2019 达观杯

文本智能信息提取挑战赛

| 答辩队伍：sk2 | 成员：刘伟棠

CONTENT

01. 赛题和数据介绍

02. 预训练语言模型

03. 信息提取方案

04. 实验结果

PART 01

赛题和数据介绍

比赛任务介绍、数据介绍

信息提取

信息抽取（information extraction），即从自然语言文本中，抽取出特定的事件或事实信息，帮助我们将海量内容自动分类、提取和重构。**文本智能抽取**是信息检索、智能问答、智能对话等人工智能应用的重要基础，它可以克服自然语言非形式化、不确定性等问题，发掘并捕获其中蕴含的有价值信息，进而用于业务咨询、决策支持、精准营销等方面，对产业界有着重要的实用意义。

输入数据

2240_7105_4246_21224_962_13514_2203_17735_16929_4531_5025_17145_9685_6601_6905_2128_21224_9460_3424_19421_5815_6601_18736_21224_6441_7230

输出数据

2240_7105_4246_21224_962_13514_2203_17735/a 16929_4531_5025_17145/b 9685_6601_6905_2128_21224_9460/c 3424_19421_5815_6601_18736/o 21224_6441_7230/a

PART 02

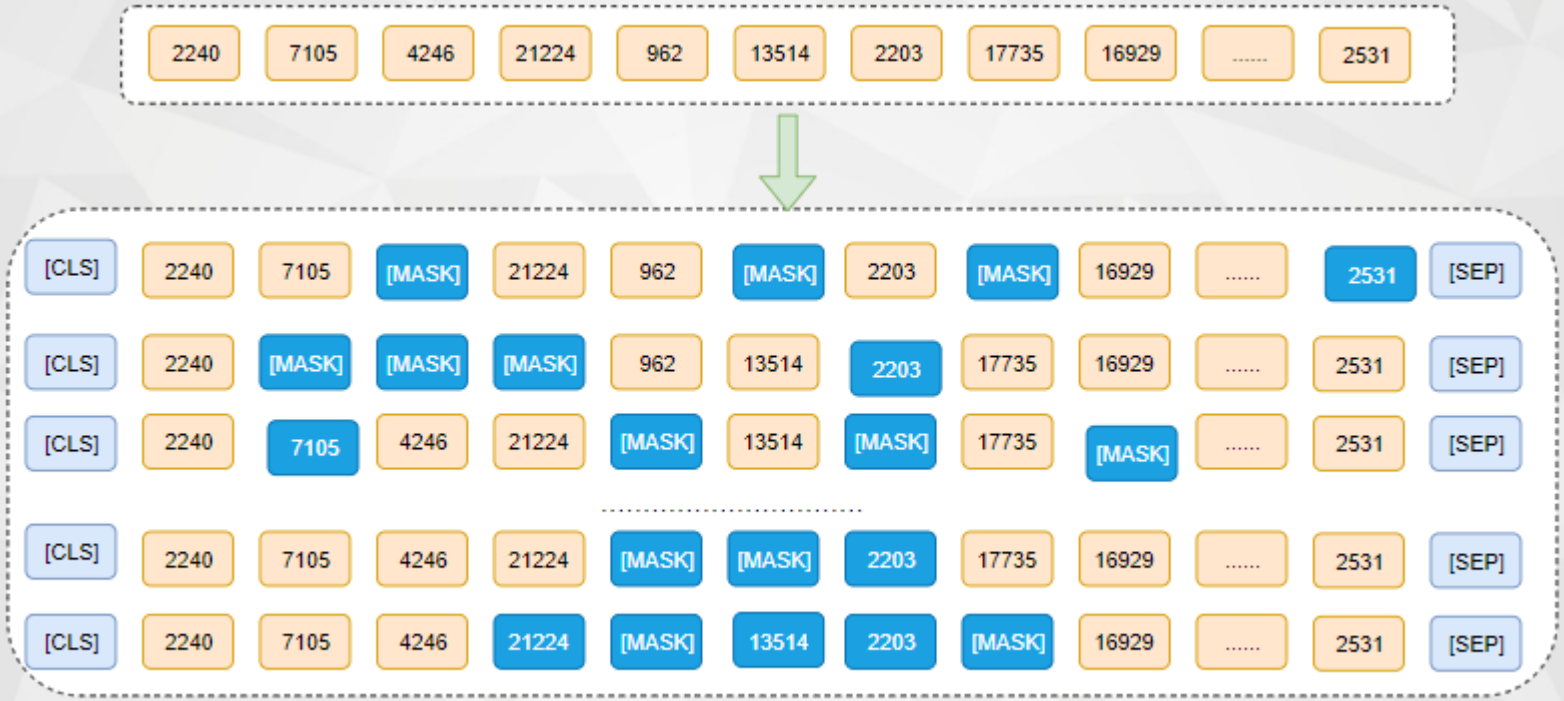
预训练语言模型

对全量数据进行语言模型训练

训练策略：8层BERT

数据

- 语料大小：916786（去重）
- 过短文本：删除长度小于5
- 动态mask：10次
- Input_length: 128
- masked_lm_prob: 0.15
- max_predictions_per_seq: 20
- Mask：
  - 80%: 使用[MASK]代替
  - 10%: 保持原token
  - 10%: 随机任意token代替



模型

{"attention_probs_dropout_prob": 0.2, "hidden_act": "gelu","hidden_dropout_prob":0.2,"num_hidden_layers": 8}

