****

**CCKS2020\_NER**

**面向中文电子病历的医疗实体抽取报告**

**队长：刘 研**

**指导老师：刘小明老师**

目录

[一、 任务概述 3](#_Toc44540120)

[1. 任务定义及描述 3](#_Toc44540121)

[1.1. 形式化定义 3](#_Toc44540122)

[1.2. 预定义类别 3](#_Toc44540123)

[2. 数据集描述 4](#_Toc44540124)

[2.1. 数据标注说明 4](#_Toc44540125)

[2.2. 数据示例 4](#_Toc44540129)

[2.3. 数据集描述 5](#_Toc44540134)

[3. 评价指标 5](#_Toc44540135)

[3.1. 严格指标 5](#_Toc44540136)

[3.2. 松弛指标 5](#_Toc44540137)

[二、 时间安排 7](#_Toc44540138)

[三、 数据处理 8](#_Toc44540139)

[1. 数据统计分析 8](#_Toc44540140)

[2. 数据标签化处理 9](#_Toc44540141)

[四、 模型算法 11](#_Toc44540142)

[1. Embedding 11](#_Toc44540143)

[2. Transformer-Encoder 11](#_Toc44540144)

[3. Bi-LSTM 13](#_Toc44540145)

[4. CRF-Decoder 13](#_Toc44540146)

[5. Bi-LSTM + CRF 14](#_Toc44540147)

[五、 模型训练 15](#_Toc44540148)

[六、 结果生成 16](#_Toc44540149)

[七、 结果分析 16](#_Toc44540150)

[八、 存在问题 17](#_Toc44540151)

[九、 总结 17](#_Toc44540152)

# 任务概述

本任务是CCKS围绕中文电子病历语义化开展的系列评测的一个延续，在CCKS 2017，2018, 2019相关评测任务的基础上进行了延伸和拓展。

本任务包括两个子任务：

1）**医疗命名实体识别**

2）**医疗事件抽取**

参赛队可同时选择两个子任务参赛，也可选择任一个子任务单独参赛。

## 任务定义及描述

本任务为中文病历医疗实体识别任务，即对于给定的一组电子病历纯文本文档，识别并抽取出与医学临床相关的实体提及（entity mention），并将它们归类到预定义类别（pre-defined categories），比如疾病、治疗、检查检验等。

相比2019命名实体识别任务我们做了如下调整：

1. 标注了新数据对训练数据进行扩充
2. 提供实体词表及大量非标注数据供参赛者使用

### 形式化定义

**输入**：

1.电子病历的自然语言文本

2.预定义类别：

**输出**：

实体提及和所属类别对的集合：

其中是出现在中的医疗实体提及（mention），和分别表示在中的起止位置，表示所属的预定义类别。要求实体提及之间不重叠，即。

### 预定义类别

预定义类别定义如下：

**1) 疾病和诊断**：医学上定义的疾病和医生在临床工作中对病因、病生理、分型分期等所作的判断。

**2) 检查：** 影像检查（X线、CT、MR、PETCT等）+造影+超声+心电图，未避免检查操作与手术操作过多冲突，不包含此外其它的诊断性操作，如胃镜、肠镜等。

**3）检验**： 在实验室进行的物理或化学检查，本期特指临床工作中检验科进行的化验，不含免疫组化等广义实验室检查

**4) 手术**： 医生在患者身体局部进行的切除、缝合等治疗，是外科的主要治疗方法。

**5) 药物**： 用于疾病治疗的具体化学物质。

**6) 解剖部位**： 指疾病、症状和体征发生的人体解剖学部位。

## 数据集描述

### 数据标注说明

词表及电子病历数据由医渡云（北京）技术有限公司编写，标注数据由医渡云公司组织专业的医学团队进行人工标注，仅限CCKS竞赛评测用。

有关数据集的详细描述及标注规范，将随数据的发布一并加以说明，任务书中不再赘述。



### 数据示例

**{**

**"originalText": "患者3月前因“直肠癌”于在我院于全麻上行直肠癌根治术（DIXON术），手术过程顺利，术后给予抗感染及营养支持治疗，患者恢复好，切口愈合良好。，术后病理示：直肠腺癌（中低度分化），浸润溃疡型，面积3.5\*2CM，侵达外膜。双端切线另送“近端”、“远端”及环周底部切除面未查见癌。肠壁一站（10个）、中间组（8个）淋巴结未查见癌。，免疫组化染色示：ERCC1弥漫（+）、TS少部分弱（+）、SYN（-）、CGA（-）。术后查无化疗禁忌后给予3周期化疗，，方案为：奥沙利铂150MG D1，亚叶酸钙0.3G+替加氟1.0G D2-D6，同时给与升白细胞、护肝、止吐、免疫增强治疗，患者副反应轻。院外期间患者一般情况好，无恶心，无腹痛腹胀胀不适，无现患者为行复查及化疗再次来院就诊，门诊以“直肠癌术后”收入院。"**

