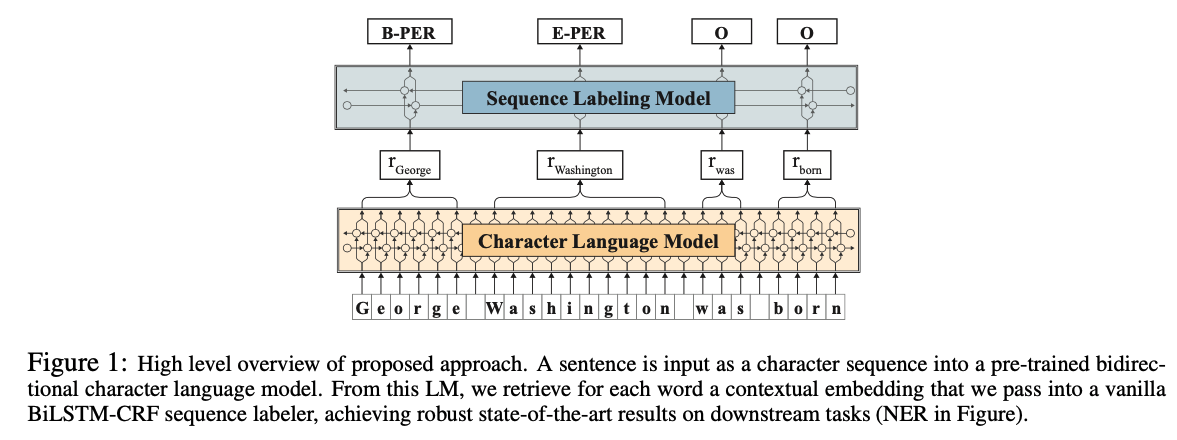
论文阅读笔记一

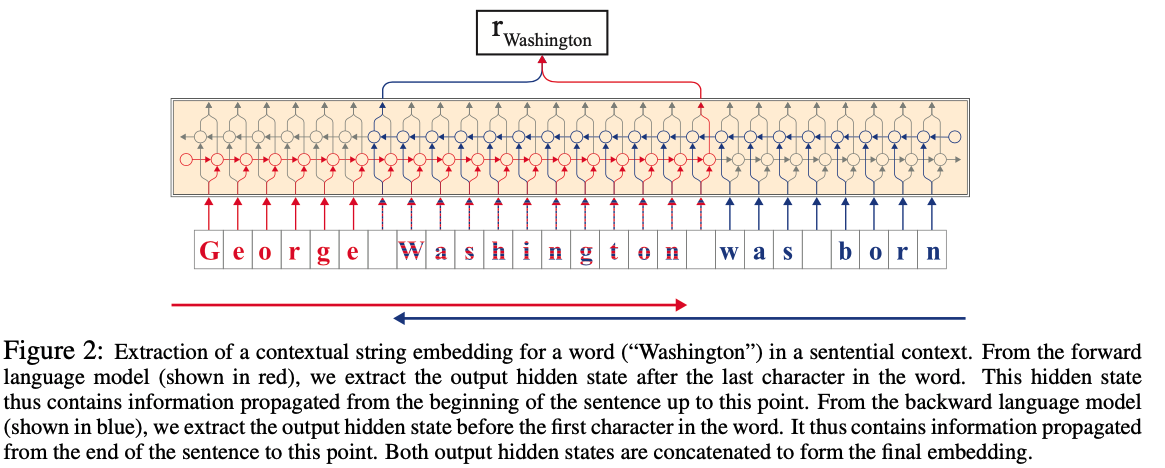
Contextual String Embeddings for Sequence Labeling

1. 贡献
2. 提出了一种从字符级神经语言模型中提取上下文单词嵌入的方法。
3. 定量评估了所提出的词嵌入模型的实用性
4. 将模型代码和预训练语言模型发布在易于使用的框架中。
5. 方法



