Question Answering

On CJRC Dataset
(Chinese judicial Reading Comprehension)

Liancheng Gong
Xinyao Han
Qiaowei Li

Contents

01 **Problem Situation** · significance · problem description · expected difficulties

02 **Descriptive Analytics** · dataset description · data cleaning · visualization

03 **Problem Solving** · Pre-trained model · Classification model · QA model

04 **Error Analysis** · Classification attempts · Out of memory · Tokenizer

05 **Future Work** · Slide window · Multi-task model · Tokenizer · Answer YES/NO

Contents

Final presentation The final presentation is an opportunity for you to showcase your work. You will probably want to follow a structure similar to:

1. Problem situation: what is the problem? Reflect on why it's important. What are the difficulties?

2. Descriptive analytics: Which data did you get? Any interesting insights/visualizations?

3. Machine Learning: How did you tackle the problem?

4. Future work: If you had infinite time, what would you envision the next steps to take are? How will you improve the results?

1

Problem Situation

❑ Significance

❑ Problem Description

❑ Expected Difficulties

Significance

Tons of cases form legal judgements (composed of summary of cases, descriptions of events, count opinions, judgement results).

Different judgement results on same case.

A time consuming process to extract relative information.s

MRC Solution

❑ Span-Extraction MRC on Chinese legal judgment documents.

Expected Difficulties

❑ Chinese corpus datasets

❑ Added ‘YES/NO’ answer to the SQuAD dataset

2

Descriptive Analytics

- ❑ dataset description
- ❑ data cleaning
- ❑ visualization

CJRC Dataset
(Chinese Judgement Reading Comprehension)

Public legal judgment documents from the “China Judgment Documents Network”

❑ Civil and criminal judgments

❑ Descriptions of 10,000 judgments issued by the Supreme People's Court of China

❑ Legal experts marked 4 to 5 Question Answer pairs

❑ Big_train_data.json, Dev_ground_truth.json, test_ground_truth.json

The datasets imitate format with SQuAD 2.0.

Case of Auction	e139eef6-fc0c-4953-acec-a83a0095ce4e.txt
Case Name	保险人代位求偿权纠纷
Case Description	经审查,原告提供的证据1-3、被告中华联合广东分公司提供的证据4-5、被告万友公司提供的证据6,各方对其真实性均没有异议,本院对其真实性予以确认综合本院采信的证据及当事人的陈述,本院认定以下事实:2015年6月1日,田x17驾驶粤A×××××号车辆与严x3驾驶的赣C×××××号重型仓栅式货车发生碰撞,造成两车不同程度损坏的交通事故交警部门作出事故认定书,认定严x3承担事故的全部责任,田x17不负事故责任粤A×××××号车辆在原告处投保了保险金额为908000元的机动车损失保险,事故发生在保险期间内事故发生后,粤A×××××号车辆的被保险人陈x18就该车辆的损失以财产保险合同纠纷起诉至佛山市禅城区人民法院案经审理,佛山市禅城区人民法院于2015年8月18日作出(2015)佛城法民二初字第1006号民事判决,查明粤A×××××号车辆经广州市华盟价格事务所有限公司评估,损失价格为241541元,陈x18支付了粤A×××××号车辆的维修费241541元、评估费9050元;本案原告在庭审中明确表示不申请重新对车辆损失进行评估鉴定并判决原告向陈x18支付粤A×××××号车辆损失保险理赔款250591元2015年10月11日,原告向陈x18赔付了250591元及诉讼费用2529元后原告提起本案之诉并查明,赣C×××××号车辆的所有人为被告万友公司,该车辆在被告中华联合广东分公司处投保了交强险,事故发生在保险期内事故发生后,被告中华联合广东分公司向该车辆的被保险人许x19赔付了2000元诉讼中,被告徐11确认其为该车辆的实际支配人,严x3是被告徐11雇请,是从事派遣工作过程中发生案涉交通事故被告徐11与被告万友公司签订《车辆挂靠合同书》,被告万友公司同意被告徐11就赣C×××××号车辆挂靠被告万友公司名下

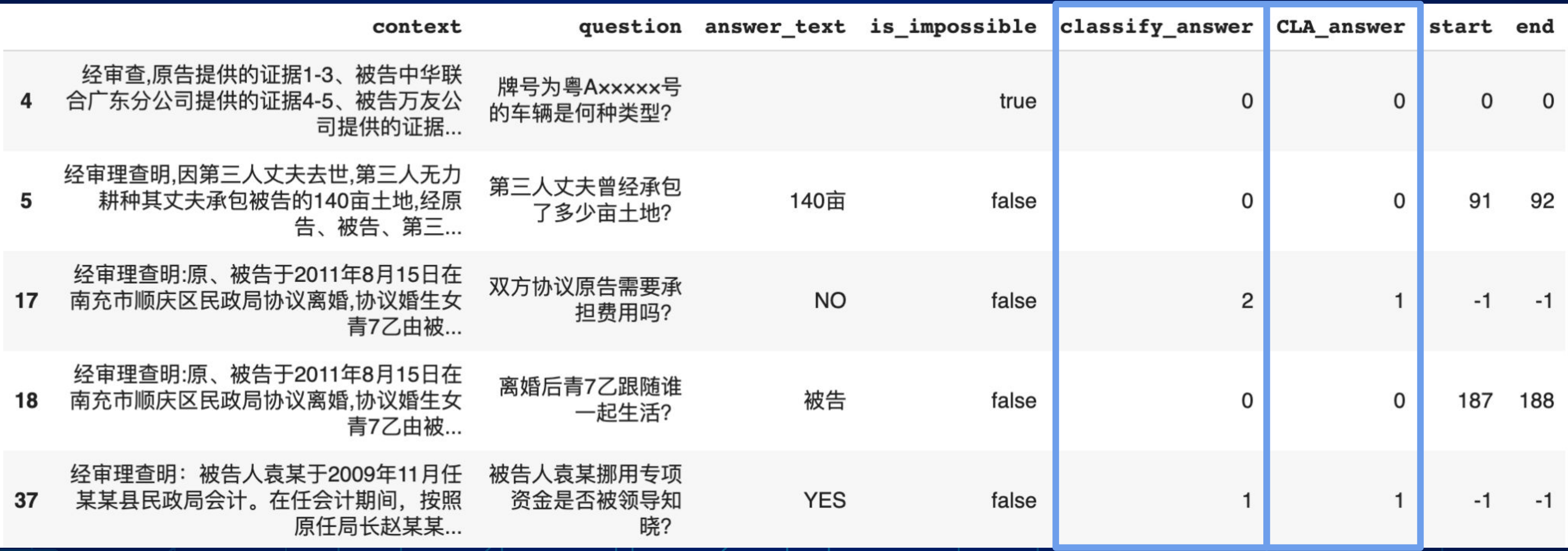
Q&A Pairs

```
{'answer_start': 153, 'text': '两车不同程度损坏'}],
    'id': 'e139eef6-fc0c-4953-acec-a83a0095ce4e.txt_001',
    'is_impossible': 'false',
    'question': '事故结果如何？'}

{'answer_start': 180, 'text': '严x3承担事故的全部责任,田x17不负事故责任'}],
    'id': 'e139eef6-fc0c-4953-acec-a83a0095ce4e.txt_002',
    'is_impossible': 'false',
    'question': '事故由谁承担什么责任？'}

{'answer_start': 233, 'text': '机动车损失保险'}],
    'id': 'e139eef6-fc0c-4953-acec-a83a0095ce4e.txt_003',
    'is_impossible': 'false',
    'question': '投保人所投保险险种？'}

{'answer_start': 225, 'text': '908000元'}],
    'id': 'e139eef6-fc0c-4953-acec-a83a0095ce4e.txt_004',
    'is_impossible': 'false',
    'question': '向原告投保的人所投保险的保险金额是多少？'}

{'answers': [],
    'id': 'e139eef6-fc0c-4953-acec-a83a0095ce4e.txt_005',
    'is_impossible': 'true',
    'question': '牌号为粤A×××××号的车辆是何种类型？'}
```

