

实战八：新冠疫情文本匹配

导师：GAUSS

目录

1/ BERT模型

2/ 项目简介

3/ 数据介绍

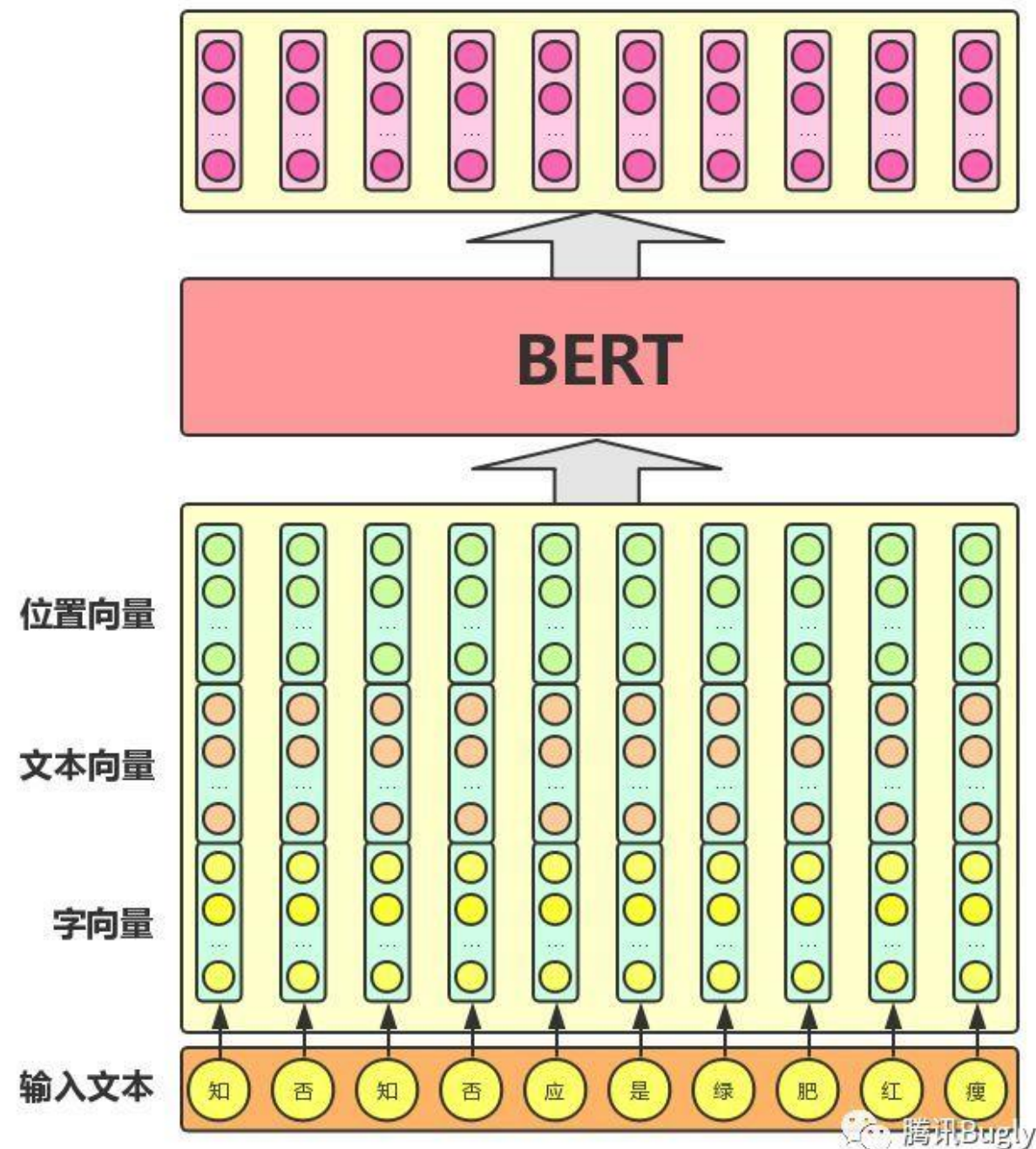
4/ 模型建模

5/ BERT模型实现文本匹配

BERT模型

BERT的输入/输出

BERT模型通过查询字向量表将文本中的每个字转换为一维向量，作为模型输入；模型输出则是输入各字对应的融合全文语义信息后的向量表示。



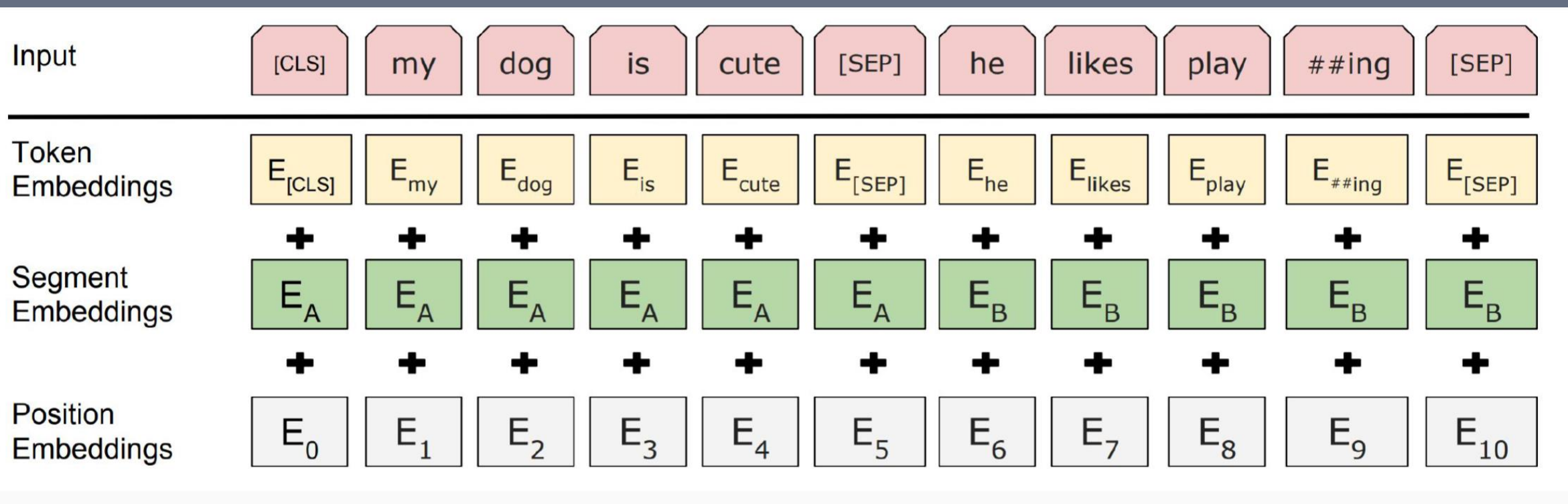
BERT的输入/输出

此外，模型输入除了**字向量**，还包含另外两个部分：

1. **文本向量**：该向量的取值在模型训练过程中自动学习，用于刻画文本的全局语义信息，并与单字/词的语义信息相融合
2. **位置向量**：由于出现在文本不同位置的字/词所携带的语义信息存在差异（比如：“我爱你”和“你爱我”），因此，BERT模型对不同位置的字/词分别附加一个不同的向量以作区分

