# 命名实体识别

## 一、数据处理

（1）数据源

数据来源主要包含两部分：

* 恒企问答数据中的问题数据
* 2019年财务与会计官方教材电子版（书籍）

（2）数据处理

* 基于人名日报语料和crf方法，训练命名实体识别模型，实现人名，地名，时间，机构。
* 基于已经训练好的命名实体识别模型和会计数据词典，对恒企问答数据进行标记，标签包括会计，人名，地名，时间，机构。
* 基于已经训练好的命名实体识别模型和会计数据词典，对财务与会计官方教材进行标记，标签包括会计，人名，地名，时间，机构。

其标记部分结果如图1所示。

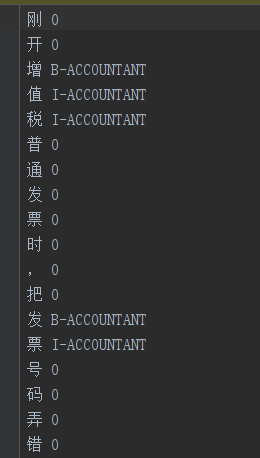


图1 部分标记数据

其中，人名，地名，时间，机构标记不变，添加会计词汇的标记，ACCOUNTANT，然后将数据分为训练数据、测试数据和验证数据。

## 二、模型训练

### 2.1基于crf的模型训练

（1）训练数据

训练数据分为两部分：

* 一个是标注后的恒企问答数据。
* 一个是标注后的财务与会计书籍数据。

（2）模型训练

* 训练参数

# Unigram  
U00:%x[-2,0]  
U01:%x[-1,0]  
U02:%x[0,0]  
U03:%x[1,0]  
U04:%x[2,0]  
U05:%x[-2,0]/%x[-1,0]  
U06:%x[-1,0]/%x[0,0]  
U07:%x[0,0]/%x[1,0]  
U08:%x[1,0]/%x[2,0]

