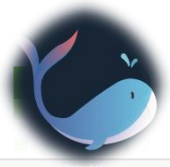
The review of Named-entity recognition (NER)

团队：Datawhale 深度学习团队
汇报人： 杨开漠
2019.11.01

目录

相关介绍

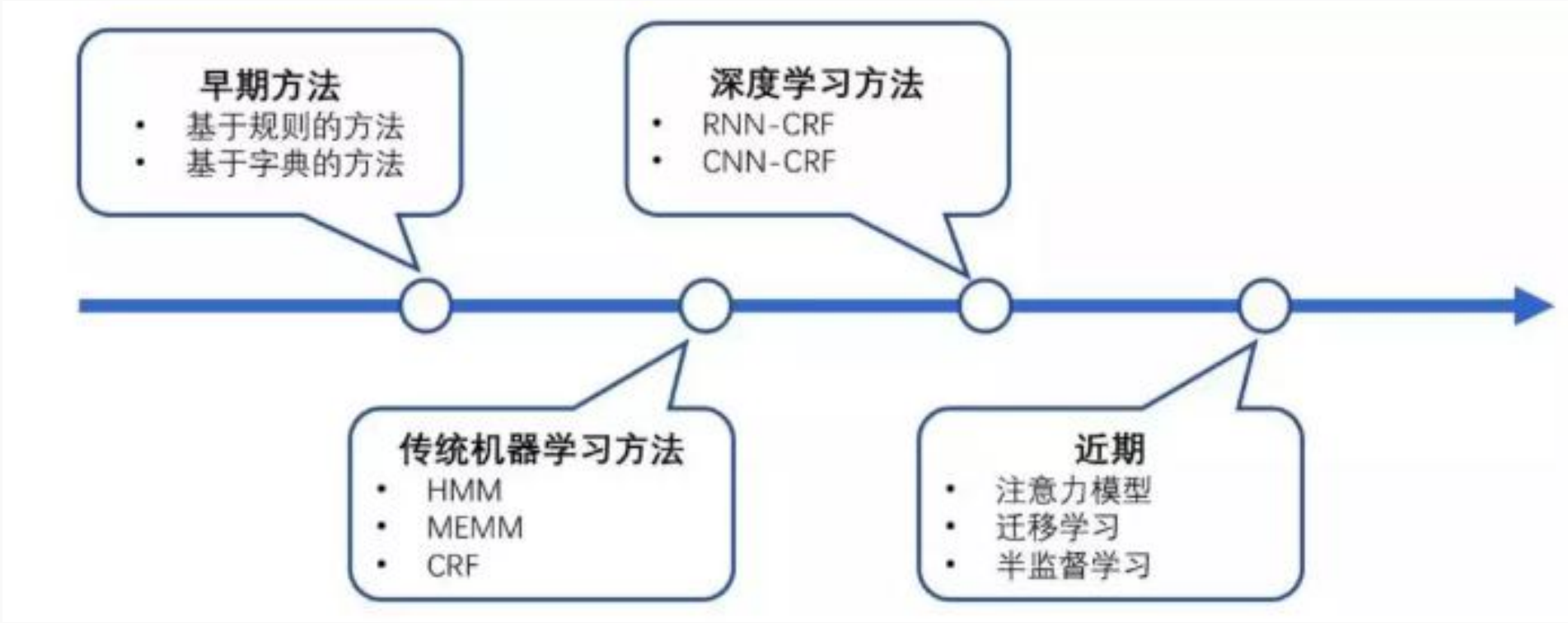
01 相关介绍



先来看看维基百科上的定义：Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

命名实体识别（Named Entity Recognition，简称NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。简单的讲，就是识别自然文本中的实体指称的边界和类别。

01 相关介绍

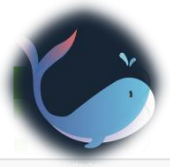


Tokens	IO	BIO	BIOES	...
5	I-TIME	B-TIME	B-TIME	
号	I-TIME	I-TIME	E-TIME	
特	I-PER	B-PER	B-PER	
朗	I-PER	I-PER	I-PER	
普	I-PER	I-PER	E-PER	
访	O	O	O	
华	I-LOC	B-LOC	S-LOC	
。	O	O	O	

数据集名称	类型	简要介绍
CCKS 2017	医疗	中文的电子病例测评相关的数据
NLPCC	口语	开放的任务型对话系统中的口语理解评测
bsson	新闻	公司提供的数据集,包含人名、地名、机构名、专有名词
dh_msra	新闻	中文命名实体识别标注数据（包括地点、机构、人物）
weibo	微博	微博常用语

