本周报告

2020/05/07 1652792 罗吉皓

本周主要针对模型和数据集进行修改调整。

机器学习模型输入维度说明

在初始 BERT 的基础上，我对于机器学习的输入向量进行进一步的改进：

通过 BERT 的输出，我们可以得到 768 维度的特征向量，该向量是通过 BERT 模型预训练得到的，可以在最大程度地保留句中词语语义信息的基础上，获取字符、词语、语句的向量表示

命名实体向量：在本实验中的命名实体共 4 个：OVTP，则在这里我们构成一个 4 维度的向量，分别对应 O，V，T，P，该分词对应的位置被定义为 1，其余为 0

依存关系向量：以候选句中词语数量作为维度组成 30 维向量，核心地理位置在核心动词位置标 1，其余都是 0

在实际实验中，依存关系向量的引入对于整体模型起到负面的作用，分析后发现，整个依存关系向量中可能存在太多的 0 起到干扰的元素，可能会影响模型的最后效果。

数据集修改

在这一次模型更新中，我借鉴了 BIESO 序列标注规则。在 BIESO 序列标注规则中，“B”表示事件触发词的开始，“I”表示事件触发词中间，“E”表示事件触发词结尾，“S”表示事件触发词为单独词语，“O”表示非事件触发词。 本实验中，主要的标注重点在地理名词标注方面，因此将原来的 OVTP 调整如下：

B-place：地理位置的开始

I-place：地理位置中间

E-place：地理位置结束

B-targetplace：核心地理位置的开始

I-targetplace：核心地理位置的中间

E-targetplace：核心地理位置的结束

B-verb：核心动词的开始

I-verb：核心动词中间

E-verb：核心动词结束

O：其他词

使用 BIESO 序列标注规则使整体标注信息更为规范。另外一方面，CRF 层在整个模型中的作用个人理解是在标签预测的同时也限制一些规则，比如说 I 一定在 B 之后，标准化模型输入也可以帮助发挥 CRF 层的作用。

相应的原来的 4 位向量变成 10 维向量

对比实验说明

对比实验方面我主要设计了两个实验，实验一中主要针对不同的特征向量提取方式，在实验中我们首先针对文本特征向量提取与否进行对比实验，在网络结构相同，参数相同的情况下进行模型训练并对比预测效果。在预训练语言模型中，除了 BERT 模型，我还选取了 ERINE 模型进行对比实验。Ernie 模型被称为中文文本预训练最强网络，他在 BERT 的基础上更注重于词法结构，语法结构，语义信息的统一建模，中文文本相对于其他语言文本来说其最大的分析难度便在于其丰富的语法结构和语义信息，ERNIE 在这一方面的强化让他在中文文本分析中更有优势。在实验中我们着重选取了两种目前最为流行的预训练语言模型，通过对比实验帮助选取最好的模型进行后续解决方案的搭建。

另外一方面我选取了目前命名实体标注方面三个标注效果比较好的网络结构 CNN_LSTM 模型，BiLSTM 模型以及 BiLSTM_CRF 三个模型进行实验，如图所示。

CNN_LSTM 模型中我主要使用了简单的卷积层累加，而后将输出结果输入后续 LSTM 网络，同时加入 Dropout 层防止过拟合。由于是序列学习问题，我在之后加入了 Time_distributed 层，帮助选取最佳的标注结果。

Bi-LSTM 网络中我通过加入双向 LSTM 网络帮助模型进一步理解上下文语境，LSTM 网络帮助上流文本信息传递到后续的学习，而 Bi-LSTM 帮助下文的信息反向传递到上文，从而帮助模型训练更为充分，比较适合像提取核心地点这种更为细粒度的标注需求。

CRF 层的介绍在上文中我已经有所提及，在 BiLSTM 模型的基础上，我去除了 Time_distributed 层而采用了 CRF 条件随机场，来对比最后的标注效果。