（3）训练结果

* 恒企问答数据训练结果

其训练模型结果如图2所示。

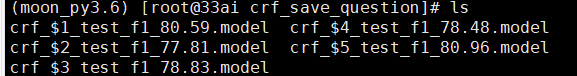


图2 问答数据训练模型

由图2可知，问答数据训练模型中，F值最高的是80.96%。用该模型，对恒企后端问答中的问题进行命名实体识别。其结果如图3所示。

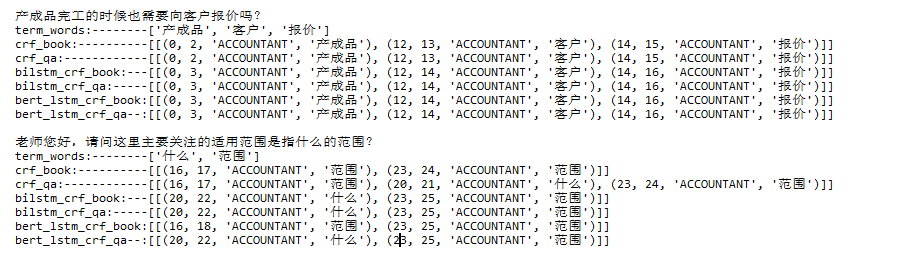


图3 命名实体识别部分结果

其中crf\_qa是命名实体识别的结果。

* 财务与会计书籍数据训练结果

其训练结果如图4所示。

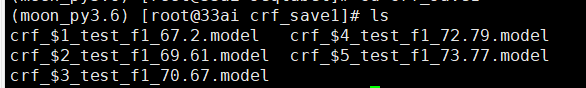


图4 书籍数据集训练模型

由图4可知，书籍数据训练模型中，F值最高的是73.77%。用该模型，对恒企后端问答中的问题进行命名实体识别。其结果如图5所示。

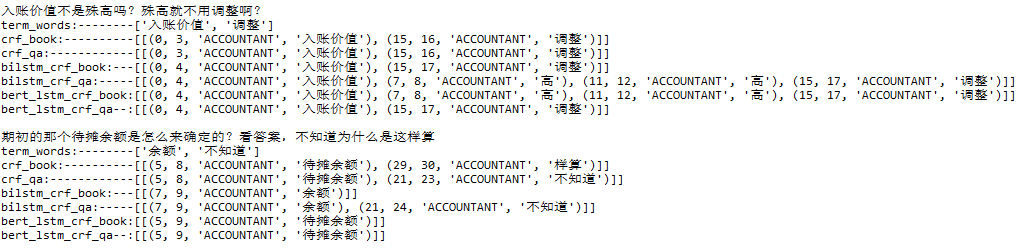


图5 命名实体识别部分结果

### 2.2基于lstm-crf的模型训练

（1）训练数据

训练数据分为两部分：

* 一个是标注后的恒企问答数据。
* 一个是标注后的财务与会计书籍数据。

（2）模型训练

* 训练参数

flags.DEFINE\_boolean("clean", True, "clean train folder")  
flags.DEFINE\_boolean("train", con\_key['trian\_type'], "Wither train the model---False")  
# configurations for the model  
flags.DEFINE\_boolean("use\_start\_end\_crf", True, "whether use start and end status in crf loss")  
flags.DEFINE\_integer("seg\_dim", 20, "Embedding size for segmentation, 0 if not used")  
flags.DEFINE\_integer("char\_dim", 100, "Embedding size for characters")  
flags.DEFINE\_integer("lstm\_dim", 100, "Num of hidden units in LSTM")  
flags.DEFINE\_string("tag\_schema", "iobes", "tagging schema iobes or iob")  
  
# configurations for training  
flags.DEFINE\_float("clip", 5, "Gradient clip")  
flags.DEFINE\_float("dropout", 0.5, "Dropout rate")  
flags.DEFINE\_integer("batch\_size", 32, "batch size")  
flags.DEFINE\_float("lr", 0.001, "Initial learning rate")  
flags.DEFINE\_string("optimizer", "adam", "Optimizer for training")  
flags.DEFINE\_boolean("pre\_emb", True, "Wither use pre-trained embedding")  
flags.DEFINE\_boolean("zeros", False, "Wither replace digits with zero")  
flags.DEFINE\_boolean("lower", True, "Wither lower case")  
  
flags.DEFINE\_integer("max\_epoch", 100, "maximum training epochs")  
flags.DEFINE\_integer("steps\_check", 100, "steps per checkpoint")  
flags.DEFINE\_string("ckpt\_path", con\_key['ckpt\_path'], "Path to save model-ckpt\_book")  
flags.DEFINE\_string("summary\_path", con\_key['summary\_path'], "Path to store summaries-summary\_book")  
flags.DEFINE\_string("log\_file", con\_key['log\_file'], "File for log-train\_book.log")  
flags.DEFINE\_string("map\_file", con\_key['map\_file'], "file for maps-maps\_book.pkl")  
flags.DEFINE\_string("vocab\_file", "vocab.json", "File for vocab")  
flags.DEFINE\_string("config\_file", "config\_file", "File for config")  
flags.DEFINE\_string("script", "conlleval", "evaluation script")  
flags.DEFINE\_string("result\_path", con\_key['result\_path'], "Path for results-result\_book")  
flags.DEFINE\_string("emb\_file", "wiki\_100.utf8", "Path for pre\_trained embedding")  
flags.DEFINE\_string("train\_file", os.path.join("data", con\_key['train\_file']), "Path for train data-train\_book.data")  
flags.DEFINE\_string("dev\_file", os.path.join("data", con\_key['dev\_file']), "Path for dev data-dev\_book.data")  
flags.DEFINE\_string("test\_file", os.path.join("data", con\_key['test\_file']), "Path for test data-test\_book.data")