BERT的输入/输出

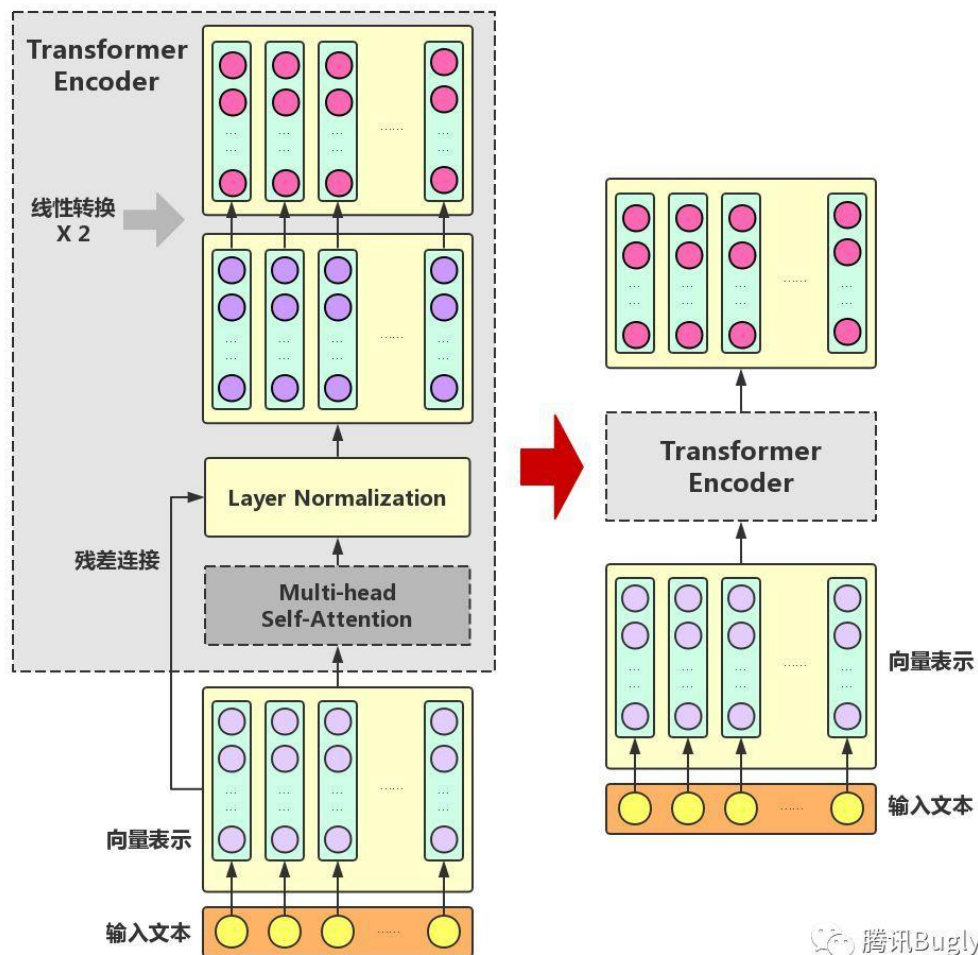
BERT最主要的组成部分便是，词向量 (token embeddings)、段向量(segment embeddings)、位置向量(position embeddings)