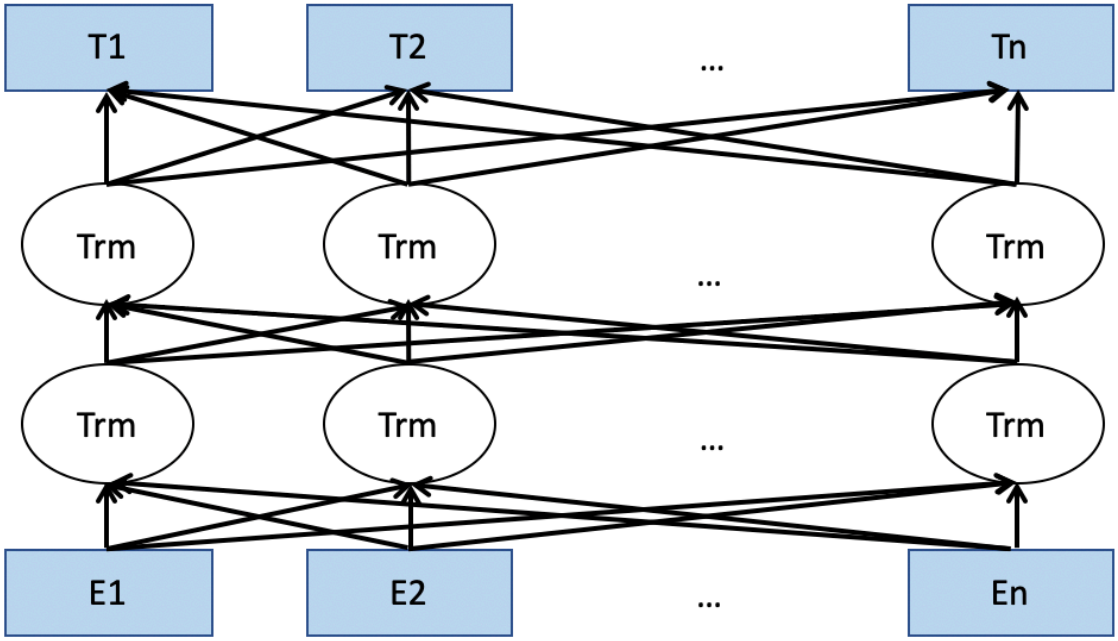
实验中三个模型 Dropout 设置为 0.5，批处理大小设置为 64，训练轮次 40 轮。

BERT 及 ERINE 介绍

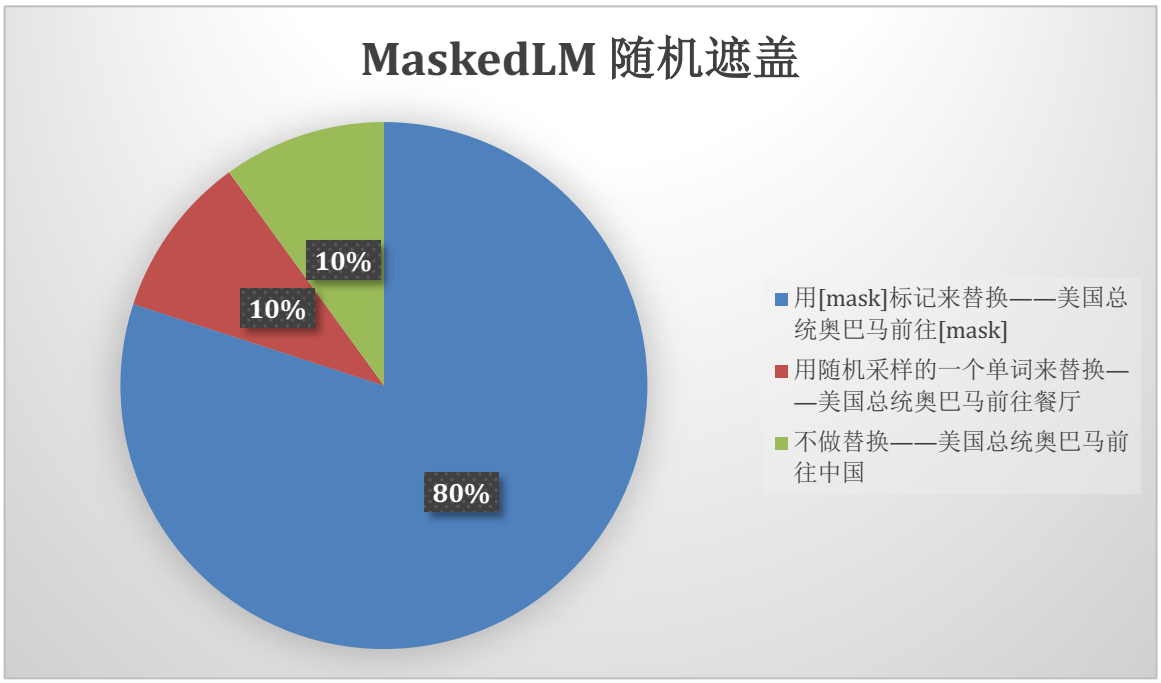
由于长新闻特征比较明显，语料量也比较大，很容易取得比较不错的结果。但是我们的语料比较少，特征不是很明显时候直接训练可能会导致模型过拟合，泛化能力很差，此时我选择使用预训练的词 Embedding 层来提高模型的泛化能力。

