UTF8gbsn
bio 在验证集上的准确率 934 单条证据质量 Doc-Start

神经循环序列标注模型

神经网络用于阅读理解和问答系统

zhaojun HAO

14 Dec 2017

目录

目录

智能问答系统

提一个问题，问答系统从可以找到的证据中搜索靠谱的答案.
问答系统要求具备阅读理解，推理等能力.

现状

深度学习用来和实现问答系统

近年来, 利用深度学习实现问答系统的模型层出不穷. 多层长短期记忆网络 (LSTM),Attention (注意力机制),条件随机场 (CRF)的应用都带来了阶段性的成果.

实现方式

- Sequence Generation (序列生成), 逐字地生成答案
- Classification/Ranking problem (分类/排序问题), 在提前给定的答案集中排名后选择

现存方法的问题

序列生成模型的问题

在一个很大的字典上选字逐字逐字生成，计算复杂度高，无法形成新词

分类/排序问题模型的问题

无法处理新问题，依赖提前给定的答案集的质量或者需要一个额外的生成答案集的部件

问答集的问题

大部分模型都是在小规模的人工生成的数据集上，不符合现实

创新

数据库

IDL (百度深度学习研究院)利用基于**百度问答**等资源并人工标注了Begin-Inside-Outside Label (’开头-内部-外部’标签)创建了叫做 **WebQA** 的数据集

模型

模型使用了序列标注的方式从开放的答案集中，通过给词定标签来提取答案. 这次 talk 要讲内容.

目录

模型图示

简单的图示说明模型，模型作用于每条证据

第一部分: 阅读问题	第二部分: 理解问题的基础上阅读证据	第三部分: 序列标注
魔镜魔镜,谁是世界上最美的女人?	王后想成为世界上最美的女人,但白雪公主才是	王后想成为世界上最美的 O O O O O O O O O O O O 女人,但白雪公主才是 O O O B I I I O O

Figure: 画图说模型

模型输出到问答系统

问题: 魔镜魔镜, 谁是世界上最美的女人?

↓

找到四条证据

1. 王后想成为世界上最美的女人, 但白雪公主才是
2. 老婆大人是最美的女人
3. 童话里的白雪公主吧
4. 当然是老娘我了!

↓

四条证据提取的答案:{白雪公主, 老婆, 白雪公主, 老娘我 }

↓ 平均投票

白雪公主

原始数据

{问题, 问题 _ID, 黄金答案, 证据, 证据 url}

例子

问题: 魔镜魔镜, 谁是世界上最美的女人?
黄金答案: 白雪公主
四条证据:

1. 王后想成为世界上最美的女人, 但白雪公主才是
2. 老婆大人是最美的女人
3. 童话里的白雪公主吧
4. 当然是老娘我了!

输入

- 问题单词的词向量，用 X_Q 表示.
- 证据单词的词向量，用 X_E 表示.
- 证据单词的E-E Common (词的证据间共存性)，用 display illustration J1 表示. 如果该单词在同一问题的其他证据内也存在，则取值为 1，否则 0.
- 证据单词的E-Q common (词的证据问题共存性)，用 display illustration J2 表示. 如果该单词在问题内也存在，则取值为 1，否则 0.