（3）训练结果

* 恒企问答数据训练结果

其训练模型结果如图6和图7所示。



图6 问答数据训练模型

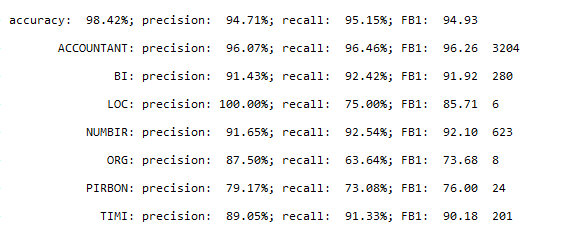


图7 训练评估结果

由图6和图7可知，问答数据训练模型中，F值是94.93%。其中可以看出，数据的不平衡性，其中会计词和数据词，占最多。用该模型，对恒企后端问答中的问题进行命名实体识别。其结果如图8所示。

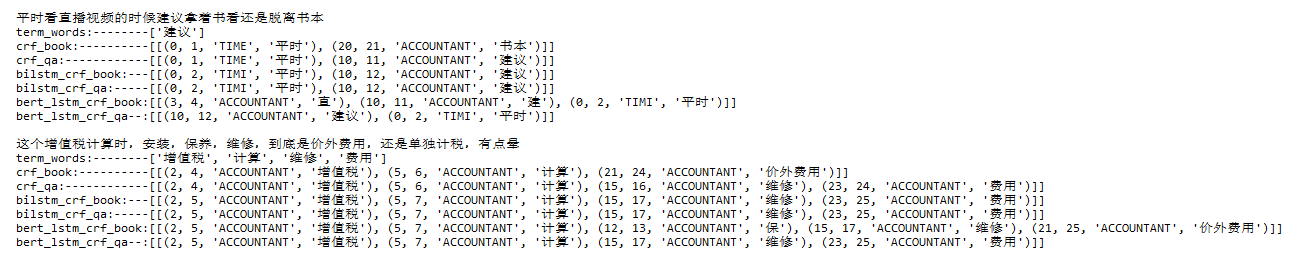


图8 命名实体识别部分结果

其中bilstm\_crf\_qa是命名实体识别的结果。

* 财务与会计书籍数据训练结果

其训练结果如图9和图10所示。



图9 书籍数据集训练模型

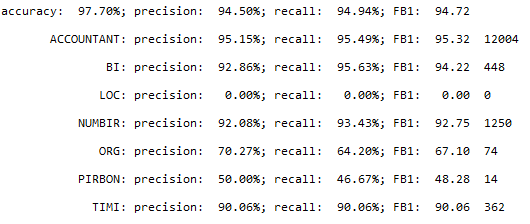


图10 训练评估结果

由图9和图10可知，问答数据训练模型中，F值是94.72%。同时，数据也存在了不平衡性。其中，会计领域的词和数字得到了识别。用该模型，对恒企后端问答中的问题进行命名实体识别。其结果如图11所示。

