[预训练语言模型专题]跨语种语言模型

原创 远皓 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的第13篇。

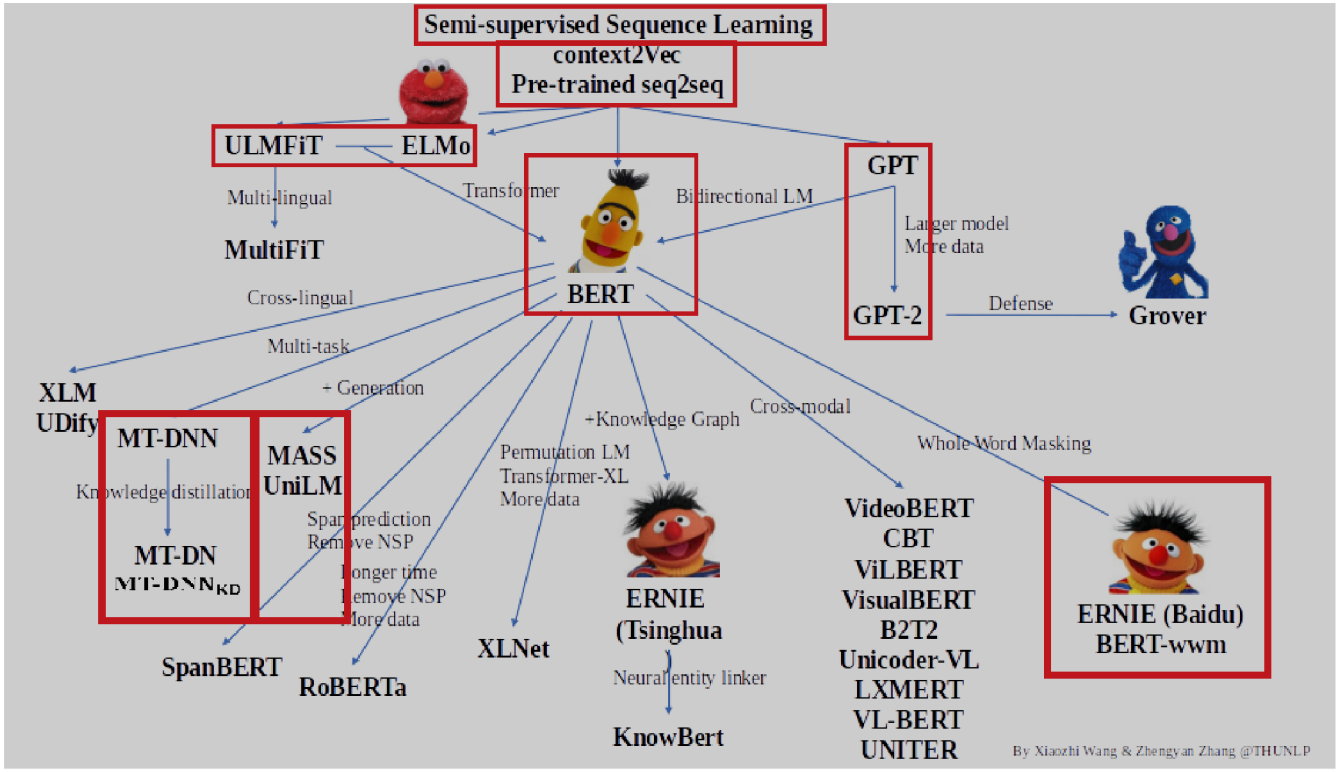
快速传送门

1-4:[萌芽时代]、[风起云涌]、[文本分类通用技巧] 、 [GPT家族]

5-8:[BERT来临]、[浅析BERT代码]、[ERNIE合集]、[MT-DNN(KD)]

9-12:[Transformer]、[Transformer-XL]、[UniLM]、[Mass-Bart]

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，今天介绍一个可能大家比较少接触的分支——跨语种语言模型（cross-lingual language model）。