训练

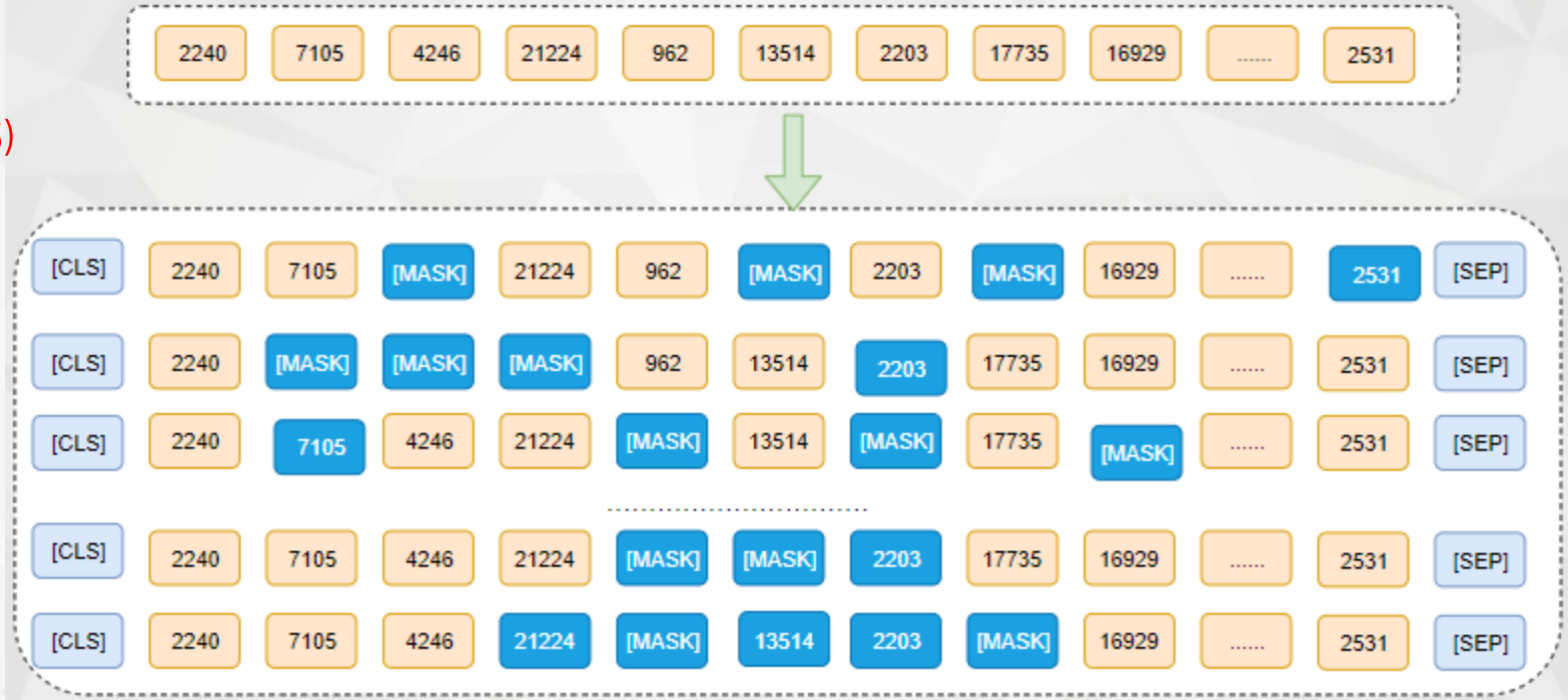
0-2epoch: 学习率 1e-4　　3-4epoch: 学习率 5e-5

开发环境： GPU:GTX1070 CPU: i5 内存：32G 系统：ubuntu16.04

注: 动态mask参考RoBerta，论文地址; https://arxiv.org/pdf/1907.11692.pdf

训练策略：12层BERT

- **数据**
  - 动态mask+混合mask：7次随机+3次n-gram(2-5)
  - 语料大小：916786（去重）
  - 过短文本：删除长度小于5
  - Input_length：128
  - masked_lm_prob:0.15
  - max_predictions_per_seq:20
  - Mask：
    - 80%:使用[MASK]代替
    - 10%:保持原token
    - 10%:随机任意token代替



- **模型**

{"attention_probs_dropout_prob":0.2,"hidden_act":"gelu","hidden_dropout_prob":0.2,"num_hidden_layers":12}