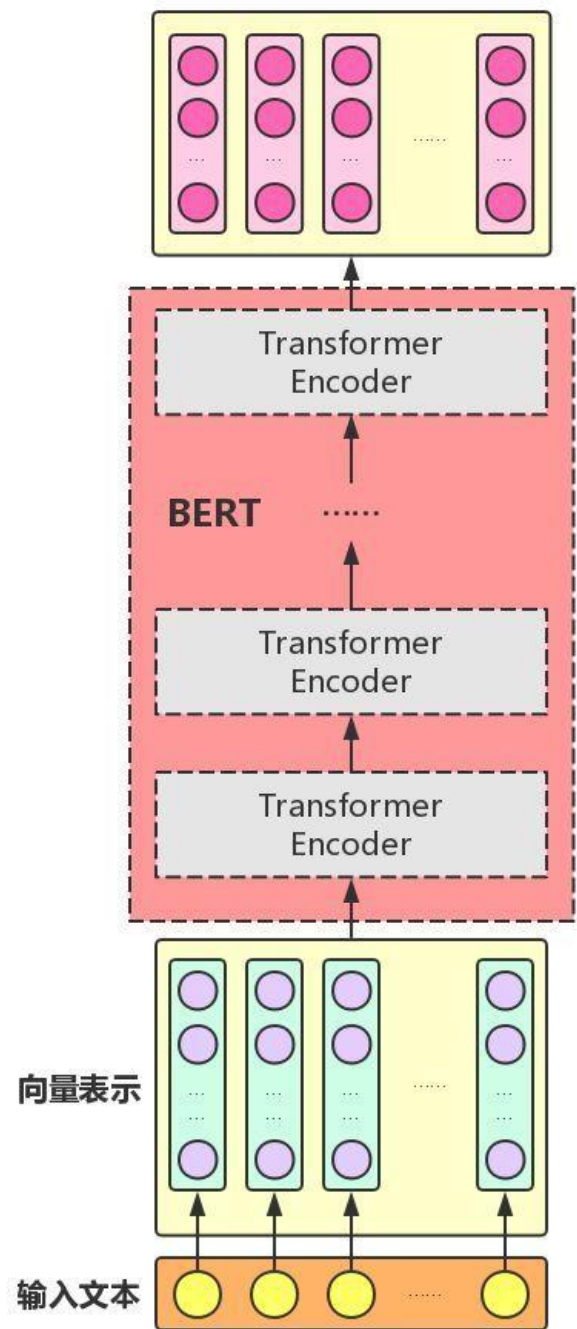
BERT模型



深度之眼
deepshare.net

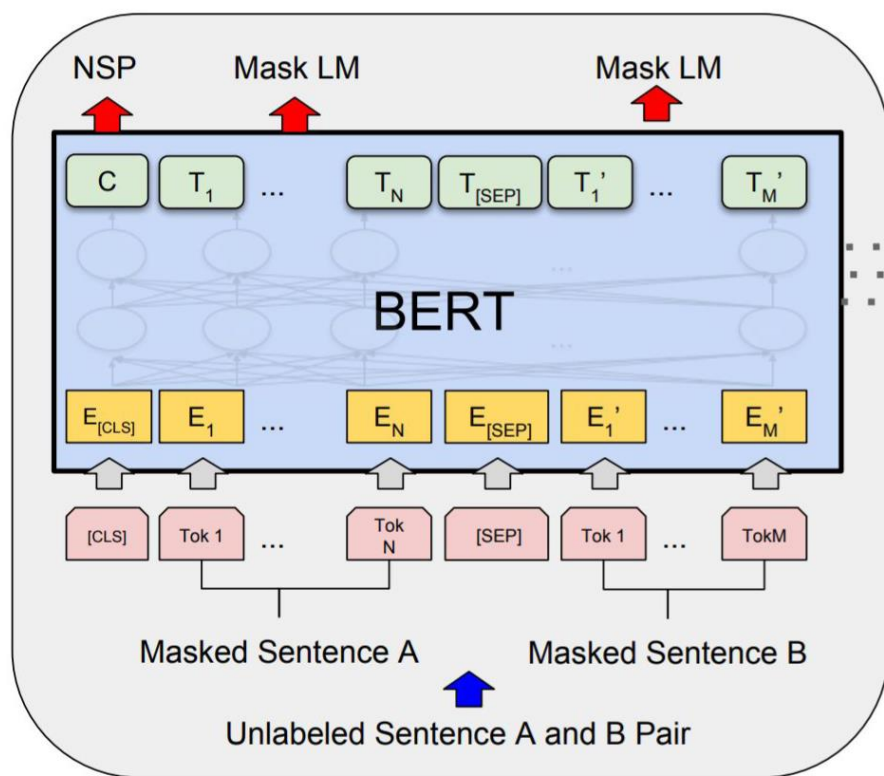


腾讯Bugly

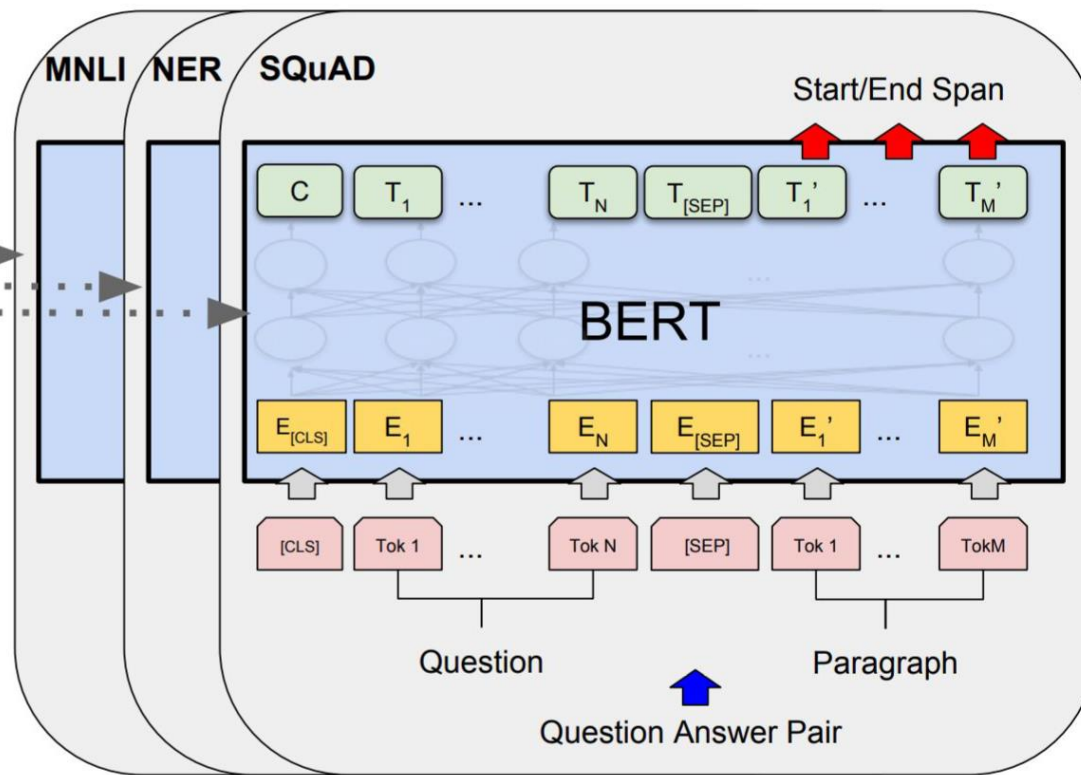


腾讯Bugly

BERT模型

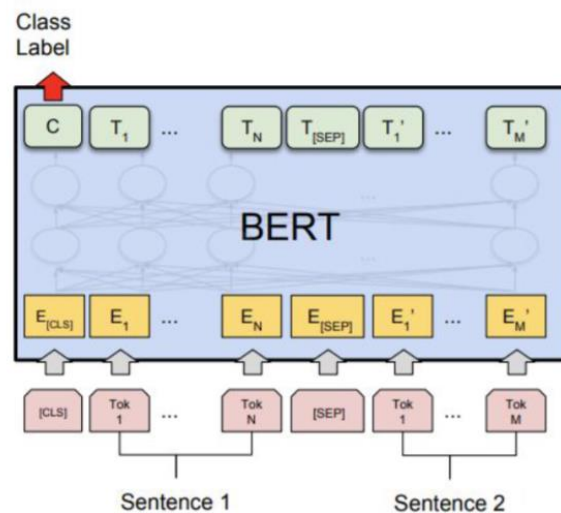


Pre-training

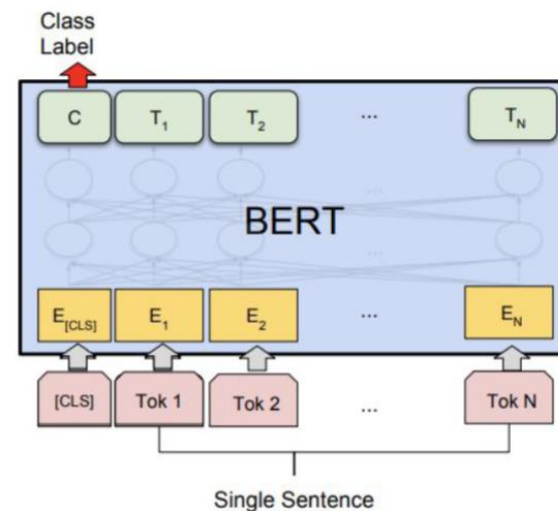


Fine-Tuning

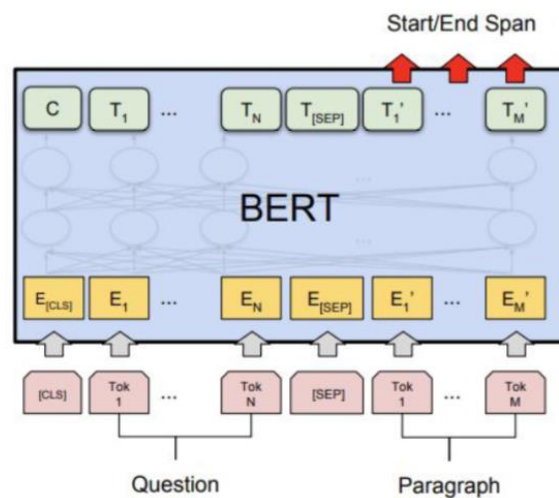
BERT模型



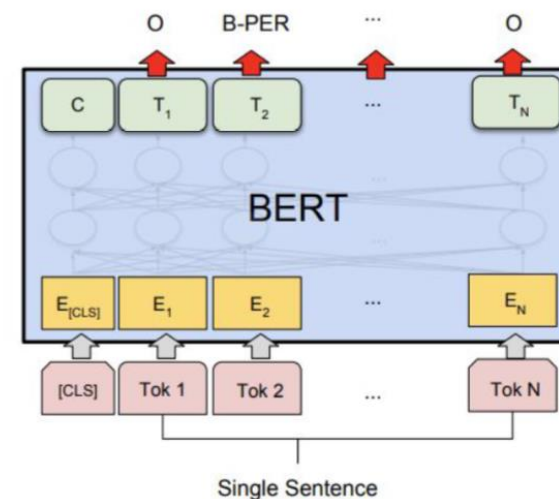
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