Q&A types in our dataset

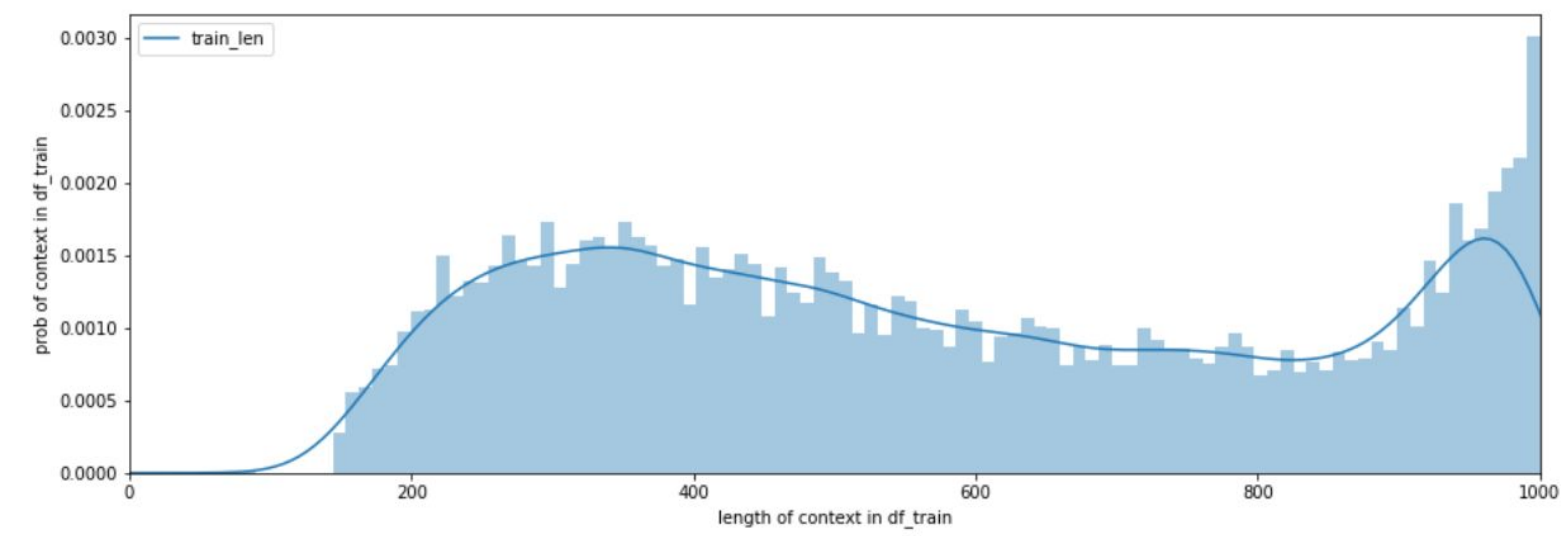
- ❑ Extract answer directly from the legal judgment documents
  - ❑ E.g. Q: 事故结果如何？ A: 两车不同程度损坏
- ❑ Yes/No answer
  - ❑ E.g. Q:双方协议原告需要承担费用吗？ A: NO
- ❑ Question unable to answer
  - ❑ E.g. Q: 牌号为粤A×××××号的车辆是何种类型？A: /

