基于 LSTM 的中文分词模型报告

一、 简介

《现代汉语》中说“词”是“最小的能独立运用的音义结合体”（沈阳 郭锐，2015），这里的“词”是中文的词，而中文作为一种孤立语和其他黏着语（如韩语、日语）以及西方屈折语（如英语、西班牙语）的主要区别在于词与词之间没有明显的边界。而想要机器理解自然语言，需要从词这一语义最小开始。因此，中文分词是许多自然语言处理任务的基础。也是自然语言处理的典型任务。

在深度学习兴起前，传统的中文分词模型主要为：基于字符匹配的机械分词方法（最大匹配、双向匹配等）、基于统计的分词方法（HMM、CRF 等）以及理解性分词。近年来，许多深度学习神经网络的模型如 RNN、CNN、GNN 等也被应用到中文分词这一任务上，分词效果相较于传统分词方法得到了显著改善。

深度神经网络的分词方法主要分为两种：基于字符的的分词方法和基于词的分词方法。我们的模型属于第一种，其基本思想是借助序列标注完成分词的任务。这类方法最初通过深层网络来自动获取任务相关特征，从而避免使用特定任务或是手工设计特征工程（Xiaoqing Zheng et al.,2013）。尽管取得了不错的分词效果，但利用的是固定窗口大小的上下文信息，忽略了一些长距离的文本信息。接下来，通过用长短期记忆网络（LSTM）替代神经网络中的隐藏层，一定程度改进了传统神经网路难以解决长距离依存关系的问题（Xinchi Chen et al.,2015）。但单向的 LSTM 只有前文信息，而前后文均可能对分词结果造成影响，于是又有论文提出双向的 LSTM 模型（Yushi Yao et al.,2016），该模型充分利用前后文信息并在训练过程中加入 dropout 以防止过拟合。在这些研究的基础上，又有更多更复杂的组合模型被提出并用于中文分词任务。

在本篇报告中，我们将通过中文分词这一任务的结果，对一个典型 LSTM 模型和在其基础上改变得到的其他模型进行比较，理解深度学习模型是怎样影响中文分词的效果。

二、 参考模型

作为对比，我们还对一个开源高准确度的中文分词包——pkuseg[1]进行了了解，该分词包简单易用，较 jieba 和 THULAC 等国内代表分词工具包在训练语料上的表现更好（Xu Sun et al.,2012；Jingjing Xu et al.,2016）。该分词包下载后用户可加载预训练好的模型，有三个不同版本，分别是基于三个训练集：MSRA[2]、CTB8[3]、WEIBO[4]，对测试集进行测试得到的数据见表 2-1。

表 2-1：不同版本 pkuseg 的预训练模型的表现

模型	召回率	准确率	F 值
pkuseg $_{\text{MSRA}}$	**0.929**	**0.947**	**0.938**
pkuseg $_{\text{CTB8}}$	0.840	0.825	0.832
pkuseg $_{\text{WEIBO}}$	0.834	0.829	0.831

从表 2-1 中可以发现，训练语料对模型分词效果影响非常明显，其中训练集 MSRA 的分词表现最好。

三、 模型和实验

[1] https://github.com/lancopku/pkuseg-python

[2] http://tcci.ccf.org.cn/conference/ 2016/pages/page05_CFPTasks.html

[3] http://sighan.cs.uchicago.edu/ bakeoff2005

[4] https://catalog.ldc.upenn.edu/ ldc2013t21

3.1 模型基本设置

基础模型的架构为 Embedding+Bi-LSTM+CRF，Embedding 采用预训练的 50 维词向量，由 Word2vec 生成。序列标注采用三标签（B:词头 I:词中 S:单字词），通过 Bi-LSTM 获取每个词对应的所有标签的概率，可取得最大概率标注序列。但这样的标注序列可能是不合实际的，因此再加上一个 CRF 层。CRF 层作用是学习标签之间的约束。仅在第二层 Bi-LSTM 后使用一个 0.4 的 dropout。每个单向 LSTM 层的 hidden size 为 64。所有实验模型的训练次数 epoch 均为 10。

模型超参数概览见表 3-1。

表 3-1：模型超参数概览

超参数	设置
learning rate	0.01
batch size	64
optimizer	Adam
Embedding	50
LSTM hidden size	64
dropout	0.4
epoch	10
Bi-LSTM layers	1

实验的因变量为：Bi-LSTM 的层数、词向量维度以及是否加入 CRF 层，所有模型的除因变量外超参数设置均与基础模型相同。

3.2 实验数据集

实验的训练集和测试集来自 SIGHAN-2004[5]竞赛。其中训练集含 86923 个分好词的中文句子，测试集包含 3985 个未分词的中文句子。作为对比，我们将训练集按 9:1 划分出验证集。

3.2 实验结果和数据分析

通过在基础模型（word embedding[50]+Bi-LSTM+CRF）上改变因变量得到不同模型在验证集和测试集上的表现见表 3-2。

表 3-2：不同模型在测试集上的表现

模型	Accuracy	Recall	Precision	F
基础模型	0.954	0.935	0.930	0.933
+1 Bi-LSTM layer	0.960	0.944	0.938	0.941
+2 Bi-LSTM layer	0.965	0.933	0.931	0.932
+0 Bi-LSTM layer &word embedding[150]	0.956	0.937	0.934	0.935
+1 Bi-LSTM layer &word embedding[150]	0.959	**0.946**	**0.943**	**0.944**
- CRF layer	0.950	0.926	0.925	0.925

在不对模型进行额外调参的的情况下，实验结果反映了增加 1 层 Bi-LSTM 明显提高了模型在训练集上的表现，因为层数的增加使得神经网络拟合状态序列类别划分的能力更强

[5] http://www.aclweb.org/anthology/ sighan.html

了；但是 Bi-LSTM 的层数并非越多越好，增加 2 层时，在验证集上的表现很好（0.965），但是在测试集上的表现比基础模型略差，很可能就是过拟合问题。理论上，高维的词向量包含的文本中词的相关信息更多。使用 150 维的词向量，的确提升了模型的分词效果，但提升空间比较小，这可能还与模型的学习能力有关。CRF 层的加入，也明显地提升了模型的分词效果。

四、 其他的想法

囿于算力，本实验的模型设计还是比较简单的。未来可以尝试的方向还有：增加隐藏层神经元数量或隐藏层数后，再用高维词向量训练；尝试 CNN+Bi-LSTM 的网络模型，通过加入一个卷积层和一个池化层来模拟组合特征并挑选最有价值的特征，在不增加模型复杂性的基础上加强对复杂特征的建模。

参考文献：

[1]沈阳,郭锐 编.现代汉语[M].高等教育出版社.62-63.2015

[2]Xiaoqing Zheng,Hanyang Chen,Tianyu Xu.Deep Learning for Chinese Segmentation and POS Tagging.EMNLP.647-657.2013

[3]Xinchi Chen,Xipeng Qiu,Chenxi Zhu,Pengfei Liu,Xuanjing Huang.Long Short-Term Memory Neural Networks for Chinese Segmentation.EMNLP.1197-1206.2015

[4]Yushi Yao,Zheng Huang.Bi-directional LSTM Recurren Neural Network for Chinese Segmentation.ICONIP(4).345-353.2016

[5] Xu Sun, Houfeng Wang, Wenjie Li. Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection. ACL. 253–262. 2012

[6] Jingjing Xu, Xu Sun. Dependency-based Gated Recursive Neural Network for Chinese Word Segmentation. ACL.567-572.2016