BERT 简介

BERT（Bidirectional Encoder Representations from Transformers）是 2018 年 google 提出来的预训练的语言模型。BERT 在原来 word2vec 模型的基础上，通过无监督的学习掌握了很多自然语言的一些语法或者语义知识，大大提升了词向量模型泛化能力，通过 MaskLM 训练和 Transformer 网络结构。编码解码帮助 Bert 在提取关系特征方面的能力大大提升，之后在做后续任务时减少了难度。

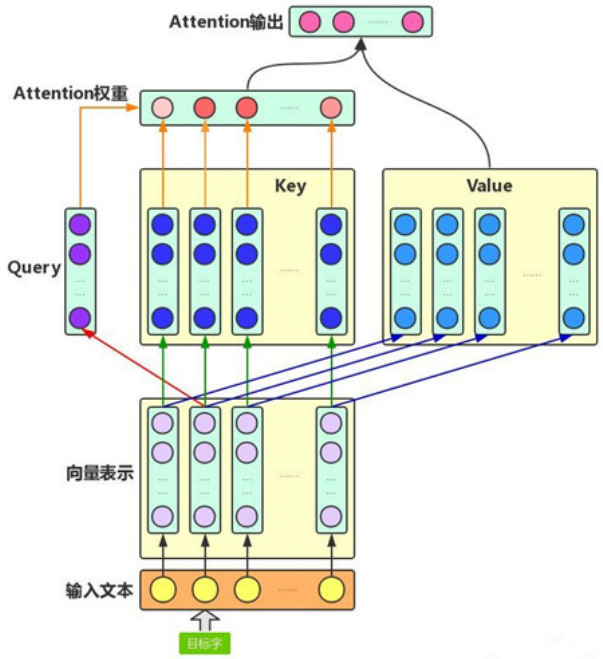


Masked LM 的引入帮助 BERT 在语言模型训练上的能力大大提升。相比起 CBOW 模型在训练的过程中把每一个分词都进行预测，MaskedLM 在训练过程中使用特殊符号随机遮盖（mask） 15%的分词作为需要进行预测的对象。而考虑到遮盖词的特殊符号，在实际 NLP 任务中其实是不存在，因此为了保证不影响后续的训练和预测任务，会按照一定的比例在需要预测的词位置上以原词或者其他随机词的方式进行遮掩。这么做的主要原因是：在后续微调任务中语句中并不会出现[MASK]标记，而且这么做的另一个好处是：预测一个词汇时，模型并不知道输入对应位置的词汇是否为正确的词汇（10%概率），这就迫使模型更多地依赖于上下文信息去预测词汇，并且赋予了模型一定的纠错能力。



无论是 RNN 递归神经网络还是 CNN 卷积神经网络，在处理 NLP 任务时其实都有缺陷：CNN 卷积神经网络由于其卷积操作并不适合序列化的文本，容易造成信息丢失，而 RNN 递归神经网络由于没有并行化，很容易超出内存限制，因此 BERT 选择使用 Transformer 模型来替代传统 RNN 递归神经网络和 CNN 卷积神经网络，用来实现机器翻译。

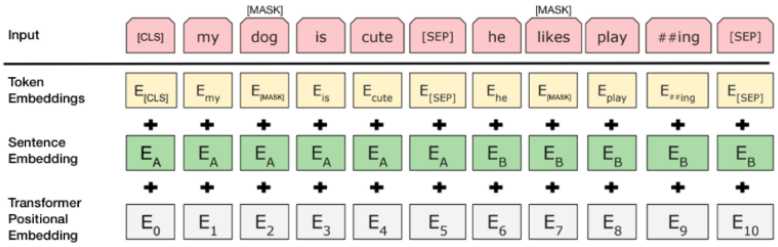
Transformer 模型由多个 encoder（编码器）和 decoder（解码器）组合而成，其中 encoder 负责接收文本作为输入，由 multi-head attention 与 feed-forward 构成， decoder 负责预测任务的结果。BERT 的核心是 Transformer 模型中的 encoder 部分，通过一次性读取整个文本序列，帮助模型能够基于单词的上下文进行学习