- **训练**

0-2 epoch:学习率 1e-4　　3-4 epoch:学习率 5e-5

开发环境： GPU:GTX1070 CPU: i5 内存：32G 系统：ubuntu16.04

PART 03

信息提取方案

介绍整个信息提取方案以及数据处理

整体方案

数据增强

对原始train数据进行数据增强处理

方案1：BERT+LSTM+CRF

基于BERT+LSTM+CRF的方案1介绍

方案2：BERT+LSTM+MDP+CRF

基于BERT+LSTM+MDP+CRF的方案2介绍

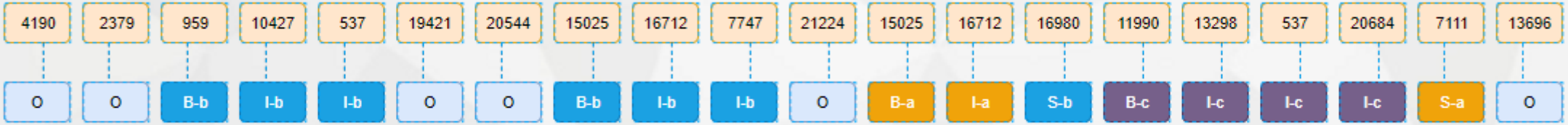
方案3：BERT+LSTM+SPAN

基于BERT+LSTM+SPAN的方案3介绍

数据增强

原始数据

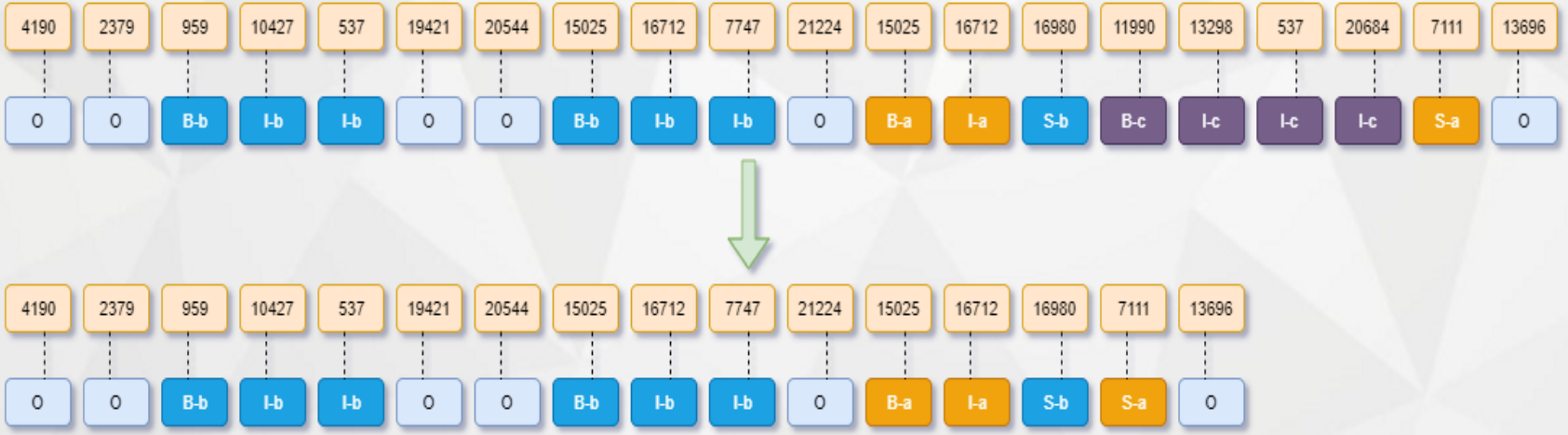