**"entities": [**

**{**

**"label\_type": "疾病和诊断",**

**"overlap": 0,**

**"start\_pos": 8,**

**"end\_pos": 11**

**},**

**{**

**"label\_type": "手术",**

**"overlap": 0,**

**"start\_pos": 21,**

**"end\_pos": 35**

**},**

**{**

**"label\_type": "疾病和诊断",**

**"overlap": 0,**

**"start\_pos": 78,**

**"end\_pos": 95**

**}**

**]**

**}**



### 数据集描述

本次评测的训练数据有：

1. 1500条标注数据
2. 1000条非标注数据。
3. 6个类别的6292个实体词词表

标注数据集统计如下表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 文本 | 疾病和诊断 | 检查 | 检验 | 手术 | 药物 | 解剖部位 | 总数 |
| 训练集 | 1500 | 6211 | 1490 | 1885 | 1327 | 2841 | 12660 | 26414 |

## 评价指标

本任务采用精确率（Precision）、召回率（Recall）以及F1-Measure作为评测指标。参赛系统的输出结果集合记为，人工标注的结果（Gold Standard）集合记为。集合元素为一个实体提及，表示为四元组，表示文档，和分别对应实体提及在文档中的起止下标，表示实体提及所属预定义类别。分别从两个层面进行评价。

### 严格指标

我们定义与严格等价，当且仅当：

基于以上等价关系，我们定义集合与的严格交集为。由此得到严格评测指标：

, ,

### 松弛指标

我们定义与松弛等价，当且仅当：

基于以上等价关系，我们定义集合与的松弛交集为。由此得到松弛评测指标：

, ,

最后，按照预定义类别的6个不同类别，对每个子类进行分开评测，共得到14个评测结果：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 疾病和诊断 | 检查 | 检验 | 手术 | 药物 | 解剖部位 | 综合 |
| 严格指标 |  |  |  |  |  |  |  |
| 松弛指标 |  |  |  |  |  |  |  |

# 时间安排

根据最新安排，比赛时间调整如下

报名时间：**3 月 20 日 - 9 月 10 日 08 时 00 分 00 秒**

签署数据使用承诺书：**3 月 20 日 - 9 月 10 日 08 时 00 分 00 秒**

测试数据发布：**9 月 15 日**

提交测试结果截止时间：**9 月 30 日 08 时 00 分 00 秒**

评测论文提交：**10 月 25 日**

CCKS会议日期(评测报告及颁奖)：**11 月 12 日 — 15 日**

**个人进度安排如下：**

**考完试当天报名参赛：6月10日**

**签署数据使用承诺书：6月15日**

**第一次提交有效结果：6月16日**

**目前成绩为：松弛指标F1值为：92.7%**

**严格指标F1值为：82.5%**

# 数据处理

## 数据统计分析

原文长度统计如下表和柱状图所示：

表3-1 原文长度统计表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 原文句长范围 | 0-199 | 200-399 | 400-599 | 600-799 | 800-999 | 1000-X | 总计 |
| 句子数 | 30 | 553 | 369 | 71 | 19 | 8 | 1050 |

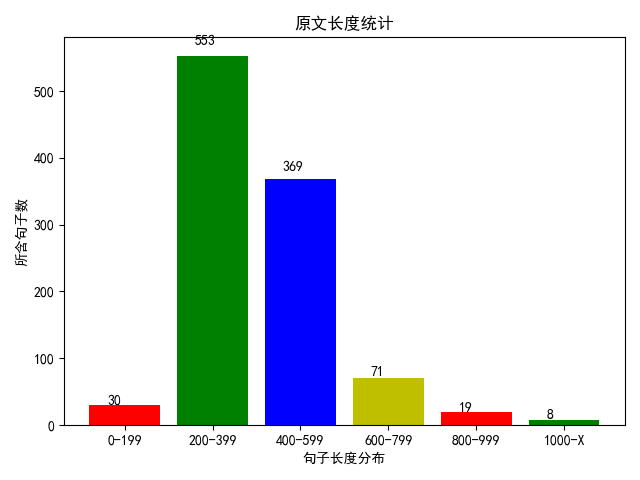


图3-1 原文长度统计柱状图

标注数据集统计如下表和柱状图所示：

表3-2 医疗实体分布统计表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 文本 | 疾病和诊断 | 检查 | 检验 | 手术 | 药物 | 解剖部位 | 总数 |
| 训练集 | 1050 | 4345 | 1002 | 1297 | 923 | 1935 | 8811 | 18313 |