BiLSTM-CRF

02 BiLSTM-CRF

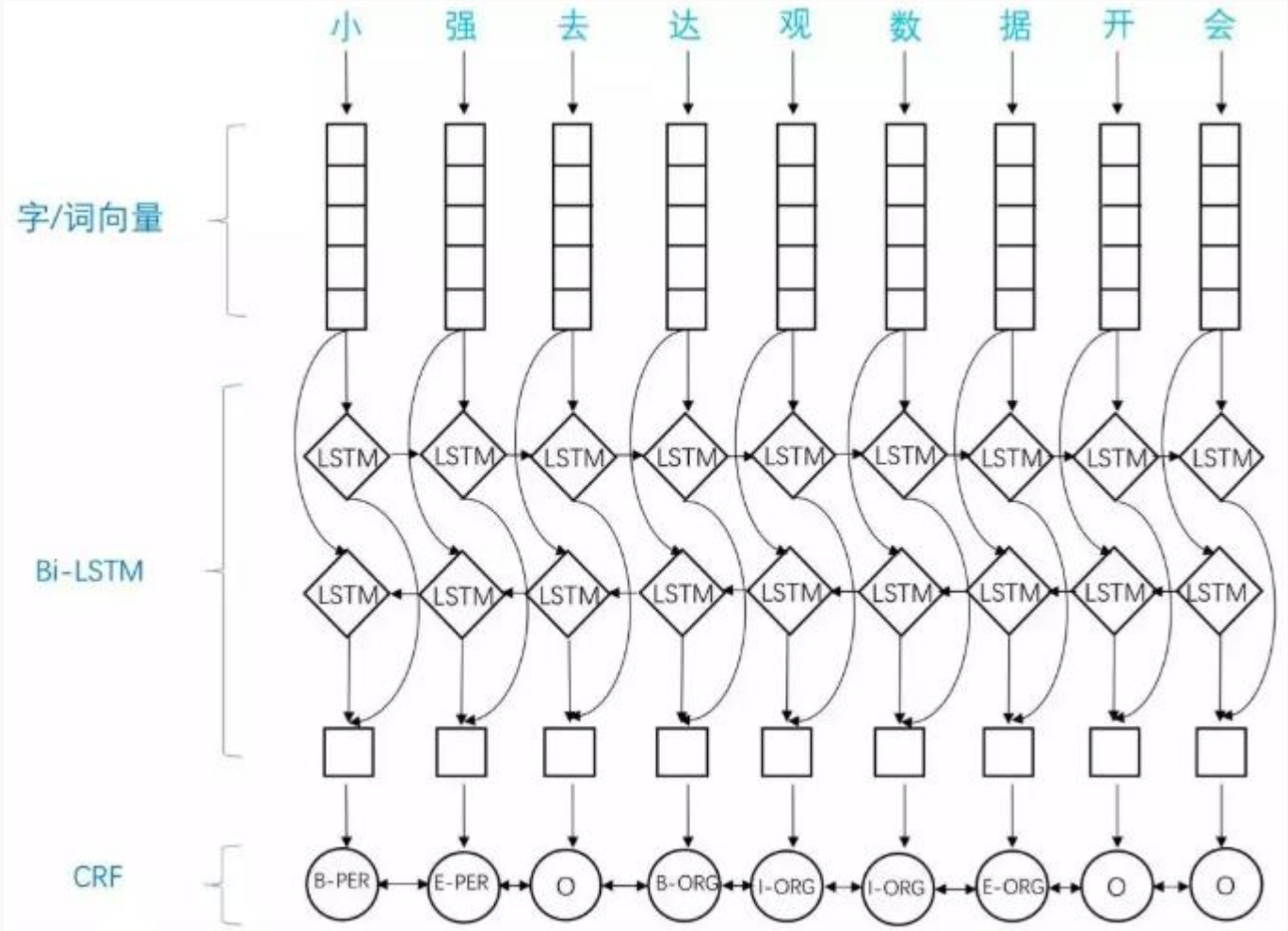


虽然传统的机器学习方法（HMM、MEMM、CRF）能够取得不错的效果，但是该方法需要依赖特征工程。

[1] Bidirectional LSTM-CRF Models for Sequence Tagging

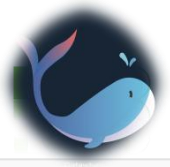
02 BiLSTM-CRF

目录



结构示意图

[1] Bidirectional LSTM-CRF Models for Sequence Tagging

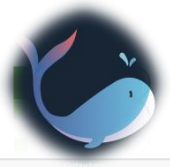


这种方法使得模型的训练成为一个端到端的过程，而非传统的pipeline，不依赖于特征工程，是一种数据驱动的方法，但网络种类繁多、对参数设置依赖大，模型可解释性差。

[1] Bidirectional LSTM-CRF Models for Sequence Tagging

IDCNN-CRF

03 IDCNN-CRF



传统 CNN 用于文本处理时所存在的劣势：只能获得局部信息

传统 RNN 用于文本处理时所存在的劣势：由于其结构，所引起的速度问题

[2] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions(2017)

