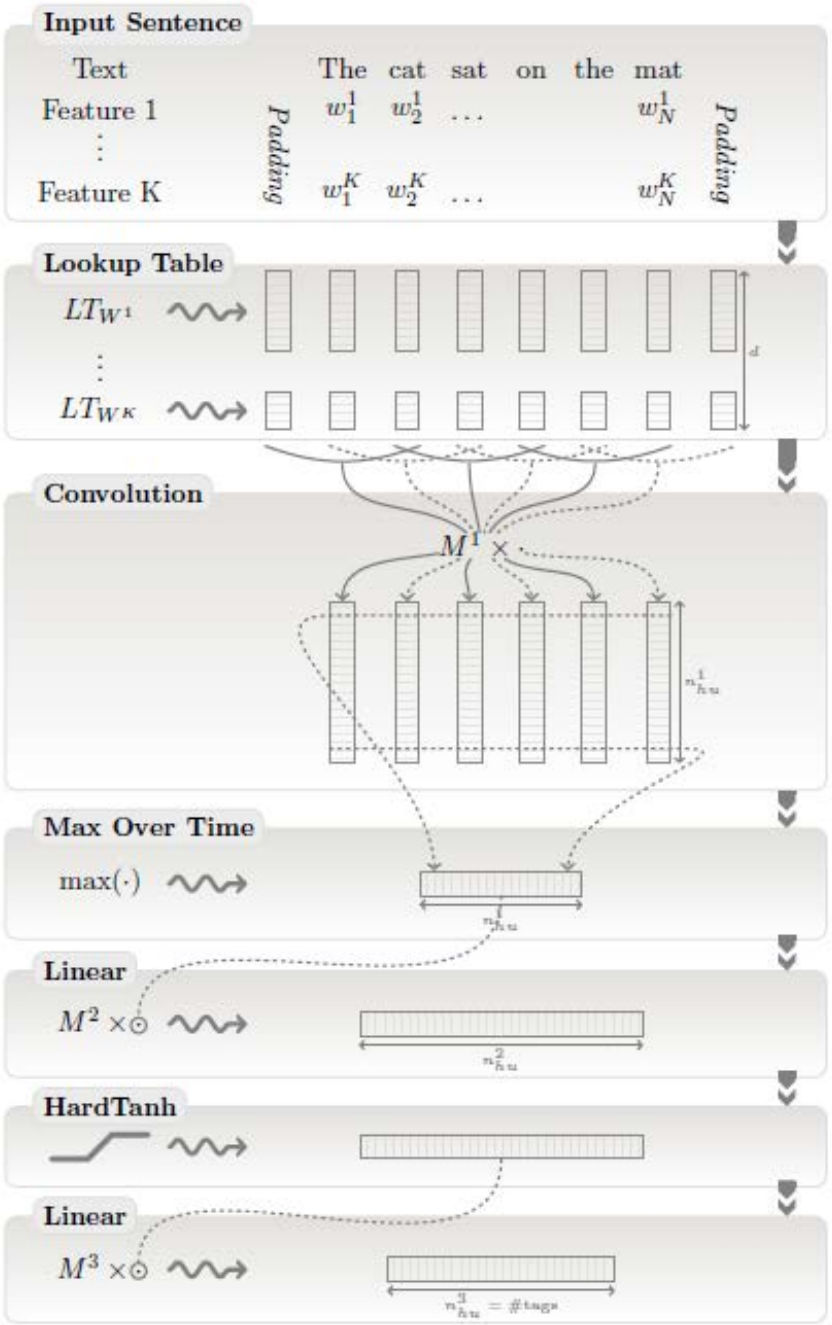


Figure 2: Sentence approach network.

神经网络方法(2)

- Huang et al.(2015)

Table 2: Comparison of tagging performance on POS, chunking and NER tasks for various models.