Attention 机制的中文名叫“注意力机制”，顾名思义，它的主要作用是让神经网络把“注意力”放在一部分输入上，即：区分输入的不同部分对输出的影响。这里，我们从增强字/词的语义表示这一角度来理解一下 Attention 机制。

我们知道，一个字/词在一篇文本中表达的意思通常与它的上下文有关。比如：光看“鹄”字，我们可能会觉得很陌生（甚至连读音是什么都不记得吧），而看到它的上下文“鸿鹄之志”后，就对它立马熟悉了起来。因此，字/词的上下文信息有助于增强其语义表示。同时，上下文中的不同字/词对增强语义表示所起的作用往往不同。比如在上面这个例子中，“鸿”字对理解“鹄”字的作用最大，而“之”字的作用则相对较小。为了有区分地利用上下文字信息增强目标字的语义表示，就可以用到 Attention 机制。

Attention 机制主要涉及到三个概念：Query、Key 和 Value。在上面增强字的语义表示这个应用场景中，目标字及其上下文的字都有各自的原始 Value，Attention 机制将目标字作为 Query、其上下文的各个字作为 Key，并将 Query 与各个 Key 的相似性作为权重，把上下文各个字的 Value 融入目标字的原始 Value 中。如下图所示，Attention 机制将目标字和上下文各个字的语义向量表示作为输入，首先通过线性变换获得目标字的 Query 向量表示、上下文各个字的 Key 向量表示以及目标字与上下文各个字的原始 Value 表示，然后计算 Query 向量与各个 Key 向量的相似度作为权重（最终形成每个目标字与其上下文的字的权重关系，权重和为 1），加权融合目标字的 Value 向量和各个上下文字的 Value 向量（其实就是做了点乘），作为 Attention 的输出，即：目标字的增强语义向量表示。



从上图中可以看出，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。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

ERINE

Google 最近提出的 BERT 模型，通过随机屏蔽 15%的字或者 word，利用 Transformer 的多层 self-attention 双向建模能力，在各项 nlp 下游任务中(如 sentence pair classification task，singe sentence classification task，question answering task) 都取得了很好的成绩。但是，BERT 模型主要是聚焦在针对字或者英文 word 粒度的完形填空学习上面，没有充分利用训练数据当中词法结构，语法结构，以及语义信息去学习建模。比如 “我要买苹果手机”，BERT 模型 将 “我”，“要”， “买”，“苹”， “果”，“手”， “机” 每个字都统一对待，随机 mask，丢失了“苹果手机” 是一个很火的名词这一信息，这个是词法信息的缺失。同时 我 + 买 + 名词 是一个非常明显的购物意图的句式，BERT 没有对此类语法结构进行专门的建模，如果预训练的语料中只有“我要买苹果手机”，“我要买华为手机”，哪一天出现了一个新的手机牌子比如栗子手机，而这个手机牌子在预训练的语料当中并不存在，没有基于词法结构以及句法结构的建模，对于这种新出来的词是很难给出一个很好的向量表示的，而 ERNIE 通过对训练数据中的词法结构，语法结构，语义信息进行统一建模，极大地增强了通用语义表示能力，在多项任务中均取得了大幅度超越 BERT 的效果

