

文章编号: 1003-0077(2024)01-0057-08

探索中文预训练模型的混合粒度编码和 IDF 遮蔽

邵云帆, 孙天祥, 邱锡鹏

(复旦大学 计算机科学技术学院, 上海 200433)

摘要: 目前大多数中文预训练语言模型采用字级别编码, 因为字符级编码序列长而产生大量计算开销。词级别编码尽管能够缓解这一问题, 但也会带来其他问题, 如词典外词、数据稀疏等。针对中文不同粒度的编码, 该文提出使用混合粒度编码的中文预训练模型。这一编码所用的词表在大规模预训练语料上得到, 因此缓解了词典外词和数据稀疏问题。为了更进一步增强模型性能, 该文提出了一种选择性的遮蔽语言建模训练策略——IDF 遮蔽。这一策略基于词在大规模预训练语料上统计的逆文档频率。实验表明, 与之前的中文预训练语言模型相比, 该文所提出方法预训练的模型在多个中文自然语言数据集上取得了更好或相当的性能, 并且能更高效地编码文本。

关键词: 中文预训练; 混合粒度编码; IDF 遮蔽

中图分类号: TP391

文献标识码: A

Exploring Chinese Pre-Training with Mixed-Grained Encoding and IDF-Masking

SHAO Yunfan, SUN Tianxiang, QIU Xipeng

(School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: Currently, most Chinese pre-trained language models adopt character-level encoding, which has a large computational overhead for long sequences. Although word-level encoding can alleviate this issue, it also brings some other issues such as out-of-vocabulary words and data sparsity. In this paper, we improve Chinese pre-trained language models with mixed-grained tokenization. The vocabulary of our encoding is obtained from large-scale corpora and thereby can alleviate the issues of out-of-vocabulary and data sparsity. To further improve the pre-training efficiency, we introduce a selectively masked language modeling method: IDF-masking, based on the inverse document frequency (IDF) collected on the pre-training corpora. The extensive experiments show that, compared with previous Chinese pre-trained language models, the proposed model can achieve better or comparable performance on various Chinese natural language processing tasks, and encode text more efficiently.

Keywords: Chinese pre-training; mixed-grained encoding; IDF-masking

0 引言

不同于英语或其他能轻易区分字词的语言, 中文句子中不包含显式的分词信息。在自然语言处理任务中, 通常用两种粒度来编码和处理中文文本, 词级别和字级别^[1-3]。然而, 如今大部分的中文预训练语言模型都采用字级别的细粒度编码方式。这会让句子被编码为很长的序列, 增加模型的计算开销。而计算开销的增加, 在预训练语言模型中更加明显。原因是预训练语言模型大都以 Transformer 作为主

要模型架构。而 Transformer 的时间和空间复杂度都是 $O(L^2)$, 其中 L 是编码句子的长度。另一方面, 词级别的粗粒度编码方式一般需结合性能良好的中文分词器。当采用词级别编码的预训练语言模型被用于不同的数据集时, 需要考虑所结合的中分词器的领域迁移问题。当数据分布在模型应用领域和分词器的训练领域中的差异过大时, 分词器的性能会有所下降, 影响预训练语言模型的编码效果。更重要的是, 词级别的粗粒度编码存在词典外词、数据稀疏以及错误传播等问题^[4-6]。这些问题进一步阻碍了粗粒度编码方式在预训练语言模型中的

收稿日期: 2022-01-25 定稿日期: 2022-04-14

基金项目: 国家自然科学基金(62022027)

直接使用。

尽管词级别编码方式有着众多不足,近年来,有许多工作显示词级别的粗粒度信息能够增强采用字级别编码的中文预训练语言模型。一种融合粗粒度信息的方式是将词级别的编码作为额外的特征输入模型^[7-9]。这一方式往往需要引入额外的参数并增加模型的计算和存储开销。另一种融合方式是利用词级别的信息改进预训练任务。一些工作尝试改进预训练中的遮蔽掩码策略,将随机遮蔽改为遮蔽句子中的粗粒度信息,如遮蔽词、短语或者实体^[1,10-11]。改进后的预训练任务能增强模型的性能,但是训练后的模型依旧用细粒度的字来编码文本。此外,在预训练时,粗粒度遮蔽的方法需要识别句子中的词、短语或实体,这通常需要一个高质量词典。在大规模预训练语料中,通常难以得到这样的词典,而用现有的词典可能会产生领域不匹配问题。