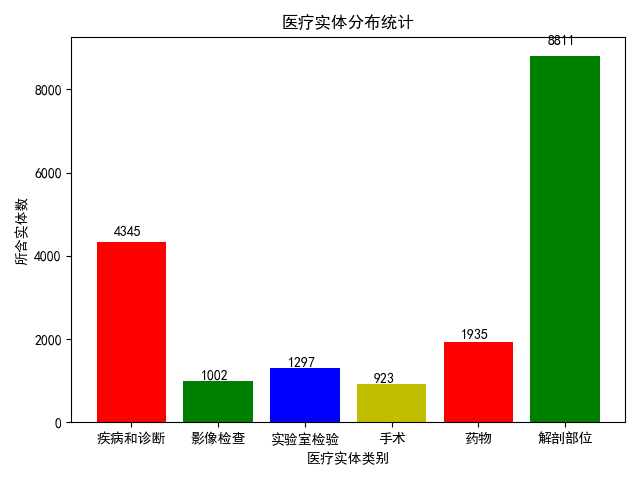


图3-2 医疗实体分布统计柱状图

## 数据标签化处理

对官方提供的task1\_train.txt文件进行分析处理，首先将数据按照json格式进行读取，存取成list集合的形式，然后使用BIO标签对医疗实体进行标注。如图3-3所示为数据标签化处理部分截图：

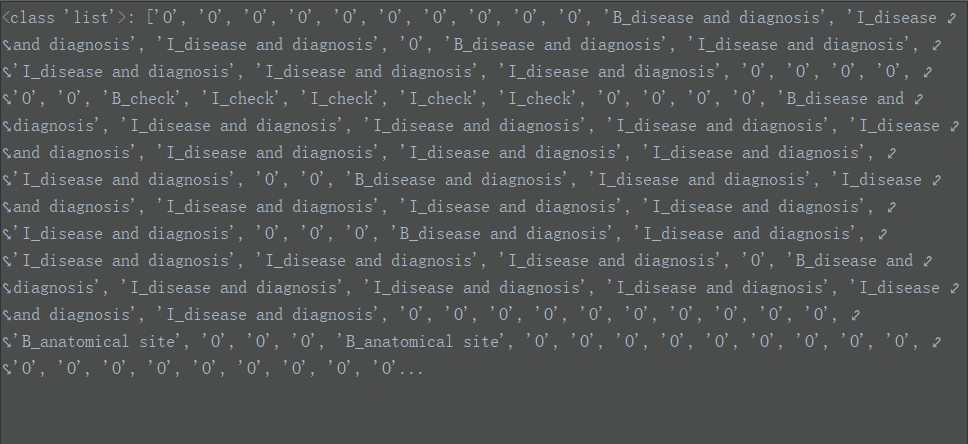
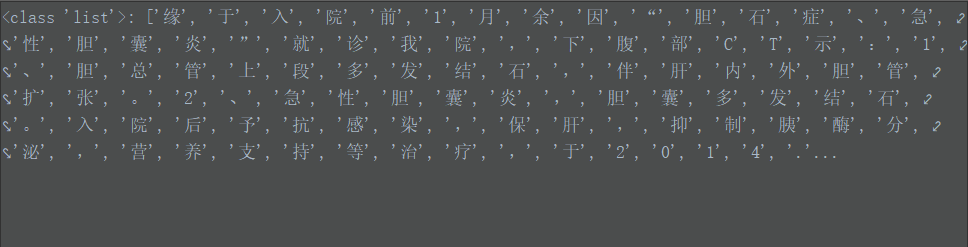
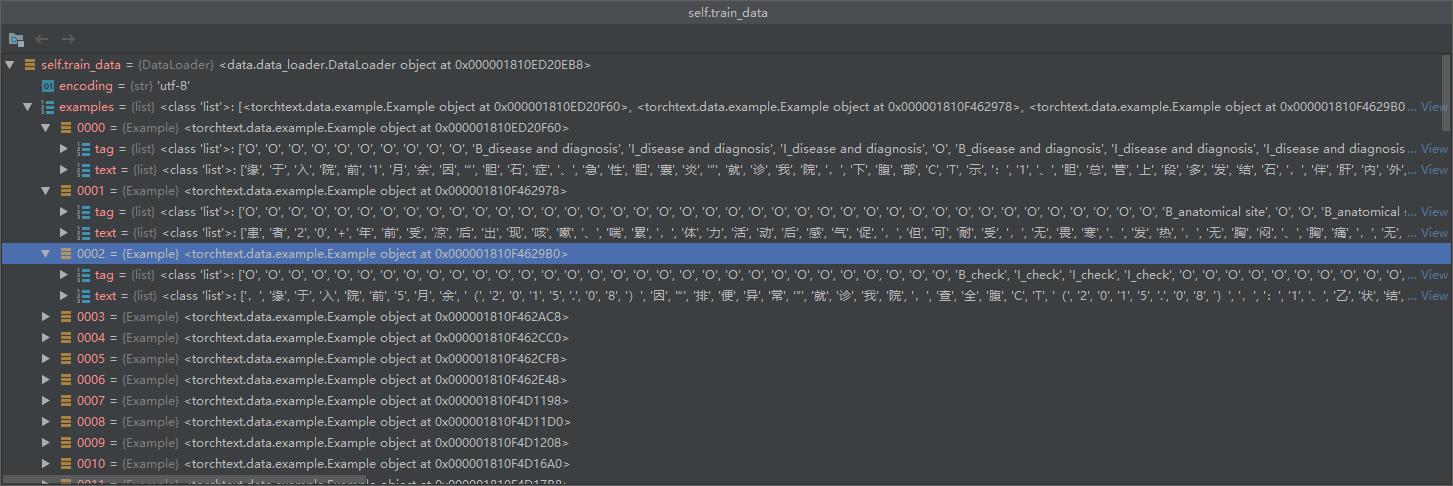