Dataset Description



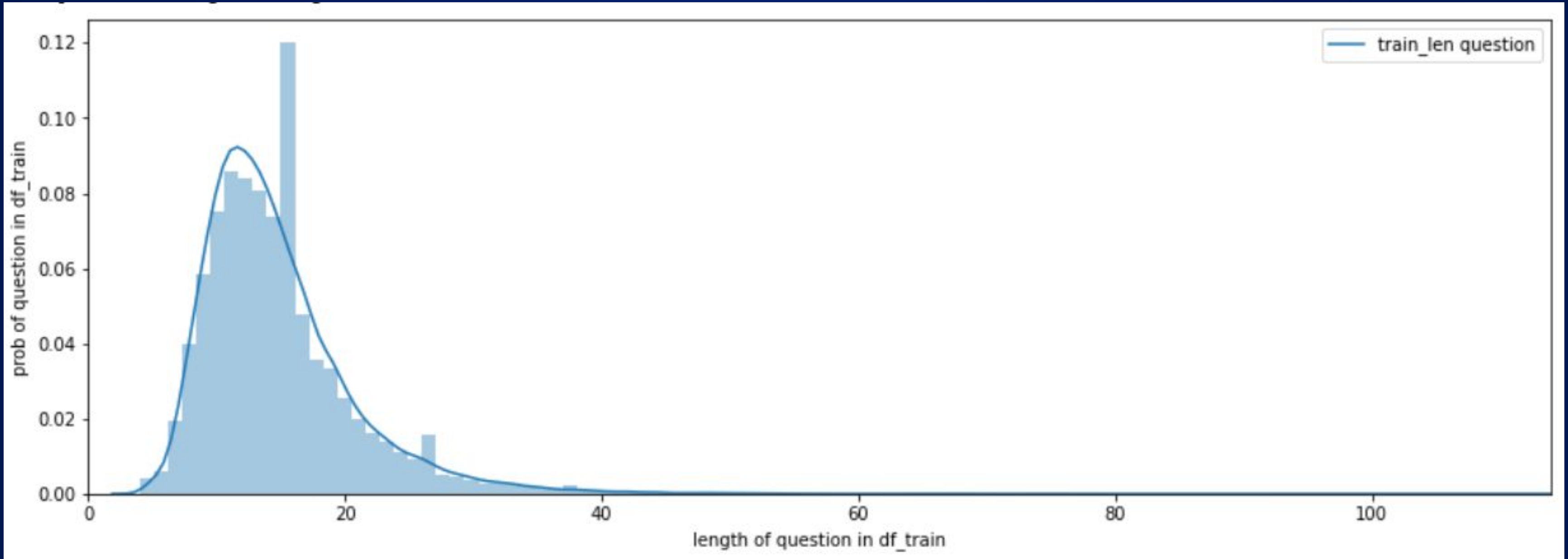
	context	question	answer_text	is_impossible	classify_answer	CLA_answer	start	end
4	经审查,原告提供的证据1-3、被告中华联合广东分公司提供的证据4-5、被告万友公司提供的证据...	牌号为粤Axxxxx号的车辆是何种类型?		true	0	0	0	0
5	经审理查明,因第三人丈夫去世,第三人无力耕种其丈夫承包被告的140亩土地,经原告、被告、第三...	第三人丈夫曾经承包了多少亩土地?	140亩	false	0	0	91	92
17	经审理查明:原、被告于2011年8月15日在南充市顺庆区民政局协议离婚,协议婚生女青7乙由被...	双方协议原告需要承担费用吗?	NO	false	2	1	-1	-1
18	经审理查明:原、被告于2011年8月15日在南充市顺庆区民政局协议离婚,协议婚生女青7乙由被...	离婚后青7乙跟随谁一起生活?	被告	false	0	0	187	188
37	经审理查明：被告人袁某于2009年11月任某某县民政局会计。在任会计期间，按照原任局长赵某某...	被告人袁某挪用专项资金是否被领导知晓?	YES	false	1	1	-1	-1

Visualization of **context** length



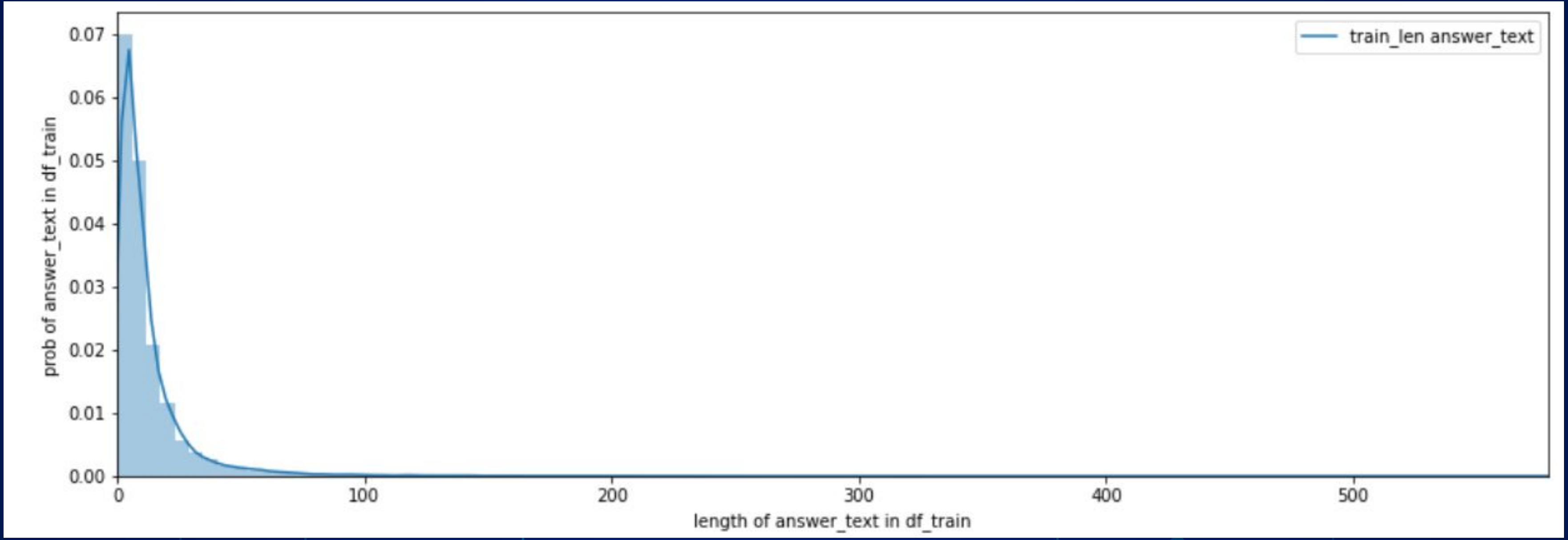
On average, **context** contains 572 word length, with minimum word length of 80, maximum word length of 1000. 75% of the data contains a context word length of 806.

Visualization of **question** length



On average, **question** contains 14 word length, with minimum word length of 4, maximum word length of 114. 75% of the data contains a context word length of 17.

Visualization of **answer** length



On average, **answer_text** contains 12 word length, with minimum word length of 0 (impossible to answer given context), maximum word length of 579. 75% of the data contains a context word length of 12.

Text truncation

- Bert only takes 512 tokens…
- We hope to have input tokens as

  [CLS] Question [SEP] Context [SEP]

- **Context:** ......Cxxxxx号重型仓栅式货车发生碰撞, 造成**两车不同程度损坏**的交通事故交警部门作出事故认证........

