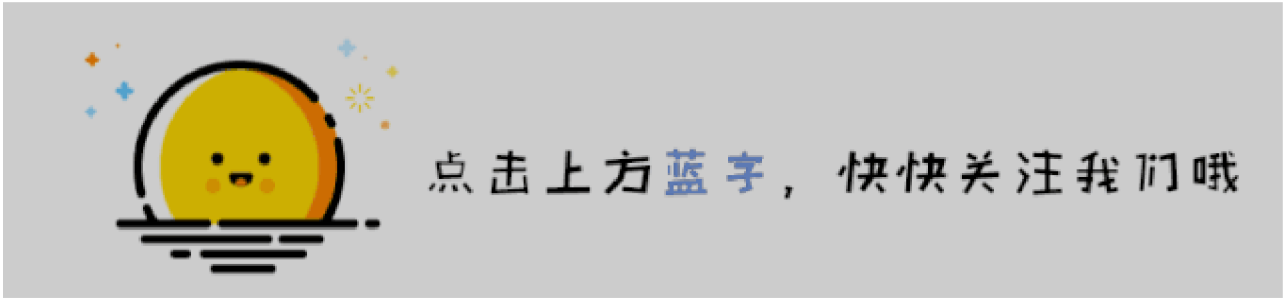
[预训练语言模型专题] 百度出品ERNIE合集，问国产预训练语言模型哪家强

原创 管扬 朴素人工智能

来自专辑

预训练语言模型

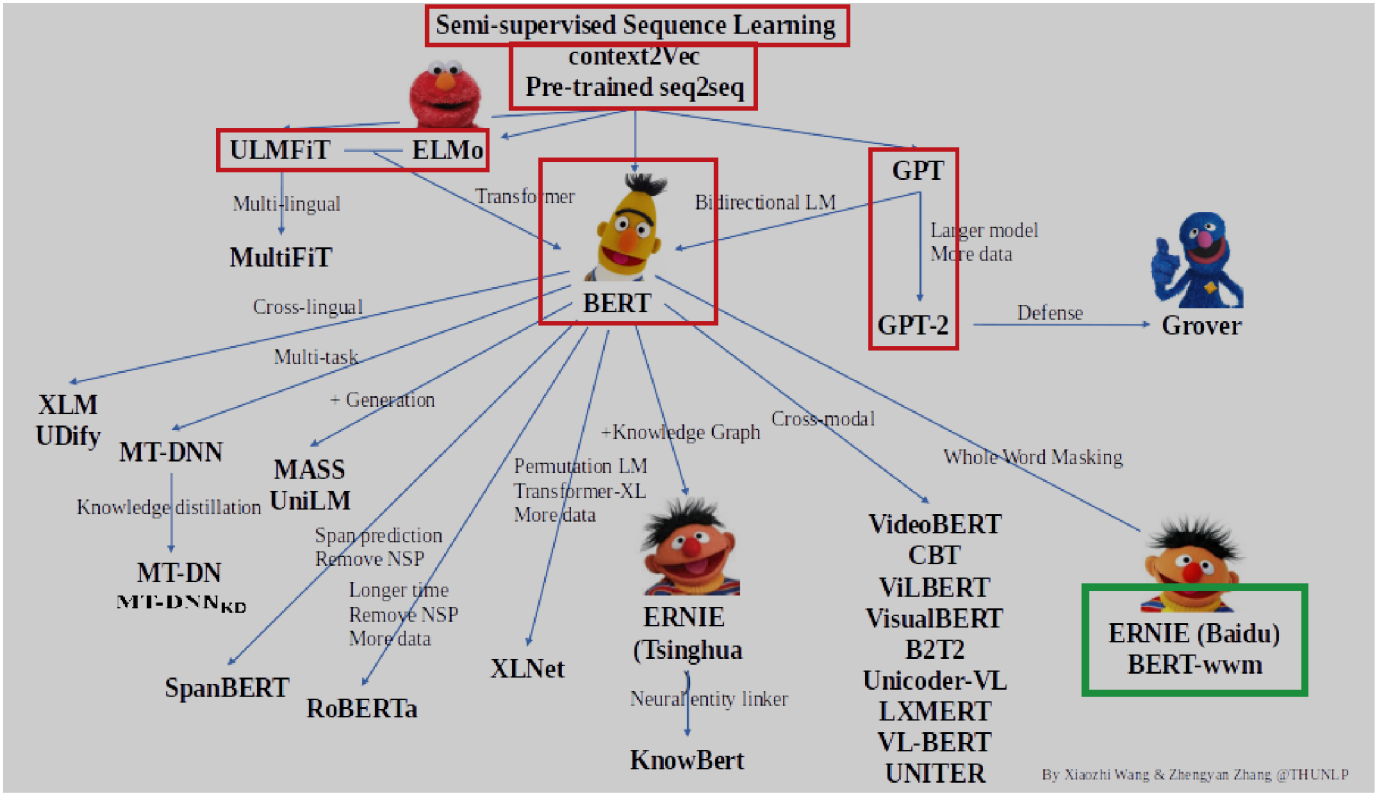


本文为预训练语言模型专题系列第七篇

系列传送门

[萌芽时代]、[风起云涌]、[文本分类通用技巧] 、 [GPT家族]、[BERT来临]、[BERT代码浅析]