字符语言模型（character language model）：利用双向LSTM隐藏层编码词向量融合上下文语义信息。

序列标注模型（sequence labeling model）：单词级词向量编码，解码过程。



字符语言模型将每个单词的字符作为输入进行建模，因为一个词在不同语境中的含义不同 ，当缺少上下文时，单词的语义便难以区分，故对单词第一个字母的前向隐藏层向量和最后一个字母的后向隐藏层向量进行拼接，得到单词的上下文词嵌入，解决一词多义问题和未登录词的问题。

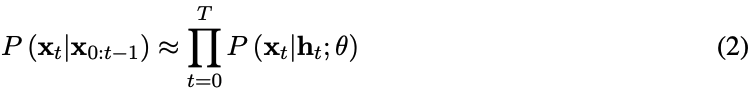
字符语言模型的训练过程：

给定字符序列

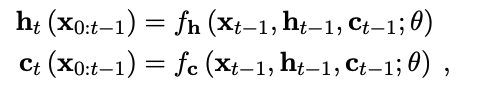
字符序列的联合概率分布为：



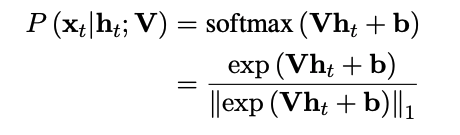
每个字符的条件概率可以近似为LSTM神经网络中隐藏层向量ht：



其中隐藏层向量ht和细胞状态ct来自LSTM模型：



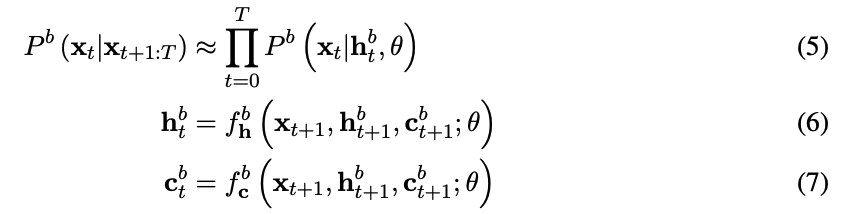
θ是模型参数，包括权值V和偏置项b。



根据语言模型的输出概率和语料库中统计所得概率分布做交叉熵计算，进行无监督的预训练。

单词的上下文表示：

将单词的第一个字符的前向向量和最后一个字符的后向向量拼接起来得到含有上下文信息的单词嵌入，其中，后向向量的表示方法如下，上标b表示后向：



拼接前向和后向向量

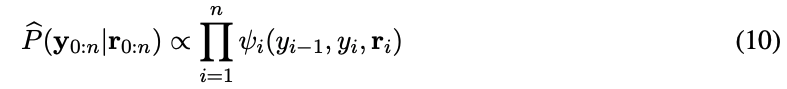