图3-3 数据标签化处理

# 模型算法

Baseline模型采用Transformer-Encoder + Bi-LSTM + CRF

Baseline模型效果： **松弛指标F1值为：92.7%**

**严格指标F1值为：82.5%**

## Embedding

采用百度预训练的词向量baidubaike.bigram\_char初始化embedding，维度为300维。使用方法为：Skip\_Gram with Negative Sampling的方法。

采用BERT预训练的词向量初始化embedding，维度为768维，词表大小为所有文本。

## Transformer-Encoder

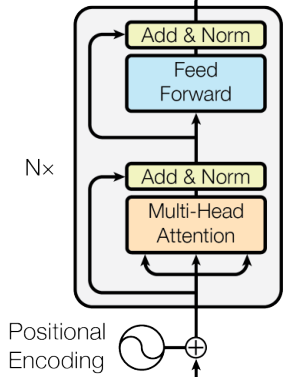


图4-1 Transformer Encoder

1. 一步的Embedding之后，此时输入数据的size=(300,batch,text\_len)
2. 由于数据中有pad，所以输入网络之前需要标记pad的位置（pad处设为True）
3. Positional Encoding：由于输入的数据经过self-attention后没有位置信息，所以需要在这里需要加上位置信息
4. Multi\_Head Attention：是一个多头的self-attention，能够跨越距离的限制更加准确地提取字符特征，以达到较好的Encoder效果
5. Feed Forward：pytorch封装后，未看出使用哪种Feed Forward，但之前看到的实现例子是通过CNN来实现的，在我的理解也是一种特征抽取，只不过两者不同，CNN是局部特征抽取，self-attention算是针对于每个字符的全局特征抽取。