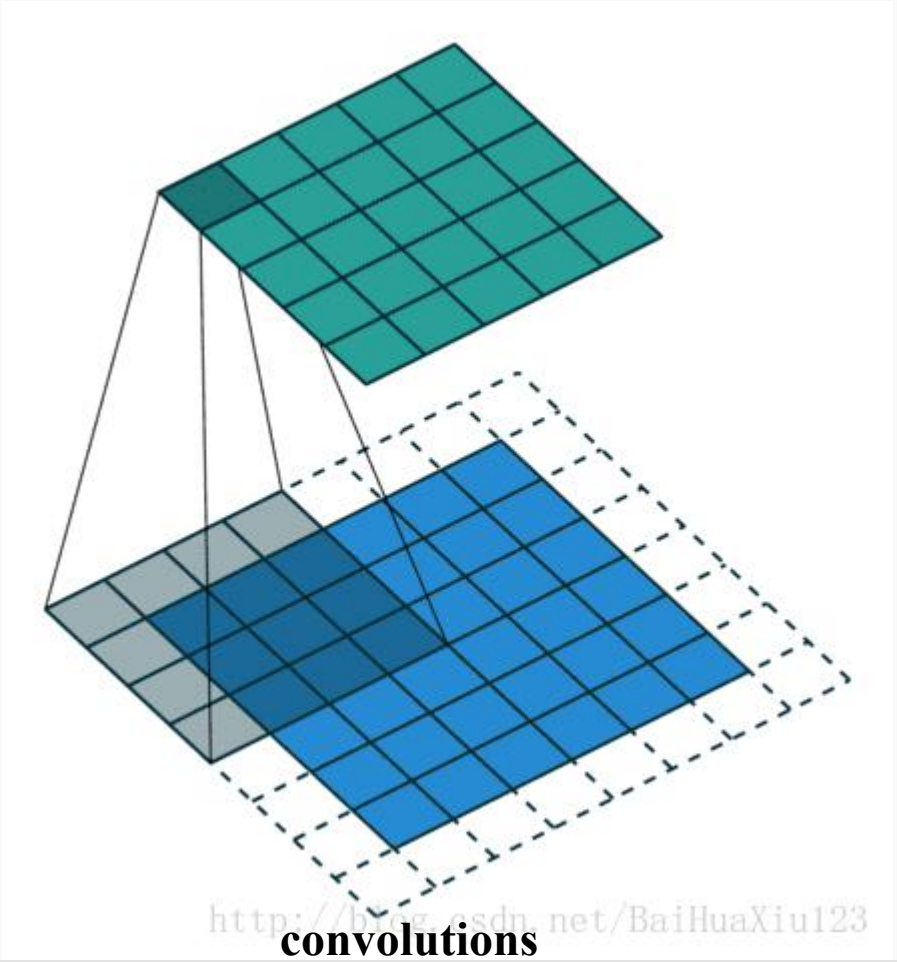
03 IDCNN-CRF

目录

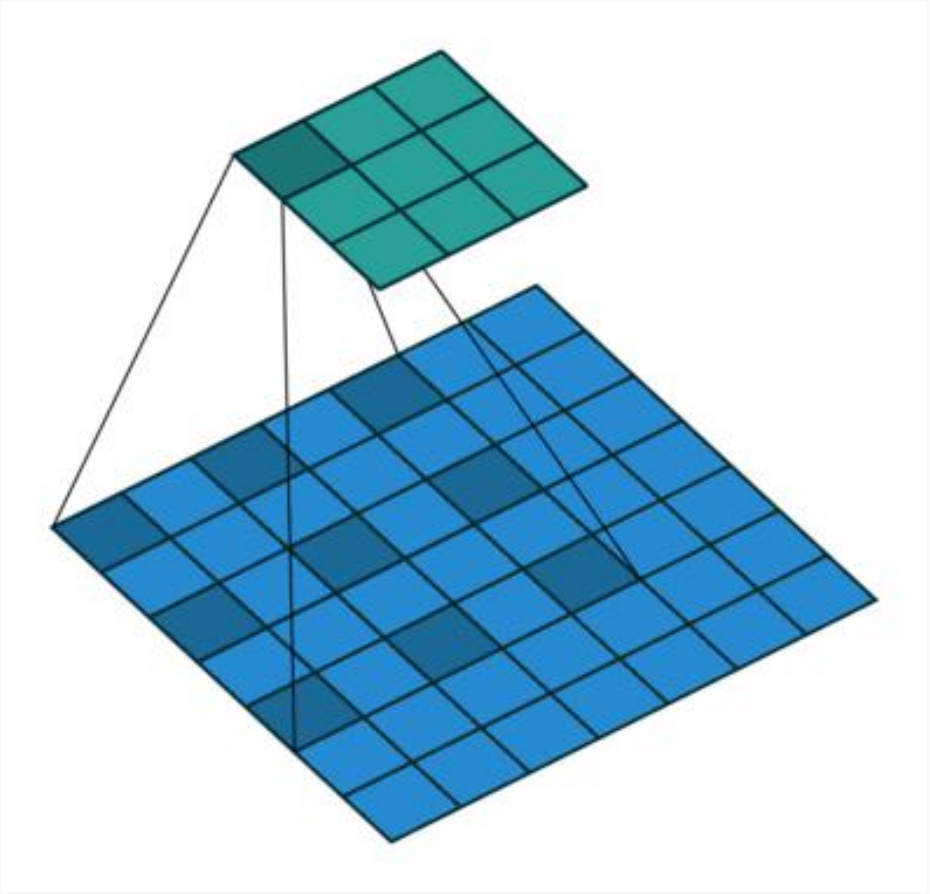
01 论文动机

02 方法介绍

03 个人点评



convolutions



dilated convolutions

[2] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions(2017)

03 IDCNN-CRF

目录

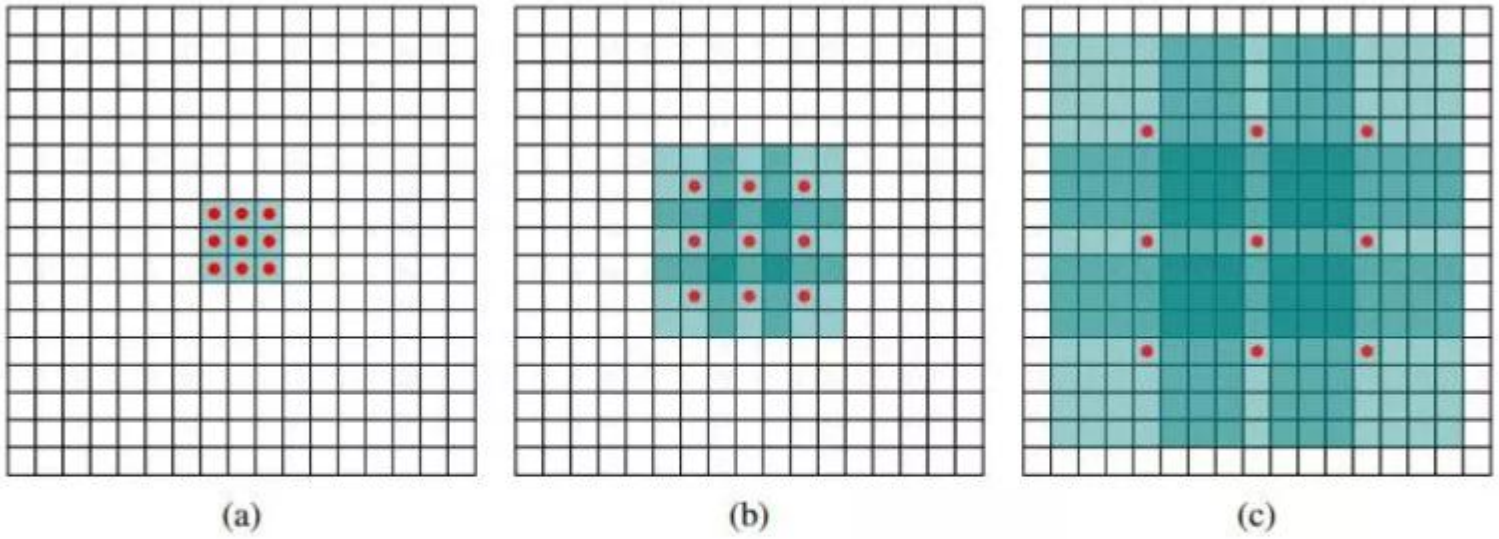


Figure 1: Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage. (a) F_1 is produced from F_0 by a 1-dilated convolution; each element in F_1 has a receptive field of 3×3. (b) F_2 is produced from F_1 by a 2-dilated convolution; each element in F_2 has a receptive field of 7×7. (c) F_3 is produced from F_2 by a 4-dilated convolution; each element in F_3 has a receptive field of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

[2] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions(2017)