- 对原始数据采用BIOS标注方案



增强方案1

多实体序列中，删除实体c，即：

- oabc → obc
- oac → oa
- obc → ob

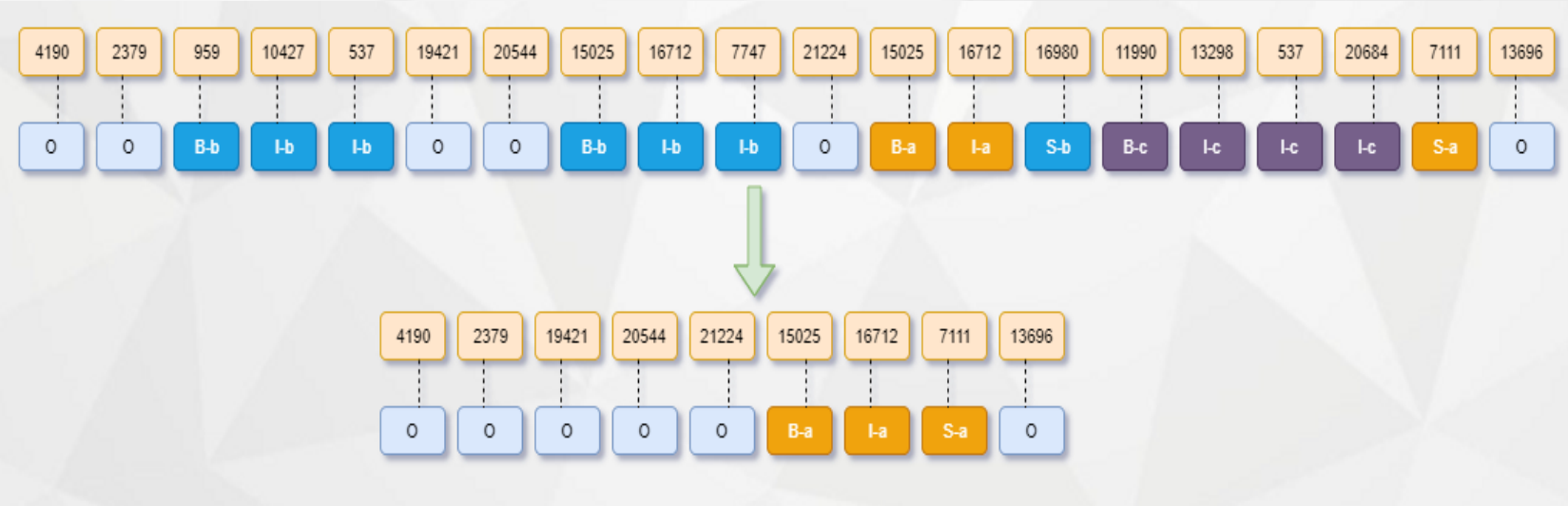


数据增强

增强方案2

包含ab实体句子中，提取a实体，即：

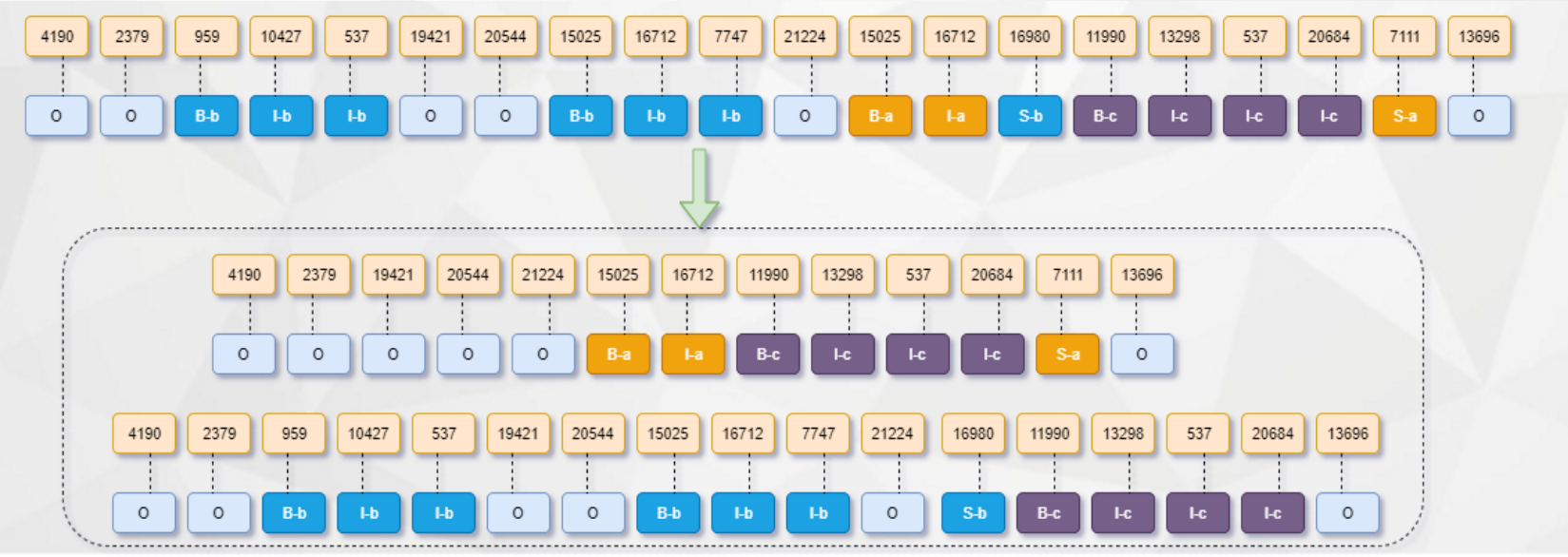
- oabc → oa
- oab → oa



增强方案3

包含abc实体句子中，提取ac和bc实体，即

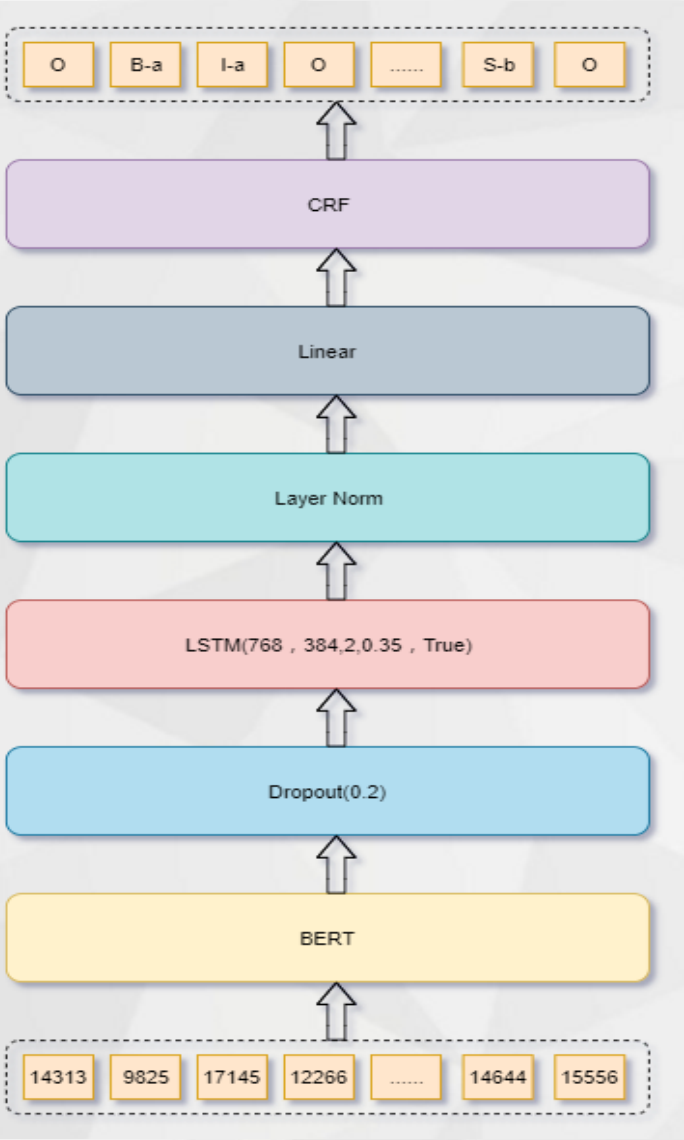
- oabc → obc
- oabc → oac



方案1：BERT+LSTM+CRF

训练策略

- 模型：8层BERT+BiLSTM+CRF
- Train length：148
- Test length：512
- Epochs：30
- Warmup-proportion：0.05
- gradient_accumulation_steps：1
- Weight_decay: 0.01
- Optimizer: adam
- Grad_clip：5.0
- 分层学习率：
  - bert_param: 1e-4
  - lstm_param: 0.001
  - crf_param: 0.001
  - linear_param: 0.001
- Lr scheduler：linear warmup + ReduceLROnPlateau（factor=0.5，patience=5）
  - 详细：当第一次满足patience时，降低lstm、crf和linear层学习为1e-4，并关闭warm-up继续训练