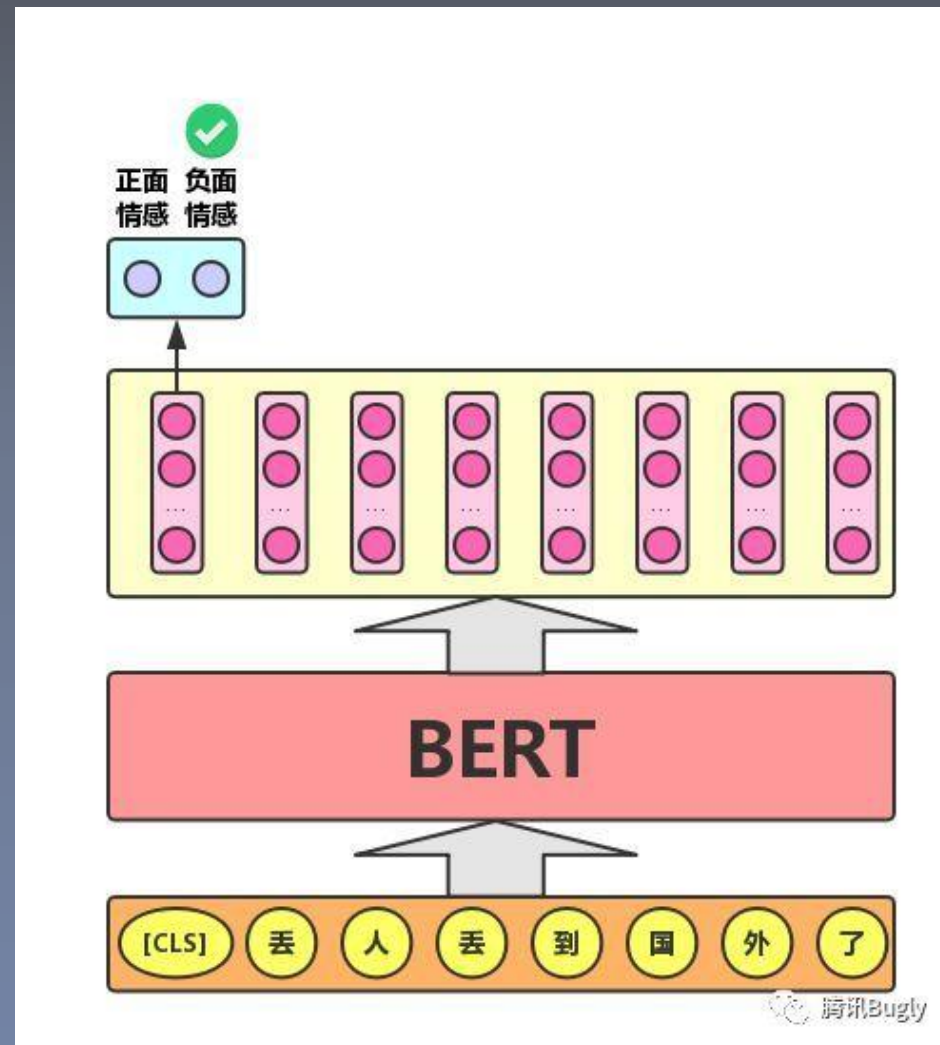
(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