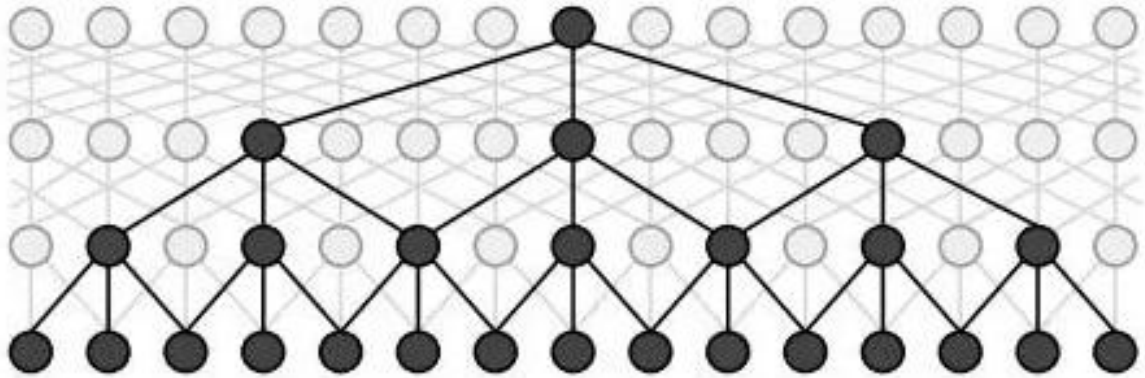
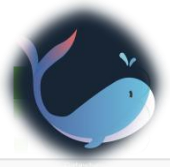


Figure 1: A dilated CNN block with maximum dilation width 4 and filter width 3. Neurons contributing to a single highlighted neuron in the last layer are also highlighted.

[2] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions(2017)

03 IDCNN-CRF

目录



01 论文动机

02 方法介绍

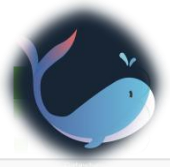
03 个人点评

CNN base方法利用空洞卷积+多层的方式实现提取整句的功能，同时也能实现并行计算加速。

[2] Fast and Accurate Entity Recognition with Iterated Dilated Convolutions(2017)

Lattice LSTM

04 Lattice LSTM



LSTM-CRF模型分析了集中变种模型的优缺点：

基于词的LSTM-CRF：一般的pipeline是对文本进行分词之后embedding后输入深度网络预测序列中单词的类别标记。但是这样的话会受限于分词那一步的表现，也就是说如果分词过程效果不好的话，会进一步影响整个NER模型的误差。而对于NER任务中，许多词都是OOV；

基于字的LSTM-CRF：那么把词输入改为字输入是不是会有所改进呢？因为字向量可以完美克服上述分词过程引入的误差。但是如果单纯采用字向量的话会丢失句子中词语之间的内在信息。

