

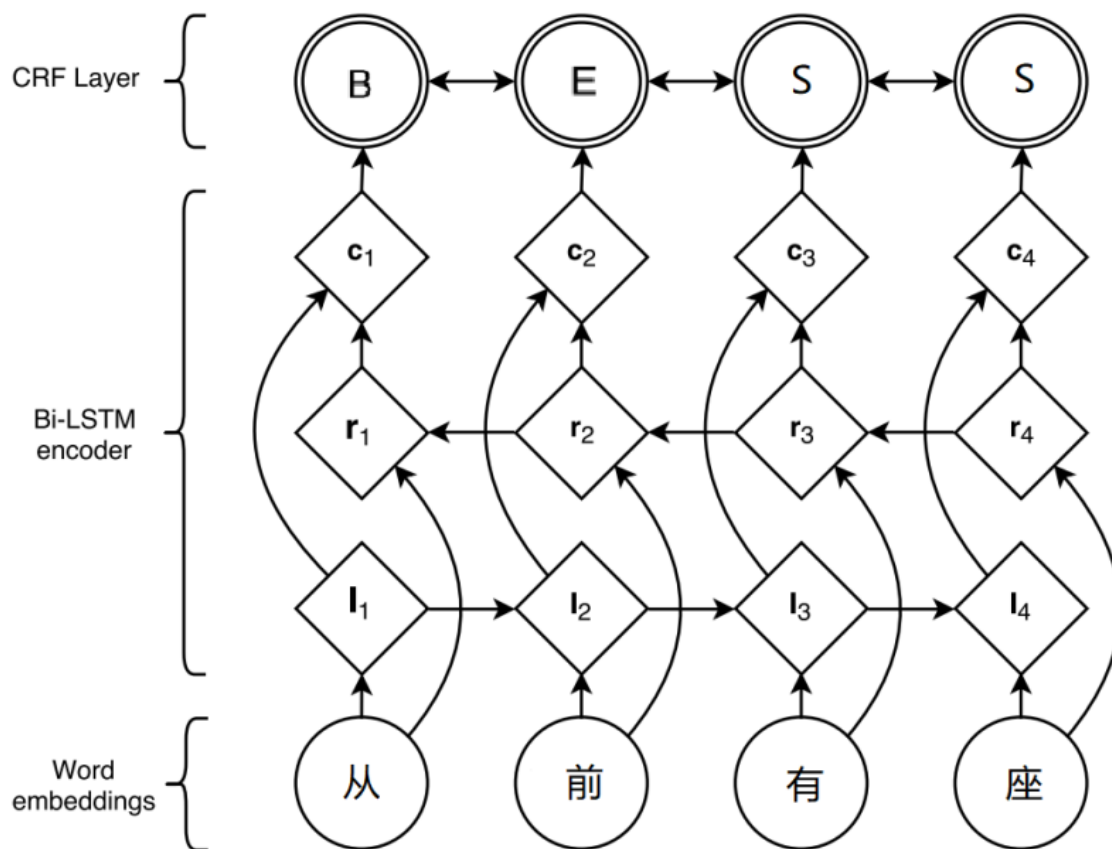
# 期中作业：中文分词+语言模型

# 作业内容

1. 使用BiLSTM+CRF分词模型，在SIGHAN Microsoft Research数据集上进行中文分词的训练和测试
2. 利用第一步分词模型对SIGHAN Microsoft Research数据集中语料进行中文分词，并用分词好的语料训练单向LSTM语言模型生成文本。

