# BERT和ERNIE谁更强？这里有一份4大场景的细致评测

量子位 2019-06-17

允中 发自 凹非寺
量子位 报道 | 公众号 QbitAI

BERT和ERNIE，NLP领域近来最受关注的2大模型究竟怎么样？

刚刚有人实测比拼了一下，结果在中文语言环境下，结果令人意外又惊喜。

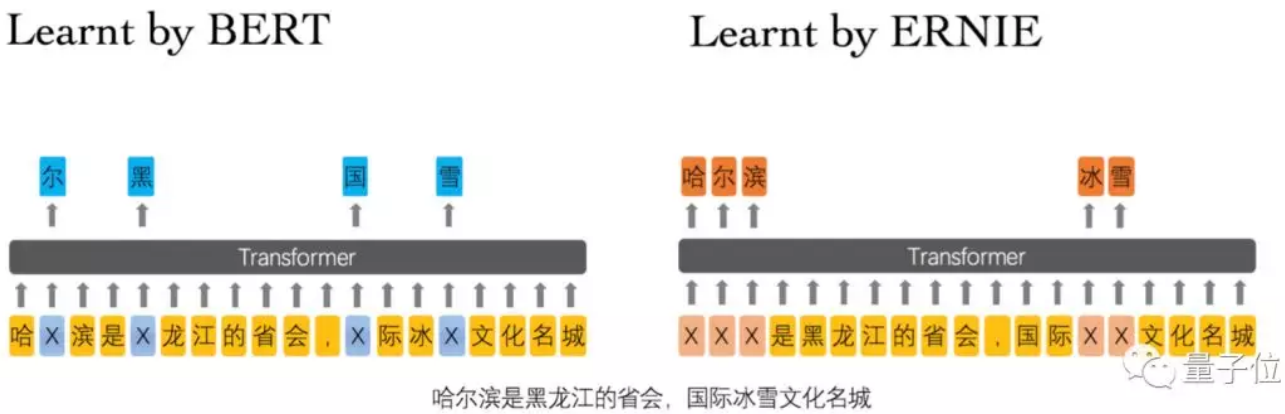
具体详情究竟如何？不妨一起围观下这篇技术评测。

## 1. 写在前面

随着2018年ELMo、BERT等模型的发布，NLP领域终于进入了“大力出奇迹”的时代。采用大规模语料上进行无监督预训练的深层模型，在下游任务数据上微调一下，即可达到很好的效果。曾经需要反复调参、精心设计结构的任务，现在只需简单地使用更大的预训练数据、更深层的模型便可解决。

随后在2019年上半年，百度的开源深度学习平台PaddlePaddle发布了知识增强的预训练模型ERNIE，ERNIE通过海量数据建模词、实体及实体关系。相较于BERT学习原始语言信号，ERNIE直接对先验语义知识单元进行建模，增强了模型语义表示能力。

简单来说，百度ERNIE采用的Masked Language Model是一种带有先验知识Mask机制。可以在下图中看到，如果采用BERT随机mask，则根据后缀“龙江”即可轻易预测出“黑”字。引入了词、实体mask之后，“黑龙江”作为一个整体被mask掉了，因此模型不得不从更长距离的依赖（“冰雪文化名城”）中学习相关性。




除此之外，百度ERNIE还引入了DLM（对话语言模型）任务，通过这种方式来学习相同回复对应的query之间的语义相似性。实验证明DLM的引入对LCQMC（文本相似度计算）系

列任务带来了较大的帮助。最终ERNIE采用多源训练数据，利用高性能分布式深度学习平台PaddlePaddle完成预训练。

## 2. 亲测

到底百度ERNIE模型所引入训练机制有没有起到作用，只有实践了以后才知道。为此，我亲自跑了BERT和ERNIE两个模型，在下面的几个场景中得到了预测结果。

### 2.1 完形填空

完形填空任务与预训练时ERNIE引入的知识先验Mask LM任务十分相似。从下图的比较中我们可以看到，ERNIE对实体词的建模更加清晰，对实体名词的预测比BERT更准确。例如BERT答案“周家人”融合了相似词语“周润发”和“家人”结果不够清晰；“市关村”不是一个已知实体；“菜菜”的词边界是不完整的。ERNIE的答案则能够准确命中空缺实体。

| 输入句子 | BERT 结果 | ERNIE 结果 | 答案 |
| --- | --- | --- | --- |
| ____对甄子丹饰演的孙悟空赞不绝口,称其为“宇宙最强美猴王”。 | 周家人 | 周润发 | 周润发 |
| 昨天，市人大代表、中关村管委会主任郭洪在参加市人大分组讨论会时透露，2014 年起，_____将牵头建立一条“京津冀大数据走廊”。 | 市关村 | 中关村 | 中关村 |
| 买菜的市民告诉记者，往年节前少有一元以及一元以下的菜价，今年一元左右的菜那么多，_____真的很便宜。 | 菜菜 | 价格 | 价格 |