词向量

将词映射成向量,one hot 向量一列代表字典里面一个字, 每个字的 one hot 向量只在这个字所在的列为 1, 其他为 0

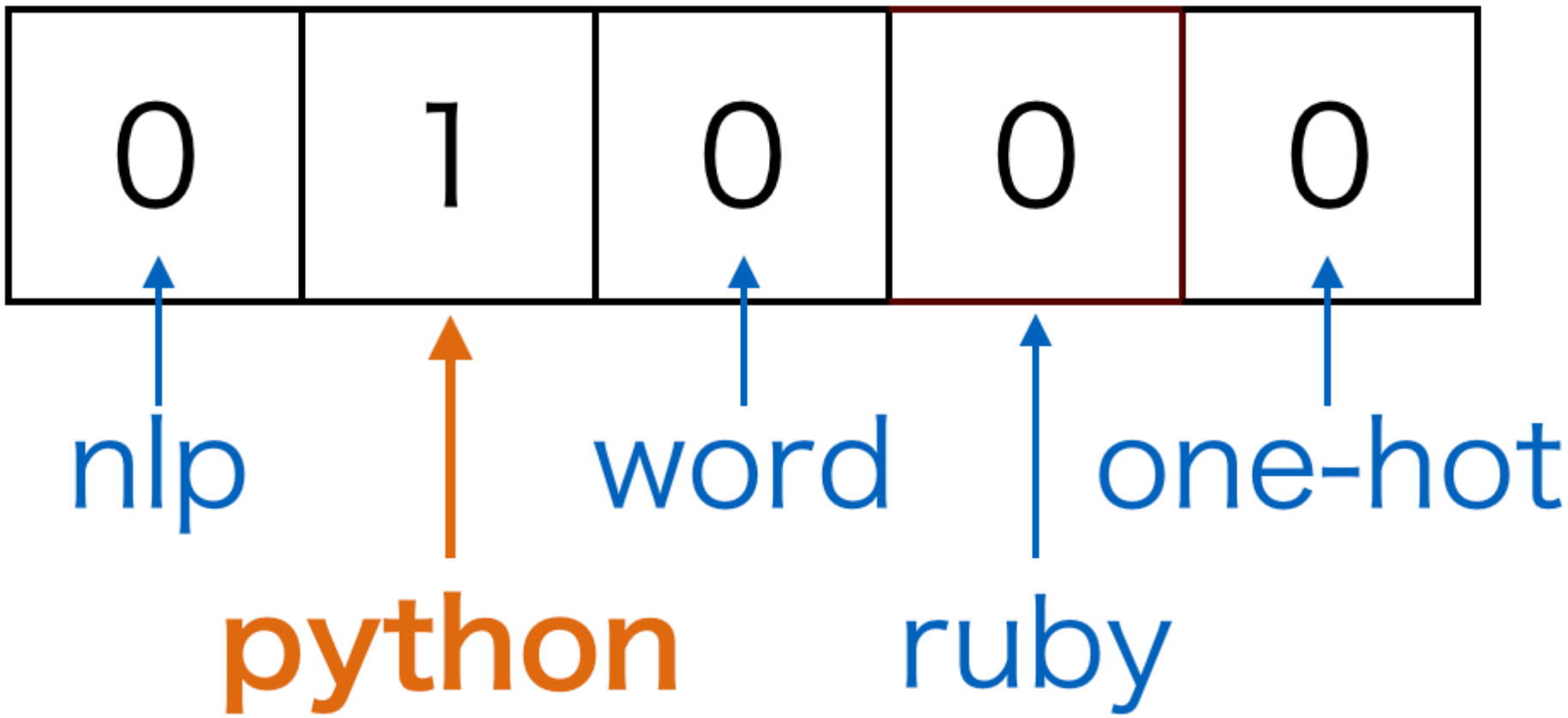

Figure: 词向量

证据中每个词都有两种额外的特征 J1 和 J2, 简单的图示说明 J1J2

问题	证据的J2特征(词是否出现问题中)	证据的J1特征(词是否出现在其他证据中)
魔镜魔镜,谁是世界上最美的女人?	王后想成为世界上最美的女 0 0 0 0 0 1 1 1 1 1 1 人,但白雪公主才是 1 0 0 0 0 0 0 0	王后想成为世界上最美的 0 0 0 0 0 0 0 0 1 1 1 女人,但白雪公主才是 1 1 0 0 0 0 0 0 1
	老婆大人是最美的女人 0 0 0 0 1 1 1 1 1 1	老婆大人是最美的女人 0 0 0 1 1 1 1 1 1 1

Figure: 画图说 J1J2

我们模型中目标输出

在此次实验中，目标变量证据中单词的’开头-内部-外部’标签也是机器生成，用 Y 表示，

- 如果证据单词出现在黄金答案开头，则用 b 表示
- 如果证据单词出现在黄金答案内部，则用 i 表示
- 否则，用 o 表示

为什么’开头-内部-外部’标签

答案往往出现在证据内部的连续位置，所以考虑用’开头-内部-外部’标签标签来标记证据序列中对应于答案的子序列的连续位置. 这样我们可以从证据中提取出来答案.

