自然语言处理
课程作业

基于BiLSTM的命名实体识别

命名实体识别

- 命名实体识别（Named Entities Recognition，NER）是指从文本中识别出各类命名实体的任务。

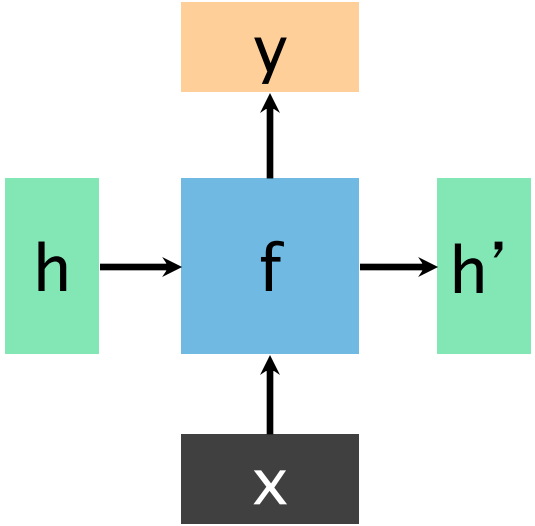
分类标签：
B：开始
I：中间
E:结束
O：其他
S：单个

小明在复旦大学上黄老师的自然语言处理课。

B E O B I I E O B I E O B I I I I I E O

Naive RNN

- 给定函数$f: h', y = f(h, x)$



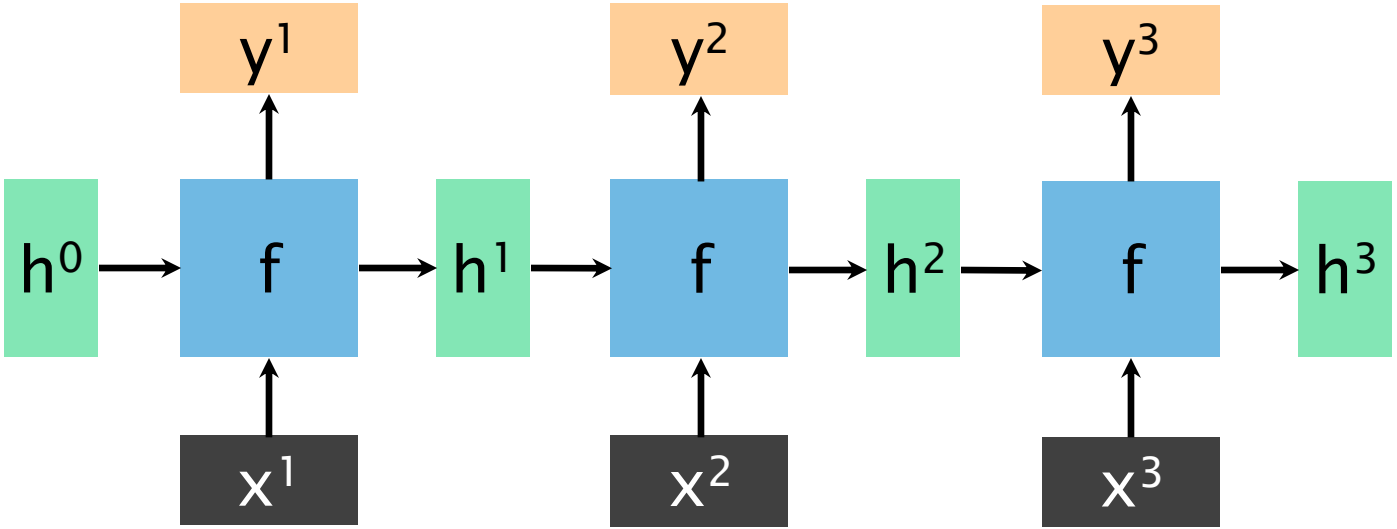
$$h' = \sigma(\ w^h\ h\ +\ w^i\ x\ )$$