感谢清华大学自然语言处理实验室对**预训练语言模型架构**的梳理，我们将沿此脉络前行，探索预训练语言模型的前沿技术，红框中为已介绍的文章，绿框中为本期介绍的文章，欢迎大家留言讨论交流。



1

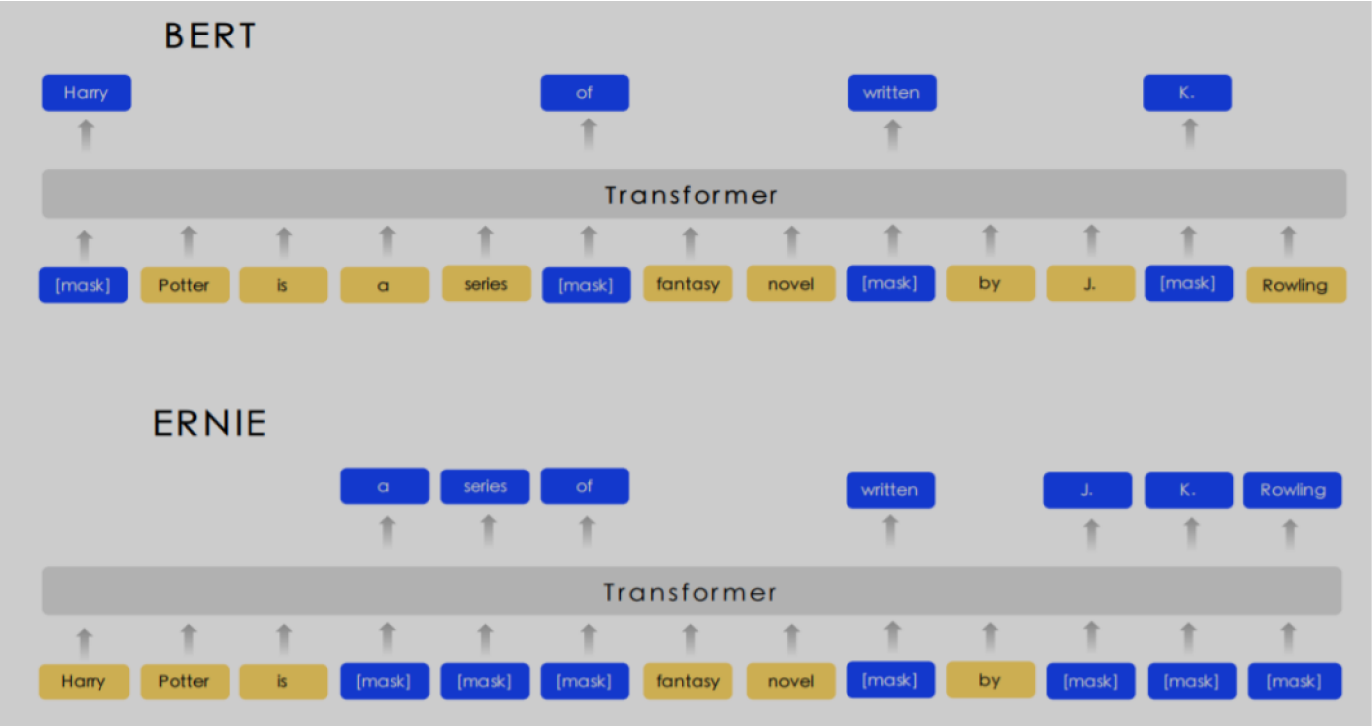
ERNIE: Enhanced Representation through Knowledge Integration (2019)