		POS	CoNLL2000	CoNLL2003
Random	Conv-CRF (Collobert et al., 2011)	96.37	90.33	81.47
	LSTM	97.10	92.88	79.82
	BI-LSTM	97.30	93.64	81.11
	CRF	97.30	93.69	83.02
	LSTM-CRF	**97.45**	93.80	84.10
	BI-LSTM-CRF	97.43	**94.13**	**84.26**
Senna	Conv-CRF (Collobert et al., 2011)	97.29	94.32	88.67 (89.59)
	LSTM	97.29	92.99	83.74
	BI-LSTM	97.40	93.92	85.17
	CRF	97.45	93.83	86.13
	LSTM-CRF	97.54	94.27	88.36
	BI-LSTM-CRF	**97.55**	**94.46**	**88.83 (90.10)**

神经网络方法(2)

- Huang et al.(2015)
- BI-LSTM+CRF+hand-drafted features
- 用了各种feature+Gazetter
  - Spelling features
  - Context features
  - Word embedding
- 训练

$$s([x]_1^T, [i]_1^T, \tilde{\theta}) = \sum_{t=1}^{T} \left([A]_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t} \right) .$$

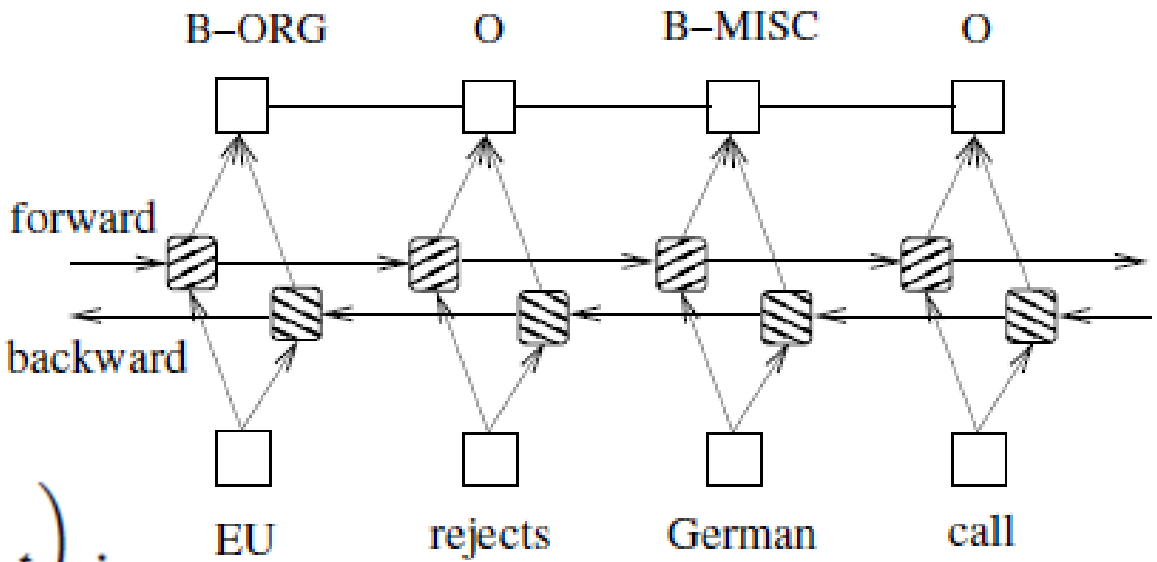
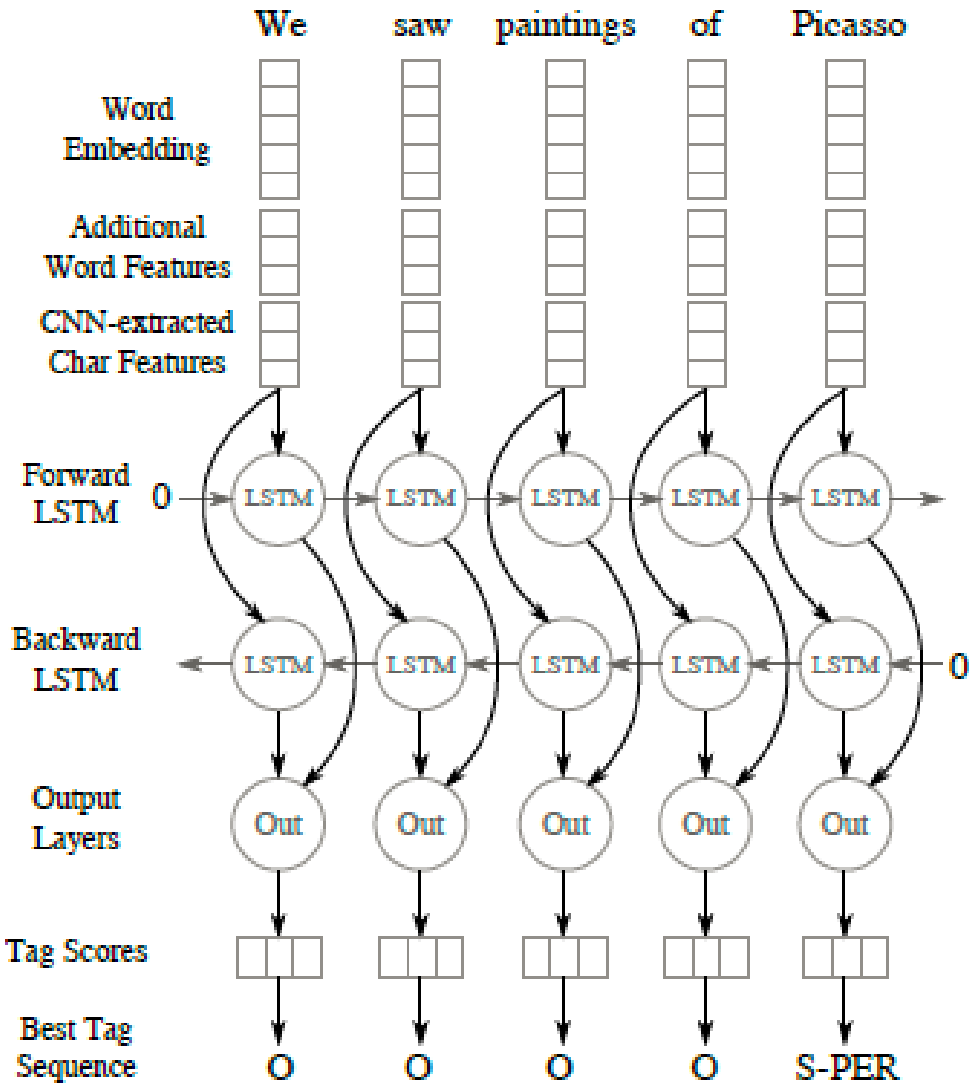