输出例子

输出为’开头-内部-外部’标签，进而把标注为"b""i" 的词提取出来作为答案.

第一部分: 阅读问题	第二部分: 理解问题的基础上阅读证据	第三部分: 序列标注
魔镜魔镜,谁是世界上最美的女人?	王后想成为世界上最美的女人,但白雪公主才是	王后想成为世界上最美的 O O O O O O O O O O O O 女人,但白雪公主才是 O O O B I I I O O

Figure: 画图说模型

模型主体结构

1. 问题训练
2. 证据训练
3. 词’开头-内部-外部’标签训练

其中前两部分为阅读理解，最后一个部分为序列标注.

模型第一部分: 问题训练

怎么做

$X_q \xrightarrow{\text{LSTM+注意力机制}} R_q$

X_q 是机器运用LSTM后提取出的问题的特征. 通俗地讲,R_q 就是对问题的理解.

为什么LSTM+注意力机制

LSTM模型可以帮助我们理解词序列的传递关系, 进而理解词序列.
但是传统的LSTM处理序列后得到的只是最后一个词的信息,注意力机制的引入相当于给词序列中各个词的信息做个线性组合, 使得我们可以综合考虑词序列中每个词的信息, 增加了对词序列的理解, 解决了序列过长带来的阅读理解障碍问题.

模型第二部分: 证据训练

第一层LSTM

$[X_E; R_q; J1; J2] \xrightarrow{\text{LSTM}} \text{H1}$

证据的词向量 X_E 结合机器提取的问题特征 R_q 和人工提取的特征 $J1, J2$, 得到该证据的每个词的第一层特征 $H1$.

第二层逆向LSTM

$H_1 \xrightarrow{\text{逆向LSTM}} \text{H2}$

证据词序列的第一层特征 H_1 作为输入, 逆向训练出该证据的第二层特征 $H2$.

第三层交叉层LSTM

$[H1 : H2] \xrightarrow{\text{LSTM}} \text{H3}$

证据词序列的第一层特征出 $H1$ 和第二层特征 $H2$ 合并作为输入, 训练出该证据的每个词的第三层特征 $H3$.

第二部分为什么这么复杂

第二部分依次使用了正向LSTM, 逆向LSTM和交叉层LSTM是因为LSTM体现的是序列的正向传递关系. 词序列中不仅后面的词受前面的词影响, 也会受到后面的词影响.

模型第三部分:CRF标签训练

怎么做

对每条证据 + 问题做阅读理解得到的 $H3$ 作为输入，加入CRF，来预测最可能的’开头-内部-外部’标签. 进而获得最可能的答案..

为什么CRF

使用CRF方法，将

1. 每个词贴’开头-内部-外部’标签的数值
2. 每个词的’开头-内部-外部’标签跳转到下一个词的’开头-内部-外部’标签的数值

转化为概率，训练出符合最大似然的’开头-内部-外部’标签序列，来标注该条证据
这里" 可能" 指的是证据的整个词序列的条件似然，而不是序列生成中的单个单词的条件似然.

目录

根据每条证据’开头-内部-外部’标签来提取答案

得到每条证据的’开头-内部-外部’标签后，我们把对应于”B” 和”I” 标签的词/字提取出来，作为每条证据对应的答案.

例子

第一部分: 阅读问题	第二部分: 理解问题的基础上阅读证据	第三部分: 序列标注
魔镜魔镜,谁是世界上最美的女人?	王后想成为世界上最美的女人,但白雪公主才是	王后想成为世界上最美的 O O O O O O O O O O O O 女人,但白雪公主才是 O O O B I I I O O

Figure: 画图说模型

这里我们提取出” 白雪公主”.