大家可能也注意到了，在上面的架构图中有**两个ERNIE**，一个是由清华大学发表在ACL上的ERNIE: Enhanced Language Representation with Informative Entities 我们以后可能会提及，另一个就是当前要介绍的这篇，由百度所出品，在本文中分享的ERNIE指 **ERNIE(Baidu)**。

有朋友也许会问，为什么它们都会起ERNIE这个名字呢？BERT和ERNIE都是著名的卡通芝麻街里的人物，而且两人是很好的朋友。甚至据小道消息（当然芝麻街出品官方是否认的），芝麻街作者对BERT和ERNIE的人物设定是一对gay友哦。咳咳，扯远了~ 不过，BERT和ERNIE的**亲密关系**是众所周知的，起这个名字，说明大家都想和BERT遥相呼应(gaoji)！

在文章摘要中，作者提出了ERNIE是通过加强**BERT的masking策略**来获取知识，包括引入实体级别的masking和词级别的masking，实验证明，ERNIE在五项中文自然语言处理任务上继BERT之后又刷新了榜单，并且在完形填空任务上有出色的表现。

ERNIE和BERT的联系确实是极为密切的，基于BERT，ERNIE提出两种加强的masking方式，分别是**entity-level masking** 和 **phrase-level masking**。由entity或phrase作为单位来做masking，一般会包含多个字或词。作者认为这样在训练的过程中，可以让模型学习到与phrase和entity相关的知识，包括实体间的关系，实体的属性，实体的类别等，帮助模型更好地泛化。



上图显示了BERT和ERNIE的masking的差异。简单来说，原来在一句句子中随机的15%token被mask，那么一个连续的词，比如“哈尔滨”，很可能其中某个字被mask掉，其他没有。而这个词实际上包含着实体连续的含义，如果只mask掉一部分，那么一来会导致这个mlm的任务过于简单，二来没能更好学到实体的知识。

为了让模型渐渐地学到更高层的知识，文章提出通过三个阶段的预训练来学习。

1. 第一个阶段是**基本的masking**，与BERT类似，用语言最基本的单元来mask。在英文层面用词，在中文层面用字来mask，这样可以获得词的基本表达，但尚未对语义知识进行建模。
2. 第二个阶段是**短语的masking**，相邻的概念层面上一组词或字符。在中文中，一般会用分词工具在将相邻的字分成词语，这样可以对稍大的语义模块来进行建模。
3. 第三个阶段是**实体的masking**，包括人名，地点，组织，产品等。这些实体会倾向于包含句子中的重要信息。通过这三个阶段的训练，向量表示就会包含更丰富的语义信息。

pre-train dataset size	mask strategy	dev Accuracy	test Accuracy
10% of all	word-level(chinese character)	77.7%	76.8%
10% of all	word-level&phrase-level	78.3%	77.3%
10% of all	word-level&phrase-leve&entity-level	78.7%	77.6%
all	word-level&phrase-level&entity-level	79.9 %	78.4%