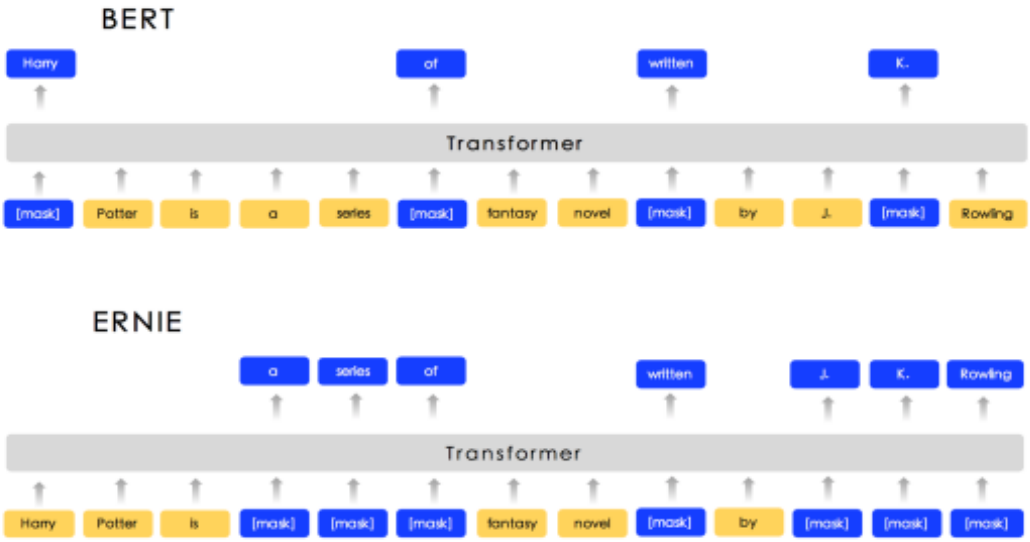
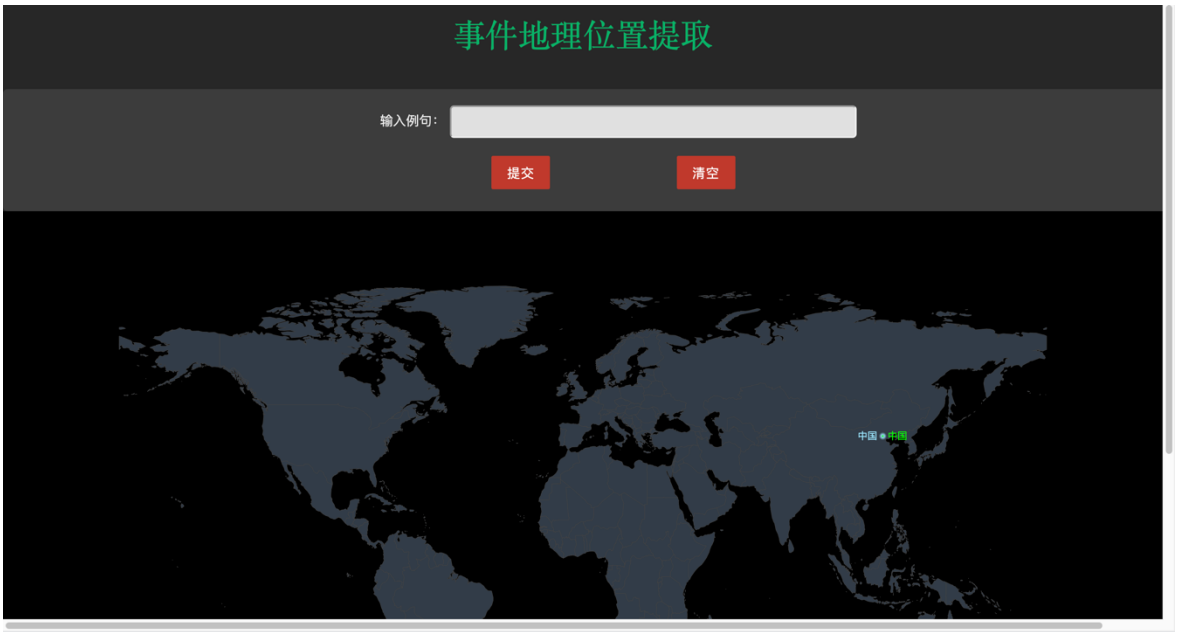