$$y = \mathrm{softmax}(\sigma(\ w^o\ h'\ ))$$

- x为当前状态下数据的输入，h表示接收到的上一个节点的输入。
- y为当前节点状态下的输出，而h'为传递到下一个节点的输出。

Recurrent Neural Network

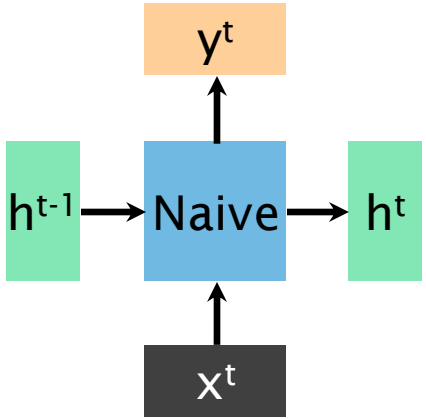
- 给定函数$f: h', y = f(h, x)$，h'和h为具有相同维度的向量

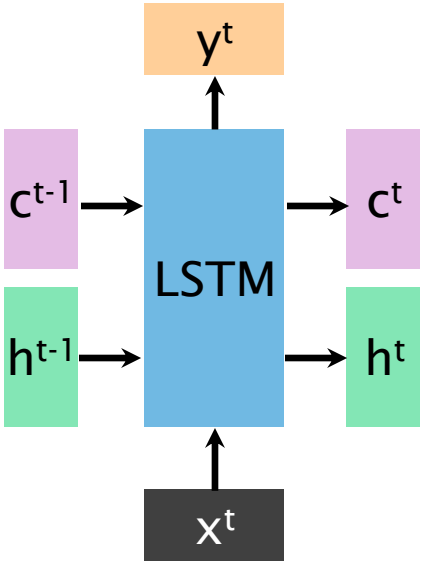


- 无论输入序列有多长，只需要一个函数 f

Long short-term memory (LSTM)