在著名的科幻电影《银河系漫游指南》里有一种叫巴别鱼[1]的神奇生物。将它塞进耳朵里你就能听懂任何语言。多语种语言模型做得事情和巴别鱼很像，人们希望这个模型能用来处理所有的语言。举个例子，大家常用的中文bert有很强的中文处理能力以及一定的英文处理能力，但基本也就只能处理这两种语言；而目前的SOTA跨语种模型XLM-RoBERTa能够处理104种语言。

> 巴别鱼，体型很小，黄色，外形像水蛭，很可能是宇宙中最奇异的事物。它靠接收脑电波的能量为生，并且不是从其携带者身上接收，而是从周围的人身上。它从这些脑电波能量中吸收所有未被人察觉的精神频率，转化成营养。然后它向携带者的思想中排泄一种由被察觉到的精神频率和大脑语言中枢提供的神经信号混合而成的心灵感应矩阵。所有这些过程的实际效果就是，如果你把一条巴别鱼塞进耳朵，你就能立刻理解以任何形式的语言对你说的任何事情。

巴别鱼剧照

数据集

训练跨语种语言模型会用到两种语料。一种是单语种（monolingual）语料，另一种是平行（parallel）语料。所谓平行语料就是源语言与译文“对齐”的语料。所谓对齐也有好几种级别，最常见的是句子级对齐，也有按词进行对齐的文本。可想而知，平行语料的获取相比于单语种语料要困难许多。如何充分借助单语种语料来提升模型能力是XLM研究的一个重点。

跨语种语言模型的评价一般有两个大方向，一个是其语义理解能力，另一个是文本生成能力。语义理解能力通常借助**XNLI**[2]数据集，它提供了15种语言的平行文本，每种语言7500对的NLI语料。文本生成通常用翻译任务来评估，感兴趣的朋友可以自己查阅相关资料。

模型

下表列出了常见的单语种和多语种预训练语言模型。接下来我们将分析其中的mBERT、XLM和XLM-R三个模型。

Model	#lgs	tokenization	L	H_m	H_{ff}	A	V	#params
$\text{BERT}_{\text{Base}}$	1	WordPiece	12	768	3072	12	30k	110M
$\text{BERT}_{\text{Large}}$	1	WordPiece	24	1024	4096	16	30k	335M
mBERT	104	WordPiece	12	768	3072	12	110k	172M
$\text{RoBERTa}_{\text{Base}}$	1	bBPE	12	768	3072	8	50k	125M
RoBERTa	1	bBPE	24	1024	4096	16	50k	355M
XLM-15	15	BPE	12	1024	4096	8	95k	250M
XLM-17	17	BPE	16	1280	5120	16	200k	570M
XLM-100	100	BPE	16	1280	5120	16	200k	570M
Unicoder	15	BPE	12	1024	4096	8	95k	250M
XLM-R Base	100	SPM	12	768	3072	12	250k	270M
XLM-R	100	SPM	24	1024	4096	16	250k	550M
GPT2	1	bBPE	48	1600	6400	32	50k	1.5B
wide-mmNMT	103	SPM	12	2048	16384	32	64k	3B
deep-mmNMT	103	SPM	24	1024	16384	32	64k	3B
T5-3B	1	WordPiece	24	1024	16384	32	32k	3B
T5-11B	1	WordPiece	24	1024	65536	32	32k	11B

常见的预训练语言模型

Multilingual Bert（mBERT）

模型来自于这论文**《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》**[3]，你没有看错，就是发表于2018年大名鼎鼎的BERT论文。

2018年11谷歌就放出了支持104种语言的多语种版本预训练模型，规格是BERT base。这个模型的较新版本是uncased版，即没有对输入文本进行规范化。使用WordPiece算法进行tokenization，词典大小是110k。其他的训练方法和普通的BERT一样，采用的是MLM和NSP两个loss，语料是Wikipedia。