[3] Chinese NER Using Lattice LSTM

04 Lattice LSTM

目录

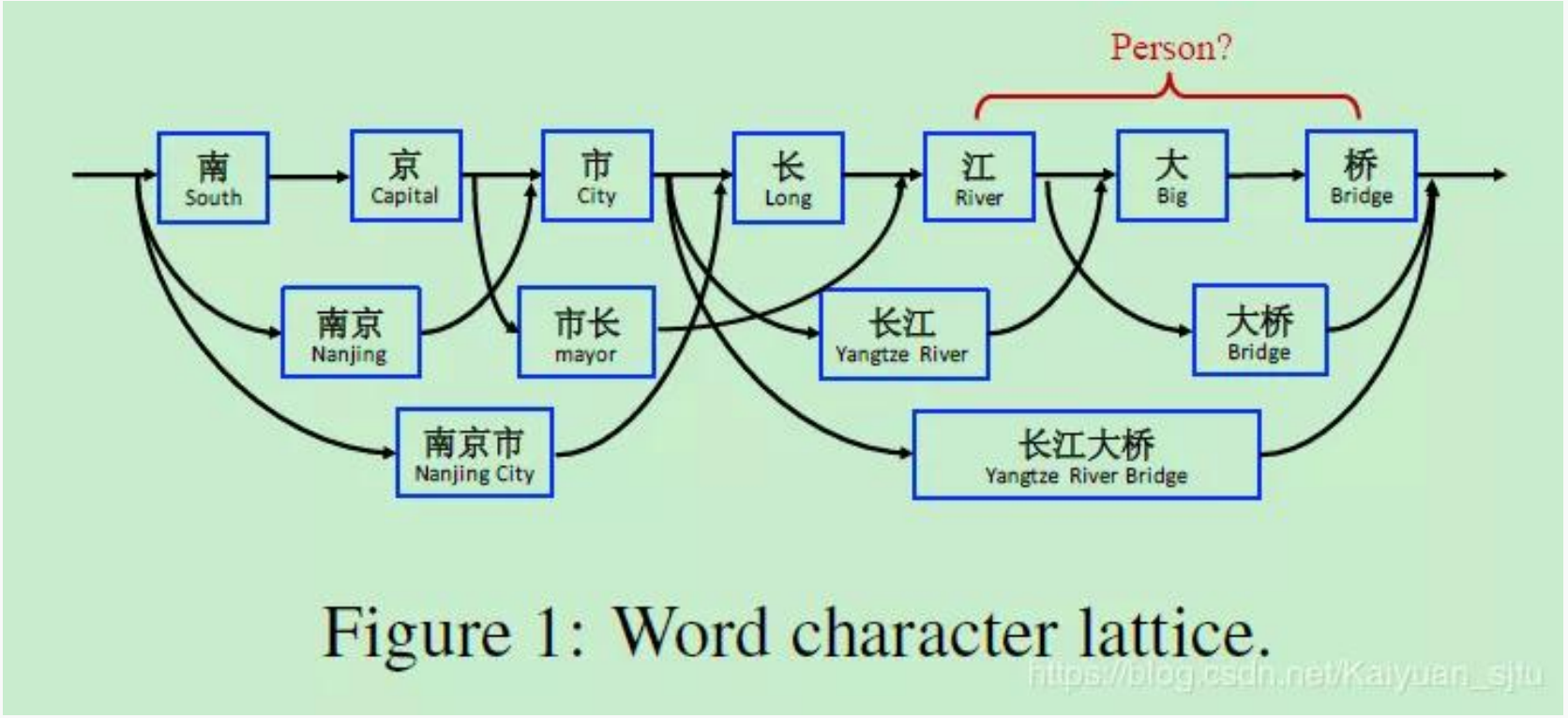


Figure 1: Word character lattice.

https://blog.csdn.net/Kaiyuan_sjtu

[3] Chinese NER Using Lattice LSTM

04 Lattice LSTM

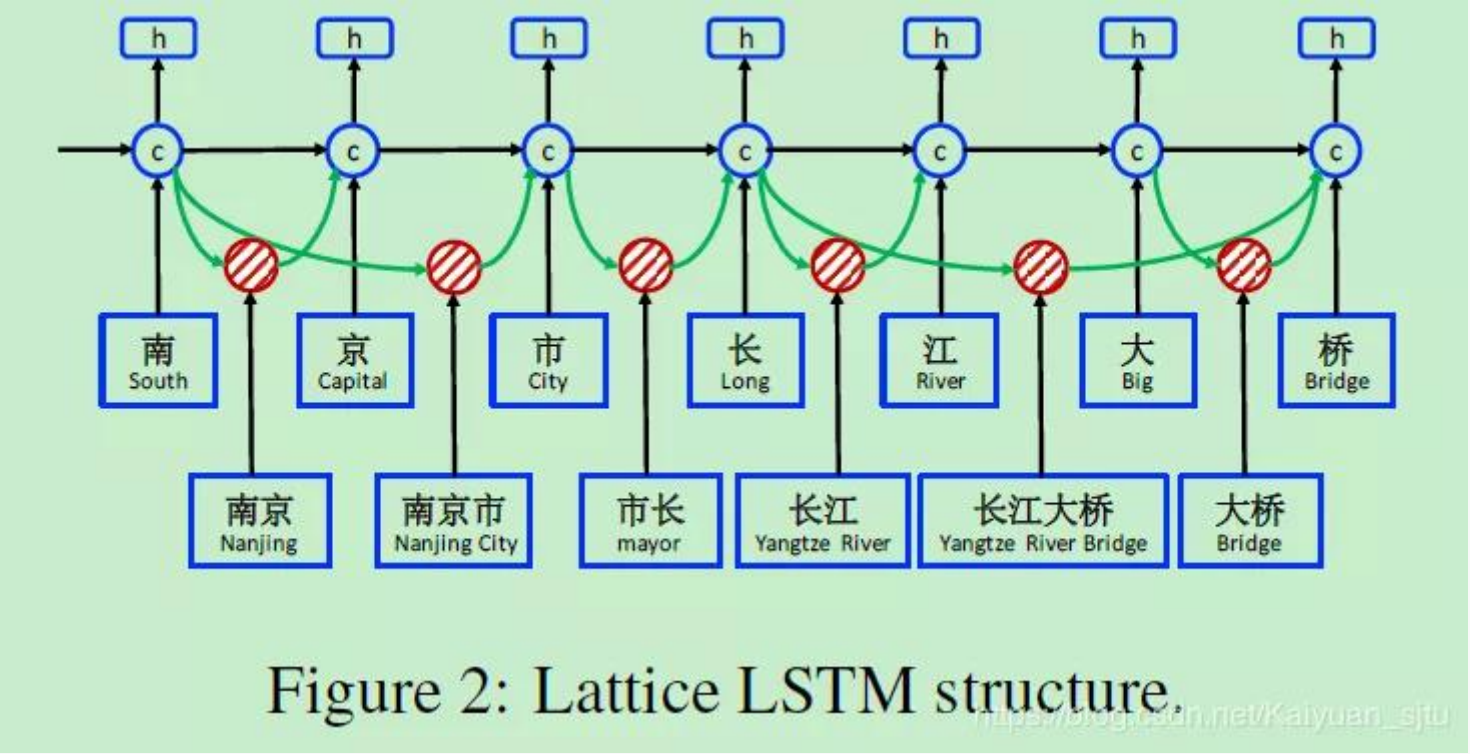


Figure 2: Lattice LSTM structure.