多条证据提取的答案投票

多条证据做display illustration”平均投票”，出现次数最多的答案就作为问题的最终答案.

例子

四条证据提取的答案:{白雪公主，老婆，白雪公主，老娘我 }
↓ 平均投票
白雪公主

图示

经过模型标注每条证据后display illustration，我们从每条证据中提取备选答案，然后display illustration投票

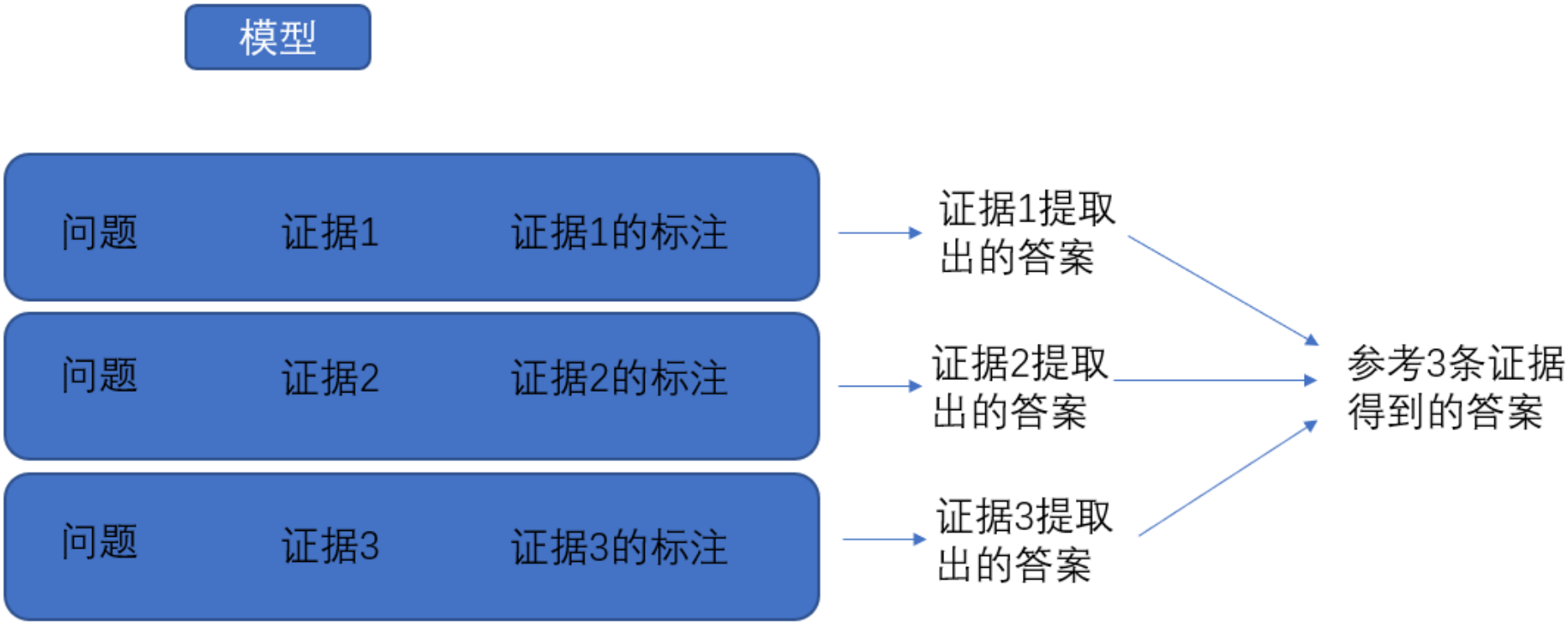

Figure: 画图说模型输出答案

目录

一般的超参数

超参数指的是模型训练前确定好的参数.