为了结合字级别和词级别编码的优点,探索高效的中文预训练模型编码粒度和编码方式。本文做出了以下改进:

首先,探索了基于混合粒度编码的中文预训练模型的有效性。在这一编码中,单字、词语和短语可以同时出现,共同编码文本。混合粒度编码的词表通过数据驱动算法直接在预训练数据中得到^[12]。相比之前的中文预训练语言模型,我们的模型编码效率更高,同时也避免了词表外词(OOV)问题,缓解了预定义词表与训练数据的领域不匹配问题。

另外,由于混合粒度的编码扩大了模型词表,可能需要用更有效率的方式预训练语言模型。为了加速预训练过程,我们提出了 IDF 遮蔽,一种基于词语逆向文档频率(IDF)信息^[13]的遮蔽策略。词语的 IDF 信息直接从预训练语料中统计得到,以缓解数据分布不一致问题。在预训练时优先遮蔽高 IDF 的词语并让模型预测。高 IDF 的词语相比随机挑选的词出现概率更低,更加难以被预测。因此,这一预训练任务能鼓励模型捕捉更充分的上下文信息,从而加速模型的收敛。

我们在多个中文自然语言处理任务中进行了实验,证明了本文方法的有效性。在 CLUE 基准数据集上超越了 BERT、RoBERTa 等预训练模型。

1 方法

1.1 混合粒度编码

相比直接使用一个训练好的中文分词器,我们

采用一元语言模型(Unigram Language Model)来获取混合粒度编码词表。更具体地,我们采用一个基于一元语言模型的无监督词表构建算法 SentencePiece^[14],在中文预训练语料上得到混合粒度的词表和每个词的出现概率。这一算法的使用让我们不用对语料进行分词处理,因此避免了因引入分词器而带来的数据分布不一致。

具体而言,我们基于句子中每个元素的出现概率都相互独立这一假设,使用一元语言模型建模语料中的句子。对于任意句子 x ,可以看成是词表中元素词元的(token)的序列 $x = (x_1, x_2, \dots, x_N)$,因此,句子的出现概率可以表示为:

$$P(x) = \prod_{i=1}^N p(x_i) \quad (1)$$

其中, N 是句子中词元个数, $p(x_i)$ 是词表中词元 x_i 的出现概率,服从条件:

$$\sum_{x_i \in V} p(x_i) = 1 \quad (2)$$

其中, V 是预先给定的固定大小的词表。词表的构建对于词信息的利用至关重要。多粒度中文预训练模型一般通过在语料上统计词频得到高频词以构建词表^[7-9, 15]。这些词表中的词一般质量较高,有清晰的词边界。然而,这些词表并未针对编码效率进行优化,若直接用于混合粒度编码,可能会出现编码后序列过长、词表过大、词表利用率低等问题。与词频统计方法不同的是,我们采用 Kudo 等^[12]提出的算法,在大规模语料上迭代式地优化词表在整个数据集上的似然 L ,从而使得到的词表更适用于混合粒度编码。具体而言,基于一元语言模型,计算整个数据集的似然 L ,并进行优化,如式(3)所示。

$$L = \sum_{k=1}^{|D|} \log \left(\sum_{x \in S(X_k)} P(x) \right) \quad (3)$$

其中, $|D|$ 为整个语料的句子数, $S(X)$ 是句子 X 的分词候选集合, $P(x)$ 的定义如式(1)。直接优化似然 L 计算量过大,因此这里使用 EM 算法并进行迭代优化。首先,算法根据语料统计得到一个大的初始词表,然后循环迭代式地逐步删除词表中的词,针对数据集似然 L 进行优化,直到词表大小符合预定义的阈值。具体的循环删词步骤为:

(1) 使用 EM 算法计算词表中每个词的概率 $p(x_i)$,使数据集似然 L 最大。

(2) 针对每个词 x_i 计算 loss_i ,表示将词 x_i 移除后似然 L 的减小量。

(3) 保留 loss_i 较高的前 $\mu\%$ (一般取 80%) 的词。

通过在大规模中文语料上使用该算法,我们得到的词表 V 包含了字、词语和短语。将该词表和中文 BERT 的词表进行对比,并重点比较了词表大小、中文字符所占比例以及编码效率。其中,编码效率我们使用编码后文本与原文的长度比作参考,所用文本来自于预训练语料中随机抽取的 5 000 条句子。结果如表 1 所示,其中, $|V|$ 表示词表大小, N_c 表示词表内中文词元的数量, $\frac{L_{enc}}{L_{raw}}$ 表示编码后文本与原文本的长度比。相比 BERT,我们的模型可以将句子编码长度大幅缩短,并且包含了更丰富的多粒度中文信息。