# 中文分词

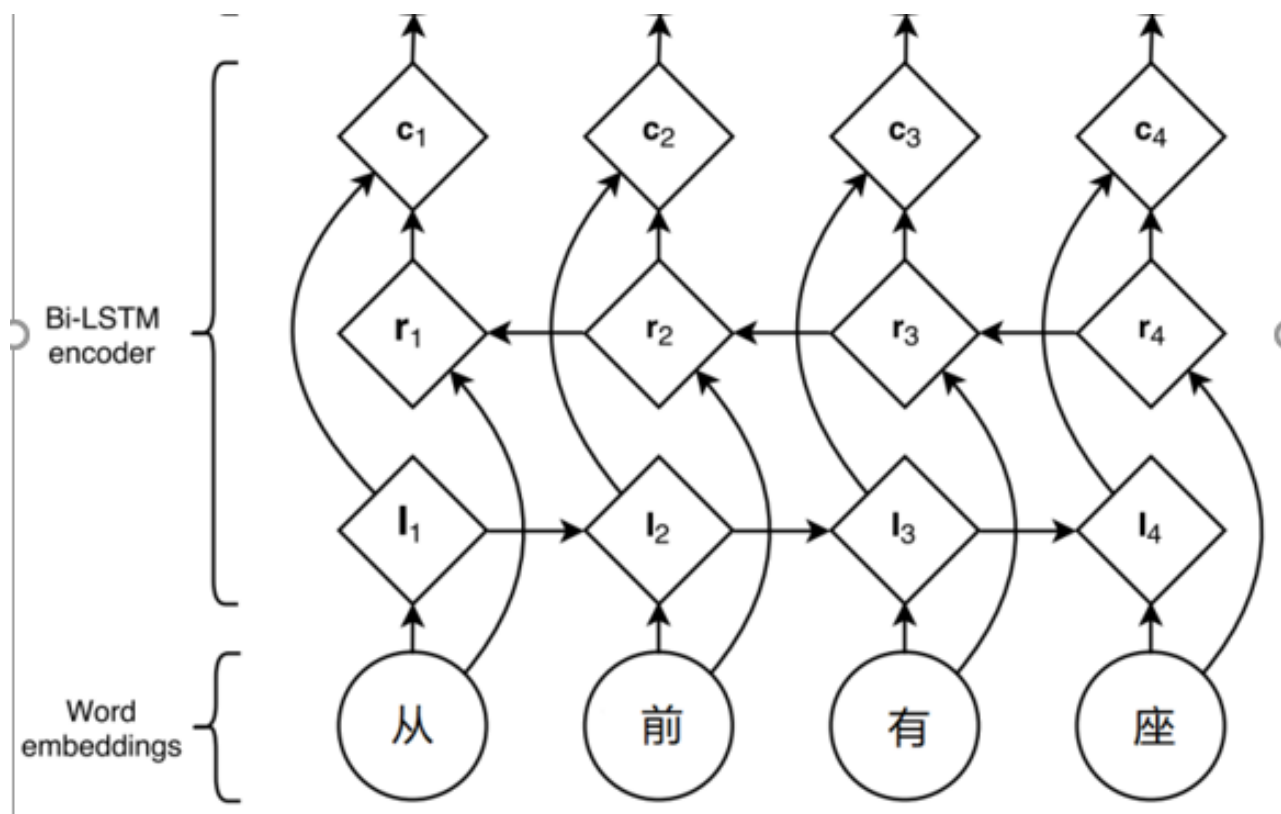
BiLSTM+CRF



# 中文分词

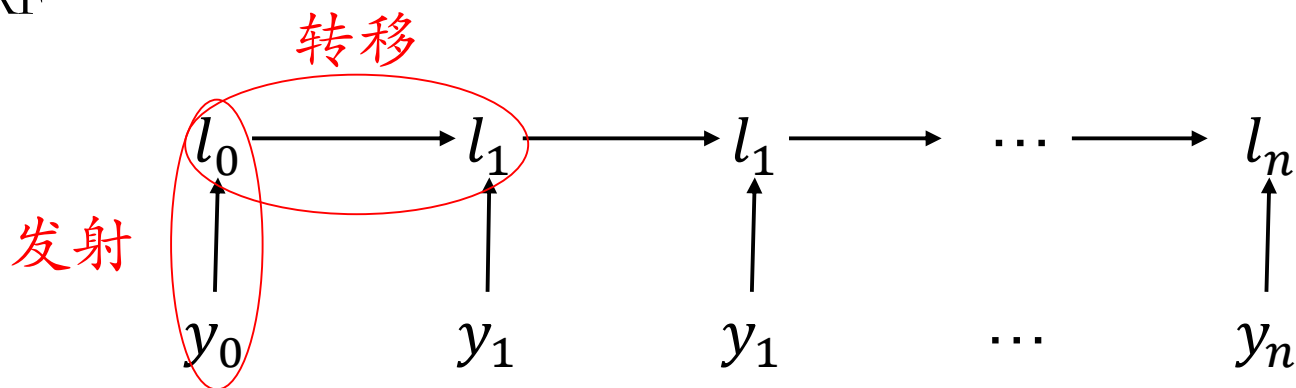
## BiLSTM

输出每个位置的字符是分词标签B,M,E,S的概率 $y_0 y_1 y_2 \dots y_n$





# 中文分词

CRF



- 对转移概率，可以设计 $V \times V$ 个特征函数 $F(l_{t-1}, l_t)$ ， $V$ 为分词标签数量（如使用B,E,M,S标签的话 $V$ 就为4），其中 $F_{i,j}(l_{t-1}, l_t) = 1$  当且仅当 $l_{t-1} = V_i$ 且 $l_t = V_j$
- 对发射概率，可以设计 $V \times V$ 个特征函数 $F(y_t, l_t)$ ， $V$ 为分词标签数量，其中 $F_{i,j}(y_t, l_t) = 1$  当且仅当 $y_t = V_i$ 且 $l_t = V_j$
- 在训练时，利用BiLSTM模型预测的概率分布 $\mathbf{y}$ 和标签 $\mathbf{l}$ 计算条件随机场的概率似然函数 $p(\mathbf{l}|\mathbf{x})$ ，以最大化似然函数作为训练的目标。
- 在测试时，得到BiLSTM模型预测的概率分布 $\mathbf{y}$ ，用训练好的条件随机场执行维特比算法解码。

# 中文分词

名称	修改日期	类型
 msr_test.utf8 测试集	2005/11/18 22:01	UTF8 文件
 msr_test_gold.utf8 测试集groundtruth	2005/8/15 7:47	UTF8 文件
 msr_training.utf8 训练集	2005/6/30 4:14	UTF8 文件