序列标注模型：对于给定输入序列，将字符训练为词向量



其中ri表示LSTM前向和后向的输出向量。

通过CRF进行解码：



1. 实验结果

作者探究了四种词向量对命名实体识别结果的影响：

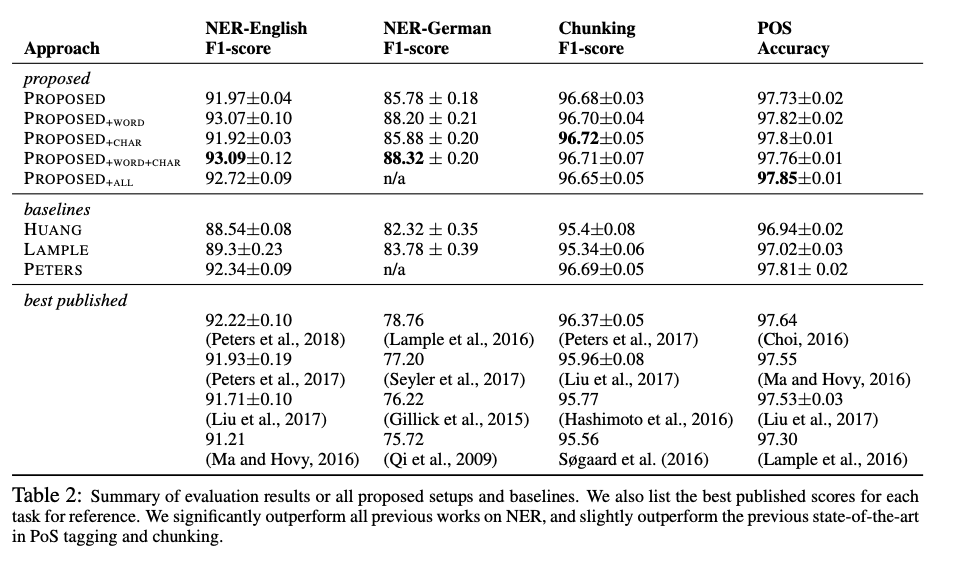
本文提出的方法

本文提出的方法+GloVe词向量

本文提出的方法+字符向量

本文提出的方法+GloVe+字符向量

实验结果如下：

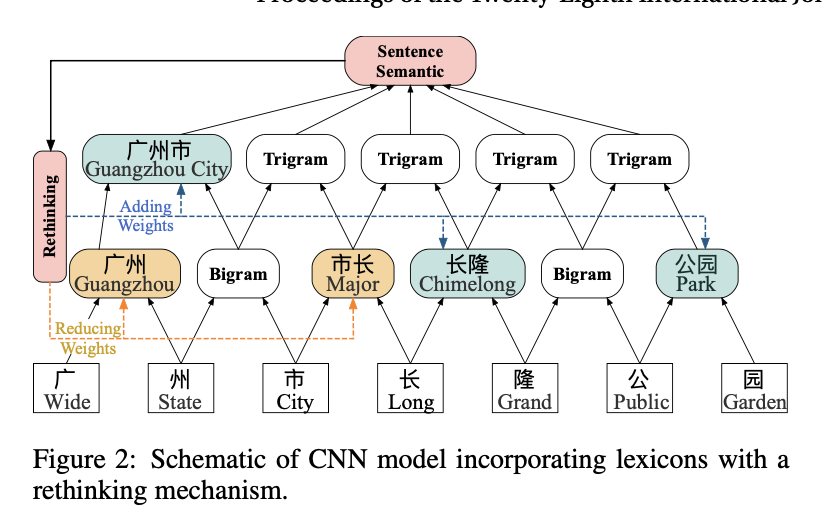


从上表可以看出，基于输入文本的词嵌入与预训练的词嵌入结合效果最佳，相较于本文提出的方法得到的91.97分，结合了单词嵌入和字符嵌入的方法提升了1.1个百分点。

论文阅读笔记二

CNN-Based Chinese NER with Lexicon Rethinking

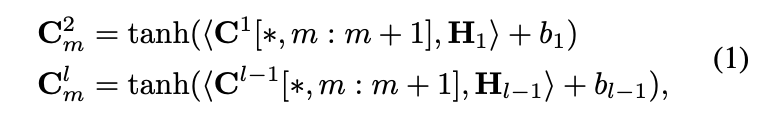
1. 主要贡献
2. 提出一种基于CNN结构的将词典信息融入到中文NER的模型，可以有效加速模型训练；
3. 设计一个rethinking机制来处理中文单词匹配矛盾问题；
4. 模型运行速度是目前SOTA模型的3.21倍。
5. 方法



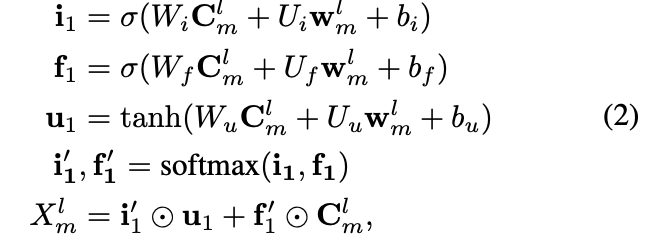
1. 基于词典的CNN

对于给定输入句子，其中表示句子中第i个字符、维度为d的词嵌入表示；表示第m个字符长度为l的单词。用CNN模型对字符进行编码，并通过attention模型融入词典信息。