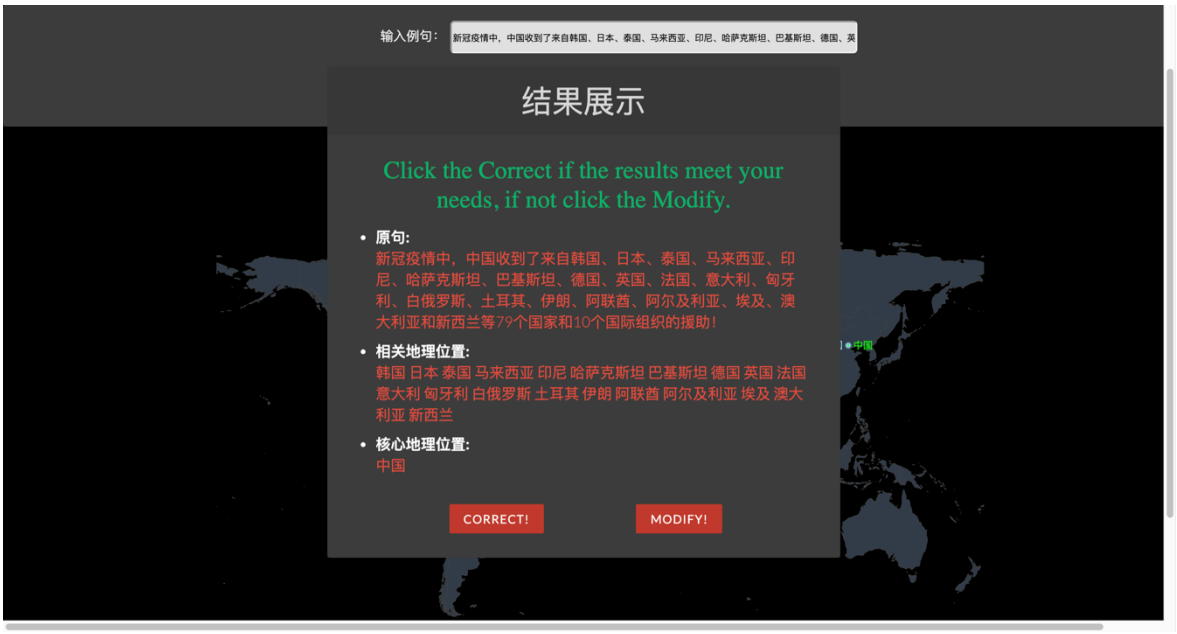
Figure 1: The different masking strategy between BERT and ERNIE

展示平台搭建

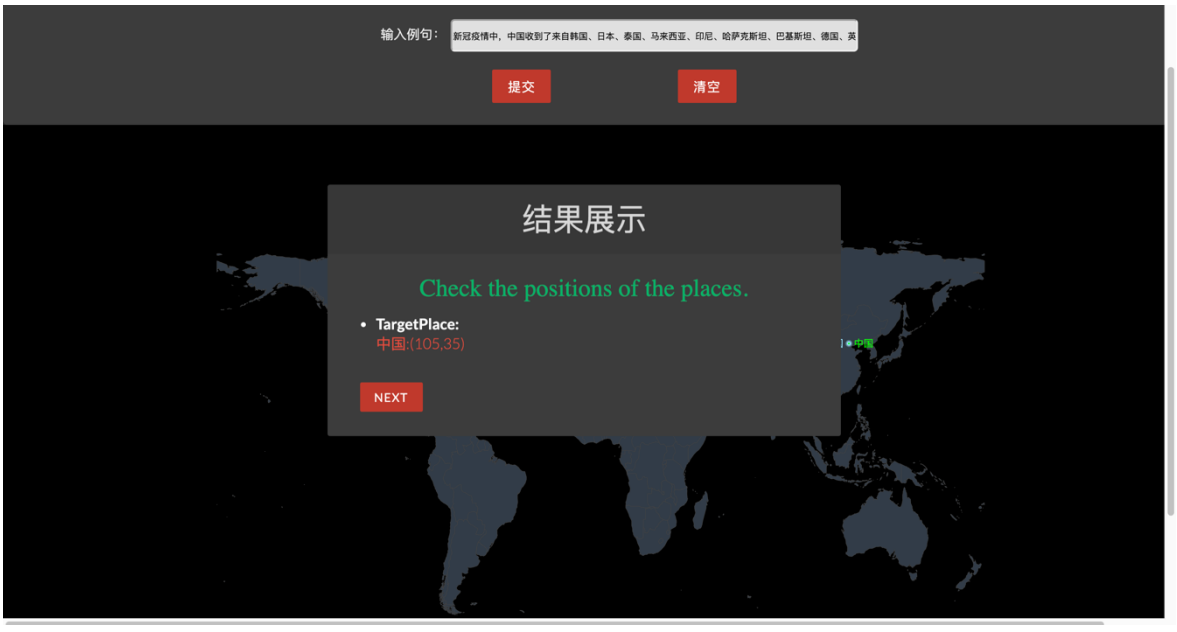
平台首页分为两个部分，如图所示，主要分为输入区和地图展示区。用户可以在输入区输入自己需要查询的新闻内容，点击提交后进入内容分析页面，输入区中的清空按钮可以帮助清空地图上的所有连线，为之后的实验做准备。地图区是主要的展示板块，相关新闻地点会在这里联系在一起，具体的下文中会有所阐述。