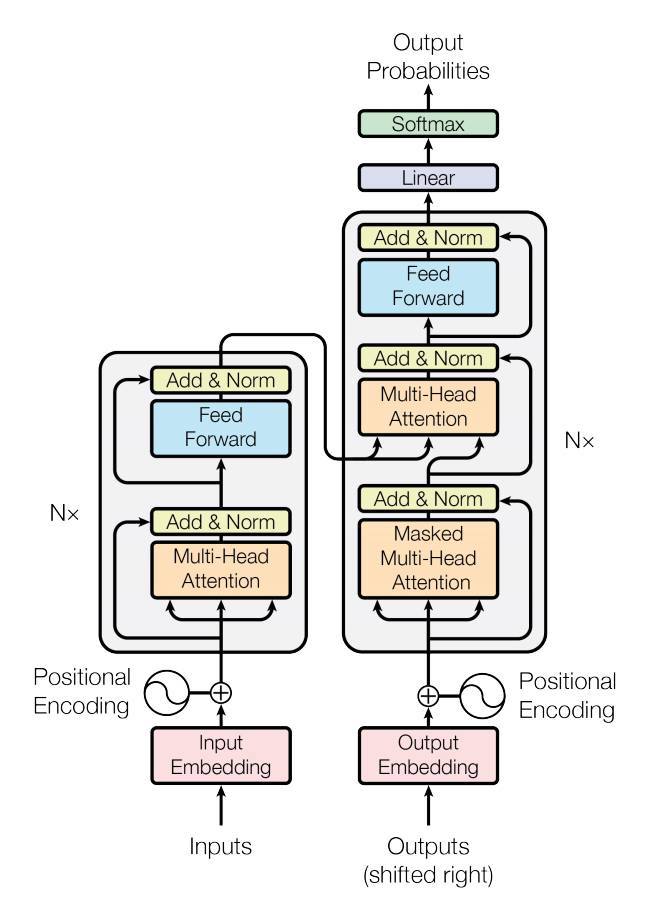


图4-2 transformer\_architecture

## Bi-LSTM

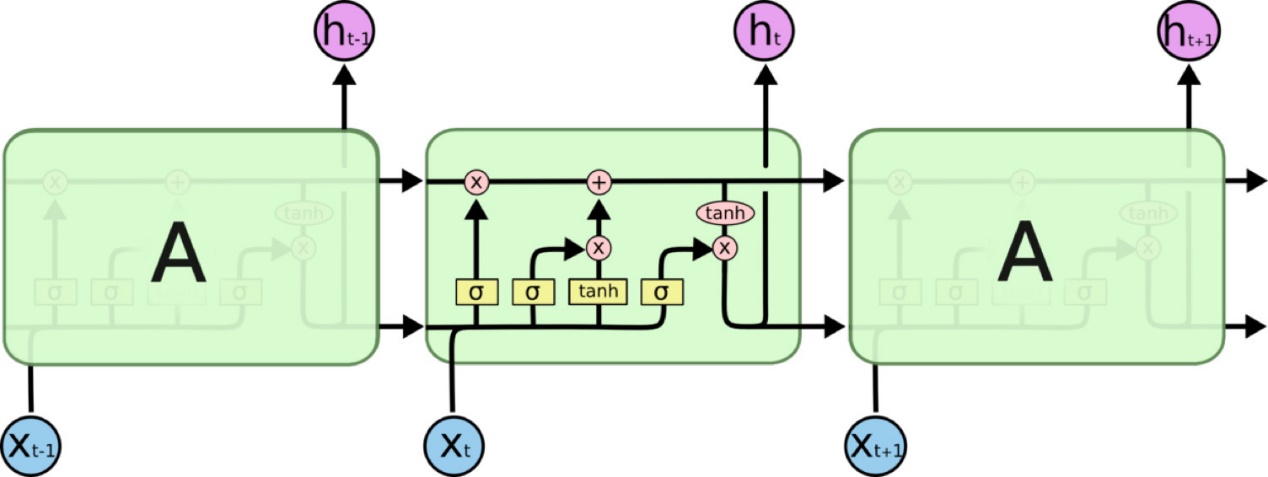


图4-3 LSTM模型

LSTM通过三个门结构（输入门，遗忘门，输出门），选择性地遗忘部分历史信息，加入部分当前输入信息，最终整合到当前状态并产生输出状态。

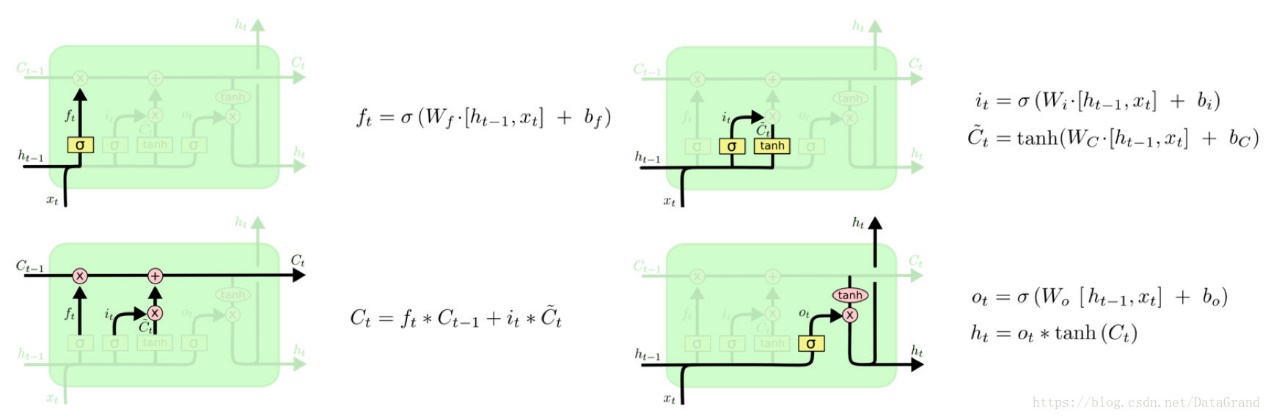


图4-4 LSTM模型概要

## CRF-Decoder

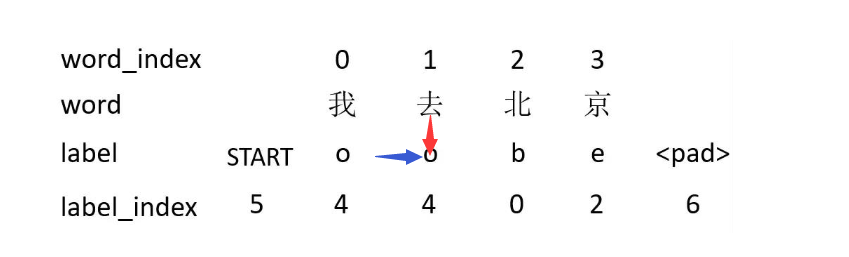


图4-5 CRF Decoder

CRF通过构建一个概率图模型，通过训练得到一个由word\_i到tag\_i的发射概率图和tag\_i-1到tag\_i的状态转移概率图。在数据足够的情况下，CRF能够很好的进行tag的预测。但可能得不到最好的效果，Decoder部分可以换用更适合的模型组合（eg: LSTM+CRF）。

## Bi-LSTM + CRF

应用于NER中的Bi-LSTM + CRF模型主要由Embedding层（主要有词向量，字向量以及一些额外特征），双向LSTM层，以及最后的CRF层构成。实验结果表明Bi-LSTM + CRF已经达到或者超过了基于丰富特征的CRF模型，成为目前基于深度学习的NER方法中的最主流模型。在特征方面，该模型继承了深度学习方法的优势，无需特征工程，使用词向量以及字符向量就可以达到很好的效果，如果有高质量的词典特征，能够进一步获得提高。