  ↑ Answer start

  Question: 事故结果如何？

- Setting:

  Full question, 90% percentile of answer length,

  # maximum **content** length 64+128 = 192
  # maximum **question** length 114
  # maximum **answer** length = 65

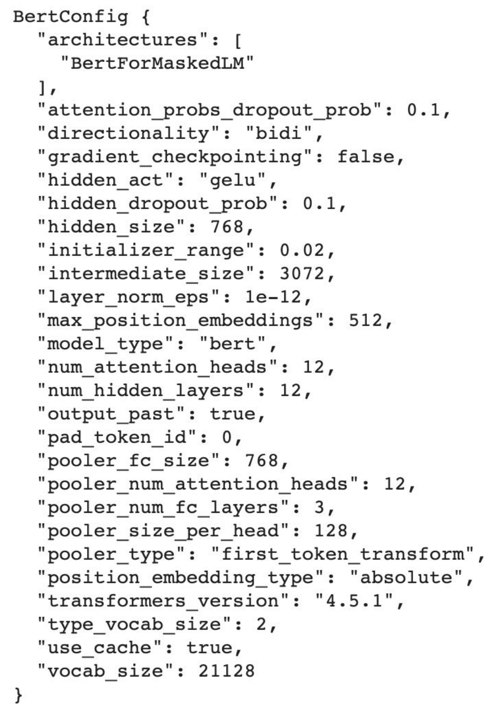
3

3 Problem Solving

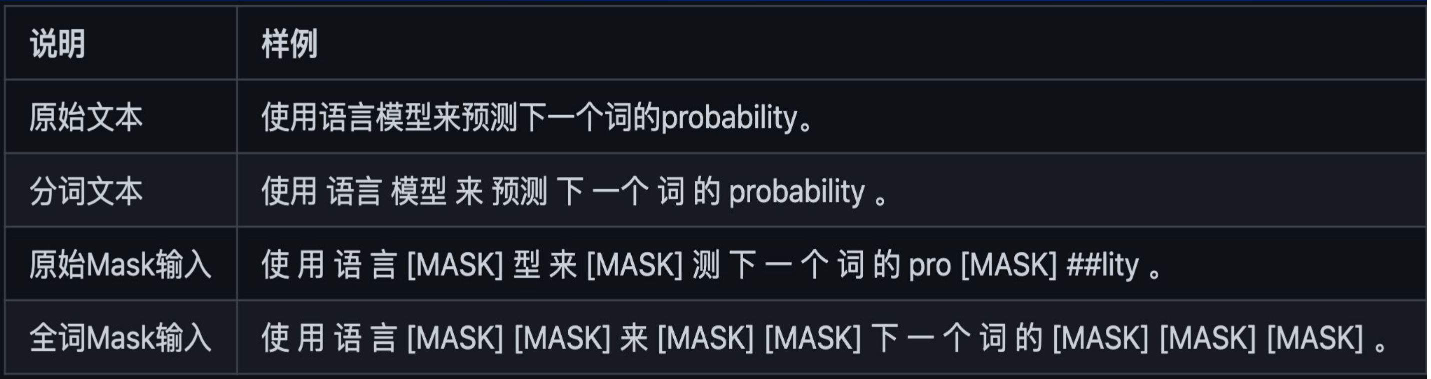
- ❑ Pre-trained model
- ❑ Classification model
- ❑ QA model