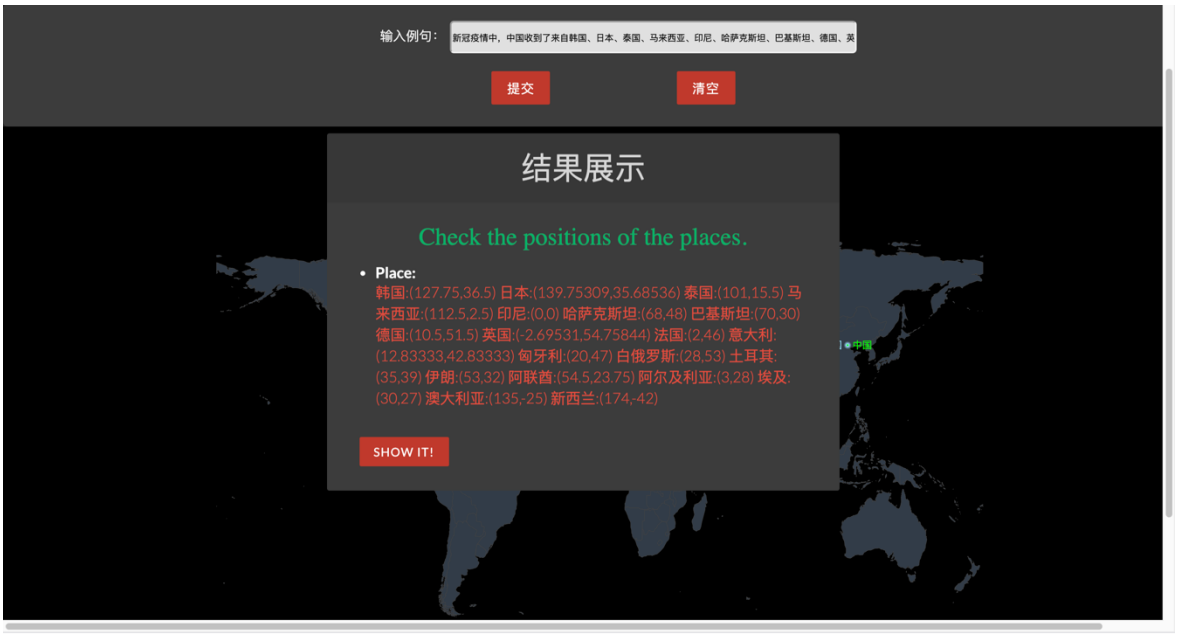


在输入相关例句后，平台经过机器学习模型的预测会将相关结果展示在页面上，如图所示，在展示平台中主要展示了预测结果，包括了例句，相关地理位置和核心地理位置，如果预测结果正确，点击下方的按钮 CORRECT 进入展示模式，如果预测错误点击 MODIFY 进入标注修改模式。



预测结果正确后，平台会计算相关地理位置的经纬度，图中展示其核心地理位置经纬度以及相关地理位置经纬度。





计算完成以后，地图上会展示核心地理位置和相关地理位置在地图上的位置，并将其连线，其中绿色表示核心地理位置，蓝色表示相关地理位置，如图所示

如果预测错误，那么平台会展示相关的标注结果，用户可以调整标注结果，相关地理位置，核心地理位置等信息帮助平台纠错，如图所示，修改完成后会进入展示界面展示正确结果，并会把相关的标注信息传递到后台作为新的数据集存储。

实验目前遇到的问题

实验目前遇到的最大的问题是数据集。数据集的大小，数据集本身的准确率对于最后的结果都有很大的影响。并且本实验中的核心地理位置的概念有很大的主观因素影响，不同的人，甚至同一个人在不同的时间点判断可能都会存在偏差。目前采用的方法是一个人（我）连续不间断的对于数据集进行标注，使用这种方法可以在一定程度上帮助减少这一部分的影响，但是还是不得不承认数据集对于模型最后的结果影响比较大，