[3] Chinese NER Using Lattice LSTM

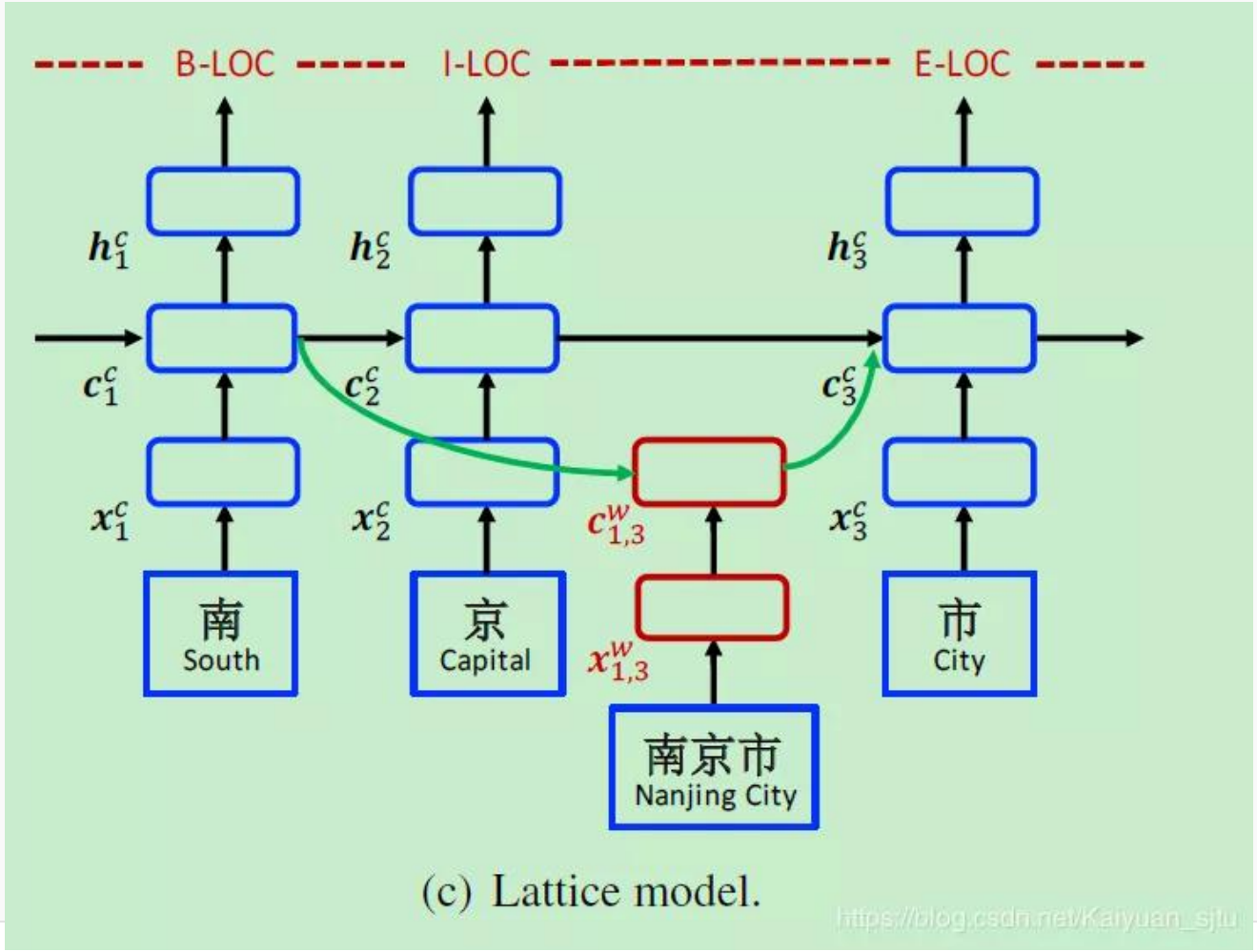
04 Lattice LSTM

目录

01 论文动机

02 方法介绍

03 个人点评

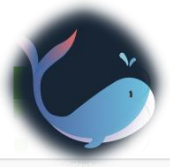


(c) Lattice model.

https://blog.csdn.net/Kaiyuan_sjtu

[3] Chinese NER Using Lattice LSTM

04 Lattice LSTM

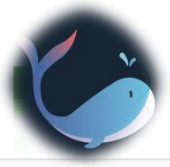


该论文的方法就是由两套LSTM子结构分别是基于字的和基于词的，然后得到基于词的状态输出之后将其加入到基于字的结构中输出预测。之后就是跟其他模型一样套上一层CRF层。