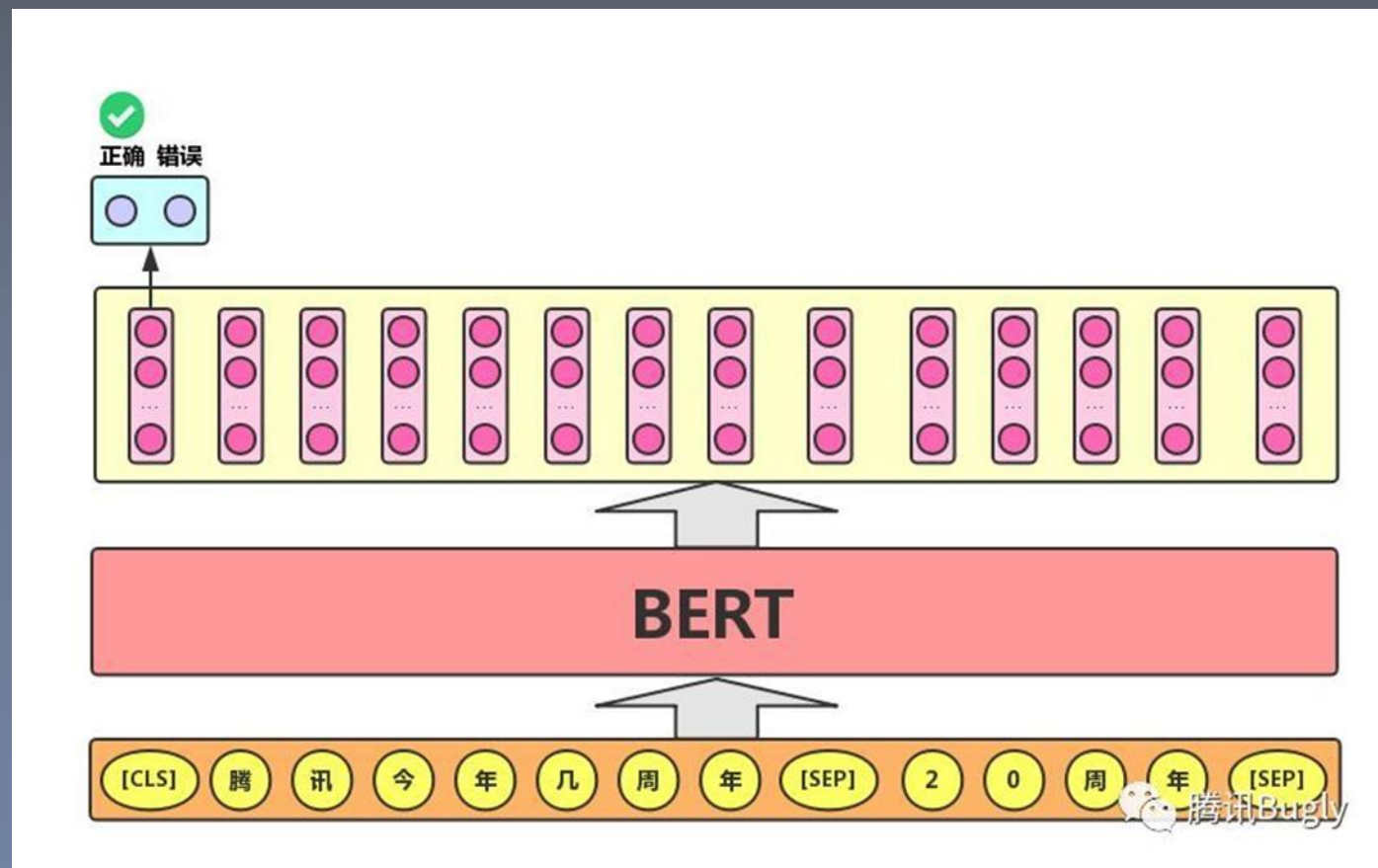
BERT应用场景

对于文本分类任务，BERT模型在文本前插入一个[CLS]符号，并将该符号对应的输出向量作为整篇文本的语义表示，用于文本分类，如下图所示。



BERT应用场景

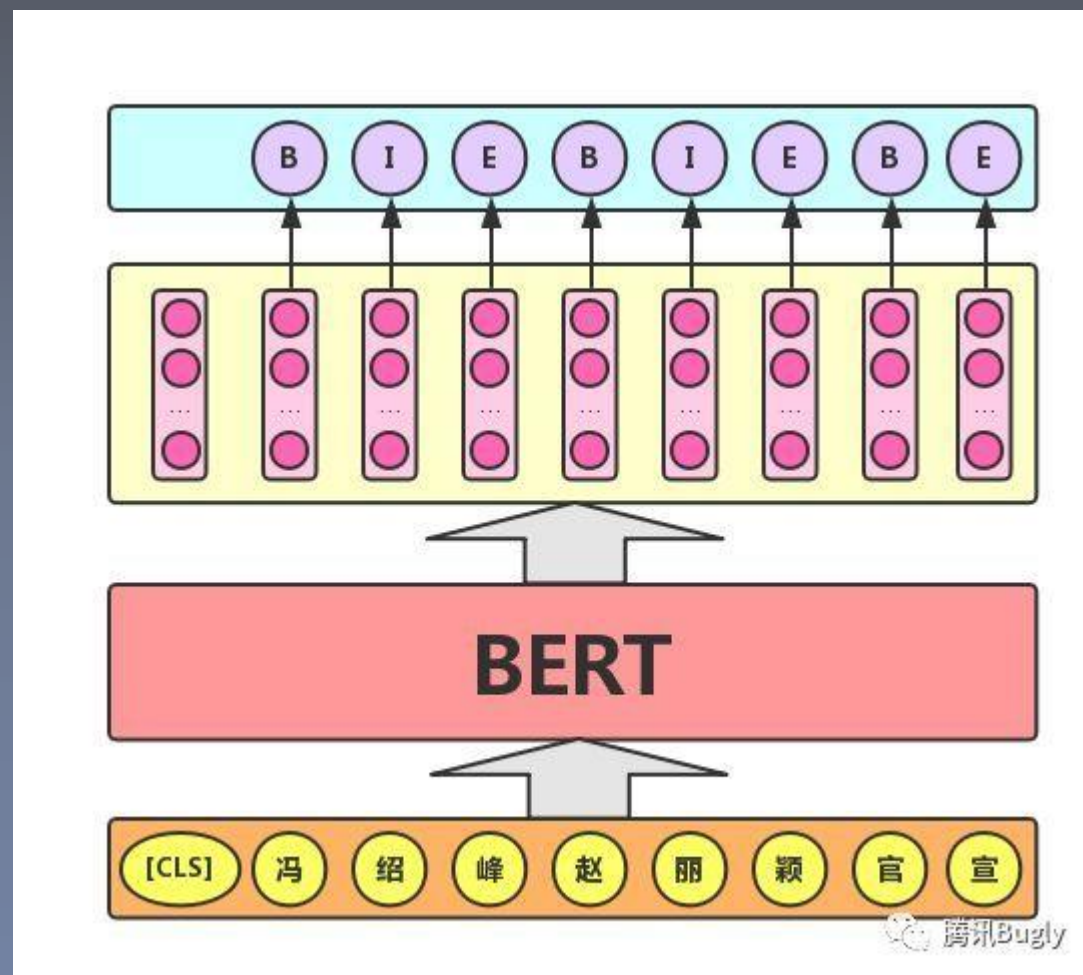
文本匹配任务：该任务的实际应用
场景包括：**问答**（判断一个问题与
一个答案是否匹配）、**语句匹配**
（两句话是否表达同一个意思）等。
对于该任务，BERT模型除了添加
[CLS]符号并将对应的输出作为文本
的语义表示，还对输入的两句话用
一个[SEP]符号作分割，并分别对两
句话附加两个不同的文本向量以作
区分。





BERT应用场景

序列标注任务：该任务的实际应用场景包括：中文分词&新词发现（标注每个字是词的首字、中间字或末字）、答案抽取（答案的起止位置）等。对于该任务，BERT模型利用文本中每个字对应的输出向量对该字进行标注（分类），如下图所示(B、I、E分别表示一个词的第一个字、中间字和最后一个字)。



项目简介

背景知识

- 面对疫情抗击，疫情知识问答应用得到普遍推广。如何通过自然语言技术将问答进行相似分类仍然是一个有价值的问题。**如识别患者相似问题，有利于理解患者真正诉求，帮助快速匹配准确答案，提升患者获得感；归纳医生相似答案，有助于分析答案规范性，保证疫情期间问诊规范性，避免误诊。**
- 本次比赛达摩院联合医疗服务机构妙健康发布疫情相似句对判定任务。比赛整理近万条真实语境下疫情相关的肺炎、支原体肺炎、支气管炎、上呼吸道感染、肺结核、哮喘、胸膜炎、肺气肿、感冒、咳血等患者提问句对，要求选手通过自然语言处理技术识别相似的患者问题。本次比赛成果将作为原子能力**助力疫情智能问答应用技术精准度提升，探索下一代医疗智能问答技术**，具有广泛的技术和公益价值。

项目介绍

比赛主打疫情相关的呼吸领域的真实数据积累，数据粒度更加细化，判定难度相比多科室文本相似度匹配更高，同时问答数据也更具时效性。本着宁缺毋滥的原则，问题的场地限制在20字以内，形成相对规范的句对。要求选手通过自然语义算法和医学知识识别相似问答和无关的问题。

数据介绍

数据介绍

本次大赛数据包括：脱敏之后的医疗问题数据对和标注数据。医疗问题涉及“肺炎”、“支原体肺炎”、“支气管炎”、“上呼吸道感染”、“肺结核”、“哮喘”、“胸膜炎”、“肺气肿”、“感冒”、“咳血”等10个病种。

数据共包含train.csv、dev.csv、test.csv三个文件，其中给参赛选手的文件包含训练集train.csv和验证集dev.csv，测试集test.csv 对参赛选手不可见。

每一条数据由Id, Category, Query1, Query2, Label构成，分别表示问题编号、类别、问句1、问句2、标签。Label表示问句之间的语义是否相同，若相同，标为1，若不相同，标为0。其中，训练集、验证集Label已知，测试集Label未知。