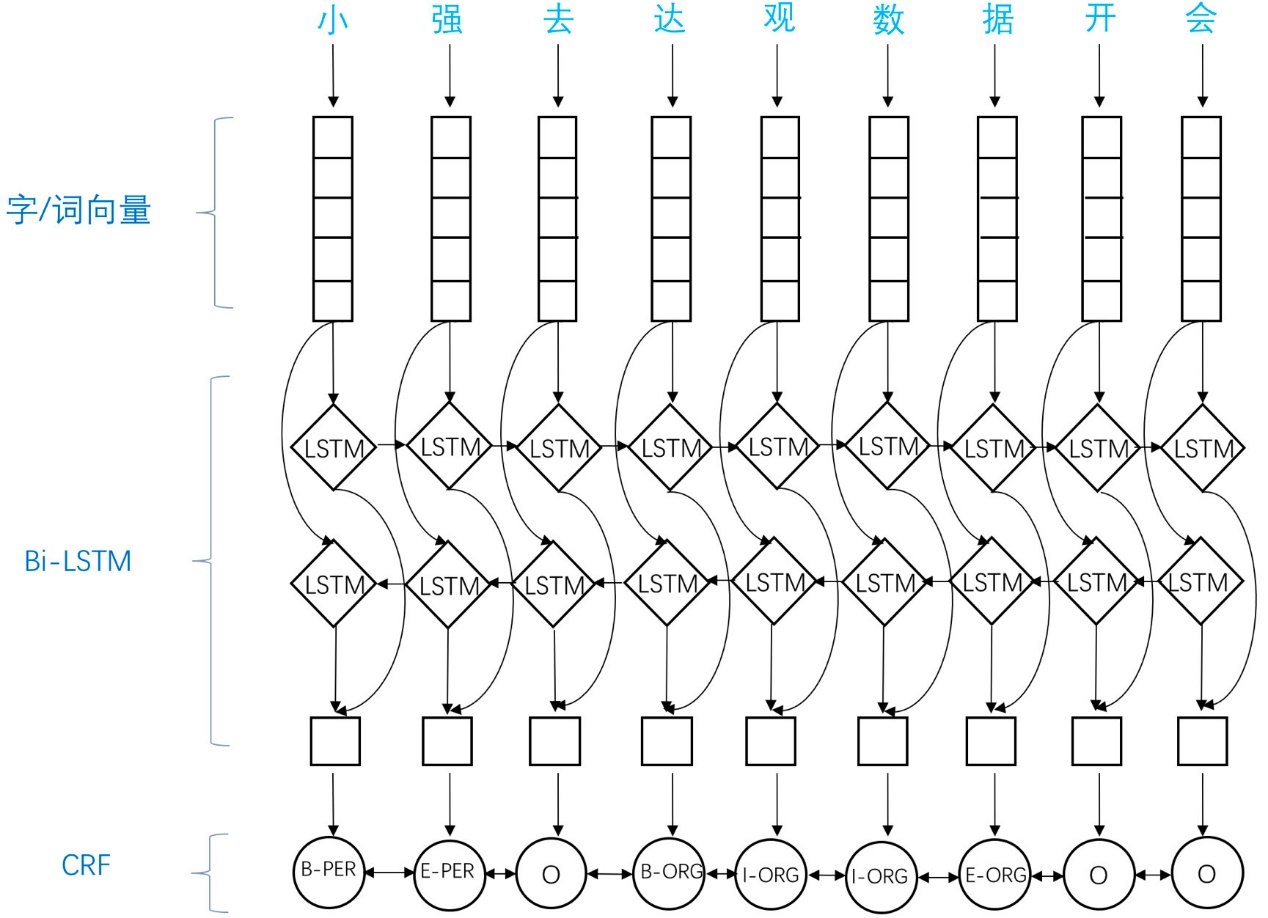
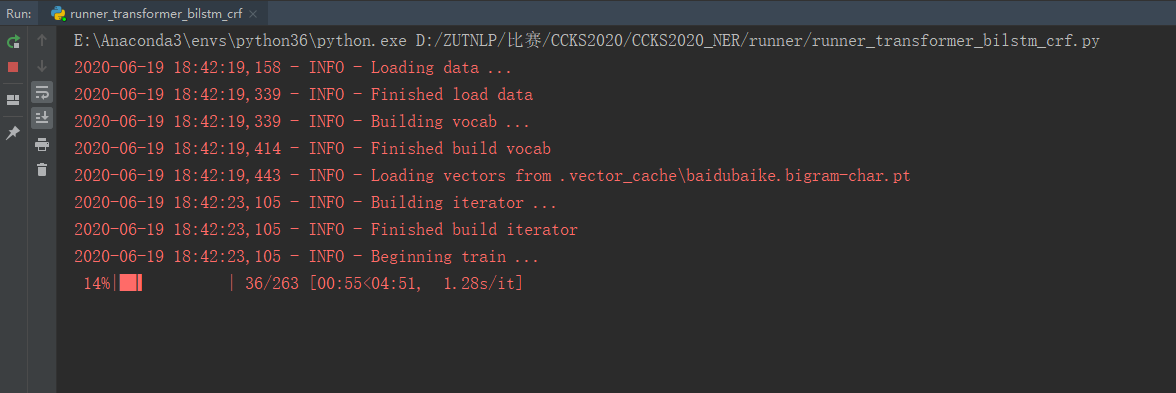


图4-6 Bi-LSTM + CRF模型

# 模型训练



1、Loading data ...

2、Building vocab ...

3、Loading vectors from .vector\_cache\baidubaike.bigram-char.pt

4、Building iterator ...

5、Beginning train ...