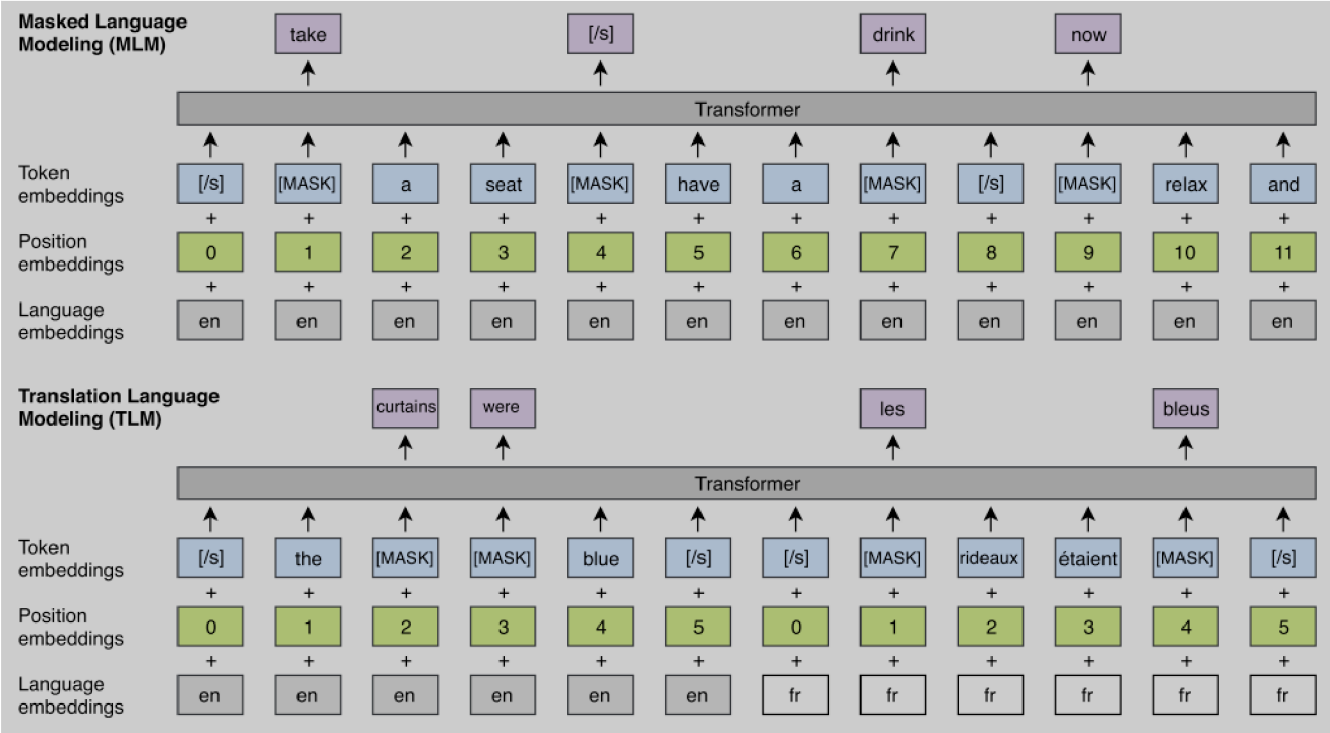
XLM

模型来自于论文**《Cross-lingual lan- guage model pretraining》**[4]，来自于FAIR，发表在NIPS2019。

XLM使用BPE算法进行tokenization，并且词典大小比mBERT更大，达到200k。论文指出Shared sub-word vocabulary对模型性能有很大的影响，在训练BPE算法的过程中他们使用了特殊的采样方式来避免低资源语种被进行字符集切分。

模型训练使用了三种不同的目标函数，在单语种语料上使用非监督的CLM和MLM。MLM就是masked language modeling，大家比较熟悉，在此就不再赘述了。CLM全称是Causal Language Modeling，简单的说就是用前面的词预测当前词，更详细的介绍大家可以参考我们之前UniLM

和GPT的文章。在平行语料上使用的目标称为Translation Language Modeling (TLM)。其训练方式如下图所示，是将平行句子拼接后随机mask，希望让模型能借助另一种语言的信息来还原出被遮蔽的词。从图中可以看出模型用language embedding替换了BERT里的type embedding，并且在做TLM任务时position embedding在两个语言间是对应的。



XLM训练示意图，增加了language embedding，TLM时position embedding在两种语言间是对应的

我们来看一下XLM在XNLI上的表现。这张表很有意思，首先对这个数据集有3种处理方式：translate-train，translate-test和直接测试，即zeroshot。第一种是把英语的MNLI数据集机器翻译成XNLI内的15种语言用于训练，在XNLI测试集上测试；第二种是把XNLI测试集的15种语言翻译成英文。本文的对照组就是上面的mBERT。