Figure 7: A BI-LSTM-CRF model.

神经网络方法(3)

- Chiu and Nichols(2015)
- CNN(character+word-level)+BI-LSTM
- CNN提取character-level的特征+4-dim lookup T (upper,lower,punctuation,other)
- Word-embedding
- Capitalization feature+lexicons
- BI-LSTM捕捉句子历史和未来的信息

$$s([x]_1^T, [i]_1^T, \tilde{\theta}) = \sum_{t=1}^{T} \left([A]_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t} \right) .$$



神经网络方法(3)

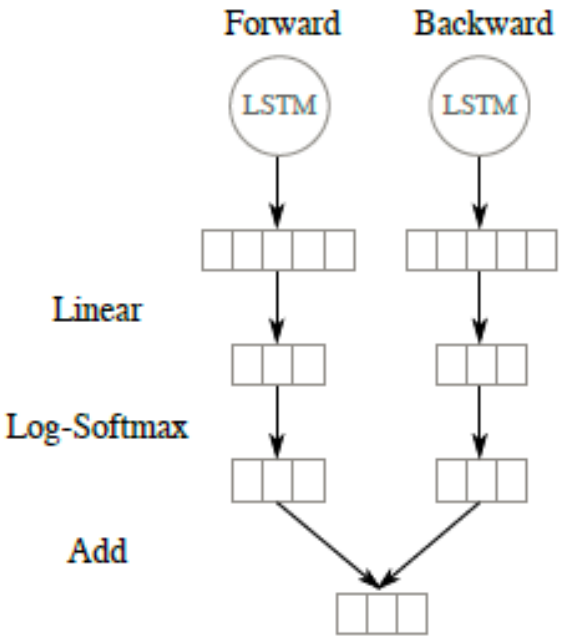


Figure 3: The output layers ("Out" in Figure 1) decode output into a score for each tag category.

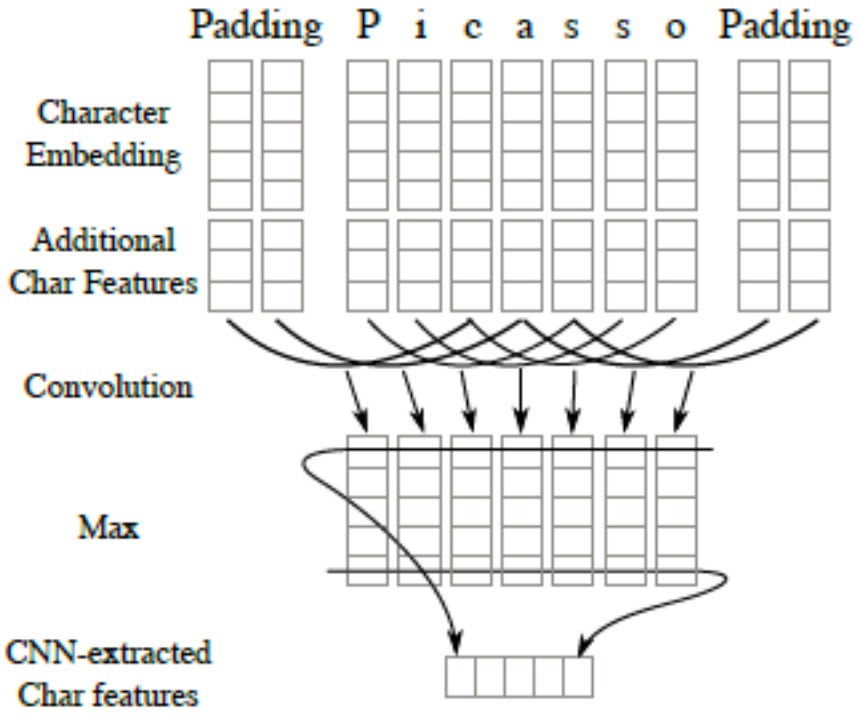


Figure 2: The convolutional neural network extracts character features from each word. The character embedding and (optionally) the character type feature vector are computed through lookup tables. Then, they are concatenated and passed into the CNN.

神经网络方法(4)

- Lample et.al(2016)
- BI-LSTM(char+word)+CRF
- 使用bi-lstm做character embedding 提取prefix and suffix information
- Dropout training
- Embedding
  - skip-n-gram (Ling et al., 2015a)
  - a variation of word2vec