在上图XNLI的结果，可以看出这三阶段的训练是有效的，同时下图显示，ERNIE基于BERT进行的改进可以稳定地提升BERT模型的效果。

Task	Metrics	Bert		ERNIE	
		dev	test	dev	test
XNLI	accuracy	78.1	77.2	79.9 (+1.8)	78.4 (+1.2)
LCQMC	accuracy	88.8	87.0	89.7 (+0.9)	87.4 (+0.4)
MSRA-NER	F1	94.0	92.6	95.0 (+1.0)	93.8 (+1.2)
ChnSentiCorp	accuracy	94.6	94.3	95.2 (+0.6)	95.4 (+1.1)
nlpcc-dbqa	mrr	94.7	94.6	95.0 (+0.3)	95.1 (+0.5)
	F1	80.7	80.8	82.3 (+1.6)	82.7 (+1.9)

编者认为，衡量一种算法的改进是否好，一方面是观察它的效果是否**有提升并且稳定**，从ERNIE的结果上来说是不错的。另一方面是这种改进的**代价是否比较小**，这一点ERNIE的思想也具有吸引力，因为改变的是预训练的mask方法，而且改动不大，对输入的语料没有影响，对后续finetune也没有影响。所以，BERT后来也结合这种改进的思路，推出了BERT-wwm，即**Whole Word Masking** 模型，在mask时对整词同时进行mask，渐渐成为了大家常用的模型之一。

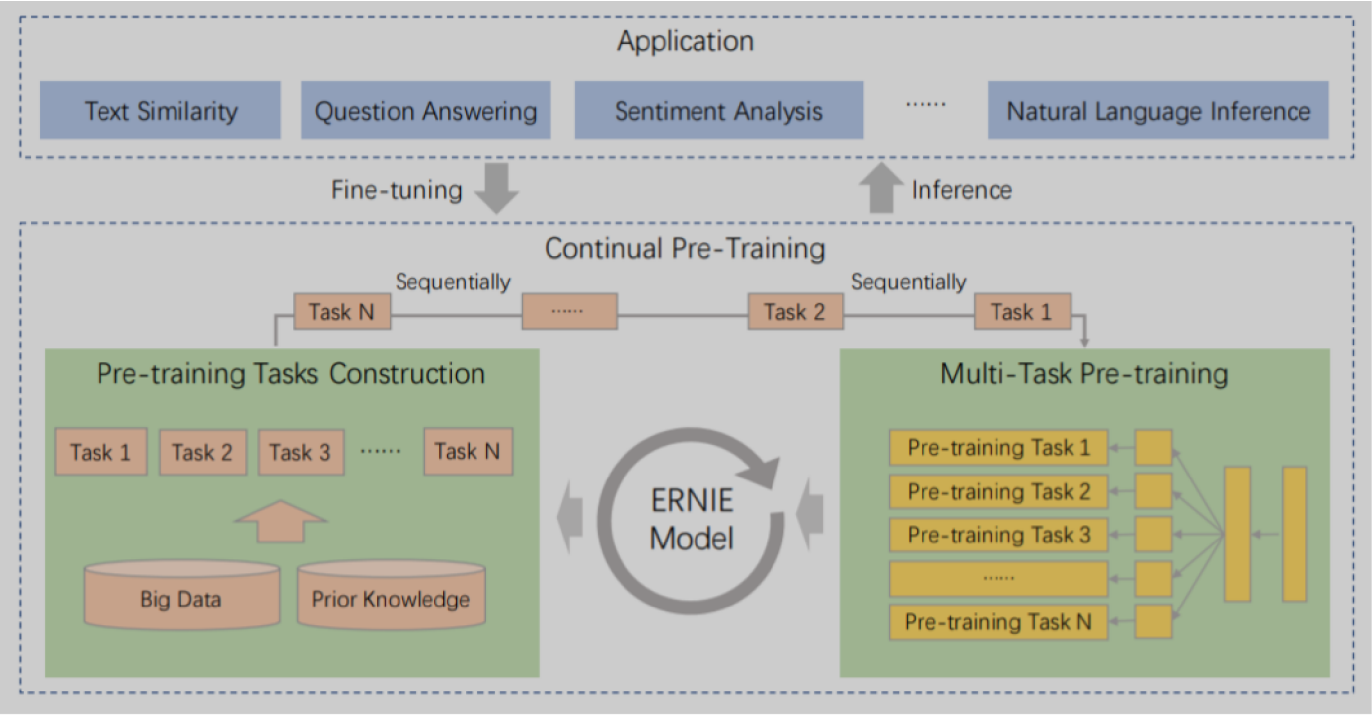
2

ERNIE 2.0: A CONTINUAL PRE-TRAINING FRAMEWORK FOR LANGUAGE UNDERSTANDING (2019)