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. [14]	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. [14]	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. [12]	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. [14]	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk [4]	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

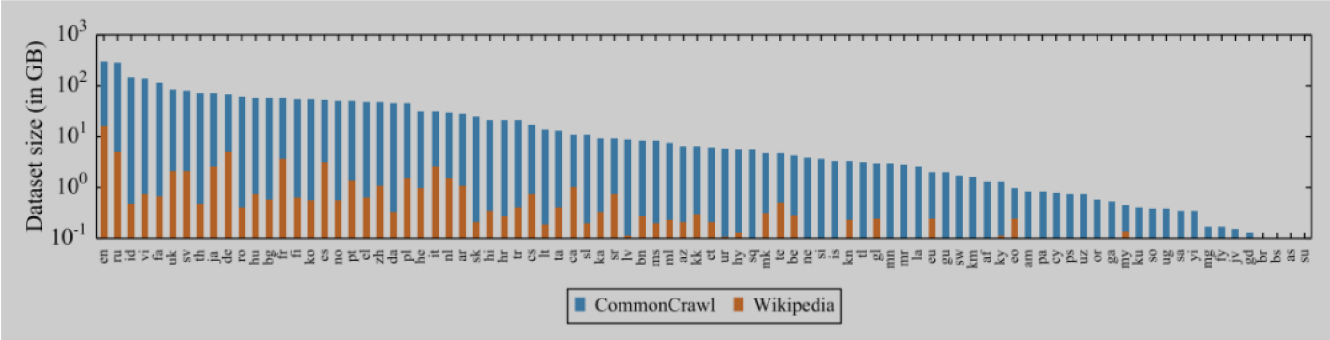
XLM在XNLI数据集上的结果

可以看到效果最好的是翻译训练集，平均精度达到了76.7%，zero-shot次之，最差的是翻译测试集。在相同的实验设定下XLM稳定优于mBERT，甚至在zero-shot下的XLM也比finetune过的mBERT强。另外MLM+TLM也稳定优于只用MLM的方式。

XLM-RoBERTa

模型来自于论文《Unsupervised Cross-lingual Representation Learning at Scale》[5]，和上文一样来自FAIR，已经被ACL 2020接收。

XLM-R使用了比XLM更大的词典，达到了250k。它也没有辜负RoBERTa的血统，使用了比Wikipedia大得多的cc100数据集。XLM-R只使用单语种语料，训练目标也只有MLM一个。



XLM-R使用的CC100数据与Wikipedia数量对比图

Tokenizer换成了sentence piece算法，在构建时也进行了采样，并且调整了系数使得各语言更加平衡。模型层面去掉了language embedding，变得更加简洁。我感觉用“重剑无锋”来形容XLM-R再合适不过了。

这篇论文总结了几个影响多语种模型的重要因素，可能会对大家有所启发：

- 当处理的语种变多的时候模型的能力会下降（嗯，符合常识）。增大模型可以一定程度对抗这种效应。
- 模型能力主要受词典大小、训练集大小、语种的采样频率影响
- 增大词典规模可以提高模型性能
- sentence piece可以提高模型的通用性

下面这张图可以让大家对这些结论有更直观的印象

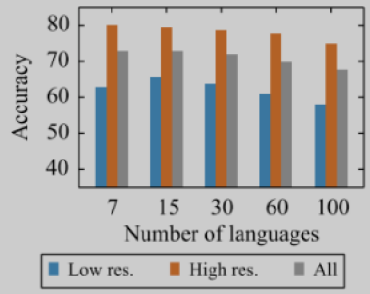


Figure 2: The transfer-interference trade-off: Low-resource languages benefit from scaling to more languages, until dilution (interference) kicks in and degrades overall performance.

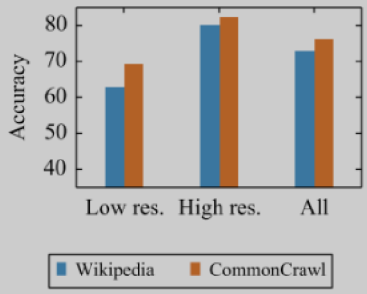


Figure 3: Wikipedia versus CommonCrawl: An XLM-7 obtains significantly better performance when trained on CC, in particular on low-resource languages.