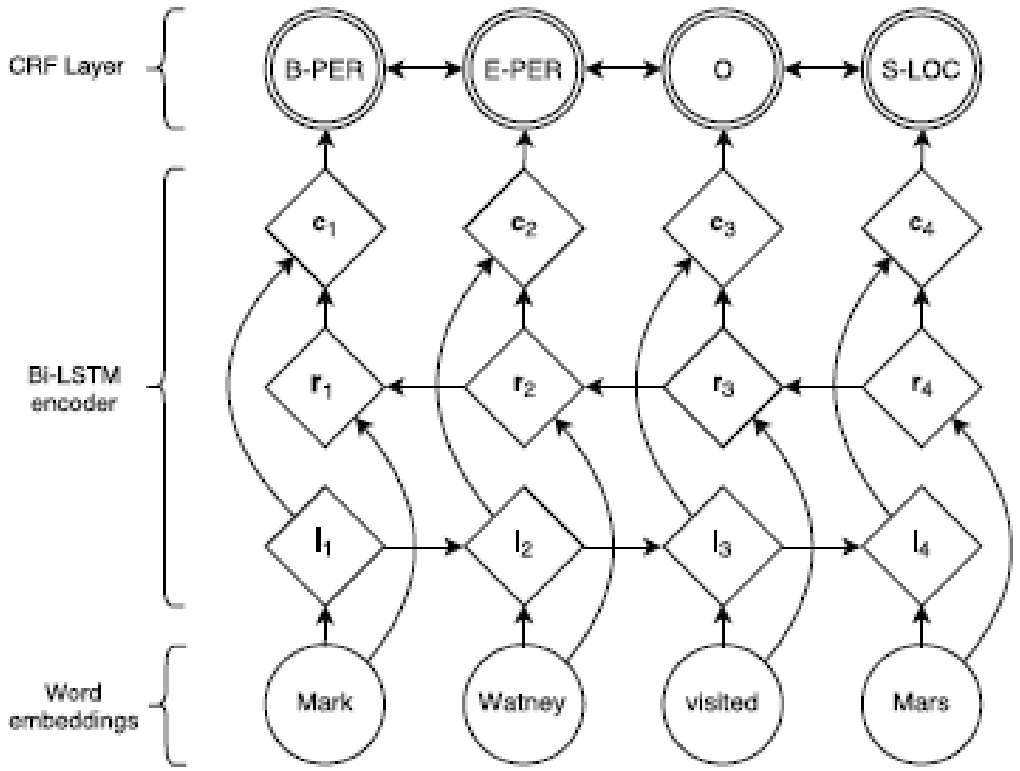


Figure 1: Main architecture of the network. Word embeddings are given to a bidirectional LSTM. $\mathbf{l}_i$ represents the word i and its left context, $\mathbf{r}_i$ represents the word i and its right context. Concatenating these two vectors yields a representation of the word i in its context, $\mathbf{c}_i$.

神经网络方法(4)

- Lample et.al(2016)
- BI-LSTM(char+word)+CRF

- 目标函数最大化对数概率

$$\log(p(\mathbf{y}|\mathbf{X})) = s(\mathbf{X},\mathbf{y}) - \log\left(\sum_{\tilde{\mathbf{y}}\in\mathbf{Y_X}} e^{s(\mathbf{X},\tilde{\mathbf{y}})}\right)$$
$$= s(\mathbf{X},\mathbf{y}) - \underset{\tilde{\mathbf{y}}\in\mathbf{Y_X}}{\text{logadd}}\, s(\mathbf{X},\tilde{\mathbf{y}}), \quad (1)$$