参数设置如下：

is\_cuda: True,

epoch: 100,

batch\_size: 4,

learning\_rate: 0.0002,

n\_hid: 200,  # the dimension of the feedforward network model in nn.TransformerEncoder

n\_layers: 2,  # the number of nn.TransformerEncoderLayer in nn.TransformerEncoder

n\_head: 2,  # the number of heads in the multiheadattention models

dropout: 0.2,  # the dropout value

vector\_win\_path:  D:/ZUTNLP/zutnlp/medical/baidubaike/baidubaike.bigram-char,

vector\_linux\_path:  /home/zutnlp/liuyan/CCKS\_EE/vector/baidubaike.bigram\_char,

em\_size: 300,  # embedding dimension     预训练模型：hidden 1024/786   word2voc：300

bi\_lstm\_hidden: 300,

loss：crf预测出目标tag的分数

optimizer：Adam

# 结果生成

1. 将训练好的模型载入
2. 将预测数据以json的方式读入内存，做成list列表
3. 遍历list列表，将每一条预测数据的原文做成tensor并送入已训练好的神经网络中去，拿到预测结果并根据标签还原成比赛所要求的的数据形式
4. 将预测结果写入result.txt文件中即可

# 结果分析

**目前成绩为：松弛指标F1值为：92.7%**

**严格指标F1值为：82.5%**

比赛榜首F1值为91%，第二名F1值为85.5%，由此看来，baseline的模型效果还算乐观，但是和第一名大佬相比，还有很大的差距，在之后的两个月中，我会多尝试其他模型，按照刘小明老师的提议，修改模型以及完善细节性问题。



# 存在问题

1. 官方提供的vocab使用导致F1值不升反降
2. 模型优劣待分析验证
3. 数据不均衡问题
4. 数据不标准问题（以逗号为首或双逗号连着）

# 总结

时间匆匆不等人，感觉像是一瞬间，大三的下学期就要过去了，回首这个学期，因为疫情的原因，我们没能和老师们如期而至的在学校见面，但是我们也在家里按部就班的上着网课，与此同时，我的实训内容也在进行着，在15周之前，我参加了2020语言与智能技术竞赛中的关系抽取任务赛道，这个任务是在2020/5/20截止提交测试结果，然后我开始了期末复习，在17周周三（6月10日），我的期末考试圆满结束，我就给CCKS2020\_NER的主办方清华大学知识工程实验室的张江涛老师发送了参赛报名的邮件，也是在18周的周一（6月15日）收到张老师的回复，成功报名并签署数据使用承诺书，然后我就对我的baseline模型开始进行设计，测试，实验工作。

通过参加本次CCKS2020\_NER的比赛，让我们学到了很多普通开发学不到的知识，接触到了面向中文电子病历的医疗实体抽取，当今社会的需求已远远不仅仅是局限于普通的应用开发，而是想让机器“活”起来，也真的实实在在把这项技术不断突破并应用到实用的生活中，对于一个特定的领域，机器也可以通过人工的方式进行训练，变得更加智能化，更加高效的服务于人类。作为一个计算机系软件工程的学生，身处信息化时代，知识日新月异，让我对IT行业的发展趋势有了更深的了解。

对于收获，这个过程确实学习到了很多关于神经网络的知识，有了很大的提升，但相比于知识的收获，又有其他方面起到了很大的作用，其中就包括问题的解决和新知识的获取途径与学习方法。（1）问题解决：要抱有一个很正常的心态去正面对待问题，因为问题总会有，也不可能解决完，只有一个很正确的态度和解决思路才是这个过程最需要掌握的东西，这点还需要继续加强，心态一定要平静。（2）新知识的获取：这学期很大一部分都是老师的知道，老师有很大的宏局观，这也是知识储备的象征，所以知识获取一面方需要老师的指导，一方面也需要靠自己学会如何搜集知识。（3）学习方法：老师给我们一个新的思路改变了以前的学习固定思维，在面对较难的问题时，可以尝试先实践或者找到已实践的，然后进一步学习，这样有时候会很大的提升你的效率，减少迷茫，也会给你很多想象，加深理解，在适当的情况下不失为一种很棒的学习思路。

在参与本次比赛的过程中，也让我学会了克服困难，学会了自学的方法，知道了一项技术的官方文档的重要性，深刻了团队共同开发的好处和团结协作的必要性，一个团队的成员在一起共同学习，营造一个良好的研究开发环境的重要性，大家有理解不透彻的地方，在一起互相帮助的温暖与开心。

在报告的最后，我谨代表我们小组全体成员对刘小明老师表示由衷的感谢！刘老师在指导过程中有重点有计划，才使得我们的学习、研究与开发工作顺利开展。路漫漫其修远兮，吾将上下而求索。我会带领着整个AI团队不忘初心，继续奋斗，坚持始终是我的信念！