Pretrained model for classification + QA

- ❑ hfl/chinese-bert-wwm
  - ❑ Whole word masking
  - ❑ configure: BertMaskedLM



```
BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "directionality": "bidi",
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "pad_token_id": 0,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "transformers_version": "4.5.1",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```



说明	样例
原始文本	使用语言模型来预测下一个词的probability。
分词文本	使用 语言 模型 来 预测 下 一个 词 的 probability 。
原始Mask输入	使 用 语 言 [MASK] 型 来 [MASK] 测 下 一 个 词 的 pro [MASK] ##lity 。
全词Mask输入	使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK] [MASK] 。

Why two models?

Simplify our task and divide into two steps:

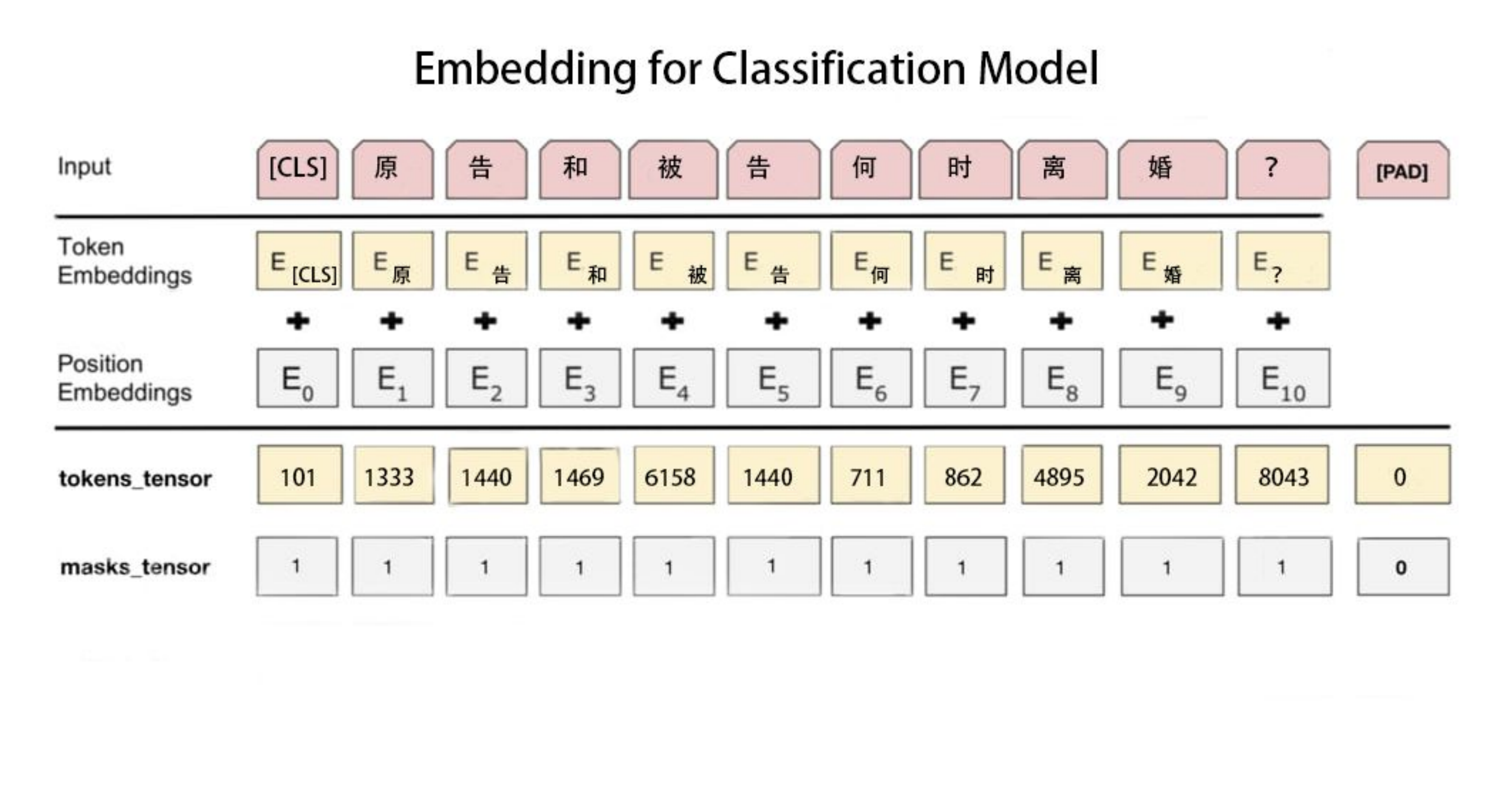
Classification and Question Answering

- First classify YES/NO or SPAN questions
- Classified result (SPAN) will feed into QA model

Classification--Data preprocessing

- ❑ Unbalanced Data: leads to biased training result
  - ❑ YES: 3,565; NO: 1,543
  - ❑ Other questions: 34,225
  - ❑ Ratio 1:6.7
- ❑ Solution: Undersampling
  - ❑ Randomly select 6,500 data from Other question set
  - ❑ Randomly merge with YES/NO set
