本周报告

> 2020/05/07 1652792 罗吉皓

本周主要做了机器学习模型的对比实验和论文的编写

对比实验

对比实验方面我主要设计了两个实验，实验一中主要针对不同的特征向量提取方式，在实验中我们首先针对文本特征向量提取与否进行对比实验，在网络结构相同，参数相同的情况下进行模型训练并对比预测效果。在预训练语言模型中，除了BERT模型，我还选取了ERINE模型进行对比实验。Ernie模型被称为中文文本预训练最强网络，他在BERT的基础上更注重于词法结构，语法结构，语义信息的统一建模，中文文本相对于其他语言文本来说其最大的分析难度便在于其丰富的语法结构和语义信息，ERNIE在这一方面的强化让他在中文文本分析中更有优势。在实验中我们着重选取了两种目前最为流行的预训练语言模型，通过对比实验帮助选取最好的模型进行后续解决方案的搭建。

另外一方面我选取了目前命名实体标注方面三个标注效果比较好的网络结构CNN_LSTM模型，BiLSTM模型以及BiLSTM_CRF三个模型进行实验。

CNN_LSTM模型中我主要使用了简单的卷积层累加，而后将输出结果输入后续LSTM网络，同时加入Dropout层防止过拟合。由于是序列学习问题，我在之后加入了Time_distributed层，帮助选取最佳的标注结果。

Bi-LSTM网络中我通过加入双向LSTM网络帮助模型进一步理解上下文语境，LSTM网络帮助上流文本信息传递到后续的学习，而Bi-LSTM帮助下文的信息反向传递到上文，从而帮助模型训练更为充分，比较适合像提取核心地点这种更为细粒度的标注需求。

CRF层的介绍在上文中我已经有所提及，在BiLSTM模型的基础上，我去除了Time_distributed层而采用了CRF条件随机场，来对比最后的标注效果。

实验中三个模型Dropout设置为0.5，批处理大小设置为64，训练轮次40轮。

对比实验结果如下所示：

模型	特征提取	T-Precision	T-Recall	T-F1Score	P-Precision	P-Recall	P-F1Score	V-Precision	V-Recall	V-F1Score
CNNLSTM	Random Init	0.2855	0.3421	0.3171	0.4643	0.4094	0.4351	0.3200	0.3107	0.3153
Bi-LSTM	Random Init	0.3134	0.2763	0.2937	0.4961	0.5039	0.5000	0.3053	0.2816	0.2929
Bi-LSTMCRF	Random Init	0.3623	0.3289	0.3448	0.5596	0.4803	0.5169	0.3793	0.3204	0.3474
CNNLSTM	Bert	0.5571	0.5342	0.5455	0.6583	0.7401	0.6968	0.5444	0.5104	0.5269
Bi-LSTM	Bert	0.5976	0.6712	0.6323	0.7514	0.7684	0.7598	0.5773	0.5833	0.5803
Bi-LSTMCRF	Bert	**0.6563**	**0.6734**	**0.6647**	0.7384	0.7175	0.7278	0.5859	**0.6042**	0.5949
CNNLSTM	ERNIE	0.5000	0.5526	0.5250	0.6640	0.6587	0.6614	0.5862	0.6602	0.6210
Bi-LSTM	ERNIE	0.6486	0.6316	0.6400	**0.7778**	**0.7778**	**0.7778**	0.6048	0.6316	0.6400
Bi-LSTMCRF	ERNIE	0.6286	0.5789	0.6027	0.7480	0.7302	0.7390	**0.6458**	0.6019	**0.6231**

实验结果与反思

对于实验一是否加载预训练语言模型其效果还是比较显著的。预加载模型对于整体实验效果的影响是比较大的。表格中的前三行为不添加与训练语言模型而直接进入网络结构进行训练，其训练效果与后面添加了语言模型的精确率平均低30%的效果，因此添加与加载语言模型进行文本预先处理的方式是能够对于精确提取句子中的核心地理位置起到帮助作用的。而对于ERNIE模型与BERT模型的对比，在实际预测效果中，ERNIE与BERT模型的差异并没有很大，但是在所有地名的查找和核心动词的ERNIE的效果确实更佳，相对来说精确率能够提升

4%，说明通过给不同的标签添加不同的权重能够提升对于特定标签预测的效果。

实验二中我们主要针对 CNN_LSTM模型，BiLSTM模型以及BiLSTM_CRF三个模型对比实验。在同样预训练模型的基础上，三个模型在9个指标的表现上相近，因此在分析的时候我着重以BERT为代表来进行相关探讨。

从三个标注类别来说，“P”即相关地理位置的标注效果要远好于“V”核心动词和“T”核心地理位置的，其原因主要在于“P”的标注范围相对比较宽泛，“T”和“V”的把vi澳洲个呼应皆为多个标签中选取一个的情况，相对准确率较低也是可以理解的。

从精度来看，相较于基于规则的动词-地名树来说，其准确率只有30-40%，而机器学习模型在他的基础上精度提高了20%-30%，效果显著，并且能够预测一些规则之外的例子，如当句子中存在多个地理位置却没有核心地理位置的情况。而BiLSTM网络的效果是要好于CNNLSTM网络效果的，说明双向通道在训练程度上确实要好于单向的LSTM，双向通道的引入帮助BiLSTM网络相较CNN提高了5%（以核心动词T为例），在网络结构上确实比较适合提取核心地点这种细粒度的标注需求。

另外一方面，可以发现加入CRF层后核心动词与核心地理位置的准确率都有所提升。在BERT预训练模型基础上，我们发现核心地理位置在BiLSTM的基础上又提高了5%左右，证明CRF在特定标签的标注上表现确实相对比较优秀，而在“P”地理位置的标注上BiLSTM的效果反而要比添加了CRF层更好，本实验的核心在于提取核心地理位置，因此在最后方案的选择上还是选择了添加CRF层来帮助标注核心地理位置。

在实验的最后我通过对于测试集进行了一些修订，在预测的过程中确实存在了很多因为“报道”之类的干扰词导致最后标注错误的情况，因此我将测试集中与“报道”相关相近的动词与地理位置进行删除后进行重新的预测，经过测试，最后的结果在原来的基础上提高了3%。

虽然目前的机器学习模型在最终的预测效果具有良好的效果，但是其中还存在着比较大的提升空间，分析后原因如下：

1. 训练集样本数量不足。在本实验中训练集仅仅为2000句政治新闻，其中涉及的地理位置较多，相对来说训练集样本较为离散，模型训练样本量较小，对训练效果造成影响。
2. 存在“报道”等标注干扰词，在之前的测试中，当删去这些干扰词以后相关的准确率提升了3%，说明这些干扰词语对于模型的训练和预测都起到了负面的影响。
3. 训练的轮数不够，训练不充分，利用BERT和ERNIE作为预训练语料模型会对最后的预测效果产生积极的影响，但是在整体训练的过程中，BERT和ERNIE的预训练本身会占用大量的时间和空间，导致每一轮的训练时间过长，如果能在性能相对更好的集群上进行多轮次的训练最后的效果可能会更好。
4. 模型本身缺陷，模型本身还存在着一定的不足，可以考虑通过改变参数，增加网络层数的方法来进行多组对比实验

大三课程项目进展

目前前端搭建完成了，后端这两个礼拜由于我编写毕业设计+电脑坏了，还没有来得及做对接，争取这两天做完。