**使用TCN-CRF模型实现命名实体识别**

胡安文 2017104077

1. **绪论及相关工作**

命名实体识别（NER,Named Entity Recognition）是自然语言处理中的一项基本而且重要的工作，是很多任务的基础，如对话系统，信息检索，机器翻译等。这个任务目的在于将文本中的词判定为限定的实体类别或者判定为非实体。常用的实体类别包括人名、地名、机构名等。

在过去的十几年中，命名实体识别领域有基于知识工程的方法，也有HMM(Hidden Markov Model)、CRF（Conditional Random Field）等概率图模型。近几年，由于神经网络的兴起，LSTM（Long Short-Term Memory）广泛应用于序列模型中，在命名实体识别领域也取得了非常好的效果。但是由于LSTM的结构特性，每一时刻的输入与上一个时刻的输出相关，使得这种结构在处理速度上比不上CNN(Convolution Neural Network)，也无法发挥GPU的优势。因此，越来越多的研究关注于如何降CNN的结构有效的应用到序列模型中，使之达到与LSTM相同的效果，同时可以极大的提高运行速度。

在论文《An empirical evaluation of generic convolutional and recurrent networks for sequence modeling》中，作者提出一种TCN（Temporal Convolution Network）的结构，该结构结合扩大式卷积和残差结构的优势，达到覆盖长度长、训练速度快的目的。该结构在多个序列任务中都效果超过LSTM或GRU(Gated Recurrent Unit)。

受该论文启发，本工作将TCN与CRF相结合，将其应用到命名实体识别中，做一个探索式的实验。

1. **方法和模型**

**2.1 扩大式卷积**

扩大式卷积（《[Multi-Scale Context Aggregation by Dilated Convolutions](http://xueshu.baidu.com/s?wd=paperuri:(906431f7e4f3b6679f265fe39c229fd8)&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http://arxiv.org/abs/1511.07122&ie=utf-8&sc_us=11132490529322257159" \t "http://xueshu.baidu.com/_blank)》）在普通卷积的基础上，设置了一种空洞卷积的结构，使得能获取到更远的特征信息。图1为扩大式卷积的结构，d代表卷积核中的间隔，d=1时和普通卷积没有区别。