  - ❑ Result: 11,608 (5,108 YES/NO+6,500 Other)

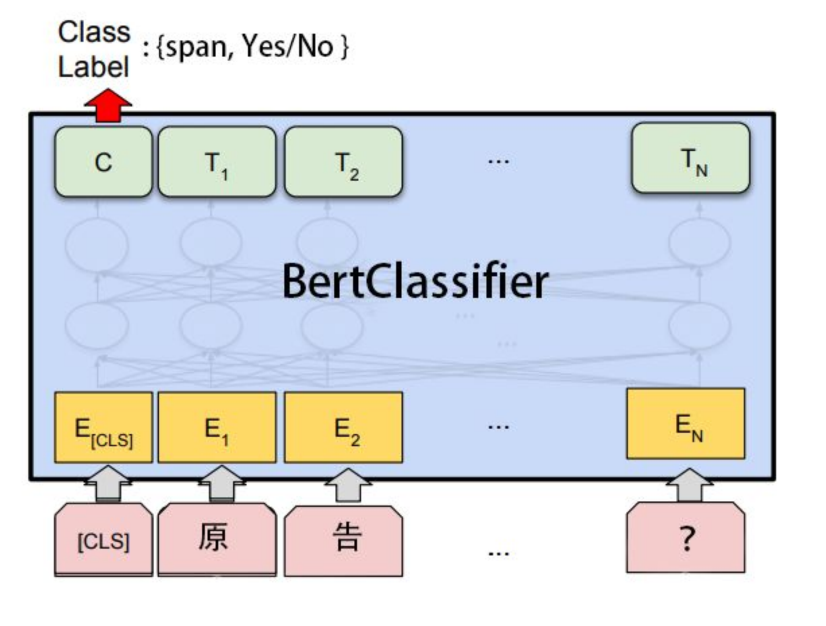
Classification--Embedding



Embedding for Classification Model

Input	[CLS]	原	告	和	被	告	何	时	离	婚	?	[PAD]
Token Embeddings	$E_{\text{[CLS]}}$	$E_{\text{原}}$	$E_{\text{告}}$	$E_{\text{和}}$	$E_{\text{被}}$	$E_{\text{告}}$	$E_{\text{何}}$	$E_{\text{时}}$	$E_{\text{离}}$	$E_{\text{婚}}$	$E_{?}$	
	+	+	+	+	+	+	+	+	+	+	+	
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	
tokens_tensor	101	1333	1440	1469	6158	1440	711	862	4895	2042	8043	0
masks_tensor	1	1	1	1	1	1	1	1	1	1	1	0

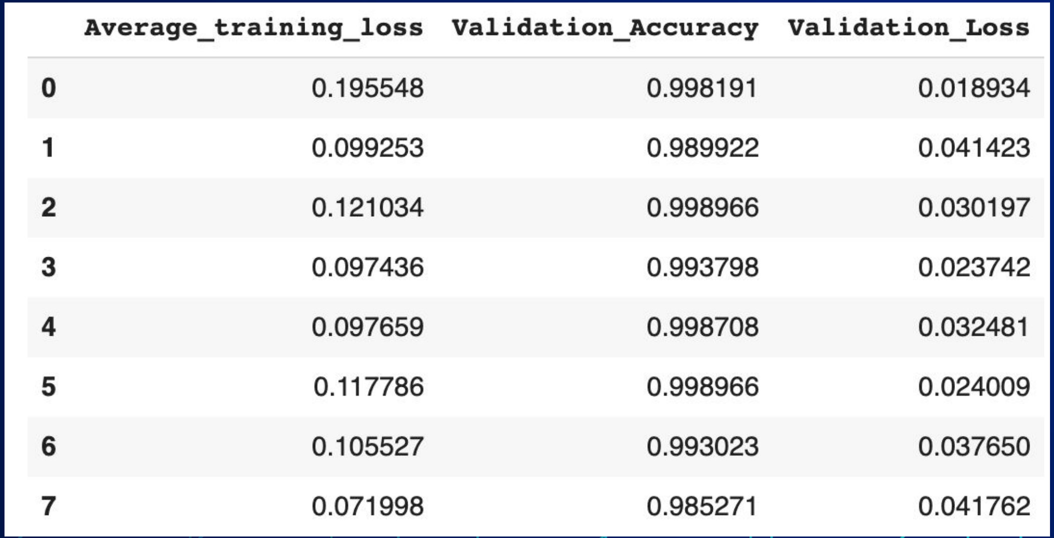
Classification--BertClassifier



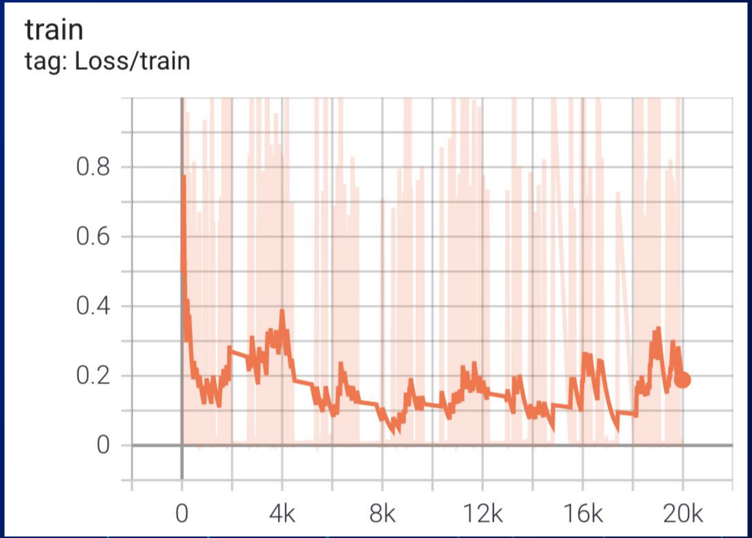
- ❑ Batch-size: 32
- ❑ Hidden-size: 768
- ❑ Num of labels: 2
- ❑ Learning rate: 6.25e-5
- ❑ Input size: (11608, 115)
- ❑ Output size: (B, N) i.e. (32,2)
- ❑ Epoch: 8
- ❑ Activation: Tanh
- ❑ Loss function: cross-entropy loss

Classification--Result+evaluation

- ❑ Test accuracy: 98.2946%
- ❑ Test loss: 0.015154



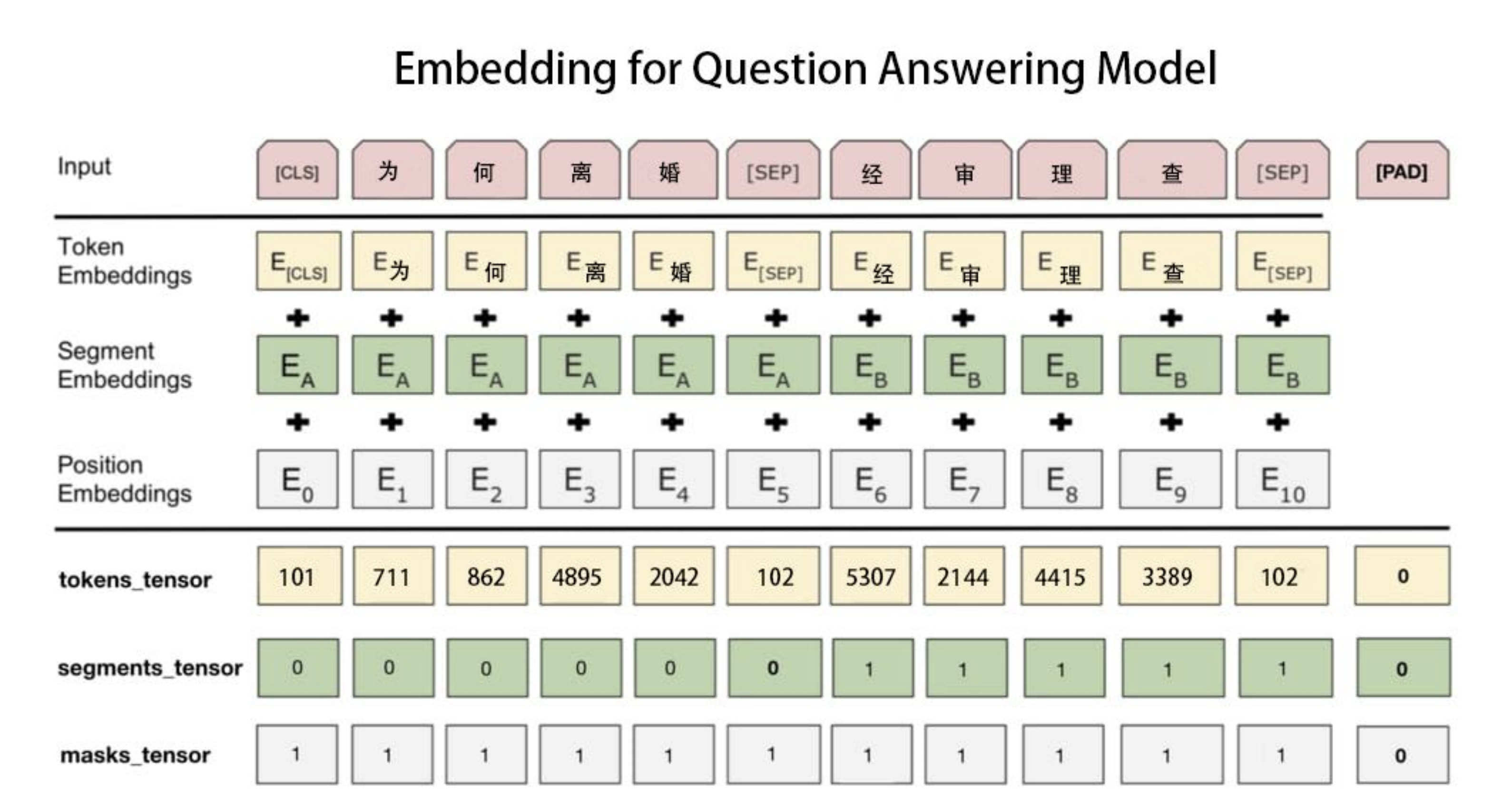
	Average_training_loss	Validation_Accuracy	Validation_Loss
0	0.195548	0.998191	0.018934
1	0.099253	0.989922	0.041423
2	0.121034	0.998966	0.030197
3	0.097436	0.993798	0.023742