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

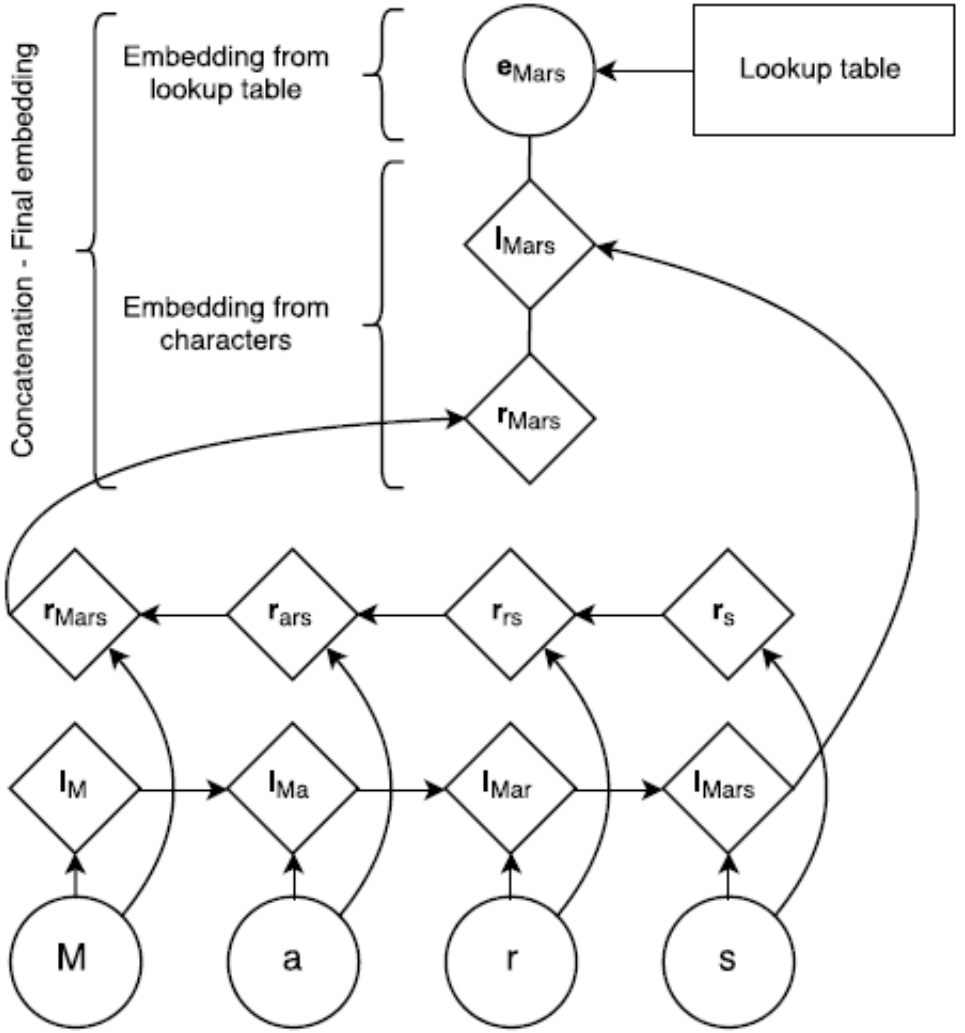
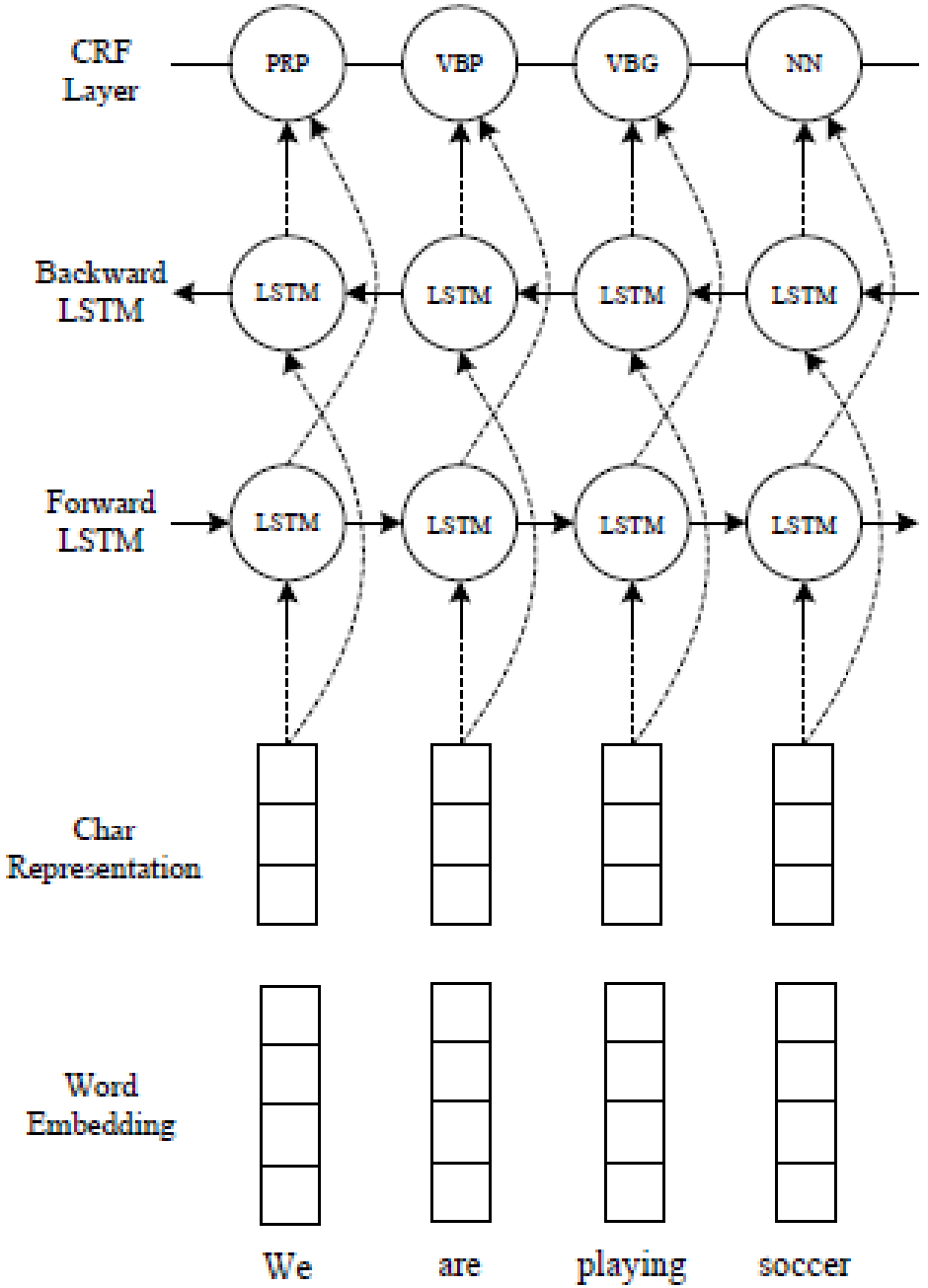