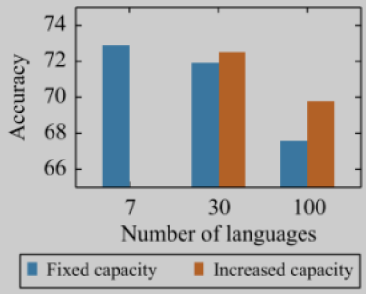


Figure 4: Adding more capacity to the model alleviates the curse of multilinguality, but remains an issue for models of moderate size.

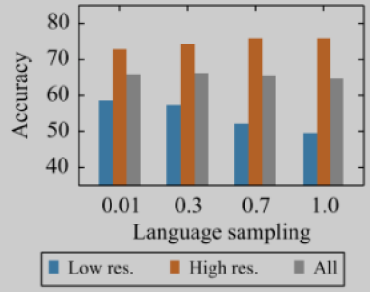


Figure 5: On the high-resource versus low-resource trade-off: impact of batch language sampling for XLM-100.

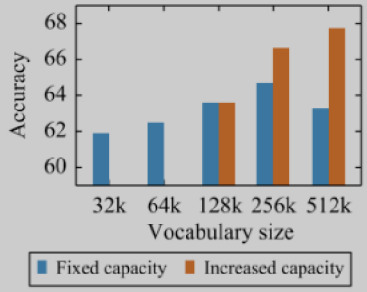


Figure 6: On the impact of vocabulary size at fixed capacity and with increasing capacity for XLM-100.

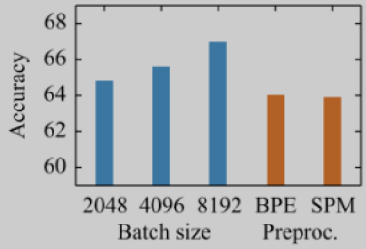


Figure 7: On the impact of large-scale training, and preprocessing simplification from BPE with tokenization to SPM on raw text data.

不同因素对XLM的影响

最后来看一下XLM-RoBERTa的实力。下表是在XNLI数据集上的结果对比，设定和XLM论文中差不多，其中Devlin et al.指的是mBERT，Lample and Conneau指的是XLM。可以看出XLM-R相比于XLM又前进了一大步。

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{Base}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{Base}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

XNLI结果对比

顺便再提一嘴，论文作者还在GLUE数据集上对比了XLM-R和XLNET、RoBERTa等单语种语言模型，XLM-R超过了BERT-large，略低于XLNET和RoBERTa。也就是说XLM-R不仅获得了多语种能

力，而且没有牺牲英文上的水平。

总结一下，从2018年的mBERT到2020年的XLM-R，跨语种预训练语言模型获得了长足的发展，地球语言范围内的巴别鱼指日可待。最近在Kaggle上正在进行一场跨语种文本分类的比赛，如果有想体验XLM最新进展的朋友可以去试试身手。

进行中的跨语言文本分类比赛

今天的文章就到这里，下期再见👋

参考资料

[1] 巴别鱼: *https://baike.baidu.com/item/%E5%B7%B4%E5%88%AB%E9%B1%BC*

[2] XNLI: Evaluating Cross-lingual Sentence Representations: *https://arxiv.org/abs/1809.05053*

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: *https://arxiv.org/abs/1810.04805*

[4] Cross-lingual lan- guage model pretraining: *https://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf*

[5] Unsupervised Cross-lingual Representation Learning at Scale: *http://arxiv.org/abs/1911.02116*

历史推荐

1.convlab2中强化学习方法之对话策略学习浅析
2.[预训练语言模型专题] Transformer-XL 超长上下文注意力模型
3.Transformer多轮对话改写实践
4.[预训练语言模型专题] BART & MASS 自然语言生成任务上的进步
5.[预训练语言模型专题] 结合HuggingFace代码浅析Transformer
6.表格问答完结篇：落地应用