表 1 不同词表的对比

词表	$ V $	N_c	$\frac{L_{enc}}{L_{raw}}$
BERT	20k	7k	0.969
Ours	60k	58k	0.565

有了混合粒度的编码词表,就可以将任意句子编码为混合粒度的词元序列。具体而言,对于一段输入文本 X ,我们找到一串词元序列 x ,使得总概率最大,如式(4)所示。

$$\begin{aligned} x^* &= \operatorname{argmax}_{x \in S(X)} P(x) \\ &= \operatorname{argmax}_{x \in S(X)} \prod_{x_i \in x} p(x_i) \end{aligned} \quad (4)$$

其中, $S(X)$ 是句子 X 的分词候选集合。集合中的每一段分词序列都由词表中的词元组成,并且词表中包含了大量细粒度的词元,即汉字和字母。因此,在词表中不存在的词,会被拆分成字序列加以表示。这样我们的分词算法就可以避免词表外词(OOV)的问题。我们采用维特比算法^[16]来得到概率最高的词元序列,避免遍历 $S(X)$ 中的所有情况,减少了分词时所需的开销。

1.2 基于 IDF 的遮蔽策略

我们观察到,混合粒度编码增加了词表大小,这可能使得预训练语言模型需要训练更多的步数才能达到原有的效果。一方面,这会引入更多参数来表示词表中的词元向量。另一方面,基于齐夫定律,词表中的词元在数据集中呈现长尾分布,存在极少的高频词元和大量的低频词元^[17]。这样高度偏斜的词频分布会导致数据稀疏的问题,即大量的词向量很少出现,也很难被训练。

为了更加系统地了解这一问题,我们在预训练

语料上进行分词,并计算了每个词的词频和逆文档频率(IDF)^[13]。对于一个词 w ,它的逆文档频率可以表示为:

$$\text{IDF}(w) = -\log P_D(w) = \log \frac{N}{N_w} \quad (5)$$

其中, $P_D(w)$ 是词 w 在文档中的出现概率, N_w 是包含词 w 的文档数, N 是语料中的总文档数。

正如图 1 中的结果所示,词频分布有着很大的方差,分布范围从 10^1 到 10^8 ,并且大部分的词都只有很低的词频(小于 10^3),显然这会导致数据稀疏问题。

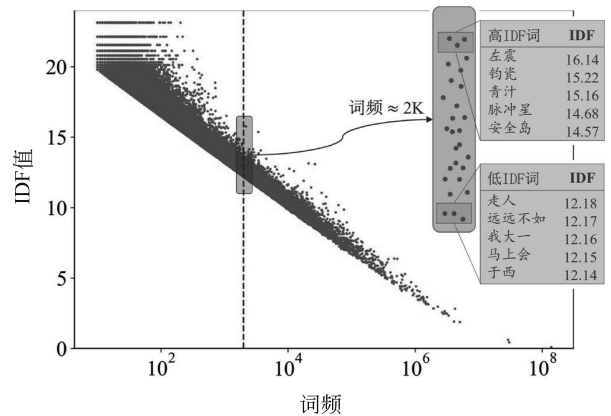


图 1 预训练语料中词的词频和逆文档频率(IDF)的分布图

为了缓解这一问题,一个简单直接的思路是在预训练时增加低频词的被遮蔽和被预测概率。然而,根据图 1,拥有相似词频的词语可能有着差异很大的数据分布,而这一分布可以利用 IDF 来描述。如在图 1 中,我们展示了词频约为 2 000 时,不同词语的 IDF 值。其中高 IDF 的词可能是实体或名词,而带有低 IDF 值的词含有的语义信息较少。这说明 IDF 值相比词频而言,在筛选低频词时更加可靠。