### 2.2 NER (命名实体识别)

在同样为token粒度的NER任务中，知识先验Mask LM也带来了显著的效果。对比MSRA-NER数据集上的F1 score表现，ERNIE与BERT分别为93.8%、92.6%。在PaddleNLP的LAC数据集上，ERNIE也取得了更好的成绩，测试集F1为92.0%，比BERT的结果90.3%提升了1.7%。分析二者在MSRA-NER测试数据中二者的预测结果。可以观察到：

1.）ERNIE对实体理解更加准确：“汉白玉”不是实体类型分类错误；

2.）ERNIE对实体边界的建模更加清晰：“美国法律所”词边界不完整，而“北大”、“清华”分别是两个机构。

Case对比：摘自MSRA-NER数据测试集中的三段句子。B_LOC/I_LOC为地点实体的标签，B_ORG/L_ORG为机构实体的标签，O为无实体类别标签。下表分别展现了 ERNIE、BERT模型在每个字上的标注结果。

| 文本 | 我 | 随 | 一 | 群 | 人 | 登 | 上 | 汉 | 白 | 玉 | 台 | 阶 | … |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ERNIE | O | O | O | O | O | O | O | O | O | O | O | O | O |
| BERT | O | O | O | O | O | O | O | B_LOC | I_LOC | I_LOC | O | O | O |



| 文本 | … | 本 | 案 | 所 | 适 | 用 | 的 | 美 | 国 | 法 | 律 | 所 | 基 | 于 | 的 | … |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ERNIE | | O | O | O | O | O | O | B_LOC | I_LOC | O | O | O | O | O | | |
| BERT | | O | O | O | O | O | O | B_ORG | I_ORG | I_ORG | I_ORG | I_ORG | O | O | | |



| 文本 | … | 有 | 一 | 两 | 个 | 考 | 上 | 北 | 大 | 清 | 华 | 的 | … |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ERNIE | | O | O | O | O | O | O | B_ORG | I_ORG | B_ORG | I_ORG | O | |
| BERT | | O | O | O | O | O | O | B_ORG | I_ORG | I_ORG | I_ORG | O | |



## 2.3 相似度

ERNIE在训练中引入的DLM能有效地提升模型对文本相似度的建模能力。因此，我们比较文本相似度任务LCQMC数据集上二者的表现。从下表的预测结果可以看出，ERNIE学习到了中文复杂的语序变化。最终ERNIE与BERT在该任务数据的预测准确率为87.4%、87.0%.

| 输入 A | 输入 B | BERT 结果 | ERNIE 结果 | Label |
|---|---|---|---|---|
| 这叫什么高跟鞋 | 这种高跟鞋叫什么呀 | 不相似 | 相似 | 相似 |
| 搞笑的电影给推荐几部 | 推荐几部搞笑的电影… | 不相似 | 相似 | 相似 |
| 电炒锅什么牌子好 | 什么牌子的电炒锅好 | 不相似 | 相似 | 相似 |



## 2.4 分类

| 输入 | BERT 结果 | ERNIE 结果 | Label |
| --- | --- | --- | --- |
| 我已经不敢告诉别人林丹是我的偶像了 | 积极 | 消极 | 消极 |
| 不是很喜欢 | 积极 | 消极 | 消极 |
| 后备箱，大的物件啥也不能放，上次去岳父家装了点菜就满了。 | 积极 | 消极 | 消极 |
| 酒店装修有些旧，电梯都是最老式的。但房间应该重新装修过，比较大也比较干净。位置极好，离汽车总站很近，对面有新一佳也很方便。 | 消极 | 积极 | 积极 |



最后，比较应用最广泛的情感分类任务。经过预训练的ERNIE能够捕捉更加细微的语义区别，这些句子通常含有较委婉的表达方式。下面展示了PaddleNLP情感分类测试集上ERNIE与BERT的打分表现：在句式“不是很...”中含有转折关系，ERNIE能够很好理解这种关系，将结果预测为“消极”。在ChnSentiCorp情感分类测试集上finetune后ERNIE的预测准确率为95.4%，高于BERT的准确率（94.3%）。

从以上数据我们可以看到，ERNIE在大部分任务上都有不俗的表现。尤其是在序列标注、完形填空等词粒度任务上，ERNIE的表现尤为突出，一点都不输给Google的BERT。有兴趣的开发者可以一试：

https://github.com/PaddlePaddle/LARK/tree/develop/ERNIE

— 完 —