- LSTM是一种特殊的RNN，能有效缓解梯度消失和爆炸问题。

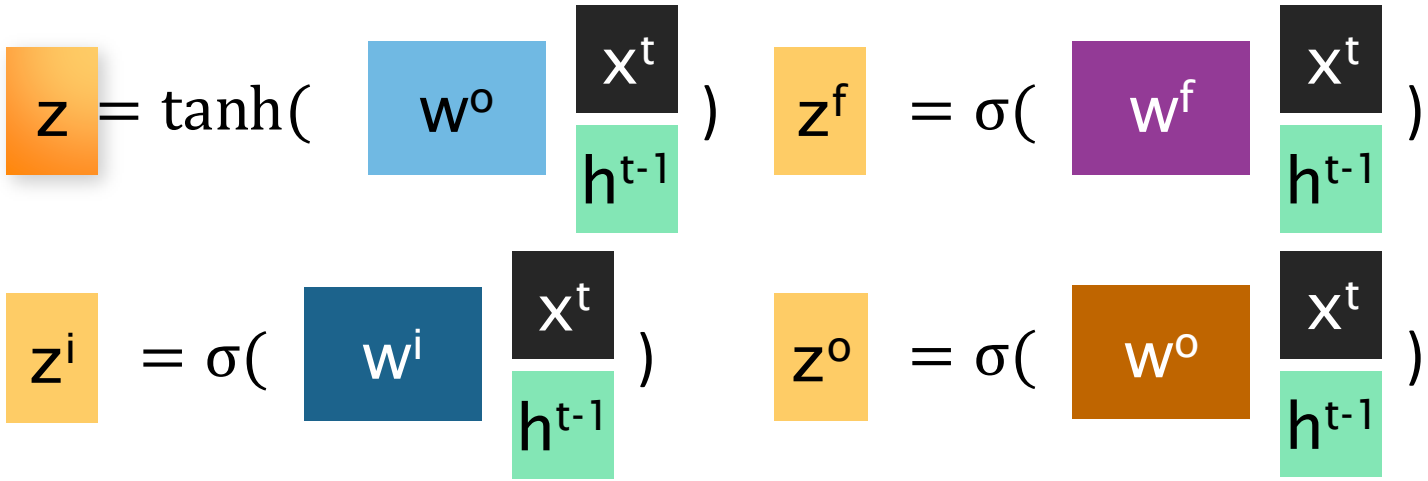




- c（长期记忆）：缓慢更新，c^t通常为上一个c^{t-1}加上一些数值。
- h（短期记忆）：快速更新，h^t在不同时间步下差异较大。

深入LSTM

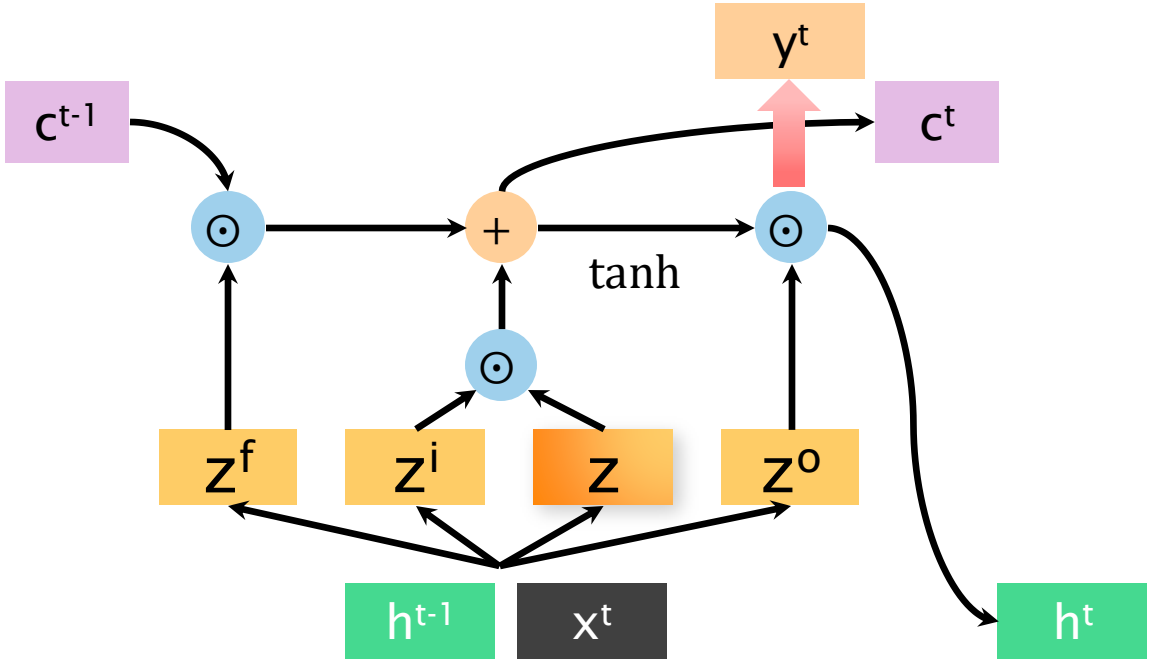
- 将LSTM的当前输入x^t与上一个h^{t-1}拼接，得到四个状态。



- z^f, z^i, z^o由拼接向量乘以权重矩阵后，过一个sigmoid函数得到0~1的数值作为门控状态。
- z过一个tanh函数得到-1~1的值，作为输入值。