# 中文分词

使用训练集（已分好词）来训练分词模型，在python代码里文件需要用utf-8编码方式打开，每行为一个已经分好词的句子，每个词用空格隔开。

“人们常说生活是一部教科书，而血与火的战争更是不可多得的教科书，她确  
“心静渐知春似海，花深每觉影生香。  
“吃屎的东西，连一捆麦也铡不动呀？  
也“严格要求自己，从一个科举出身的进士成为一个伟大的民主主义者，进而成为  
“征而未用的耕地和有收益的土地，不准荒芜。  
“这首先是个民族问题，民族的感情问题。  
我扔了两颗手榴弹，他一下子出溜下去。  
“废除先前存在的所有制关系，并不是共产主义所独具的特征。  
“这个案子从始至终我们都没有跟法官接触过，也没有跟原告、被告接触过。  
“你只有把事情做好，大伙才服你。  
“那阵子，条件虽艰苦，可大家热情高着呢，什么活都抢着干，谁都争着多  
“哎呀，怎么也不来告诉我一声？  
“种菜，也有烦恼，那是累的时候；另外，大棚菜在降价。  
“种田要有个明白账，投本要赚利润是起码的道理。  
“成长的岁月，传统、激进、前卫，都曾经历过。  
“为什么大家在火车上什么话都敢讲，到了单位就不出声了呢？  
虽然一路上队伍里肃静无声，但“不要失望，让我们共同迎战艾滋病！  
“在现代化的战舰上，不存在技术简单的岗位。  
“当时我很伤心，认为这辈子算完了。  
“我的孩子天资还不错，但学习成绩一般，小动作较多，老师不是特别喜欢，也  
“鲨鱼软骨素——人类抗癌、防癌、治癌的新希望。  
也关于“教育要面向现代化，面向世界，面向未来”；“我们要实现现代化，关键  
禹尔有老乡拥上来想看“大官”，立即会遭到“闪开！”  
“往常排队又累又急，输密码、点钞票也不愿旁人看见。  
“目前我国的残疾人多分布在石山区、荒漠区、黄土高原区、地方病高发区，比