这篇ERNIE 2.0也是百度出品，没有在架构图中，但有一定的关联，我们就一起介绍一下。这篇文章出了以后，大家习惯叫之前的那篇为ENRIE 1.0。

文中提出了一种预训练学习框架，可以通过多任务构造和学习的方法，提取训练语料中的多种语义信息，实验表明，ERNIE2.0在16个任务上超过了BERT和XLNet。

文章指出，当前的预训练语言模型大多是基于词语和句子之间的**共现关系**来进行建模训练的，然而实际上训练语料中包含很多其他的，包括词汇，句法，语义的有价值的信息值得去关注。比如命名实体识别会包含实体概念的知识；句子的顺序和距离，会包含文本结构的信息；段落级别的相似性和语句间论述关系，会包含语言逻辑信息等等。为了挖掘出训练语料中的丰富信息，作者们设计了一种预训练框架ERNIE 2.0来对多任务进行**连续增量**的训练。



ERNIE 2.0脱胎于传统的预训练+Finetune框架，但又有所区别。相比于传统做法，它不是在少量预训练任务上完成的，而是通过不断引入新的预训练任务，帮助模型持续地对知识进行增量学习。

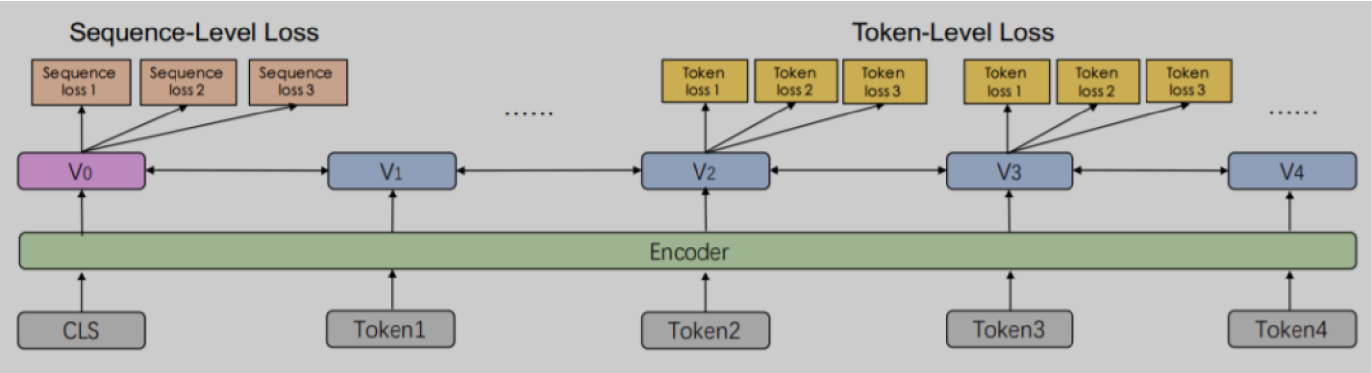
持续学习的过程分为两个阶段，首先，利用先验知识和大量数据，构建无监督预训练的任务；接着，通过多任务训练增量地更新ERNIE的模型。

- 预训练模型任务的构建，包括词级别的任务，结构级别的任务，语义级别的任务。所有这些的任务都是不需要依赖大量人类标注的自监督或弱监督数据。
- 多任务训练时，ERNIE会通过一种增量学习的范式来训练所有的这些任务。先用一个简单的任务来预训练模型，然后一个一个添加新的任务。每添加一个任务都会同时训练新任务和原来任务，以确保从之前任务学到的知识不被遗忘。通过这样使知识增量积累，以达成在新任务上的更好效果。