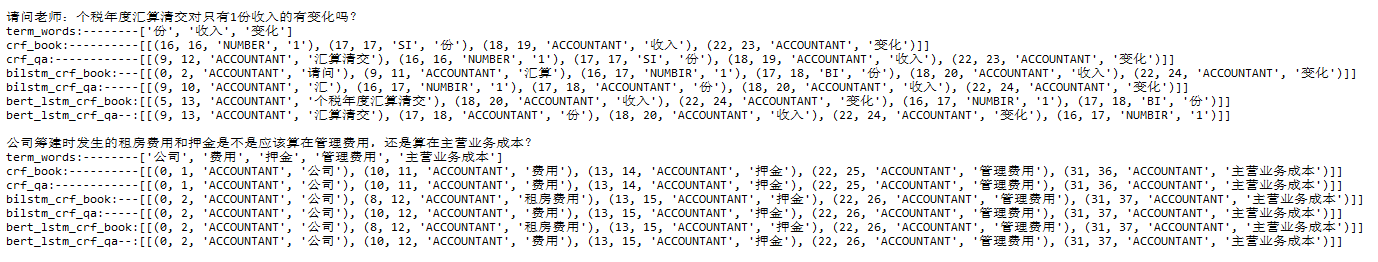


图11 命名实体识别部分结果

其中bilstm\_crf\_book是命名实体识别的结果。

### 2.3基于bert-lstm-crf的模型训练

（1）训练数据

训练数据分为两部分：

* 一个是标注后的恒企问答数据。
* 一个是标注后的财务与会计书籍数据。

（2）模型训练

* 训练参数

if os.name == 'nt':  
 bert\_path = r'E:\cbackup\Desktop\code\github\nlp\bert-kbqa\NER\_BERT\_BiLSTM\_CRF\bert\chinese\_L-12\_H-768\_A-12'  
 root\_path = r'E:\cbackup\Desktop\code\github\nlp\bert-kbqa\NER\_BERT\_BiLSTM\_CRF'  
else:  
 bert\_path = '/data/menghao/yun2space/NER\_BERT\_BiLSTM\_CRF/bert/bert-wwm-ext/'  
 root\_path = '/data/menghao/yun2space/NER\_BERT\_BiLSTM\_CRF/'  
print('a')  
flags.DEFINE\_string(  
 "data\_dir", os.path.join(root\_path, 'NERdata'),  
 "The input datadir.",  
)  
  
flags.DEFINE\_string(  
 "bert\_config\_file", os.path.join(bert\_path, 'bert\_config.json'),  
 "The config json file corresponding to the pre-trained BERT model."  
)  
  
flags.DEFINE\_string(  
 "task\_name", 'ner', "The name of the task to train."  
)  
  
flags.DEFINE\_string(  
 "output\_dir", os.path.join(root\_path, 'output\_book'),  
 "The output directory where the model checkpoints will be written."  
)  
  
## Other parameters  
flags.DEFINE\_string(  
 "init\_checkpoint", os.path.join(bert\_path, 'bert\_model.ckpt'),  
 "Initial checkpoint (usually from a pre-trained BERT model)."  
)  
  
flags.DEFINE\_bool(  
 "do\_lower\_case", True,  
 "Whether to lower case the input text."  
)  
  
flags.DEFINE\_integer(  
 "max\_seq\_length", 128,  
 "The maximum total input sequence length after WordPiece tokenization."  
)  
  
flags.DEFINE\_boolean('clean', True, 'remove the files which created by last training')  
  
flags.DEFINE\_bool("do\_train", True, "Whether to run training."  
)  
flags.DEFINE\_bool("use\_tpu", False, "Whether to use TPU or GPU/CPU.")  
  
flags.DEFINE\_bool("do\_eval", True, "Whether to run eval on the dev set.")  
  
flags.DEFINE\_bool("do\_predict", True, "Whether to run the model in inference mode on the test set.")  
  
flags.DEFINE\_integer("train\_batch\_size", 32, "Total batch size for training.")  
  
flags.DEFINE\_integer("eval\_batch\_size", 8, "Total batch size for eval.")  
  
flags.DEFINE\_integer("predict\_batch\_size", 8, "Total batch size for predict.")  
  
flags.DEFINE\_float("learning\_rate", 5e-5, "The initial learning rate for Adam.")  
  
flags.DEFINE\_float("num\_train\_epochs", 15.0, "Total number of training epochs to perform.")  
flags.DEFINE\_float('droupout\_rate', 0.5, 'Dropout rate')  
flags.DEFINE\_float('clip', 5, 'Gradient clip')  
flags.DEFINE\_float(  
 "warmup\_proportion", 0.1,  
 "Proportion of training to perform linear learning rate warmup for. "  
 "E.g., 0.1 = 10% of training.")  
  
flags.DEFINE\_integer("save\_checkpoints\_steps", 1000,  
 "How often to save the model checkpoint.")  
  
flags.DEFINE\_integer("iterations\_per\_loop", 1000,  
 "How many steps to make in each estimator call.")  
  
flags.DEFINE\_string("vocab\_file", os.path.join(bert\_path, 'vocab.txt'),  
 "The vocabulary file that the BERT model was trained on.")  
  