1. 优化器类型: Neural Recurrent Sequence Labeling Model (神经循环序列标注模型)是有监督学习, 损失函数体现了训练出的模型参数下真实结果出现的概率的负 log 值. 因为损失函数是超越的, 所以我们用数值方法来改进模型参数, 使得模型的预测更加贴近真实结果
2. μ : 优化器的学习率, 或者优化器的初始学习率.
3. epoch 次数: 模型训练时遍历全部样本的次数. 遍历少了, 模型训练不充分, 遍历多了, 模型过拟合.
4. batch 大小: 每次优化的 batch 的大小. 模型每步训练都会拿一个 batch 的数据作为输入. 主要出于两方面考虑, 我们选择 batch 大小:
    1. 计算机算力
    2. 数据的方差

神经循环序列标注模型中超参

神经循环序列标注模型中的超参数包括:

- D：词向量维度
- D1：J1 嵌入的维度
- D2：J2 嵌入的维度
- L：’开头-内部-外部’标签的数量

超参数选取

超参选择

选取超参

In [2]:

```
###################### 选取超参 ######################
############################################## start
D = 64
D_J1 = 2
D_J2 = 2
D1 = 2
D2 = 2
L = 3
init_scale = 0.1
learning_rate = 0.1
max_grad_norm = 5
num_layers = 1
max_epoches = 2
keep_prob = 0.95
batch_size = 5
########################################## end
```

目录

评价标准

根据上一章的方法, 得到每个问题的最终答案. 这里我们没有使用 precision,recall 和 F1 score. 用精确匹配的准确率作为评价指标.
总体来看, 此次实验的模型在训练集 (不到 80%) 和验证集 (不到 60%) 上准确率不高. 按照重要性排序, 原因可能是:

- 超参选取不好
- 模型设计不足
- 单条证据质量不高
- 词嵌入模型效率待定

超参没有充分选取

一般我们会对超参进行多次选择, 甚至 grid search. 但是此次实验中没有充分选取超参.

模型设计不足

1. ’开头-内部-外部’标签不准确:’开头-内部-外部’标签是，经过 jieba 分词或否，机器标注的. 在一条证据中答案出现处都标注了"B""I" 而没有考虑上下文.
2. 对问题下的全部证据一视同仁，没有计算 confidence scores 来区分证据的质量. 这点不如分类/排序问题方法中的每条证据的可信度.

单条证据质量不高

将每条证据对应的机器标注的’开头-内部-外部’标签序列转化为答案后，发现空答案在在训练集总体和验证集总体上的比率都在在百分之三十左右. 机器标注的’开头-内部-外部’标签在训练集上的准确率

In [9]:
```
f= open('golden_answers_valid.txt', 'a',encoding='utf-8')
num_empty = 0
for index_batch in range(num_batches):
    index_in_batch = 0
    for  y_ in  batches_Y[index_batch ]:
        true_seq_len = (y_ >-1).sum()
        selected = [E_list[index_batch * batch_size+index_in_batch][i] for i in range(true_seq_len) if( y_[i]> 0 ) ]
        if( len(selected) < 1):
            num_empty +=1
        index_id = ID_list[ index_batch * batch_size+index_in_batch]
        f.write("{}\t{}\n".format(index_id, ''.join(selected)))
        index_in_batch += 1

f.close()
ratio = float(num_empty)/ (num_batches*batch_size)
print('机器从证据中标注的bio标签所对应的%d条答案中,空答案的比率为 :  %.2f%%'%(num_batches*batch_size ,float(100*ratio)) )
```

机器从证据中标注的bio标签所对应的29900条答案中,空答案的比率为 : 68.89%

Figure: 机器标注的’开头-内部-外部’标签在验证集上的准确率

词嵌入模型不准确

此次实验中，嵌入模不是提前准备的型，而是在训练过程中训练出的. 因为 corpus 比较小，所以词嵌入模型不能很好得反应词之间的相似度.

目录

可能的改进方向

- 尝试不同的超参
- 引入更复杂的’开头-内部-外部’标签标注方法
- 计算问题下不同的证据的可信度
- 引入更加快可靠的词嵌入模型

结束语

谢谢大家的时间!

谢谢皓天在实验中的帮助和建议
谢谢鹏飞和欣雨配合预演讲，帮助改进 PPT