因此,我们提出 IDF 遮蔽策略来加速预训练过程,并增强模型的表达能力。具体的处理流程如图 2 所示,我们首先在预训练语料中统计得到大规模词表以及对应的 IDF 值。在预训练时,我们首先利用现有分词器将句子分为由 n 个词组成的序列,然后从词表中得到每个出现词的 IDF 值,进一步从中取 IDF 值最大的 k 个词作为候选集;最后,在候选集中随机采样 m 个词作为遮蔽词。由于 IDF 词表一般远大于混合粒度的编码词表,遮蔽词可能横跨多个混合粒度词元序列。为了减少分词器的错误传播,我们选择能覆盖遮蔽词的连续混合粒度词元序列进行遮蔽,即以词粒度选择遮蔽范围,以模型编码粒度进行遮蔽。因为 IDF 词表的数据从预训练

语料中得到,缓解了数据分布不一致的问题。另外, IDF 遮蔽还增加了预训练任务的难度。相比随机遮蔽策略和简单的词级别遮蔽策略, IDF 遮蔽策略让模型更加难以利用局部信息预测遮蔽位置的词,进而使得模型建模更丰富的、跨度更大的上下文信息,从而增强了模型的表达能力^[11]。



图2 混合粒度编码的模型
使用 IDF 遮蔽训练时的数据处理流程

2 实验设置

2.1 预训练设置

为了公平对比,我们使用大规模开源数据进行预训练,处理后数据集大小为 13 GB,包含九百万个文档。对于混合粒度编码的词表,设置词表大小为 6 万,用 SentencePiece 在预训练语料中随机采样的 2 000 万个句子进行一元语言模型的训练。对于 IDF 遮蔽需要的低频词表和 IDF 值信息,我们首先用现有的中文分词器,将预训练语料分词,然后统计词频和 IDF 值,略去词频小于 10 的词。最后得到的 IDF 词表包含了大约 100 万个词。我们采用 BERT-base 架构,即 12 层的双向 Transformer 编码器,768 维隐向量,前馈网络中间层维度为 3 072,12 个注意力头,每个注意力向量的维度为 64。我们从头开始预训练,以公平对比,并使用 Adam 优化器训练模型 100 万步,最大句子长度为 512,一个批次有 384 个句子。我们利用学习率先热身(Warmup)再衰减策略,学习率从零开始线性增加 1 万步,达到峰值 10^{-4} ,然后线性衰减到零。遵循 RoBERTa 的超参数设置,我们将 Adam 的 beta 设为 (0.9, 0.98),参数衰减(Weight Decay)设为 0.01。在预训练时,我们将遮蔽比例设为 15%,并采用 IDF 遮蔽策略。

其中, k 为句子总词数的 30%, m 为句子总词数的 15%。如当句子中包含 512 个词时,候选集大小 $k=154$,遮蔽词数量 $m=77$ 。整个预训练在 8 块英伟达 RTX3090 显卡(显存为 24 GB)上完成,需要 7 天左右的时间。

2.2 下游任务设置

2.2.1 数据集

为了测试不同领域以及不同句长分布下模型的性能,我们选用以下数据集进行实验:

CLUE 中文语言理解基准分类数据集^[18]: 包含了 6 个数据集,语义相似度匹配数据集 AFQMC,文本分类数据集 IFLYTEK 和 TNEWS,自然语言推理数据集 CMNLI,代词消歧数据集 WSC,以及论文关键词匹配数据集 CSL。与 Zhang 等相同,我们在 TNEWS 和 CSL 上使用了数据增强^[8]。此外,我们还在情感分析数据集 ChnSentiCorp^[19],篇章级文本分类 THUCNEWS^[20]句对语义匹配数据集 LCQMC^[21]上测试了我们的模型。

对于已有分词标注的任务(如词性标注,命名实体识别,基于 Span 的阅读理解等),在使用混合粒度编码时,直接使用传统方式微调会由于分词不一致问题产生性能损失。针对这些任务,可以采用以下方法:①改进解码方式,让模型预测词边界;②直接使用字粒度编码微调;③配合其他字粒度编码器同时使用;这些改进能缓解分词不一致问题,我们将混合粒度编码在这些任务中的扩展作为未来工作。

2.2.2 微调实验设置

对于基准模型,我们选择几个有代表性的主流中文预训练语言模型进行比较,以证明本文所提出的方法的有效性。为了与其他不同的编码方式进行比较,选取的模型有字级别、词级别以及多粒度融合字词的中文预训练语言模型:

BERT: 为了充分比较,我们选取几种不同的 BERT 实现:①由谷歌发布的中文版 BERT 模型^[22],在中文 wiki 上使用随机遮蔽训练,在本文中用 BERT(Google)表示;②由 Cui 等发布的基于整词遮蔽训练的 BERT-wwm^[2],在本文中简称为 BERT;③由 Zhang 和 Li 得到的词级别 BERT^[8],简称 BERT(word)。

RoBERTa: 我们使用 Cui 等发布的 RoBERTa-wwm-ext 进行实验^[2]。这一模型使用谷歌 BERT 初始化,在大规模语料上训练 100 万步得到。

AMBERT: 由 Zhang 等提出的多粒度编码模

型^[8],分别使用两个参数共享的编码器处理细粒度和粗粒度分词的文本。

LICHEE: 通过改进词嵌入层融合文本的多粒度信息的预训练方法^[15]。作者将这一方法在不同结构的预训练模型上进行了应用。我们选用 LICHEE-BERT 进行对比。

Lattice-BERT: 通过改进注意力机制以编码 Lattice 结构的多粒度中文预训练模型^[9]。

我们对所有模型都使用统一的微调流程,针对每个数据集,仅使用该数据集的训练集进行微调。在每个数据集上,都进行有限的网格搜索。其中,设置批量大小为{16,32},学习率为{1,2,3,5} $\times 10^{-5}$,训练周期在除了 WSC 的数据集上取{5,8},在 WSC 上固定为 80 个周期。对于 CLUE 分类数据集,

本文取开发集正确率最高的模型在测试集上进行预测,并汇报测试集的正确率。对于其他数据集,我们运行 5 次相同设置的实验,并同时汇报最佳和平均正确率。

3 实验结果与分析

3.1 主实验结果

我们汇报了基准模型以及我们的模型分别在 CLUE 分类数据测试集上的性能,结果列在表 2 中。对于其他数据集,我们的结果列在表 3 中,其中,“†”表示本文根据原论文复现后的结果,括号中的数字为平均值,括号前的数字为最高值。

表 2 不同模型在 CLUE 基准测试中测试集上的正确率

(单位: %)

模型	参数量	TNEWS	IFLYTEK	AFQMC	CMNLI	WSC	CSL	Avg.
BERT(Google)	108M	66.99	60.29	73.70	79.69	71.03	83.50	72.53
BERT	108M	67.14	59.43	74.07	80.42	74.48	84.20	73.29
RoBERTa	108M	67.63	60.31	74.04	80.51	76.90	84.70	74.02
AMBERT	176M	68.58	59.73	73.87	81.87	78.28	85.70	74.67
LICHEE-BERT	270M	67.94	60.94	73.65	—	81.03	84.51	—
Lattice-BERT	100M	68.38	62.85	74.77	81.13	82.41	84.03	75.60
BERT(word)	165M	68.20	59.96	73.48	79.97	75.52	85.17	73.72
Ours	132M	69.01	61.58	73.50	79.97	78.28	84.97	74.82

表 3 不同模型在三个数据集上的正确率对比

(单位: %)

模型	ChnSentiCorp	THUCNews	LCQMC
BERT	95.4(95.0)	97.8(97.6)	87.0(86.8)
RoBERTa	95.6(94.9)	97.8(97.5)	86.4(86.1)
Lattice-BERT [†]	95.6(95.3)	97.8(97.6)	87.0(86.5)
Ours	96.2(96.0)	98.0(97.9)	87.7(87.5)

如表 2 和表 3 所示,混合粒度编码相比单一粒度的中文预训练模型,如 BERT 和 RoBERTa 等,有着明显的性能优势。如表 2 所示,而对比其他多粒度编码的预训练模型,如 AMBERT、LICHEE-BERT 和 Lattice-BERT 等,也有很强的竞争力。同时,混合粒度编码减短了输入长度,提高了模型推理效率。

3.2 混合粒度编码的有效性

混合粒度编码的一大优势是相比细粒度的编码,在保持模型的表达能力不变的同时,能够缩短编

码序列的长度,并能减少模型的计算量。我们随机从数据集中抽取不同句子长度的文本,并计算这些文本用不同编码方式所需要的模型计算量,即 FLOPs。结果如图 3 所示,混合粒度编码有效减少了计算开销,在长文本上效果尤为明显。