tf.flags.DEFINE\_string("master", None, "[Optional] TensorFlow master URL.")  
flags.DEFINE\_integer(  
 "num\_tpu\_cores", 8,  
 "Only used if `use\_tpu` is True. Total number of TPU cores to use.")  
flags.DEFINE\_string('data\_config\_path', os.path.join(root\_path, 'data.conf'),  
 'data config file, which save train and dev config')  
# lstm parame  
flags.DEFINE\_integer('lstm\_size', 128, 'size of lstm units')  
flags.DEFINE\_integer('num\_layers', 1, 'number of rnn layers, default is 1')  
flags.DEFINE\_string('cell', 'lstm', 'which rnn cell used')

（3）训练结果

* 恒企问答数据训练结果

其训练模型结果如图12和图13所示。

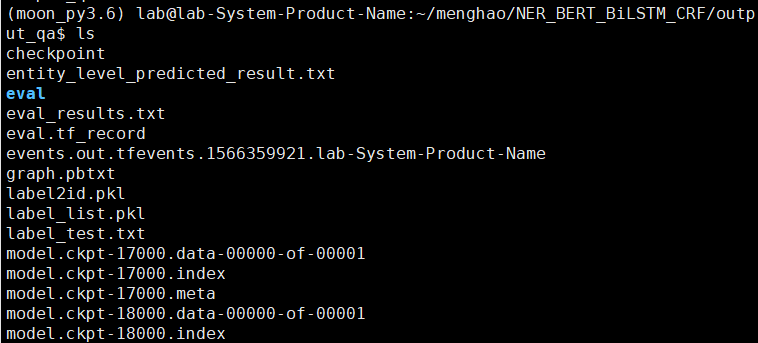


图12 问答数据训练模型

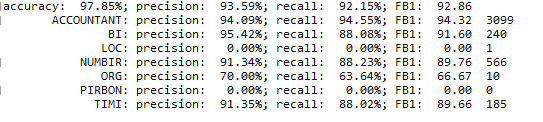


图13 训练评估结果

由图12和图13可知，问答数据训练模型中，F值是92.86%。其中可以看出，数据的不平衡性，其中会计词和数据词，占最多。用该模型，对恒企后端问答中的问题进行命名实体识别。其结果如图14所示。