[3] Chinese NER Using Lattice LSTM

Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism

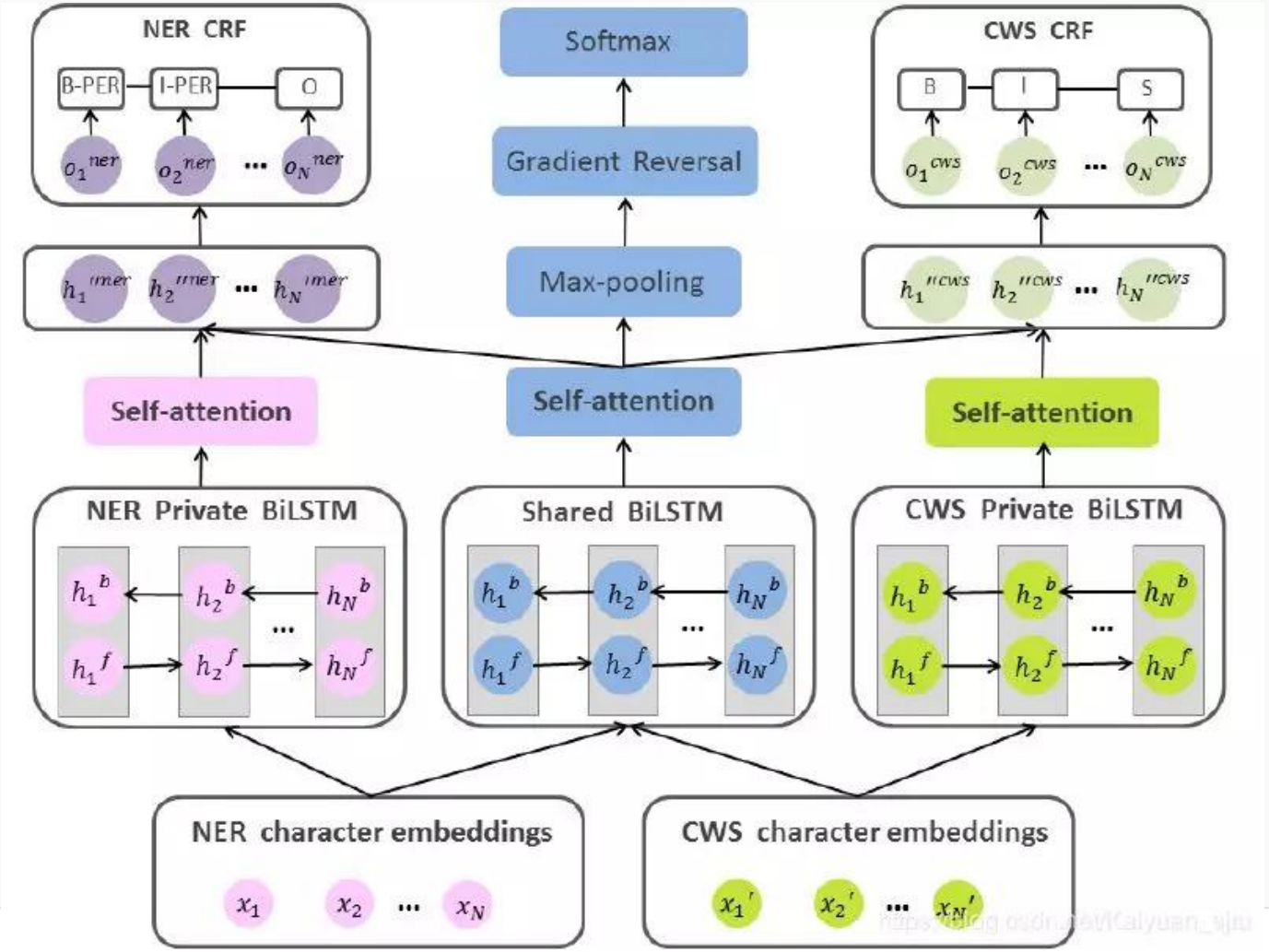
05 ATL for NER



对于中文NER任务，只有极少量的注释数据可用。中文NER任务和中文分词（CWS）任务有许多相似的单词边界。每项任务都有特殊性。但是，中国NER的现有方法要么不利用CWS的字边界信息，要么不能过滤CWS的具体信息。

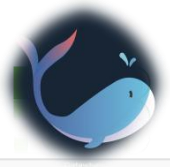
[4] Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism

05 ATL for NER



[4] Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism

05 ATL for NER



针对中文分词任务与命名实体识别任务之间的共性，引入了对抗性机制来获取中文分词任务中的边界信息，有过滤掉可能对命名实体识别任务产生影响的中文分词任务具体信息。

同时，为了解决长距离依赖问题，引入了自注意力机制。

[4] Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism

敬请各位大佬批评指正