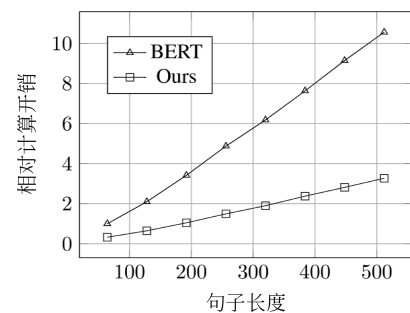


图 3 字级别编码方式(BERT)和混合粒度编码(Ours)在不同长度下的相对计算开销比较

我们进一步在不同数据集上对比不同编码方式的平均编码长度,并分别记录推理时间,结果如表 4 所示,相比字粒度编码,混合粒度编码缩小了文本的

平均编码长度,并加快了推理速度。

表 4 不同编码方式在不同数据集上的平均编码长度和推理时间比较

数据集	平均编码长度		推理时间/s	
	BERT	Ours	BERT	Ours
TNEWS	39.3	24.4	19.8	15.2
IFLYTEK	276.5	164.0	16.2	15.5
AFQMC	28.6	16.5	9.0	5.7

表 5 不同编码方式和遮蔽方式在 CLUE 基准测试中验证集上的正确率(除特殊标明模型使用整词遮蔽训练)

(单位: %)

模型	参数量/M	TNEWS	IFLYTEK	AFQMC	CMNLI	WSC	CSL	Avg.
BERT	108	67.14	59.45	73.77	80.92	80.95	84.03	74.38
RoBERTa	108	67.42	60.80	74.30	80.70	85.50	85.70	75.74
OurBERT(char)	108	67.10	59.40	73.82	80.82	76.97	81.63	73.81
+IDF 遮蔽	108	68.11	60.95	73.73	80.93	80.82	84.31	74.94
Our BERT(mixed)	132	68.08	60.60	73.66	80.58	81.91	84.97	74.97
+IDF 遮蔽	132	68.57	62.83	74.14	80.61	85.85	84.77	76.13

3.3 IDF 遮蔽的有效性

为了更进一步地证明 IDF 遮蔽策略的有效性,我们分别使用 IDF 遮蔽策略以及整词遮蔽策略预训练了语言模型,最终结果见表 5。配合混合粒度编码,我们的模型最终能在 CLUE 验证集上取得 76.13% 的平均性能。而单独在字级别预训练模型上使用 IDF 遮蔽,增强了预训练任务的难度,从而也增强了模型的性能(表 5)。

我们选取出优化了不同步数的模型在 TNEWS 数据集上进行微调。结果如图 4 所示,当训练步数相同时,使用了 IDF 遮蔽策略的模型的下游任务性能比整词遮蔽策略优化的模型要好,并且随着训练步数的增加 IDF 遮蔽策略将持续地提升模型的性能,在训练 1M 步后依然保持了性能增长趋势,并且大幅超过了整词遮蔽策略训练的模型。

为了更细致地比较 IDF 遮蔽策略和整词遮蔽策略的不同,我们随机采样了约 2 000 万个词的预训练语料(约 200 MB),分别使用整词遮蔽和 IDF 遮蔽选择遮蔽词,并将被遮蔽的词数按词频进行统计,结果如图 5 所示。相比整词遮蔽,IDF 遮蔽倾向于选择词频在 $10^2 \sim 10^6$ 的中低频词进行掩码,而显著减少高频词(词频高于 10^8)的掩码次数。如图 5

为了进一步比较不同的编码方式,我们去除了 IDF 遮蔽,使用和 BERT 相同的整词遮蔽进行训练和比较。如表 5 所示,我们的模型“Our BERT(mixed)”就编码效率而言优于同等数据量训练的 BERT 模型。为了公平比较,用同样的数据集和超参复现了 BERT,结果见“Our BERT(char)”。可以看到 IDF 遮蔽相比整词遮蔽能同时为字级别和混合粒度编码的预训练模型带来性能提升。当 IDF 遮蔽配合混合粒度编码时,所带来的提升更加明显。

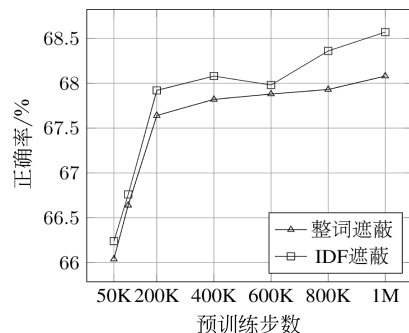


图 4 使用不同遮蔽方式训练的模型在 TNEWS 验证集上的性能随预训练步数的变化情况

所示,IDF 遮蔽改变了不同词频的词被遮蔽的概率,而非完全忽略高频词。因为 IDF 遮蔽仅在句内进行 IDF 排序和选择,而语料中存在很多仅由高频词组成的句子,这时 IDF 遮蔽从句中选择的遮蔽词(其在句内排序 IDF 很高)在整个语料上 IDF 很低,属于高频词。