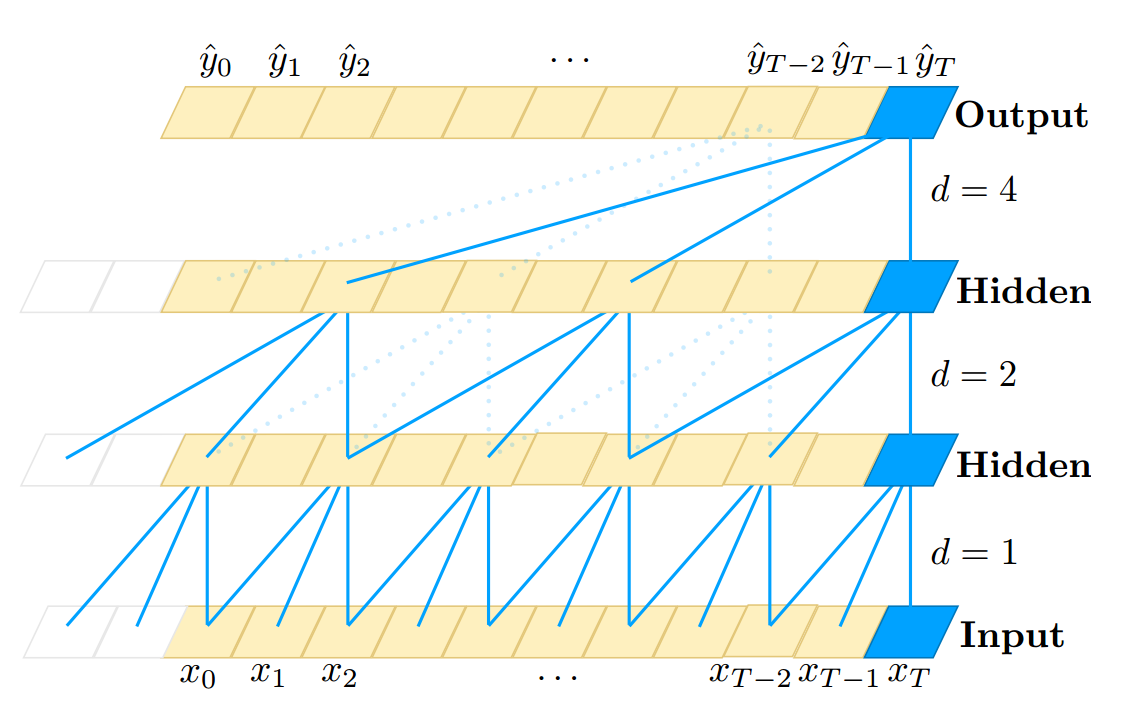


图1

**2.2 残差结构**

残差结构提出于《[Deep Residual Learning for Image Recognition](http://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90" \t "http://xueshu.baidu.com/_blank)》，该结构通过将输入向量与输出向量相加，达到加快参数的训练的目的，从而大大减少模型的训练时间。图2为残差结构的示意图。当输入向量维度与输出向量维度不一致时，通过对1×1的卷积操作调整输入向量的维度。本实验中的一个残差块中包含两层结构一致的扩大式卷积。

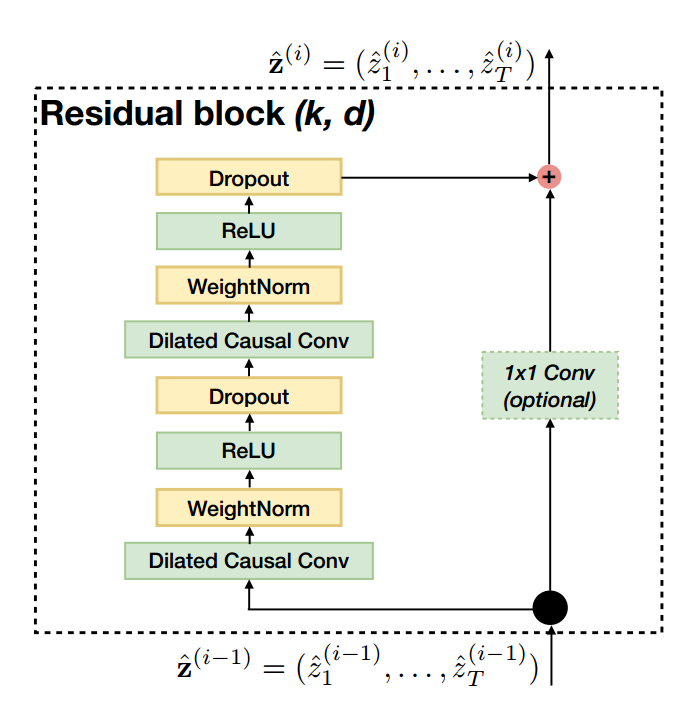


图2

**2.3 条件随机场**

条件随机场被广泛应用于序列标注任务，如词性识别，实体识别等。该算法通过引入一个状态转移矩阵，计算整个序列的条件概率。状态转移矩阵如图3所示。

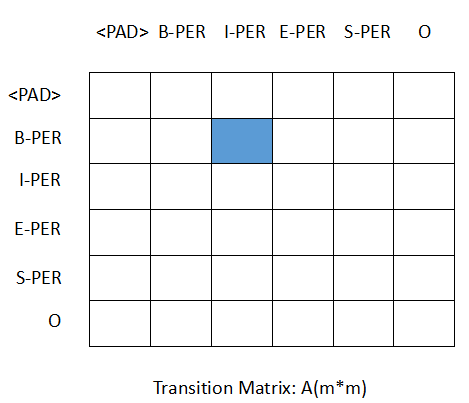


图3

**2.4 TCN-CRF**

TCN-CRF整体结构如图4所示。通过堆叠残差结构的个数，使得输出向量包含更远距离的信息。残差结构中的扩大式卷积核的间隔呈以2为底的指数增长，这样可以大大减少残差块的个数。经过所有的残差结构后将输出向量送入CRF层计算所有可能输出序列的得分，训练过程中通过极大似然估计训练参数，预测时使用维特比算法选取得分最高的序列。