# 中文分词

用训练好的分词模型在测试集上测试结果，测试集中每行文本为一个未分词的句子。

扬帆远东做与中国合作的先行

希腊的经济结构较特殊。

海运业雄踞全球之首，按吨位计占世界总数的17%。

另外旅游、侨汇也是经济收入的重要组成部分，制造业规模相对较小。

多年来，中希贸易始终处于较低的水平，希腊几乎没有在中国投资。

十几年来，改革开放的中国经济高速发展，远东在崛起。

瓦西里斯的船只中有40%驶向远东，每个月几乎都有两三条船停靠中国港口。

他感受到了中国经济发展的大潮。

他要与中国人合作。

他来到中国，成为第一个访华的大船主。

访问归来，他对中国发展充满信心，他向希腊海运部长介绍了情况，提出了两国在海运、造船业方面合作

1995年10月，希腊海运部长访华时，他根据“船长”的建议与中方探讨了在海运、造船方面合作的

“船长”本人还与几个船主联合起来准备与我远洋公司建立合资企业。

“船长”常说，要么不干，干就要争第一。

他拥有世界最大的私人集装箱船队，也要做与中国合作的先行。

找准人生价值的坐标

王思斌，男，1949年10月生。

北京大学社会学系教授、博士生导师、系主任，中国社会学学会副会长、中国社会工作教育协会副会长兼秘

多年来，一直从事农村社会学、组织社会学方面的研究，主要著述有《社会学概论》（合编）、《经济体制转轨与

《社会学概论》（合编）、《经济体制转轨与



# 中文分词

使用训练集的groundtruth对分词模型在测试集上的分词结果进行评估。  
评估指标为F1:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