深入LSTM

- 四个状态在LSTM中的使用：



$$c^t = z^f \odot c^{t-1} + z^i \odot z$$

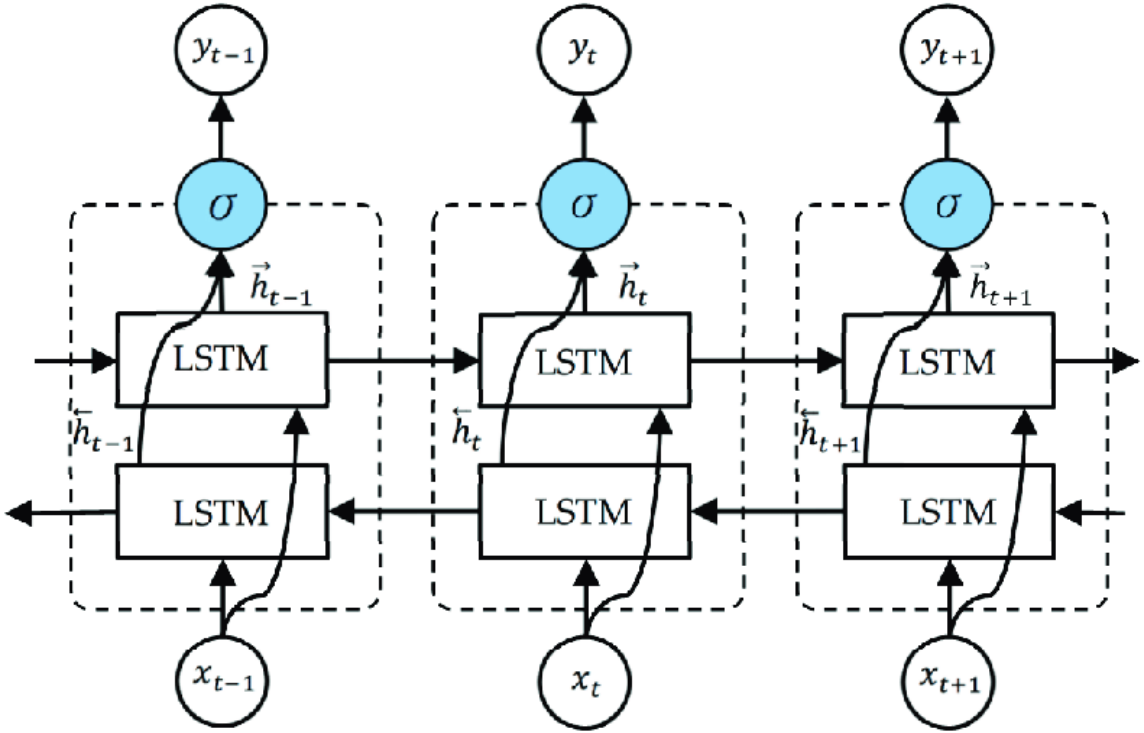
$$h^t = z^o \odot \tanh(c^t)$$

$$y^t = \sigma(W' h^t)$$

- 忘记阶段：z^f作为忘记门控，控制哪些需要留哪些需要忘。
- 选择记忆阶段：z^i作为门控，对输入x^t进行选择性记忆。
- 输出阶段：z^o控制哪些内容被输出。

BiLSTM (Bi-directional LSTM)

- LSTM无法捕捉双向依赖，因此引入BiLSTM



基于BiLSTM的命名实体识别

- 数据集
  - CHisIEC 古文命名实体识别数据集
    - tangxuemei1995/CHisIEC: CHisIEC An Information Extraction Corpus for Ancient Chinese History (github.com)
  - 数据预处理方法：
    - 将每个汉字映射到一个index和一个char embedding
    - 每个label映射到一个index
    - 考虑embedding table中不存在的OOV汉字如何处理