4 相关工作

4.1 多种粒度编码的中文预训练模型

现存的中文预训练语言模型通常将句子以字序列的方式编码。然而,这会使得长句子难以被表示,

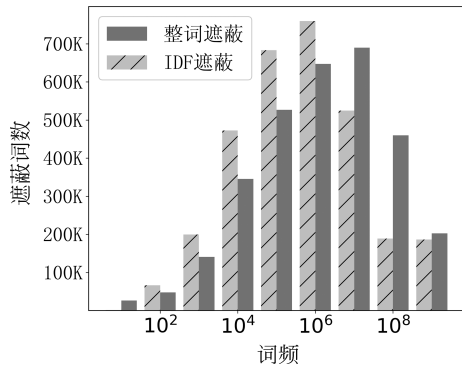


图5 不同遮蔽方式中被遮掩词在语料中的词频分布

会产生很大的计算开销。同时,这也忽略了中文丰富的粗粒度信息。为了融合粗粒度信息到中文预训练语言模型中,Zhang 等提出了使用两个参数共享的编码器分别建模粗粒度和细粒度的信息^[8]。而 Diao 等提出使用一个额外的编码器单独建模 N 元 (N -Gram)信息,并逐层融合到细粒度编码的预训练语言模型中^[7]。然而,这些额外引入的编码器会引入大量的参数,额外增加了计算和存储开销。Li 等将文本进行细粒度编码的同时,加入了 Lattice 信息作为模型的输入,并提出了表示 Lattice 结构的相对位置编码,使得模型能同时建模多种粒度的文本^[9]。然而,Li 等的模型结构使得输入序列变长,相对位置编码也存在较大的计算开销,大大增加了模型的计算量。相比之前的工作,我们提出的混合粒度编码能将句子表示为较短的序列,增加了模型的编码效率。

4.2 词信息增强的遮蔽语言建模

相比仅使用随机遮蔽的预训练任务,有一系列工作显示,遮蔽文本中的词、短语和实体进行训练,将会增强预训练模型的性能^[1, 9-11]。例如,谷歌在 BERT 的基础上,使用整词遮蔽训练了 BERT-wwm。Joshi 等人提出 SpanBERT,随机遮蔽连续的词元序列,从而进一步增加任务难度^[10-12]。ERNIE 为了融合知识到模型中,提出遮蔽句子中的实体。针对融合了 Lattice 信息的模型,Li 等提出先将文本分成几个互相没有 Lattice 信息连边的几个分割,然后以分割为单元遮蔽^[9]。这样的模型无法直接利用相连的 Lattice 信息进行预测,增加了预测难度。ERNIE-Gram 提出在训练细粒度编码的语言模型时,可以同时让模型预测文本中的粗粒度信息,从而增强模型的性能^[22]。这样模型仍然建模细粒度分词的文本,在预训练时使用整词遮蔽后,模

型需同时预测被遮蔽词以及词的细粒度信息(词元序列)。然而,这些方法在遮蔽前往往需要得到文本中的词语或实体等粗粒度信息。这时一般需要一个预先给定的大规模词库或实体库。因此,预训练语料和词库之间可能会出现数据分布不一致的问题。并且,这些方法在遮蔽时没有考虑到数据稀疏的问题,低频词的出现相对较少且比较稀疏。相比之前的方法,IDF 遮蔽策略使用了在预训练语料上得到的统计信息,并且让模型重点学习文本中的稀有词,从而缓解了数据不一致和数据稀疏的问题,更进一步地增强了模型。

5 总结与展望

本文探索了中文预训练中应用混合粒度编码的可能性。相比之前中文预训练模型广泛采用的细粒度编码,我们提出的混合粒度编码有效提升了编码效率。而相比词级别等粗粒度的编码方式,本文的方法有效缓解了词表外词(OOV)以及数据稀疏等问题。搭配上所提出的 IDF 遮蔽策略,我们预训练的混合粒度编码语言模型在多个数据集上取得了有竞争力的性能,并且减少了大量模型计算量。

