

第一赛季:

目标: 实体抽取(序列标注)

实验 1: 损失函数

模型	Bidirectional LSTM + Softmax	Bidirectional LSTM + CRF
引用	-	-
分数	68	74

总结: 在抽取有特定组合的标签, CRF 精度高于 LSTM。

实验 2: 特征提取层

模型	BILSTM+FC	CNN+FC	BILSTM+CNN+FC	BILSTM+CNN+POOL
引用	-	-	-	End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF
分数	74	63	67	-

总结: 加入卷积层效果变差, 不太清楚是不是没有用池化的原因

实验 3: 长度截取

模型	BILSTM+FC	
长度	128	全部
现象	得分先上升后下降	得分先上升后稳定
分数	74	71

实验 4: 层数

模型	BILSTM+FC	
层数	2	8
分数	74	61

第二赛季

目标: 关系抽取

方案 1: 通过第一赛季模型获取所有实体在最后一层特征。然后对同一文本的实体两两组合, 经过 AttentionLayer 得到 $len(e_1) * len(e_2) * C$ 的特征图, 将特征图进行分类。

结果: F1= 0.0281351。

原因: 阴阳比失衡(比例 300: 250000)

方案 2: 对文本进行 512 截取, 忽略实体长度大于 320 的关系, 然后对每一个句子进行分类。

阴阳比例降至 300: 12000, 但训练集有 603(0.78%)个样本会被忽略。

结果: F1=0。

方案 3: 在方案 2 基础上忽略关系类别与实体类别组合不一致的样本, 阴阳比例降至 9.7, 但训练级有 1829+603(3.1%)个样本会被忽略(代码存在 bug, xlsx 文件数据可能存在问题, 统计可能不准确)。

训练策略: 若选取 n 个阳性样本用于训练, 则用分类器对 $9.7n$ 阴性样本预测一遍后按照阴性分数从低往高进行排序, 从前 $3n * dilute$ 个样本中随机取 $3 * n$ 个阴性样本用于训练。

不同模型和 dilute 的性能表现。

模型	bert_12_768	bert_2_384	bert_12_768	bert_2_384
dilute	1	1	3	3
训练集	57	82	65	72
测试集	53	37	64	42
官方	-	-	63	-

参考 Deep Residual Learning for Weakly-Supervised Relation Extraction 添加 Resnet 模型

模型	bert_2_384	bert_2_385+resnet6
dilute	3	3
训练集	72	-
测试集	42	75
官方	-	56

总结:

1. 数据集存在错误的阴性样本, dilute=1 时增加错误阴性样本的训练的概率, 导致模型不收敛或者过拟合。
2. Bert 模型层数越深越难训练但泛化能力越强。
3. Resnet 能增加模型的泛化能力。