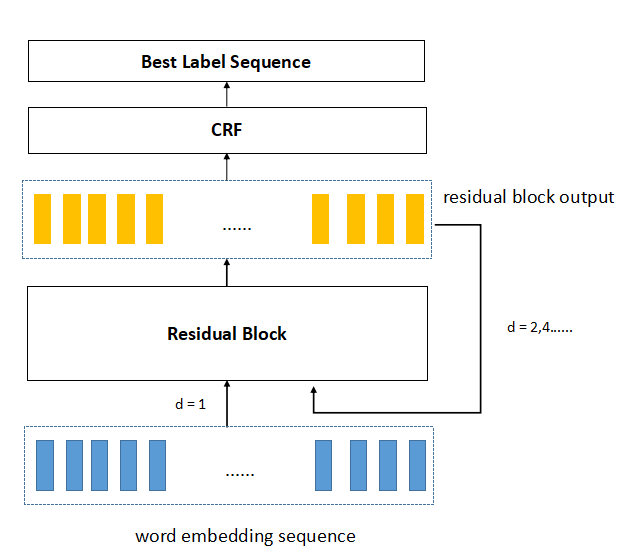


图4

1. **实验设计及结果分析**

**3.1 数据集**

本次实验采用Conll-2003数据集，该数据集由路透社发布，包含1996年8月20日至1997年8月19日的英文新闻。该语料中的命名实体识别数据集的分布如图5所示。训练集共包含946篇新闻，14041个句子，验证集包含216篇新闻，3250个句子，测试集包含231篇新闻，3453个句子。该数据包含四种实体类型，即人名，地名，机构名，其他实体。该数据集中标注编码采用BIO（Begin,Inside,Outside）编码形式，之前研究表明BIOES（Begin,Inside,Outside,End,Single）编码形式优于BIO编码，所以本次实验训练也采用BIOES编码，测试时会将BIOES编码转化为BIO编码，仍采用原始的与数据集一同发布的基于BIO编码的评测脚本进行评测。

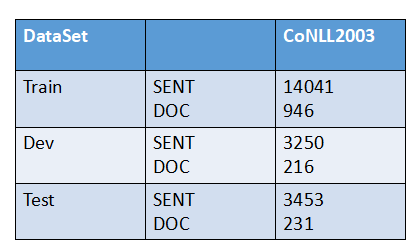


图5

**3.2 超参数设置**

本次实验基于上数据集分别做了句子级实体识别以及篇章级实体识别。句子级实体识别以句子为单位作为模型的输入，忽略一篇文章中句子间的关联。篇章级实体识别以文章为单位作为模型的输入，考虑一篇文章中句子的语义相关性。两种模型相关超参数如表1所示。

由于篇章级序列长度远大于句子级序列长度，所以在卷积核大小相同的情况下，篇章级需要堆叠更多的残差结构。

|  |  |  |
| --- | --- | --- |
| 超参数 | 句子级模型 | 篇章级模型 |
| 卷积核大小（filter size） | 5 | 5 |
| 残差结构个数 | 4 | 7 |
| batch size | 8 | 2 |
| 迭代轮数(epoch) | 100 | 100 |
| 优化器(optimizer) | SGD | SGD |
| 梯度修剪(gradient clip) | 0.35 | 0.35 |
| 丢失率(dropout rate) | 0.5 | 0.5 |

表1

**3.3实验结果**

实验结果如表2所示。初步实验显示，TCN\_CRF模型表现还是不如BILSTM\_CRF模型。

|  |  |
| --- | --- |
| 模型 | F1 |
| BILSTM-CRF | 0.8980 |
| TCN-CRF(SENT) | 0.8464 |
| TCN-CRG(DOC) | 0.8320 |

表2

1. **总结**

本次实验结果表明TCN的结构在实体识别领域效果不理想。

由于参考论文中未直接在实体识别领域进行实验，本次实验的超参数参考了其中比较相近的任务——词语级别的语言模型，所以超参数的设置还有待进一步实验。

本次实验在CPU上进行，所以也无法充分发挥TCN结构的计算优势，但仍然相比于BILSTM节省了一半的时间。

1. **参考文献**

1.He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015:770-778.

2.Yu, Fisher, Koltun, Vladlen. Multi-Scale Context Aggregation by Dilated Convolutions[J]. 2015.

3.Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv:1803.01271, 2018.

3.Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. 2016:260-270.