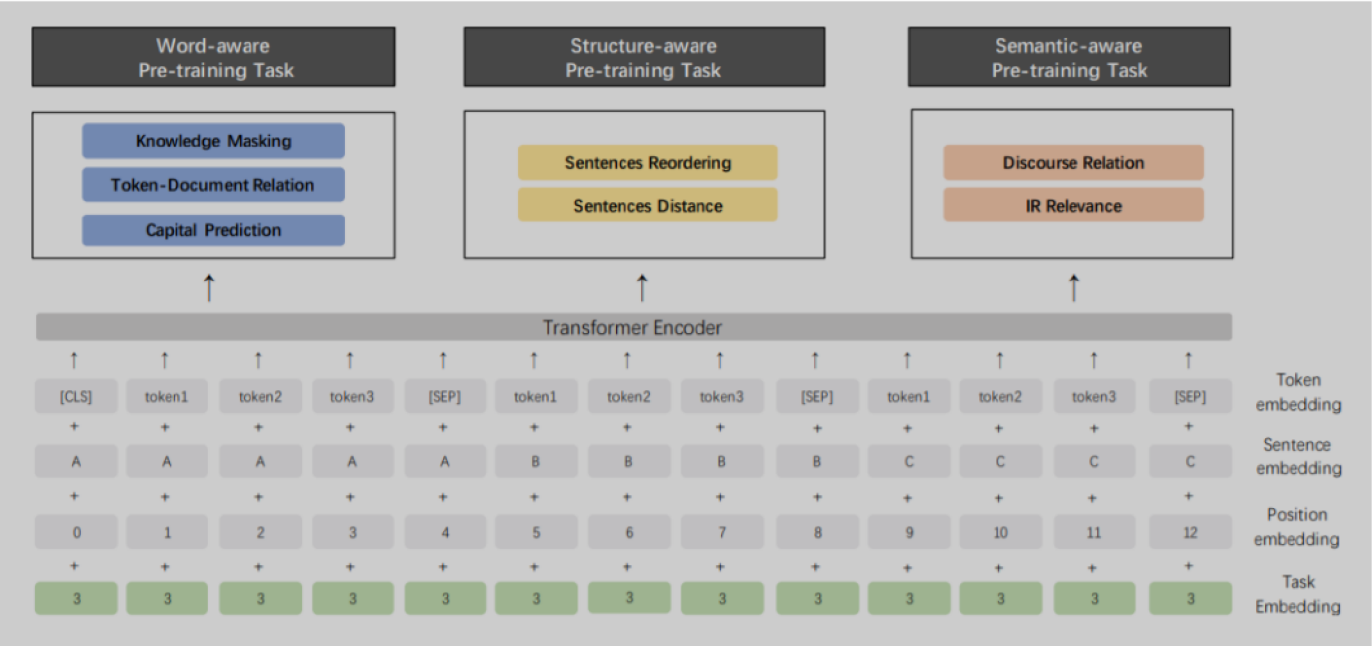
总结一下ERNIE的预训练方式。首先，任务会**串行地构造**不同的预训练任务，并一个个连续地加入训练中，这样可以保持任务训练的灵活性；其次，当前已加入训练的任务会进行**并行的多任务训练**，这样使模型不致于忘记之前训练过的任务，比如这样：

(task1)->(task1,task2)->(task1,task2,task3)->...->(task1，task2,...,taskN)

跟BERT一样，对于句级别的loss，模型一般会通过[CLS] token的表达来学习，而每个token的loss，则会通过对应的token的表示去学习，多个任务用不同的loss来学习。



Transformer层也和BERT的一样，有一点不同的是，因为模型会对多任务进行持续增量的学习，所以会在Embedding中引入**Task Embedding**，以对不同的任务提供特定特征。