数据示例

示例

编号：0

类别：肺炎

问句1：肺部发炎是什么原因引起的？

问句2：肺部发炎是什么引起的

标签:1

编号：1

类别：肺炎

问句1：肺部发炎是什么原因引起的？

问句2：肺部炎症有什么症状

标签:0

数据示例

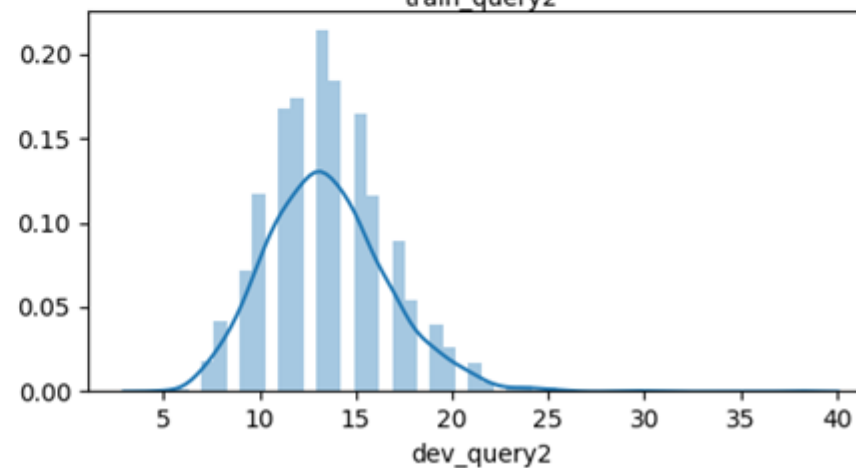
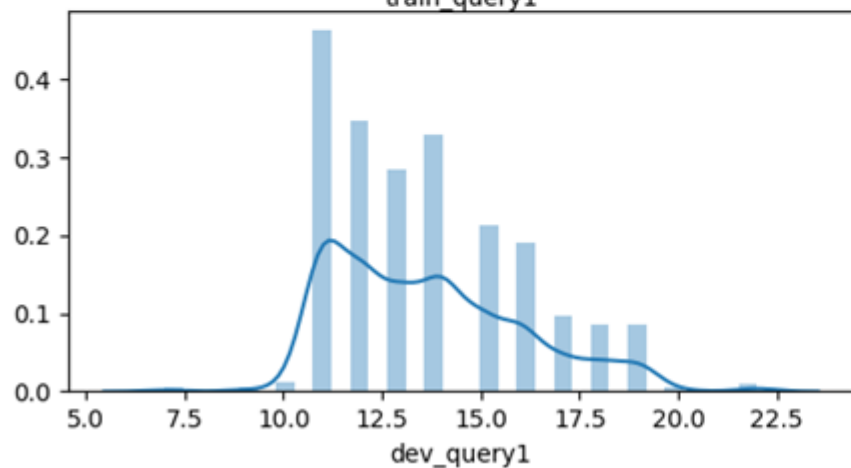
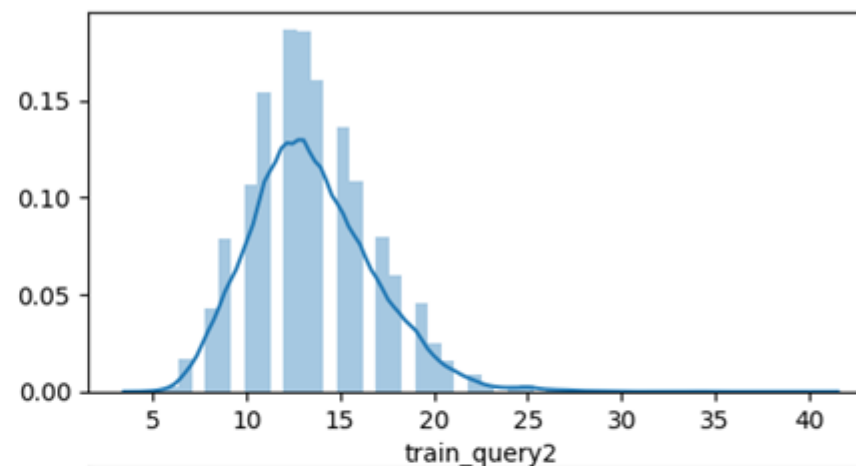
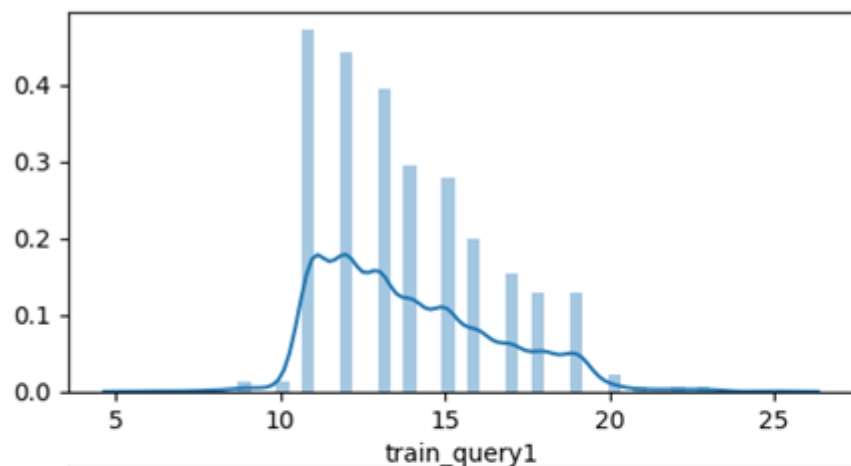
query1	query2	label
剧烈运动后咯血,是怎么回事?	剧烈运动后咯血是什么原因?	1
剧烈运动后咯血,是怎么回事?	剧烈运动后为什么会咯血?	1
剧烈运动后咯血,是怎么回事?	剧烈运动后咯血, 应该怎么处理?	0
剧烈运动后咯血,是怎么回事?	剧烈运动后咯血, 需要就医吗?	0
剧烈运动后咯血,是怎么回事?	剧烈运动后咯血, 是否很严重?	0