4	0.097659	0.998708	0.032481
5	0.117786	0.998966	0.024009
6	0.105527	0.993023	0.037650
7	0.071998	0.985271	0.041762



train

tag: Loss/train

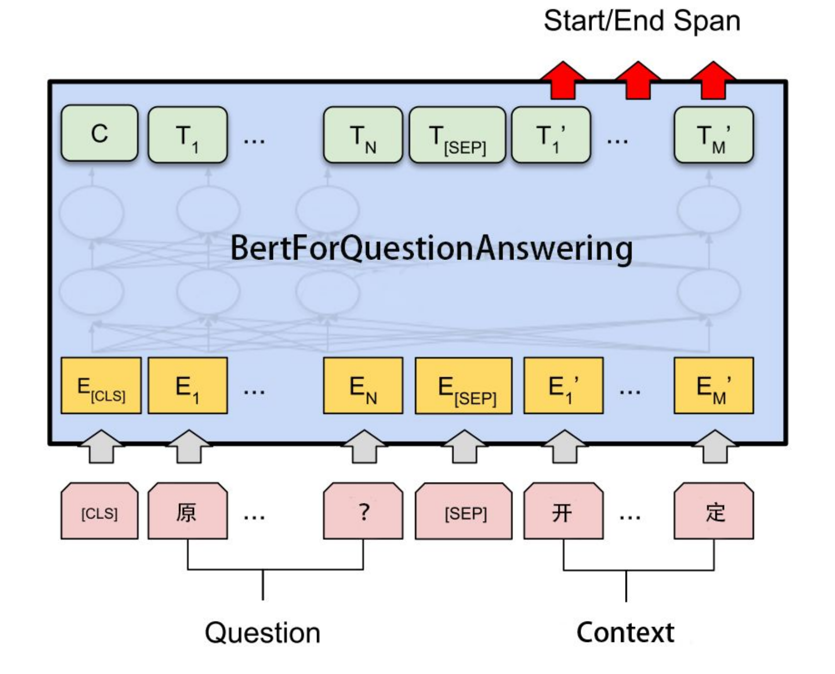
QA--Embedding



Embedding for Question Answering Model

Input	[CLS]	为	何	离	婚	[SEP]	经	审	理	查	[SEP]	**[PAD]**
Token Embeddings	$E_{[CLS]}$	$E_{为}$	$E_{何}$	$E_{离}$	$E_{婚}$	$E_{[SEP]}$	$E_{经}$	$E_{审}$	$E_{理}$	$E_{查}$	$E_{[SEP]}$	
	+	+	+	+	+	+	+	+	+	+	+	
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	
	+	+	+	+	+	+	+	+	+	+	+	
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	
tokens_tensor	101	711	862	4895	2042	102	5307	2144	4415	3389	102	**0**
segments_tensor	0	0	0	0	0	**0**	1	1	1	1	1	**0**
masks_tensor	1	1	1	1	1	1	1	1	1	1	1	**0**

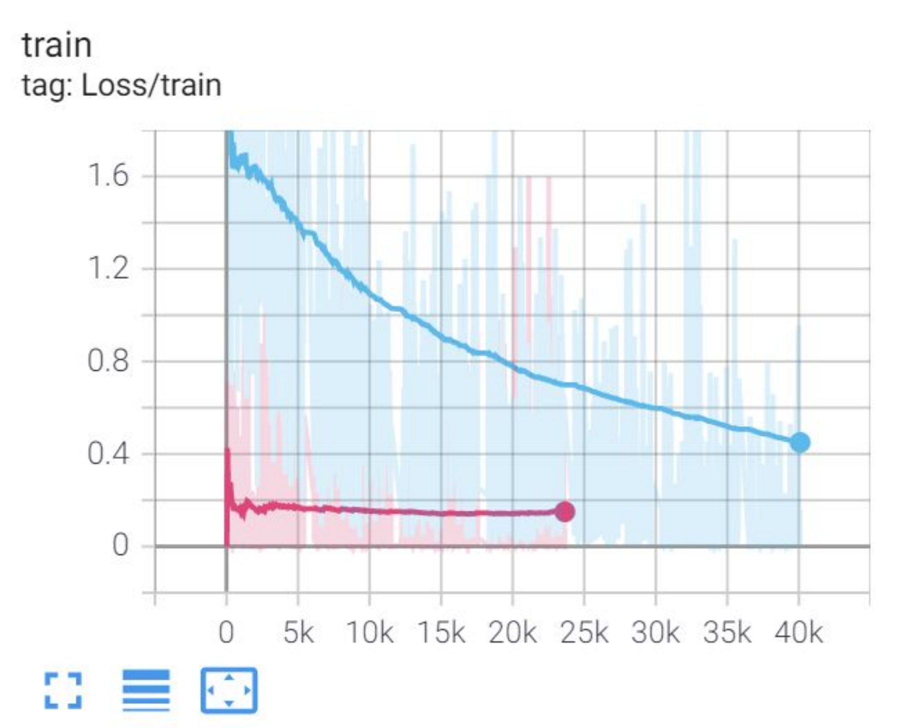
QA--BertForQuestionAnswering



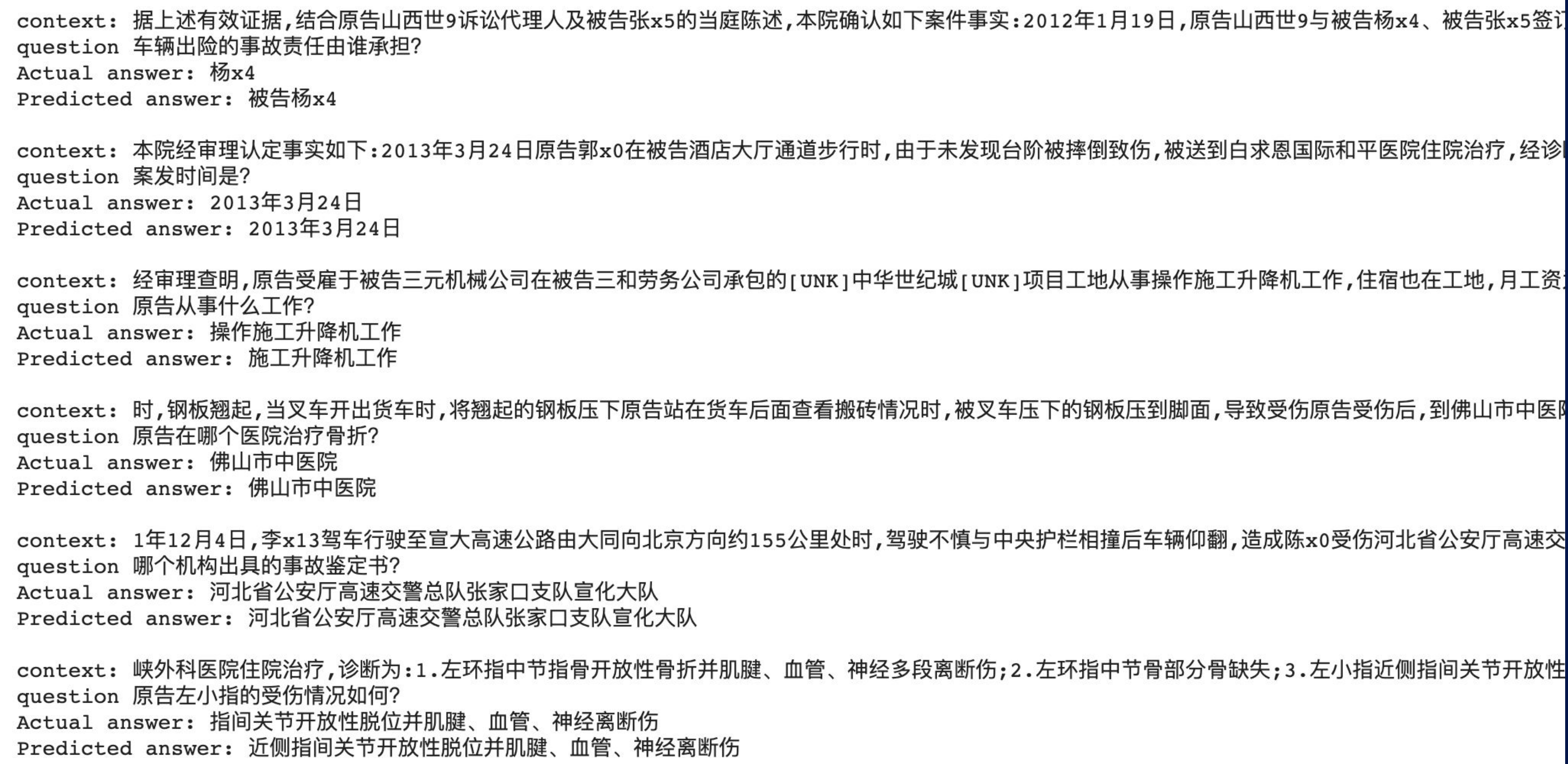
- ❑ Batch-size: 6
- ❑ Hidden-size: 768
- ❑ Two logits
- ❑ Learning rate: 6.25e-5
- ❑ Input size: (32,570, 259)
- ❑ Output size: (B, N) i.e. (6,259)
- ❑ Epoch: 10
- ❑ Activation: Tanh
- ❑ Loss function: cross-entropy loss

QA--Result+evaluation

- ❑ Test Accuracy: 73.4109%
- ❑ Test Loss: 1.863349
- ❑ F1-score: 0.81197



DEMO