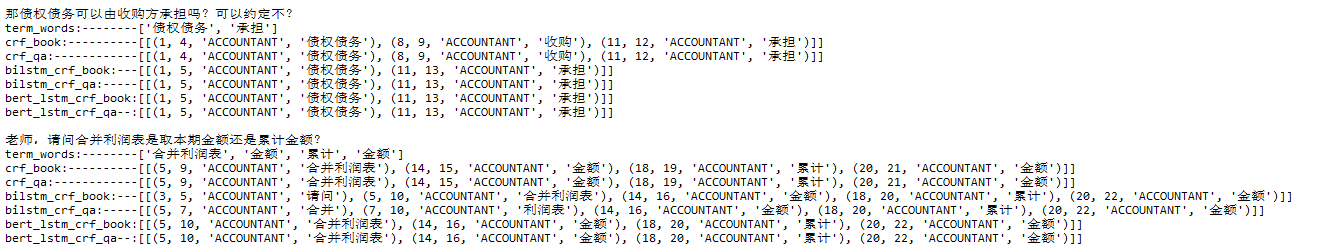


图14 命名实体识别部分结果

其中bert\_lstm\_crf\_qa是命名实体识别的结果。

* 财务与会计书籍数据训练结果

其训练结果如图15和图16所示。

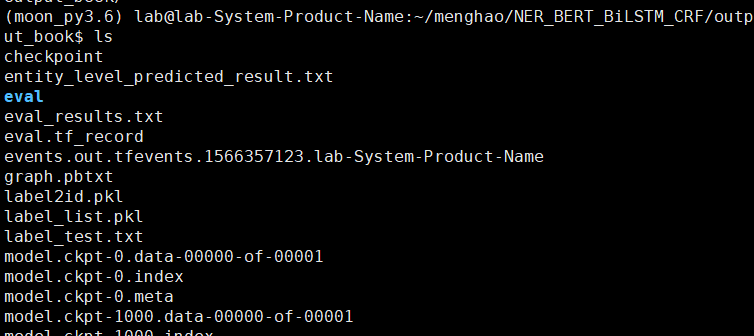


图15 书籍数据集训练模型

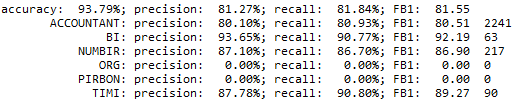


图16 训练评估结果

由图15和图16可知，问答数据训练模型中，F值是81.85%。同时，数据也存在了不平衡性。其中，会计领域的词和数字得到了识别。用该模型，对恒企后端问答中的问题进行命名实体识别。其结果如图5所示。

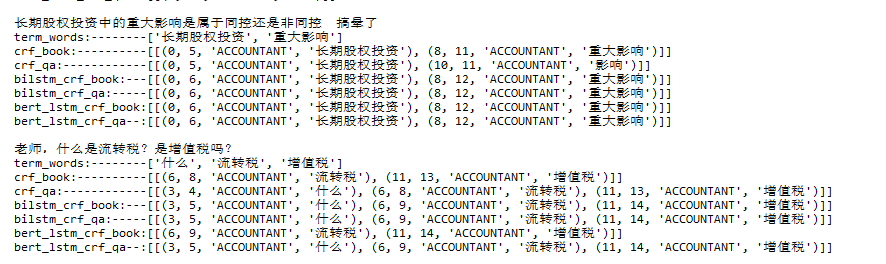


图5 命名实体识别部分结果

其中bert\_lstm\_crf\_book是命名实体识别的结果。

## 三、测试分析

（1）测试数据

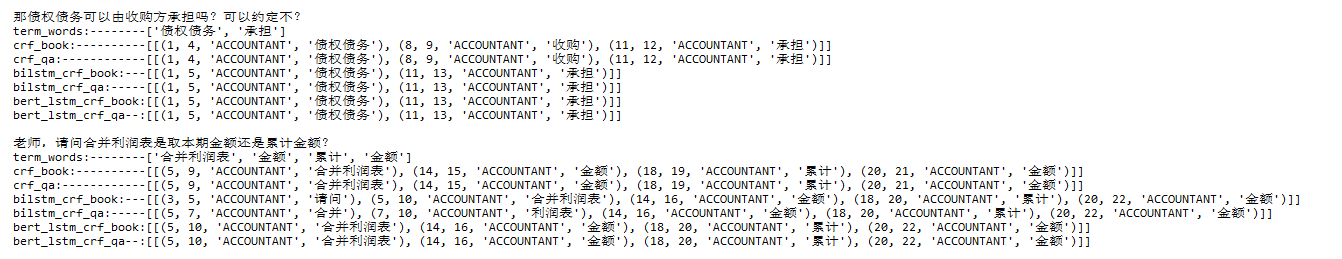
从恒企数据后台，最新的1000条问答数据。

（2）测试内容

基于已经训练好的模型，对1000条问答数据中的问题进行命名实体识别。

（2）结果分析

其测试的部分结果如下图所示。



测试部分结果图

* 数据问题

老师您好，请问这里主要关注的适用范围是指什么的范围？

term\_words:--------['什么', '范围']

crf\_book:----------[[(16, 17, 'ACCOUNTANT', '范围'), (23, 24, 'ACCOUNTANT', '范围')]]

crf\_qa:------------[[(16, 17, 'ACCOUNTANT', '范围'), (20, 21, 'ACCOUNTANT', '什么'), (23, 24, 'ACCOUNTANT', '范围')]]

bilstm\_crf\_book:---[[(20, 22, 'ACCOUNTANT', '什么'), (23, 25, 'ACCOUNTANT', '范围')]]

bilstm\_crf\_qa:-----[[(20, 22, 'ACCOUNTANT', '什么'), (23, 25, 'ACCOUNTANT', '范围')]]

bert\_lstm\_crf\_book:[[(16, 18, 'ACCOUNTANT', '范围'), (23, 25, 'ACCOUNTANT', '范围')]]

bert\_lstm\_crf\_qa--:[[(20, 22, 'ACCOUNTANT', '什么'), (23, 25, 'ACCOUNTANT', '范围')]]