CNN使用维度为d×l的矩阵作为卷积核，l表示卷积核窗口大小，；CNN使用不同大小窗口的卷积核进行特征提取，l为2时抽取bigram特征，大小为l时表示l-gram特征。



最后句子的特征表示为，计算过程如下：

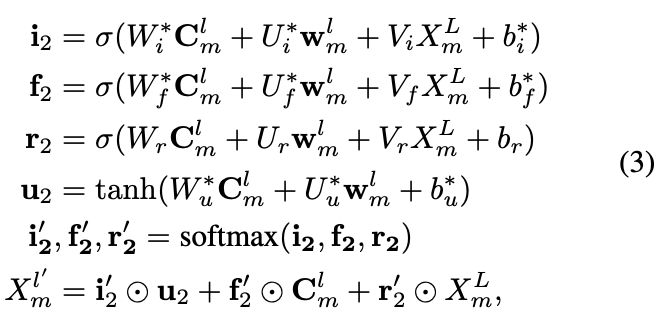


其中，W，U，b是注意力模块的参数。该模块可以并行处理字符和单词。

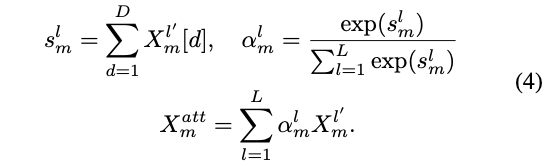
1. Rethinking机制

CNN分层机制使得底层的单词信息无法处理高层的单词，rerethinking机制处理词典中的单词匹配冲突问题。

本文将CNN顶层特征作为高层特征，通过给每一层CNN添加一个反馈层来调整单词的权值，计算过程为：

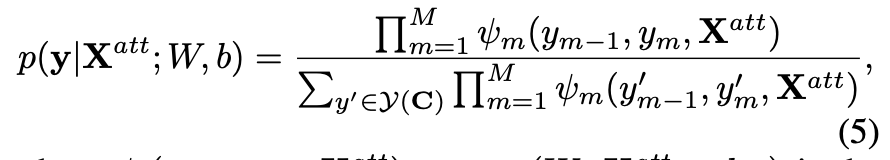