```
context: 据上述有效证据,结合原告山西世9诉讼代理人及被告张x5的当庭陈述,本院确认如下案件事实:2012年1月19日,原告山西世9与被告杨x4、被告张x5签
question 车辆出险的事故责任由谁承担?
Actual answer: 杨x4
Predicted answer: 被告杨x4

context: 本院经审理认定事实如下:2013年3月24日原告郭x0在被告酒店大厅通道步行时,由于未发现台阶被摔倒致伤,被送到白求恩国际和平医院住院治疗,经诊
question 案发时间是?
Actual answer: 2013年3月24日
Predicted answer: 2013年3月24日

context: 经审理查明,原告受雇于被告三元机械公司在被告三和劳务公司承包的[UNK]中华世纪城[UNK]项目工地从事操作施工升降机工作,住宿也在工地,月工资
question 原告从事什么工作?
Actual answer: 操作施工升降机工作
Predicted answer: 施工升降机工作

context: 时,钢板翘起,当叉车开出货车时,将翘起的钢板压下原告站在货车后面查看搬砖情况时,被叉车压下的钢板压到脚面,导致受伤原告受伤后,到佛山市中医
question 原告在哪个医院治疗骨折?
Actual answer: 佛山市中医院
Predicted answer: 佛山市中医院

context: 1年12月4日,李x13驾车行驶至宣大高速公路由大同向北京方向约155公里处时,驾驶不慎与中央护栏相撞后车辆仰翻,造成陈x0受伤河北省公安厅高速交
question 哪个机构出具的事故鉴定书?
Actual answer: 河北省公安厅高速交警总队张家口支队宣化大队
Predicted answer: 河北省公安厅高速交警总队张家口支队宣化大队

context: 峡外科医院住院治疗,诊断为:1.左环指中节指骨开放性骨折并肌腱、血管、神经多段离断伤;2.左环指中节骨部分骨缺失;3.左小指近侧指间关节开放性
question 原告左小指的受伤情况如何?
Actual answer: 指间关节开放性脱位并肌腱、血管、神经离断伤
Predicted answer: 近侧指间关节开放性脱位并肌腱、血管、神经离断伤
```

4

Error Analysis

- ❑ Classification attempts
- ❑ Out of memory
- ❑ Tokenizer

Error Analysis I

Classification approach 1

Input: [CLS] Question [SEP] Context [SEP]

Output: three classes YES/NO/SPAN

↓

Classification approach 2

Input: [CLS] Question

Output: two classes i.e. whether it belongs to YES/NO question type

Bad Performance!!!

- Much longer context
- Questions unable to answer
- Not enough data

Error Analysis II

- ❑ Cuda out of memory:
  - ❑ Reduce batch size in training: 32 → 6
  - ❑ Set torch no_grad() in validation and test

Error Analysis III

- Unmatched tokenizer between question and context
  - Answers longer than cut context
  - Answer tokenization different from context
- Example:
  - Answer tokens: ‘2016’ ‘年’ ‘5’ ‘月’
  - Context tokens: ‘严’ ‘x1’ ‘#2016’ ‘年’

5

Future Work

- □ Slide window
- □ Multi-task model
- □ Tokenizer
- □ Answer YES/NO

Future Work

- ❑ Train: Slide window
  - ❑ Current: fixed window 64+[start]+128
  - ❑ Slide window: length 300
- ❑ Multitask Model: combine two models into one
- ❑ Different tokenizer between answer and context
- ❑ Answer YES/NO questions

Thank you!

Any Questions?