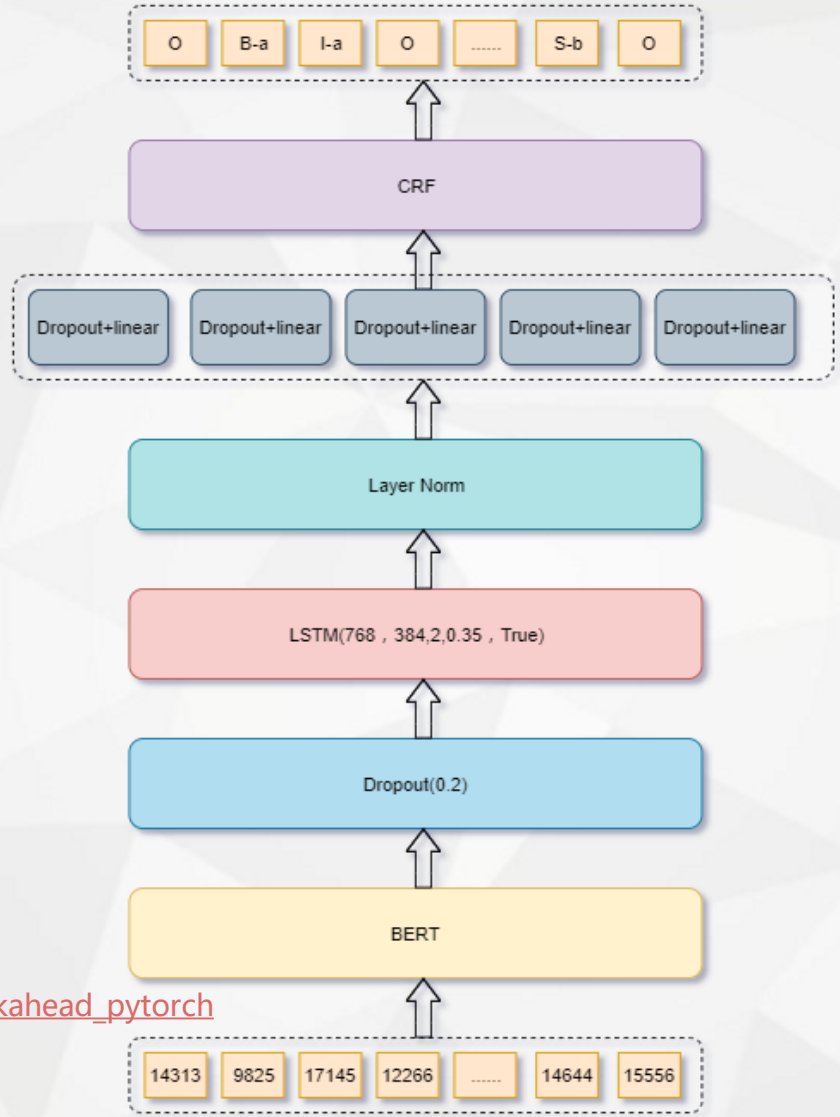


开发环境： GPU:GTX1070 CPU: i5 内存：32G 系统：ubuntu16.04

方案2：BERT+LSTM+MDP+CRF

训练策略

- 模型：12层BERT+BiLSTM+MDP+CRF，dropout rate：0.5
- 输入长度：128
- Epochs：30
- Warmup-proportion：0.05
- gradient_accumulation_steps：1
- Weight_decay: 0.01
- Optimizer: adam + Lookahead
- Grad_clip：5.0
- 分层学习率：
  - bert_param: 1e-4
  - lstm_param: 0.001
  - crf_param: 0.001