在所有的训练模型中，课本的数据中，很少有什么，而问答数据集较多，因此，基于问答的数据集能识别出来，课本的数据集无法识别。

期初的那个待摊余额是怎么来确定的？看答案，不知道为什么是这样算

term\_words:--------['余额', '不知道']

crf\_book:----------[[(5, 8, 'ACCOUNTANT', '待摊余额'), (29, 30, 'ACCOUNTANT', '样算')]]

crf\_qa:------------[[(5, 8, 'ACCOUNTANT', '待摊余额'), (21, 23, 'ACCOUNTANT', '不知道')]]

上述例子中，样算在书籍数据集中描述的较多，因此能够识别出来。

老师小规模纳税人代开的专票，是必须交税吗？

term\_words:--------['小规模纳税人']

crf\_book:----------[[(2, 7, 'ACCOUNTANT', '小规模纳税人'), (8, 10, 'ACCOUNTANT', '代开的')]]

crf\_qa:------------[[(2, 7, 'ACCOUNTANT', '小规模纳税人')]]

bilstm\_crf\_book:---[[(2, 8, 'ACCOUNTANT', '小规模纳税人'), (8, 10, 'ACCOUNTANT', '代开'), (11, 13, 'ACCOUNTANT', '专票')]]

bilstm\_crf\_qa:-----[[(2, 8, 'ACCOUNTANT', '小规模纳税人')]]

bert\_lstm\_crf\_book:[[(2, 8, 'ACCOUNTANT', '小规模纳税人'), (11, 13, 'ACCOUNTANT', '专票')]]

bert\_lstm\_crf\_qa--:[[(2, 8, 'ACCOUNTANT', '小规模纳税人')]]

* 模型优势

请问老师：个税年度汇算清交对只有1份收入的有变化吗？

term\_words:--------['份', '收入', '变化']

crf\_book:----------[[(16, 16, 'NUMBER', '1'), (17, 17, 'SI', '份'), (18, 19, 'ACCOUNTANT', '收入'), (22, 23, 'ACCOUNTANT', '变化')]]

crf\_qa:------------[[(9, 12, 'ACCOUNTANT', '汇算清交'), (16, 16, 'NUMBER', '1'), (17, 17, 'SI', '份'), (18, 19, 'ACCOUNTANT', '收入'), (22, 23, 'ACCOUNTANT', '变化')]]

bilstm\_crf\_book:---[[(0, 2, 'ACCOUNTANT', '请问'), (9, 11, 'ACCOUNTANT', '汇算'), (16, 17, 'NUMBIR', '1'), (17, 18, 'BI', '份'), (18, 20, 'ACCOUNTANT', '收入'), (22, 24, 'ACCOUNTANT', '变化')]]

bilstm\_crf\_qa:-----[[(9, 10, 'ACCOUNTANT', '汇'), (16, 17, 'NUMBIR', '1'), (17, 18, 'ACCOUNTANT', '份'), (18, 20, 'ACCOUNTANT', '收入'), (22, 24, 'ACCOUNTANT', '变化')]]

bert\_lstm\_crf\_book:[[(5, 13, 'ACCOUNTANT', '个税年度汇算清交'), (18, 20, 'ACCOUNTANT', '收入'), (22, 24, 'ACCOUNTANT', '变化'), (16, 17, 'NUMBIR', '1'), (17, 18, 'BI', '份')]]

bert\_lstm\_crf\_qa--:[[(9, 13, 'ACCOUNTANT', '汇算清交'), (17, 18, 'ACCOUNTANT', '份'), (18, 20, 'ACCOUNTANT', '收入'), (22, 24, 'ACCOUNTANT', '变化'), (16, 17, 'NUMBIR', '1')]]

