百度ERNIE，中文任务全面超越BERT

机器学习算法与Python学习 2019-05-24

本文转自机器之心

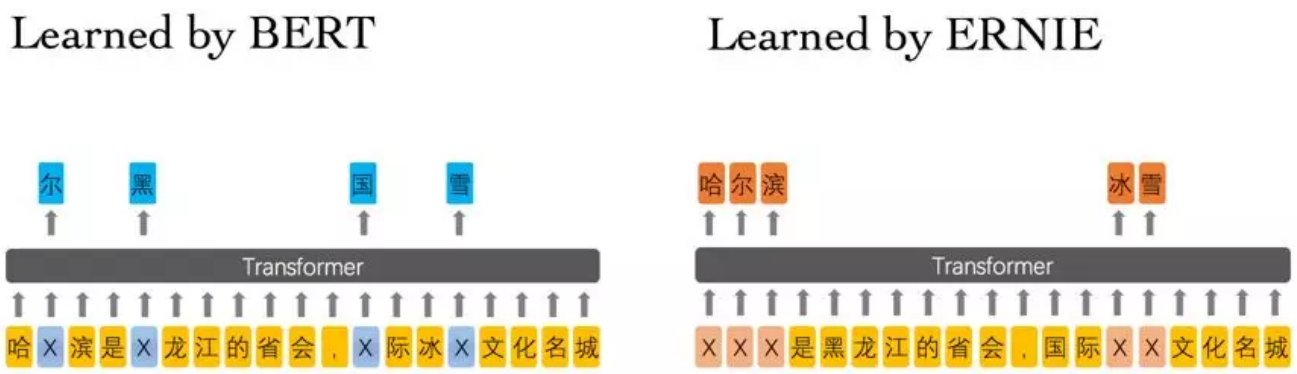
ERNIE Github 项目地址：https://github.com/PaddlePaddle/LARK/tree/develop/ERNIE

近年来，无监督文本的深度神经网络预训练模型大幅提升了各个 NLP 任务的效果。早期的工作聚焦于上下文无关的词向量建模，而之后提出的 Cove，ELMo，GPT 等模型，构建了语句级的语义表示。Google 近期提出的 BERT 模型，通过预测屏蔽的词，利用 Transformer 的多层 self-attention 双向建模能力，取得了更好的效果。

无论是稍早提出的 Cove、Elmo、GPT, 还是能力更强的 BERT 模型，其建模对象主要聚焦在原始语言信号上，较少利用语义知识单元建模。这个问题在中文方面尤为明显，例如，BERT 在处理中文语言时，通过预测汉字进行建模，模型很难学出更大语义单元的完整语义表示。例如，对于乒 [mask] 球，清明上 [mask] 图，[mask] 颜六色这些词，BERT 模型通过字的搭配，很容易推测出掩码的字信息，但没有显式地对语义概念单元 (如乒乓球、清明上河图) 以及其对应的语义关系进行建模。

设想如果能够让模型学习到海量文本中蕴含的潜在知识，势必会进一步提升各个 NLP 任务效果。因此百度提出了基于知识增强的 ERNIE 模型。ERNIE 模型通过建模海量数据中的实体概念等先验语义知识，学习真实世界的语义关系。具体来说，ERNIE 模型通过对词、实体等语义单元的掩码，使得模型学习完整概念的语义表示。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。

举个例子：



哈尔滨是黑龙江的省会，国际冰雪文化名城

- Learned by BERT ： 哈 [mask] 滨是 [mask] 龙江的省会，[mask] 际冰 [mask] 文化名城。
- Learned by ERNIE：[mask] [mask] [mask] 是黑龙江的省会，国际 [mask] [mask] 文化名城。

在 BERT 模型中，通过『哈』与『滨』的局部共现，即可判断出『尔』字，模型没有学习与『哈尔滨』相关的知识。而 ERNIE 通过学习词与实体的表达，使模型能够建模出『哈尔滨』与『黑龙江』的关系，学到『哈尔滨』是『黑龙江』的省会以及『哈尔滨』是个冰雪城市。

ERNIE 模型本身保持基于字特征输入建模，使得模型在应用时不需要依赖其他信息，具备更强的通用性和可扩展性。相对词特征输入模型，字特征可建模字的组合语义，例如建模红色，绿色，蓝色等表示颜色的词语时，通过相同字的语义组合学到词之间的语义关系。

此外，ERNIE 的训练语料引入了多源数据知识。除了百科类文章建模，还对新闻资讯类、论坛对话类数据进行学习，这里重点介绍下论坛对话建模。对于对话数据的学习是语义表示的重要途径，往往相同回复对应的 Query 语义相似。基于该假设，ERINE 采用 DLM（Dialogue Language Model）建模 Query-Response 对话结构，将对话 Pair 对作为输入，引入 Dialogue Embedding 标识对话的角色，利用 Dialogue Response Loss 学习对话的隐式关系，通过该方法建模进一步提升模型语义表示能力。