Figure 4: The character embeddings of the word "Mars" are given to a bidirectional LSTMs. We concatenate their last outputs to an embedding from a lookup table to obtain a representation for this word.

神经网络方法(5)

- Ma and Hovy(2016)
- End-to-end Sequence Labeling
- Bi-directional LSTM-CNNs-CRF
- 没有特征工程，不需要数据预处理
- CNN for Character-level Representation
- BI-LSTM for past and future context
- CRF for sequence labeling



神经网络方法(5)

- Ma and Hovy(2016)
- Word embedding: GloVe 100-dim
- Charcater embedding:随机初始化某个区间
- Weight Matrices： 随机初始化某个区间
- Bias Vectors： 0
- Optimizer: SGD
- Early stopping
- Fine Tuning the embeddings
- Dropout training
- Random Search 其他超参

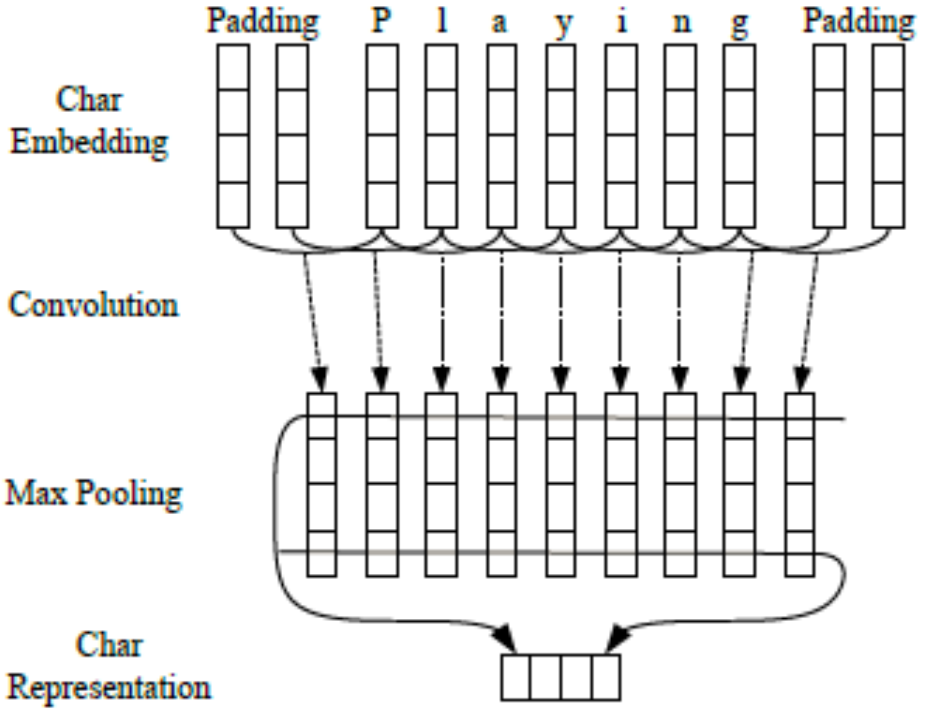


Figure 1: The convolution neural network for extracting character-level representations of words. Dashed arrows indicate a dropout layer applied before character embeddings are input to CNN.

最新进展

- Attention-based
  - Attending to Characters in Neural Sequence Labeling Models(Rei et al. 2016)
- Transfer learning
  - Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks (Yang et al. 2017)
- Semi-supervised
  - Semi-supervised sequence tagging with bidirectional language models(Peters et al. 2017)
- External knowledge
  - Joint Named Entity Recognition and Disambiguation(Luo et al. 2015)
- Best for Now
  - Contextual String Embeddings for Sequence Labeling(Akbik et al. 2018)

数据集

- CoNLL 2003 dataset
- MUC-7 dataset,
- BP Track NER dataset

		CoNLL 2003
Training	sentence # token #	14987 204567
Validation	sentence # token #	3466 51578
Test	sentence # token #	3684 46666
Label		9

模型效果

Table 1. NER evaluation results of previous state-of-art method in CoNLL2003 dataset

author	model	F1
Huang et al.(2015)	CRF	83.02
McCallum and Li(2003)	CRF+lexicons	84.04
Collobert et al.(2011)	CNN+NN+hand-drafted features	89.59
Huang et al.(2015)	BI-LSTM+CRF+ hand-drafted features	90.10
Chiu and Nichols(2015)	CNN(char+word)+BI-LSTM	90.77
Luo et.al(2015)	CRF+entity linking	91.20
Lample et.al(2016)	BI-LSTM(char+word)+CRF	90.94
Ma and Hovy(2016)	CNN(char+word)+BI-LSTM+CRF	91.21
Yang et.al(2017)	Transfer learning	91.26
Peters et al. (2018)	BiLSTM-CRF+ELMo	92.22
Akbik et al.(2018)	Contextual embedding+BI-LSTM +CRF(char + word)	93.09

系统效果

Table 2. NER evaluation results of state-of-art NER systems in CoNLL2003 dataset

System	Precision	Recall	F1
Stanford	95.1	78.3	85.9
UIUC	91.2	90.5	90.8
Nerel	86.8	89.5	88.2
JERL	91.5	91.4	91.2

引用

- R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research (JMLR).
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging.CoRR, abs/1508.01991.
- Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In Proceedings of NAACL-2016, San Diego, California, USA, June.
- Xuezhe Ma and Eduard Hovy. 2015. Efficient inner-to-outer greedy algorithm for higher-order labeled dependency parsing. In Proceedings of the EMNLP-2015, pages 1322–1328, Lisbon, Portugal, September.
- ……