  - linear_param: 0.001
- Lr scheduler：linear warmup + ReduceLROnPlateau（factor=0.5，patience=5）
  - 详细：当第一次满足patience时，降低lstm、crf和linear层学习为1e-4，并关闭warm-up继续训练



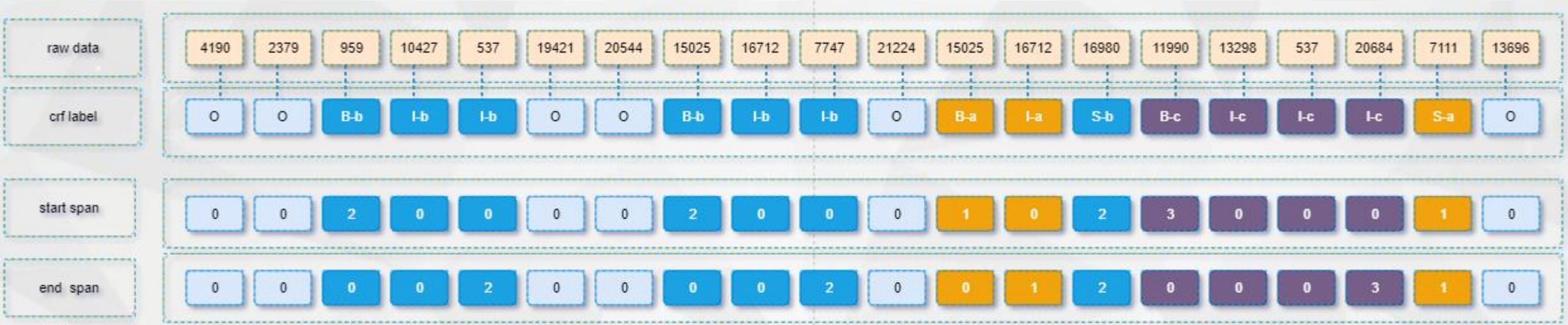
注：1. MDP表示 multi-sample dropout，论文地址：https://arxiv.org/pdf/1905.09788.pdf
2. Lookahead 优化器，论文地址：https://arxiv.org/abs/1907.08610 code：https://github.com/lonePatient/lookahead_pytorch