通过卷积核抽取l-gram特征的同时更新句子中每个字符的表示，通过attention机制将字符在不同卷积尺度下的特征进行融合：



最后得到，该向量讲作为CRF层解码和预测的输入。

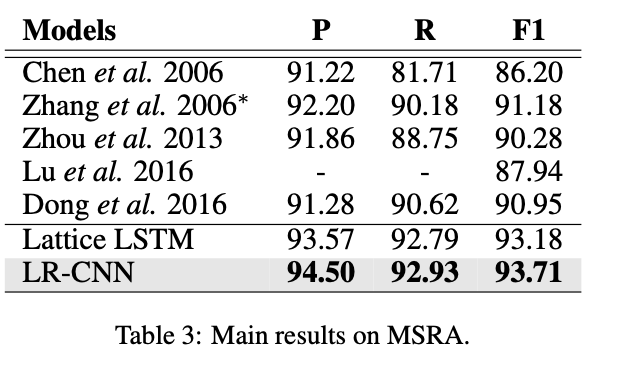
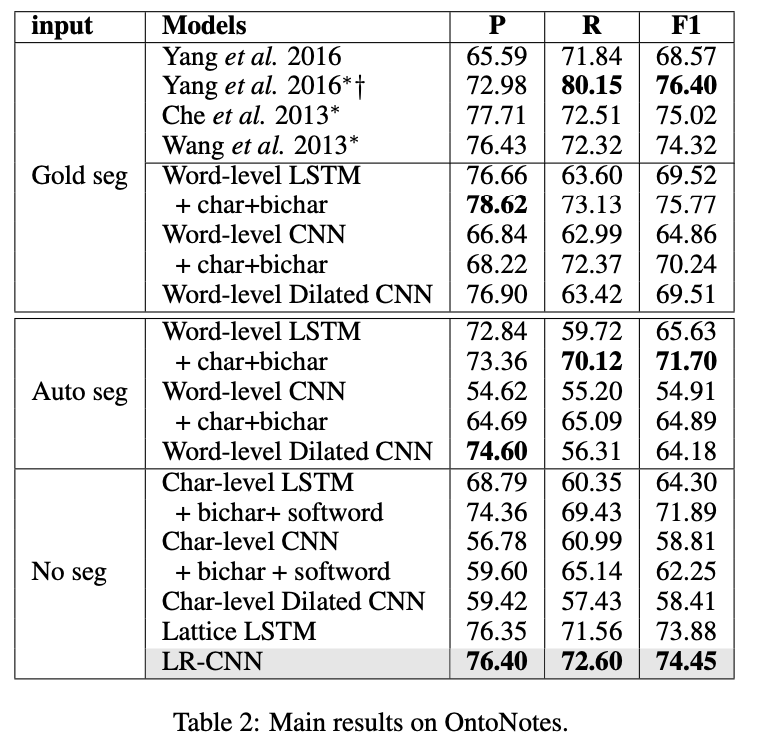
CRF解码

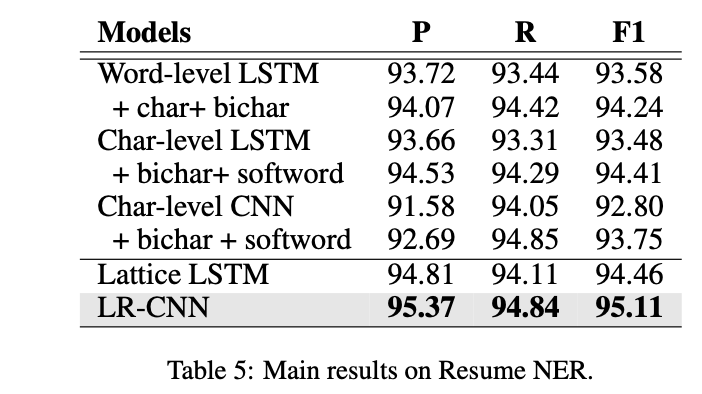
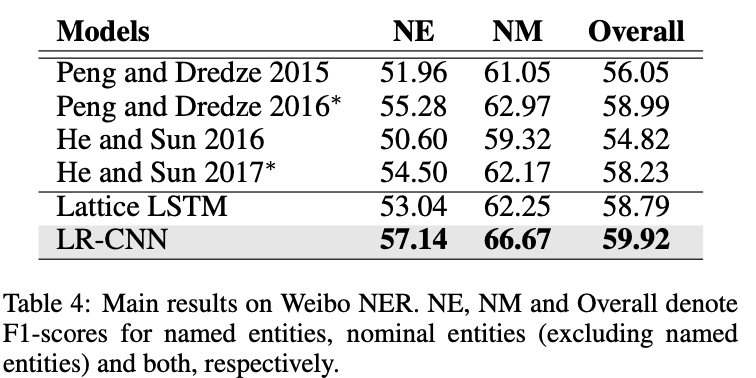


1. 实验结果

数据集：OntoNotes、MSRA、weibo和简历数据集

Baseline：LSTM、CNN、lattice LSTM等





作者分析了LR-RNN在不同数据集上的结果，证明LR-CNN的实验效果，同时分析了解码阶段的计算效率。作者将该模型与lattice-LSTM进行了对比，LR-CNN在处理短句上更具优势，随着句子商都的增加，LR-CNN速度变化比较稳定。