未来工作包括混合粒度编码应用范围的拓展以及预训练遮蔽策略的进一步改进。混合粒度编码能够明显提升编码器的编码效率,但对于需要细粒度分词的下游任务不够友好。未来可以将混合粒度编码器与细粒度的解码器相结合,进一步扩展模型的应用深度和广度。另一方面,IDF 遮蔽策略根据词频信息调节了不同词汇的遮蔽频率,未来可以基于更丰富的数据集和其他先验信息更系统地选择遮蔽词和遮蔽方式,以加快模型的预训练过程并提升训练效果。

参考文献

- [1] SUN Y, WANG S, LI Y K, et al. ERNIE: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv: 1904.09223, 2019.
- [2] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. arXiv preprint arXiv: 1906.08101, 2019.
- [3] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C]// Proceedings of EMNLP, 2020: 657-668.
- [4] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]// Proceedings of ACL, 2018: 1554-1564.

- [5] LI X, YAN H, QIU X, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of ACL, 2020: 6836-6842.
- [6] LI X, MENG Y, SUN X, et al. Is word segmentation necessary for deep learning of Chinese representations? [C]//Proceedings of ACL, 2019: 3242-3252.
- [7] DIAO S, BAI J, SONG Y, et al. ZEN: Pre-training Chinese text encoder enhanced by n -gram representations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing: Findings, 2020: 4729-4740.
- [8] ZHANG X, LI H. AMBERT: Apre-trained language model with multi-grained tokenization[J]. arXiv pre-print arXiv: 2008.11869, 2020.
- [9] LAI Y, LIU Y, FENG Y, et al. Lattice-BERT: Leveraging multi-granularity representations in Chinese pre-trained language models [C]//Proceedings of NAACL-HLT, 2021: 1716-1731.
- [10] JOSHI M, CHEN D, LIU Y, et al. SpanBERT: Improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.
- [11] LEVINE Y, LENZ B, LIEBER O, et al. PMI-masking: Principled masking of correlated spans[C]//Proceedings of 9th International Conference on Learning Representations, 2021.
- [12] KUDO T. Subword regularization: Improving neural network translation models with multiple subword candidates[C]//Proceedings of ACL, 2018: 66-75.
- [13] JONES K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, (28)1: 11-21.
- [14] KUDO T, RICHARDSON J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing[C]//Proceedings EMNLP: System Demonstrations, 2018: 66-71.
- [15] GUO W, ZHAO M, ZHANG L, et al. LICHEE: Improving language model pre-training with multi-grained tokenization[C]//Proceedings of the Association for Computational Linguistics, Online Event, 2021: 1383-1392.
- [16] VITERBI A J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm [J]. IEEE Transactions on Information Theory, 1967, 13(2): 260-269.
- [17] ZIPF G K. Human behavior and the principle of least effort[M]. Ravenio Books, 1949.
- [18] XU L, HU H, ZHANG X, et al. CLUE: A Chinese language understanding evaluation benchmark NLP Chinese corpus: Large scale Chinese corpus for NLP [C]//Proceedings of COLING, 2020: 4762-4772.
- [19] TAN S, ZHANG J. An empirical study of sentiment analysis for Chinese documents[J]. Expert Syst. Appl. 2008, 34(4): 2622-2629.
- [20] LI J, SUN M. Scalable term selection for text categorization [C]//Proceedings of EMNLP, 2007: 774-782.
- [21] LIU X, CHEN Q, DENG C, et al. LCQMC: A large-scale Chinese question matching corpus [C]//Proceedings of COLING, 2018: 1952-1962.
- [22] XIAO D, LI Y K, ZHANG H, et al. ERNIE-gram: Pre-training with explicitly n -gram masked language modeling for natural language understanding [C]//Proceedings of NAACL-HLT, 2021: 1702-1715.
- [23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of Annual Conference on Neural Information Processing Systems, 2017: 5998-6008.
- [24] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of NAACL-HLT, 2019: 4171-4186.



邵云帆(1997—), 硕士研究生, 主要研究领域为自然语言处理。
E-mail: yfshao19@fudan.edu.cn



孙天祥(1997—), 博士研究生, 主要研究领域为自然语言处理。
E-mail: txsun19@fudan.edu.cn



邱锡鹏(1983—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理。
E-mail: xpqiu@fudan.edu.cn