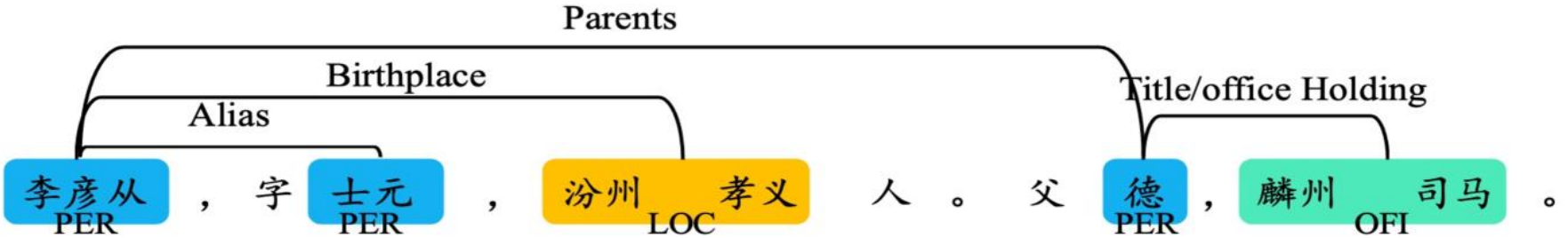
基于BiLSTM的命名实体识别

- 测试集格式

B：Begin，命名实体的开始
I：Inner，命名实体的中间的内容
E：End，命名实体的结束
S：Single，单字命名实体
O: Other，不是命名实体

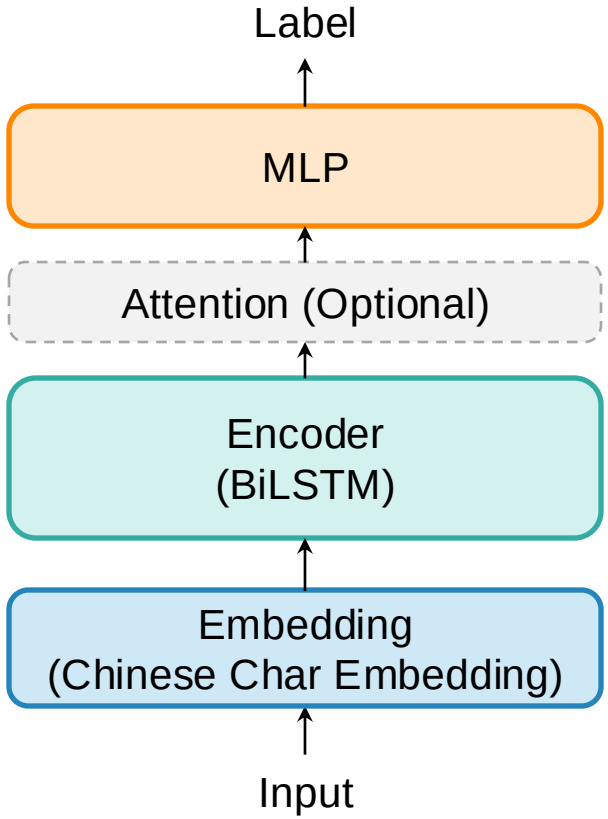
Other 标签的字符非常多，需要考虑如何处理类别不均衡的情况。

此外，实体还需要被划分为PER（人）、LOC（地）、OFI（官）和BOOK（书）四类



基于BiLSTM的命名实体识别

- 模型要求：
  - 使用 PyTorch 2.0 以上版本实现
  - 模型中必须包含 BiLSTM 结构，且该结构的参数量需占模型除 embedding 层以外的绝大多数。
    - 即，不能在里面插入一个大模型，然后LSTM只作为最后分类器的一个小部件。



基于BiLSTM的命名实体识别

- 实验内容（不需要排列组合所有变量，验证集测试）
  - Word embedding方式（尝试不同embedding的效果，可用提供版或者自己寻找更好的版本）
  - Encoder（BiLSTM）
  - BiLSTM后接一层Attention测试效果（可选）
  - 使用网格搜索或随机搜索寻找最佳超参数（可选，超参至少包含学习率和batch size）
  - 比较你所知的模型训练技巧对F1-macro值的影响（可选，训练技巧包括dropout、模型初始化方式、优化器等）

基于BiLSTM的命名实体识别

- 考核方式（DDL 10月13日23:59）
  - 学号-姓名.ipynb
  - ipynb 文件包含模型训练过程的输出日志
    - 包含代码各个部分的说明
    - 包含对Accuracy与F1结果的分析
    - 包含模型结构对F1结果影响的分析
  - 测试集 F1-macro
    - 合格线：0.6
    - 合格线以上，鼓励创新性探索，无需继续过分关注F1-macro指标。如果能找到有效提升F1-macro值的方法，需要在ipynb文件里分析。
  - 严禁在测试集上训练、篡改实验数据等学术不端行为

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \text{Recall} = \frac{\text{TP}}{\text{P}}$$

$$\text{F1} = \frac{2 * precision * recall}{precision + recall}$$

评分标准

- 准时提交文件(有运行过程): 40分
- 测试集分数达到0.6以上: 60分
- 文件中含有相关注释: 70分
- 对模型进行优化改进: 80分
- 对训练过程进行对比分析: 90分
- 有独立思考和发现，性能达到更优，提出改进方向等: 91-100分
- 迟交: 1天内分数*0.6，1天后为0分(如有任何特殊情况，请提前说明，无充分证明时不接受事后解释)

联系我们

- 邮件:
  - 助教-郭虹麟：hlguo24@m.fudan.edu.cn
  - 助教-叶俊杰：jjye23@m.fudan.edu.cn
  - 助教-金森杰：sjjin22@m.fudan.edu.cn