最后，我们来看看在不同任务上的结果对比。下图中可以看到，ERNIE 2.0 在大量任务中力压当时 state-of-art的模型。

Task(Metrics)	*BASE model*		*LARGE model*				
	Test		Dev			Test	
	BERT	ERNIE 2.0	BERT	XLNet	ERNIE 2.0	BERT	ERNIE 2.0
CoLA (Matthew Corr.)	52.1	**55.2**	60.6	63.6	**65.4**	60.5	**63.5**
SST-2 (Accuracy)	93.5	**95.0**	93.2	95.6	**96.0**	94.9	**95.6**
MRPC (Accurary/F1)	84.8/88.9	**86.1/89.9**	88.0/-	89.2/-	**89.7/-**	85.4/89.3	**87.4/90.2**
STS-B (Pearson Corr./Spearman Corr.)	87.1/85.8	**87.6/86.5**	90.0/-	91.8/-	**92.3/-**	87.6/86.5	**91.2/90.6**
QQP (Accuracy/F1)	89.2/71.2	**89.8/73.2**	91.3/-	91.8/-	**92.5/-**	89.3/72.1	**90.1/73.8**
MNLI-m/mm (Accuracy)	84.6/83.4	**86.1/85.5**	86.6/-	**89.8/-**	89.1/-	86.7/85.9	**88.7/88.8**
QNLI (Accuracy)	90.5	**92.9**	92.3	93.9	**94.3**	92.7	**94.6**
RTE (Accuracy)	66.4	**74.8**	70.4	83.8	**85.2**	70.1	**80.2**
WNLI (Accuracy)	**65.1**	**65.1**	-	-	-	65.1	**67.8**
AX(Matthew Corr.)	34.2	**37.4**	-	-	-	39.6	**48.0**
Score	78.3	**80.6**	-	-	-	80.5	**83.6**

在中文的任务上，ERNIE 2.0 也是表现出色，全线飘红。

Task	Metrics	BERT_{BASE}		ERNIE 1.0_{BASE}		ERNIE 2.0_{BASE}		ERNIE 2.0_{LARGE}	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
CMRC 2018	EM/F1	66.3/85.9	-	65.1/85.1	-	69.1/88.6	-	**71.5/89.9**	-
DRCD	EM/F1	85.7/91.6	84.9/90.9	84.6/90.9	84.0/90.5	88.5/93.8	88.0/93.4	**89.7/94.7**	**89.0/94.2**
DuReader	EM/F1	59.5/73.1	-	57.9/72.1	-	61.3/74.9	-	**64.2/77.3**	-
MSRA-NER	F1	94.0	92.6	95.0	93.8	95.2	93.8	**96.3**	**95.0**
XNLI	Accuracy	78.1	77.2	79.9	78.4	81.2	79.7	**82.6**	**81.0**
ChnSentiCorp	Accuracy	94.6	94.3	95.2	95.4	95.7	95.5	**96.1**	**95.8**
LCQMC	Accuracy	88.8	87.0	89.7	87.4	**90.9**	**87.9**	**90.9**	**87.9**
BQ Corpus	Accuracy	85.9	84.8	86.1	84.8	86.4	85.0	**86.5**	**85.2**
NLPCC-DBQA	MRR/F1	94.7/80.7	94.6/80.8	95.0/82.3	95.1/82.7	95.7/84.7	95.7/85.3	**95.9/85.3**	**95.8/85.8**

总结一下，ERNIE 2.0 是一个连续增量学习的框架，通过增量构造并加入预训练语言模型任务，连续并行地进行多任务学习，取得了较好的效果。框架提供的思路和对任务的分类都很有启发。但个人感觉美中不足的是，这篇文章缺少abalation study，而且预训练的过程又相对繁琐，模型使用了7个以上的预训练任务，但是读者无法了解，哪些预训练任务是相对重要或者有用的。总之，这篇文章还是很有意思的，值得读者们借鉴。

未完待续

本期的论文就给大家分享到这里，感谢大家的阅读和支持，下期我们会给大家带来其他预训练语言模型的介绍，敬请大家期待！

欢迎关注朴素人工智能，这里有很多最新最热的论文阅读分享，有问题或建议可以在公众号下留言。

往期推荐

- 性能媲美BERT却只有其1/10参数量？| 近期最火模型ELECTRA解析
- 微软统一预训练语言模型UniLM 2.0解读
- 叫我如何相信你？聊一聊语言模型的校准
- 十分钟了解文本分类通用技巧