开发环境： GPU:GTX1070 CPU: i5 内存：32G 系统：ubuntu16.04

方案3：BERT+LSTM+SPAN

数据处理

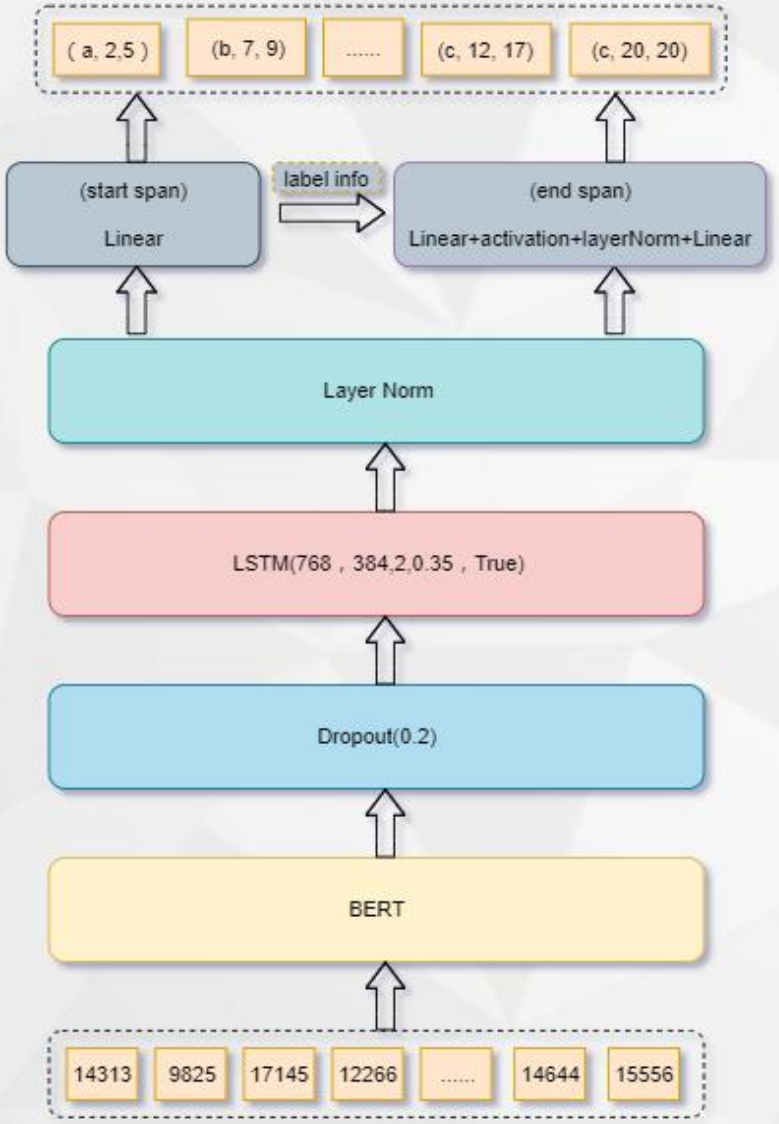
- 数据中存在嵌套关系。
- 常规：bios编码，无法得到嵌套输出
- 指针输出：
  - 类似SQUAD等阅读数据集的构造数据
  - 对每个序列构造start span和end span



方案3：BERT+LSTM+SPAN

训练策略

- 模型：12层BERT+LSTM+SPAN
- 输入长度：128
- Epochs：50
- Warmup-proportion：0.05
- gradient_accumulation_steps：1
- Weight_decay: 0.01
- Optimizer: adam + Lookahead
- Grad_clip：5.0
- 分层学习率：
  - bert_param: 1e-4
  - lstm_param: 0.0005
  - crf_param: 0.0005
  - linear_param: 0.0005
- Lr scheduler： linear warmup + ReduceLROnPlateau （factor=0.5， patience=5）
  - 详细：当第一次满足patience时，降低lstm、crf和linear层学习为1e-4，并关闭warm-up继续训练



开发环境： GPU:GTX1070 CPU: i5 内存：32G 系统：ubuntu16.04

PART 04

实验结果

各个方案在a榜和b榜的实验结果

实验结果

- 5折CV结果提交：

LSTM+CNN

时间：7.17 ~ 8.17

BERT+LSTM+CRF

时间：8.14 ~ 8.31

BERT+LSTM+SPAN

时间：8.31

B榜

最终得分

提交：BERT+LSTM+CRF和BERT+LSTM+MDP+CRF


2019
感谢聆听！ Thanks!