评估指标

本次比赛测评指标为准确率，计算公式为：

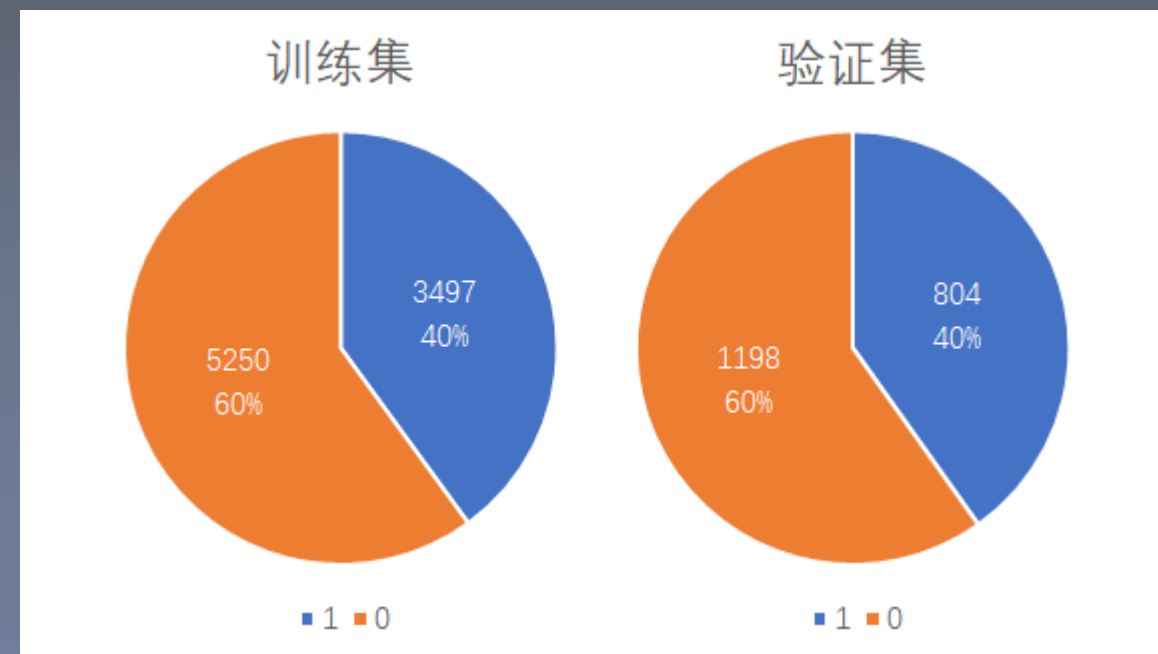
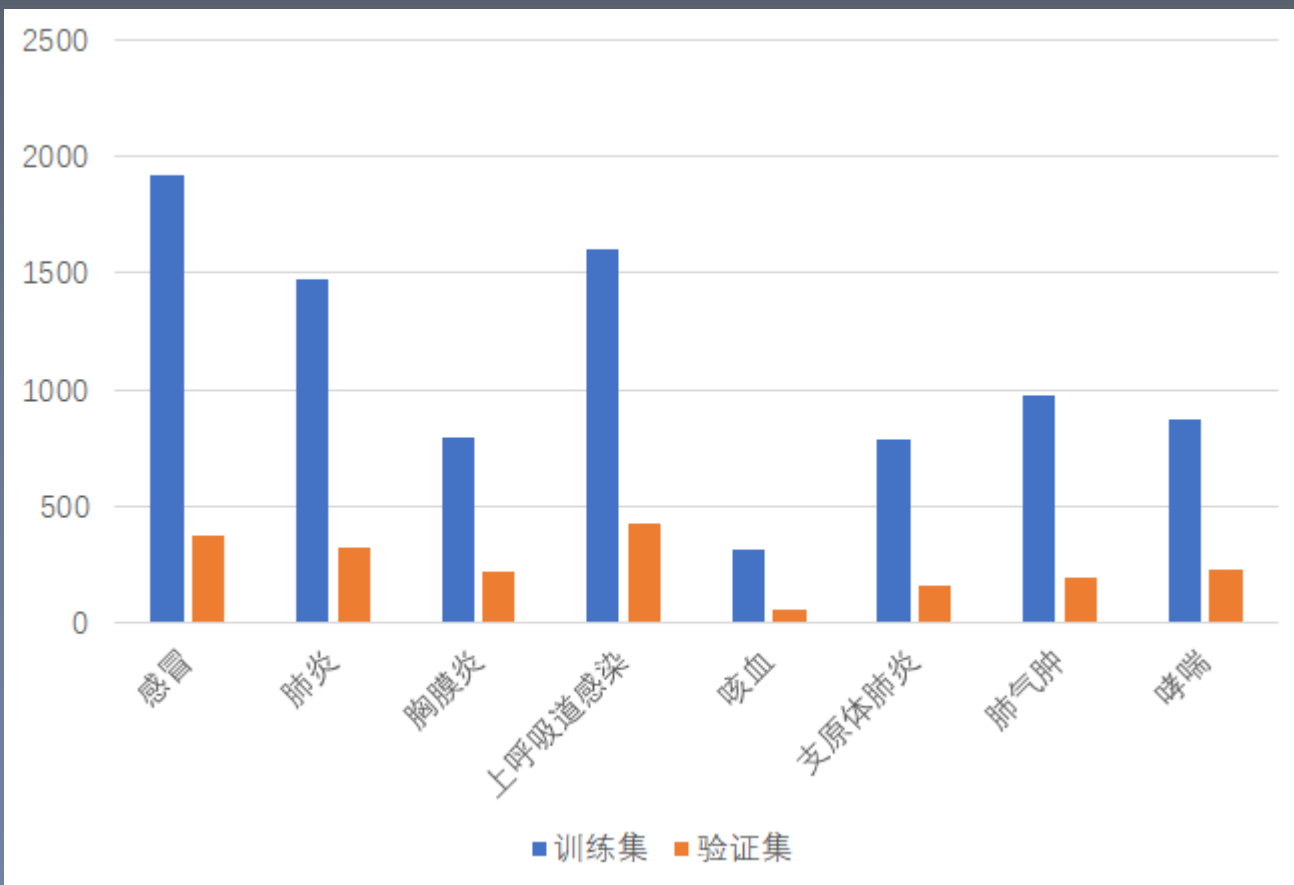
准确率 = 正确预测数目 / 总问题对数目