$precision$  = 分对词数/预测分词结果的总词数

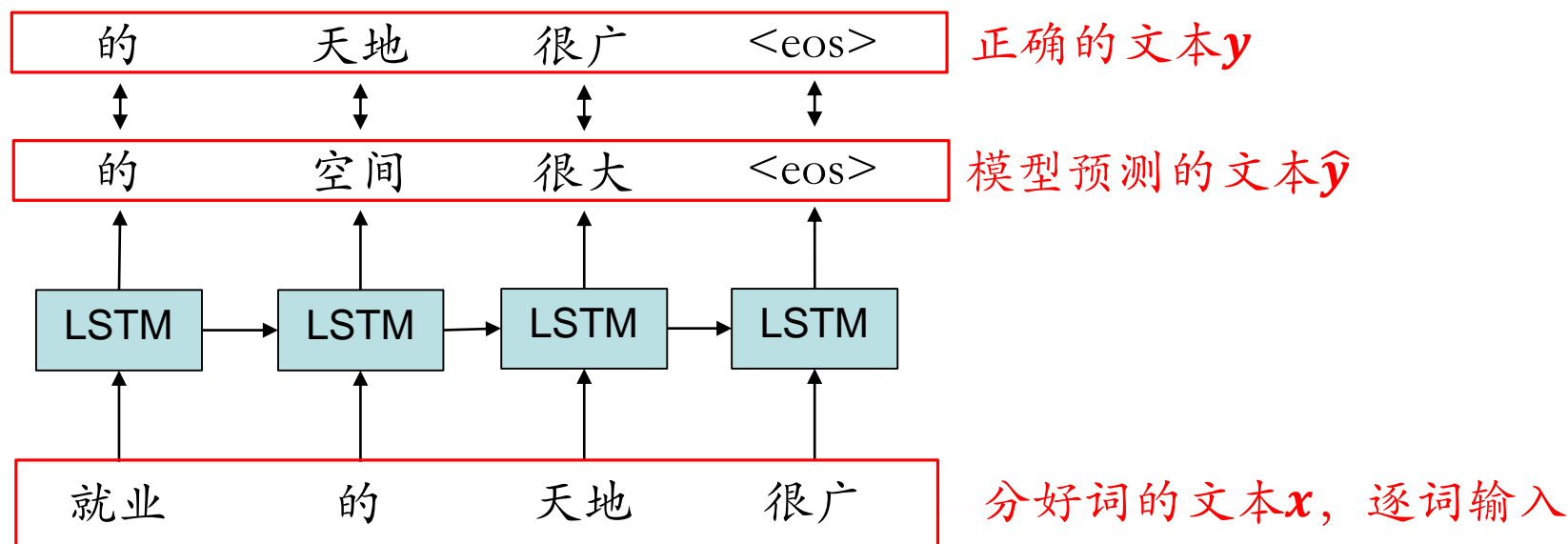
$recall$  = 分对词数/正确分词结果的总词数

最终模型评分为测试集中所有句子F1分数的平均值

# 语言模型

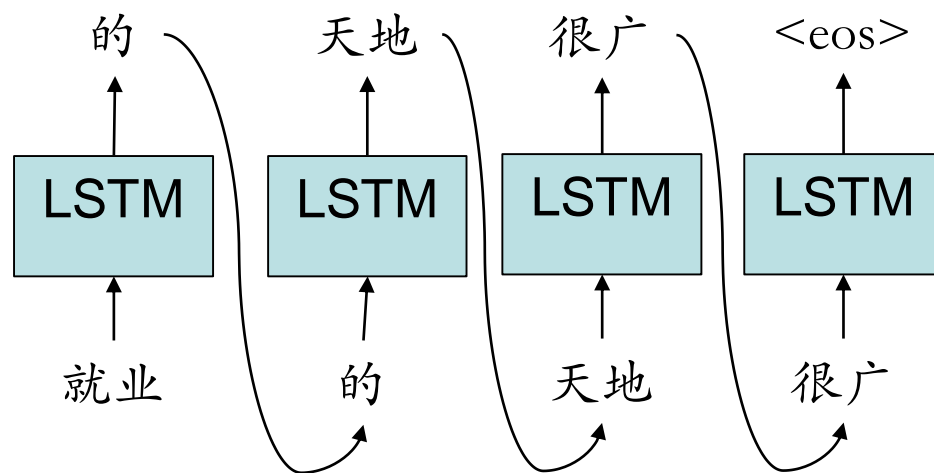
使用第一步训练好的分词模型，对训练集和测试集中的文本重新重新分词，并训练简单的LSTM语言模型，利用训练好的模型生成文本。

训练的时候，使用模型预测的文本概率 $\hat{y}$ 与正确的文本 $y$ 逐词的交叉熵作为损失函数。



# 语言模型

测试时，使用上一时刻LSTM模型输出的结果 $y_{t-1}$ 作为当前时刻的输入 $x_t$ ，直到输出结果为结束符号<eos>，则文本生成结束。在第0时刻，输入任意一个词，观察以这个词开头会生成什么文本。



# 关键时间节点

- 作业提交邮箱：[sysucsters@163.com](mailto:sysucsters@163.com)，使用压缩文件提交，命名格式：姓名+学号+中文分词or语言模型
- 中文分词模型代码/报告提交时间：2020年11月30日前
- 语言模型代码/报告提交时间：2020年12月30前

# 提交材料（基本要求）

1. 完整代码
2. 代码Readme（代码环境，如何运行你的代码进行训练和测试）
3. 测试集上中文分词结果，每行对应一个句子，分好的词用空格隔开，确保TA能够在本地利用你的分词结果得到你报告中的F1分数。
4. 报告：研究内容，研究方案，核心代码讲解，实验步骤设计，实验结果（至少需要汇报中文分词在测试集上的F1分数和语言模型的文本生成样例）

# 提交材料（加分项）

1. 代码关键模块有原创部分
2. 模型结构和方法上的创新（可参考现有论文工作）。
3. 实验部分进行更多分析和讨论（使用预训练词向量，更换特征函数，模型超参数组合，不同目标函数的选择等）
4. 用英文写作报告。
5. 所有你实现的加分项需要在报告中另开一个章节进行概述。

# Reminder

- 报告不能抄袭
- 代码可参考开源，可使用基础深度学习框架（pytorch, tensorflow, keras等），不能使用高级框架，例如openNMT, allenNLP，不能抄袭。
- 不要迟交，项目工程量较大，请尽早开始动手。