NER MODEL

当前存在一些问题：

1. 数据集标注问题
2. 取样和测试方式
3. 人造数据集无法提升模型真实性能
4. 模型对标签为U的字识别能力显著低于其他标签
5. 模型容易混淆标签为O和S的字

解决方法和统计结果：

数据集标注问题：基本解决

取样和测试方式：

1. 取样要从原本的随机取样替换成按人取样
2. 使用k-fold corss validation进行测试 ($k = 5$)
3. 人造数据集只用来训练不用来测试 (效果不明显)

一些对模型提升的尝试:

1. 改变crf层的权重或初始学习率
2. 在crf层上设计限制转移矩阵 (每一个训练批次结束之后都重新设置一次)
3. 使用预训练word2vec代替模型里的nn.Embedding

```
..... super(BiLSTM, self).__init__()  
..... # 采用预训练的 embedding  
..... if (LSTMConfig.pre_trained == True):  
.....     self.embedding = nn.Embedding.from_pretrained(weight)  
..... else:  
.....     self.embedding = nn.Embedding(vocab_size, emb_size)
```

统计结果

	Precision	Recall	F1	整个句子的 准确率	含U的句子 准确率	不含有U的句 子的准确率
BiLSTM-CRF	0.966	0.967	0.966	0.873	0.329	0.884
设置转移矩阵	0.968	0.968	0.968	0.884	0.456	0.894
使用预训练的 word vector	0.962	0.964	0.963	0.859	0.193	0.874

可以尝试的方向：

1. 使用预训练的word vector代替模型中的nn.Embedding

(考虑问题：需要选取适当的语料库)

2. 考虑人工创建包含U标签的训练集（可能问题：可能无法提高
在真实数据上的性能）

3. 词汇增强（Lattice LSTM）