数据不同，模型不同，能够识别出汇算清交等术语词，在于模型的区别。

公司筹建时发生的租房费用和押金是不是应该算在管理费用，还是算在主营业务成本？

term\_words:--------['公司', '费用', '押金', '管理费用', '主营业务成本']

crf\_book:----------[[(0, 1, 'ACCOUNTANT', '公司'), (10, 11, 'ACCOUNTANT', '费用'), (13, 14, 'ACCOUNTANT', '押金'), (22, 25, 'ACCOUNTANT', '管理费用'), (31, 36, 'ACCOUNTANT', '主营业务成本')]]

crf\_qa:------------[[(0, 1, 'ACCOUNTANT', '公司'), (10, 11, 'ACCOUNTANT', '费用'), (13, 14, 'ACCOUNTANT', '押金'), (22, 25, 'ACCOUNTANT', '管理费用'), (31, 36, 'ACCOUNTANT', '主营业务成本')]]

bilstm\_crf\_book:---[[(0, 2, 'ACCOUNTANT', '公司'), (8, 12, 'ACCOUNTANT', '租房费用'), (13, 15, 'ACCOUNTANT', '押金'), (22, 26, 'ACCOUNTANT', '管理费用'), (31, 37, 'ACCOUNTANT', '主营业务成本')]]

bilstm\_crf\_qa:-----[[(0, 2, 'ACCOUNTANT', '公司'), (10, 12, 'ACCOUNTANT', '费用'), (13, 15, 'ACCOUNTANT', '押金'), (22, 26, 'ACCOUNTANT', '管理费用'), (31, 37, 'ACCOUNTANT', '主营业务成本')]]

bert\_lstm\_crf\_book:[[(0, 2, 'ACCOUNTANT', '公司'), (8, 12, 'ACCOUNTANT', '租房费用'), (13, 15, 'ACCOUNTANT', '押金'), (22, 26, 'ACCOUNTANT', '管理费用'), (31, 37, 'ACCOUNTANT', '主营业务成本')]]

bert\_lstm\_crf\_qa--:[[(0, 2, 'ACCOUNTANT', '公司'), (10, 12, 'ACCOUNTANT', '费用'), (13, 15, 'ACCOUNTANT', '押金'), (22, 26, 'ACCOUNTANT', '管理费用'), (31, 37, 'ACCOUNTANT', '主营业务成本')]]

标注中没有标注出租房费用，而在数据为课本，同时非单一crf的方法下，识别出了租房费用。

收到税务局退回来的当时多交的附加税怎么做账

term\_words:--------['税务局', '附加税', '做账']

crf\_book:----------[[(2, 3, 'ACCOUNTANT', '税务'), (5, 6, 'ACCOUNTANT', '退回'), (9, 10, 'TIME', '当时')]]

crf\_qa:------------[[(2, 4, 'ACCOUNTANT', '税务局'), (9, 10, 'TIME', '当时'), (14, 16, 'ACCOUNTANT', '附加税'), (19, 20, 'ACCOUNTANT', '做账')]]

bilstm\_crf\_book:---[[(2, 5, 'ACCOUNTANT', '税务局'), (9, 11, 'TIMI', '当时')]]

bilstm\_crf\_qa:-----[[(2, 5, 'ACCOUNTANT', '税务局'), (9, 11, 'TIMI', '当时'), (14, 17, 'ACCOUNTANT', '附加税'), (19, 21, 'ACCOUNTANT', '做账')]]

bert\_lstm\_crf\_book:[[(2, 5, 'ACCOUNTANT', '税务局'), (5, 6, 'ACCOUNTANT', '退'), (14, 17, 'ACCOUNTANT', '附加税'), (9, 11, 'TIMI', '当时')]]

bert\_lstm\_crf\_qa--:[[(2, 5, 'ACCOUNTANT', '税务局'), (14, 17, 'ACCOUNTANT', '附加税'), (19, 21, 'ACCOUNTANT', '做账'), (9, 11, 'TIMI', '当时')]]

基本的crf，使用book数据，税务局为识别出来。