文本长度





标签分布



模型建模

模型框架

The framework of model



- Bidirectional RNNs (GRU、LSTM)
- CNN
- Transformer
- Capsule等

模型选择

- ESIM
- Infer Sentence
- DSSM

- BERT
- Roberta

BERT模型实现文本匹配

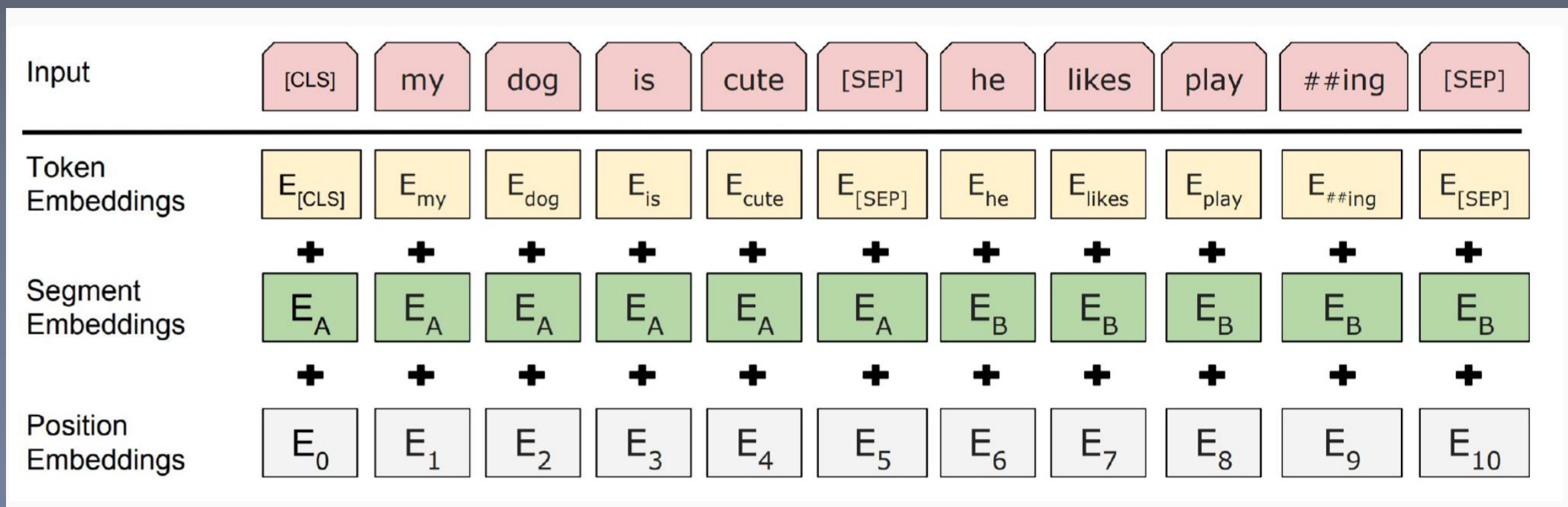
步骤



深度之眼
deepshare.net



数据输入



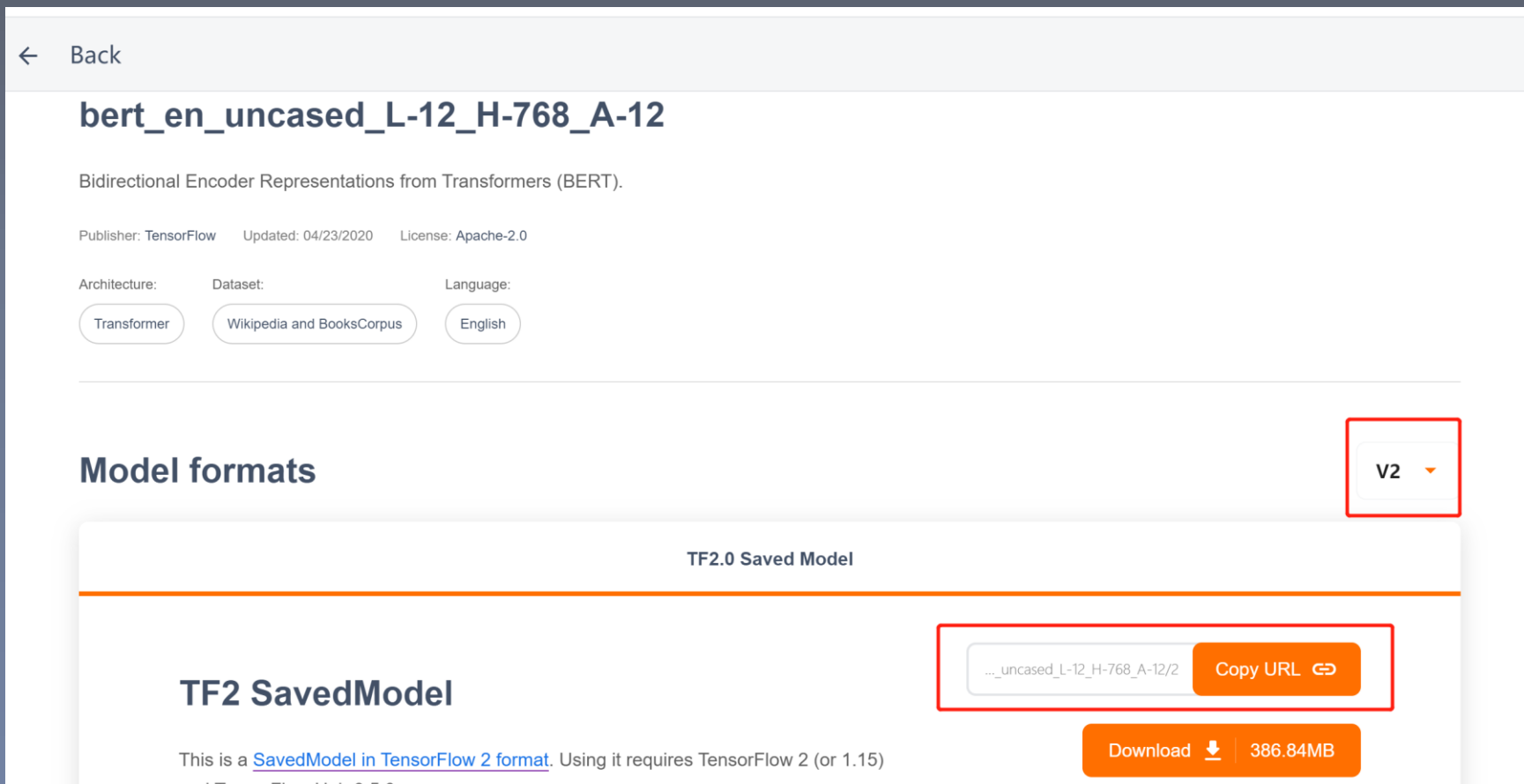
BERT模型构建

Bert模型

该模型的地址如下：

https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/2

其中，末尾的 2 为该模型的版本号。



← Back

bert_en_uncased_L-12_H-768_A-12

Bidirectional Encoder Representations from Transformers (BERT).

Publisher: TensorFlow Updated: 04/23/2020 License: Apache-2.0

Architecture: Dataset: Language:

Transformer Wikipedia and BooksCorpus English

Model formats

V2

TF2.0 Saved Model

TF2 SavedModel

This is a [SavedModel in TensorFlow 2 format](#). Using it requires TensorFlow 2 (or 1.15) and TensorFlow Hub 0.5.0 or newer.

Copy URL

Download 386.84MB

BERT模型构建

```
max_seq_length = 128 # Your choice here.

input_word_ids = tf.keras.layers.Input(shape=(max_seq_length,), dtype=tf.int32,
                                         name="input_word_ids")

input_mask = tf.keras.layers.Input(shape=(max_seq_length,), dtype=tf.int32,
                                     name="input_mask")

segment_ids = tf.keras.layers.Input(shape=(max_seq_length,), dtype=tf.int32,
                                     name="segment_ids")

bert_layer = hub.KerasLayer("https://tfhub.dev/tensorflow/bert_en_uncased_L-
12_H-768_A-12/2", trainable=True)

pooled_output, sequence_output = bert_layer([input_word_ids, input_mask, segment
_ids])
```

参考

<https://cloud.tencent.com/developer/article/1389555>

<https://github.com/zzy99/epidemic-sentence-pair>

总结

本节小结

Summary

实战八：新冠 疫情文本匹配

项目简介

BERT模型

数据介绍

模型建模

BERT模型实现文本匹配

结语

——我 说——

看过千万代码，不如实践一把！





深度之眼
deepshare.net

联系我们：

电话：18001992849

邮箱：service@deepshare.net

QQ：2677693114



公众号



客服微信