ERNIE 对实体概念知识的学习以及训练语料的扩展，增强了模型语义表示能力。为验证 ERNIE 的知识学习能力，研究者利用几道有趣的填空题对模型进行了考察。实验将段落中的实体知识去掉，让模型推理其答案。

	ERNIE预测	BERT预测	答案
2006年9月，___与张柏芝结婚，两人婚后育有两儿子——大儿子Lucas谢振轩，小儿子Quintus谢振南；2012年5月，二人离婚。	谢霆锋	谢振轩	谢霆锋
戊戌变法，又称百日维新，是___、梁启超等维新派人士通过光绪帝进行的一场资产阶级改良。	康有为	孙世昌	康有为
高血糖则是由于___分泌缺陷或其生物作用受损，或两者兼有引起。糖尿病时长期存在的高血糖，导致各种组织，特别是眼、肾、心脏、血管、神经的慢性损害、功能障碍。	胰岛素	糖糖内	胰岛素
___是中国神魔小说的经典之作，达到了古代长篇浪漫主义小说的巅峰，与《三国演义》《水浒传》《红楼梦》并称为中国古典四大名著。	西游记	《小》	西游记
相对论是关于时空和引力的理论，主要由___创立。	爱因斯坦	卡尔斯所	爱因斯坦
向日葵，因花序随___转动而得名。	太阳	日阳	太阳
___是太阳系八大行星中体积最大、自转最快的行星，从内向外的第五颗行星。它的质量为太阳的千分之一，是太阳系中其它七大行星质量总和的2.5倍。	木星	它星	木星
地球表面积5.1亿平方公里，其中71%为___，29%为陆地，在太空上看地球呈蓝色。	海洋	海空	海洋

可以看到 ERNIE 在基于上下文知识推理能力上表现的更加出色。

对于知识推理能力，ERNIE 在自然语言推断任务上做了进一步实验。XNLI 由 Facebook 和纽约大学的研究者联合构建，旨在评测模型多语言的句子理解能力。目标是判断两个句子的关系（矛盾、中立、蕴含）。ERNIE 与 Google 公布的 BERT 进行了比较：

	开发集准确率		测试集准确率	
	均值	方差	均值	方差
BERT	78.1%	0.0038	77.2%	0.0026
ERNIE	79.9%(+1.8%)	0.0041	78.4%(+1.2%)	0.0040

实验表明，ERNIE 模型相较于 BERT，在语言推断效果上更胜一筹。

多个公开的中文数据集上的进一步效果验证显示，相较 BERT，ERNIE 模型均取得了更好的效果：

1. 语义相似度任务 LCQMC

LCQMC 是哈尔滨工业大学在自然语言处理国际顶会 COLING2018 构建的问题语义匹配数据集，其目标是判断两个问题的语义是否相同。

	开发集准确率		测试集准确率	
	均值	方差	均值	方差
BERT	88.8%	0.0029	87.0%	0.0060
ERNIE	89.7%(**+0.9%**)	0.0021	87.4%(**+0.4%**)	0.0019

2. 情感分析任务 ChnSentiCorp

ChnSentiCorp 是中文情感分析数据集，其目标是判断一段话的情感态度。

	开发集准确率		测试集准确率	
	均值	方差	均值	方差
BERT	94.6%	0.0027	94.3%	0.0058
ERNIE	95.2%(**+0.6%**)	0.0012	95.4%(**+1.1%**)	0.0044

3. 命名实体识别任务 MSRA-NER

MSRA-NER 数据集由微软亚研院发布，其目标是命名实体识别，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名等。

	开发集F1		测试集F1	
	均值	方差	均值	方差
BERT	94.0%	0.0024	92.6%	0.0024
ERNIE	95.0%(+1.0%)	0.0027	93.8%(+1.2%)	0.0031

4. 检索式问答匹配任务 NLPCC-DBQA

NLPCC-DBQA 是由国际自然语言处理和中文计算会议 NLPCC 于 2016 年举办的评测任务，其目标是选择能够回答问题的答案。

	开发集MRR		测试集MRR	
	均值	方差	均值	方差
BERT	94.7%	0.0022	94.6%	0.0014
ERNIE	95.0%(+0.3%)	0.0011	95.1%(+0.5%)	0.0011
	开发集F1		测试集F1	
	均值	方差	均值	方差
BERT	80.7%	0.0158	80.8%	0.0158
ERNIE	82.3%(+1.6%)	0.0103	82.7%(+1.9%)	0.0136

研究团队表示，此次技术突破将被应用于多种产品和场景，进一步提升用户体验。未来百度将在基于知识融合的预训练模型上进一步深入研究。例如使用句法分析或利用其他任务的弱监督信号进行建模。此外，百度也会将该思路推广到其他语言，在其他语言上进一步验证。

百度自然语言处理（Natural Language Processing，NLP）以『理解语言，拥有智能，改变世界』为使命，研发自然语言处理核心技术，打造领先的技术平台和创新产品，服务全球用户，让复杂的世界更简单。

推荐阅读

Python高级技巧：lazy property

20 个超棒的数据科学 Python 库

华为突遭谷歌釜底抽薪！官方安卓不再支持华为手机